Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

DISSERTATION

Modeling novel bioinformatics approaches to investigate bioactive substance production based on genomics and transcriptomics

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
unter der Leitung von

Univ.Prof. Mag. Dr.rer.nat. Robert Mach
Mag.rer.nat. Dr.rer.nat. Christian Derntl

E166
Institut für Verfahrenstechnik, Umwelttechnik und Technische Biowissenschaften

Eingereicht an der
Technischen Universität Wien
Fakultät für Technische Chemie

von

Mag.pharm. Gabriel Alexander Vignolle
00738272

Wien, am 17.12.2021

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table of contents

Kurzfassung ……………………………………………………………………………3

Abstract ……………………………………………………………………………...5

Introduction ……………………………………………………………………………7

Aims ……………………………………………………………………………..11

Conclusions ……………………………………………………………………………...12

References for Introduction ……………………………………………………..13

Peer reviewed publication:
"Novel approach in whole genome mining and transcriptome analysis reveal conserved RiPPs in *Trichoderma* spp."
Published in *BMC Genomics* ……………………………………………………16

Peer reviewed publication:
"FunOrder: A robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution."
Published in *PLOS Computational Biology* ……………………………………..53

Peer reviewed publication:
"FunOrder 2.0 – a fully automated method for the identification of co-evolved genes."
Submitted to *PLOS Computational Biology* ……………………………………...110

Peer reviewed publication:
"Genome Sequence of the Black Yeast-Like Strain *Aureobasidium pullulans* var. *aubasidani* CBS 100524."
Published in *Microbiology Resource Announcement* ……………………………131

Peer reviewed publication:
"Genome sequencing of *Wardomyces moseri*: a rare but cosmopolitan fungus with an outstanding secondary metabolite production potential."
Submitted to *BMC Genomics* ……………………………………………………..133

Appendix ……………………………………………………………………………200

Acknowledgements ……………………………………………………………………235

Curriculum Vitae ……………………………………………………………………236

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Genome-Mining- und Bioinformatik-Technologien sind in der heutigen Zeit für die Suche nach neuartigen Sekundärmetaboliten (SM) unverzichtbar geworden. SM sind eine große Gruppe von Verbindungen mit unterschiedlichen Strukturen und Eigenschaften. Sie werden meist von Enzymen produziert, deren entsprechende Gene im Genom kolokalisiert sind und in biosynthetischen Genclustern (BGC) organisiert sind. Die Identifizierung und Suche von BGC ist ein Schlüsselaspekt der Naturstoffbioinformatik geworden. Darüber hinaus ist die Entdeckung neuer SM-Klassen in den Genomen von Pilzen, sogenannter „Dark Matter"-BGC, ein Gegenstand derzeitiger Forschung. In dieser Dissertation wurden verschiedene Themen mit dem Ziel behandelt, den Nachweis und die Analyse exotischer Biosynthesewege von SM zu erleichtern. Diese verschiedenen Themen besitzen als gemeinsamen roten Faden die Suche und Beschreibung von SM-BGC.

Diese Doktorarbeit umfasst mehrere veröffentlichte und eingereichte Studien, die in einer angemessenen Reihenfolge thematisch geordnet sind. Das erste angesprochene Thema ist die Identifizierung neuer BGC in Pilzen. Zu diesem Zweck wurde eine neue Methode zum Analysieren von Pilzgenomen eingeführt, diese detektiert ribosomal synthetisierte und posttranslational modifizierte Peptide (RiPPs) durch Kombination und Anpassung vorhandener Werkzeuge, gefolgt von einer umfangreichen manuellen Kurierung basierend auf der Identifizierung konservierter Domänen, (vergleichende) phylogenetische Analysen und durch die Anwendung von RNASeq-Daten. RiPPs sind eine sehr vielfältige Gruppe von SM und wurden vor kurzem in Pilzgenomen eingehend untersucht. Gene, die an der Biosynthese von RiPPs in Pilzen beteiligt sind, wie für viele andere SM, sind in BGC gepackt. Die vorliegende Veröffentlichung ist der erste Bericht über das Potenzial der Pilzgattung *Trichoderma* zur Produktion von RiPPs. Erwähnenswert ist, dass die mit dieser neuartigen Methode entdeckten Cluster, Gene beinhalten die Enzyme kodieren für den Biosyntheseweg für neuartige uncharakterisierte Pilz-RiPPs.

Neben dem Aspekt, nach neuartigen BGCs zu suchen, war die eingehende Analyse der gefundenen BGC ein Ziel. BGC können sogenannte Gap-Gene enthalten, die nicht an der Biosynthese des SM beteiligt sind. Gap-Gene von Genen zu unterscheiden, die an der Biosynthese beteiligt sind, ist eine langwierige, teure und mühsame Aufgabe. Diesem Thema widmeten sich zwei Studien, von denen die erste das Functional Order Tool (FunOrder) als halbautomatische

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Methode zur Identifizierung koevolutionär verknüpfter Gene in BGC vorstellte. Die Ergebnisse legen nahe, dass die Koevolution von Proteinfamilien für die Differenzierung von Gap-Genen von biosynthetisch aktiven Genen genutzt werden kann. In der anschließenden Studie wird das verbesserte und vollautomatisierte FundOrder 2 vorgestellt, bei dem frühere Einschränkungen durch die Einführung einer vollautomatisierten und verbesserten Bestimmung von koevolvierten Genen behoben wurden. Der automatisierte Nachweis koevolvierender Gene verwendet mehrere mathematische Indizes, um die optimale Anzahl von Gengruppen in den FunOrder-Daten zu bestimmen und die Implementierung von k-Means-Clustering basierend auf den ersten drei Hauptkomponenten (PC) einer Hauptkomponentenanalyse (PCA) bestimmt diese. FunOrder 2 kann als wesentliche Verbesserung gegenüber seinem Vorgänger angesehen werden, insbesondere durch die automatisierte Analyse ohne Bias und die Anpassung an größere Datenbanken.

Im weiterer Folge wird Sequenzierung, Assemblierung und Analyse neuartiger uncharakterisierter Pilzarten thematisiert, mit dem Hauptfokus auf die Suche und Analyse ihres SM-Produktionspotenzials. Vier Genome wurden sequenziert und in zwei Studien präsentiert, die das letzte Thema dieser Arbeit behandeln. Zunächst wird die Genomsequenz des schwarzen hefeähnlichen Pilz *Aureobasidium pullulans* var. *aubasidani* CBS 100524, mit industrieller Relevanz durch ausgeschiedene extrazelluläre Polysaccharide, vorgestellt und kurz beschrieben. Darauf folgt eine Studie, die eine eingehende vergleichende Genomanalyse und die phylogenetische Reklassifizierung von drei sequenzierten *Wardomyces moseri* Stämmen durchführt. W. Gams beschrieb den Ascomyceten *W. moseri* erstmals 1995. Während einer phylogenetischen Studie im Jahr 2016 wurde *W. moseri* als phylogenetisch fehlplaziert beschrieben und sollte daher neu bewertet werden. Das metabolische Potenzial dieses historischen Pilzes wurde analysiert und seine Taxonomie neu bewertet, indem die Genome des Ex-Isotyp-Stamms *W. moseri* CBS 164.80 und zwei Isolate von der anderen Seite der Welt, *W. moseri* TUCIM 5827 und TUCIM 5799, sequenziert wurden. Es konnte gezeigt werden, wie historische Stämme aus bereits bestehenden Stamm-Sammlungen für die Suche nach neuartigen Naturstoffen benutzt werden können.

Im Anhang aufgeführt sind abschließend interdisziplinäre Studien, die aus Kooperationen mit verschiedenen Arbeitsgruppen hervorgegangen sind.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Abstract

Genome mining and bioinformatics technologies have become essential to the discovery process of novel secondary metabolites (SMs). SMs are a vast group of compounds with different structures and properties. Enzymes whose corresponding genes are co-localized in the genome, organized in biosynthetic gene clusters (BGCs), readily produce them. The identification and search of BGCs is a key aspect of natural product bioinformatics. Further, the detection of novel SM classes in the genomes of fungi, so termed “dark-matter” BGCs, is an ongoing subject of research. In this thesis, various topics were addressed for the ultimate goal to facilitate the detection and analysis of exotic biosynthetic pathways of SMs. These different subjects are connected by the search for and description of SM BGCs.

This thesis encloses several published and submitted studies and orders them thematically. The first issue addressed is the identification of novel BGCs in fungi, a novel method to mine fungal genomes for ribosomally synthesized and post-translationally modified peptides (RiPPs) by combining and adapting existing tools followed by extensive manual curation based on conserved domain identification, (comparative) phylogenetic analysis, and RNASeq data was introduced for this purpose. RiPPs are a highly diverse group of SM and have been recently started to be studied in more depth in fungal genomes. Genes involved in the biosynthesis of fungal RiPPs, as for many other SMs, are packed in BGCs. The presented publication is the first report of the potential of the fungal genus *Trichoderma* to produce RiPPs and the clusters detected by this novel method encode genes that ultimately lead to novel uncharacterized fungal RiPPs.

Besides the aspect to search for novel BGCs, the in depth analysis of detected BGCs was a target. BGCs may contain so-called gap genes, which are not involved in the biosynthesis of the SM. To differentiate gap genes from genes involved in the biosynthesis is a lengthy, expensive and arduous task. This topic was addressed by two studies the first describing and introducing the Functional Order tool (FunOrder), as a semi-automated method for the identification of co-evolutionary linked genes in BGCs. The results suggest that protein family co-evolution can be leveraged for the differentiation of gap genes from genes involved in the biosynthesis of a SM. In the subsequent study, the improved and fully automated FunOrder 2 is presented, where previous limitations were address by introducing a fully automated and enhanced determination of co-evolved genes. The automated detection of co-evolving genes uses several mathematical indices

to determine the optimal number of gene groups in the FunOrder output and the implementation of k-means clustering based on the first three principal components (PC) of a principal component analysis (PCA) detects them. FunOrder 2 can be seen as a major improvement over its predecessor, especially considering the unbiased automated analysis and the adaptation to larger databases.

The last theme is the topic of sequencing, assembly and analysis of novel uncharacterized fungal species primarily for the search and analysis of their slumbering SM production potential. Four genomes have been sequenced included in two studies that address the final topic in this thesis. First, the genome sequence of the black yeast-like strain *Aureobasidium pullulans* var. *aubasidani* CBS 100524 with industrial relevance due to excreted extracellular polysaccharides is introduced and briefly described. This is followed by a study performing an in depth comparative genomic analysis and phylogenetic replacement of three sequenced *Wardomyces moseri* strains. W. Gams first described the ascomycete *W. moseri* in 1995. During a phylogenetic study in 2016 *W. moseri* was suggested to be phylogenetically misplaced and should therefore be re-evaluated. The metabolic potential of this historic fungus was analyzed and its taxonomy re-evaluated, by sequencing the genomes of the ex-isotype strain *W. moseri* CBS 164.80 and two isolates from the opposite side of the world, *W. moseri* TUCIM 5827 and TUCIM 5799. It could be demonstrated how historic strains from already existing collections can be used for the search of novel natural products.

Finally listed in the appendix, are interdisciplinary studies fruited from collaborations with different working groups.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Introduction

Secondary metabolites (SMs) are a diverse group of compounds with different chemical structures and properties which are found in all domains of life, but are predominantly studied in bacteria, fungi, and plants (1). SMs are not essential for the survival and growth of an organism but can be advantageous under particular environmental circumstances, for instance antibiotics under competitive conditions, pigments to tolerate radiation, and toxins as either defensive or virulence factors (2, 3). SMs can be grouped into different classes based on their biosynthetic pathways and chemical structures. In fungi, the two main classes are non-ribosomal peptides (e.g. the antibiotic penicillin (4) or the immunosuppressant cyclosporine (5)) and polyketides (e.g. the mycotoxin aflatoxin (6) or the cholesterol-lowering drug lovastatin (7)). Further SM classes are alkaloids, terpenes, melanins (8, 9), and ribosomally synthesized and post-translationally modified peptides (RiPPs) (10, 11). The genes encoding the enzymes responsible for the production of SMs are spatially organized in biosynthetic gene clusters (BGCs) in many cases (12, 13). SMs from fungal sources have been used for therapeutic purposes and to promote and preserve the human well-being already since ancient times (14-16). Fungal SMs and chemically modified variants are widely used as antibiotics, immunomodulators and anti-cancer drugs (17). The study of the secondary metabolism of fungi, especially from understudied strains and genera, holds the promise for much needed novel antibiotics, pharmaceuticals, and most recently also precursors for the synthesis of innovative plastics (18).

In the last decades, genome mining and bioinformatics have played a crucial role and became an essential tool in the discovery of novel natural products. Especially the detection and classification of BGCs has contributed to the ongoing unraveling of biochemical space. Several databases (e.g. minimum information about a biosynthetic gene cluster (MIBiG) repository (19, 20)) and tools have been developed for the analysis and detection of BGCs. Some developed software for the discovery of BGCs are antiSMASH (21-23), Cassis/CASSIS and SMIPS (24), SMURF (25), TOUCAN, a supervised learning framework capable of predicting BGCs on amino acid sequences (26), and DeepBGC, a unrestricted machine learning approach using deep neural networks (27). These programs can be used for the identification of BGCs in fungi, Cassis/CASSIS and SMIPS (24), and SMURF (25) have been developed especially for this purpose. The position as gold standard for BGC detection and definition is currently held by antiSMASH (23) in both

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

bacterial and fungal genomes. AntiSMASH uses a rule based approach for the definition of BGCs, it detects core biosynthetic enzymes and by applying a greedy-approach it includes surrounding genes into the newly defined BGC (23). This possibly will result in overlaps or combinations of closely situated clusters. Nevertheless, the genes within the predicted BGCs are defined as core biosynthetic genes, additional biosynthetic genes, transport-related genes, regulatory genes, and other genes based on profile hidden Markov models by the antiSMASH tool. As for other types of BGCs, fungal RiPP BGCs (28) are detected by antiSMASH with a rule based approach based primarily on the ustiloxin B cluster of *Aspergillus flavus* (23, 29). This restriction was addressed in this thesis by introducing a novel method for detecting the precursor peptides of RiPPs within fungal genomes. RiPPs are a rapidly increasing group of natural products that can be classified in several different compound classes [reviewed in (10, 11, 30)]. A more in depth description of the biosynthesis of RiPPs and the structure of RiPP BGCs in fungi can be found in the first chapter of this thesis.

A major limitation in the discovery of yet undescribed SMs is the fact that most BGCs are inactive under standard laboratory conditions, as they do not serve a purpose for the organisms then. Currently, different approaches are followed to circumvent this difficulty (31, 32). Untargeted approaches aim to induce the expression of any SM. To this end, biotic and abiotic stresses are applied, or global regulators and regulatory mechanisms are manipulated (33). These strategies may lead to the discovery of novel compounds, whose corresponding genes have to be subsequently identified (32). The targeted approach would be to manipulate the genes within the BGC or, if no molecular tools are available for the organism from which it originates, to follow a heterologous expression strategy by introducing the essential genes in an established host organism. Apart from core enzymes, BGCs may also contain genes encoding for transporters (34), transcription factors (35), or resistance genes (36). While their gene products are not directly involved in the biosynthesis of a SM, they are still essential for the production in the organism. In contrast, only the biosynthetic genes and a selection of other essential genes (e.g. transporters) are necessary for heterologous expression [reviewed in (37)]. Adjacent to essential genes for the production of SMs, BGCs can contain so-called gap genes. Gap genes are not involved directly or indirectly in the biosynthesis, export or gene activation of the production of a SM. The inclusion of gap genes in targeted approaches can lead to valuable time spent futilely in the laboratory without meaningful results. For this purpose, the detection of genes involved in the biosynthesis

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

of a SM and differentiation from gap genes is advantageous. This information can be obtained by the exploration of transcriptome data since the genes essential for SM production within a BGC are typically co-expressed with each other but not with the gap genes (38). Then again, this demands the knowledge of expression conditions and does not work for silent BGCs. In general, BGCs are suggested to undergo a distinct and faster evolution than the rest of the genome, based on different mechanisms and genetic drivers (39-45). This suggests that protein family co-evolution can be used to distinguish gap genes from essential genes in the BGC. In other words, genes involved in the biosynthetic process of a certain SM share a similar evolutionary background and can therefore be considered co-evolutionary linked. The two chapters included in this thesis describing and introducing the Functional Order tool (FunOrder), as a semi-automated method for the identification of co-evolutionary linked genes in BGCs, and FunOrder 2, as the fully automated and enhanced software package, capitalize on this hypothesis. More details on this subject can be found in both chapters.

As previously introduced, novel uncharacterized fungal genomes might harbor the potential for the production of novel drug lead compounds (46). This hypothesis is addresses by sequencing, assembly and analysis of yet uncharacterized fungal species primarily for the search and analysis of their slumbering SM production potential. Plant-associated endo- and epiphytic fungi are considered to be among the most prolific SM producers (16, 47-49). Consequently, many new fungi have been isolated from the phyllosphere with the aim to find novel SMs. In the recent years, the search area was broadened towards more extreme environments such as marine or arctic habitats (39). These efforts and the further sampling from host associated fungi have led to the discovery of manifold diverse species, which were described and classified, but remained understudied in respect to their secondary metabolism due to the sheer quantity of new isolates (16). The first genome presented is from *Aureobasidium pullulans* var. *aubasidani* strain CBS 100524. *A. pullulans* is a black yeast-like ascomycete with industrial importance due to its extracellular polysaccharides (50). The main exopolysaccharide of *A. pullulans* var. *aubasidani* strain CBS 100524 is aubasidan instead of pullulan (51, 52). This strain was previously isolated from plant exudates of a *Betula* sp. from the Leningrad region, Russia (51). Based on a previous multilocus analysis, *A. pullulans* var. *aubasidani* strain CBS 100524 and *A. pullulans* var. *pullulans* EXF-150 are part of the same phylogenetic group (52). The second group of sequenced and analyzed genomes were from the species *Wardomyces moseri*. The ascomycete *W. moseri* was first isolated from a dead petiole of *Mauritia minor* in Colombia in 1980. Walter Gams described the fungus in 1995 and named it after his mentor Meinhard Moser (CBS 164.80) (53). This fungus forms sporodochium-like structures and aggregates conidia loosely in slimy masses. *W. moseri* was described already in 1995 as an unusual *Wardomyces* species, because of its easily released conidia. Later, Sandoval-Denis *et al.* showed that the large subunit (LSU) rRNA gene and the internally transcribed spacer (ITS) sequences of *W. moseri* clustered among the Xylariales but not with the genus *Wardomyces* (54). The fungal order of Xylariales (Ascomycota) holds a large number of symbionts, saprotrophs, a variety of isolated endophytes, and plant pathogens (47, 49, 55). *W. moseri* appears related to members of the *Amphisphaeriaceae* and *Clypeosphaeriaceae*. Based on these findings, *W. moseri* was suggested to be re-examined regarding its taxonomic assignment. To date, there is only one more preprint mentioning this fungus indicating again the apparent misclassification (56). The metabolic potential of this historic fungus was analyzed and its taxonomy re-evaluated, by sequencing the genomes of the ex-isotype strain *W. moseri* CBS 164.80 and two isolates from the opposite side of the world, *W. moseri* TUCIM 5827 and TUCIM 5799.

As a final remark, very large supplemental files (some surpassing hundreds of pages) connected to the aforementioned studies are deposited online and can be found following the respective links for the repositories or the journal. This is necessary especially but not exclusively for the sequenced genomes, their annotation and gene prediction. Nevertheless, the genomes are publicly available in the NCBI National Center for Biotechnology Information repository.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Aims

The major aim of this PhD was the modeling and establishment of novel approaches to investigate secondary metabolite (SM) production in primarily fungal microorganisms. The first endeavor to address this aim was to attempt to discover a method for the identification and search of fungal ribosomally synthesized and post-translationally modified peptide (RiPP) biosynthetic gene clusters (BGCs). To achieve this first objective existing bioinformatics tools, the combination of said tools and elucidation of novel pipelines using available genomic and transcriptomic data were applied.

A central target to achieve the major aim was the development of a novel strategy in analyzing detected fungal BGCs with the ultimate goal to decide which genes should be included in heterologous expression efforts. For this purpose, a new software package leveraging protein family co-evolution was developed to facilitate future studies.

Another aim of this thesis, as stated by the PhD program TU Wien bioactive, was to sequence and analyze fungal genomes, focusing on their SM production potential. To this end, fungal strains were chosen for sequencing based on their respective taxonomic placement.

Conclusions

The five included peer reviewed publications make evident that the specified aims were fully addressed during the work on this thesis. A new method for the detection of ribosomally synthesized and post-translationally modified peptide (RiPP) precursors in fungal genomes was introduced. A new software package was written and validated for the differentiation of essential biosynthetic genes from gap genes within fungal biosynthetic gene clusters (BGCs). Finally, the genomes of four fungal strains were sequenced and analyzed with a special focus on their secondary metabolism.

The first presented publication is the first report of the potential of the fungal genus *Trichoderma* to produce RiPPs and the clusters detected by the presented novel method might eventually lead to the discovery of uncharacterized fungal RiPPs. This study ultimately led to the funded FWF-Project Nr. P 34036 “Identification and Characterization of Novel Fungal RiPPs”.

The second and third study focused on a new software package, the Functional Order (FunOrder) tool, which aims to distinguish gap genes from biosynthetic genes within fungal BGCs. The results of these two studies indicate that protein family co-evolution can be leveraged for the differentiation of gap genes from genes involved in the biosynthesis of a secondary metabolite (SM).

The two final publications address the aim of sequencing and analysis of uncharacterized fungal strains. Four whole fungal genomes including its mitochondria were successfully sequenced and it could be demonstrated how historic strains from already existing strain collections can be used for the exploration of novel natural products. Further, considering the presented phylogenetic evidence, the species *Wardomyces moseri* was suggested to be placed in the phylogenetic family *Sporocadaceae*.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

References for Introduction

1. Thirumurugan D, Cholarajan A, Raja SSS, Vijayakumar R. An Introductory Chapter: Secondary Metabolites. In: Vijayakumar R, Raja SSS, editors. Secondary Metabolites - Sources and Applications. London, UK: IntechOpen Limited; 2018.
2. Malik VS. Microbial secondary metabolism. Trends in Biochemical Sciences. 1980;5(3):68-72.
3. Keller NP, Turner G, Bennett JW. Fungal secondary metabolism — from biochemistry to genomics. Nature Reviews Microbiology. 2005;3(12):937-47.
4. van den Berg MA, Westerlaken I, Leeflang C, Kerkman R, Bovenberg RA. Functional characterization of the penicillin biosynthetic gene cluster of Penicillium chrysogenum Wisconsin54-1255. Fungal Genet Biol. 2007;44(9):830-44.
5. Weber G, Schörgendorfer K, Schneider-Scherzer E, Leitner E. The peptide synthetase catalyzing cyclosporine production in *Tolypocladium niveum* is encoded by a giant 45.8-kilobase open reading frame. Current Genetics. 1994;26:120-5.
6. Kensler TW, Roebuck BD, Wogan GN, Groopman JD. Aflatoxin: a 50-year odyssey of mechanistic and translational toxicology. Toxicol Sci. 2011;120 Suppl 1:S28-48.
7. Mulder KC, Mulinari F, Franco OL, Soares MS, Magalhaes BS, Parachin NS. Lovastatin production: From molecular basis to industrial process optimization. Biotechnol Adv. 2015;33(6 Pt 1):648-65.
8. Gomez BL, Nosanchuk JD. Melanin and fungi. Curr Opin Infect Dis. 2003;16(2):91-6.
9. Wheeler MH, Bell AA. Melanins and their importance in pathogenic fungi. Curr Top Med Mycol. 1988;2:338-87.
10. Luo S, Dong SH. Recent Advances in the Discovery and Biosynthetic Study of Eukaryotic RiPP Natural Products. Molecules. 2019;24(8).
11. Montalban-Lopez M, Scott TA, Ramesh S, Rahman IR, van Heel AJ, Viel JH, et al. New developments in RiPP discovery, enzymology and engineering. Nat Prod Rep. 2020.
12. Osbourn A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends in Genetics. 2010;26(10):449-57.
13. Tran PN, Yen MR, Chiang CY, Lin HC, Chen PY. Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. Appl Microbiol Biotechnol. 2019;103(8):3277-87.
14. Hoffmeister D, Keller NP. Natural products of filamentous fungi: enzymes, genes, and their regulation. Natural product reports. 2007;24(2):393-416.
15. Bassett EJ, Keith MS, Armelagos GJ, Martin DL, Villanueva AR. Tetracycline-labeled human bone from ancient Sudanese Nubia (A.D. 350). Science. 1980;209(4464):1532-4.
16. Hyde KD, Xu J, Rapior S, Jeewon R, Lumyong S, Niego AGT, et al. The amazing potential of fungi: 50 ways we can exploit fungi industrially. Fungal Diversity. 2019;97(1):1-136.
17. Alberti F, Foster GD, Bailey AM. Natural products from filamentous fungi and production by heterologous expression. Applied microbiology and biotechnology. 2017;101(2):493-500.
18. Newman DJ, Cragg GM, Kingston DGI. Chapter 5 - Natural Products as Pharmaceuticals and Sources for Lead Structures**Note: This chapter reflects the opinions of the authors, not necessarily those of the US Government. In: Wermuth CG, Aldous D, Raboisson P, Rognan D, editors. The Practice of Medicinal Chemistry (Fourth Edition). San Diego: Academic Press; 2015. p. 101-39.
19. Epstein SC, Charkoudian LK, Medema MH. A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences. Stand Genomic Sci. 2018;13:16.
20. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research. 2019;48(D1):D454-D8.

21. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47(W1):W81-w7.

22. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45(W1):W36-W41.

23. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema Marnix H, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research. 2021;49(W1):W29-W35.

24. Wolf T, Shelest V, Nath N, Shelest E. CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. Bioinformatics. 2016;32(8):1138-43.

25. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal genetics and biology : FG & B. 2010;47(9):736-41.

26. Almeida H, Palys S, Tsang A, Diallo AB. TOUCAN: a framework for fungal biosynthetic gene cluster discovery. NAR Genomics and Bioinformatics. 2020;2(4).

27. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 2019;47(18):e110.

28. Kloosterman AM, Medema MH, van Wezel GP. Omics-based strategies to discover novel classes of RiPP natural products. Current Opinion in Biotechnology. 2021;69:60-7.

29. Umemura M, Nagano N, Koike H, Kawano J, Ishii T, Miyamura Y, et al. Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in Aspergillus flavus. Fungal Genet Biol. 2014;68:23-30.

30. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep. 2013;30(1):108-60.

31. Brakhage AA, Schroeckh V. Fungal secondary metabolites - strategies to activate silent gene clusters. Fungal genetics and biology : FG & B. 2011;48(1):15-22.

32. Atanasov AG, Zotchev SB, Dirsch VM, Orhan IE, Banach M, Rollinger JM, et al. Natural products in drug discovery: advances and opportunities. Nature Reviews Drug Discovery. 2021.

33. Wiemann P, Keller NP. Strategies for mining fungal natural products. Journal of industrial microbiology & biotechnology. 2014;41(2):301-13.

34. Wang DN, Toyotome T, Muraosa Y, Watanabe A, Wuren T, Bunsupa S, et al. GliA in Aspergillus fumigatus is required for its tolerance to gliotoxin and affects the amount of extracellular and intracellular gliotoxin. Medical mycology. 2014;52(5):506-18.

35. Derntl C, Rassinger A, Srebotnik E, Mach RL, Mach-Aigner AR. Identification of the Main Regulator Responsible for Synthesis of the Typical Yellow Pigment Produced by *Trichoderma reesei*. Appl Environ Microbiol. 2016;82(20):6247-57.

36. Schrettl M, Carberry S, Kavanagh K, Haas H, Jones GW, O'Brien J, et al. Self-protection against gliotoxin--a component of the gliotoxin biosynthetic cluster, GliT, completely protects Aspergillus fumigatus against exogenous gliotoxin. PLoS Pathog. 2010;6(6):e1000952.

37. Anyaogu DC, Mortensen UH. Heterologous production of fungal secondary metabolites in Aspergilli. Frontiers in Microbiology. 2015;6(77).

38. Tai Y, Liu C, Yu S, Yang H, Sun J, Guo C, et al. Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (Camellia sinensis). BMC Genomics. 2018;19(1):616.

39. Keller NP. Fungal secondary metabolism: regulation, function and drug discovery. Nat Rev Microbiol. 2019;17(3):167-80.

40. Lind AL, Wisecaver JH, Lameiras C, Wiemann P, Palmer JM, Keller NP, et al. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. PLOS Biology. 2017;15(11):e2003583.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

41. Rokas A, Wisecaver JH, Lind AL. The birth, evolution and death of metabolic gene clusters in fungi. Nature reviews Microbiology. 2018;16(12):731-44.
42. Palmer JM, Keller NP. Secondary metabolism in fungi: does chromosomal location matter? Current opinion in microbiology. 2010;13(4):431-6.
43. Hoogendoorn K, Barra L, Waalwijk C, Dickschat JS, van der Lee TAJ, Medema MH. Evolution and Diversity of Biosynthetic Gene Clusters in Fusarium. Frontiers in microbiology. 2018;9:1158.
44. Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: How natural selection drives chemical innovation. Proceedings of the National Academy of Sciences. 2008;105(12):4601-8.
45. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. PLOS Computational Biology. 2014;10(12):e1004016.
46. Higginbotham SJ, Arnold AE, Ibañez A, Spadafora C, Coley PD, Kursar TA. Bioactivity of Fungal Endophytes as a Function of Endophyte Taxonomy and the Taxonomy and Distribution of Their Host Plants. PLOS ONE. 2013;8(9):e73192.
47. Helaly SE, Thongbai B, Stadler M. Diversity of biologically active secondary metabolites from endophytic and saprotrophic fungi of the ascomycete order Xylariales. Natural product reports. 2018;35(9):992-1014.
48. Ancheeva E, Daletos G, Proksch P. Bioactive Secondary Metabolites from Endophytic Fungi. Curr Med Chem. 2020;27(11):1836-54.
49. Becker K, Stadler M. Recent progress in biodiversity research on the Xylariales and their secondary metabolism. The Journal of Antibiotics. 2021;74(1):1-23.
50. Rekha M, Sharma CP. Pullulan as a promising biomaterial for biomedical applications: a perspective. Trends Biomater Artif Organs. 2007;20(2):116-21.
51. Yurlova NA, de Hoog GS. A new variety of Aureobasidium pullulans characterized by exopolysaccharide structure, nutritional physiology and molecular features. Antonie Van Leeuwenhoek. 1997;72(2):141-7.
52. Zalar P, Gostinčar C, de Hoog GS, Uršič V, Sudhadham M, Gunde-Cimerman N. Redefinition of Aureobasidium pullulans and its varieties. Studies in Mycology. 2008;61:21-38.
53. Grams W. An unusual species of Wardomyces (Hyphomycetes). Beih Sydowia X. 1995:67-72.
54. Sandoval-Denis M, Guarro J, Cano-Lira JF, Sutton DA, Wiederhold NP, de Hoog GS, et al. Phylogeny and taxonomic revision of Microascaceae with emphasis on synnematous fungi. Stud Mycol. 2016;83:193-233.
55. Franco MEE, Wisecaver JH, Arnold AE, Ju Y-M, Slot JC, Ahrendt S, et al. Secondary metabolism drives ecological breadth in the Xylariaceae. bioRxiv. 2021:2021.06.01.446356.
56. Milan CS, Kevin DH, Sajeewa SNM, Marc S, Jones EBG, Itthayakorn P, et al. Taxonomy, Phylogeny, Molecular Dating and Ancestral State Reconstruction of Xylariomycetidae (Sordariomycetes). Fungal Diversity. 2021.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Vignolle *et al. BMC Genomics* (2020) 21:258
https://doi.org/10.1186/s12864-020-6653-6

BMC Genomics

METHODOLOGY ARTICLE Open Access

Novel approach in whole genome mining and transcriptome analysis reveal conserved RiPPs in *Trichoderma* spp

Gabriel A. Vignolle, Robert L. Mach, Astrid R. Mach-Aigner and Christian Derntl*

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Abstract

Background: Ribosomally synthesized and post-translationally modified peptides (RiPPs) are a highly diverse group of secondary metabolites (SM) of bacterial and fungal origin. While RiPPs have been intensively studied in bacteria, little is known about fungal RiPPs. In Fungi only six classes of RiPPs are described. Current strategies for genome mining are based on these six known classes. However, the genes involved in the biosynthesis of theses RiPPs are normally organized in biosynthetic gene clusters (BGC) in fungi.

Results: Here we describe a comprehensive strategy to mine fungal genomes for RiPPs by combining and adapting existing tools (e.g. antiSMASH and RiPPMiner) followed by extensive manual curation based on conserved domain identification, (comparative) phylogenetic analysis, and RNASeq data. Deploying this strategy, we could successfully rediscover already known fungal RiPPs. Further, we analysed four fungal genomes from the *Trichoderma* genus. We were able to find novel potential RiPP BGCs in *Trichoderma* using our unconventional mining approach.

Conclusion: We demonstrate that the unusual mining approach using tools developed for bacteria can be used in fungi, when carefully curated. Our study is the first report of the potential of *Trichoderma* to produce RiPPs, the detected clusters encode novel uncharacterized RiPPs. The method described in our study will lead to further mining efforts in all subdivisions of the fungal kingdom.

Keywords: Genome mining, RiPP, *Trichoderma*, Ascomycota, Basidiomycota, Secondary metabolism

Background

Secondary metabolites (SMs) from fungal sources have played a crucial role in improving human health not only since the discovery of Penicillin, but even in prehistoric times [1, 2]. These natural products and chemically modified variants are widely used as antibiotics, immunomodulators and anti-cancer drugs [3]. Generally well-known examples of fungal SMs belong to two main classes. They are either polyketides (e.g. the mycotoxin aflatoxin and the cholesterol-lowering drug lovastatin) or non-ribosomal peptides (e.g. the antibiotic penicillin and the immunosuppressant cyclosporine). However, also other SM classes are present in fungi: e.g. terpenes, melanins [4, 5], and ribosomally synthesized and post-translationally modified peptides (RiPPs). RiPPs are a rapid growing group of natural products that can be classified in more than 20 different compound classes. Please refer to the reviews by Arnison, P. G. et al. and Luo, S. & Dong, S. H [6, 7]. Small peptides are of increasing interest due to unique bioactive properties aiming at "undruggable" diseases and successfully eradicating anti-biotic resistant microorganisms [8]. The many applications of natural cyclic peptides, including potent lipid-lowering effects of fungal cyclic peptides, are reviewed by Abdalla, M. A. & McGaw, L. J [9].

---

* Correspondence: christian.derntl@tuwien.ac.at
Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Gumpendorfer Strasse 1a, 1060 Wien, Austria

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

It is important to differentiate RiPPs from fungal Kexin-like proteinase (KEX2)-processed repeat proteins called KEPs. KEPs are small secreted peptides that do not undergo post-translational modifications, their precursor peptide is cleaved by different proteases and then released by exocytosis [10]. As described by Le Marquer et al., many of these KEPs are putative sexual pheromones but may also play other important roles.

Biosynthesis of RiPPs follows a very straight forward production pathway (Fig. 1). A precursor peptide consisting of a leader, a core and a follower amino acid sequence is synthesized by the ribosome. The subsequent post-translational modifications of the core sequence are mediated by modifying enzymes as specified by the leader and follower sequences. After removal of the leader and the follower sequences, the finished bioactive RiPP is released. Many RiPPs undergo a cyclisation step that stabilizes them, reduces their toxicity, improves binding affinity and selectivity. These properties make cyclized RiPPs very attractive candidates for drug development. This labels fungal RiPPs the potential next generation therapeutics [11]. However, only six different classes of RiPPs are described in fungi, yet. Two classes are found in basidiomycetes, i.e. the amatoxins and phallotoxins in the genus *Amanita,* and the borosins with selective nematotoxic activity in *Omphalotus olearius*. RiPPs produced by ascomycetes are the dikaritins and are classified as ustiloxins, asperipins, phomopsins and epichloëcyclins [7].

The genes encoding for the biosynthetic enzymes for SMs are often arranged in individual clusters named biosynthetic gene cluster (BGC), regardless of the class of SM [1]. This organization of clusters is also given for the previously described fungal RiPPs ustiloxins, phomopsins, amatoxins, phallotoxins, borosins and asperipins [7]. This clustered organization is one important feature for the in silico identification of BGCs. Recent advances in next generation sequencing (NGS) lead to the publication of more and more high-quality full genomes from various fungal species and genera such as *Aspergillus flavus* or various *Trichoderma* spp. [12, 13]. Today, fungi represent a vast and generally untapped pool for new lead compounds with pharmaceutical and agricultural applications [14]. However, efforts in genome mining for the search of RiPP BGCs, that encode for the machinery responsible to produce secondary metabolites, have thus far been focused on bacterial genomes due to the lack of a large database of fungal RiPPs [11, 15]. Therefore, most bioinformatic tools available are tailored to mine bacterial genomes for RiPPs.

The current online version of antiSMASH ver. 5.0 includes the identification of RiPP clusters in fungal genomes based on the query sequence (YVIPID) of the putative precursor peptide sequence of phomopsin and

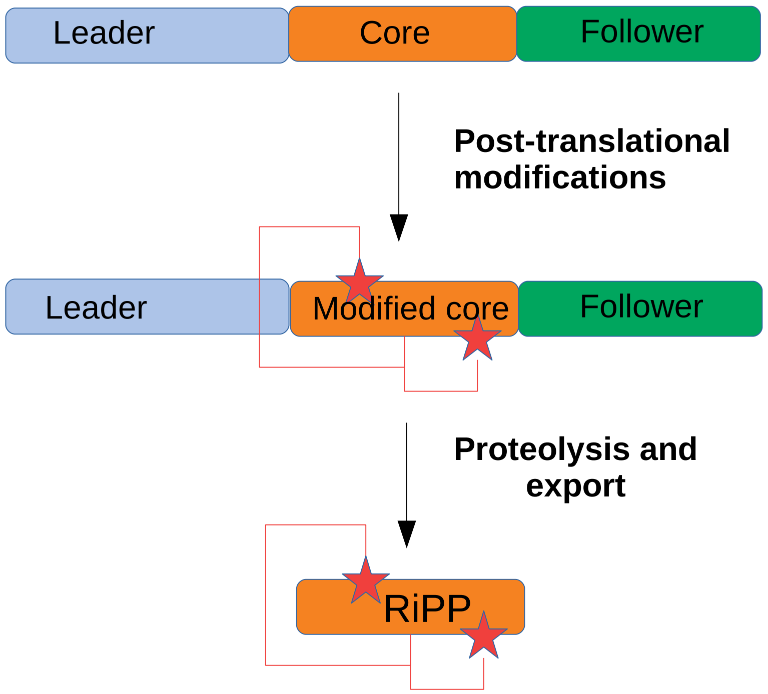


Fig. 1 General RiPP biosynthetic pathway. The leader and follower peptide direct the modifications (e.g. addition of functional groups, indicated by stars, or formation of additional bonds, indicated by the connective lines) on the core peptide. After removal of the leader and follower sequence the mature RiPP is released. The figure is an adaptation of the original figure in [6]

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

ustYa/ustYb together with the *ustA* precursor peptide of the ustiloxin cluster [16]. This approach, although being restrictive in its potential to detect novel classes, will aid in the mining for ustiloxin and phomopsin like RiPPs in fungi. Previously, this approach was able to detect 94 putative RiPP precursor peptides in *Aspergillus* spp. This led to the discovery of structurally new cyclic peptides (Asperipins) even though the clusters exhibit high homology to the ustiloxin clusters [7, 17, 18]. We reason that a broader, unconventional forward approach for the detection of putative precursor peptides can be achieved by utilisation and adaptation of bioinformatic tools developed for bacteria. This approach might lead to the discovery of novel fungal RiPPs with potentially new applications and unknown adjacent modifying enzymes. These novel enzymes and the identified precursor peptides can furthermore be used to identify more homologous RiPP BGCs across the fungal kingdom as it was done for the ustiloxin cluster, thereby broadening our search parameters for novel RiPP BGCs.

Trichoderma spp. are mesophilic ascomycetes and part of the sordariomycetes, one of the largest classes within their division. The genus *Trichoderma* contains mycoparasitic, saprophytic and opportunistically pathogenic fungi. *T. reesei* is a well-studied saprobe and used industrially for the production of cellulases and hemicellulases [12]. *T. harzianum* is a ubiquitous species with agricultural applications, the opportunistically pathogenic *T. citrinoviride* is often isolated as endophyte and *T. brevicompactum* is a producer of antifungal metabolites [12, 19–21]. All mentioned *Trichoderma* species contain various classes of BGCs, Type 1 polyketide synthetases (T1pks), nonribosomal peptide synthases (NRPSs), terpene BGCs, fatty acid BGCs and various combined and putative clusters.

In this study we demonstrate in silico that by combining antiSMASH [22], the ClusterFinder algorithm and a full HMMer analysis a large set of putative SM BGCs can be identified. After cross-referencing the individual results, we predicted potential RiPP precursor peptides. These sequences were further refined by using previously published RNASeq data [23] and thereby providing a comprehensive highly probable in silico prediction backed up with genomic and transcriptional data.

Results

Diversity of secondary metabolite gene clusters in *Trichoderma* spp. and known fungal RiPP producers

First, we compared the biosynthetic gene clusters diversity of nine randomly chosen *Trichoderma* species for which high quality genomes were available. To this end, they were all mined with the command line version of antiSMASH ver. 4.3.0 [22]. We also mined the genomes of *A. flavus* and *Amanita phalloides* in which fungal RiPPs were previously described. The results of the mining with the command line version of antiSMASH are shown in Table 1. The total number of SM BGCs ranges from 11 for the *A. phalloides* genome to 186 found in the SM producer *A. flavus*. There was neither Type 3 pks clusters found in the *Trichoderma* spp. nor any siderophore or indole clusters. Notably, antiSMASH ver. 4.3.0 [22] does not yet include the search for fungal RiPP clusters. The web based antiSMASH ver. 5.0 [16] contains this feature, and was able to detect the ustiloxin B cluster in the *A. flavus* genome, but no other fungal RiPP clusters were found in the mined genomes. Nevertheless, the *Trichoderma* spp. already display a high potential to produce a diverse range of SMs, based on the antiSMASH results.

Next, we calculated the average nucleotide identity (ANI) for each strain against each other (Fig. 2). Within the *Trichoderma* spp. there are three distinct clusters detectable based on the ANI value and the computed dendrogram when applying 85% ANI as cutoff. The first containing *T. harzianum*, *T. atrobrunneum* and *T. virens;* the second *T. arundinaceum* and *T. brevicompactum*; the third *T. reesei*, *T. koningii* and *T. citrinoviride*. Based on these findings *T. reesei* and one high quality genome from each cluster were chosen to be mined for putative RiPP precursor genes namely *T. harzianum*, *T. citrinoviride* and *T. brevicompactum*.

The RiPPMiner standalone tool detects fungal RiPP precursors

As a prerequisite for our analysis, we needed to test the applicability of the RiPPMiner [24] software to recognize precursor peptides of fungal RiPPs. To this end, we tested the software on known precursor peptides of fungal RiPPs extracted from the UniProt database, namely precursors for α-amanitin (A8W7M4), β-amanitin (ABW87785), phallacidin (ABW87771), phalloidin (ABW87787) and 75 diverse known precursor peptides (see Additional file 1). RiPPMiner was able to recognize all of them as precursor peptides, even though it classified them into bacterial groups and predicted improper structure models (Additional file 1). This is a consequence of the used model for these predictions; the model is based on a manually curated database of known precursor peptides of bacterial RiPPs. The precursor peptides for α-amanitin and β-amanitin were predicted to be lassopeptides, whereas phallacidin and phalloidin had no RiPP class prediction (Additional file 1). Next, we evaluated the potential of the RiPPMiner to detect RiPP precursors in known RiPP BGCs and distinguish them from functional polypeptides. To this end, we extracted the sequences from the ustiloxin B BGC from *A. flavus* using the web based antiSMASH ver. 5.0 [16] output and subjected them to an analysis using the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Table 1 Prediction of SM BCGs using antiSMASH ver. 4.3.0

Species	Total BGC	NRPS	T1pks	T3pks	Sid.[a]	Ter[b]	Ind[c]	Mix[d]	Other	fatty acid	putativ
T. reesei	80	7	9	0	0	6	0	3	4	1	50
T. citrinoviride	85	8	8	0	0	5	0	4	6	2	52
T. harzianum	129	5	19	0	0	7	0	8	8	2	80
T. brevicompactum	96	9	14	0	0	5	0	6	6	2	54
T. asperellum	92	5	9	0	0	7	0	6	4	2	59
T. arundinaceum	117	9	14	0	0	8	0	11	7	2	66
T. atrobrunneum	114	9	18	0	0	5	0	8	7	2	65
T. koningii	69	7	9	0	0	4	0	2	4	2	41
T. virens	145	16	14	0	0	10	0	8	11	2	84
A. flavus	186	11	19	2	1	10	4	10	12	3	114
A. phalloides	11	0	1	0	0	5	0	0	1	1	3

[a]Siderophore, [b]Terpene, [c]Indole, [d]T1pks-NRPS/Mix

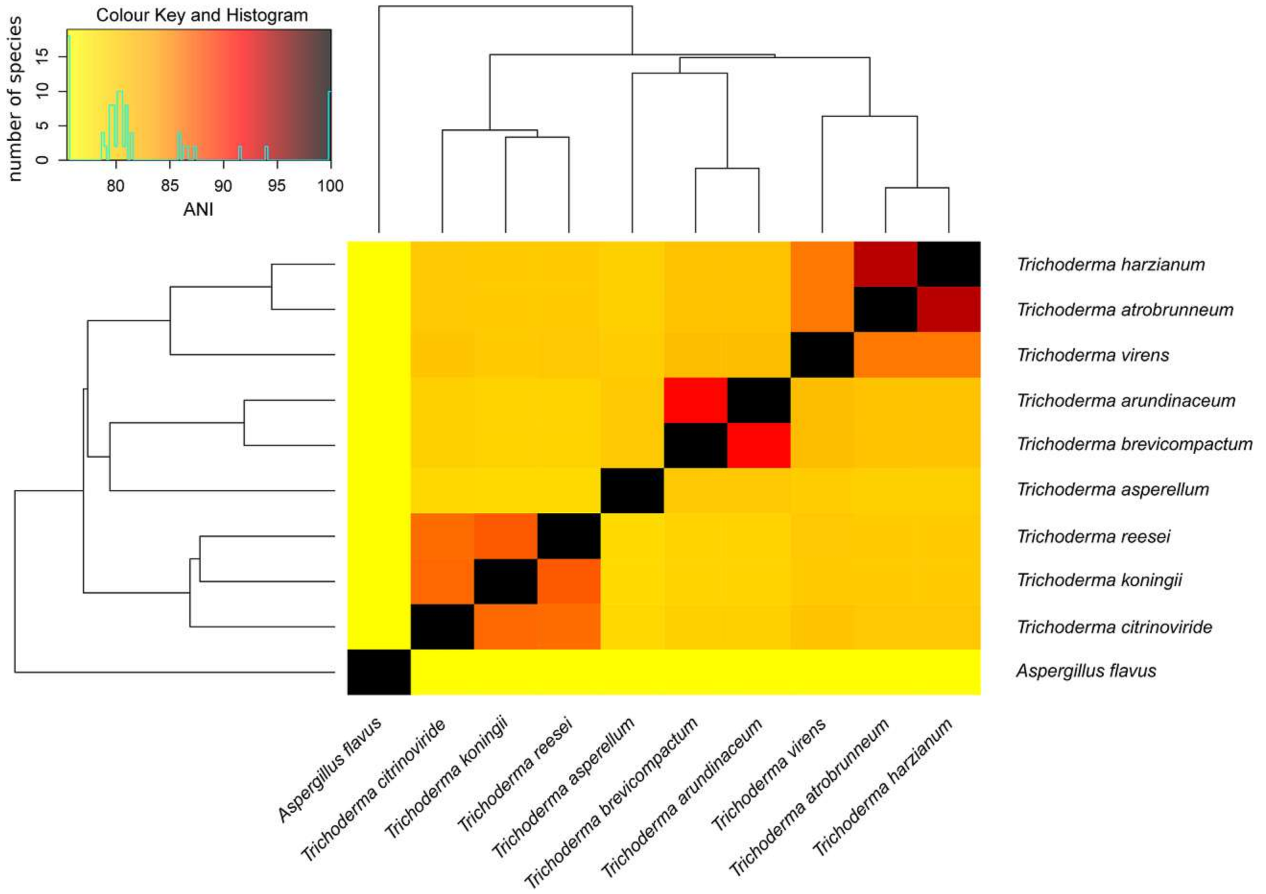


Fig. 2 Heatmap of the average nucleotide identity (ANI) between the analyzed fungal species. The respective ANI value is represented by the color gradient. In addition, a histogram indicates the number of species with that certain ANI value. The dendrogram in the heatmap is computed with the complete linkage method to find similar clusters based on the Euclidean distance, representing a whole genome phylogeny

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

RiPPMiner software (Additional file 2). The RiPPMiner was able to detect the verified RiPP precursor (#INPUT 13, AFLA_095020) but misclassified it as cyanobactin. The RiPPMiner predicted three additional input sequences (AFLA_094930, AFLA_094970, AFLA_095000) as RiPP precursors (Additional File 2). These are hypothetical proteins without predicted conserved regions [17]. These results demonstrate that the RiPPMiner software is able to identify fungal RiPP precursors, although it was designed to predict bacterial RiPPs. This leads to misclassifications and false positives.

Genome mining of *Trichoderma* spp. for putative RiPP precursors

As we could verify the in principal applicability of the RiPPMiner for the identification of fungal RiPP precursors, we proceeded with the search for RiPPs in *Trichoderma* spp. As shown in Fig. 3, the amino acid query sequences were extracted from the results from antiSMASH ver. 4.3.0 from the "putative clusters" found by the ClusterFinder algorithm. Only genes without classification from antiSMASH were chosen as query sequences. This means that core biosynthetic genes, additional biosynthetic genes, transport-related genes and regulatory genes were not included in the RiPP prediction. The prediction was performed with the standalone version of RiPPMiner. The results of the RiPP mining procedure for the four *Trichoderma* spp. are shown in Table 2. For *T. harzianum* 23% of the query sequences were predicted to be putative precursor RiPPs. In the *T. reesei* genome 15% of the query sequences were recognized as putative precursor peptides by the RiPPMiner, for *T. citrinoviride* 17% of the query sequences and in the *T. brevicompactum* genome 22% of the query sequences were predicted as putative RiPP precursor peptides (Table 2).

All amino acid sequences predicted to be RiPPs were manually inspected. This included aligning the sequences using Blastp v2.9.0+ [25] against the non-redundant protein database and a manually curated database of fungal proteomes to refine the search. Sequences with highly conserved active domains found in the Conserved Domain Database (CDD) [26] were removed, as well as classified sequences such as transcription factors, enzymes and ribosomal proteins. After manual inspection the sequences of *T. harzianum* were reduced to a final set of 222 sequences, *T. citrinoviride* was reduced to 110 and *T. brevicompactum* to 92. For *T. reesei* the genes for putative precursor sequences were furthermore compared to RNASeq data, and based on our analysis of the alignments to these genes, those without RNASeq data mapping to them were discarded as false positives. After further manual curation of the BGCs *T. reesei* was left with a final set of 6 putative RiPP precursor peptide genes.

RiPP analysis by maximum likelihood method

We then inferred a maximum likelihood (ML) phylogenetic tree based on the putative precursor RiPP peptides from *T. reesei*, *T. citrinoviride*, *T. harzianum* and *T. brevicompactum*, the known fungal RiPP precursor peptides α-amanitin (A8W7M4), β-amanitin (ABW87785), phallacidin (ABW87771) and phalloidin (ABW87787), in order to find evolutionary linked sequences and to detect possible precursor peptide families. The analysis involved a total of 434 amino acid sequences, with sequence lengths ranging from 27 to 150 amino acids. Following the multiple sequence

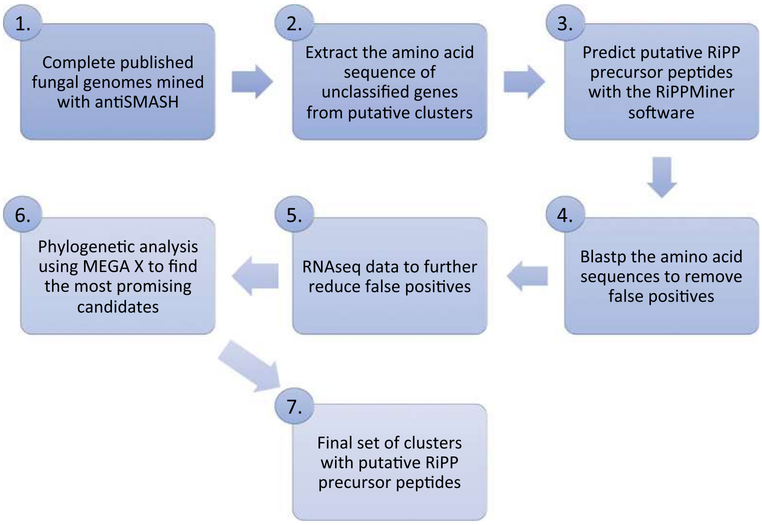

1.
Complete published fungal genomes mined with antiSMASH
2.
Extract the amino acid sequence of unclassified genes from putative clusters
3.
Predict putative RiPP precursor peptides with the RiPPMiner software
4.
Blastp the amino acid sequences to remove false positives
5.
RNAseq data to further reduce false positives
6.
Phylogenetic analysis using MEGA X to find the most promising candidates
7.
Final set of clusters with putative RiPP precursor peptides

Fig. 3 Schematic diagram of the pipeline used to discover possible RiPP precursor peptides

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Table 2 Number of predicted RiPPs (and subclasses) found by RippMiner

	T. reesei	*T. citirinoviride*	*T. harzianum*	*T. brevicompactum*
Query sequences	690	759	1099	518
Total predicted RiPPs	108	131	258	118
Cyanobactin	34	41	77	39
LanthipeptideB	10	7	18	4
LanthipeptideC	0	0	2	0
Lassopeptide	4	3	9	3
Linaridin	3	5	7	4
Microcin	1	5	5	0
Bacterial head to tail	0	0	1	2
Thiopeptide	1	0	0	0
Auto inducing peptide	0	1	0	0
NONE	54	69	140	66
After manual inspection	6	110	222	92

The protein sequences extracted from the predicted SM BGCs using antiSMASH ver. 4.3.0 (query sequences) were analyzed with the RiPPMiner software for *T. reesei*, *T. citrinoviride*, *T. harzianum*, and *T. brevicompactum*. False positives were removed via manual inspection

alignment computed with muscle [27], a total of 231 relevant positions were extracted for all sequences. These were used in the final data set to infer the phylogenetic distance corrected for multiple substitutions based on the substitution-rate matrices. The ML tree with the highest log likelihood (– 101,279.17) (Additional file 3) was used to extract the sub-trees including the putative RiPP precursor peptides from *T. reesei* and those including known precursor peptides of fungal RiPPs extracted from the UniProt database (Additional file 4). The branch lengths of the ML sub-tree are proportional to the relative distance between the sequences measured in the number of substitutions per site. As expected, the sequences of α-amanitin (A8W7M4), β-amanitin (ABW87785) and phalloidin (ABW87787) clustered together defining an own clade (Fig. 4). Within this amanitin/phalloidin sub-tree two sequences clustered closely together, one from the *T. reesei* set and one from *T. citrinoviride* (Additional file 1). These two putative RiPP precursor peptides were both defined by the RiPPMiner to be cyanobactins and have a bootstrap value of 0.98, making it highly likely that they are sisters to each other. Additionally, the *Trichoderma* sequences within this clade all showed high similarities to the putative structural toxin protein of *Eutypa lata* (UCREL1), the structural toxin protein RtxA of *Aspergillus oryzae* and *Metarhizium rileyi* in the Blastp output (Additional file 5).

We identified another outstandingly interesting subtree within the ML-tree. The amino acid sequence from *T. reesei* found in the BGCs 50 on contig 16 of the genome in the open reading frame 123 clusters within a conserved clade with the bootstrap values 0.64–0.89 (Fig. 5). The clade consists of one putative precursor RiPP peptide sequence from *T. harzianum*, *T. citrinoviride*, *T. brevicompactum* and *T. reesei* respectively. We called this clade the *Trichoderma*-putative-RiPP clade because the subtree of these sequences resembles the dendrogram in the heatmap, representing a whole genome phylogeny of the *Trichoderma* genus (Fig. 2). Within the other extracted sub-trees, the *T. reesei* putative RiPP precursor peptide sequences cluster with different sequences from each set of putative RiPP precursors namely *T. harzianum*, *T. citrinoviride* and *T. brevicompactum*. The sequences from the *T. harzianum*, *T. citrinoviride* and *T. brevicompactum* sets also made up own clades, these were not considered in the exploratory analysis due to the lack of RNASeq data for these specific strains and therefore the unknown high amount of false positive predicted RiPP precursor peptide sequences.

Analysis of the putative RiPP cluster 55 of *T. reesei*

Based on the phylogenetic and exploratory analyses of the putative RiPP precursor peptide sequences, we decided to perform a detailed analysis of a possible novel RiPP cluster found in *T. reesei*, namely cluster 55. Cluster 55 contains the putative RiPP precursor peptide that clustered in the ML tree in the amanitin/phalloidin clade (Fig. 4). This putative RiPP precursor peptide from *T. reesei* has a sister in *T. citrinoviride* with a high bootstrap value (Fig. 4). Furthermore, the RNASeq data showed that the putative RiPP precursor peptide from cluster 55 is transcribed at low levels. These findings highly suggest that this putative RiPP precursor peptide is indeed present in the genome of *T. reesei*. First, we manually annotated all genes in cluster 55 (as predicted by antiSMASH) by performing a Blastp v2.9.0+ [25] search against the non-redundant protein database, the conserved region finder and a manually curated database (Additional file 5). The results are visualized in Fig. 6.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

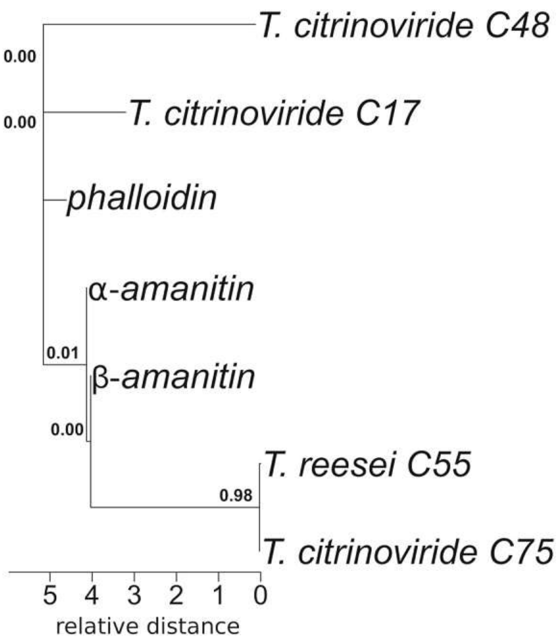


Fig. 4 The extracted subtree showing the amanitin/phalloidin clade from a maximum likelihood phylogenetic tree. The ML phylogenetic tree was inferred based on 434 amino acid sequences. The evolutionary history was inferred by using the Maximum Likelihood method and Dayhoff w/ freq. Model. The tree with the highest log likelihood (− 101,279.17) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The final dataset consisted of a total of 231 sites. The analyses were conducted in MEGA X [27].

Gene D was classified as a putative major facilitator superfamily (MFS) general substrate transporter. Notably, the same kind of transporter is found in the ustiloxin B cluster of *A. flavus* [17]. Adjacent to the gene of the putative RiPP precursor peptide a sulfatase gene is encoded (Gene F). Additionally, a putative hydrolase (Gene L), acid phosphatase (Gene N), cytochrome P450 (Gene P) and peptidases (Gene S) are found. This gives the cluster 55 the potential arsenal of enzymes needed for posttranslational modification and transport of the finished putative RiPP. Notably, these and further genes of *T. reesei cluster* 55 have homologs in *T. citrinoviride* cluster 75 (Additional files 6, 7 and 8).

RiPP precursor peptide analysis

Next, we performed a more in-depth analysis of the putative RiPP precursor peptide sequence found in cluster 55 of *T. reesei* (Fig. 6). The analysis involved a multiple sequence alignment with the 20 top hits from the Blastp output using the ClustalW algorithm and was performed with PRALINE [28] (Fig. 7). The putative RiPP precursor peptide from *T. reesei* is 109 amino acids long. There was no O-glycosylation potential predicted with NetOGlyc4.0 [29] and only a single low potential N-glycosylation site could be detected at the asparagine in position 33 with NetNGlyc1.0 [30]. There was no N-myristoylation site found nor was there a C-terminus appropriate for peroxisomal import detected. To detect DNA motif binding sites NsitePred [31] was used, only low probability motifs were found (below 0.264) not giving the precursor peptide the ability to bind DNA effectively. These analyses were performed to exclude the possibility of the putative precursor peptide being involved in transcriptional regulation. To determine the core sequence of the putative precursor peptide firstly the possible posttranslational modifications based on the enzymes found in the cluster 55 were evaluated. The adjacent sulfatase gene strongly suggests a sulfated residue. To detect an appropriate sulfating site within the peptide the Sulfinator [32] application was used. It found the Tyrosine in position 96 to be the only possible sulfated site within the peptide. This suggests residue 96 to lie within the core sequence. The RiPPMiner software predicted the core sequence to be residues 91 to 99, comprising the core sequence KKAHPYEEP (Fig. 7). The start of this putative core sequence is a typical peptidase cut site (KK) only found once in the putative precursor peptide. Tyrosine 101 (2 residues after the C-terminal end of the predicted core sequence) is a predicted phosphorylation site according to NetPhos3.1 [33]. This might suggest a possible activation site for further processing of the core peptide sequence. Furthermore, the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

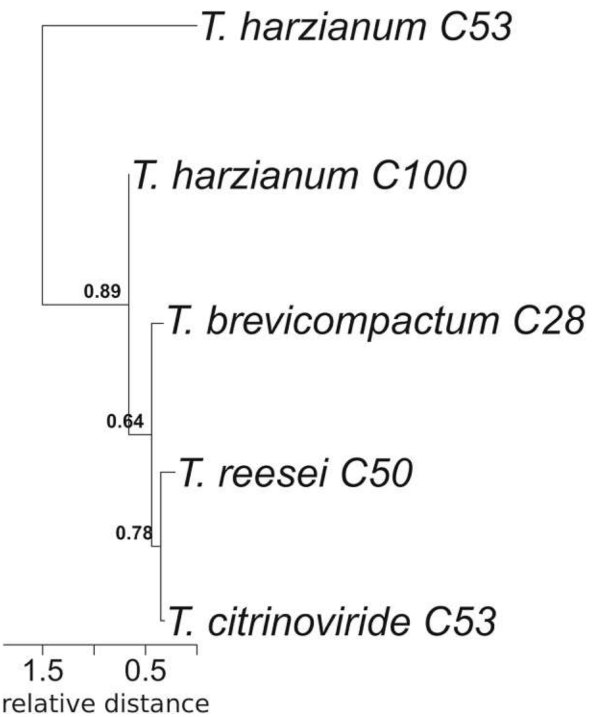


Fig. 5 The extracted subtree showing the *Trichoderma*-putative-RiPP clade from a maximum likelihood phylogenetic tree. The ML phylogenetic tree was inferred based on 434 amino acid sequences. The evolutionary history was inferred by using the Maximum Likelihood method and Dayhoff w/freq. Model. The tree with the highest log likelihood (– 101,279.17) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The final dataset consisted of a total of 231 sites. The analyses were conducted in MEGA X [27].

predicted core sequence is highly conserved, the main part is not predicted to be part of either an alpha-helix nor a beta-sheet, and the amino acids of the possible predicted core sequence are mainly hydrophilic (Fig. 7), making them easily accessible for enzymatic posttranslational modification. These findings further support the peptide from residues 91 to 99 to be the core sequence or at least part of the core of the putative RiPP precursor peptide.

Discussion

Using antiSMASH ver. 5.0 [16] for the search of BGCs returned one identified fungal RiPP cluster in the genome

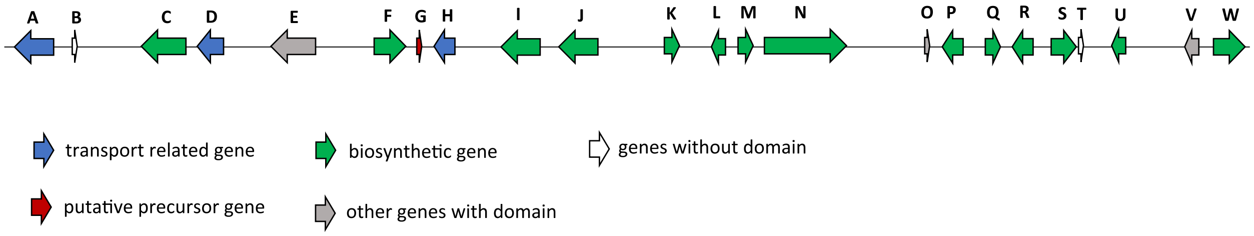


Fig. 6 Schematic representation of biosynthetic gene cluster 55 of *T. reesei*. The gene cluster is located on scaffold 19 (571843–660,892 nt) and contains 22 predicted genes and two possible pseudogenes. Gene A is a putative general substrate transporter, position B is a possible pseudogene, gene C a glycosyltransferase from the family 1, gene D is a putative MFS general substrate transporter, gene E is a HET-domain-containing protein, gene F is a sulfatase, gene G is a putative RiPP precursor peptide, gene H is a putative amino acid transporter, gene I is a chitinase, gene J is a putative GMC oxidoreductase, gene K is a casein kinase II alpha subunit, gene L is a putative alpha/beta-hydrolase, gene M is a GroES-like protein, gene N is an acid phosphatase, gene O is a hypothetical protein, gene P encodes for a cytochrome P450, gene Q is a putative NAD (P)-binding protein, gene R is a family 54 glycoside hydrolase, gene S is a putative carboxypeptidase S, gene T is a possible pseudogene, gene U is a putative class I glutamine amidotransferase-like protein, gene V is a PTH11-type GPCR and gene W is a GMC oxidoreductase. The gene annotations were manually curated and based on a Blastp v2.9.0+ (protein-protein BLAST) [25] search against the non-redundant protein database, the conserved region finder and a manually curated database (Additional file 5)

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

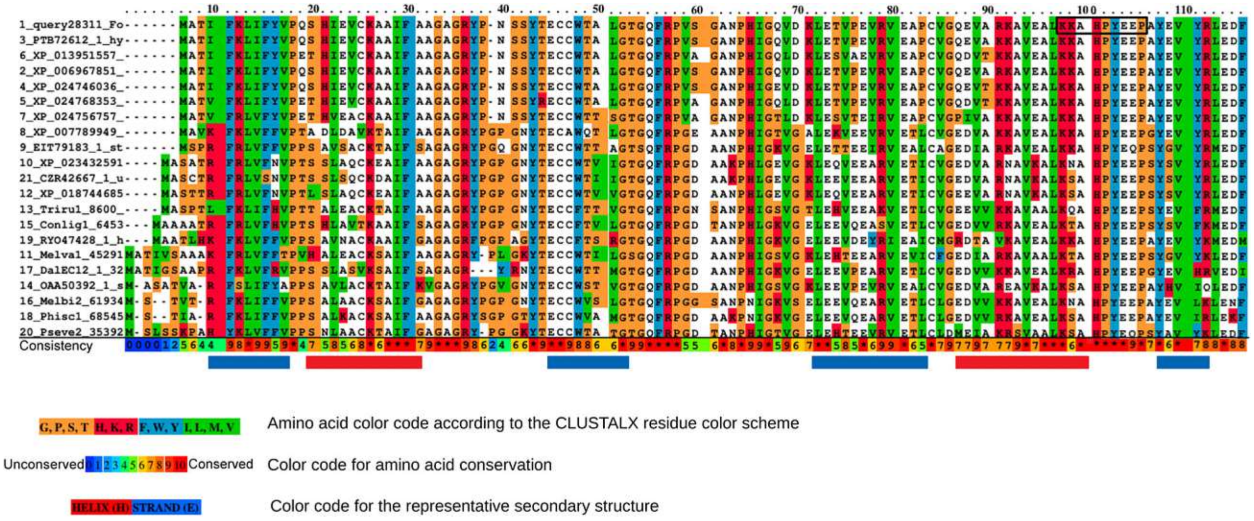


Fig. 7 Multiple sequence alignment of the putative RiPP precursor peptide from cluster 55. The putative RiPP precursor peptide from *T. reesei* was aligned to the 20 top hits from the Blastp output using the ClustalW algorithm and was performed with PRALINE [28]. The 21 aligned amino acid sequences are colored according to the ClustalX residue color-scheme. The row labeled 'Consistency' is color-coded based on the amino acid conservation performed by PRALINE, 0 representing the least conserved alignment position colored in blue up to 10 marked by an asterix in red. Below the Consistency row the blue and red colored blocks stand for the representative secondary structure prediction using DSSP and PSIPRED. The *β*-Strand predictions are colored in blue and the red colored blocks are the *α*-Helix predictions. The predicted putative core peptide sequence is indicated by a black frame

of *A. flavus*. This was expected since the underlying search for fungal RiPP clusters in antiSMASH ver. 5.0 [16] is based on the ustiloxin B cluster from *A. flavus*. In contrast, the search for BGCs in the *Trichoderma* spp. and the *A. phalloides* genomes yielded the same results using the last two versions of antiSMASH (ver. 4.3.0 [22] and ver. 5.0 [16]) and no predictions of fungal RiPP clusters. Our unconventional approach found a total of 615 potential RiPP precursor peptides in the 4 mined *Trichoderma* genomes. Notably, the results from our approach were obtained by using tools designed for bacterial sequences. This procedure would strongly benefit from a database of fungal RiPPs that could be integrated in the RiPPMiner software. Consequently, these findings have to be carefully manually inspected and thereafter verified by RNA sequencing data to reduce false positives, as we did for the *T. reesei* results in this study. As we have shown for *T. reesei* after careful inspection of the results we could reduce our set of potential RiPP precursor peptides from 108 to 6.

One of these predicted putative RiPP precursor peptides is found in the *Trichoderma*-putative-RiPP clade, suggesting the existence of a conserved putative RiPP precursor peptide within the *Trichoderma* genus. Another putative novel fungal RiPP cluster in the *T. reesei* genome is cluster 55 (Fig. 6). Its precursor peptide sequence clustered in the amanitin/phalloidin clade together with a sequence from *T. citrinoviride*. Furthermore, the putative precursor peptide sequences within this clade all showed high similarities to the putative structural toxin protein of *E. lata* (UCREL1), the structural toxin protein RtxA of *A. oryzae* and structural toxin protein RtxA of *M. rileyi* in the Blastp output. The putative precursor peptide found in this cluster shows in the potential core sequence a predicted sulfatation site similar to the one found in the known fungal RiPP precursor peptide α-amanitin. (Fig. 7) Our results largely support the hypothesis that fungal genomes contain biosynthetic gene clusters for RiPPs that might be a vast untapped source for possible new lead compounds with yet undescribed potential applications. Further in vitro and in vivo investigations are needed to be able to predict a preliminary biosynthetic pathway for the described RiPP clusters and to definitively classify these six clusters found in silico as novel fungal RiPP clusters in *T. reesei*.

Conclusion

In this study we describe a novel, unconventional mining approach for the search for RiPPs in fungi. While this method offers new possibilities it also demands a rather long hands on time to refine the search, due to the lack of automatization. However, we could successfully find previously known fungal RiPPs and predict several putative novel RiPPs in the genus *Trichoderma*.

In the fight against the rising threat of multiresistant pathogenic strains, fungal RiPPs represent an indispensable new armament of possible diverse lead compounds. Our study is the first report of the potential of *Trichoderma* to produce RiPPs and might pave the way

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

for further studies on fungal RiPPs in *Trichoderma*. The method described in our study will lead to further mining efforts in all subdivisions of the fungal kingdom.

Methods

Extraction of RNA and sequencing

The RNASeq data used in this study was generated in a previous study by Derntl et al. [23]. Therein the wild-type like *T. reesei* strain QM6a Δ*tmus53* strain [34] cultivated in Mandels-Andreotti medium [35] containing 1% carboxy methyl cellulose as carbon source. After 48 h of solid-state incubation at 30 °C, the RNA was isolated using the RNeasy Plant Mini Kit (Quiagen) and libraries were prepared using a TruSeq Stranded mRNA Sample Prep Kit including poly (A) enrichment (Illumina). The libraries were sequenced on a NextSeq500 instrument (Illumina) with paired-end 75 nt long reads [23].

Full genomes

The genomes of *T. asperellum* CBS 433.97 (assembly Trias v. 1.0; BioSample accession: SAMN00769595), *T. virens* Gv29–8 (assembly TRIVI v2.0; BioSample accession: SAMN02744059), *T. arundinaceum* (assembly Trichoderma_arundinaceum_IBT40837_contigs; BioSample accession: SAMN06320351), *T. reesei* QM6a (assembly v2.0; BioSample accession: SAMN02746107), *T. citrinoviride* (assembly Trici v4.0; BioSample accession: SAMN05369575), *T. harzianum* CSB 226.95 (assembly Triha v1.0; BioSample accession: SAMN00761861), *T. atrobrunneum* (assembly ASM343991v1; BioSample accession: SAMN08325511), *T. brevicompactum* (assembly Trichoderma_brevicompactum_IBT40841_contigs; BioSample accession: SAMN06320626) and *T. koningii* (assembly JCM_1883_assembly_v001; BioSample accession: SAMD00028335) were downloaded from the NCBI database. Furthermore, the genomes of *Amanita phalloides* (assembly ASM198338v1; BioSample accession: SAMN05444494) and *Aspergillus flavus* NRRL3357 (assembly JCVI-afl1-v2.0; BioSample accession: SAMN05591370) were downloaded from the NCBI database, to evaluate our mining procedure.

Genome mining

The command line version of antiSMASH ver. 4.3.0 [22] was used to mine the selected genomes for secondary metabolite biosynthetic gene clusters with following specifications in order to yield the best results for the fungal genomes. The taxon was specified with the option --taxon to be of fungal origin, the --clusterblast, –-subclusterblast and --knownclusterblast options were used to compare the identified clusters against a database of antiSMASH-predicted clusters, known subclusters that synthesize precursors and known gene clusters from the MIBiG database [36] respectively. The --smcogs option enables a search for BGCs of orthologous SM groups. Furthermore, the ClusterFinder algorithm was activated with the --inclusive option for additive cluster discovery. In parallel a genome wide HMMer analysis was performed by specifying the --full-hmmer option and the active site finder module with the --asf option. The results for *T. reesei*, *A. flavus* and *A. phalloides* were then cross referenced with the online version of antiSMASH ver. 5.0 [16] that includes the identification of fungal RiPP clusters.

To verify the presence of similar precursor peptides within the *Trichoderma* genus four full genomes were chosen based on their average nucleotide identity (ANI) calculated with a fast alignment-free implementation for computing whole-genome ANI between genomes called fastANI [37]. The choice which *Trichoderma* spp. were to be mined for RiPPs was based on their average nucleotide identity (ANI). Within the putative clusters, when applying an 85% ANI cutoff, of the chosen genomes the amino acid sequences of the genes classified as “other genes” were extracted and concatenated in a single file. Core biosynthetic genes, additional biosynthetic genes, transport-related genes and regulatory genes were not included. The extracted sequences were then analyzed using the standalone version of RiPPMiner [24] to predict possible RiPPs within the genomes. The method of RiPP prediction was tested on known precursor peptides of fungal RiPPs extracted from the UniProt database, namely α-amanitin (A8W7M4), β-amanitin (ABW87785), phallacidin (ABW87771) and phalloidin (ABW87787) and 75 diverse known precursor peptides (Additional file 1).

All extracted amino acid sequences, that were predicted as putative RiPP precursor peptides by the RiPPMiner software, were blasted using Blastp v2.9.0+ (protein-protein BLAST) [25] against the non-redundant protein database (All non-redundant GenBank CDS translations, PDB, SwissProt, PIR, PRF excluding environmental samples from WGS projects) to refine the search and a manually curated database (e.g. KEPs could be identified and removed). Sequences with highly conserved active domains were removed from the total set, as well as classified sequences such as transcription factors, enzymes and ribosomal proteins. Only hypothetical proteins, small secreted cysteine rich proteins of unknown function (SSCRP) and sequences without considerable similarities were kept. The refined putative RiPP precursor peptides and the known precursor peptides of fungal RiPPs as reference were aligned with MUSCLE and a Nearest-Neighbor-Interchange (NNI) tree with 100 Bootstraps using the Jones-Taylor-Thornton (JTT) model was inferred by using the maximum likelihood method and Dayhoffw/freq. Model. The analysis was conducted with the MEGA X software platform [27].

Further analysis, visualizations and exploratory data analysis were carried out in R v3.6.0 [38] with the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

following packages: phangorn v2.5.4 [39]; ape v5.3 [40]; ggplot2 v3.1.1 [41]; ggtree v1.17.1 [42]; gplots v3.0.1.1 [43]; stats v3.6.0.

Transcriptome analysis

The raw RNASeq paired-end reads were aligned to the *Trichoderma reesei* QM6a genome (assembly v2.0; BioSample accession: SAMN02746107) without using predefined annotations. This was done following the protocol for mapping RNASeq reads with a 2-pass procedure described by Dobin and Gingeras with the software STAR v 2.7.0c [44]. The alignments were visualized with IGV v2.5.3 (Integrative Genomics Viewer) [45]. This procedure was chosen to reduce false positive putative precursor peptide gene calls. Putative precursor peptide genes to which RNASeq data aligned were considered true positives. A schematic diagram depicting the overall scheme of the pipeline used to discover and curate possible RiPP precursor peptides is illustrated in Fig. 3.

RiPP precursor peptide analysis

The most highly likely putative RiPP precursor peptide from *T. reesei* was aligned to the 20 top hits from the Blastp output using the ClustalW algorithm and was performed with PRALINE [28]. Furthermore, the peptide sequence was analyzed with NetOGlyc 4.0 [29] to predict O glycosylation sites, NetNGlyc 1.0 [30] to find possible N glycosylation sites, NsitePred [31] to evaluate if there are probable DNA motif binding sites, NMT [46] was used to recognize glycine N-myristoylation sites of fungi and to detect if the C-terminus is appropriate for peroxisomal import, NetPhos 3.1 [33] to predict phosphorylation sites and ExPASy – Sulfinator [32] to find appropriate sulfatation sites within the peptide. Furthermore, the conservation scoring was performed with PRALINE and the secondary structure prediction was performed using the Define Secondary Structure of Proteins (DSSP) algorithm and PSI-blast based secondary structure PREDiction (PSIPRED).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-6653-6.

Additional file 1. The output of the RiPPMiner prediction for known precursor peptides of fungal RiPPs extracted from the UniProt database.

Additional file 2. The output of the RiPPMiner prediction for the Ustiloxin B cluster extracted from the antiSMASH output.

Additional file 3. Full maximum likelihood (ML) phylogenetic tree. The ML phylogenetic tree was inferred based on 434 amino acid sequences.

Additional file 4. The extracted sub-trees including the putative RiPP precursor peptides from *T. reesei* and those including known precursor peptides of fungal RiPPs extracted from the UniProt database.

Additional file 5. Blastp v2.9.0+ output for Cluster 55, containing the predicted RiPP precursors with similarities to amatoxins.

Additional file 6. Comparison of biosynthetic gene cluster 55 of *T. reesei* and biosynthetic gene cluster 75 of *T. citrinoviride.*

Additional file 7. Blastp v2.9.0+ output for cluster 75 of *T. citrinoviride*, containing the predicted RiPP precursor.

Additional file 8. Gene annotations for cluster 75 of *T. citrinoviride*, manually curated and based on a Blastp v2.9.0+ (protein-protein BLAST) [25] search against a manually curated database, including protein accessions.

Abbreviations

ANI: Average nucleotide identity; BGC: Biosynthetic gene cluster; DSSP: Define Secondary Structure of Proteins; JTT: Jones-Taylor-Thornton; KEP: KEX2-processed repeat proteins; KEX2: Fungal Kexin-like proteinases (killer expression); ML: Maximum likelihood; NNI: Nearest-Neighbor-Interchange; NRPS: Non-ribosomal peptide synthetase; PSIPRED: PSI-blast based secondary structure PREDiction; RiPPs: Ribosomally synthesized and post-translationally modified peptides; SM: Secondary metabolite; SSCRP: Small secreted cysteine rich proteins; T1pks: Type I Polyketide synthase

Acknowledgements

Not applicable.

Authors' contribution

GAV and CD conceptualized, designed the study, and drafted the manuscript. GAV generated the datasets, performed the data analysis, generated the figures. ARMA and RLM revised and contributed to the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by a grant from the Austrian Science Fund (FWF): P29556 given to RM, and the PhD program TU Wien bioactive.

Availability of data and materials

The datasets analyzed during the current study are publicly available in the NCBI National Center for Biotechnology Information repository, https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/. The datasets supporting the conclusions of this article are included within the article and its additional files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 29 November 2019 Accepted: 4 March 2020
Published online: 27 March 2020

References

1. Hoffmeister D, Keller NP. Natural products of filamentous fungi: enzymes, genes, and their regulation. Nat Prod Rep. 2007;24(2):393–416.
2. Bassett EJ, Keith MS, Armelagos GJ, Martin DL, Villanueva AR. Tetracycline-labeled human bone from ancient Sudanese Nubia (a.D. 350). Science. 1980; 209(4464):1532–4.
3. Alberti F, Foster GD, Bailey AM. Natural products from filamentous fungi and production by heterologous expression. Appl Microbiol Biotechnol. 2017; 101(2):493–500.
4. Gomez BL, Nosanchuk JD. Melanin and fungi. Curr Opin Infect Dis. 2003; 16(2):91–6.
5. Wheeler MH, Bell AA. Melanins and their importance in pathogenic fungi. Curr Top Med Mycol. 1988;2:338–87.
6. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, et al. Ribosomally synthesized and post-translationally modified peptide natural

products: overview and recommendations for a universal nomenclature. Nat Prod Rep. 2013;30(1):108–60.

7. Luo S, Dong SH. Recent Advances in the Discovery and Biosynthetic Study of Eukaryotic RiPP Natural Products. Molecules. 2019;24:8.
8. Tsomaia N. Peptide therapeutics: targeting the undruggable space. Eur J Med Chem. 2015;94:459–70.
9. Abdalla MA, McGaw LJ. Natural Cyclic Peptides as an Attractive Modality for Therapeutics: A Mini Review. Molecules. 2018;23:8.
10. Le Marquer M, San Clemente H, Roux C, Savelli B, Freidit Frey N. Identification of new signalling peptides through a genome-wide survey of 250 fungal secretomes. BMC Genomics. 2019;20:64.
11. Hetrick KJ, van der Donk WA. Ribosomally synthesized and post-translationally modified peptide natural product discovery in the genomic era. Curr Opin Chem Biol. 2017;38:36–44.
12. Kubicek CP, Steindorff AS, Chenthamara K, Manganiello G, Henrissat B, Zhang J, et al. Evolution and comparative genomics of the most common *Trichoderma* species. BMC Genomics. 2019;20(1):485.
13. Nierman WC, Yu J, Fedorova-Abrams ND, Losada L, Cleveland TE, Bhatnagar D, et al. Genome Sequence of *Aspergillus flavus* NRRL 3357, a Strain That Causes Aflatoxin Contamination of Food and Feed. Genome Announc. 2015;3:2.
14. Theobald S, Vesth TC, Rendsvig JK, Nielsen KF, Riley R, de Abreu LM, et al. Uncovering secondary metabolite evolution and biosynthesis using gene cluster networks and genetic dereplication. Sci Rep. 2018;8(1):17957.
15. Hu D, Gao C, Sun C, Jin T, Fan G, Mok KM, et al. Genome-guided and mass spectrometry investigation of natural products produced by a potential new actinobacterial strain isolated from a mangrove ecosystem in Futian, Shenzhen, China. Sci Rep. 2019;9(1):823.
16. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47(W1):W81–W7.
17. Umemura M, Nagano N, Koike H, Kawano J, Ishii T, Miyamura Y, et al. Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in *Aspergillus flavus*. Fungal Genet Biol. 2014;68:23–30.
18. Nagano N, Umemura M, Izumikawa M, Kawano J, Ishii T, Kikuchi M, et al. Class of cyclic ribosomal peptide synthetic genes in filamentous fungi. Fungal Genet Biol. 2016;86:58–70.
19. Chaverri P, Branco-Rocha F, Jaklitsch W, Gazis R, Degenkolb T, Samuels GJ. Systematics of the *Trichoderma harzianum* species complex and the re-identification of commercial biocontrol strains. Mycologia. 2015;107(3):558–90.
20. Park Y-H, Chandra Mishra R, Yoon S, Kim H, Park C, Seo S-T, et al. Endophytic *Trichoderma citrinoviride* isolated from mountain-cultivated ginseng (Panax ginseng) has great potential as a biocontrol agent against ginseng pathogens. J Ginseng Res. 2019;43(3):408–20.
21. Shentu X, Zhan X, Ma Z, Yu X, Zhang C. Antifungal activity of metabolites of the endophytic fungus *Trichoderma brevicompactum* from garlic. Braz J Microbiol. 2014;45(1):248–54.
22. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45(W1):W36–41.
23. Derntl C, Kluger B, Bueschl C, Schuhmacher R, Mach RL, Mach-Aigner AR. Transcription factor Xpp1 is a switch between primary and secondary fungal metabolism. Proc Natl Acad Sci U S A. 2017;114(4):E560–E9.
24. Agrawal P, Khater S, Gupta M, Sain N, Mohanty D. RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. Nucleic Acids Res. 2017;45(W1): W80–W8.
25. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
26. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017;45(D1):D200–D3.
27. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547–9.
28. Pirovano W, Feenstra KA, Heringa J. PRALINE™: a strategy for improved multiple alignment of transmembrane proteins. Bioinformatics. 2008;24(4):492–7.
29. Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KTBG, et al. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J. 2013;32(10):1478–88.
30. Gupta R, Jung E, Brunak S. Prediction of N-glycosylation sites in human proteins, vol. 46; 2004. p. 203–6.
31. Chen K, Mizianty MJ, Kurgan L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. Bioinformatics. 2011;28(3):331–41.
32. Monigatti F, Gasteiger E, Bairoch A, Jung E. The Sulfinator: predicting tyrosine sulfation sites in protein sequences. Bioinformatics. 2002;18(5): 769–70.
33. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999;294(5):1351–62.
34. Steiger MG, Vitikainen M, Uskonen P, Brunner K, Adam G, Pakula T, et al. Transformation system for *Hypocrea jecorina* (*Trichoderma reesei*) that favors homologous integration and employs reusable bidirectionally selectable markers. Appl Environ Microbiol. 2011;77(1):114–21.
35. Mandels M. Applications of cellulases. Biochem Soc Trans. 1985;13(2):414–6.
36. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum information about a biosynthetic gene cluster. Nat Chem Biol. 2015;11(9):625–31.
37. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114.
38. Rc T. R: a language and environment for statistical computing; 2019.
39. Schliep K, Potts AJ, Morrison DA, Grimm GW. Interwining phylogenetic trees and networks. Methods Ecol Evol. 2017;8:1212–20.
40. PES K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018;35:526–8.
41. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.
42. Yu G, Smith D, Zhu H, Guan Y, Tsan-Yuk T. Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8(1):28–36.
43. Warnes G., Bolker B, Bonebakker L, Gentleman R, Liaw WH, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data. 2019.
44. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 2015;51:1–9.
45. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
46. Eisenhaber F, Eisenhaber B, Kubina W, Maurer-Stroh S, Neuberger G, Schneider G, et al. Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-pi, NMT and PTS1. Nucleic Acids Res. 2003;31(13):3631–4.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 1

```
#INPUT		1		sp|P85421|AAMAT_AMAPH Alpha-amanitin proprotein (Fragment)
OS=Amanita phalloides OX=67723 PE=1 SV=2

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT		2		sp|P0CU56|ANT_AMAPH Antamanide OS=Amanita phalloides OX=67723
PE=1 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT		3		sp|A0A067SLB9|AAMA1_GALM3 Alpha-amanitin proprotein 1 OS=Galerina
marginata (strain CBS 339.88) OX=685588 GN=AMA1-1 PE=1 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT		4		sp|H2E7Q6|AAMA2_GALM3 Alpha-amanitin proprotein 2 OS=Galerina
marginata (strain CBS 339.88) OX=685588 GN=AMA1-2 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT		5		sp|P0CU60|CYAD_AMAPH Cycloamanide D OS=Amanita phalloides
OX=67723 PE=1 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT		6		sp|P0CU59|CYAC_AMAPH Cycloamanide C OS=Amanita phalloides
OX=67723 PE=1 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

```
#INPUT          7          sp|A8W7M4|AAMAT_AMABI Alpha-amanitin proprotein OS=Amanita
bisporigera OX=87325 GN=AMA1 PE=3 SV=1

Predicted RiPP Class:    Lassopeptide

MODEL          1
Cleavage Site:   14
Leader Peptide: MSDINATRLPIWGI
Core Peptide:    GCNPCVGDDVTTLLTRGEALC
Predicted Crosslinks:     1,8,(Gly-Asp);2,21,(Cys-Cys);5,0,(Cys-Cys);
SMILES
          N3CC(=O)NC(CS4)C(=O)NC(CC(=O)N)C(=O)N(CCC1)C1C(=O)NC(CS5)C(=O)NC(C(C)C)C
(=O)NCC(=O)NC(CC3(=O))C(=O)NC(CC(=O)O)C(=O)NC(C(C)C)C(=O)NC(C(C)O)C(=O)NC(C(C)O
)C(=O)NC(CC(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=O)NCC(=O)NC
(C(CC(=O)O))C(=O)NC(C)C(=O)NC(CC(C)C)C(=O)NC(CS4)C(=O)O

MODEL          2
Cleavage Site:   14
Leader Peptide: MSDINATRLPIWGI
Core Peptide:    GCNPCVGDDVTTLLTRGEALC
Predicted Crosslinks:     1,9,(Gly-Asp);2,21,(Cys-Cys);5,0,(Cys-Cys);
SMILES
          N3CC(=O)NC(CS4)C(=O)NC(CC(=O)N)C(=O)N(CCC1)C1C(=O)NC(CS5)C(=O)NC(C(C)C)C
(=O)NCC(=O)NC(CC(=O)O)C(=O)NC(CC3(=O))C(=O)NC(C(C)C)C(=O)NC(C(C)O)C(=O)NC(C(C)O
)C(=O)NC(CC(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=O)NCC(=O)NC
(C(CC(=O)O))C(=O)NC(C)C(=O)NC(CC(C)C)C(=O)NC(CS4)C(=O)O

MODEL          3
Cleavage Site:   25
Leader Peptide: MSDINATRLPIWGIGCNPCVGDDVT
Core Peptide:    TLLTRGEALC
Predicted Crosslinks:     1,7,(Thr-Glu);
SMILES
          N3C(C(C)O)C(=O)NC(CC(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N
)C(=O)NCC(=O)NC(C(CC3(=O)))C(=O)NC(C)C(=O)NC(CC(C)C)C(=O)NC(CS)C(=O)O


#INPUT          8          sp|P0CU58|CYAB_AMAPH Cycloamanide B proprotein OS=Amanita
phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          9          sp|U5L406|AAMA1_AMAEX Alpha-amanitin proprotein 1 OS=Amanita
exitialis OX=262245 GN=AMA PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

```
#INPUT          10        sp|U5L3X2|AAMA2_AMAEX Alpha-amanitin proprotein 2 OS=Amanita
exitialis OX=262245 GN=AMA PE=2 SV=1

Predicted RiPP Class:     Lassopeptide

MODEL           1
Cleavage Site:   25
Leader Peptide: MSDINATRLPIWGIGCNPCVGDDVT
Core Peptide:    SVLTRGEA
Predicted Crosslinks:      1,7,(Ser-Glu);
SMILES
            N3C(CO)C(=O)NC(C(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=
O)NCC(=O)NC(C(CC3(=O)))C(=O)NC(C)C(=O)O

MODEL           2
Cleavage Site:   14
Leader Peptide: MSDINATRLPIWGI
Core Peptide:    GCNPCVGDDVTSVLTRGEA
Predicted Crosslinks:      1,8,(Gly-Asp);2,5,(Cys-Cys);
SMILES
            N3CC(=O)NC(CS4)C(=O)NC(CC(=O)N)C(=O)N(CCC1)C1C(=O)NC(CS4)C(=O)NC(C(C)C)C
(=O)NCC(=O)NC(CC3(=O))C(=O)NC(CC(=O)O)C(=O)NC(C(C)C)C(=O)NC(C(C)O)C(=O)NC(CO)C(
=O)NC(C(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=O)NCC(=O)NC(C(C
C(=O)O))C(=O)NC(C)C(=O)O

MODEL           3
Cleavage Site:   14
Leader Peptide: MSDINATRLPIWGI
Core Peptide:    GCNPCVGDDVTSVLTRGEA
Predicted Crosslinks:      1,9,(Gly-Asp);2,5,(Cys-Cys);
SMILES
            N3CC(=O)NC(CS4)C(=O)NC(CC(=O)N)C(=O)N(CCC1)C1C(=O)NC(CS4)C(=O)NC(C(C)C)C
(=O)NCC(=O)NC(CC(=O)O)C(=O)NC(CC3(=O))C(=O)NC(C(C)C)C(=O)NC(C(C)O)C(=O)NC(CO)C(
=O)NC(C(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=O)NCC(=O)NC(C(C
C(=O)O))C(=O)NC(C)C(=O)O


#INPUT          11        sp|U5L408|AAMA5_AMAEX Alpha-amanitin proprotein 5 OS=Amanita
exitialis OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:     NONE


#INPUT          12        sp|P0CU65|PHAD3_AMAPH Phalloidin proprotein OS=Amanita phalloides
OX=67723 PE=1 SV=1

The Input Peptide sequence is predicted as RiPP!
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

```
Predicted RiPP Class:    NONE


#INPUT          13        sp|P0CU57|CYAA_AMAPH Cycloamanide A OS=Amanita phalloides
OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          14        sp|A0A023IWM8|AAMA1_AMARI Alpha-amanitin proprotein OS=Amanita
rimosa OX=580330 GN=AMA PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          15        sp|A0A023IWK7|AAMAT_AMAPL Alpha-amanitin proprotein OS=Amanita
pallidorosea OX=1324310 GN=AMA PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          16        sp|U5L3K1|AMAN2_AMAEX Amanexitide proprotein 2 OS=Amanita exitialis
OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          17        sp|A0A023IWG4|AAMAT_AMAFU Alpha-amanitin proprotein OS=Amanita
fuliginea OX=67708 GN=AMA PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          18        sp|A8W7M7|PHAT1_AMABI Phallacidin proprotein 1 OS=Amanita bisporigera
OX=87325 GN=PHA1_1 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE
```

#INPUT 19 sp|U5L3J5|AMAN1_AMAEX Amanexitide proprotein 1 OS=Amanita exitialis OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 20 sp|P0CU64|PHAD2_AMAPH Phalloidin proprotein OS=Amanita phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 21 sp|P0CU63|PHAD1_AMAPH Phalloidin proprotein OS=Amanita phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 22 sp|A0A023UBX6|PHAT1_AMAEX Phallacidin proprotein 1 (Fragment) OS=Amanita exitialis OX=262245 GN=PHA3 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 23 sp|A0A023IWE3|AAMA1_AMAFL Alpha-amanitin proprotein OS=Amanita fuligineoides OX=580329 GN=AMA PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 24 sp|U5L397|PHAT2_AMAEX Phallacidin proprotein 2 OS=Amanita exitialis OX=262245 GN=PHA PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 25 sp|U5L3M7|BAMAT_AMAEX Beta-amanitin proprotein OS=Amanita exitialis OX=262245 PE=2 SV=1

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

```
The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          26        sp|A0A023UCA6|AAMA2_AMAFL Alpha-amanitin proprotein (Fragment)
OS=Amanita fuligineoides OX=580329 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          27        sp|D6CFW3|BAMA3_AMAPH Beta-amanitin proprotein OS=Amanita
phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          28        sp|A0A023IWE2|BAMAT_AMAPL Beta-amanitin proprotein OS=Amanita
pallidorosea OX=1324310 PE=3 SV=1

Predicted RiPP Class:    Lassopeptide

MODEL           1
Cleavage Site:  14
Leader Peptide: MSDINATRLPIWGI
Core Peptide:   GCDPCVGDDVTAVLTRGEA
Predicted Crosslinks:    1,8,(Gly-Asp);2,5,(Cys-Cys);
SMILES
          N3CC(=O)NC(CS4)C(=O)NC(CC(=O)O)C(=O)N(CCC1)C1C(=O)NC(CS4)C(=O)NC(C(C)C)C
(=O)NCC(=O)NC(CC3(=O))C(=O)NC(CC(=O)O)C(=O)NC(C(C)C)C(=O)NC(C(C)O)C(=O)NC(C)C(=
O)NC(C(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=O)NCC(=O)NC(C(C
C(=O)O))C(=O)NC(C)C(=O)O

MODEL           2
Cleavage Site:  14
Leader Peptide: MSDINATRLPIWGI
Core Peptide:   GCDPCVGDDVTAVLTRGEA
Predicted Crosslinks:    1,9,(Gly-Asp);2,5,(Cys-Cys);
SMILES
          N3CC(=O)NC(CS4)C(=O)NC(CC(=O)O)C(=O)N(CCC1)C1C(=O)NC(CS4)C(=O)NC(C(C)C)C
(=O)NCC(=O)NC(CC(=O)O)C(=O)NC(CC3(=O))C(=O)NC(C(C)C)C(=O)NC(C(C)O)C(=O)NC(C)C(=
O)NC(C(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=O)NCC(=O)NC(C(C
C(=O)O))C(=O)NC(C)C(=O)O

MODEL           3
Cleavage Site:  25
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

```
Leader Peptide: MSDINATRLPIWGIGCDPCVGDDVT
Core Peptide:    AVLTRGEA
Predicted Crosslinks:     1,7,(Ala-Glu);
SMILES
            N3C(C)C(=O)NC(C(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=O
)NCC(=O)NC(C(CC3(=O)))C(=O)NC(C)C(=O)O


#INPUT          29        sp|A8W7P1|BAMA2_AMAPH Beta-amanitin proprotein (Fragment)
OS=Amanita phalloides OX=67723 GN=AMA2 PE=3 SV=1

Predicted RiPP Class:     Lassopeptide

MODEL           1
Cleavage Site:   14
Leader Peptide: MSDINATRLPIWGI
Core Peptide:    GCDPCIGDDVTILLTRGE
Predicted Crosslinks:     1,8,(Gly-Asp);2,5,(Cys-Cys);
SMILES
            N3CC(=O)NC(CS4)C(=O)NC(CC(=O)O)C(=O)N(CCC1)C1C(=O)NC(CS4)C(=O)NC(C(C)CC)
C(=O)NCC(=O)NC(CC3(=O))C(=O)NC(CC(=O)O)C(=O)NC(C(C)C)C(=O)NC(C(C)O)C(=O)NC(C(C)
CC)C(=O)NC(CC(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=O)NCC(=O)
NC(C(CC(=O)O))C(=O)O

MODEL           2
Cleavage Site:   14
Leader Peptide: MSDINATRLPIWGI
Core Peptide:    GCDPCIGDDVTILLTRGE
Predicted Crosslinks:     1,9,(Gly-Asp);2,5,(Cys-Cys);
SMILES
            N3CC(=O)NC(CS4)C(=O)NC(CC(=O)O)C(=O)N(CCC1)C1C(=O)NC(CS4)C(=O)NC(C(C)CC)
C(=O)NCC(=O)NC(CC(=O)O)C(=O)NC(CC3(=O))C(=O)NC(C(C)C)C(=O)NC(C(C)O)C(=O)NC(C(C)
CC)C(=O)NC(CC(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)N)C(=O)NCC(=O)
NC(C(CC(=O)O))C(=O)O

MODEL           3
Cleavage Site:   25
Leader Peptide: MSDINATRLPIWGIGCDPCIGDDVT
Core Peptide:    ILLTRGE
Predicted Crosslinks:     1,7,(Ile-Glu);
SMILES
            N3C(C(C)CC)C(=O)NC(CC(C)C)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CCCN=C(N)
N)C(=O)NCC(=O)NC(C(CC3(=O)))C(=O)O


#INPUT          30        sp|A8W7N5|MSD7_AMABI MSDIN-like toxin proprotein 7 (Fragment)
OS=Amanita bisporigera OX=87325 GN=MSD7 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:     NONE
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

```
#INPUT	31	sp|U5L3J9|MSD8_AMAEX MSDIN-like toxin proprotein 8 OS=Amanita
exitialis OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT	32	sp|A0A023IWK3|MSD2_AMAFU MSDIN-like toxin proprotein 2 OS=Amanita
fuliginea OX=67708 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT	33	sp|A0A023IWE0|MSD1_AMAFL MSDIN-like toxin proprotein 1 OS=Amanita
fuligineoides OX=580329 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT	34	sp|U5L3X0|MSD1_AMAEX MSDIN-like toxin proprotein 1 OS=Amanita
exitialis OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT	35	sp|A0A023IWG1|MSD3_AMAFL MSDIN-like toxin proprotein 3 OS=Amanita
fuligineoides OX=580329 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT	36	sp|U5L409|MSD2_AMAEX MSDIN-like toxin proprotein 2 OS=Amanita
exitialis OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

```
#INPUT          37        sp|A8W7N4|MSD6_AMABI MSDIN-like toxin proprotein 6 OS=Amanita
bisporigera OX=87325 GN=MSD6 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          38        sp|A0A023IWI4|MSD2_AMAFL MSDIN-like toxin proprotein 2 OS=Amanita
fuligineoides OX=580329 PE=3 SV=1

Predicted RiPP Class:    Lassopeptide

MODEL          1
Cleavage Site:   24
Leader Peptide: MSDINATRLPHLVRYPPYVGDGTD
Core Peptide:    LTLNRGEK
Predicted Crosslinks:     1,7,(Leu-Glu);
SMILES
            N3C(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CC(C)C)C(=O)NC(CC(=O)N)C(=O)NC(CCCN=C(
N)N)C(=O)NCC(=O)NC(C(CC3(=O)))C(=O)NC(C(CCCN))C(=O)O

MODEL          2
Cleavage Site:   23
Leader Peptide: MSDINATRLPHLVRYPPYVGDGT
Core Peptide:    DLTLNRGEK
Predicted Crosslinks:     1,8,(Asp-Glu);
SMILES
            N3C(CC(=O)O)C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CC(C)C)C(=O)NC(CC(=O)N)
C(=O)NC(CCCN=C(N)N)C(=O)NCC(=O)NC(C(CC3(=O)))C(=O)NC(C(CCCN))C(=O)O

MODEL          3
Cleavage Site:   14
Leader Peptide: MSDINATRLPHLVR
Core Peptide:    YPPYVGDGTDLTLNRGEK
Predicted Crosslinks:     1,7,(Tyr-Asp);
SMILES
            N3C(CC1=C(C=C(O)C=C1))C(=O)N(CCC1)C1C(=O)N(CCC1)C1C(=O)NC(CC1=C(C=C(O)C
=C1))C(=O)NC(C(C)C)C(=O)NCC(=O)NC(CC3(=O))C(=O)NCC(=O)NC(C(C)O)C(=O)NC(CC(=O)O)
C(=O)NC(CC(C)C)C(=O)NC(C(C)O)C(=O)NC(CC(C)C)C(=O)NC(CC(=O)N)C(=O)NC(CCCN=C(N)N
)C(=O)NCC(=O)NC(C(CC(=O)O))C(=O)NC(C(CCCN))C(=O)O


#INPUT          39        sp|A0A023IWE1|MSD4_AMAPH MSDIN-like toxin proprotein 4 OS=Amanita
phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

#INPUT 40 sp|A0A023IWG3|BAMAT_AMAFL Beta-amanitin proprotein OS=Amanita fuligineoides OX=580329 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 41 sp|A0A023IWM6|MSD1_AMARI MSDIN-like toxin proprotein 1 OS=Amanita rimosa OX=580330 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 42 sp|A0A023IWM4|MSD1_AMAFU MSDIN-like toxin proprotein 1 OS=Amanita fuliginea OX=67708 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 43 sp|A8W7P0|MSD12_AMABI MSDIN-like toxin proprotein 12 OS=Amanita bisporigera OX=87325 GN=MSD12 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 44 sp|A0A023IWK5|MSD2_AMARI MSDIN-like toxin proprotein 2 OS=Amanita rimosa OX=580330 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 45 sp|A8W7N1|MSD3_AMABI MSDIN-like toxin proprotein 3 (Fragment) OS=Amanita bisporigera OX=87325 GN=MSD3 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 46 sp|A8W7N2|MSD4_AMABI MSDIN-like toxin proprotein 4 (Fragment) OS=Amanita bisporigera OX=87325 GN=MSD4 PE=3 SV=1

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

```
The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          47        sp|U5L3J7|MSD7_AMAEX MSDIN-like toxin proprotein 7 OS=Amanita
exitialis OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          48        sp|U5L396|MSD6_AMAEX MSDIN-like toxin proprotein 6 OS=Amanita
exitialis OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          49        sp|U5L3M6|MSD5_AMAEX MSDIN-like toxin proprotein 5 OS=Amanita
exitialis OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          50        sp|A8W7N0|MSD2_AMABI MSDIN-like toxin proprotein 2 (Fragment)
OS=Amanita bisporigera OX=87325 GN=MSD2 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          51        sp|U5L3M8|MSD3_AMAEX MSDIN-like toxin proprotein 3 OS=Amanita
exitialis OX=262245 PE=2 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          52        sp|A8W7N6|MSD8_AMABI MSDIN-like toxin proprotein 8 (Fragment)
OS=Amanita bisporigera OX=87325 GN=MSD8 PE=3 SV=2

Predicted RiPP Class:    Lassopeptide

MODEL           1
```

```
Cleavage Site:   15
Leader Peptide: MSDINTARLPCIGFL
Core Peptide:    GIPSVGDDIEMVLRHG
Predicted Crosslinks:     1,7,(Gly-Asp);
SMILES
        N3CC(=O)NC(C(C)CC)C(=O)N(CCC1)C1C(=O)NC(CO)C(=O)NC(C(C)C)C(=O)NCC(=O)NC
(CC3(=O))C(=O)NC(CC(=O)O)C(=O)NC(C(C)CC)C(=O)NC(C(CC(=O)O))C(=O)NC(CCSC)C(=O)NC
(C(C)C)C(=O)NC(CC(C)C)C(=O)NC(CCCN=C(N)N)C(=O)NC(CC1=C(NC=N1))C(=O)NCC(=O)O

MODEL          2
Cleavage Site:   15
Leader Peptide: MSDINTARLPCIGFL
Core Peptide:    GIPSVGDDIEMVLRHG
Predicted Crosslinks:     1,8,(Gly-Asp);
SMILES
        N3CC(=O)NC(C(C)CC)C(=O)N(CCC1)C1C(=O)NC(CO)C(=O)NC(C(C)C)C(=O)NCC(=O)NC
(CC(=O)O)C(=O)NC(CC3(=O))C(=O)NC(C(C)CC)C(=O)NC(C(CC(=O)O))C(=O)NC(CCSC)C(=O)NC
(C(C)C)C(=O)NC(CC(C)C)C(=O)NC(CCCN=C(N)N)C(=O)NC(CC1=C(NC=N1))C(=O)NCC(=O)O

MODEL          3
Cleavage Site:   17
Leader Peptide: MSDINTARLPCIGFLGI
Core Peptide:    PSVGDDIEMVLRHG
Predicted Crosslinks:     1,8,(Pro-Glu);
SMILES
        N3(CCC1)C1C(=O)NC(CO)C(=O)NC(C(C)C)C(=O)NCC(=O)NC(CC(=O)O)C(=O)NC(CC(=O
)O)C(=O)NC(C(C)CC)C(=O)NC(C(CC3(=O)))C(=O)NC(CCSC)C(=O)NC(C(C)C)C(=O)NC(CC(C)C)C
(=O)NC(CCCN=C(N)N)C(=O)NC(CC1=C(NC=N1))C(=O)NCC(=O)O


#INPUT          53        sp|A8W7N8|MSD10_AMABI MSDIN-like toxin proprotein 10 OS=Amanita
bisporigera OX=87325 GN=MSD10 PE=3 SV=2

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:     NONE


#INPUT          54        sp|A0A023IWK4|MSD3_AMAPH MSDIN-like toxin proprotein 3 OS=Amanita
phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:     NONE


#INPUT          55        sp|A8W7N9|MSD11_AMABI MSDIN-like toxin proprotein 11 (Fragment)
OS=Amanita bisporigera OX=87325 GN=MSD11 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Predicted RiPP Class: NONE

#INPUT 56 sp|A8W7P2|MSD1_AMAPH MSDIN-like toxin proprotein a (Fragment) OS=Amanita phalloides OX=67723 GN=MSDa PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 57 sp|A8W7N3|MSD5_AMABI MSDIN-like toxin proprotein 5 (Fragment) OS=Amanita bisporigera OX=87325 GN=MSD5 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 58 sp|A8W7N7|MSD9_AMABI MSDIN-like toxin proprotein 8 (Fragment) OS=Amanita bisporigera OX=87325 GN=MSD9 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 59 sp|A0A023IWG2|MSD6_AMAPH MSDIN-like toxin proprotein 6 OS=Amanita phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 60 sp|A8W7M9|MSD1_AMABI MSDIN-like toxin proprotein 1 (Fragment) OS=Amanita bisporigera OX=87325 GN=MSD1 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

#INPUT 61 sp|A0A023IWI5|MSD5_AMAPH MSDIN-like toxin proprotein 5 OS=Amanita phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class: NONE

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

```
#INPUT          62        sp|A0A023IWM5|MSD2_AMAPH MSDIN-like toxin proprotein 2 OS=Amanita
phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          63        sp|A0A023IWI6|BAMAT_AMAFU Beta-amanitin proprotein OS=Amanita
fuliginea OX=67708 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          64        sp|A0A023UA23|PHAT_AMAFU Phallacidin proprotein (Fragment)
OS=Amanita fuliginea OX=67708 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          65        sp|A0A023UCC1|PHAT_AMAFL Phallacidin proprotein (Fragment)
OS=Amanita fuligineoides OX=580329 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          66        sp|A0A023IWM7|BAMAT_AMARI Beta-amanitin proprotein OS=Amanita
rimosa OX=580330 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          67        sp|A8W7P3|PHAT_AMAOC Phalloidin proprotein (Fragment) OS=Amanita
ocreata OX=235532 GN=PHD PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          68        sp|A0A023UBY3|PHAT_AMAPH Phallacidin proprotein 1 (Fragment)
OS=Amanita phalloides OX=67723 GN=PHA1 PE=3 SV=1
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

```
The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          69        sp|A0A023IWD9|MSD4_AMAEX MSDIN-like toxin proprotein 4 OS=Amanita
exitialis OX=262245 PE=2 SV=2

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          70        sp|A0A023IWK6|BAMA1_AMAPH Beta-amanitin proprotein OS=Amanita
phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          71        sp|A8W7M6|PHAT2_AMABI Phallacidin proprotein 1 (Fragment) OS=Amanita
bisporigera OX=87325 GN=PHA1_2 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          72        sp|A0A023UBA8|PHAT_AMARI Phallacidin proprotein (Fragment)
OS=Amanita rimosa OX=580330 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          73        sp|A0A023IWI8|PHAT_AMAPL Phallacidin proprotein OS=Amanita
pallidorosea OX=1324310 GN=PHA PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          74        sp|P0CU61|CYAE_AMAPH Cycloamanide E proprotein OS=Amanita
phalloides OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE
```

```
#INPUT          75      sp|P0CU62|CYAF_AMAPH Cycloamanide F proprotein OS=Amanita phalloides
OX=67723 PE=3 SV=1

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:   NONE
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

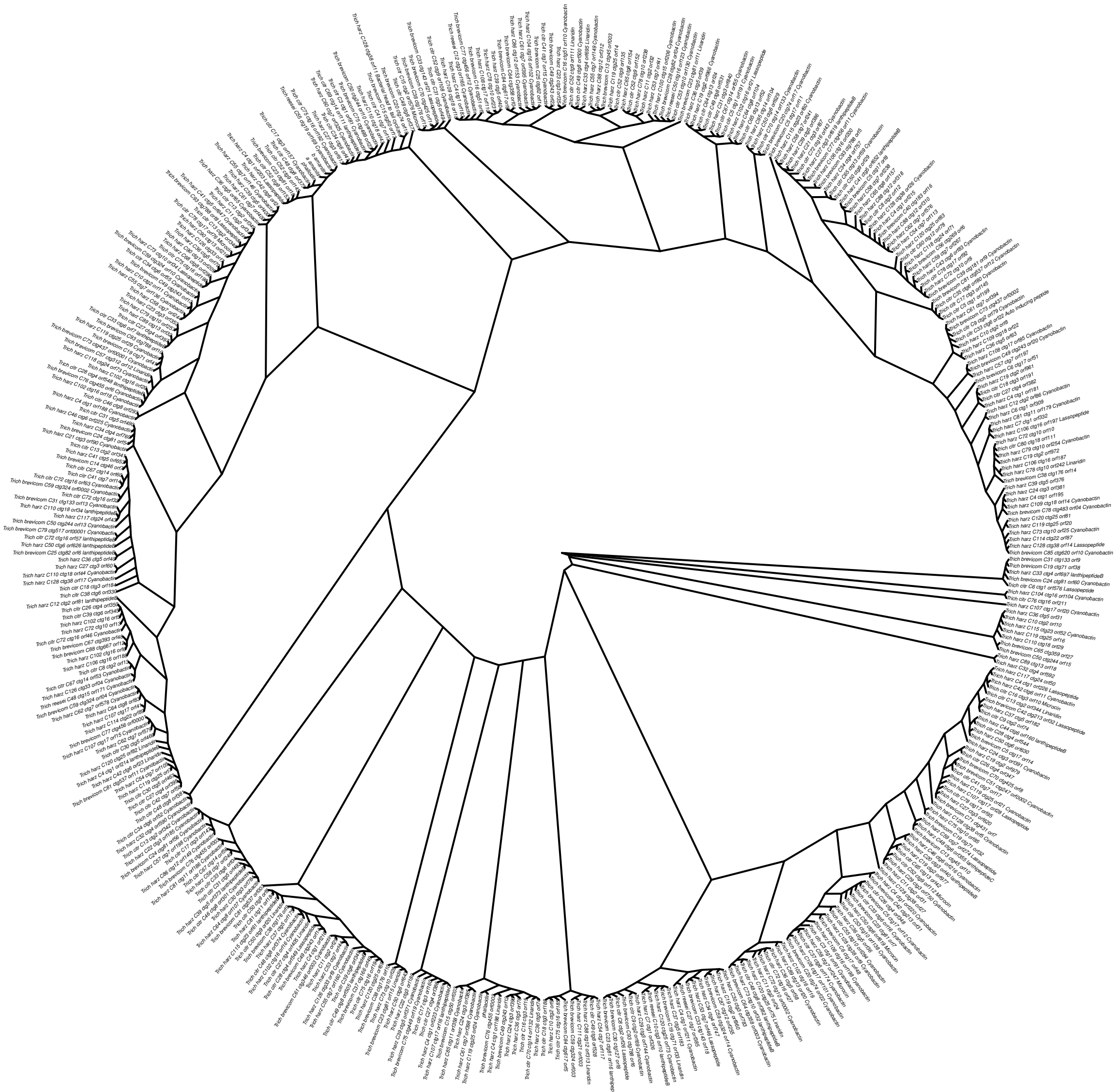
Additional File 2

```
#INPUT	1	AFLA_094900_1627:3671 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT	2	AFLA_094910_5519:7213 Reverse

The Input Peptide sequence is not predicted as RiPP!

#INPUT	3	AFLA_094920_8996:10049 Reverse

The Input Peptide sequence is not predicted as RiPP!

#INPUT	4	AFLA_094930_10952:11494 Forward

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT	5	AFLA_094940_11820:12633 Reverse

The Input Peptide sequence is not predicted as RiPP!

#INPUT	6	AFLA_094950_12983:14794 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT	7	AFLA_094960_15001:16869 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT	8	AFLA_094970_16928:17447 Reverse

The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:	NONE


#INPUT	9	AFLA_094980_17572:18341 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT	10	AFLA_094990_18786:19695 Reverse

The Input Peptide sequence is not predicted as RiPP!

#INPUT	11	AFLA_095000_19986:20251 Forward
```

```
The Input Peptide sequence is predicted as RiPP!

Predicted RiPP Class:    NONE


#INPUT          12        AFLA_095010_20853:22342 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT          13        AFLA_095020_22443:22987 Reverse

Predicted RiPP Class:    Cyanobactin


#INPUT          14        AFLA_095030_23859:25923 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT          15        AFLA_095040_26031:27483 Reverse

The Input Peptide sequence is not predicted as RiPP!

#INPUT          16        AFLA_095050_27748:29397 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT          17        AFLA_095060_29612:30961 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT          18        AFLA_095070_31085:32875 Reverse

The Input Peptide sequence is not predicted as RiPP!

#INPUT          19        AFLA_095080_35067:35960 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT          20        AFLA_095090_35976:36893 Forward

The Input Peptide sequence is not predicted as RiPP!

#INPUT          21        AFLA_095100_37200:38432 Reverse

The Input Peptide sequence is not predicted as RiPP!

#INPUT          22        AFLA_095110_38515:39246 Forward

The Input Peptide sequence is not predicted as RiPP!
```

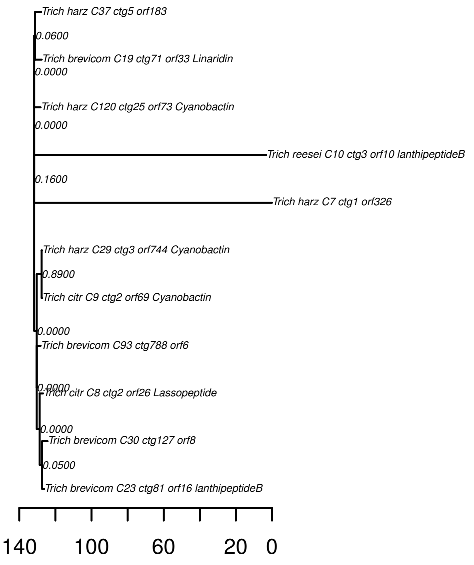
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

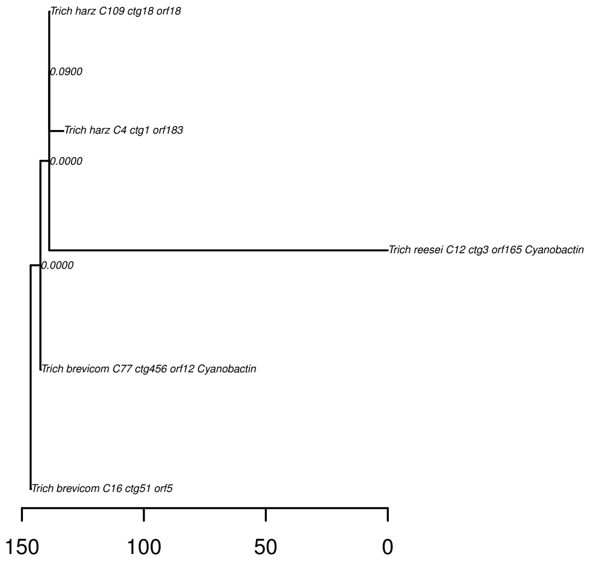
Additional File 3

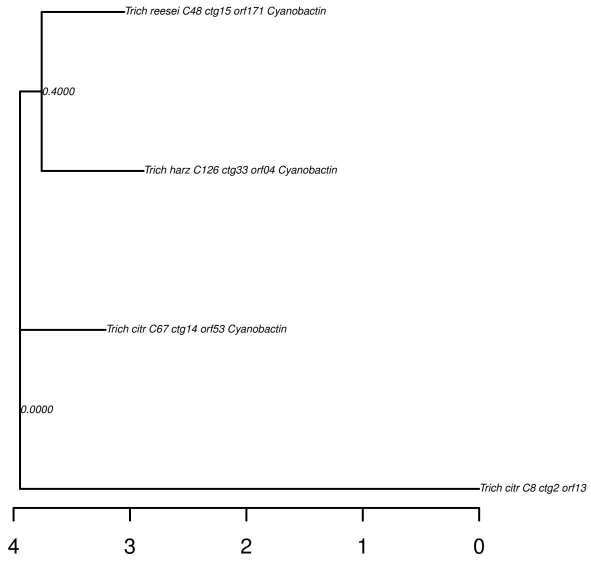


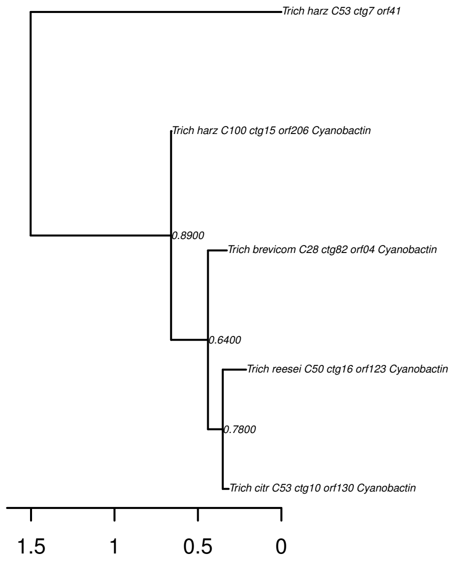
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

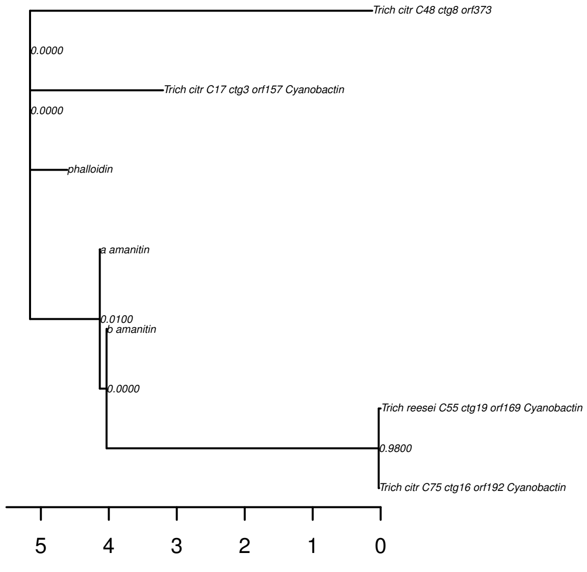
Additional File 4

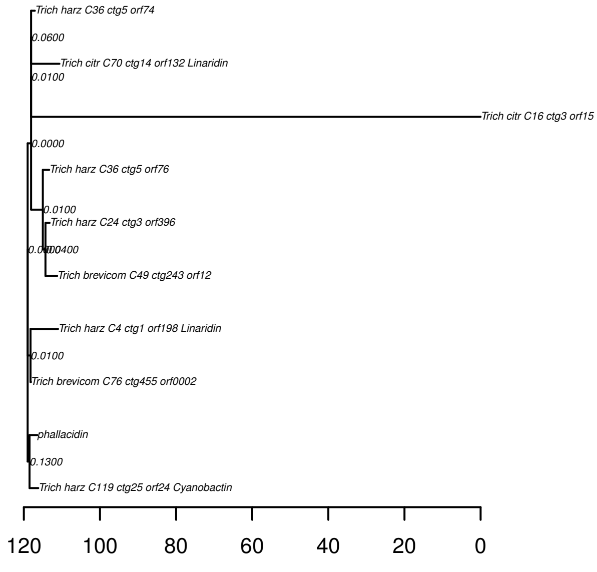












Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

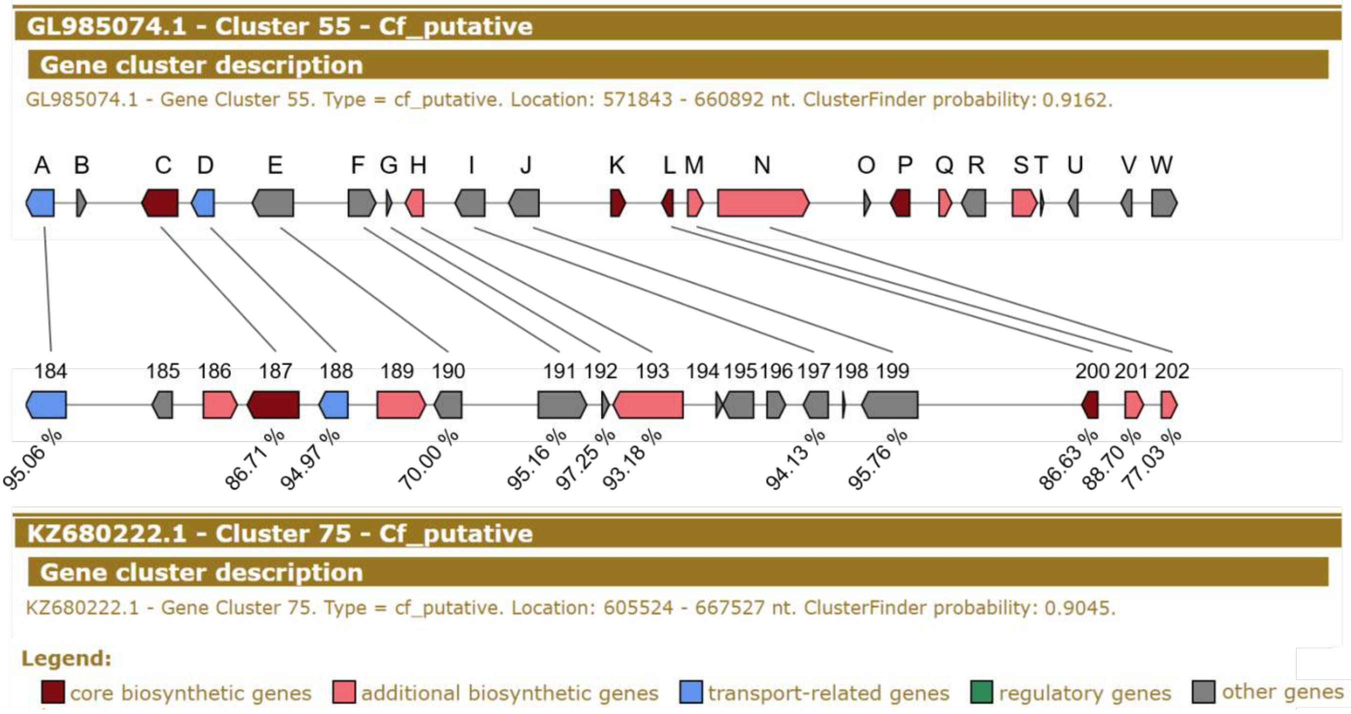
Additional File 5

This is a very large file and can be found online following this link:

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6653-6#Sec17

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 6



GL985074.1 - Cluster 55 - Cf_putative

Gene cluster description

GL985074.1 - Gene Cluster 55. Type = cf_putative. Location: 571843 - 660892 nt. ClusterFinder probability: 0.9162.

KZ680222.1 - Cluster 75 - Cf_putative

Gene cluster description

KZ680222.1 - Gene Cluster 75. Type = cf_putative. Location: 605524 - 667527 nt. ClusterFinder probability: 0.9045.

Additional file 6. Comparison of biosynthetic gene cluster 55 of *T. reesei* and biosynthetic gene cluster 75 of *T. citrinoviride.* The gene cluster of *T. citrinoviride* is located on scaffold KZ680222.1 (605524-667527 nt) and contains 16 predicted genes and three possible pseudogenes. The schematic representation of the two clusters was extracted directly from the antiSMASH results. The gene annotations were manually curated and based on a Blastp v2.9.0+ (protein-protein BLAST) (25) search against a manually curated database (Figure 6, Additional file 7). The gene designations of the *T. reesei* cluster 55 genes is the same as in Figure 6. The number of the open reading frames (orf) assigned by antiSMASH in the *T. citrinoviride* cluster 75 begin are indicated above the genes. All annotations and protein accession numbers for their corresponding orf can be found in Additional file 8. Orf 184 encodes a general substrate transporter, orf 185 a possible pseudogene, orf 186 a aldehyde dehydrogenase, orf 187 a glycosyltransferase family 1 protein, orf 188 encodes for a major facilitator superfamily (MFS) general substrate transporter, orf 189 a carbon-nitrogen hydrolase, orf 190 encodes for a Heterokaryon incompatibility protein, orf 191 a sulfatase, orf 192 the putative RiPP precursor peptide with the Location 636556 – 636948 nt, orf 193 a amino acid transporter, orf 194 a possible pseudogene, orf 195 a hypothetical protein, orf 196 a putative fungal transcription protein, orf 197 a carbohydrate-binding module family 1 protein, orf 198 a possible pseudogene, orf 199 a GMC oxidoreductase, orf 200 a alpha/beta-hydrolase, orf 201 a GroES-like protein and orf 202 encodes for a putative 3-hydroxyisobutyrate dehydrogenase.

The lines between genes of the two clusters indicate homology. The percentages beneath the *T. citrinoviride* genes represent the sequence similarities between the two homologous genes. Genes O – S, U W have homologs in *T. citrinoviride* (Additional file 5) at the corresponding location, but this is not depicted here, because antiSMASH did not predict these genes to be part of the BGC in *T. citrinoviride*.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Additional File 7

This is a very large file and can be found online following this link:

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6653-6#Sec17

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 8

Cluster 75 from T. citrinoviride KZ680222.1 Location 605524 - 667527

# orf 184
XP_024746026.1 general substrate transporter 95.06% sequence similarity with T. reesei

# orf 185
no hits

# orf 186
XP_024746029.1 aldehyde dehydrogenase [52.88% sequence similarity with T. reesei]

# orf 187
XP_024746030.1 glycosyltransferase family 1 protein 86.71% sequence similarity with T. reesei

# orf 188
XP_024746031.1 MFS general substrate transporter 94.97% sequence similarity with T. reesei

# orf 189
XP_024746033.1 carbon-nitrogen hydrolase [76.77% sequence similarity with T. reesei]

# orf 190
XP_024746034.1 Heterokaryon incompatibility protein 70.00% sequence similarity with T. reesei

# orf 191
XP_024746035.1 sulfatase 95.16% sequence similarity with T. reesei

# orf 192
XP_024746036.1 putative RiPP precursor peptide 97.25% sequence similarity with T. reesei Location: 636556 - 636948

# orf 193
XP_024746038.1 amino acid transporter 93.18% sequence similarity with T. reesei

# orf 194
no hits

# orf 195
XP_024746039.1 hypothetical protein [26.65% sequence similarity with T. reesei]

# orf 196

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

XP_024746040.1 hypothetical protein / fungal transcription protein [26.96% sequence similarity with T. reesei]

# orf 197
XP_024746041.1 carbohydrate-binding module family 1 protein 94.133% sequence similarity with T. reesei

# orf 198
no hits

# orf 199
XP_024746042.1 GMC oxidoreductase 95.760% sequence similarity with T. reesei

# orf 200
XP_024746046.1 alpha/beta-hydrolase 86.634% sequence similarity with T. reesei

# orf 201
XP_024746047.1 GroES-like protein 88.703% sequence similarity with T. reesei

# orf 202
XP_024746048.1 hypothetical protein / 3-hydroxyisobutyrate dehydrogenase 77.032% sequence similarity with T. reesei

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

PLOS COMPUTATIONAL BIOLOGY

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

RESEARCH ARTICLE

FunOrder: A robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution

Gabriel A. Vignolle, Denise Schaffer, Leopold Zehetner, Robert L. Mach, Astrid R. Mach-Aigner, Christian Derntl*

Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Vienna, Austria

* christian.derntl@tuwien.ac.at

Abstract

Secondary metabolites (SMs) are a vast group of compounds with different structures and properties that have been utilized as drugs, food additives, dyes, and as monomers for novel plastics. In many cases, the biosynthesis of SMs is catalysed by enzymes whose corresponding genes are co-localized in the genome in biosynthetic gene clusters (BGCs). Notably, BGCs may contain so-called gap genes, that are not involved in the biosynthesis of the SM. Current genome mining tools can identify BGCs, but they have problems with distinguishing essential genes from gap genes. This can and must be done by expensive, laborious, and time-consuming comparative genomic approaches or transcriptome analyses. In this study, we developed a method that allows semi-automated identification of essential genes in a BGC based on co-evolution analysis. To this end, the protein sequences of a BGC are blasted against a suitable proteome database. For each protein, a phylogenetic tree is created. The trees are compared by treeKO to detect co-evolution. The results of this comparison are visualized in different output formats, which are compared visually. Our results suggest that co-evolution is commonly occurring within BGCs, albeit not all, and that especially those genes that encode for enzymes of the biosynthetic pathway are co-evolutionary linked and can be identified with FunOrder. In light of the growing number of genomic data available, this will contribute to the studies of BGCs in native hosts and facilitate heterologous expression in other organisms with the aim of the discovery of novel SMs.

Author summary

The discovery and description of novel fungal secondary metabolites promises novel antibiotics, pharmaceuticals, and other useful compounds. A way to identify novel secondary metabolites is to express the corresponding genes in a suitable expression host. Consequently, a detailed knowledge or an accurate prediction of these genes is necessary. In fungi, the genes are co-localized in so-called biosynthetic gene clusters. Notably, the clusters may also contain genes that are not necessary for the biosynthesis of the secondary

OPEN ACCESS

Citation: Vignolle GA, Schaffer D, Zehetner L, Mach RL, Mach-Aigner AR, Derntl C (2021) FunOrder: A robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution. PLoS Comput Biol 17(9): e1009372. https://doi.org/10.1371/journal.pcbi.1009372

Editor: Samuel V. Scarpino, Northeastern University, UNITED STATES

Received: April 13, 2021

Accepted: August 23, 2021

Published: September 27, 2021

Copyright: © 2021 Vignolle et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The FunOrder tool, the relevant database, and the sequences and the FunOrder output of the negative control GCs and the positive control BGCs are available in the GitHub repository (https://github.com/gvignolle/FunOrder). We have also used Zenodo to assign a DOI to the repository: 10.5281/zenodo.5118984.

Funding: This study was supported by the Austrian Science Fund (FWF, https://www.fwf.ac.at/) [P 29556 to RM, P 34036 to CD] and TU Wien (https://www.tuwien.at/) [PhD program TU Wien bioactive]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

metabolites, so-called gap genes. We developed a method to detect co-evolved genes within the clusters and demonstrated that essential genes are co-evolving and can thus be differentiated from the gap genes. This adds an additional layer of information, which can support researchers with their decisions on which genes to study and express for the discovery of novel secondary metabolites.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Secondary metabolites (SMs) are a diverse group of compounds with a plethora of different chemical structures and properties which are found in all domains of life, but are predominantly studied in bacteria, fungi, and plants [1]. SMs are not necessary for the basic survival and growth of an organism but can be beneficial under certain conditions. For example, pigments help to sustain radiation, antibiotics help in competitive situations, and toxins can serve as defensive compounds or as virulence factors [2,3]. Notably, many SMs are used by humankind as drugs and pharmaceuticals, pigments and dyes, sweeteners and flavours, and most recently also as precursors for the synthesis of plastics [4]. The study of the secondary metabolism holds the promise for novel antibiotics, pharmaceuticals and other useful compounds [5].

A major hinderance in the discovery of yet undescribed SMs is the fact that most SMs are not produced under standard laboratory conditions, as they do not serve a purpose for the organisms then. Currently, different strategies are followed to circumvent this problem [6,7]. Untargeted approaches aim to induce the expression of any SM. To this end, biotic and abiotic stresses are applied, or global regulators and regulatory mechanisms are manipulated [8]. These strategies may lead to the discovery of novel compounds, whose corresponding genes have to be identified subsequently by time-consuming and expensive methods [7]. An extreme example are the aflatoxins, major food contaminants with serious toxicological effects [9]. It took over 40 years from the discovery of the aflatoxins as the causal agent of "turkey X" disease in the 1950s [10] until the corresponding genes were finally described in 1995 [11]. Targeted SM discovery approaches aim to induce the production of specific SMs by either overexpressing genes in the native host or by heterologous expression in another organism [12]. The targeted approaches, also called reverse strategy or bottom-up strategy allows a direct connection of SMs to the corresponding genes and does not rely on the inducibility of SM production in the native host. Inherently, the bottom-up approach is depending on modern genomics and accurate gene prediction tools [13].

In bacteria and fungi, the genes responsible for the biosynthesis of a certain SM are often co-localized in the genome, forming so called biosynthetic gene clusters (BGCs) [14,15]. The BGCs consists of one or more core genes, several tailoring enzymes, and genes involved in regulation and transport. As all these genes are essential for the production of a SM in the native host, we will refer to them as "essential genes" in this study. The core genes are responsible for assembling the basic chemical scaffold, which is further modified by the tailoring enzymes yielding the final SM [16]. We refer to the core genes and the tailoring genes as "biosynthetic genes" in this study. Depending on the class of the produced SM, the core genes differ. In fungi, the main SM classes are polyketides (e.g. the cholesterol-lowering drug lovastatin [17]

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

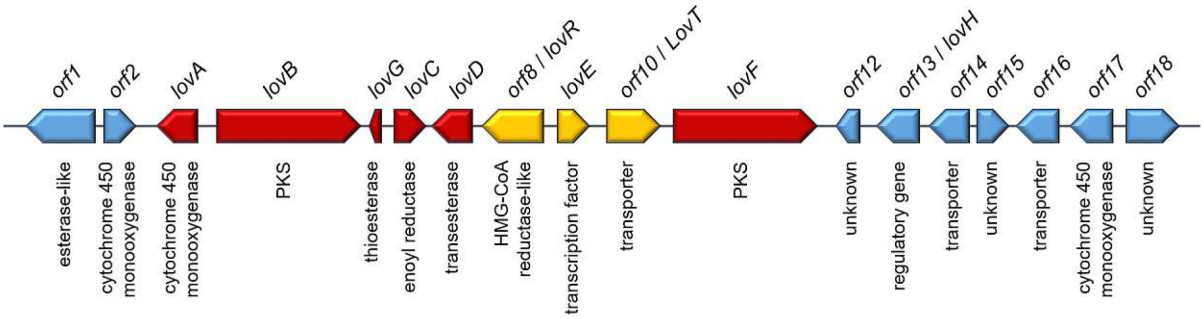


Fig 1. Schematic representation of the lovastatin BGC from *Aspergillus terreus* (lov). In red the biosynthetic genes for SM production, in gold the further essential genes, and in blue the genes not involved in the biosynthetic pathway.

https://doi.org/10.1371/journal.pcbi.1009372.g001

and the mycotoxin aflatoxin [9]) and non-ribosomal peptides (e.g. the immunosuppressant cyclosporine [18] and the antibiotic penicillin [19]), with polyketide synthases (PKS) or non-ribosomal peptide synthetases (NRPS) as core enzymes, respectively. Other SM classes are terpenoids, alkaloids, melanins [20,21], and ribosomally synthesized and posttranslationally modified peptides (RiPPs) [22,23], whose corresponding genes may also be organized in BGCs. As mentioned, BGCs may also contain genes encoding for transporters [24], transcription factors [25], or resistance genes [26]. While their gene products are not directly involved in the biosynthesis of a SM they are still essential for the biosynthesis; we will call them „further essential genes”in the following and differentiate them from the „biosynthetic genes“. The biosynthetic genes and the further essential genes are both necessary for the biosynthesis of a SM in the native organisms. In contrast, only the biosynthetic genes and a selection of the further essential genes (e.g. transporters) are necessary for heterologous expression [reviewed in [27]]. Notably, fungal BGCs often also contain genes that are not necessary for the production of a SM, the so-called gap genes. The gap genes are not involved in the biosynthesis, regulation, or transport of the SM, but have an unrelated function (Fig 1). We would like to stress here, that this cannot be predicted based only on the class of the gene product. For instance, a gene encoding for a transporter in the aflatoxin BGC was reported to have no significant role in aflatoxin secretion [28].

As mentioned, the bottom-up approach for SM discovery is depending on modern genomics and the accurate prediction of genes and BGCs. Each important gene missing in the prediction is detrimental for obvious reasons, whereas each unnecessarily considered gap gene makes the study of a BGC more complicated and complex, and the construction and transformation processes for heterologous expression more challenging. Currently, several BGC prediction tools are available for fungi. Some tools for genome mining are antiSMASH [29], CASSIS and SMIPS [30], SMURF [31], TOUCAN, a supervised learning framework capable of predicting BGCs on amino acid sequences [32], and DeepBGC, an unrestricted machine learning approach using deep neural networks [33]. These tools are effective and successful in finding and predicting BGCs based solely on genomic data. AntiSMASH uses a rule-based approach to identify BGCs based on the identification of core or signature enzymes and applies a greedy approach to extend a cluster on either side. This may result in overlaps or combinations of closely situated clusters. However, the genes within the predicted BGCs are classified into core biosynthetic genes, additional biosynthetic genes, transport-related genes, regulatory genes, and other genes based on profile hidden Markov models by the antiSMASH tool. The BGC prediction method of CASSIS and SMIPS is based on the principle that the promoter regions of genes in a BGC contain one or more shared motif, as they are co-expressed and presumably regulated by the same regulatory factors and/or mechanisms [30].

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

As mentioned above, the class of an enzyme may be a good indication for a potential involvement in the biosynthesis of a SM but does not guarantee a correct prediction. This problem can be solved by the analysis of transcriptome data because the genes necessary for SM production within a BGC are normally co-expressed with each other but not with the gap genes [34]. Notably, this demands the knowledge of expression conditions and does not work for silent BGCs. However, it is an obvious advantage to have as much information as possible about a BGC before studying it in the native host or performing heterologous expression for a bottom-up approach for SM discovery.

We speculate that a comparative genomics analysis focusing on the evolutionary history of the genes in a BGC might be a feasible alternative to a transcriptomics analysis in fungi for the following reasons. In general, BGCs are suggested to undergo a distinct and faster evolution than the rest of the genome, based on different mechanisms and genetic drivers [16,35–40]. In bacteria, the evolution of BGCs is strongly influenced by the strong occurrence of horizontal gene transfer in these group of microorganism [39]. Medema et al. performed a large-scale computational analysis of bacterial BGCs and found that many BGCs consist of sub-clusters. These sub-clusters encode for enzymes that work together to form a distinct chemical structure. Notably, this sub-clusters were described as "independent evolutionary entities" and the contained genes are co-evolving. The authors suggested a "bricks and mortar" model. Therein, different sub-clusters, the "bricks" form different chemical building blocks for a secondary metabolite. Additional genes within the BGCs are encoding for enzymes that combine the building blocks, and fulfil other functions such as tailoring, regulation and transport. These individual genes are the "mortar" in the "brick and mortar" model [40]. The "bricks" correspond to what we term "biosynthetic genes" and the "mortar" to our "further essential genes". Through horizontal gene transfer, the "bricks" can be easily exchanged and recombined to form novel BGCs and secondary metabolites[40]. Notably, not all bacterial BGCs are composed of exchangeable sub-units but some BGCs keep a stable architecture over a long time [40].

In fungi, three molecular evolutionary processes were suggested to be responsible for shaping the BGCs in a recent study, i.e., functional divergence, horizontal gene transfer, and *de novo* assembly [41]. Rokas et al. define functional divergence as a "process by which homologous BGCs, through the accumulation of genetic changes, gradually diverge in their functions changes" [41] and horizontal gene transfer as a "process by which an entire BGC from the genome of one organism is transferred and stably integrated into the genome of another through non-reproduction related mechanisms" [41]. This implies in both cases, that fungal BGCs are staying intact. Further, the genes are suggested to undergo a co-evolution which is faster than the rest of the genome [41]. Medema's "brick and mortar" model would more or less correspond to what Rokas et al. describe as "*de novo* assembly". This is defined as a "process by which an entire BGC is evolutionarily assembled through the recruitment and relocation of native genes, duplicates of native genes, and horizontally acquired genes" [41]. Notably, Rokas et al. state that this is the"least well-documented evolutionary process involved in the generation of fungal chemodiversity" [41], suggesting that in known and described fungal BGCs functional divergence and horizontal gene transfer are the two main evolutionary process, during which BGCs are staying intact and genes undergo a similar evolution. Further, we hypothesize that especially the biosynthetic genes in a BGC are co-evolutionary linked by the selection pressure to keep the biosynthetic pathway intact. Notably, a co-evolution analysis is a laborious and time-consuming task because a phylogenetic tree has to be calculated for each gene and then the trees compared to each other manually [42]. Recently, a method for the detection of co-evolution in bacterial BGCs was developed with the aim to identify sub-clusters [43]. That method is based on the detection of orthologous genes that are present in close

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

vicinity in many BGCs. This method is working unsupervised but requires a large set of BGCs as input [43].

In this study we describe a method (FunOrder) that allows a fast, semi-automated co-evolution analysis using individual BGCs as input. Based on this analysis and the assumption that the essential genes undergo a shared or similar evolution, FunOrder aims to identify essential genes in BGCs. To this end, we constructed a database of fungal proteomes as basis for the identification of co-evolutionary linked genes in ascomycetes. We determine the thresholds for the detection of co-evolution within different control gene sets. Then, we evaluated FunOrder and tested the underlying hypothesis, whether essential genes within a BGC could be identified based on the principle of co-evolution. We demonstrated the robustness and the applicability of the FunOrder method by analysing different control gene sets, including empirically validated BGCs and evaluated our method using stringent statistical tests.

Material and methods

Construction of a fungal proteome database

In this study we aim to identify co-evolutionary linked genes in ascomycetes. As the basis for the detection of co-evolution is a suitable database [42], we compiled an empirically optimized database consisting of 134 fungal proteomes from mainly ascomycetes and from two basidiomycetes for this method (Table 1). The two basidiomycete proteomes were included for the off chance of analysing gene clusters that do not originate from ascomycetes. The database covers the complete ascomycetes phylum and was iteratively tested and optimized for the detection of co-evolution in ascomycetes. The sequences were downloaded from the National Center for Biotechnology Information (NCBI) database and the Joint Genome Institute (JGI) [44]. A short identifier, unique in the database for each proteome, was introduced to enable multiple pairwise tree comparisons by the treeKO application [45]. A custom Perl script was used for removing duplicated entries in the database. The database is deposited in the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984).

Workflow

The workflow for the FunOrder method is depicted in Fig 2. First, the sequences of the BGC to be analysed are fed into the software bundle. FunOrder accepts a single file in either genbank file format or fasta format as input. The input files contain BGCs predicted by tools such as antiSMASH [29] or DeepBGC [33]. In case a genbank file is provided, a python script (Genbank to FASTA by Cedar McKay and Gabrielle Rocap, University of Washington) is called to extract the amino acid sequence of the genes in the BGC and create a fasta file. The multi-fasta file is then split into individual fasta files each containing a single protein sequence. These are placed in a subfolder created for the analysis of the BGC. Each file is named either after the position of the gene in the BGC or after the respective protein sequence description. This varies from the input file and the varying annotations used (If needed this can be changed in the script following the instructions of Genbank to FASTA by Cedar McKay and Gabrielle Rocap, University of Washington). Each header of the query sequences is tagged with the identifier "query" at the beginning of the header. The individual sequences are compared to the empirically optimized proteome database (Table 1) by a sequence similarity search using blastp 2.8.1+ (Protein-Protein BLAST) [133]. The output of this search is saved in a file with the ". tab" extension. Additionally, an optional remote search of the non-redundant National Center for Biotechnology Information (NCBI) protein database can be performed, yielding a file with the "ncbi.tab" extension. This allows a preliminary manual analysis of the input sequences and facilitates subsequent annotations of the BGCs.

Table 1. Fungal proteomes included in the empirically optimized database.

Organism	Source Database	Identifier	Reference
Acremonium chrysogenum	JGI	AcCh	[46]
Alternaria alternata	NCBI	AlAl	[47]
Alternaria arborescens	NCBI	AlAr	[48]
Alternaria gaisen	NCBI	AlGa	[49]
Alternaria sp. MG1	NCBI	AlSp	[50]
Alternaria tenuissima	NCBI	AlTe	[49]
Amanita muscaria	NCBI	AmMu	[51]
Amorphotheca resinae	JGI	AmRe	[52]
Arthrobotrys oligospora	JGI	ArOl	[53]
Arthroderma benhamiae	JGI	ArBe	[54]
Ascobolus immersus	JGI	AsIm	[55]
Aspergillus costaricaensis	NCBI	AsCo	[56]
Aspergillus fijiensis	NCBI	AsFi	[56]
Aspergillus flavus	NCBI	AsFl	[57]
Aspergillus fumigatus	NCBI	AsFu	[58]
Aspergillus homomorphus	NCBI	AsHo	[56]
Aspergillus ibericus	NCBI	AsIb	[56]
Aspergillus japonicus	NCBI	AsJa	[56]
Aspergillus niger	NCBI	AsNi	[59]
Aspergillus oryzae	NCBI	AsOr	[60]
Aspergillus phoenicis	NCBI	AsPh	[61]
Aspergillus terreus	NCBI	AsTe	[62]
Blumeria graminis	JGI	BlGr	[63]
Botryosphaeria dothidea	JGI	BoDo	[64]
Botrytis cinerea	NCBI	BoCi	[65]
Botrytis elliptica	NCBI	BoEl	[66]
Botrytis galanthina	NCBI	BoGa	[66]
Botrytis hyacinthi	NCBI	BoHy	[66]
Botrytis paeoniae	NCBI	BoPa	[66]
Botrytis porri	NCBI	BoPo	[66]
Botrytis tulipae	NCBI	BoTu	[66]
Cadophora sp.	JGI	CaSp	[67]
Capronia semiimmersa	JGI	CaSe	[68]
Chaetomium globosum	JGI	ChGl	[69]
Choiromyces venosus	JGI	ChVe	[55]
Cladonia grayi	JGI	ClGr	[70]
Cladophialophora bantiana	JGI	ClBa	[68]
Cladophialophora carrionii	JGI	ClCa	[68]
Cladophialophora immunda	JGI	ClIm	[68]
Cochliobolus heterostrophus	JGI	CoHe	[71]
Cochliobolus victoriae	JGI	CoVi	[72]
Colletotrichum nymphaeae	JGI	CoNy	[73]
Colletotrichum orchidophilum	JGI	CoOr	[74]
Colletotrichum salicis	JGI	CoSa	[73]
Colletotrichum simmondsii	JGI	CoSi	[73]
Colletotrichum tofieldiae	JGI	CoTo	[75]
Coniosporium apollinis	JGI	CoAp	[68]

(*Continued*)

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar. The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 1. (Continued)

Organism	Source Database	Identifier	Reference
Coniosporium apollinis CBS 100218	JGI	Capo	[68]
Corynespora cassiicola	JGI	CoCa	[76]
Daldinia eschscholzii	JGI	DaEs	[77]
Diaporthe ampelina	JGI	DiAm	[78]
Diplodia seriata	JGI	DiSe	[78]
Erysiphe necator	JGI	ErNe	[79]
Eutypa lata	NCBI	EuLa	[80]
Exophiala aquamarina	JGI	ExAq	[68]
Exophiala dermatitidis	JGI	ExDe	[68]
Exophiala oligosperma	JGI	ExOl	[68]
Exophiala spinifera	JGI	ExSp	[68]
Exophiala xenobiotica	JGI	ExXe	[68]
Fonsecaea monophora	JGI	FoMo	[81]
Fusarium fujikuroi	NCBI	FuFu	[82]
Fusarium graminearum	NCBI	FuGr	[83]
Fusarium oxysporum	NCBI	FuOx	[84]
Fusarium proliferatum	NCBI	FuPr	[85]
Fusarium pseudograminearum	NCBI	FuPs	[86]
Fusarium verticillioides	NCBI	FuVe	[83]
Gaeumannomyces graminis	JGI	GaGr	[87]
Glonium stellatum	JGI	GlSt	[88]
Hypoxylon sp. EC38	JGI	HyEC	[77]
*Hypoxylon sp.*CO27	JGI	Hysp	[77]
Magnaporthe grisea	JGI	MaGr	[89]
Magnaporthiopsis poae	JGI	MaPo	[87]
Meliniomyces bicolor	JGI	MeBi	[52]
Meliniomyces variabilis	JGI	MeVa	[52]
Metarhizium acridum	NCBI	MeAc	[90]
Metarhizium album	NCBI	MeAl	[91]
Metarhizium anisopliae	NCBI	MeAn	[91]
Metarhizium brunneum	NCBI	MeBr	[91]
Metarhizium guizhouense	NCBI	MeGu	[91]
Metarhizium majus	NCBI	MeMa	[91]
Metarhizium rileyi	NCBI	MeRi	[92]
Metarhizium robertsii	NCBI	MeRo	[90]
Monacrosporium haptotylum	JGI	MoHa	[93]
Morchella importuna	JGI	MoIm	[94]
[Nectria] haematococca	NCBI	NeHa	[95]
Nectria haematococca	JGI	NeHa	[95]
Neurospora crassa	JGI	NeCr2	[96]
Neurospora crassa FGSC	JGI	NeCr	[97]
Neurospora tetrasperma	JGI	NeTe	[98]
Oidiodendron maius	JGI	OiMa	[51]
Ophiostoma piceae	JGI	OpPi	[99]
Paecilomyces variotii	JGI	PaVa	[100]
Panaeolus cyanescens	NCBI	PaCy	[101]
Paracoccidioides brasiliensis	JGI	PaBr	[102]

(*Continued*)

Table 1. (Continued)

Organism	Source Database	Identifier	Reference
Penicillium camemberti	NCBI	PeCa	[103]
Penicillium chrysogenum	NCBI	PeCh	[104]
Penicillium digitatum	NCBI	PeDi	[105]
Penicillium expansum	NCBI	PeEx	[106]
Penicillium nalgiovense	NCBI	PeNa	[107]
Penicillium oxalicum	NCBI	PeOx	[108]
Penicillium roqueforti	NCBI	PeRo	[103]
Penicillium rubens Wisconsin	NCBI	PeRu	[109]
Penicillium vulpinum	JGI	PeVu	[107]
Periconia macrospinosa	JGI	PeMa	[67]
Pestalotiopsis fici	NCBI	PeFi	[110]
Phaeoacremonium aleophilum	JGI	PhAl	[111]
Phaeomoniella chlamydospora	JGI	PhCh	[78]
Phialocephala scopiformis	JGI	PhSc	[112]
Pneumocystis jirovecii	JGI	PnJi	[113]
Pseudogymnoascus destructans	JGI	PsDe	[114]
Pseudomassariella vexata	JGI	PsVe	[115]
Rhizoctonia solani	NCBI	RhSo	[116]
Saccharomyces arboricola	NCBI	SaAr	[117]
Saccharomyces cerevisiae	NCBI	SaCe	[118]
Terfezia boudieri	JGI	TeBo	[55]
Tolypocladium ophioglossoides	NCBI	ToOp	[119]
Tolypocladium paradoxum	NCBI	ToPa	[120]
Trichoderma arundinaceum	NCBI	TrAr	[121]
Trichoderma asperellum	NCBI	TrAs	[122]
Trichoderma atroviride	NCBI	TrAt	[123]
Trichoderma citrinoviride	NCBI	TrCi	[122]
Trichoderma harzianum	NCBI	TrHa	[124]
Trichoderma longibrachiatum	NCBI	TrLo	[125]
Trichoderma reesei	NCBI	TrRe	[126]
Trichoderma virens	NCBI	TrVi	[123]
Trichophyton rubrum	JGI	TrRu	[127]
Tuber aestivum var. urcinatum	JGI	TuAe	[55]
Tuber magnatum	JGI	TuMa	[55]
Venturia inaequalis	JGI	VeIn	[128]
Verruconis gallopava	JGI	VeGa	[68]
Verticillium dahliae	JGI	VeDa	[129]
Xylona heveae	JGI	XyHe	[130]
Zymoseptoria brevis	JGI	ZyBr	[131]
Zymoseptoria pseudotritici	JGI	ZyPs	[132]

The sequences were downloaded from the National Center for Biotechnology Information (NCBI) database or the Joint Genome Institute (JGI). The identifiers were used in the FunOrder software package.

https://doi.org/10.1371/journal.pcbi.1009372.t001

Next, the top 20 results of the blastp analysis are extracted and combined with the query sequence for each gene. A custom Perl script removes potential duplicate entries based on sequence identity. Using emma, a multiple sequence alignment of these protein sequences is

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

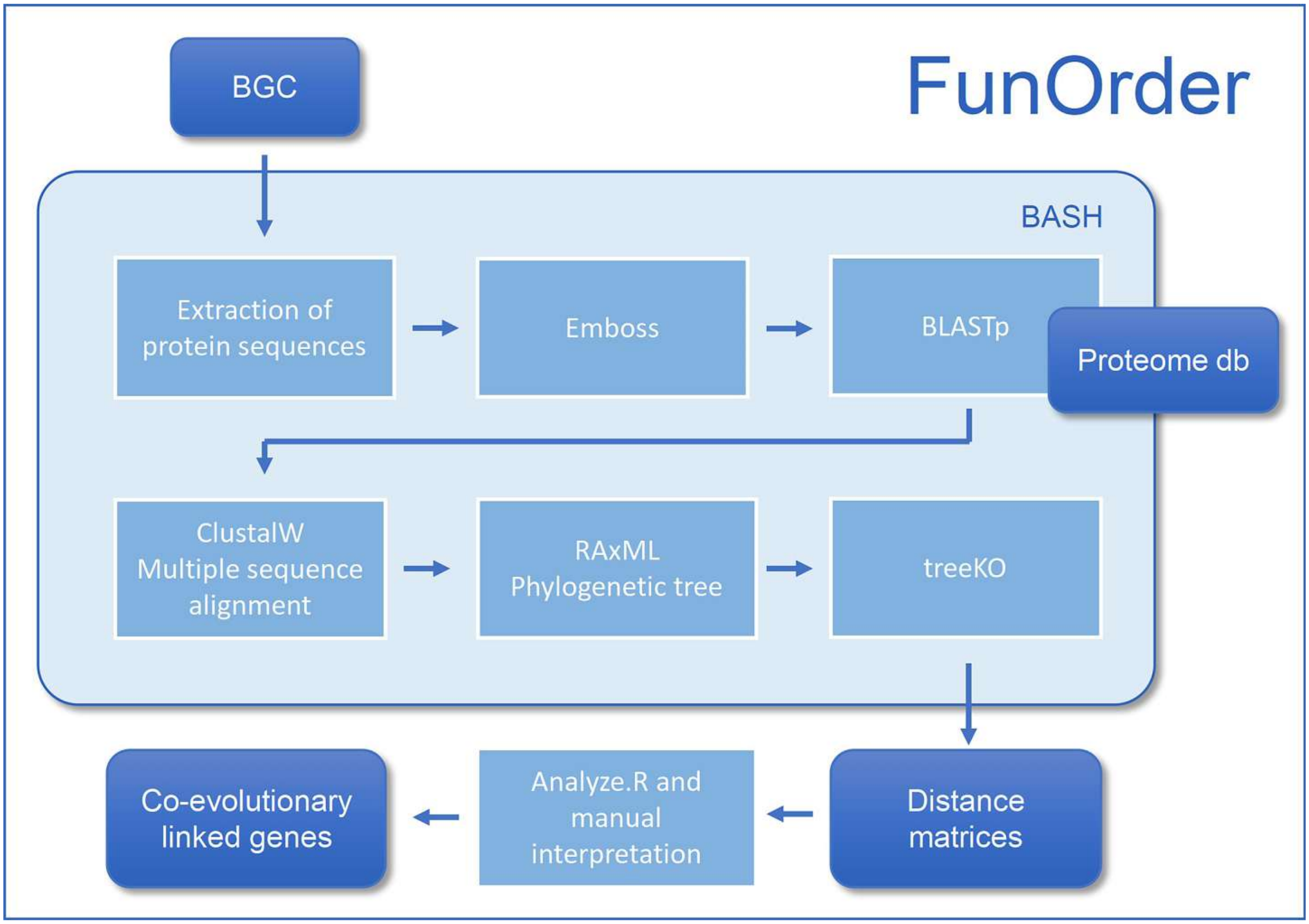

BGC
FunOrder
BASH
Extraction of protein sequences
Emboss
BLASTp
Proteome db
ClustalW Multiple sequence alignment
RAxML Phylogenetic tree
treeKO
Co-evolutionary linked genes
Analyze.R and manual interpretation
Distance matrices

Fig 2. Schematic representation of the workflow of FunOrder.

https://doi.org/10.1371/journal.pcbi.1009372.g002

calculated based on the ClustalW [134] algorithm, and a dendrogram computed. Based on the multiple sequence alignment, 100 rapid Bootstraps and a subsequent search for the best-scoring maximum likelihood (ML) tree are performed using RAxML (Randomized Axelerated Maximum Likelihood) [135]. The phylogenetic trees are computed using the LG amino acid substitution model. Furthermore, a standard ascertainment bias correction by Paul O. Lewis is performed. At this stage, we have obtained a phylogenetic tree (within the context of our empirically optimized database) for each protein of the input BGC.

To estimate if and to what extent the different genes within a BGC are co-evolved, the strict distance and speciation distance among the ML trees of the individual genes are calculated using the TreeKO algorithm [45]. This tool was designed for automated tree comparison and was already suggested to be used for the detection of co-evolution in protein families [45]. The tool compares the topology of different trees; a distance of 0 in both distance measures represents identical trees. In this context, a higher similarity between the different trees of the individual genes points towards a shared evolution. The strict distance is a weighted Robinson-Foulds (RF) distance measure that penalizes dissimilarities in evolutionarily important events such as gene losses and gene duplications; it has been suggested to be more significant in the detection of co-evolution than the evolutionary distance [45]. In contrast, the evolutionary or

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

speciation distance is computed without taking evolutionary exceptions, such as duplication events or different species content of the two compared trees into account and infers shared "speciation history" based solely on topology without considering branch lengths and only considering shared species of the compared trees. Therefore, an evolutionary distance of 0 does not necessarily describe identical trees but shared "speciation history" of shared species. All pairwise strict and evolutionary distances are combined into matrices which are used as input for an R script [136–140].

In this R-script, first, the strict and evolutionary distances are summed up to a third combined distance matrix combining the information about co-evolution and shared speciation into a single measure. In our experience, this measure can be helpful to detect genes that share little co-evolution with the core-enzymes but are still essential for the biosynthesis, which is reflected in a shared speciation. The evolutionary distance is not directly part of the output of FunOrder as is not intended to be used for the detection of co-evolution. Second, the strict and the combined distance matrices are visualized as heatmaps with a dendrogram computed with the complete linkage method, to find similar clusters in these data sets. Next, the Euclidean distance within the matrices is computed and clustered using Ward's minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of the dissimilarities before cluster updating, for the two distance matrices separately, with scaled input data [141]. Lastly, a principal component analysis (PCA) is performed on the two distance matrices and the score plot of the first two principal components visualized, respectively. These outputs enable the adoption of a larger view on the distance measures and thereby allow the analysis of co-evolution within the BGC from different perspectives. We describe in a following subchapter how to interpret these visualisations.

The software bundle is written in the BASH (Bourn Again Shell) environment and includes all necessary subprograms. As BASH is the default shell-language of all Linux distributions and MacOS, FunOrder can run on these two operation systems. The FunOrder software package is deposited in the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984). Notably, the software package includes scripts adapted to the use on servers and for the integration in various pipelines; details on these can be found in the ReadMe file on the GitHub repository. FunOrder requires some dependencies e.g., RAxML (Randomized Axelerated Maximum Likelihood) [135] and the EMBOSS (The European Molecular Biology Open Software Suite) package [142], for details and links to all dependencies please refer to the ReadMe file on the GitHub repository.

Compilation of benchmark gene clusters (GCs)

To test and evaluate the applicability of the FunOrder method, we used different control and test gene (or protein) sets. The sequences of all test and control sets are deposited in the GitHub repository https://github.com/gvignolle/FunOrder (doi: 10.5281/zenodo.5118984). The first set of negative control gene clusters (GCs) were 42 completely randomly generated synthetic GCs, which were created with a custom BASH script. Therein, ATGC strings of random composition and length were translated to amino acid strings using transeq from the EMBOSS package and the asterisks were removed. The second set of negative controls were 60 random GCs which were created by subsampling randomly the fungal proteome database with a Perl script from the MEME suit [143]. For each random GC a different seed number was given to guarantee non repetitive GCs, each random GC contained 3–10 randomly chosen protein sequences in a random order. These negative control GCs were subsampled from different genomes to maximize the randomness and use gene clusters that should not contain co-evolved genes.

Table 2. Empirically characterized biosynthetic gene clusters used as positive controls.

Product—BGC	Organism	MIBiG id	Reference(s)
2-Pyridon-Desmethylbassianin (dmb)	*Beauveria bassiana*	BGC0001136	[145]
Aflatoxin (afl)	*Aspergillus flavus*	BGC0000008	[146,147]
Botrydial (bot)	*Botrytis cinera*	BGC0000631	[148,149]
Cephalosporin (cef)	*Acremonium chrysogenum*	BGC0000317	[150]
Compactin (mlc)	*Penicillium citrinum*	BGC0000039	[151,152]
Cyclosporin (cyc2)	*Beauveria felina*	BGC0001565	[18,153–155]
Destruxin (dtxs)	*Metarhizium robertsii*	BGC0000337	[156]
Fumagillin (fma)	*Aspergillus fumigatus*	BGC0001067	[157]
Fumitremorgin (ftm)	*Aspergillus fumigatus*	-	[158–161]
Fumonisin (fum1)	*Fusarium oxysporum*	BGC0000063	[162]
Fumonisin (fum2)	*Fusarium verticilloides*	BGC0000062	[163–170]
Fusaric acid (FUB)	*Fusarium fujikuroi*	-	[171]
Ilicicolin H (ili)	*Neonectaria sp.* DH2	BGC0002035	[172]
Leporin (lep)	*Aspergillus flavus*	BGC0001445	[173]
Lovastatin (lov)	*Aspergillus terreus*	-	[17,62,174]
Mycophenolic acid (mpa1)	*Penicillium brevicompactum*	BGC0000104	[175–180]
Mycophenolic acid (mpa2)	*Penicillium roqueforti*	BGC0001360	[181]
Mycophenolic acid (mpa3)	*Penicillium roqueforti*	BGC0001677	[182]
Paxillin (pax)	*Penicillium paxilli*	BGC0001082	[183]
Penicillin (pen1)	*Penicillium chrysogenum*	BGC0000404	[184]
Penicillin (pen2)	*Penicillium chrysogenum*	BGC0000405	[19]
Pestheic acid (pta)	*Pestalotiopsis fici*	BGC0000121	[185]
Pneumocandin (GL)	*Glaera Iozoyensis*	BGC0001035	[186–188]
Sorbicillinol (sor1)	*Penicillium rubens*	BGC0001404	[189,190]
Sorbicillinol (sor2)	*Trichoderma reesei*	-	[191]
Tenellin (ten)	*Beauveria bassiana*	BGC0001049	[192,193]
Terrein (ter)	*Aspergillus terreus*	BGC0000161	[194]
Tetramic acid (tas)	*Hapsidospora irregularis*	-	[195]
Ustiloxin B (ust)	*Aspergillus flavus*	-	[196]
Xanthocillin (xan)	*Aspergillus fumigatus*	BGC0001990	[197]

https://doi.org/10.1371/journal.pcbi.1009372.t002

We used a set of 30 empirically well characterized BGCs from a broad range of different genera (Table 2) as positive controls. The BGC sequences were downloaded from NCBI or the MIBiG (Minimum information about a biosynthetic gene cluster) database [144]. The sequences are available at the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984). All BGCs were manually inspected for correctness and completeness based on the respective literature (S1 Table, references in Table 2). We further added 2 genes on each side of the BGC to mimic the greedy gain performed by antiSMASH, if possible (sequences available) and applicable (only few or no gap genes present). Next, we defined the class of each gene (biosynthetic gene, further essential gene, gap, or extra gene) according to the described function of the enzymes in the literature (S1 Table).

Further, we compiled 10 protein sets containing the sequences of enzymes of conserved metabolic pathways from organisms that were not included in the proteome database, termed „Biosynthetic_pathways", or „BioPath"(S2 Table; sequences deposited at the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984)). As we anticipate a strong co-evolution among the corresponding genes, we used these sets as positive controls for co-evolution in general. Finally, we subsampled the genomes of organisms that were not

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

included in the proteome database for 30 random loci containing 8 to 10 genes (S3 Table; sequences available at the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984)). We termed this control set „sequential GCs". This set should represent the random degree of co-evolution based only on genomic vicinity. Notably, due to the randomness of the sampling, the sequential GCs may also contain evolutionary linked genes.

Calculation of MEM and determination of thresholds for co-evolution

As the thresholds for the strict and/or evolutionary distance for the analysis of protein co-evolution are database dependent, we needed to define these thresholds manually. To this end, we performed a manual comparison of the phylogenetic trees of genes anticipated to be co-evolved and of not presumably co-evolved genes. As positive control datasets (anticipated co-evolution), we used the essential genes within the positive control BGCs. As negative control data set (anticipated to not have co-evolved), we used the genes in the random GCs. For the manual tree comparisons, we considered the topology (defined in S4 Table), branch lengths, number of nodes, and shared leaves of the trees and calculated the manual evaluation measure (MEM) according to the definitions in S5 Table. We calculated the MEM for each gene tree pair of the positive and the negative control data sets (S6 and S7 Tables, respectively). The measure ranges from 3 (same) to 0 (no shared leaves). The MEM values of each pair-wise tree comparison were then manually reconciled with the corresponding strict and the combined distance measures obtained from the treeKO analysis and the subsequent R script, respectively. The procedure is exemplary described for the 2-Pyridon-Desmethylbassianin (dmb) BGC from *Beauveria bassiana* in S1 File. Based on these manual comparisons, we defined the threshold values for strict and combined distances in the following: two genes are considered as co-evolved if the strict distance value is less than 0.7 or if the combined distance is equal to or less than 60 percent of the maximum value in the combined distance matrix of the analysed set.

Calculation of the Internal co-evolutionary quotient (ICQ)

The internal co-evolutionary quotient (ICQ) expresses how many genes in a GC or proteins in a protein set are co-evolved according to the previously defined threshold for strict and combined distances within the distance matrices of an analysed GC (or protein set). To calculate the ICQ, each protein is compared with every other protein. The total number of all possible pairwise comparisons is $2^* [d^*(d-1)]$ for d proteins. The ICQ was calculated using Eq 1, resulting in values between 0 and 1, with 1 representing no co-evolved genes, and 0 representing that most genes are co-evolved with each other in the insert GC.

$$ICQ = 1 - \left\{ \frac{g}{2 * [d * (d - 1)]} \right\} \qquad \text{Eq 1}$$

ICQ = internal co-evolutionary quotient; g = number of strict distances < 0.7 and combined distances $< =$ (0.6 * max value of the combined distance matrix) in all matrices (visualized in the heatmaps); d = number of genes in the GC.

Manual interpretation of the FunOrder output

The FunOrder outputs three different visualizations (heatmap, dendrogram, PCA) each of the strict and combined distance matrices among the genes (or proteins) of an inserted GC (or protein set). These visualizations need to be interpreted manually. For the manual interpretation, we first searched for genes that clustered together with the core enzyme(s) in any of the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

three visualisations of the strict distance. The definition of the clusters needs to be performed carefully keeping the biological background (gene predictions) in mind. For instance, a cluster containing typical tailoring enzymes (e.g., hydrolases, P450 cytochrome oxidases, FAD-containing enzymes, etc.) and/or further essential genes (e.g., transcription factors or transporters) make sense, whereas clusters containing a lot of genes encoding for unknown genes and/or genes that are unlikely to be involved in the biosynthesis of a secondary metabolite) do not make sense. Next, clustering in the visualizations of the combined distances is considered. As the combined distance also contains information about the speciation history, it may be used to add further genes to the list of "detected genes". Notably, this needs to be critically evaluated and decided on a case-to-case basis, taking the gene predictions into account. Please also refer to S2 File for a detailed step-by-step description of the interpretation procedure, the exemplary analysis of the lovastatin BGC from *A*. *terreus* in the results, and S3 File and S4 File for the exemplary analysis of two unknown BGCs.

Performance evaluation

To test the robustness of FunOrder, we analysed 42 completely randomly generated synthetic GCs. To test whether the FunOrder method can be used to detect co-evolution within GCs (or protein sets), we calculated the ICQ for different control sets and compared the results in a kernel density plot. To evaluate the performance of the FunOrder method regarding its capability to identify presumably co-evolved essential genes (as defined in S1 Table) and to distinguish them from (presumably not co-evolved) gap genes and genes outside of the BGC via the detection of co-evolution, we performed a manual interpretation of 30 empirically characterized BGCs (Table 2) as described above. Genes that clustered together with the core enzyme(s) according to the procedure described above were considered as „detected". Then we counted the total number of (1a) detected essential genes or (1b) detected biosynthetic genes, (2a) not detected essential genes or (2b) not detected biosynthetic genes, (3) detected gap and extra genes, and (4) not detected gap or extra genes in all BGCs, and defined (1a or 1b) as true positives (TP), (2a or 2b) as false negatives (FN), (3) as false positives (FP), and (4) as true negatives (TN). The values were used for a final statistical evaluation of FunOrder as suggested by Chicco and Jurman [198].

Results and discussion

Applicability of FunOrder for the detection of co-evolution

First, we analyzed the 42 synthetic negative control GCs with the FunOrder software. We could not find any sequence similarities with the empirically optimized fungal proteome database, demonstrating the robustness of the FunOrder method towards non-biological random amino acid sequences. Consequently, the 42 synthetic negative control GCs were not considered in the following.

Next, we performed FunOrder analyses of different control GCs and protein sets and calculated the internal co-evolutionary quotients (ICQs) using Eq 1. The ICQ is a value for the relative amount of co-evolutionary relations among the genes (or proteins) in a given GC or protein set. An ICQ of 0 means that most genes (or proteins) are co-evolved with each other. An ICQ of 1 means, that no co-evolution can be detected using the defined thresholds. As negative control for co-evolution, we used 60 randomly assembled negative control GCs (random GCs, S8 Table). The random GCs were compiled by subsampling different proteomes, to minimize the chance of random, unwanted co-evolution in the clusters. As positive control for co-evolution we used 10 protein sets from conserved metabolic pathways of different ascomycetes (S2 Table), termed „Biosynthetic pathways", or „BioPath". Given, that the proteins are part of

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

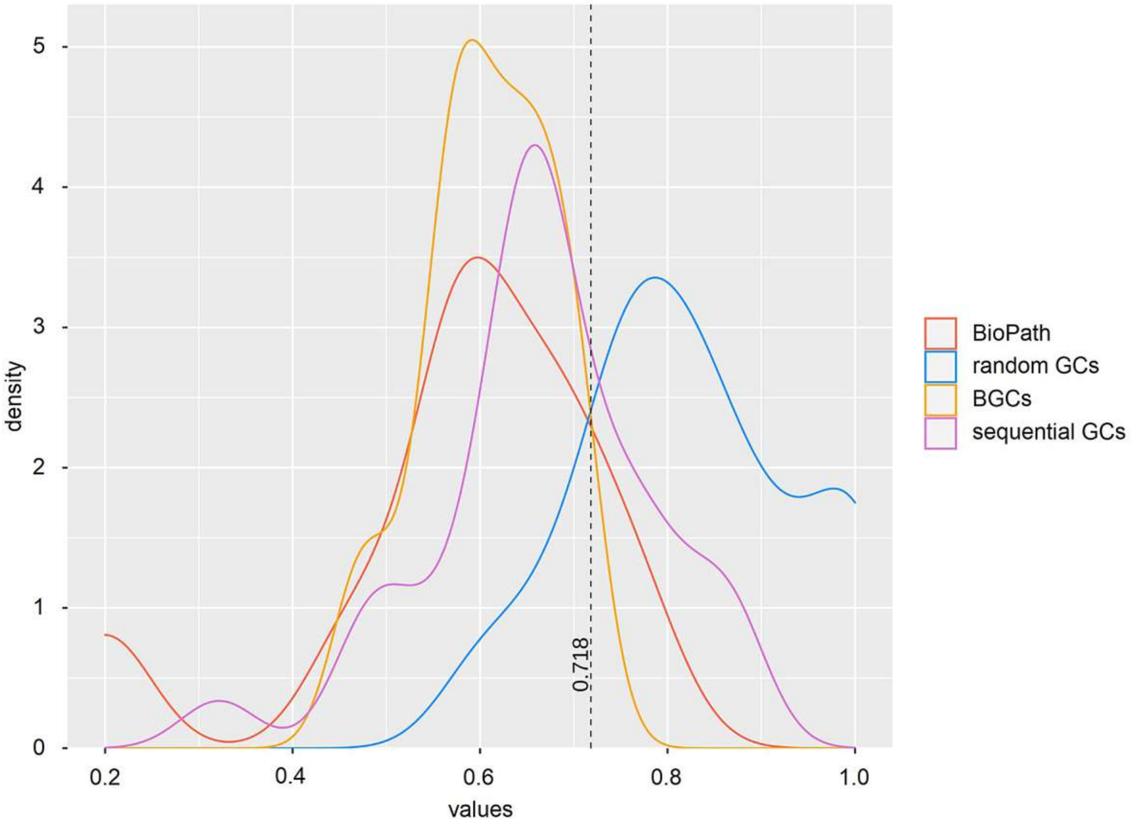


Fig 3. Kernel density plot of the ICQ values for co-evolutionary linked enzymes of different control sets. BioPath, protein sets of conserved biosynthetic pathways of the primary metabolism (S2 Table); random GCs, randomly assembled protein sets from 134 fungal proteomes (Table 1); BGCs, previously empirically characterized fungal BGCs (Table 2); sequential GCs, co-localized genes from random loci of different ascomycetes (S3 Table).

https://doi.org/10.1371/journal.pcbi.1009372.g003

the conserved primary metabolism and that their enzymatic functions are interrelated, we can assume a high level of internal co-evolution among the proteins within these protein sets. As control for the basic co-evolutionary value of co-localized (or sequential) genes, we used 30 random genetic loci containing 8 to 10 genes (S3 Table). We termed this control set „sequential GCs“. As test set for BGCs of the secondary metabolism in ascomycetes we used 30 empirically characterized BGCs (Table 2, S1 Table), also termed positive control BGCs.

We compared the ICQs of the different sets in an ANOVA (S5 File) and in a kernel density plot (Fig 3). We found that the ICQs for the random GCs were significantly different from all the other sets, demonstrating that the workflow of the FunOrder method can be used to detect co-evolution, that the ICQ is a meaningful measure to represent the content of co-evolutionary relationships within a GC or protein set, and that the manually defined thresholds for strict and combined distances are applicable to define co-evolution within GC or proteins sets. Based on these results, we defined the threshold of the ICQ for biologically relevant co-evolution within a GC as the point of intersect between the random GCs and the BGCs (0.718). GCs with an ICQ above this threshold do not contain significantly more co-evolutionary connections among the contained genes than randomly assembled GCs.

To our surprise, we could not detect a statistically significant difference between the sequential GCs and the positive control GCs. However, the maxima for the BioPath proteins and the BGC are at the same value and the shape of the corresponding density plot is

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

remarkably similar (Fig 3), whereas the maximum of the sequential GC is shifted towards the random GCs and the shape of the curve is different to the two positive control sets (Fig 3). These results indicate, that using only the absolute values of strict and combined distance may not be enough to distinguish co-evolutionary linked genes within the context of co-localized genes, but that the distances need to be assessed and interpreted in a case-by-case scenario considering the biological background and context of the analyzed GC.

Exemplary analysis of the lovastatin BGC (lov)

The FunOrder method allows the detection of co-evolved genes within a set of genes or proteins. As mentioned, we speculate that essential genes in BGCs are co-evolving and can therefore be differentiated from gap genes. In this context, the application of FunOrder might be used to detect the essential or at least the biosynthetic genes in BGCs. The software package of the FunOrder method calculates two distance matrices for the proteins within an input GC representing the evolutionary similarities (based on pair-wise comparisons of the phylogenetic trees using the treeKO tool [45]). First, we tried to use the previously defined thresholds for the strict and combined distances to automatically detect the co-evolutionary relations in BGCs. As insinuated above, this proofed not to be a successful strategy (not shown). We speculate, that the evolutionary similarities or distances among neighbouring genes are highly location specific and that the absolute values are therefore not meaningful as general thresholds. However, as the underlying strategy and method is clearly able to detect co-evolution (Fig 3), we speculated that the obtained data may need to be represented in different forms and/or reduced. Consequently, we added the following data visualizations to the FunOrder pipeline. The strict and combined distances are visualized in a heatmap and clustered by higher similarities (complete linkage method). Next, the Euclidean distances within the scaled distance matrices are calculated and clustered (hierarchical clustering) using the Wards minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of dissimilarities before cluster updating. The clustering is visualized in dendrograms. Finally, the principal components of the data are represented in a score plot. Here, we exemplary describe the manual interpretation of these visualizations (S6 File and Fig 4) with the aim to detect co-evolution within the lovastatin BGC of *A*. *terreus* (lov, Fig 1). Please refer also to the step-by-step description on how to interpret the FunOrder output in S2 File.

For the analysis of the lovastatin BGC, we first had a look at the heatmap representing the strict distance matrix (S6 File). Therein all biosynthetic genes (*lovA-D*, *F*, *G*; Fig 1, red arrows) are clustering together with each other and with the gap gene *orf1*, although not all inter-gene distances were below the previously defined threshold (S6 File, heatmaps). This demonstrates again that, evaluating only the numerical values (regardless of the concrete thresholds) is not enough for a thorough analysis of a BGC. It is necessary to consider the distances within the genomic context by comparing all provided visualisations. The biosynthetic genes of lovastatin (*lovA-D*, *F*, *G*) also formed distinct clusters in the dendrograms and in the PCA of the strict distance (S6 File and Fig 4A) In our experience, it was often helpful to additionally take the combined distance values into consideration to get a more comprehensive picture of the BGC. As mentioned before, the combined distance also considers speciation history. In the case of the lovastatin BGC, *orf10* and *extra03* clustered together with *lovA*, *B*, *D*, *F*, *G* in the PCA of the combined distance (Fig 4B). The gene *orf10* encodes for an MFS (major facilitator superfamily) transporter, which warrants adding it to the „detected genes"; the transporter is actually necessary for the export of lovastatin [17] (Fig 1). The gene *extra03* is predicted to encode for an alpha-glucuronidase (AguA) which is involved in the hydrolysis of xylan. Therefore, the clustering only in combined distance matrix does not justify classifying the gene *extra03* as

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

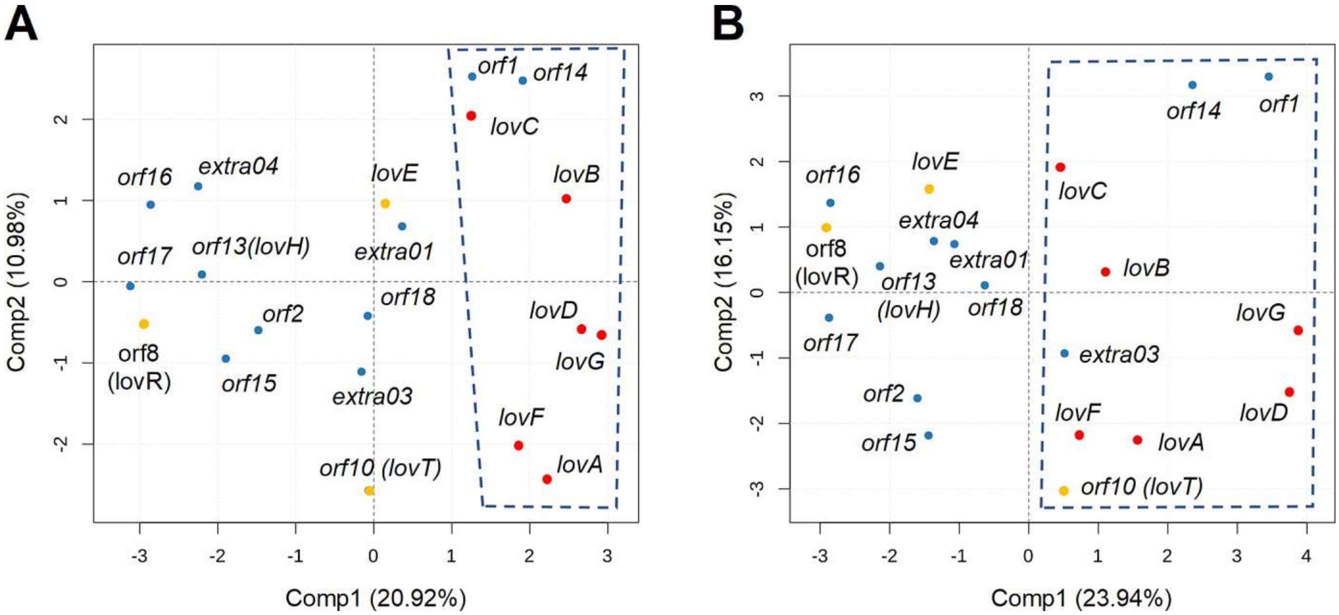


Fig 4. A selection of the standard output of the FunOrder analysis of the lovastatin BGC (lov). Score plots of the first two principal components from a PCA performed on the strict distance matrix (A) and on the combined distance matrix (B). The biosynthetic genes and the further essential genes are indicated in red and gold, respectively. Clusters in the PCA are indicated by the dashed boxes.

https://doi.org/10.1371/journal.pcbi.1009372.g004

„detected“. The other two „further essential genes“, *lovE and orf8* did not cluster together with the biosynthetic genes in any visualizations of the distance matrices (Fig 4 and S6 File)). LovE is a transcription factor and the main regulator of the lovastatin cluster [17] and essential for the lovastatin biosynthesis in the native organism, although it is not directly part of the biosynthetic pathway. The gene *orf8* encodes for a 3-hydroxy-3-methylglutaryl coenzyme-A (HMG-CoA) reductase, which is the target of statins [199] and in this case is conveying self-resistance to lovastatin [200]. These results suggest that these two genes did not undergo the same evolutionary process as the biosynthetic genes. This is in accordance with the „brick and mortar”model suggested by Medema et al. [40]. The biosynthetic genes represent a co-evolving „brick“, that is integrated into the biological context of *A*. *terreus* via the „mortar”that are the further essential genes.

This exemplary analysis demonstrates how the different data output formats of the software package need to be considered and compared manually, to decide on which genes are co-evolutionary linked and likely to be involved in the biosynthesis of a secondary metabolite. When considering only one output, one might get a distorted view of the analysed BGC. Notably, we did not intend to leave this step up to automation, because the human (expert or child) pattern recognition and mind still outperforms artificial intelligence (AI) algorithms and machine learning algorithms in this regard [201]. Please also refer to S3 File and S4 File in which we describe the analysis of two yet undescribed BGCs.

Speed and scalability of the software

As the empirically optimized proteome database contained only 134 fungal proteomes, we were able to use the blastp algorithm for sequence similarity search. The analysis of the lovastatin BGC of *A*. *terreus* (lov) with 17 genes, took 1 h 19 m 48 sec real time using 22 threads on an Ubuntu Linux system with 128 GB DDR4 RAM. The same analysis took 6 h 54 m 50 sec real time using 3 threads and 5 h 48 m 50 sec using 4 threads on a Linux Mint Laptop,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

demonstrating that the analysis of such a large cluster as the lovastatin cluster is fast and feasible. The number of threads can be defined, to increase the scalability and the overall performance.

Performance evaluation

Up to this point, we demonstrated that the FunOrder method can be used to detect the overall level of internal co-evolutionary relations within a GC or set of proteins. We demonstrated that similar levels of co-evolutionary relations occur among the genes in BGCs and among proteins of conserved metabolic pathways of the primary metabolism, and that these positive control sets can be distinguished from negative control GC, containing randomly stringed together proteins from different organisms with a threshold of 0.718 for the ICQ (Fig 3). Further, we showed that the values of strict and combined distances need to be visualized in different forms and then interpreted manually to detect co-evolution of individual genes within fungal BGCs. Next, we aimed to test, whether the detection of co-evolved genes is indeed a useful approach to identify the essential genes in fungal BGCs. To this end, we analysed the 30 empirically verified BGCs (Table 2) as described for the lovastatin cluster before. We looked for genes that are co-evolutionary linked with the core biosynthetic gene. These genes were considered as "detected". The "detected" genes sets were compared to the previously empirically obtained set of essential genes and classified the genes in true positives (TP), false negatives (FN), false positives (FP), or true negatives (TN) (S1 Table). To test and evaluate, how well FunOrder is performing in detecting either all essential or just the biosynthetic genes, we determined two different sets of TP and FN. TPs were either all detected essential genes, or all detected biosynthetic genes. Accordingly, FNs were either all not detected essential genes or all not detected biosynthetic genes (S1 Table). In both cases, FPs were all detected gap and extra genes, and TNs were all not detected gap and extra genes (S1 Table) because it makes biologically no sense to define a „detected"further essential gene as a FP, even when defining detected biosynthetic genes as TP. For an initial performance estimation, we calculated the percentages of detected essential and biosynthetic genes (S1 Table) and compiled them in a kernel density plot (Fig 5). More than 75% of all essential genes and biosynthetic genes were found to be co-evolving using the FunOrder method in 13 and 16 BGCs (out of 30 BGCs), respectively. The curves in the density plot also differ at high percentages; nearly all (above 90%) biosynthetic genes could be detect in more cases than nearly all essential genes. These two observations point in the direction, that especially the biosynthetic genes share a more coherent co-evolutionary history and can thus be identified by looking for co-evolved genes in BGCs. Obviously, not all essential genes in all BGCs are co-evolving and/or can be detected as co-evolved with this method. This is at least partly based on the biological background. Each BGC has a unique evolutionary background and needs to be interpreted individually. The FunOrder method offers additional information about co-evolution for already defined BGCs and may be useful in deciding which genes might be most relevant when studying a BGC.

For a stringent statistical evaluation, we calculated the normalized Matthews correlation coefficient (normMCC) and other classical metrics and global metrics (Table 3) as indicated by Chicco and Jurman [198] based on the previously defined TP, FN, FP, and TN (S1 Table). To determine the degree of balance between positive and negative controls we calculated the no-information error rate ni which is best for balanced test sets with the value 0.5. The obtained values of 0.5084 and 0.5444 allowed for the usage and confirmed the validity of the classical metrics such as F1 score and Accuracy. The FunOrder method displays overall high metrics in identifying essential and/or biosynthetic genes in a BGC. Despite the differences between biosynthetic and essential genes in Fig 5, we could not detect strong differences in the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

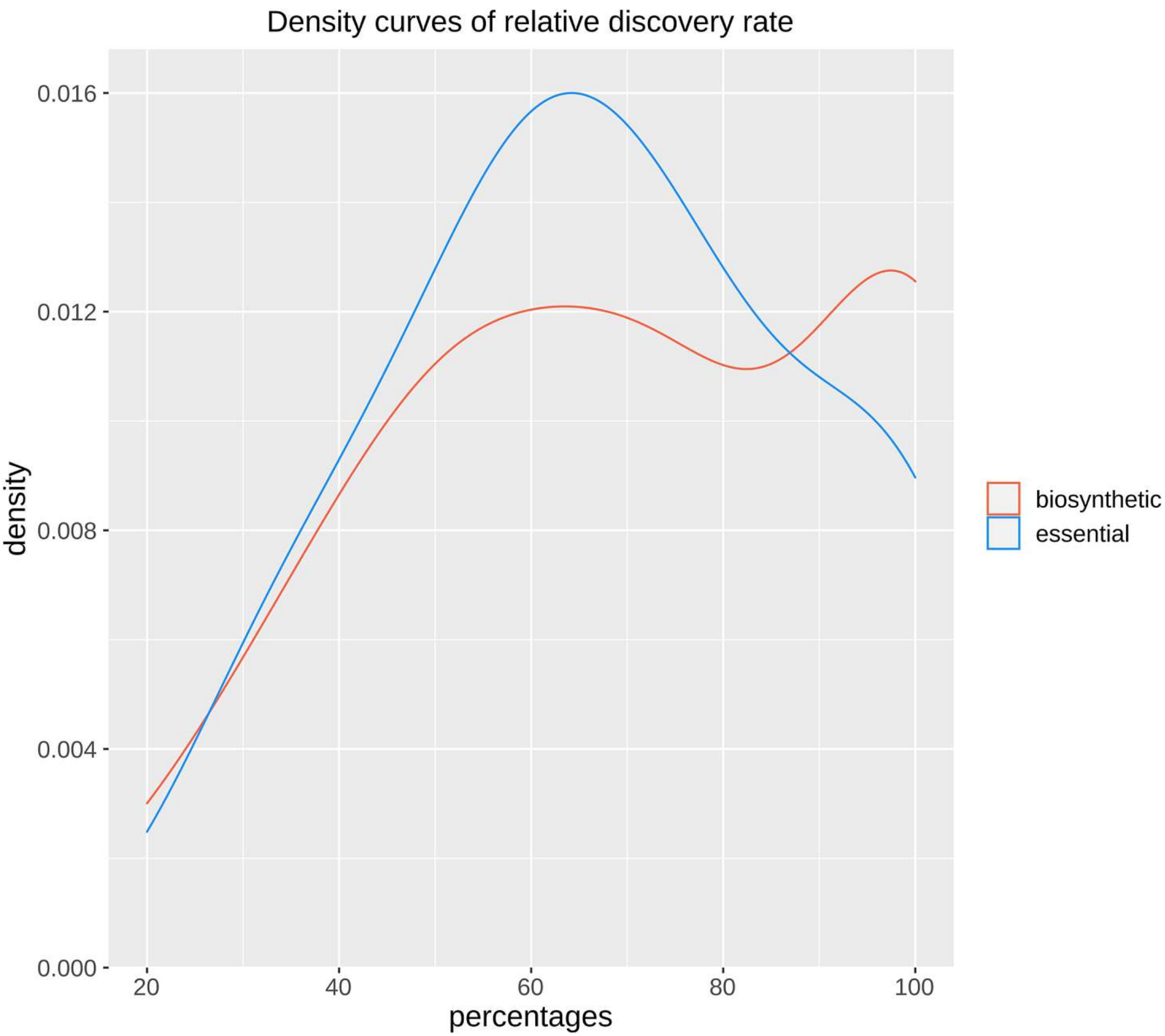


Fig 5. Kernel density plots of the relative discovery rate of essential or biosynthetic genes in 30 tested fungal BGCs.

https://doi.org/10.1371/journal.pcbi.1009372.g005

overall statistical assessment. FunOrder can be used to detect essential and biosynthetic genes in a BGC based on protein family co-evolution with a accuracy of 0.7215 and 0.743, respectively.

Concluding remarks

The FunOrder method was created to identify the essential genes in a BGC and distinguish them from gap genes based on the hypothesis that the essential genes are co-evolutionary linked. We evaluated this method and simultaneously tested the underlying hypothesis using different control sets of genes and proteins, respectively. We observed on the one hand that

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table 3. Statistical evaluation of the performance of FunOrder in detecting relevant genes in BGCs.

	essential genes	biosynthetic genes
Sensitivity	0.6349	0.6615
Specificity	0.8112	0.8112
Precision	0.7766	0.7457
Negative Predictive Value	0.6823	0.7412
False Positive Rate	0.1888	0.1888
False Discovery Rate	0.2234	0.2543
False Negative Rate	0.3651	0.3385
Accuracy	0.7215	0.743
F1 Score	0.6986	0.7011
Matthews Correlation Coefficient	0.4524	0.4797
Normalized Matthews Correlation Coefficient	0.7262	0.73985
No-information error rate ni	0.5084	0.5444

https://doi.org/10.1371/journal.pcbi.1009372.t003

co-evolutionary linkage in fungal BGCs is commonly occurring—especially within the biosynthetic genes, and on the other hand that the FunOrder method can be used to detect the biosynthetic genes within BGCs and to some extent also the further essential genes. We would like to stress that this method is delivering data on co-evolution, that needs to be critically evaluated and interpreted keeping the biological background in mind, and that FunOrder is not to be considered a stand-alone tool but meant to deliver supplementary data about co-evolution within predefined BGCs.

During the testing and evaluation, we encountered several cases of ambiguous results, where the different visualizations clustered different genes together. One way to handle such ambiguous results is to critically assess the results by considering the gene predictions. We further suggest adding and/or removing genes at the edges of the BGC and re-running the analysis. This might change the clustering behaviour and clarify the results. Alternatively, homologous BGCs from other fungi may be analysed by FunOrder and the clustering of the corresponding genes compared to the initial BGC.

The basis but also limitation for the method is the database [42]. Here we used a specific set of proteomes (Table 1) and were thus able to detect co-evolved genes in ascomycetes. Notably, the underlying strategy and workflow of FunOrder can be adapted to analysing genomic regions in other phyla, orders, or even kingdoms by using different databases. In case a larger database is integrated into the software package, alternative search algorithms, such as DIAMOND [202] or HMMER (similarity search using hidden Markov models) [203] might be used instead of blastp to enhance the performance. Nevertheless, each novel database, even if only one single proteome would be introduced in an existing database, will have to be verified and validated.

In this study, we looked for genes that share the same or a similar evolutionary background with the core genes of BGCs and could demonstrate that FunOrder is a fast and powerful method that can support scientists to decide which genes of a BGC are promising study objects. Notably, the application of this method is not limited to fungal BGC. It can be used for any applications where information of a shared co-evolution can contribute to a better understanding. FunOrder with the existing ascomycete database might already be used for a genome wide analysis of co-evolving transcription factors or detection of functionally connected protein-protein interactions [42]. As a future perspective, FunOrder might be even used for the analysis of total proteomes to detect evolutionary linked genes.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Supporting information

S1 Table. Empirically tested BGCs used as control set in this study.
(XLSX)

S2 Table. Protein sets of conserved metabolic pathways of the primary metabolism.
(XLSX)

S3 Table. Sequential GCs used in this study.
(XLS)

S4 Table. Definition of topology.
(PDF)

S5 Table. Parameters used to calculate the manual evaluation measure (MEM).
(PDF)

S6 Table. Calculation of MEM values for positive control BGCs.
(XLSX)

S7 Table. Calculation of MEM values for negative control GCs.
(XLSX)

S8 Table. Random GCs used in this study.
(XLSX)

S1 File. Exemplary MEM analysis of the dmb BGC.
(PDF)

S2 File. Step-by-step explanation for the manual interpretation of the FunOrder output.
(PDF)

S3 File. Exemplary interpretation of the FunOrder output of an unknown fungal BGC 1.
(PDF)

S4 File. Exemplary interpretation of the FunOrder output of an unknown fungal BGC 2.
(PDF)

S5 File. ANOVA for the ICQ values of the control and tests GCs and protein sets, respectively.
(PDF)

S6 File. FunOrder output of the Lovastatin BGC from *A. terreus* (lov).
(PDF)

Acknowledgments

We thank Dr. Ekaterina Shelest for reviewing a preprint of this study on https://www.biorxiv.org/ and providing us with helpful inputs and suggestions. We thank Matthias Schmal for testing the FunOrder installation and performance on a Linux Mint Laptop.

Author Contributions

Conceptualization: Gabriel A. Vignolle, Christian Derntl.

Data curation: Gabriel A. Vignolle, Denise Schaffer, Leopold Zehetner.

Formal analysis: Gabriel A. Vignolle, Leopold Zehetner.

Funding acquisition: Robert L. Mach, Christian Derntl.

Investigation: Gabriel A. Vignolle, Denise Schaffer.

Methodology: Gabriel A. Vignolle, Christian Derntl.

Project administration: Christian Derntl.

Resources: Robert L. Mach, Astrid R. Mach-Aigner.

Software: Gabriel A. Vignolle.

Supervision: Robert L. Mach, Astrid R. Mach-Aigner, Christian Derntl.

Validation: Gabriel A. Vignolle, Leopold Zehetner, Christian Derntl.

Visualization: Gabriel A. Vignolle, Christian Derntl.

Writing – original draft: Gabriel A. Vignolle, Christian Derntl.

Writing – review & editing: Denise Schaffer, Robert L. Mach, Astrid R. Mach-Aigner.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

References

1. Thirumurugan D, Cholarajan A, Raja SSS, Vijayakumar R. An Introductory Chapter: Secondary Metabolites. In: Vijayakumar R, Raja SSS, editors. Secondary Metabolites—Sources and Applications. London, UK: IntechOpen Limited; 2018.
2. Malik VS. Microbial secondary metabolism. Trends in Biochemical Sciences. 1980; 5(3):68–72. https://doi.org/10.1016/0968-0004(80)90071-7.
3. Keller NP, Turner G, Bennett JW. Fungal secondary metabolism—from biochemistry to genomics. Nature Reviews Microbiology. 2005; 3(12):937–47. https://doi.org/10.1038/nrmicro1286 PMID: 16322742
4. Alberti F, Foster GD, Bailey AM. Natural products from filamentous fungi and production by heterologous expression. Applied microbiology and biotechnology. 2017; 101(2):493–500. Epub 2016/12/15. https://doi.org/10.1007/s00253-016-8034-2 PMID: 27966047; PubMed Central PMCID: PMC5219032.
5. Newman DJ, Cragg GM, Kingston DGI. Chapter 5—Natural Products as Pharmaceuticals and Sources for Lead Structures**Note: This chapter reflects the opinions of the authors, not necessarily those of the US Government. In: Wermuth CG, Aldous D, Raboisson P, Rognan D, editors. The Practice of Medicinal Chemistry (Fourth Edition). San Diego: Academic Press; 2015. p. 101–39.
6. Brakhage AA, Schroeckh V. Fungal secondary metabolites—strategies to activate silent gene clusters. Fungal genetics and biology: FG & B. 2011; 48(1):15–22. https://doi.org/10.1016/j.fgb.2010.04.004 PMID: 20433937.
7. Atanasov AG, Zotchev SB, Dirsch VM, Orhan IE, Banach M, Rollinger JM, et al. Natural products in drug discovery: advances and opportunities. Nature Reviews Drug Discovery. 2021. https://doi.org/10.1038/s41573-020-00114-z PMID: 33510482
8. Wiemann P, Keller NP. Strategies for mining fungal natural products. Journal of industrial microbiology & biotechnology. 2014; 41(2):301–13. https://doi.org/10.1007/s10295-013-1366-3 PMID: 24146366.
9. Kensler TW, Roebuck BD, Wogan GN, Groopman JD. Aflatoxin: a 50-year odyssey of mechanistic and translational toxicology. Toxicol Sci. 2011; 120 Suppl 1:S28–48. https://doi.org/10.1093/toxsci/kfq283 PMID: 20881231; PubMed Central PMCID: PMC3043084.
10. Blount W. Turkey "X" disease. Turkeys. 1961; 9(2):52–5.
11. Yu J, Chang PK, Cary JW, Wright M, Bhatnagar D, Cleveland TE, et al. Comparative mapping of aflatoxin pathway gene clusters in *Aspergillus parasiticus* and *Aspergillus flavus*. Applied and environmental microbiology. 1995; 61(6):2365–71. https://doi.org/10.1128/aem.61.6.2365-2371.1995 PMID: 7793957; PubMed Central PMCID: PMC167508.
12. Soldatou S, Eldjarn GH, Huerta-Uribe A, Rogers S, Duncan KR. Linking biosynthetic and chemical space to accelerate microbial secondary metabolite discovery. FEMS microbiology letters. 2019; 366 (13). https://doi.org/10.1093/femsle/fnz142 PMID: 31252431
13. Craney A, Ahmed S, Nodwell J. Towards a new science of secondary metabolism. The Journal of antibiotics. 2013; 66(7):387–400. https://doi.org/10.1038/ja.2013.25 PMID: 23612726

14. Osbourn A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends in Genetics. 2010; 26(10):449–57. https://doi.org/10.1016/j.tig.2010.07.001 PMID: 20739089

15. Tran PN, Yen MR, Chiang CY, Lin HC, Chen PY. Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. Appl Microbiol Biotechnol. 2019; 103(8):3277–87. Epub 2019/03/13. https://doi.org/10.1007/s00253-019-09708-z PMID: 30859257; PubMed Central PMCID: PMC6449301.

16. Keller NP. Fungal secondary metabolism: regulation, function and drug discovery. Nat Rev Microbiol. 2019; 17(3):167–80. Epub 2018/12/12. https://doi.org/10.1038/s41579-018-0121-1 PMID: 30531948; PubMed Central PMCID: PMC6381595.

17. Mulder KC, Mulinari F, Franco OL, Soares MS, Magalhaes BS, Parachin NS. Lovastatin production: From molecular basis to industrial process optimization. Biotechnol Adv. 2015; 33(6 Pt 1):648–65. Epub 2015/04/15. https://doi.org/10.1016/j.biotechadv.2015.04.001 PMID: 25868803.

18. Weber G, Schörgendorfer K, Schneider-Scherzer E, Leitner E. The peptide synthetase catalyzing cyclosporine production in *Tolypocladium niveum* is encoded by a giant 45.8-kilobase open reading frame. Current Genetics. 1994; 26:120–5. https://doi.org/10.1007/BF00313798 PMID: 8001164

19. van den Berg MA, Westerlaken I, Leeflang C, Kerkman R, Bovenberg RA. Functional characterization of the penicillin biosynthetic gene cluster of *Penicillium chrysogenum* Wisconsin54-1255. Fungal genetics and biology: FG & B. 2007; 44(9):830–44. Epub 2007/06/06. https://doi.org/10.1016/j.fgb.2007.03.008 PMID: 17548217.

20. Nosanchuk JD, Stark RE, Casadevall A. Fungal Melanin: What do We Know About Structure? Front Microbiol. 2015; 6:1463. Epub 2016/01/07. https://doi.org/10.3389/fmicb.2015.01463 PMID: 26733993; PubMed Central PMCID: PMC4687393.

21. Wheeler MH, Bell AA. Melanins and their importance in pathogenic fungi. Curr Top Med Mycol. 1988; 2:338–87. https://doi.org/10.1007/978-1-4612-3730-3_10 PMID: 3288360.

22. Luo S, Dong SH. Recent Advances in the Discovery and Biosynthetic Study of Eukaryotic RiPP Natural Products. Molecules. 2019; 24(8). Epub 2019/04/21. https://doi.org/10.3390/molecules24081541 PMID: 31003555; PubMed Central PMCID: PMC6514808.

23. Montalban-Lopez M, Scott TA, Ramesh S, Rahman IR, van Heel AJ, Viel JH, et al. New developments in RiPP discovery, enzymology and engineering. Nat Prod Rep. 2020. Epub 2020/09/17. https://doi.org/10.1039/d0np00027b PMID: 32935693.

24. Wang DN, Toyotome T, Muraosa Y, Watanabe A, Wuren T, Bunsupa S, et al. GliA in *Aspergillus fumigatus* is required for its tolerance to gliotoxin and affects the amount of extracellular and intracellular gliotoxin. Medical mycology. 2014; 52(5):506–18. Epub 2014/05/23. https://doi.org/10.1093/mmy/myu007 PMID: 24847038.

25. Derntl C, Rassinger A, Srebotnik E, Mach RL, Mach-Aigner AR. Identification of the Main Regulator Responsible for Synthesis of the Typical Yellow Pigment Produced by *Trichoderma reesei*. Applied and environmental microbiology. 2016; 82(20):6247–57. https://doi.org/10.1128/AEM.01408-16 PMID: 27520818.

26. Schrettl M, Carberry S, Kavanagh K, Haas H, Jones GW, O'Brien J, et al. Self-protection against gliotoxin—a component of the gliotoxin biosynthetic cluster, GliT, completely protects Aspergillus fumigatus against exogenous gliotoxin. PLoS pathogens. 2010; 6(6):e1000952. https://doi.org/10.1371/journal.ppat.1000952 PMID: 20548963; PubMed Central PMCID: PMC2883607.

27. Anyaogu DC, Mortensen UH. Heterologous production of fungal secondary metabolites in *Aspergilli*. Frontiers in microbiology. 2015; 6(77). https://doi.org/10.3389/fmicb.2015.00077 PMID: 25713568

28. Chang PK, Yu J, Yu JH. aflT, a MFS transporter-encoding gene located in the aflatoxin gene cluster, does not have a significant role in aflatoxin secretion. Fungal genetics and biology: FG & B. 2004; 41(10):911–20. Epub 2004/09/03. https://doi.org/10.1016/j.fgb.2004.06.007 PMID: 15341913.

29. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019; 47(W1):W81–w7. Epub 2019/04/30. https://doi.org/10.1093/nar/gkz310 PMID: 31032519; PubMed Central PMCID: PMC6602434.

30. Wolf T, Shelest V, Nath N, Shelest E. CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. Bioinformatics. 2016; 32(8):1138–43. Epub 2015/12/15. https://doi.org/10.1093/bioinformatics/btv713 PMID: 26656005; PubMed Central PMCID: PMC4824125.

31. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal genetics and biology: FG & B. 2010; 47(9):736–41. https://doi.org/10.1016/j.fgb.2010.06.003 PMID: 20554054; PubMed Central PMCID: PMC2916752.

32. Almeida H, Palys S, Tsang A, Diallo AB. TOUCAN: a framework for fungal biosynthetic gene cluster discovery. NAR Genomics and Bioinformatics. 2020; 2(4). https://doi.org/10.1093/nargab/lqaa098 PMID: 33575642

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

33. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic acids research. 2019; 47(18):e110. https://doi.org/10.1093/nar/gkz654 PMID: 31400112; PubMed Central PMCID: PMC6765103.

34. Tai Y, Liu C, Yu S, Yang H, Sun J, Guo C, et al. Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (*Camellia sinensis*). BMC Genomics. 2018; 19(1):616. Epub 2018/08/17. https://doi.org/10.1186/s12864-018-4999-9 PMID: 30111282; PubMed Central PMCID: PMC6094456.

35. Lind AL, Wisecaver JH, Lameiras C, Wiemann P, Palmer JM, Keller NP, et al. Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. PLOS Biology. 2017; 15(11): e2003583. https://doi.org/10.1371/journal.pbio.2003583 PMID: 29149178

36. Rokas A, Wisecaver JH, Lind AL. The birth, evolution and death of metabolic gene clusters in fungi. Nature reviews Microbiology. 2018; 16(12):731–44. Epub 2018/09/09. https://doi.org/10.1038/s41579-018-0075-3 PMID: 30194403.

37. Palmer JM, Keller NP. Secondary metabolism in fungi: does chromosomal location matter? Current opinion in microbiology. 2010; 13(4):431–6. Epub 2010/07/16. https://doi.org/10.1016/j.mib.2010.04.008 PMID: 20627806; PubMed Central PMCID: PMC2922032.

38. Hoogendoorn K, Barra L, Waalwijk C, Dickschat JS, van der Lee TAJ, Medema MH. Evolution and Diversity of Biosynthetic Gene Clusters in *Fusarium*. Frontiers in microbiology. 2018; 9:1158. Epub 2018/06/21. https://doi.org/10.3389/fmicb.2018.01158 PMID: 29922257; PubMed Central PMCID: PMC5996196.

39. Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: How natural selection drives chemical innovation. Proceedings of the National Academy of Sciences. 2008; 105(12):4601–8. https://doi.org/10.1073/pnas.0709132105 PMID: 18216259

40. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. PLOS Computational Biology. 2014; 10(12):e1004016. https://doi.org/10.1371/journal.pcbi.1004016 PMID: 25474254

41. Rokas A, Mead ME, Steenwyk JL, Raja HA, Oberlies NH. Biosynthetic gene clusters and the evolution of fungal chemodiversity. Natural product reports. 2020; 37(7):868–78. https://doi.org/10.1039/c9np00045c PMID: 31898704

42. Ochoa D, Pazos F. Practical aspects of protein co-evolution. Frontiers in Cell and Developmental Biology. 2014; 2(14). https://doi.org/10.3389/fcell.2014.00014 PMID: 25364721

43. Del Carratore F, Zych K, Cummings M, Takano E, Medema MH, Breitling R. Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. Communications Biology. 2019; 2(1):83. https://doi.org/10.1038/s42003-019-0333-6 PMID: 30854475

44. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 2014; 42(Database issue):D26–31. Epub 2013/11/15. https://doi.org/10.1093/nar/gkt1069 PMID: 24225321; PubMed Central PMCID: PMC3965075.

45. Marcet-Houben M, Gabaldon T. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. Nucleic Acids Res. 2011; 39(10):e66. Epub 2011/02/22. https://doi.org/10.1093/nar/gkr087 PMID: 21335609; PubMed Central PMCID: PMC3105381.

46. Terfehr D, Dahlmann TA, Specht T, Zadra I, Kurnsteiner H, Kuck U. Genome Sequence and Annotation of *Acremonium chrysogenum*, Producer of the beta-Lactam Antibiotic Cephalosporin C. Genome announcements. 2014; 2(5). https://doi.org/10.1128/genomeA.00948-14 PMID: 25291769; PubMed Central PMCID: PMC4175204.

47. Nguyen HD, Lewis CT, Lévesque CA, Gräfenhan T. Draft Genome Sequence of *Alternaria alternata* ATCC 34957. Genome announcements. 2016; 4(1). Epub 2016/01/16. https://doi.org/10.1128/genomeA.01554-15 PMID: 26769939; PubMed Central PMCID: PMC4714121.

48. Hu J, Chen C, Peever T, Dang H, Lawrence C, Mitchell T. Genomic characterization of the conditionally dispensable chromosome in *Alternaria arborescens* provides evidence for horizontal gene transfer. BMC Genomics. 2012; 13(1):171. https://doi.org/10.1186/1471-2164-13-171 PMID: 22559316

49. Armitage AD, Cockerton HM, Sreenivasaprasad S, Woodhall J, Lane CR, Harrison RJ, et al. Genomics Evolutionary History and Diagnostics of the *Alternaria alternata* Species Group Including Apple and Asian Pear Pathotypes. Frontiers in microbiology. 2020; 10(3124). https://doi.org/10.3389/fmicb.2019.03124 PMID: 32038562

50. Lu Y, Ye C, Che J, Xu X, Shao D, Jiang C, et al. Genomic sequencing, genome-scale metabolic network reconstruction, and in silico flux analysis of the grape endophytic fungus *Alternaria* sp. MG1. Microb Cell Fact. 2019; 18(1):13. Epub 2019/01/27. https://doi.org/10.1186/s12934-019-1063-7 PMID: 30678677; PubMed Central PMCID: PMC6345013.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

51. Kohler A, Kuo A, Nagy LG, Morin E, Barry KW, Buscot F, et al. Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nat Genet. 2015; 47(4):410–5. Epub 2015/02/24. https://doi.org/10.1038/ng.3223 PMID: 25706625.

52. Martino E, Morin E, Grelet GA, Kuo A, Kohler A, Daghino S, et al. Comparative genomics and transcriptomics depict ericoid mycorrhizal fungi as versatile saprotrophs and plant mutualists. New Phytol. 2018; 217(3):1213–29. Epub 2018/01/10. https://doi.org/10.1111/nph.14974 PMID: 29315638.

53. Yang J, Wang L, Ji X, Feng Y, Li X, Zou C, et al. Genomic and proteomic analyses of the fungus *Arthrobotrys oligospora* provide insights into nematode-trap formation. PLoS Pathog. 2011; 7(9): e1002179. Epub 2011/09/13. https://doi.org/10.1371/journal.ppat.1002179 PMID: 21909256; PubMed Central PMCID: PMC3164635.

54. Burmester A, Shelest E, Glöckner G, Heddergott C, Schindler S, Staib P, et al. Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. Genome Biol. 2011; 12 (1):R7. Epub 2011/01/21. https://doi.org/10.1186/gb-2011-12-1-r7 PMID: 21247460; PubMed Central PMCID: PMC3091305.

55. Murat C, Payen T, Noel B, Kuo A, Morin E, Chen J, et al. Pezizomycetes genomes reveal the molecular basis of ectomycorrhizal truffle lifestyle. Nat Ecol Evol. 2018; 2(12):1956–65. Epub 2018/11/14. https://doi.org/10.1038/s41559-018-0710-4 PMID: 30420746.

56. de Vries RP, Riley R, Wiebenga A, Aguilar-Osorio G, Amillis S, Uchima CA, et al. Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus *Aspergillus*. Genome Biology. 2017; 18(1):28. https://doi.org/10.1186/s13059-017-1151-0 PMID: 28196534

57. Nierman WC, Yu J, Fedorova-Abrams ND, Losada L, Cleveland TE, Bhatnagar D, et al. Genome Sequence of *Aspergillus flavus* NRRL 3357, a Strain That Causes Aflatoxin Contamination of Food and Feed. Genome announcements. 2015; 3(2). https://doi.org/10.1128/genomeA.00168-15 PMID: 25883274; PubMed Central PMCID: PMC4400417.

58. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, et al. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature. 2005; 438(7071):1151–6. Epub 2005/12/24. https://doi.org/10.1038/nature04332 PMID: 16372009.

59. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, et al. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nature biotechnology. 2007; 25 (2):221–31. https://doi.org/10.1038/nbt1282 PMID: 17259976.

60. Zhao G, Yao Y, Qi W, Wang C, Hou L, Zeng B, et al. Draft genome sequence of *Aspergillus oryzae* strain 3.042. Eukaryotic cell. 2012; 11(9):1178–. https://doi.org/10.1128/EC.00160-12 PMID: 22933657.

61. Vesth TC, Nybo JL, Theobald S, Frisvad JC, Larsen TO, Nielsen KF, et al. Investigation of inter- and intraspecies variation through genome sequencing of *Aspergillus* section *Nigri*. Nature Genetics. 2018; 50(12):1688–95. https://doi.org/10.1038/s41588-018-0246-1 PMID: 30349117

62. Savitha J, Bhargavi SD, Praveen VK. Complete Genome Sequence of Soil Fungus *Aspergillus terreus* (KM017963), a Potent Lovastatin Producer. Genome Announc. 2016; 4(3). Epub 2016/06/11. https://doi.org/10.1128/genomeA.00491-16 PMID: 27284150; PubMed Central PMCID: PMC4901219.

63. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, et al. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science. 2010; 330 (6010):1543–6. Epub 2010/12/15. https://doi.org/10.1126/science.1194573 PMID: 21148392.

64. Marsberg A, Kemler M, Jami F, Nagel JH, Postma-Smidt A, Naidoo S, et al. *Botryosphaeria dothidea*: a latent pathogen of global importance to woody plant health. Mol Plant Pathol. 2017; 18(4):477–88. Epub 2016/09/30. https://doi.org/10.1111/mpp.12495 PMID: 27682468; PubMed Central PMCID: PMC6638292.

65. Van Kan JA, Stassen JH, Mosbach A, Van Der Lee TA, Faino L, Farmer AD, et al. A gapless genome sequence of the fungus *Botrytis cinerea*. Mol Plant Pathol. 2017; 18(1):75–89. Epub 2016/02/26. https://doi.org/10.1111/mpp.12384 PMID: 26913498; PubMed Central PMCID: PMC6638203.

66. Valero-Jiménez CA, Veloso J, Staats M, van Kan JAL. Comparative genomics of plant pathogenic *Botrytis* species with distinct host specificity. BMC Genomics. 2019; 20(1):203. https://doi.org/10.1186/s12864-019-5580-x PMID: 30866801

67. Knapp DG, Németh JB, Barry K, Hainaut M, Henrissat B, Johnson J, et al. Comparative genomics provides insights into the lifestyle and reveals functional heterogeneity of dark septate endophytic fungi. Sci Rep. 2018; 8(1):6321. Epub 2018/04/22. https://doi.org/10.1038/s41598-018-24686-4 PMID: 29679020; PubMed Central PMCID: PMC5910433.

68. Teixeira MM, Moreno LF, Stielow BJ, Muszewska A, Hainaut M, Gonzaga L, et al. Exploring the genomic diversity of black yeasts and relatives (*Chaetothyriales, Ascomycota*). Stud Mycol. 2017; 86:1–28.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Epub 2017/03/30. https://doi.org/10.1016/j.simyco.2017.01.001 PMID: 28348446; PubMed Central PMCID: PMC5358931.

69. Cuomo CA, Untereiner WA, Ma LJ, Grabherr M, Birren BW. Draft Genome Sequence of the Cellulolytic Fungus *Chaetomium globosum*. Genome announcements. 2015; 3(1). https://doi.org/10.1128/genomeA.00021-15 PMID: 25720678; PubMed Central PMCID: PMC4342419.

70. Armaleo D, Müller O, Lutzoni F, Andrésson Ó S, Blanc G, Bode HB, et al. The lichen symbiosis re-viewed through the genomes of *Cladonia grayi* and its algal partner *Asterochloris glomerata*. BMC Genomics. 2019; 20(1):605. Epub 2019/07/25. https://doi.org/10.1186/s12864-019-5629-x PMID: 31337355; PubMed Central PMCID: PMC6652019.

71. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, et al. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen *Dothideomycetes* fungi. PLoS pathogens. 2012; 8(12):e1003037. https://doi.org/10.1371/journal.ppat.1003037 PMID: 23236275; PubMed Central PMCID: PMC3516569.

72. Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, Otillar R, et al. Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. PLoS Genet. 2013; 9 (1):e1003233. Epub 2013/01/30. https://doi.org/10.1371/journal.pgen.1003233 PMID: 23357949; PubMed Central PMCID: PMC3554632.

73. Baroncelli R, Amby DB, Zapparata A, Sarrocco S, Vannacci G, Le Floch G, et al. Gene family expansions and contractions are associated with host range in plant pathogens of the genus *Colletotrichum*. BMC Genomics. 2016; 17:555. Epub 2016/08/09. https://doi.org/10.1186/s12864-016-2917-6 PMID: 27496087; PubMed Central PMCID: PMC4974774.

74. Baroncelli R, Sanz-Martin JM, Rech GE, Sukno SA, Thon MR. Draft Genome Sequence of *Colletotrichum sublineola*, a Destructive Pathogen of Cultivated Sorghum. Genome announcements. 2014; 2 (3). https://doi.org/10.1128/genomeA.00540-14 PMID: 24926053; PubMed Central PMCID: PMC4056296.

75. Hacquard S, Kracher B, Hiruma K, Münch PC, Garrido-Oter R, Thon MR, et al. Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. Nat Commun. 2016; 7:11362. Epub 2016/05/07. https://doi.org/10.1038/ncomms11362 PMID: 27150427; PubMed Central PMCID: PMC4859067.

76. Lopez D, Ribeiro S, Label P, Fumanal B, Venisse JS, Kohler A, et al. Genome-Wide Analysis of *Corynespora cassiicola* Leaf Fall Disease Putative Effectors. Front Microbiol. 2018; 9:276. Epub 2018/03/20. https://doi.org/10.3389/fmicb.2018.00276 PMID: 29551995; PubMed Central PMCID: PMC5840194.

77. Wu W, Davis RW, Tran-Gyamfi MB, Kuo A, LaButti K, Mihaltcheva S, et al. Characterization of four endophytic fungi as potential consolidated bioprocessing hosts for conversion of lignocellulose into advanced biofuels. Appl Microbiol Biotechnol. 2017; 101(6):2603–18. Epub 2017/01/13. https://doi.org/10.1007/s00253-017-8091-1 PMID: 28078400.

78. Morales-Cruz A, Amrine KC, Blanco-Ulate B, Lawrence DP, Travadon R, Rolshausen PE, et al. Distinctive expansion of gene families associated with plant cell wall degradation, secondary metabolism, and nutrient uptake in the genomes of grapevine trunk pathogens. BMC Genomics. 2015; 16(1):469. Epub 2015/06/19. https://doi.org/10.1186/s12864-015-1624-z PMID: 26084502; PubMed Central PMCID: PMC4472170.

79. Jones L, Riaz S, Morales-Cruz A, Amrine KC, McGuire B, Gubler WD, et al. Adaptive genomic structural variation in the grape powdery mildew pathogen, *Erysiphe necator*. BMC Genomics. 2014; 15 (1):1081. Epub 2014/12/10. https://doi.org/10.1186/1471-2164-15-1081 PMID: 25487071; PubMed Central PMCID: PMC4298948.

80. Blanco-Ulate B, Rolshausen PE, Cantu D. Draft Genome Sequence of the Grapevine Dieback Fungus *Eutypa lata* UCR-EL1. Genome announcements. 2013; 1(3). Epub 2013/06/01. https://doi.org/10.1128/genomeA.00228-13 PMID: 23723393; PubMed Central PMCID: PMC3668001.

81. Bombassaro A, de Hoog S, Weiss VA, Souza EM, Leão AC, Costa FF, et al. Draft Genome Sequence of *Fonsecaea monophora* Strain CBS 269.37, an Agent of Human Chromoblastomycosis. Genome Announc. 2016; 4(4). Epub 2016/07/30. https://doi.org/10.1128/genomeA.00731-16 PMID: 27469960; PubMed Central PMCID: PMC4966464.

82. Bashyal BM, Rawat K, Sharma S, Kulshreshtha D, Gopala Krishnan S, Singh AK, et al. Whole Genome Sequencing of *Fusarium fujikuroi* Provides Insight into the Role of Secretory Proteins and Cell Wall Degrading Enzymes in Causing Bakanae Disease of Rice. Front Plant Sci. 2017; 8:2013-. https://doi.org/10.3389/fpls.2017.02013 PMID: 29230233.

83. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, et al. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. Science. 2007; 317(5843):1400–2. Epub 2007/09/08. 317/5843/1400 [pii] https://doi.org/10.1126/science.1143708 PMID: 17823352.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

84. Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature. 2010; 464(7287):367–73. Epub 2010/03/20. https://doi.org/10.1038/nature08850 PMID: 20237561; PubMed Central PMCID: PMC3048781.

85. Niehaus EM, Münsterkötter M, Proctor RH, Brown DW, Sharon A, Idan Y, et al. Comparative "Omics" of the *Fusarium fujikuroi* Species Complex Highlights Differences in Genetic Potential and Metabolite Synthesis. Genome Biol Evol. 2016; 8(11):3574–99. Epub 2017/01/04. https://doi.org/10.1093/gbe/evw259 PMID: 28040774; PubMed Central PMCID: PMC5203792.

86. Gardiner DM, Benfield AH, Stiller J, Stephen S, Aitken K, Liu C, et al. A high-resolution genetic map of the cereal crown rot pathogen *Fusarium pseudograminearum* provides a near-complete genome assembly. Mol Plant Pathol. 2018; 19(1):217–26. Epub 2016/11/27. https://doi.org/10.1111/mpp.12519 PMID: 27888554; PubMed Central PMCID: PMC6638115.

87. Okagaki LH, Nunes CC, Sailsbery J, Clay B, Brown D, John T, et al. Genome Sequences of Three Phytopathogenic Species of the *Magnaporthaceae* Family of Fungi. G3 (Bethesda). 2015; 5(12):2539–45. Epub 2015/09/30. https://doi.org/10.1534/g3.115.020057 PMID: 26416668; PubMed Central PMCID: PMC4683626.

88. Peter M, Kohler A, Ohm RA, Kuo A, Krützmann J, Morin E, et al. Ectomycorrhizal ecology is imprinted in the genome of the dominant symbiotic fungus *Cenococcum geophilum*. Nat Commun. 2016; 7:12662. Epub 2016/09/08. https://doi.org/10.1038/ncomms12662 PMID: 27601008; PubMed Central PMCID: PMC5023957.

89. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, et al. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature. 2005; 434(7036):980–6. Epub 2005/04/23. https://doi.org/10.1038/nature03449 PMID: 15846337.

90. Gao Q, Jin K, Ying SH, Zhang Y, Xiao G, Shang Y, et al. Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M*. *acridum*. PLoS Genet. 2011; 7(1):e1001264. Epub 2011/01/22. https://doi.org/10.1371/journal.pgen.1001264 PMID: 21253567; PubMed Central PMCID: PMC3017113.

91. Hu X, Xiao G, Zheng P, Shang Y, Su Y, Zhang X, et al. Trajectory and genomic determinants of fungal-pathogen speciation and host adaptation. Proc Natl Acad Sci U S A. 2014; 111(47):16796–801. Epub 2014/11/05. https://doi.org/10.1073/pnas.1412662111 PMID: 25368161; PubMed Central PMCID: PMC4250126.

92. Binneck E, Lastra CCL, Sosa-Gómez DR. Genome Sequence of *Metarhizium rileyi*, a Microbial Control Agent for Lepidoptera. Microbiology Resource Announcements. 2019; 8(36):e00897–19. https://doi.org/10.1128/MRA.00897-19 PMID: 31488537

93. Meerupati T, Andersson K-M, Friman E, Kumar D, Tunlid A, Ahrén D. Genomic Mechanisms Accounting for the Adaptation to Parasitism in Nematode-Trapping Fungi. PLOS Genetics. 2013; 9(11): e1003909. https://doi.org/10.1371/journal.pgen.1003909 PMID: 24244185

94. Tan H, Kohler A, Miao R, Liu T, Zhang Q, Zhang B, et al. Multi-omic analyses of exogenous nutrient bag decomposition by the black morel *Morchella importuna* reveal sustained carbon acquisition and transferring. Environ Microbiol. 2019; 21(10):3909–26. Epub 2019/07/18. https://doi.org/10.1111/1462-2920.14741 PMID: 31314937.

95. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J, et al. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet. 2009; 5(8):e1000618. Epub 2009/08/29. https://doi.org/10.1371/journal.pgen.1000618 PMID: 19714214; PubMed Central PMCID: PMC2725324.

96. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, et al. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature. 2003; 422(6934):859–68. Epub 2003/04/25. https://doi.org/10.1038/nature01554 PMID: 12712197.

97. Baker SE, Schackwitz W, Lipzen A, Martin J, Haridas S, LaButti K, et al. Draft Genome Sequence of Neurospora crassa Strain FGSC 73. Genome announcements. 2015; 3(2):e00074–15. https://doi.org/10.1128/genomeA.00074-15 PMID: 25838471.

98. Ellison CE, Stajich JE, Jacobson DJ, Natvig DO, Lapidus A, Foster B, et al. Massive changes in genome architecture accompany the transition to self-fertility in the filamentous fungus *Neurospora tetrasperma*. Genetics. 2011; 189(1):55–69. Epub 2011/07/14. https://doi.org/10.1534/genetics.111.130690 PMID: 21750257; PubMed Central PMCID: PMC3176108.

99. Haridas S, Wang Y, Lim L, Massoumi Alamouti S, Jackman S, Docking R, et al. The genome and transcriptome of the pine saprophyte *Ophiostoma piceae*, and a comparison with the bark beetle-associated pine pathogen *Grosmannia clavigera*. BMC Genomics. 2013; 14:373. Epub 2013/06/04. https://doi.org/10.1186/1471-2164-14-373 PMID: 23725015; PubMed Central PMCID: PMC3680317.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

100. Urquhart AS, Mondo SJ, Mäkelä MR, Hane JK, Wiebenga A, He G, et al. Genomic and Genetic Insights Into a Cosmopolitan Fungus, *Paecilomyces variotii* (*Eurotiales*). Front Microbiol. 2018; 9:3058. Epub 2019/01/09. https://doi.org/10.3389/fmicb.2018.03058 PMID: 30619145; PubMed Central PMCID: PMC6300479.

101. Reynolds HT, Vijayakumar V, Gluck-Thaler E, Korotkin HB, Matheny PB, Slot JC. Horizontal gene cluster transfer increased hallucinogenic mushroom diversity. Evolution Letters. 2018; 2(2):88–101. https://doi.org/10.1002/evl3.42 PMID: 30283667

102. Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, Bailão AM, et al. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet. 2011; 7 (10):e1002345. Epub 2011/11/03. https://doi.org/10.1371/journal.pgen.1002345 PMID: 22046142; PubMed Central PMCID: PMC3203195.

103. Cheeseman K, Ropars J, Renault P, Dupont J, Gouzy J, Branca A, et al. Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nat Commun. 2014; 5:2876. Epub 2014/01/11. https://doi.org/10.1038/ncomms3876 PubMed Central PMCID: PMC3896755. PMID: 24407037

104. Specht T, Dahlmann TA, Zadra I, Kürnsteiner H, Kück U. Complete Sequencing and Chromosome-Scale Genome Assembly of the Industrial Progenitor Strain P2niaD18 from the *Penicillin Producer* Penicillium chrysogenum. Genome announcements. 2014; 2(4). Epub 2014/07/26. https://doi.org/10.1128/genomeA.00577-14 PMID: 25059858; PubMed Central PMCID: PMC4110216.

105. Marcet-Houben M, Ballester A-R, de la Fuente B, Harries E, Marcos JF, González-Candelas L, et al. Genome sequence of the necrotrophic fungus *Penicillium digitatum*, the main postharvest pathogen of citrus. BMC Genomics. 2012; 13(1):646. https://doi.org/10.1186/1471-2164-13-646 PMID: 23171342

106. Ballester AR, Marcet-Houben M, Levin E, Sela N, Selma-Lázaro C, Carmona L, et al. Genome, Transcriptome, and Functional Analyses of *Penicillium expansum* Provide New Insights Into Secondary Metabolism and Pathogenicity. Mol Plant Microbe Interact. 2015; 28(3):232–48. Epub 2014/10/23. https://doi.org/10.1094/MPMI-09-14-0261-FI PMID: 25338147.

107. Nielsen JC, Grijseels S, Prigent S, Ji B, Dainat J, Nielsen KF, et al. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species. Nat Microbiol. 2017; 2:17044. Epub 2017/04/04. https://doi.org/10.1038/nmicrobiol.2017.44 PMID: 28368369.

108. Liu G, Zhang L, Wei X, Zou G, Qin Y, Ma L, et al. Genomic and Secretomic Analyses Reveal Unique Features of the Lignocellulolytic Enzyme System of *Penicillium decumbens*. PloS one. 2013; 8(2): e55185. https://doi.org/10.1371/journal.pone.0055185 PMID: 23383313

109. van den Berg MA, Albang R, Albermann K, Badger JH, Daran JM, Driessen AJ, et al. Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. Nature biotechnology. 2008; 26(10):1161–8. https://doi.org/10.1038/nbt.1498 PMID: 18820685.

110. Wang X, Zhang X, Liu L, Xiang M, Wang W, Sun X, et al. Genomic and transcriptomic analysis of the endophytic fungus *Pestalotiopsis fici* reveals its lifestyle and high potential for synthesis of natural products. BMC Genomics. 2015; 16(1):28. Epub 2015/01/28. https://doi.org/10.1186/s12864-014-1190-9 PMID: 25623211; PubMed Central PMCID: PMC4320822.

111. Blanco-Ulate B, Rolshausen P, Cantu D. Draft Genome Sequence of the Ascomycete *Phaeoacremonium aleophilum* Strain UCR-PA7, a Causal Agent of the Esca Disease Complex in Grapevines. Genome announcements. 2013; 1(3). Epub 2013/07/03. https://doi.org/10.1128/genomeA.00390-13 PMID: 23814032; PubMed Central PMCID: PMC3695428.

112. Walker AK, Frasz SL, Seifert KA, Miller JD, Mondo SJ, LaButti K, et al. Full Genome of *Phialocephala scopiformis* DAOMC 229536, a Fungal Endophyte of Spruce Producing the Potent Anti-Insectan Compound Rugulosin. Genome announcements. 2016; 4(1). Epub 2016/03/08. https://doi.org/10.1128/genomeA.01768-15 PMID: 26950333; PubMed Central PMCID: PMC4767923.

113. Cissé OH, Pagni M, Hauser PM. *De novo* assembly of the *Pneumocystis jirovecii* genome from a single bronchoalveolar lavage fluid specimen from a patient. mBio. 2012; 4(1):e00428–12. Epub 2012/12/28. https://doi.org/10.1128/mBio.00428-12 PMID: 23269827; PubMed Central PMCID: PMC3531804.

114. Chibucos MC, Crabtree J, Nagaraj S, Chaturvedi S, Chaturvedi V. Draft Genome Sequences of Human Pathogenic Fungus *Geomyces pannorum* Sensu Lato and Bat White Nose Syndrome Pathogen *Geomyces* (*Pseudogymnoascus*) *destructans*. Genome announcements. 2013; 1(6). Epub 2013/12/21. https://doi.org/10.1128/genomeA.01045-13 PMID: 24356829; PubMed Central PMCID: PMC3868853.

115. Mondo SJ, Dannebaum RO, Kuo RC, Louie KB, Bewick AJ, LaButti K, et al. Widespread adenine N6-methylation of active genes in fungi. Nat Genet. 2017; 49(6):964–8. Epub 2017/05/10. https://doi.org/10.1038/ng.3859 PMID: 28481340.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

116. Cubeta MA, Thomas E, Dean RA, Jabaji S, Neate SM, Tavantzis S, et al. Draft Genome Sequence of the Plant-Pathogenic Soil Fungus *Rhizoctonia solani* Anastomosis Group 3 Strain Rhs1AP. Genome Announc. 2014; 2(5). Epub 2014/11/02. https://doi.org/10.1128/genomeA.01072-14 PMID: 25359908; PubMed Central PMCID: PMC4214984.

117. Liti G, Ba ANN, Blythe M, Müller CA, Bergström A, Cubillos FA, et al. High quality *de novo* sequencing and assembly of the *Saccharomyces arboricolus* genome. BMC Genomics. 2013; 14(1):69. https://doi.org/10.1186/1471-2164-14-69 PMID: 23368932

118. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996; 274(5287):546, 63–7. Epub 1996/10/25. https://doi.org/10.1126/science.274.5287.546 PMID: 8849441.

119. Quandt CA, Bushley KE, Spatafora JW. The genome of the truffle-parasite *Tolypocladium ophioglossoides* and the evolution of antifungal peptaibiotics. BMC Genomics. 2015; 16(1):553. Epub 2015/07/29. https://doi.org/10.1186/s12864-015-1777-9 PMID: 26215153; PubMed Central PMCID: PMC4517408.

120. Quandt CA, Patterson W, Spatafora JW. Harnessing the power of phylogenomics to disentangle the directionality and signatures of interkingdom host jumping in the parasitic fungal genus *Tolypocladium*. Mycologia. 2018; 110(1):104–17. Epub 2018/06/05. https://doi.org/10.1080/00275514.2018.1442618 PMID: 29863984.

121. Proctor RH, McCormick SP, Kim HS, Cardoza RE, Stanley AM, Lindo L, et al. Evolution of structural diversity of trichothecenes, a family of toxins produced by plant pathogenic and entomopathogenic fungi. PLoS Pathog. 2018; 14(4):e1006946. Epub 2018/04/13. https://doi.org/10.1371/journal.ppat.1006946 PMID: 29649280; PubMed Central PMCID: PMC5897003.

122. Druzhinina IS, Chenthamara K, Zhang J, Atanasova L, Yang D, Miao Y, et al. Massive lateral transfer of genes encoding plant cell wall-degrading enzymes to the mycoparasitic fungus *Trichoderma* from its plant-associated hosts. PLoS Genet. 2018; 14(4):e1007322. Epub 2018/04/10. https://doi.org/10.1371/journal.pgen.1007322 PMID: 29630596; PubMed Central PMCID: PMC5908196.

123. Kubicek CP, Herrera-Estrella A, Seidl-Seiboth V, Martinez DA, Druzhinina IS, Thon M, et al. Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma*. Genome Biol. 2011; 12(4):R40. Epub 2011/04/20. https://doi.org/10.1186/gb-2011-12-4-r40 PMID: 21501500; PubMed Central PMCID: PMC3218866.

124. Baroncelli R, Piaggeschi G, Fiorini L, Bertolini E, Zapparata A, Pè ME, et al. Draft Whole-Genome Sequence of the Biocontrol Agent *Trichoderma harzianum* T6776. Genome announcements. 2015; 3(3):e00647–15. https://doi.org/10.1128/genomeA.00647-15 PMID: 26067977.

125. Xie BB, Qin QL, Shi M, Chen LL, Shu YL, Luo Y, et al. Comparative genomics provide insights into evolution of trichoderma nutrition style. Genome Biol Evol. 2014; 6(2):379–90. Epub 2014/02/01. https://doi.org/10.1093/gbe/evu018 PMID: 24482532; PubMed Central PMCID: PMC3942035.

126. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, et al. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nature biotechnology. 2008; 26(5):553–60. Epub 2008/05/06. https://doi.org/10.1038/nbt1403 PMID: 18454138.

127. Martinez DA, Oliver BG, Gräser Y, Goldberg JM, Li W, Martinez-Rossi NM, et al. Comparative genome analysis of *Trichophyton rubrum* and related dermatophytes reveals candidate genes involved in infection. mBio. 2012; 3(5):e00259–12. Epub 2012/09/07. https://doi.org/10.1128/mBio.00259-12 PMID: 22951933; PubMed Central PMCID: PMC3445971.

128. Deng CH, Plummer KM, Jones DAB, Mesarich CH, Shiller J, Taranto AP, et al. Comparative analysis of the predicted secretomes of Rosaceae scab pathogens *Venturia inaequalis* and *V. pirina* reveals expanded effector families and putative determinants of host range. BMC Genomics. 2017; 18(1):339. Epub 2017/05/04. https://doi.org/10.1186/s12864-017-3699-1 PMID: 28464870; PubMed Central PMCID: PMC5412055.

129. Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BP, et al. Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. PLoS Pathog. 2011; 7(7):e1002137. Epub 2011/08/11. https://doi.org/10.1371/journal.ppat.1002137 PMID: 21829347; PubMed Central PMCID: PMC3145793.

130. Gazis R, Kuo A, Riley R, LaButti K, Lipzen A, Lin J, et al. The genome of *Xylona heveae* provides a window into fungal endophytism. Fungal Biol. 2016; 120(1):26–42. Epub 2015/12/24. https://doi.org/10.1016/j.funbio.2015.10.002 PMID: 26693682.

131. Grandaubert J, Bhattacharyya A, Stukenbrock EH. RNA-seq-Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. G3 (Bethesda). 2015; 5(7):1323–33. Epub 2015/04/29. https://doi.org/10.1534/g3.115.017731 PMID: 25917918; PubMed Central PMCID: PMC4502367.

132. Stukenbrock EH, Christiansen FB, Hansen TT, Dutheil JY, Schierup MH. Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. Proceedings of the National Academy of Sciences. 2012; 109(27):10954. https://doi.org/10.1073/pnas.1201403109 PMID: 22711811

133. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25 (17):3389–402. Epub 1997/09/01. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694; PubMed Central PMCID: PMC146917.

134. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 1994; 22(22):4673–80. Epub 1994/11/11. https://doi.org/10.1093/nar/22.22.4673 PMID: 7984417; PubMed Central PMCID: PMC308517.

135. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30(9):1312–3. Epub 2014/01/24. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623; PubMed Central PMCID: PMC3998144.

136. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2019.

137. Kucheryavskiy S. mdatools: Multivariate Data Analysis for Chemometrics. 2019.

138. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: Various R Programming Tools for Plotting Data. 2019.

139. Wickham H, Hester J, Francois R. readr: Read Rectangular Text Data. 2018.

140. Fox J, Weisberg S. An {R} Companion to Applied Regression. Sage. 2019.

141. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification. 2014; 31(3):274–95. https://doi.org/10.1007/s00357-014-9161-z

142. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000; 16(6):276–7. Epub 2000/05/29. https://doi.org/10.1016/s0168-9525(00)02024-2 PMID: 10827456.

143. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research. 2009; 37(Web Server issue):W202–8. https://doi.org/10.1093/nar/gkp335 PMID: 19458158; PubMed Central PMCID: PMC2703892.

144. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research. 2019; 48(D1): D454–D8. https://doi.org/10.1093/nar/gkz882 PMID: 31612915

145. Heneghan MN, Yakasai AA, Williams K, Kadir KA, Wasil Z, Bakeer W, et al. The programming role of trans-acting enoyl reductases during the biosynthesis of highly reduced fungal polyketides. Chemical Science. 2011; 2(5):972–9. https://doi.org/10.1039/C1SC00023C

146. Ehrlich KC, Chang PK, Yu J, Cotty PJ. Aflatoxin biosynthesis cluster gene *cypA* is required for G aflatoxin formation. Appl Environ Microbiol. 2004; 70(11):6518–24. Epub 2004/11/06. https://doi.org/10.1128/AEM.70.11.6518-6524.2004 PMID: 15528514; PubMed Central PMCID: PMC525170.

147. Bhatnagar D, Cary JW, Ehrlich K, Yu J, Cleveland TE. Understanding the genetics of regulation of aflatoxin production and *Aspergillus flavus* development. Mycopathologia. 2006; 162(3):155–66. Epub 2006/09/01. https://doi.org/10.1007/s11046-006-0050-9 PMID: 16944283.

148. Porquier A, Morgant G, Moraga J, Dalmais B, Luyten I, Simon A, et al. The botrydial biosynthetic gene cluster of Botrytis cinerea displays a bipartite genomic structure and is positively regulated by the putative Zn(II)2Cys6 transcription factor BcBot6. Fungal Genet Biol. 2016; 96:33–46. Epub 2016/10/22. https://doi.org/10.1016/j.fgb.2016.10.003 PMID: 27721016.

149. Pinedo C, Wang CM, Pradier JM, Dalmais B, Choquer M, Le Pecheur P, et al. Sesquiterpene synthase from the botrydial biosynthetic gene cluster of the phytopathogen *Botrytis cinerea*. ACS Chem Biol. 2008; 3(12):791–801. Epub 2008/11/28. https://doi.org/10.1021/cb800225v PMID: 19035644; PubMed Central PMCID: PMC2707148.

150. Hamed RB, Gomez-Castellanos JR, Henry L, Ducho C, McDonough MA, Schofield CJ. The enzymes of β-lactam biosynthesis. Natural product reports. 2013; 30(1):21–107. Epub 2012/11/09. https://doi.org/10.1039/c2np20065a PMID: 23135477.

151. Abe Y, Suzuki T, Ono C, Iwamoto K, Hosobuchi M, Yoshikawa H. Molecular cloning and characterization of an ML-236B (compactin) biosynthetic gene cluster in *Penicillium citrinum*. Mol Genet Genomics. 2002; 267(5):636–46. Epub 2002/08/13. https://doi.org/10.1007/s00438-002-0697-y PMID: 12172803.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

152. Abe Y, Ono C, Hosobuchi M, Yoshikawa H. Functional analysis of mlcR, a regulatory gene for ML-236B (compactin) biosynthesis in *Penicillium citrinum*. Mol Genet Genomics. 2002; 268(3):352–61. Epub 2002/11/19. https://doi.org/10.1007/s00438-002-0755-5 PMID: 12436257.

153. Hoffmann K, Schneider-Scherzer E, Kleinkauf H, Zocher R. Purification and Characterization of Eucaryotic Alanine Racemase Acting as Key Enzyme in Cyclosporin Biosynthesis. Biological Chemistry. 1994; 269(17):12710–4. PMID: 8175682

154. Yang X, Feng P, Yin Y, Bushley K, Spatafora JW, Wang C. Cyclosporine Biosynthesis in *Tolypocladium inflatum* Benefits Fungal Adaptation to the Environment. Molecular Biology and Physiology. 2018; 9(5). https://doi.org/10.1128/mBio.01211-18 PMID: 30279281

155. Weber G, Leitner E. Disruption of the cyclosporin synthetase gene of *Tolypocladium niveum*. Current Genetics. 1994; 26:461–7. https://doi.org/10.1007/BF00309935 PMID: 7874740

156. Wang B, Kang Q, Lu Y, Bai L, Wang C. Unveiling the biosynthetic puzzle of destruxins in *Metarhizium* species. Proc Natl Acad Sci U S A. 2012; 109(4):1287–92. Epub 2012/01/11. https://doi.org/10.1073/pnas.1115983109 PMID: 22232661; PubMed Central PMCID: PMC3268274.

157. Lin H-C, Chooi Y-H, Dhingra S, Xu W, Calvo AM, Tang Y. The Fumagillin Biosynthetic Gene Cluster in *Aspergillus fumigatus* Encodes a Cryptic Terpene Cyclase Involved in the Formation of β-trans-Bergamotene. Journal of the American Chemical Society. 2013; 135(12):4616–9. https://doi.org/10.1021/ja312503y PMID: 23488861

158. Kato N, Suzuki H, Takagi H, Uramoto M, Takahashi S, Osada H. Gene disruption and biochemical characterization of verruculogen synthase of *Aspergillus fumigatus*. Chembiochem. 2011; 12(5):711–4. Epub 2011/03/16. https://doi.org/10.1002/cbic.201000562 PMID: 21404415.

159. Kato N, Suzuki H, Takagi H, Asami Y, Kakeya H, Uramoto M, et al. Identification of cytochrome P450s required for fumitremorgin biosynthesis in *Aspergillus fumigatus*. Chembiochem. 2009; 10(5):920–8. Epub 2009/02/20. https://doi.org/10.1002/cbic.200800787 PMID: 19226505.

160. Maiya S, Grundmann A, Li SM, Turner G. Improved tryprostatin B production by heterologous gene expression in *Aspergillus nidulans*. Fungal Genet Biol. 2009; 46(5):436–40. Epub 2009/04/18. https://doi.org/10.1016/j.fgb.2009.01.003 PMID: 19373974.

161. Kato N, Suzuki H, Okumura H, Takahashi S, Osada H. A point mutation in *ftmD* blocks the fumitremorgin biosynthetic pathway in *Aspergillus fumigatus* strain Af293. Biosci Biotechnol Biochem. 2013; 77(5):1061–7. Epub 2013/05/08. https://doi.org/10.1271/bbb.130026 PMID: 23649274.

162. Proctor RH, Busman M, Seo JA, Lee YW, Plattner RD. A fumonisin biosynthetic gene cluster in *Fusarium oxysporum* strain O-1890 and the genetic basis for B versus C fumonisin production. Fungal genetics and biology: FG & B. 2008; 45(6):1016–26. Epub 2008/04/01. https://doi.org/10.1016/j.fgb.2008.02.004 PMID: 18375156.

163. Zaleta-Rivera K, Xu C, Yu F, Butchko RA, Proctor RH, Hidalgo-Lara ME, et al. A Bidomain Nonribosomal Peptide Synthetase Encoded by FUM14 Catalyzes the Formation of Tricarballylic Esters in the Biosynthesis of Fumonisins. Biochemistry 2006; 45:2561–9. https://doi.org/10.1021/bi052085s PMID: 16489749

164. Butchko RA, Plattner RD, Proctor RH. Deletion Analysis of FUM Genes Involved in Tricarballylic Ester Formation during Fumonisin Biosynthesis. J Agric Food Chem. 2006; 54:9398−404. https://doi.org/10.1021/jf0617869 PMID: 17147424

165. Butchko RA, Plattner RD, Proctor RH. FUM13 Encodes a Short Chain Dehydrogenase/Reductase Required for C-3 Carbonyl Reduction during Fumonisin Biosynthesis in *Gibberella moniliformis*. J Agric Food Chem 2003; 51:3000−6. https://doi.org/10.1021/jf0262007 PMID: 12720383

166. Butchko RA, Plattner RD, Proctor RH. FUM9 is required for C-5 hydroxylation of fumonisins and complements the meitotically defined Fum3 locus in *Gibberella moniliformis*. Appl Environ Microbiol. 2003; 69(11):6935–7. Epub 2003/11/07. https://doi.org/10.1128/AEM.69.11.6935-6937.2003 PMID: 14602658; PubMed Central PMCID: PMC262316.

167. Brown DW, Butchko RA, Busman M, Proctor RH. The *Fusarium verticillioides* FUM gene cluster encodes a Zn(II)2Cys6 protein that affects FUM gene expression and fumonisin production. Eukaryot Cell. 2007; 6(7):1210–8. Epub 2007/05/08. https://doi.org/10.1128/EC.00400-06 PMID: 17483290; PubMed Central PMCID: PMC1951116.

168. Du L, Zhu X, Gerber R, Huffman J, Lou L, Jorgenson J, et al. Biosynthesis of sphinganine-analog mycotoxins. J Ind Microbiol Biotechnol. 2008; 35(6):455–64. Epub 2008/01/25. https://doi.org/10.1007/s10295-008-0316-y PMID: 18214562.

169. Proctor RH, Desjardins AE, Plattner RD, Hohn TM. A Polyketide Synthase Gene Required for Biosynthesis of Fumonisin Mycotoxins in *Gibberella fujikuroi* Mating Population A. Fungal Genetics and Biology 1999; 27:100–12. https://doi.org/10.1006/fgbi.1999.1141 PMID: 10413619

170. Lia Y, Lou L, Cerny RL, Butchko RA, Proctor RH, Shen Y, et al. Tricarballylic ester formation during biosynthesis of fumonisin mycotoxins in *Fusarium verticillioides*. Mycology. 2013; 4(4):179–86. Epub

2014/03/04. https://doi.org/10.1080/21501203.2013.874540 PMID: 24587959; PubMed Central PMCID: PMC3933019.

171. Studt L, Janevska S, Niehaus EM, Burkhardt I, Arndt B, Sieber CM, et al. Two separate key enzymes and two pathway-specific transcription factors are involved in fusaric acid biosynthesis in *Fusarium fujikuroi*. Environ Microbiol. 2016; 18(3):936–56. Epub 2015/12/15. https://doi.org/10.1111/1462-2920.13150 PMID: 26662839.

172. Lin X, Yuan S, Chen S, Chen B, Xu H, Liu L, et al. Heterologous Expression of Ilicicolin H Biosynthetic Gene Cluster and Production of a New Potent Antifungal Reagent, Ilicicolin J. Molecules. 2019; 24 (12). Epub 2019/06/21. https://doi.org/10.3390/molecules24122267 PMID: 31216742; PubMed Central PMCID: PMC6631495.

173. Cary JW, Uka V, Han Z, Buyst D, Harris-Coward PY, Ehrlich KC, et al. An *Aspergillus flavus* secondary metabolic gene cluster containing a hybrid PKS-NRPS is necessary for synthesis of the 2-pyridones, leporins. Fungal Genet Biol. 2015; 81:88–97. Epub 2015/06/09. https://doi.org/10.1016/j.fgb.2015.05.010 PMID: 26051490.

174. Manzoni M, Rollini M. Biosynthesis and biotechnological production of statins by filamentous fungi and application of these cholesterol-lowering drugs. Appl Microbiol Biotechnol. 2002; 58(5):555–64. Epub 2002/04/17. https://doi.org/10.1007/s00253-002-0932-9 PMID: 11956737.

175. Zhang W, Cao S, Qiu L, Qi F, Li Z, Yang Y, et al. Functional characterization of MpaG', the O-methyltransferase involved in the biosynthesis of mycophenolic acid. Chembiochem. 2015; 16(4):565–9. Epub 2015/01/30. https://doi.org/10.1002/cbic.201402600 PMID: 25630520.

176. Zhang W, Du L, Qu Z, Zhang X, Li F, Li Z, et al. Compartmentalized biosynthesis of mycophenolic acid. Proc Natl Acad Sci U S A. 2019; 116(27):13305–10. Epub 2019/06/19. https://doi.org/10.1073/pnas.1821932116 PMID: 31209052; PubMed Central PMCID: PMC6613074.

177. Hansen BG, Genee HJ, Kaas CS, Nielsen JB, Regueira TB, Mortensen UH, et al. A new class of IMP dehydrogenase with a role in self-resistance of mycophenolic acid producing fungi. BMC Microbiol. 2011; 11:202. Epub 2011/09/20. https://doi.org/10.1186/1471-2180-11-202 PMID: 21923907; PubMed Central PMCID: PMC3184278.

178. Hansen BG, Salomonsen B, Nielsen MT, Nielsen JB, Hansen NB, Nielsen KF, et al. Versatile enzyme expression and characterization system for *Aspergillus nidulans*, with the *Penicillium brevicompactum* polyketide synthase gene from the mycophenolic acid gene cluster as a test case. Appl Environ Microbiol. 2011; 77(9):3044–51. Epub 2011/03/15. https://doi.org/10.1128/AEM.01768-10 PMID: 21398493; PubMed Central PMCID: PMC3126399.

179. Regueira TB, Kildegaard KR, Hansen BG, Mortensen UH, Hertweck C, Nielsen J. Molecular basis for mycophenolic acid biosynthesis in *Penicillium brevicompactum*. Appl Environ Microbiol. 2011; 77 (9):3035–43. Epub 2011/03/15. https://doi.org/10.1128/AEM.03015-10 PMID: 21398490; PubMed Central PMCID: PMC3126426.

180. Hansen BG, Mnich E, Nielsen KF, Nielsen JB, Nielsen MT, Mortensen UH, et al. Involvement of a natural fusion of a cytochrome P450 and a hydrolase in mycophenolic acid biosynthesis. Appl Environ Microbiol. 2012; 78(14):4908–13. Epub 2012/05/01. https://doi.org/10.1128/AEM.07955-11 PMID: 22544261; PubMed Central PMCID: PMC3416377.

181. Del-Cid A, Gil-Duran C, Vaca I, Rojas-Aedo JF, Garcia-Rico RO, Levican G, et al. Identification and Functional Analysis of the Mycophenolic Acid Gene Cluster of *Penicillium roqueforti*. PLoS One. 2016; 11(1):e0147047. Epub 2016/01/12. https://doi.org/10.1371/journal.pone.0147047 PMID: 26751579; PubMed Central PMCID: PMC4708987.

182. Gillot G, Jany JL, Dominguez-Santos R, Poirier E, Debaets S, Hidalgo PI, et al. Genetic basis for mycophenolic acid production and strain-dependent production variability in *Penicillium roqueforti*. Food Microbiol. 2017; 62:239–50. Epub 2016/11/28. https://doi.org/10.1016/j.fm.2016.10.013 PMID: 27889155.

183. Scott B, Young CA, Saikia S, McMillan LK, Monahan BJ, Koulman A, et al. Deletion and gene expression analyses define the paxilline biosynthetic gene cluster in *Penicillium paxilli*. Toxins (Basel). 2013; 5(8):1422–46. Epub 2013/08/21. https://doi.org/10.3390/toxins5081422 PMID: 23949005; PubMed Central PMCID: PMC3760044.

184. Fierro F, Garcia-Estrada C, Castillo NI, Rodriguez R, Velasco-Conde T, Martin JF. Transcriptional and bioinformatic analysis of the 56.8 kb DNA region amplified in tandem repeats containing the penicillin gene cluster in *Penicillium chrysogenum*. Fungal Genet Biol. 2006; 43(9):618–29. Epub 2006/05/23. https://doi.org/10.1016/j.fgb.2006.03.001 PMID: 16713314.

185. Xu X, Liu L, Zhang F, Wang W, Li J, Guo L, et al. Identification of the first diphenyl ether gene cluster for pestheic acid biosynthesis in plant endophyte *Pestalotiopsis fici*. Chembiochem. 2014; 15(2):284–92. Epub 2013/12/05. https://doi.org/10.1002/cbic.201300626 PMID: 24302702.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

186. Chen L, Yue Q, Zhang X, Xiang M, Wang C, Li S, et al. Genomics-driven discovery of the pneumocandin biosynthetic gene cluster in the fungus *Glarea lozoyensis*. BMC Genomics. 2013; 14(339). https://doi.org/10.1186/1471-2164-14-339 PMID: 23688303

187. Chen L, Li Y, Yue Q, Loksztejn A, Yokoyama K, Felix EA, et al. Engineering of New Pneumocandin Side-Chain Analogues from *Glarea lozoyensis* by Mutasynthesis and Evaluation of Their Antifungal Activity. ACS Chem Biol. 2016; 11(10):2724–33. Epub 2016/10/22. https://doi.org/10.1021/acschembio.6b00604 PMID: 27494047; PubMed Central PMCID: PMC5502478.

188. Chen L, Yue Q, Li Y, Niu X, Xiang M, Wang W, et al. Engineering of *Glarea lozoyensis* for exclusive production of the pneumocandin B0 precursor of the antifungal drug caspofungin acetate. Appl Environ Microbiol. 2015; 81(5):1550–8. Epub 2014/12/21. https://doi.org/10.1128/AEM.03256-14 PMID: 25527531; PubMed Central PMCID: PMC4325176.

189. Salo O, Guzman-Chavez F, Ries MI, Lankhorst PP, Bovenberg RAL, Vreeken RJ, et al. Identification of a Polyketide Synthase Involved in Sorbicillin Biosynthesis by *Penicillium chrysogenum*. Appl Environ Microbiol. 2016; 82(13):3971–8. Epub 2016/04/24. https://doi.org/10.1128/AEM.00350-16 PMID: 27107123; PubMed Central PMCID: PMC4907180.

190. Guzman-Chavez F, Salo O, Nygard Y, Lankhorst PP, Bovenberg RAL, Driessen AJM. Mechanism and regulation of sorbicillin biosynthesis by *Penicillium chrysogenum*. Microb Biotechnol. 2017; 10(4):958–68. https://doi.org/10.1111/1751-7915.12736 PMID: 28618182; PubMed Central PMCID: PMC5481523.

191. Derntl C, Guzman-Chavez F, Mello-de-Sousa TM, Busse HJ, Driessen AJM, Mach RL, et al. In Vivo Study of the Sorbicillinoid Gene Cluster in *Trichoderma reesei*. Frontiers in microbiology. 2017; 8:2037. https://doi.org/10.3389/fmicb.2017.02037 PMID: 29104566; PubMed Central PMCID: PMC5654950.

192. Heneghan MN, Yakasai AA, Halo LM, Song Z, Bailey AM, Simpson TJ, et al. First heterologous reconstruction of a complete functional fungal biosynthetic multigene cluster. Chembiochem. 2010; 11(11):1508–12. Epub 2010/06/25. https://doi.org/10.1002/cbic.201000259 PMID: 20575135.

193. Halo LM, Heneghan MN, Yakasai AA, Song Z, Williams K, Bailey AM, et al. Late Stage Oxidations during the Biosynthesis of the 2-Pyridone Tenellin in the Entomopathogenic Fungus *Beauveria bassiana*. J Am Chem Soc. 2008; 130:17988–96. https://doi.org/10.1021/ja807052c PMID: 19067514

194. Zaehle C, Gressler M, Shelest E, Geib E, Hertweck C, Brock M. Terrein biosynthesis in *Aspergillus terreus* and its impact on phytotoxicity. Chem Biol. 2014; 21(6):719–31. Epub 2014/05/13. https://doi.org/10.1016/j.chembiol.2014.03.010 PMID: 24816227.

195. Kakule TB, Zhang S, Zhan J, Schmidt EW. Biosynthesis of the Tetramic Acids Sch210971 and Sch210972. Organic Letters. 2015; 17(10):2295–7. https://doi.org/10.1021/acs.orglett.5b00715 PMID: 25885659

196. Umemura M, Nagano N, Koike H, Kawano J, Ishii T, Miyamura Y, et al. Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in *Aspergillus flavus*. Fungal Genet Biol. 2014; 68:23–30. Epub 2014/05/21. https://doi.org/10.1016/j.fgb.2014.04.011 PMID: 24841822.

197. Lim FY, Won TH, Raffa N, Baccile JA, Wisecaver J, Rokas A, et al. Fungal Isocyanide Synthases and Xanthocillin Biosynthesis in Aspergillus fumigatus. mBio. 2018; 9(3). Epub 2018/05/31. https://doi.org/10.1128/mBio.00785-18 PMID: 29844112; PubMed Central PMCID: PMC5974471.

198. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020; 21(1):6. Epub 2020/01/04. https://doi.org/10.1186/s12864-019-6413-7 PMID: 31898477; PubMed Central PMCID: PMC6941312.

199. Frishman WH, Rapier RC. Lovastatin: an HMG-CoA reductase inhibitor for lowering cholesterol. The Medical clinics of North America. 1989; 73(2):437–48. Epub 1989/03/01. https://doi.org/10.1016/s0025-7125(16)30681-2 PMID: 2645482.

200. Hutchinson CR, Kennedy J, Park C, Kendrew S, Auclair K, Vederas J. Aspects of the biosynthesis of non-aromatic fungal polyketides by iterative polyketide synthases. Antonie van Leeuwenhoek. 2000; 78(3):287–95. https://doi.org/10.1023/a:1010294330190 PMID: 11386351

201. Gauglitz G. Artificial vs. human intelligence in analytics. Analytical and bioanalytical chemistry. 2019; 411(22):5631–2. https://doi.org/10.1007/s00216-019-01972-2 PMID: 31240356

202. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature methods. 2015; 12(1):59–60. https://doi.org/10.1038/nmeth.3176 PMID: 25402007

203. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011; 39(Web Server issue):W29–W37. Epub 05/18. https://doi.org/10.1093/nar/gkr367 PMID: 21593126.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Representative calculation of the manual evaluation measure (MEM) and comparison of the results to the FunOrder output for determination of thresholds based on the 2-Pyridon-Desmethylbassianin (dmb) BGC from *Beauveria bassiana*

Two phylogenetic trees, each representing a gene within a cluster in the context of our empirically optimized database, were compared. For each tree comparison we first determined if there were similar leaves (similar Species) between the two trees. If yes, branch length differences, node-differences, branch colours and overall topology between the leaves and the query were determined.
The branch lengths were measured, and the differences then calculated. The nodes between a species and the query were counted and compared to the number of nodes of the other phylogenetic tree. The branch colour describes the similarity to the most common node based on the Robinson-Foulds (RF) distance. 0 to 40% similarity was defined as “yellow”, 40 – 66,6% similarity was defined as “green” and the rest was defined as “blue”. For each of the four measures average pairwise distances were determined. If the trees contained more than two similar leaves, another average would be calculated of the resulted average measures, called manual evaluation measure (MEM). These MEMs (the higher the MEM the higher the similarity) (S6 Table) were put together in matrices to calculate heatmaps, dendrograms and PCA to evaluate the FunOrder output based on the treeKO algorithm (lower distances indicated higher similarities).

Table 1 FunOrder strict matrix

	dmbS	*dmbA*	*dmbB*	*dmbC*
dmbS	0	0.524	0.864	0.672
dmbA	0.524	0	0.793	0.533
dmbB	0.864	0.793	0	0.807
dmbC	0.672	0.533	0.807	0

Table 2 FunOrder evol matrix

	dmbS	*dmbA*	*dmbB*	*dmbC*
dmbS	0	0.12	0	0.343
dmbA	0.153	0	0	0
dmbB	0	0	0	0.103
dmbC	0.322	0	0.103	0

The 2-Pyridon-Desmethylbassianin (dmb) BGC from *Beauveria bassiana* (BGC0001136) consists of 4 genes (*dmbA*, *dmbB*, *dmbC*, *dmbS*). According to literature *dmbS* and *dmbC* are needed for the production of 2-Pyridon-Desmethylbassianin (Heneghan, Yakasai et al. 2011).

We can now compare the highest MEMs to the corresponding strict distances from the FunOrder output to determine if they are comparable. The highest MEM was calculated for the *dmbA:dmbC* comparison (MEM =2.61), in the FunOrder output for the strict distance this comparison (0.533) is next to lowest (Table 1 and S6 Table). The evolutionary distance for this comparison is 0 (Table 2). Next the the *dmbA:dmbS* comparison (MEM = 2.53) is the lowest in the strict matrix but has the evolutionary distance (0.12 and 0.153) (The differences between the two values are created due to the treeKO algorithm and the decision which tree to use as reference). This clarifies the strength of the MEMs, that they consider evolutionary history in one value, and the introduction of the combined distance measure, where speciation is considered with the strict distance as background. The comparison *dmbS:dmbC* had a MEM of 2.4 and a strict distance of 0.672. When comparing the clustering of the Ward`s minimum variance on the unscaled data, we can further observe similar clustering (Figure 1 A and B).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

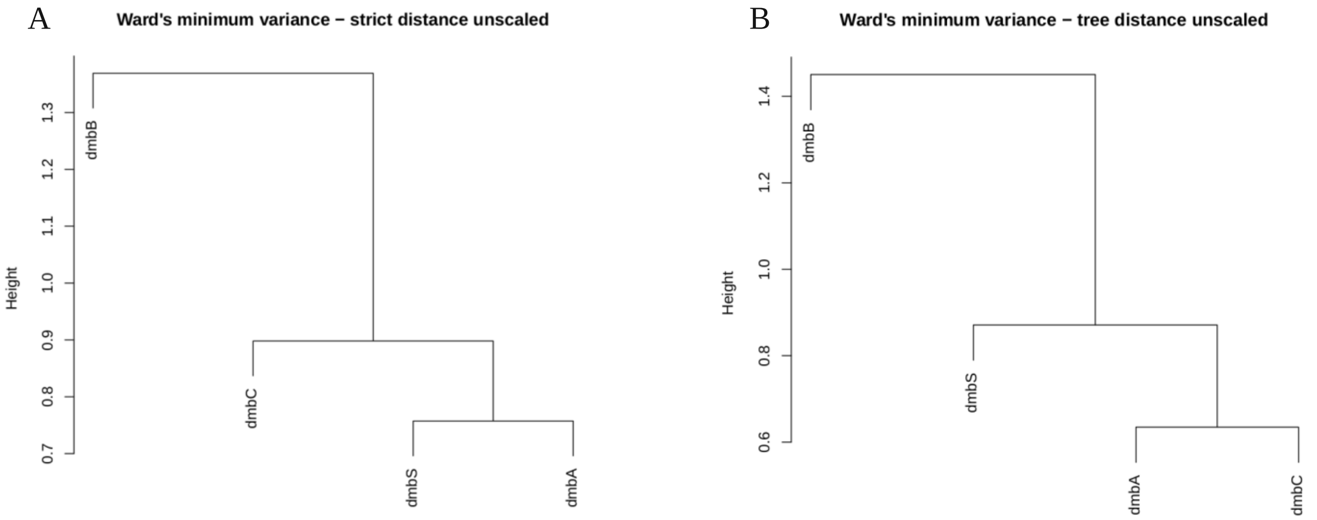


Figure 1 A – Standard output of the FunOrder analysis of the 2-Pyridon-Desmethylbassianin BGC of *Beauveria bassiana* (BGC0001136) (dmb). Dendrogram based on the Euclidean distance within the unscaled strict distance matrix clustered using Ward's minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of the dissimilarities before cluster updating. **B** – Dendrogram based on the Euclidean distance within the unscaled MEM matrix of the 2-Pyridon-Desmethylbassianin BGC of *Beauveria bassiana* (BGC0001136) (dmb) clustered using Ward's minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of the dissimilarities before cluster updating.

In this example, we can see clear similar clustering (Figure 1) between the dendrograms based on the MEM values and those based on the strict distances. Further, we can see how strict distance values below 0.7 reflect manual determination and refinement of co-evolution (Table 1, Table 2 and S6 Table). We can further see how the manual determination of co-evolution takes speciation history into account and therefore the validity of the introduction of the combined distance, which resembles the manual comparison. We further compared Heatmaps and PCAs and performed these comparisons for all analysed positive control BGCs and negative control gene clusters.

References:

Heneghan, M. N., A. A. Yakasai, K. Williams, K. A. Kadir, Z. Wasil, W. Bakeer, K. M. Fisch, A. M. Bailey, T. J. Simpson, R. J. Cox and C. M. Lazarus (2011). "The programming role of trans-acting enoyl reductases during the biosynthesis of highly reduced fungal polyketides." Chemical Science **2**(5).

Standard Operation Procedure for the Interpretation of the FunOrder Results

1) The internal co-evolution quotient (ICQ) gives information about the total co-evolution within an insert BGC. If the ICQ is above 0.718, the content of the co-evolution is not significantly different to randomly assembled GCs. Such BGCs must be interpreted with extreme caution. It might be worth trying to add or remove some gene from the edges of the BGC and re-run the FunOrder analysis.
2) The heatmap based on the strict distances is representing the calculated raw data (treeKO output) and can be used for an initial, general overview. Genes with a shared co-evolution may cluster together and form distinct clusters in the heatmap (regardless of the absolute values) and the corresponding dendogram. Such clusters may be a good first indication for co-evolution but are not a necessity.
3) Next, the dendrogram based on the euclidean distances within the scaled strict distance matrix is inspected. In the context of BGCs, it is sensible to look for genes that cluster together with the core enzyme(s) in the dendrogram. Notably, the dendrogram is still a representation of the complete data set. Clustering (or the absence of clustering) may not only be caused by co-evolution but also by potential generated noise.
4) Therefore, the final and crucial step to detect co-evolving genes is to consider the PCA-plot of the strict distance. First, the percentage described by the principal components must be compared and taken into account for the clustering. For example, if PC1 (x-axis) describes 50% of the data and PC2 (y-axis) describes 10% of the data, longer vertical distances between genes are allowed, because of the stronger horizontal impact. Genes that cluster together with the core enzyme(s) are highly likely to share a similar co-evolution.
5) Genes that are clustering together with the core enzyme(s) in any of the three visualisations are considered ‘detected’ and can be anticipated to share a similar co-evolution.
6) Finally, the steps 2 – 4 are repeated with the visualizations of the combined distances. This may add further genes to the pool of ‘detected’ genes.

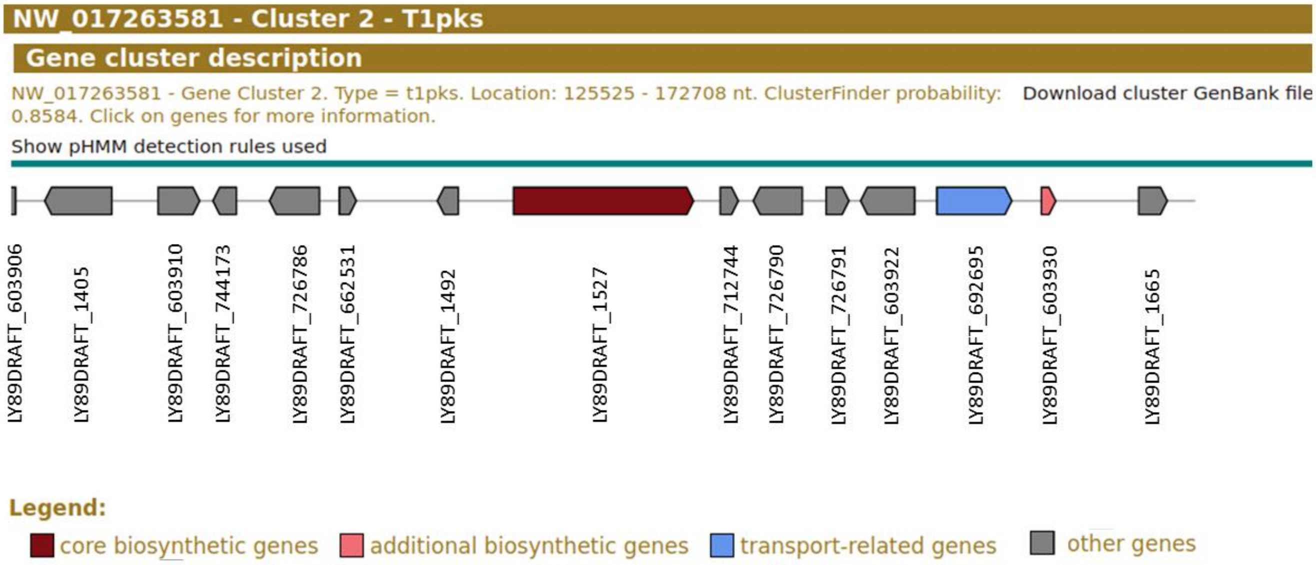
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

S3 File

Phialocephala scopiformis biosynthetic gene cluster analysis with FunOrder

To give an example for the FunOrder analysis of an undescribed biosynthetic gene cluster (BGC), we chose a putative Type I polyketide synthase (T1pks) BGC from the fungal conifer needle endophyte *Phialocephala scopiformis* (1) (located on scaffold NW_017263581, 125525-172708 nt). This cluster was predicted with antiSMASH 4.3.0 (2) (Figure 1) and the output was directly analyzed with FunOrder.



NW_017263581 - Cluster 2 - T1pks

Gene cluster description

NW_017263581 - Gene Cluster 2. Type = t1pks. Location: 125525 - 172708 nt. ClusterFinder probability: 0.8584. Click on genes for more information.

Download cluster GenBank file

Show pHMM detection rules used

Figure 1 Screenshot of the cluster defined by antiSMASH.

The first step of the analysis was to inspect the internal co-evolutionary quotient (ICQ) calculated for this specific cluster. The ICQ was 0.5727, which is below the previously defined threshold for relevant co-evolution detected of 0.718. We therefore continued with the inspection of the heatmap based on the strict distance matrix (Figure 2). The color key in the heatmap is a direct visualization of the values of the strict distance, they are clustered based on a calculated dendrogram based on the complete linkage method. We observed a first indication of which genes might share a potential co-evolution with the core enzyme LY89DRAFT_1527 (marked as LY89DRAFT_1527_T1PKS in all figures). The inspection of figure 2 indicated LY89DRAFT_1492 (annotated as hypothetical protein and after a sequence similarity search with blastp (3) against the non redundant protein database revealed as putative serine hydrolase), LY89DRAFT_603930 (annotated as NAD(P)-binding protein and a smCOG short chain dehydrogenase/reductase) and LY89DRAFT_603910 (annotated as type 1 phosphodiesterase/nucleotide pyrophosphatase) as sharing relatively lower distance values among each other.

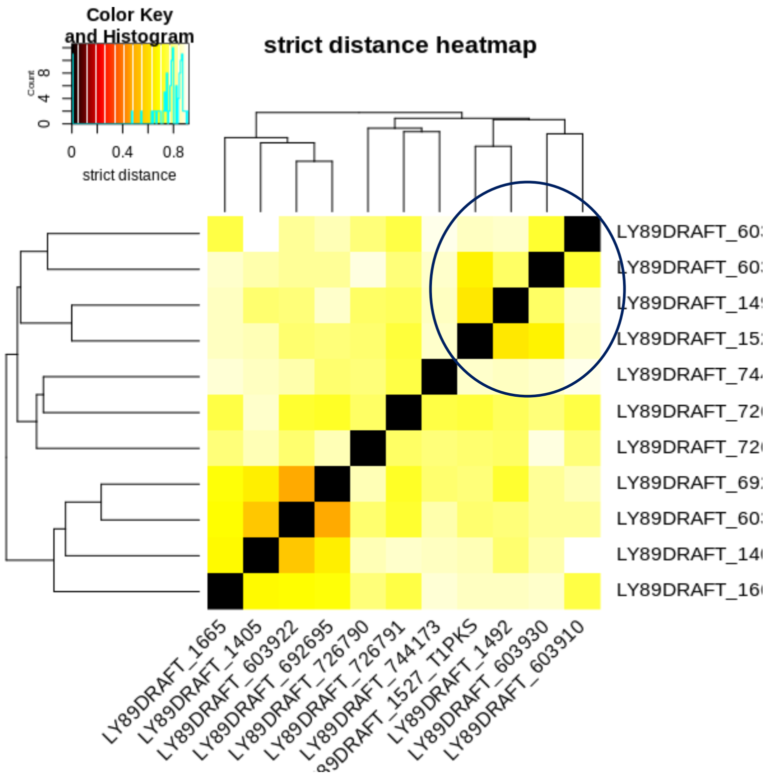


Figure 2 Standard output of the analysis of the putative T1pks BGC of *Phialocephala scopiformis* (located on scaffold NW_017263581, 125525-172708 nt). Heatmap of the strict distance matrix. The clustering mentioned in the text is indicated by a blue circle.

Next we examined the dendrogram (Figure 3) based on the Euclidean distances within the scaled strict distance matrix clustered using Ward´s minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of the dissimilarities before cluster updating. Again, we looked for the core enzyme LY89DRAFT_1527. This enabled us to determine that LY89DRAFT_1492 seems to share the strongest similarity in strict distances with LY89DRAFT_1527. Clustering relatively close to the T1pks were LY89DRAFT_603930 and LY89DRAFT_603910. This clustering considered the complete strict

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

distance matrix, including potential noise, which could distort the detection of true co-evolution within the BGC.

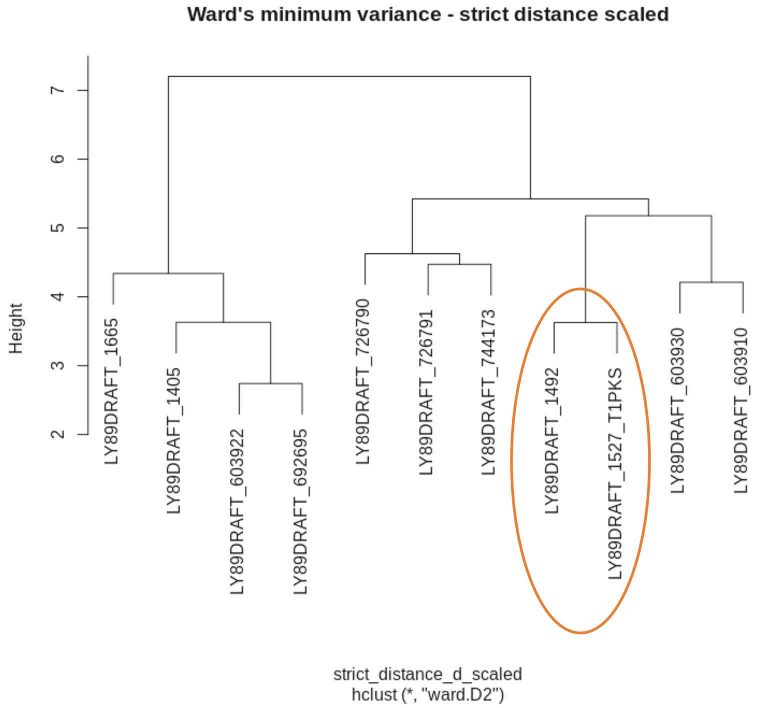


Figure 3 Standard output of the analysis of the putative T1pks BGC of *Phialocephala scopiformis* (located on scaffold NW_017263581, 125525-172708 nt). Dendrogram based on the Euclidean distances within the scaled strict distance matrix clustered using Ward´s minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of the dissimilarities before cluster updating. The clustering mentioned in the text is indicated by an orange circle.

We moved on to evaluate the score plot of the first two principal components (PC) of the principal component analysis (PCA) performed on the strict distance matrix (Figure 4 A). After inspecting the explained percentage of variance from each PC (indicated as Comp 1 and Comp 2 in Figure 4), we observed a clear clustering of the core enzyme LY89DRAFT_1527 with LY89DRAFT_1492 and LY89DRAFT_603930. Whereas LY89DRAFT_603910 clearly clustered with a different group of genes. This clustering pattern is further supported by the score plot of the first two PC of the PCA performed on the combined distance matrix (Figure 4 B).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

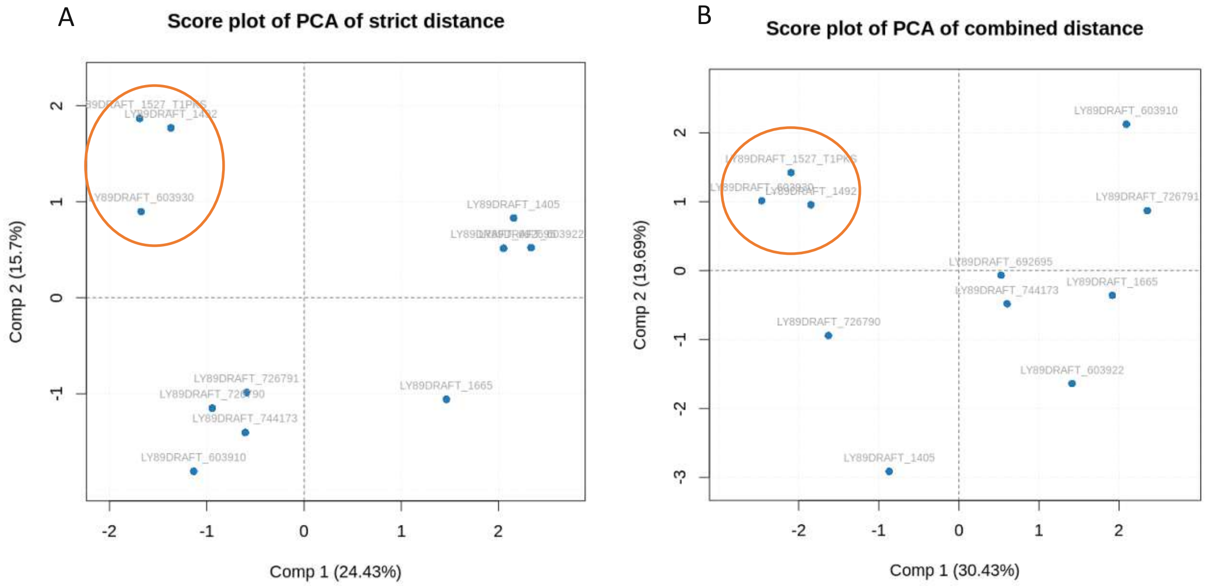


Figure 4 Standard output of the analysis of the putative T1pks BGC of *Phialocephala scopiformis* (located on scaffold NW_017263581, 125525-172708 nt). A - Score plot of the first two principal components (PC) of the principal component analysis (PCA) performed on the strict distance matrix. The clustering mentioned in the text is indicated by an orange circle. B - Score plot of the first two PC of the PCA performed on the combined distance matrix. The clustering mentioned in the text is indicated by an orange circle.

This lead to the hypothesis, that the T1pks LY89DRAFT_1527 with the putative serine hydrolase LY89DRAFT_1492 and the putative short chain dehydrogenase/reductase LY89DRAFT_603930 are responsible for the biosynthesis of the secondary metabolite (SM) encoded in this BGC, because they exhibit a shared co-evolution based on the FunOrder analysis. This hypothesis would have to be verified by corresponding *in-vitro*/*in-vivo* methods.

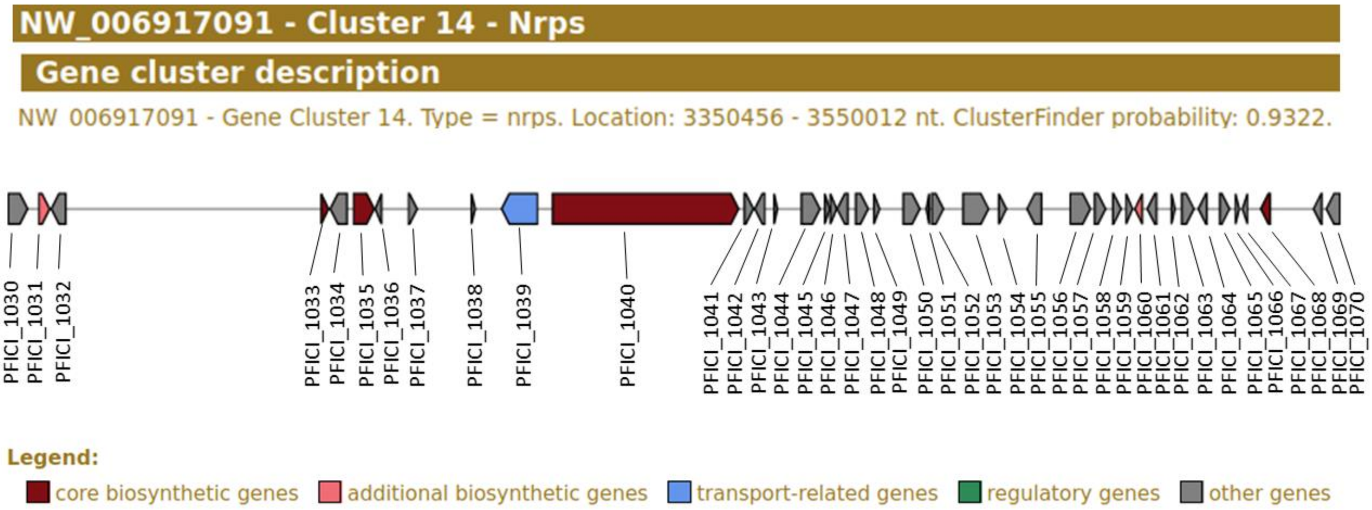
References:

1. Walker AK, Frasz SL, Seifert KA, Miller JD, Mondo SJ, LaButti K, et al. Full Genome of Phialocephala scopiformis DAOMC 229536, a Fungal Endophyte of Spruce Producing the Potent Anti-Insectan Compound Rugulosin. Genome Announc. 2016;4(1).
2. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45(W1):W36-W41.
3. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421-.

S4 File

Pestalotiopsis fici biosynthetic gene cluster analysis with FunOrder

To give an example for the FunOrder analysis of an undescribed biosynthetic gene cluster (BGC), we chose a putative Non-ribosomal peptide synthetase (NRPS) BGC from *Pestalotiopsis fici* (1) (located on scaffold NW_006917091, 3350456 - 3550012 nt). This cluster was predicted with antiSMASH 4.3.0 (2) (Figure 1) and the output was directly analyzed with FunOrder.



NW_006917091 - Cluster 14 - Nrps

Gene cluster description

NW_006917091 - Gene Cluster 14. Type = nrps. Location: 3350456 - 3550012 nt. ClusterFinder probability: 0.9322.

Legend:

■ core biosynthetic genes ■ additional biosynthetic genes ■ transport-related genes ■ regulatory genes ■ other genes

Figure 1 Screenshot of the cluster defined by antiSMASH.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

The first step of the analysis was to inspect the internal co-evolutionary quotient (ICQ) calculated for this specific cluster. The ICQ was 0.5908, which is below the previously defined threshold for relevant co-evolution detected of 0.718. We therefore continued with the inspection of the heatmap based on the strict distance matrix (Figure 2). The color key in the heatmap is a direct visualization of the values of the strict distance, they are clustered based on a calculated dendrogram based on the complete linkage method. We observed a first indication of which genes might share a potential co-evolution with the core enzyme PFICI_01040 (marked as PFICI_01040_NRPS in all figures). In this case, there was no significant clustering detectable with the core enzyme PFICI_01040. PFICI_01040 appeared to share mostly high strict distances to the other genes in the BGC.

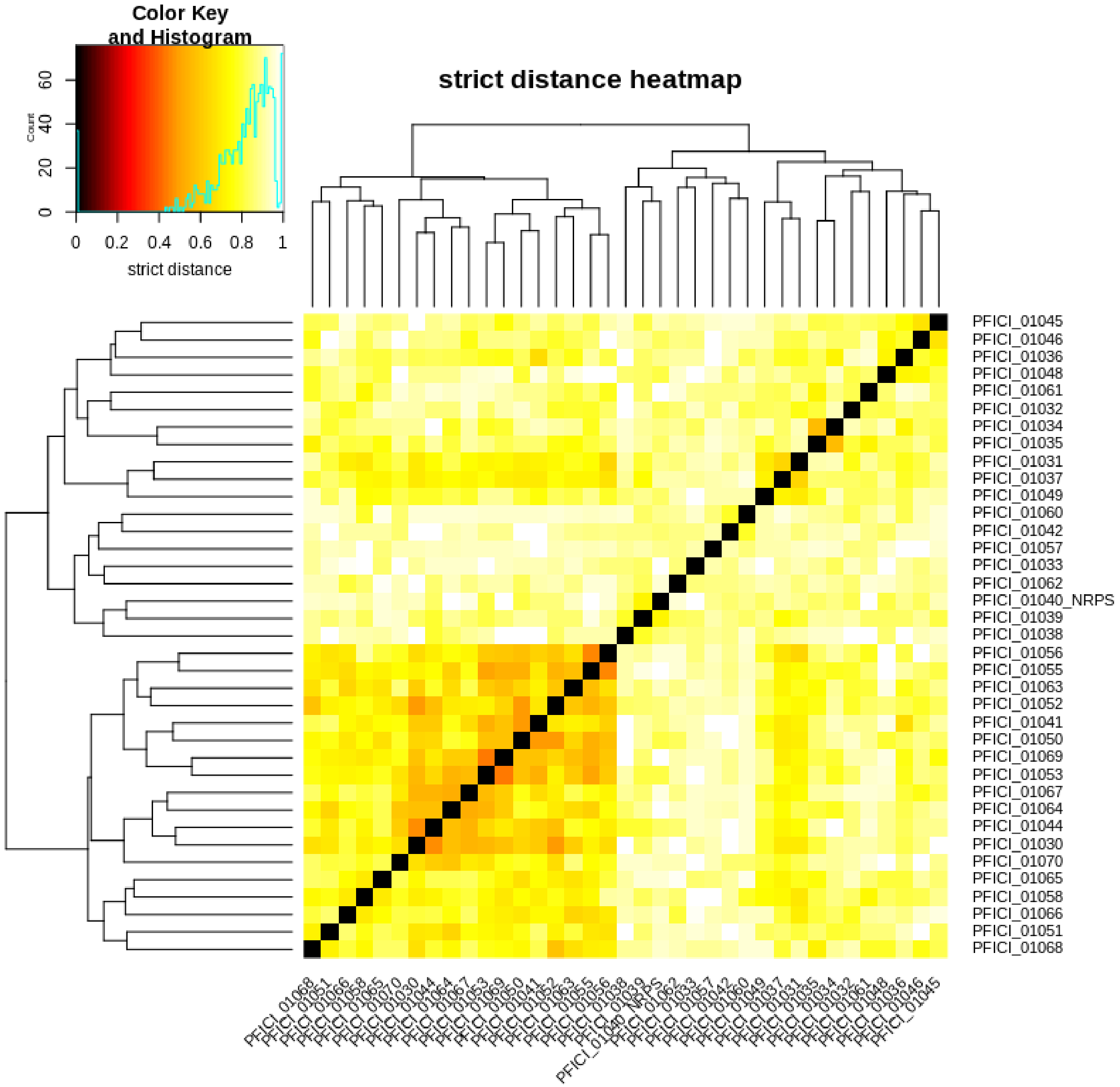


Figure 2 Standard output of the analysis of putative NRPS BGC from *Pestalotiopsis fici* (located on scaffold NW_006917091, 3350456 - 3550012 nt). Heatmap of the strict distance matrix.

Next we examined the dendrogram (Figure 3) based on the Euclidean distances within the scaled strict distance matrix clustered using Ward´s minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of the dissimilarities before cluster updating. Again, we looked for the core enzyme PFICI_01040. This enabled us to determine that PFICI_01040 seems to share the strongest co-evolution within the BGC with PFICI_01039 (annotated as hypothetical protein and recognized by antiSMASH as putative multi drug transporter) and PFICI_01038 (annotated as hypothetical protein and revealed to contain a putative DNA binding domain based on a sequence similarity search using blastp (3) against the non redundant protein database and conserved domain database). This clustering considered the complete strict distance matrix, including potential noise, which could distort the detection of true co-evolution within the BGC.

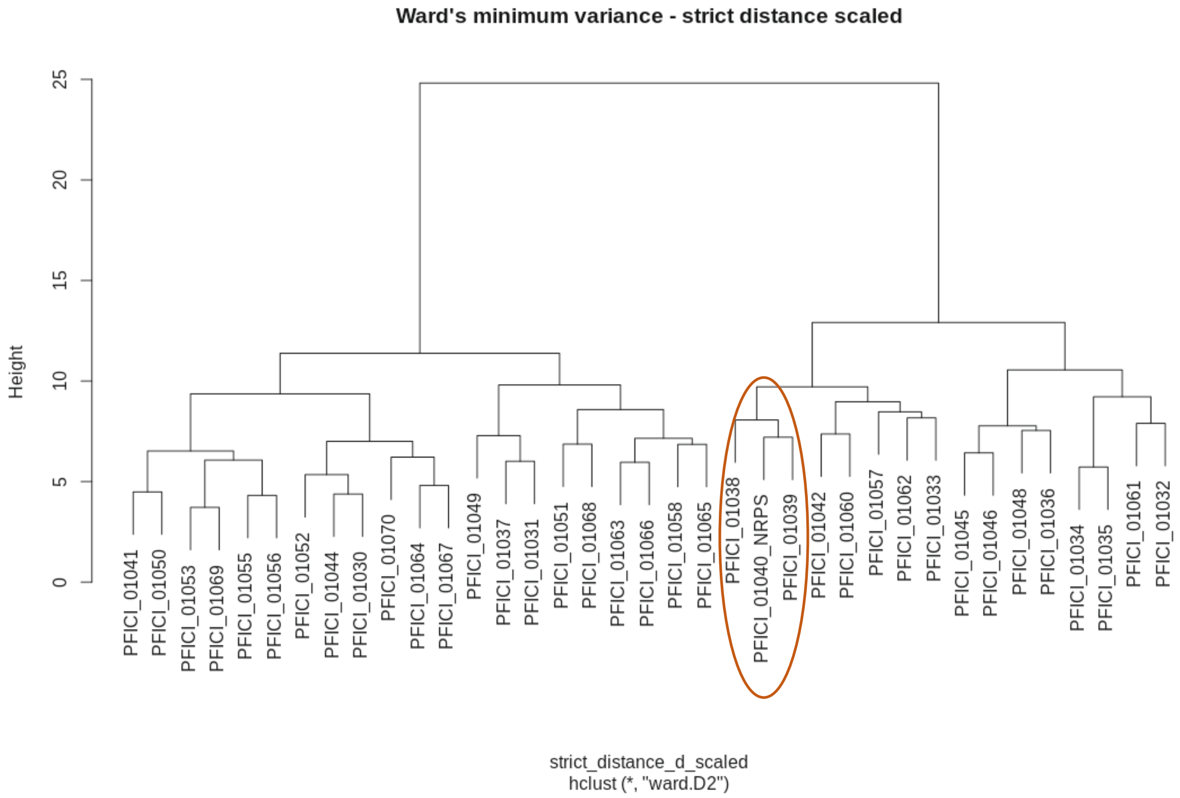


Figure 3 Standard output of the analysis of the putative NRPS BGC from *Pestalotiopsis fici* (located on scaffold NW_006917091, 3350456 - 3550012 nt). Dendrogram based on the Euclidean distances within the scaled strict distance matrix clustered using Ward´s minimum variance method aiming at finding compact spherical clusters, with the implemented squaring of the dissimilarities before cluster updating. The clustering mentioned in the text is indicated by an orange circle.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

We moved on to evaluate the score plot of the first two principal components (PC) of the principal component analysis (PCA) performed on the strict distance matrix (Figure 4). After inspecting the explained percentage of variance from each PC (indicated as Comp 1 and Comp 2 in Figure 4), we observed 6 genes clustering with the core enzyme PFICI_01040 (Table 1). As non-ribosomal peptides (NRP) produced by NRPS can undergo several modifications after their synthesis (4), the number of genes clustering is no surprise.

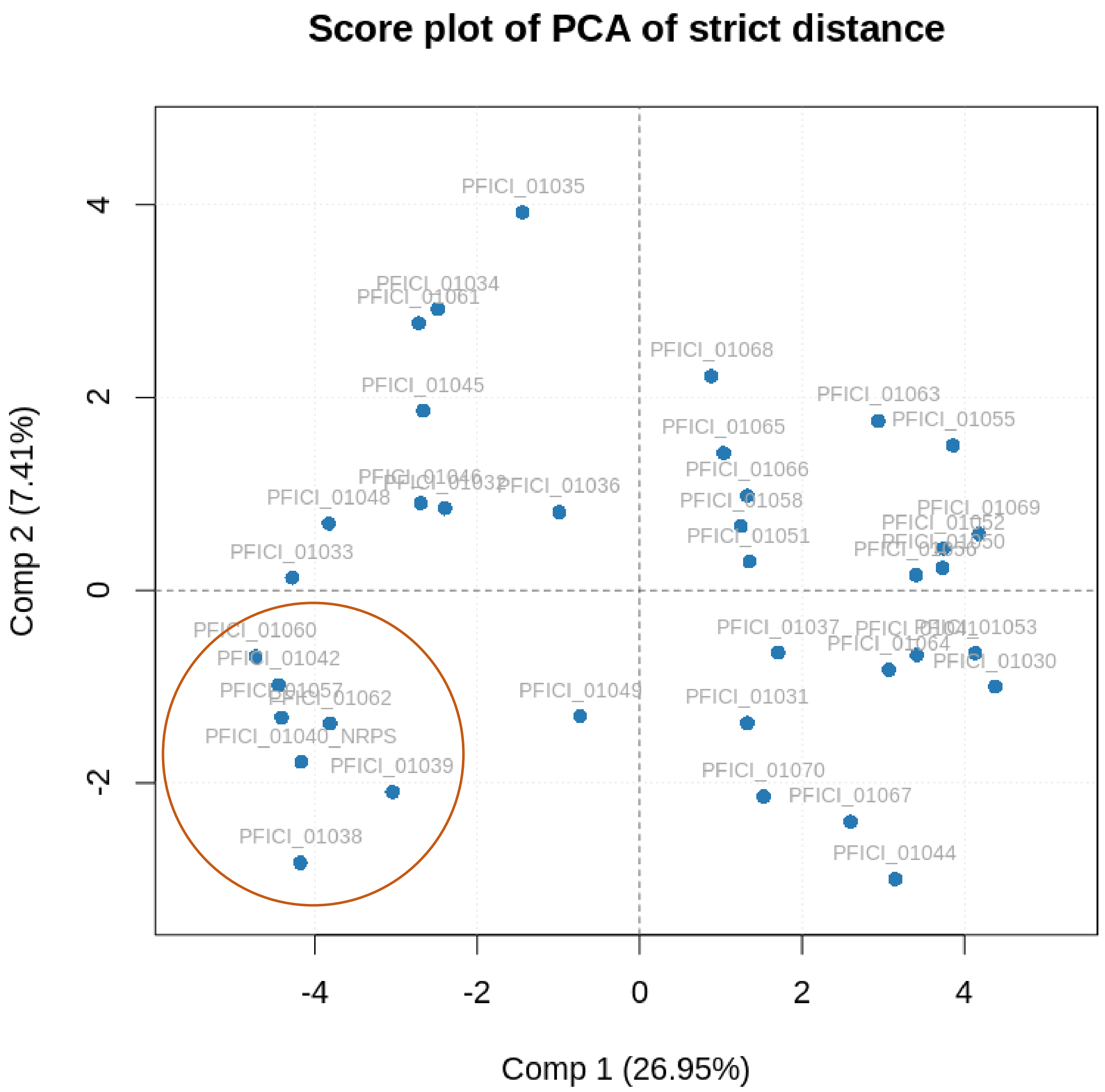


Figure 4 Standard output of the analysis of the putative NRPS BGC from *Pestalotiopsis fici* (located on scaffold NW_006917091, 3350456 - 3550012 nt). Score plot of the first two principal components (PC) of the principal component analysis (PCA) performed on the strict distance matrix. The clustering mentioned in the text is indicated by an orange circle.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Table 1 Genes clustering in the Score plot of the first two principal components (PC) of the principal component analysis (PCA) performed on the strict distance matrix with the core enzyme PFICI_01040.

Gene Locus-tag	**Annotation**	**Manual annotation**
PFICI_01040	Hypothetical protein	Putative NRPS
PFICI_01038	Hypothetical protein	Putative DNA binding domain containing protein
PFICI_01039	Hypothetical protein	putative multi drug transporter
PFICI_01062	Hypothetical protein	Putative Hydrophobic surface binding protein
PFICI_01057	Hypothetical protein	Putative feruloyl esterase
PFICI_01042	Hypothetical protein	Putative Peroxidase
PFICI_01060	Hypothetical protein	Putative Inositol monophosphatase

Nevertheless, this led to the hypothesis, that the NRPS PFICI_01040 produces a NRP that might be finally excreted by PFICI_01039, because they exhibit a shared co-evolution based on the FunOrder analysis. Further we hypothesized that PFICI_01038 might be involved in the regulation of the transcription of the NRPS gene. The enzymes encoded by the genes PFICI_01057 and PFICI_01042 may well play a part in the modification of the NRP. Besides, it could be possible that the NRPS already produces the final compound. This possibility was supported by the overall high strict distances shared by the core enzyme with the other genes of the BGC. These hypotheses would have to be verified by corresponding *in-vitro*/*in-vivo* methods.

References:

1. Wang X, Zhang X, Liu L, Xiang M, Wang W, Sun X, et al. Genomic and transcriptomic analysis of the endophytic fungus Pestalotiopsis fici reveals its lifestyle and high potential for synthesis of natural products. BMC Genomics. 2015;16:28.
2. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45(W1):W36-W41.
3. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421-.
4. Le Govic Y, Papon N, Le Gal S, Bouchara J-P, Vandeputte P. Non-ribosomal Peptide Synthetase Gene Clusters in the Human Pathogenic Fungus Scedosporium apiospermum. Frontiers in Microbiology. 2019;10(2062).

Statistical analysis of the Internal Co-evolutionary Quotient (ICQ)

All the statistical tests were performed in the R environment (1). The Shapiro-Wilk test used below was used to check for the normality of the ICQ data sets (table 1). Normality assumptions underlie outlier detection hypothesis tests. If the p-value is above the set alpha significance value (0.01) then the null hypothesis is not discarded. In other words it can be considered a normal distribution.

Table 1 Shapiro-Wilk normality tests.

ICQ data set	**p-value**
random GC	0.01901
BioPath	0.119
BGC	0.228
sequential GC	0.2093

Table 2 Levene's Test for Homogeneity of Variance (center = median) performed on the ICQ data sets from the analysis of the BioPath, BGCs, random GCs and sequential GCs.

	Df	**F value**	**Pr(>F)**
ICQ data sets	3	2.1335	0.09933

From the output in table 2, it can be seen that the p-value was not less than the significance level of 0.05. This means that there was no evidence to suggest that the variance is statistically significantly different for the data sets. Levene’s test is an alternative to Bartlett’s test when the data is not normally distributed.

Table 3 Computed one-way ANOVA test the analysis of variance performed on the ICQ values from the analysis of the BioPath, BGCs, random GCs and sequential GCs.

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
ICQ data sets	3	1.267	0.4224	33.45	6.11e-16
Residuals	125	1.579	0.0126		

The output in table 3 includes the columns F value and Pr(>F) corresponding to the p-value of the test. As the p-value is less than the significance level 0.05, we could conclude that there are significant differences between the ICQ data sets in the model summary. We could therefore continue to perform an analysis of variance (ANOVA). In one-way ANOVA test, a significant p-value indicates that some of the ICQ data sets means are different, but we don’t know which pairs of the ICQ data sets are different. It is possible to perform multiple pairwise-comparison, to determine if the mean difference between specific pairs are statistically significant. As the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

ANOVA test was significant, we could compute Tukey HSD (Tukey Honest Significant Differences), for performing multiple pairwise-comparison between the means of the ICQ data sets. It can be seen from the output in table 4 that only the differences are significant with an adjusted p-value lower than 0.05.

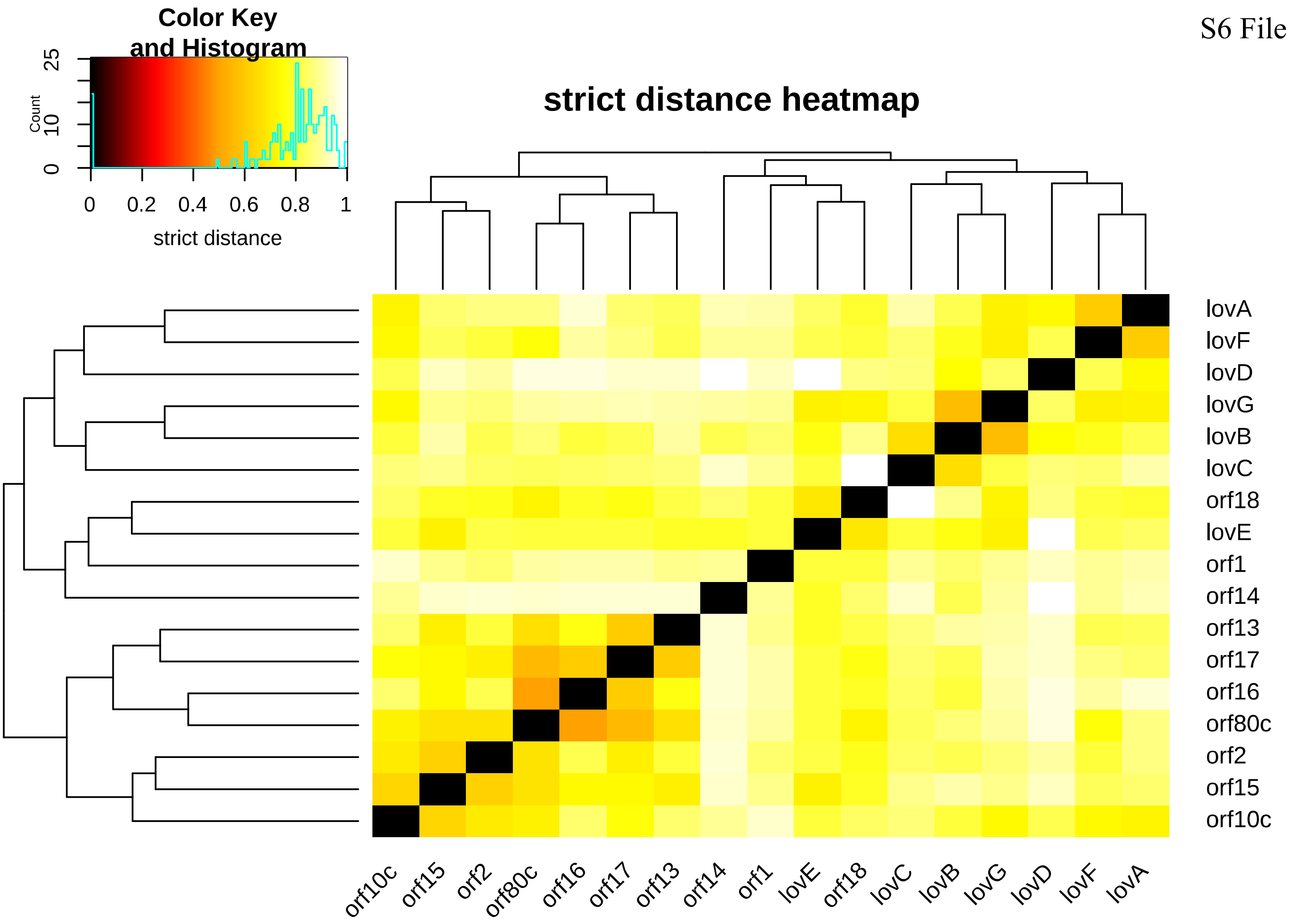
Table 4 Tukey multiple comparisons of means based on an ANOVA performed on the ICQ values from the analysis of the BioPath, BGCs, random GCs and sequential GCs with a 95% family-wise confidence level.

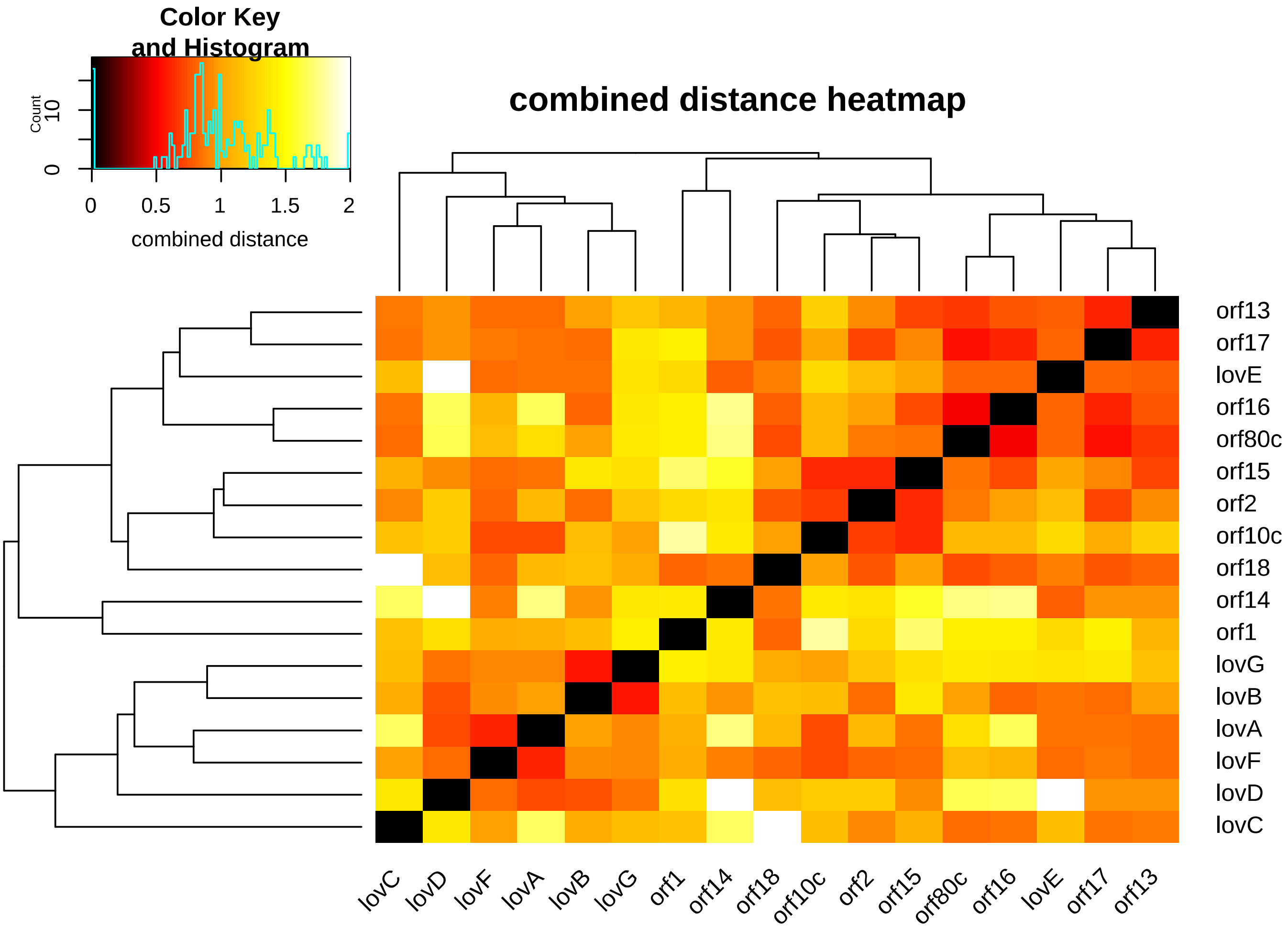
comparison	**diff**	**lwr**	**upr**	**p adj**
BioPath-BGC	-0.02870077	-0.13554943	0.07814789	0.8971149
random GC-BGC	0.2120088	0.14657762	0.27743997	**0**
sequential GC-BGC	0.05503562	-0.02116633	0.13123758	0.2416214
random GC-BioPath	0.24070957	0.14076179	0.34065734	**0**
sequential GC-BioPath	0.08373639	-0.02357184	0.19104462	0.1818826
sequential GC-random GC	-0.15697317	-0.22315216	-0.09079419	**0.0000001**

References:

1. R Core Team. (2019) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

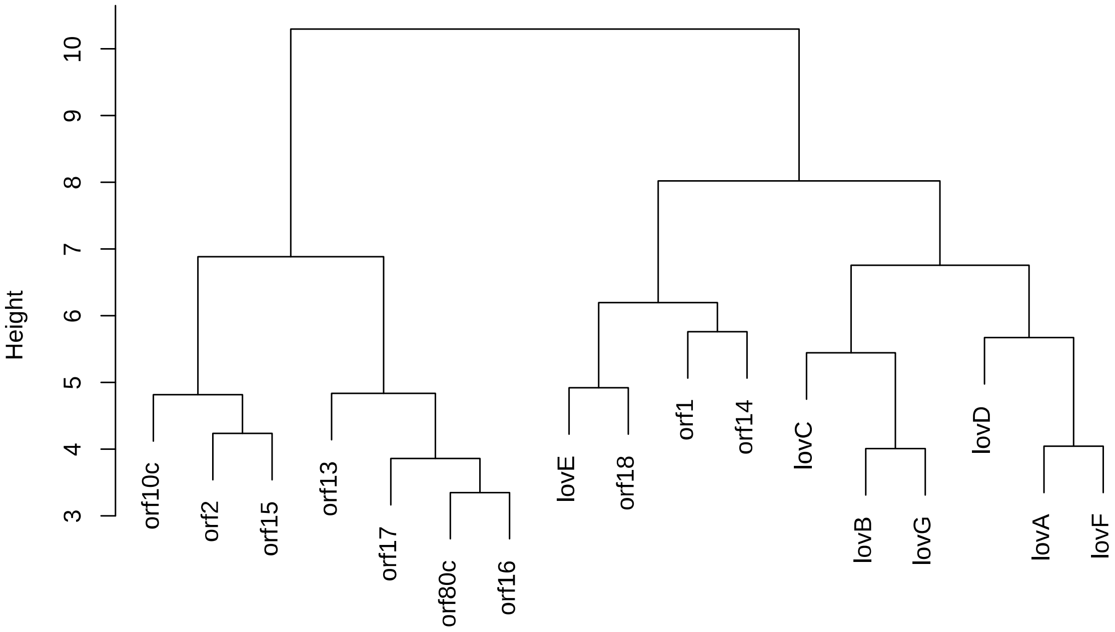
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.



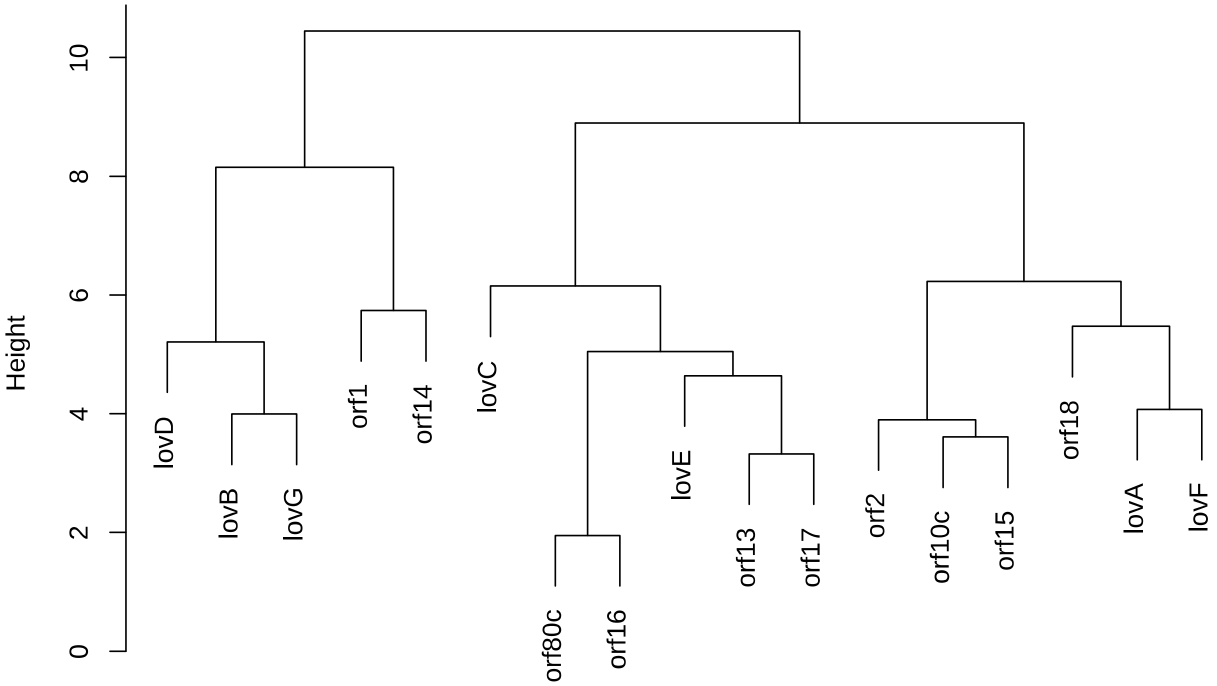


Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

S6 File

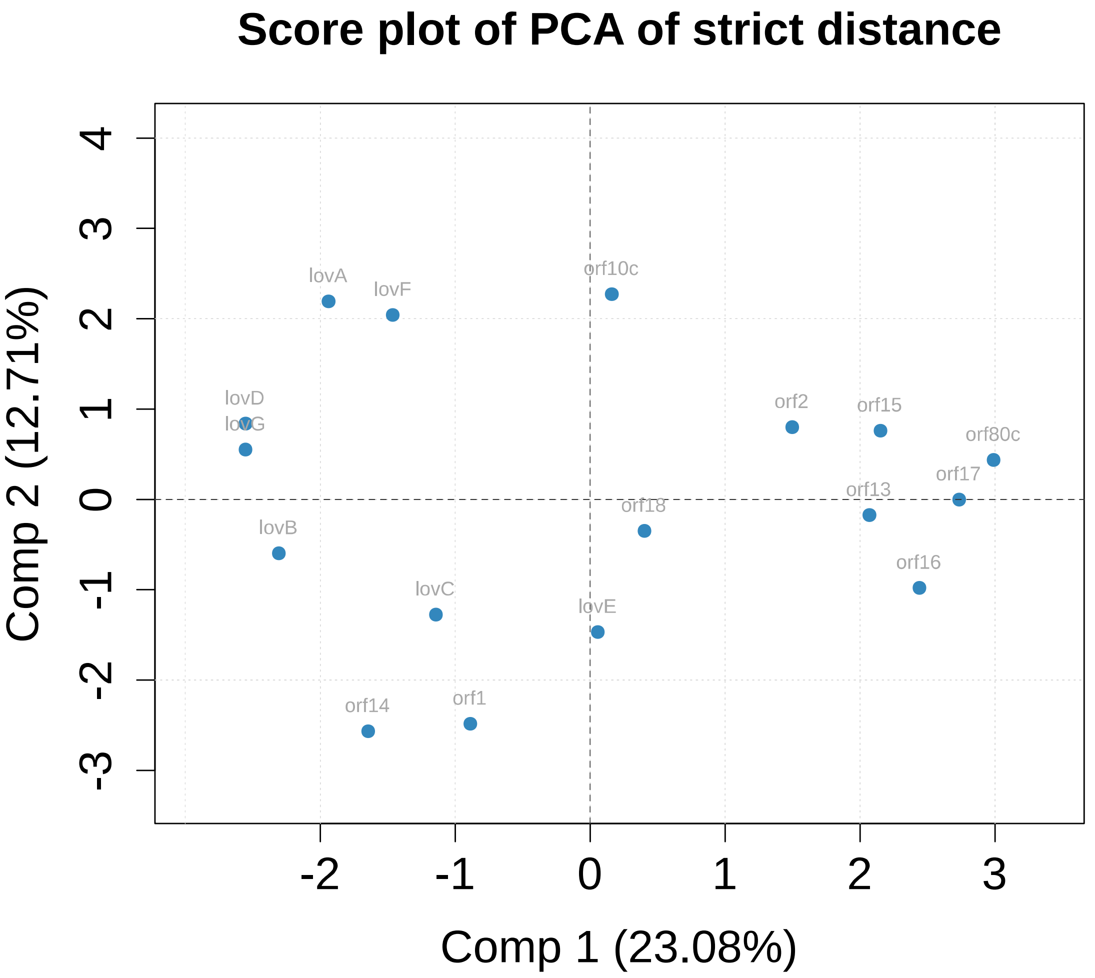

Ward's minimum variance - strict distance scaled

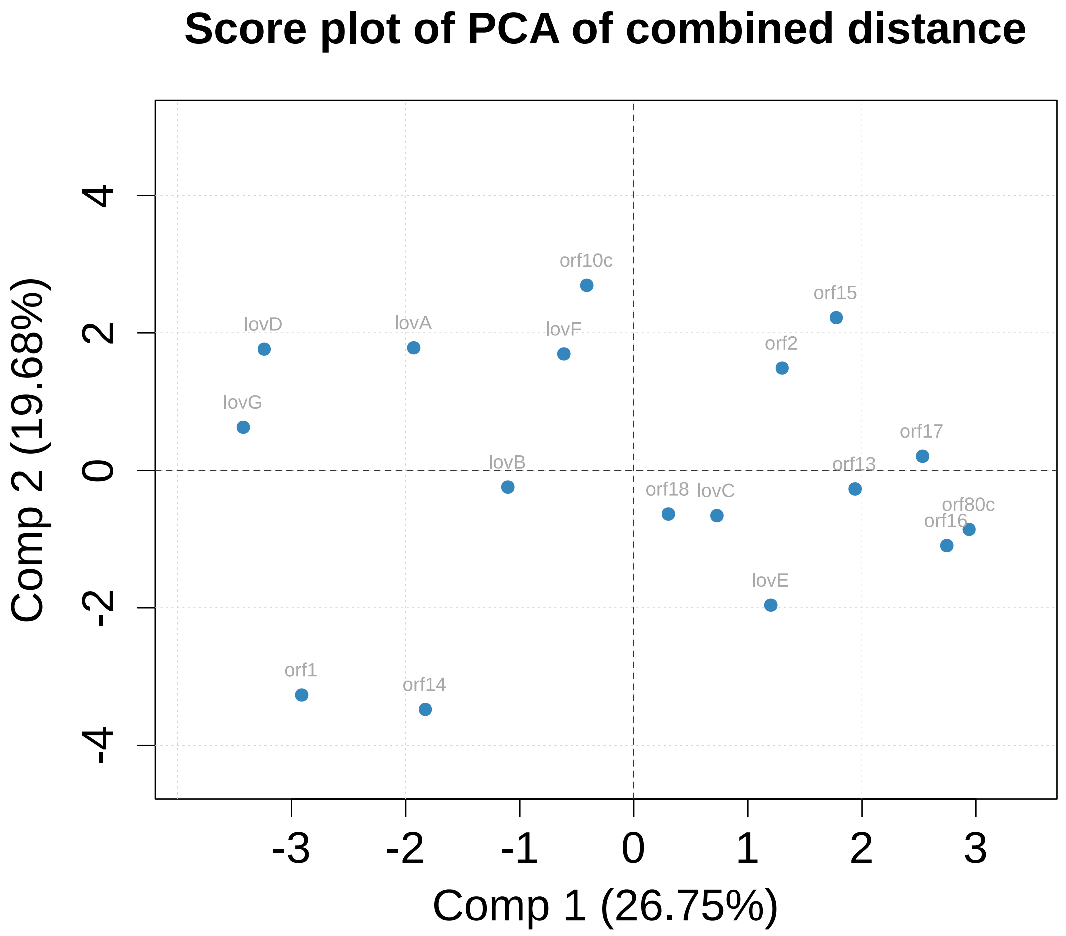
strict_distance_d_scaled
hclust (*, "ward.D2")


Ward's minimum variance - combined distance scaled

combined_distance_d_scaled
hclust (*, "ward.D2")

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
Bibliothek Your knowledge hub
TU WIEN





Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

S1 Table

The first table of this file (31 rows and 28 columns) can be found online with following link: https://doi.org/10.1371/journal.pcbi.1009372.s001

	TP	**FN**	**TN**	**FP**
1) essential genes	153	88	189	44
2) biosynthetic genes	129	66	189	44

	essential genes	**biosynthetic genes**
Sensitivity	0,6349	0,6615
Specificity	0,8112	0,8112
Precision	0,7766	0,7457
Negative Predictive Value	0,6823	0,7412
False Positive Rate	0,1888	0,1888
False Discovery Rate	0,2234	0,2543
False Negative Rate	0,3651	0,3385
Accuracy	0,7215	0,743
F1 Score	0,6986	0,7011
Matthews Correlation Coefficient	0,4524	0,4797
Normalized Matthews Correlation Coefficient	0,7262	0,73985
No-information error rate ni	0,5084	0,5444

S2 Table

Pathway	**Species**	**ICQ**
AAA_lysin_biosynthesis	Vanderwaltozyma_polyspora	0,46429
citrate_cycle	Lachancea_thermotolerans	0,77273
ergocalciferol_biosynthesis	Podospora_anserina	0,65934
Glycolysis	Kluyveromyces_marxianus	0,57778
Histidin_biosynthesis	Verticillium_alfalfae	0,54762
IMP_biosynthesis	Naumovozyma_castellii	0,6
Neurosporaxanthin_biosynthesis	Baudoinia_panamericana	0,71795
Pentosephosphate_pathway	Ashbya_gossypii	0,58929
Pyrimidine_biosynthesis	Marssonina_brunnea	0,67582
terpenoid_backbone_biosynthesis	Paracoccidioides_brasiliensis	0,2

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

S3 Table

Species	**ICQ**
Alectoria_fallacina	0,652778
Aplosporella_prunicola	0,321429
Ascodesmis_nigricans	0,589286
Ascoidea_rubescens	0,6
Aulographum_hederae	0,666667
Candida_albicans_L26	0,619048
Cephellophora_europea	0,488889
Cryomyces_minteri	0,666667
Dactylellina_haptotyla	0,55
Dactyrella_cylindrospora	0,866667
Drechslerella_brochopaga	0,714286
Eremomyces_bilateralis	0,714286
Heterodermia_speciosa	0,5
Kalaharituber_pfeilii	NA
Lasiodiplodia_theobromae	0,652778
Neofusicoccum_parvum	0,655556
Neolecta_irregularis	0,644444
Orbilia_oligospora	0,866667
Phialophora_americana	0,476191
Piedraia_hortae	0,75
Polychaeton_citri	0,688889
Pseudovirgaria_hyperparasitica	0,666667
Pyronema_omphalodes	0,642857
Rhinocladiella_mackenziei	0,642857
Rhizodiscina_lignyota	0,777778
Saccharata_proteae	0,666667
Saitoella_complicata	0,738095
Taphrina_deformans	0,785714
Tirmania_nivea	0,857143
Yamadazyma_tenuis	0,8

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Table. Definition of topology.

Topology	Definition
same	min. 8 similar species, same topology with only little exceptions, colour 70-100%
very similar	min.5 similar species, similar topology, colour min. 70%
similar	distance < 2, colour min. 50%
somewhat similar	either 1 or 2 similar species with distances < 0.5 and nodes <3, or more species but only little similarities
different	no similarities or only 1 similar species

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

S5 Table. Parameters used to calculate the manual evaluation measure (MEM).

ΔBranch lenghth	ΔNodes	Color	Topology	MEM
0 – 0.5	0	blue	same	3
0.5 - 1	1	-	very similar	2.5
1 – 1.5	2	green	Similar	2
1.5 - 2	3	-	somewhat similar	1.5
> 2	> 4	yellow	different	1

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

S6 Table

This is a very large file and can be found online following this link:

https://doi.org/10.1371/journal.pcbi.1009372.s006

S7 Table

This is a very large file and can be found online following this link:

https://doi.org/10.1371/journal.pcbi.1009372.s007

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

S8 Table

Random BGC	ICQ
1	0,8
2	0,66
3	0,64
4	0,83
5	0,66
6	1
7	0,66
8	1
9	0,66
10	1
11	0,81
12	0,75
13	0,6
14	0,66
15	0,66
16	1
17	0,83
18	0,81
19	0,73
20	0,7
21	1
22	0,95
23	0,92
24	0,62
25	0,68
26	0,86
27	0,65
28	0,77
29	0,74
30	0,5
31	0,61
32	0,83
33	0,8
34	0,75
35	0,75
36	0,75
37	0,83
38	0,66
39	0,9
40	0,76
41	0,55
42	0,83
43	0,58
44	0,76
45	0,76
46	0,59
47	0,75
48	0,85
49	0,5

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

S8 Table

50	0,83
51	0,76
52	0,83
53	0,7
54	0,8
55	0,65
56	0,81
57	0,73
58	0,75
59	0,7
60	1

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

FunOrder 2.0 – a fully automated method for the identification of co-evolved genes

Gabriel A. Vignolle[1], Robert L. Mach[1], Astrid R. Mach-Aigner[1], Christian Derntl[1,*]

[1] Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Vienna, 1060, Austria

* Corresponding author: christian.derntl@tuwien.ac.at

ABSTRACT

Coevolution is an important biological process that shapes interacting species or even proteins – may it be physically interacting proteins or consecutive enzymes in a metabolic pathway. The detection of co-evolved proteins may contribute to a better understanding of biological systems. Previously, we developed a semi-automated method, termed FunOrder, for the detection of co-evolved genes from an input gene or protein set. We demonstrated the usability and applicability of FunOrder by identifying essential genes in a biosynthetic gene cluster from different ascomycetes. A major drawback of this original method was the need for a manual assessment, which may create a user bias and prevents a high-throughput application. Here we present a fully automated version of this method termed FunOrder 2. To fully automatize the method, we used several mathematical indices to determine the optimal number of clusters in the FunOrder output, and a subsequent k-means clustering based on the first three principal components of a principal component analysis of the FunOrder output. Further, we replaced the BLAST with the DIAMOND tool, which enhanced speed and allows the future integration of larger proteome databases. The introduced changes slightly decreased the sensitivity of this method, which is outweighed by enhanced overall speed and specificity. Additionally, the changes lay the foundation for future high-throughput applications of FunOrder 2 in different phyla to help answer different biological questions.

AUTHOR SUMMARY

Coevolution is a process, which arises between different species or even different proteins that interact with each other. Any change occurring in one partner must be met by a corresponding change in the other partner to maintain the interaction throughout evolution. These interactions may occur in symbiotic relationships or between rivaling species. Within an organism, consecutive enzymes of metabolic pathways are also subjected to coevolution. We developed a fully automated method, FunOrder 2, for the detection of co-evolved proteins, which may contribute to a better understanding of protein interactions within an organism. We demonstrate that this method can be used to identify essential genes of the secondary metabolism of fungi, but FunOrder 2 may also be used to detect pathogenicity factors or remains of horizontal gene transfer next to many other biological systems that were shaped by coevolution.

INTRODUCTION

Every form of life known to humankind is subjected to evolution. This process shapes and forms all biological systems on macroscopic and molecular level. Thus, understanding and detecting evolutionary processes substantially contributes to understanding life forms and life itself. An important evolutionary process is the so-called coevolution. This is defined as a "process of reciprocal evolutionary change that occurs between pairs of species or among groups of species as they interact with one another" (2). This definition can be extended to interacting proteins (3), may it be physical interactions or

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

may it be consecutive actions in a metabolic pathway. In this regard, coevolution describes a similar evolutionary process with a similar evolutionary history among interacting proteins and the corresponding genes.

In a previous study, we described a semi-automated method for the identification of coevolutionary linked genes, named FunOrder (1). Therein, the protein sequences of an input set of proteins are blasted against an empirically optimized proteome database. The top 20 results of each search are then compared in a multisequence alignment and a phylogenetic tree is calculated for each input protein. Next, the phylogenetic trees of all proteins are compared pairwise using the treeKO tool. This tool calculates how similar two trees are, and in thus how similar the evolutionary history of two proteins is. The treeKO tool calculates two distances, the strict distance and the speciation distance. Notably, the strict distance had previously been suggested to be more suitable for the detection of coevolution in protein families than the speciation (or evolutionary) distance (4). However, we combined the two distance values to a third measure, the combined distance, in order to consider also the speciation history in the FunOrder method. The strict and the combined distances of all pairwise comparisons were then compiled in two matrices and visualized as heatmaps, dendrograms and two principal component analyses (PCA) were performed. In the final step of this method, the user needed to assess these different visualizations of the underlying data to detect co-evolved proteins (Fig 1A). Please also refer to the original study for a detailed description of this method (1).

Previously, we demonstrated the functionality and applicability of this method by identifying essential genes in biosynthetic gene clusters (BGCs) of ascomycetes (1). Fungal BGCs contain genes whose corresponding enzymes catalyze the biosynthesis of secondary metabolites (SMs) (5). SMs are a vast group of compounds with different structures and properties that are not necessary for the normal growth of an organism but can be beneficial under certain conditions (6). Notably, many SMs also have medicinal or other useful purposes, such as dyes, food additives, and as monomers for novel plastics (7). However, we can classify the genes in a BGC into biosynthetic genes, further essential genes, and gap genes. The biosynthetic genes encode for enzymes that are directly involved in the biosynthesis of the SM, while the further essential genes encode for transporters (8), transcription factors (9), or resistance genes (10). In contrast, gap genes are not involved in the biosynthesis of the SM despite being co-localized in the BGC (11). Both, the biosynthetic genes and the further essential genes are necessary for the biosynthesis of a SM in the native organisms (12). We could use FunOrder to detect theses essential genes, because they share a similar evolutionary background in many fungal BGCs (1). The FunOrder method contributes to a better understanding of fungal BGCs by adding an additional layer of information. This may support users in the decision which genes should be considered for detailed studies in the laboratory. Importantly, the application of FunOrder is not limited to BGCs from ascomycetes but may be useful to answer any biological problem that contains molecular coevolution of genes or proteins. Notably, this requires the compilation and evaluation of a suitable proteome database.

The obvious major shortcoming of the original FunOrder method is the final manual assessment which prevents full automation and high-throughput analyses. Further, the very sensitive but slow BLAST algorithm (13) limits the size of the proteome database used in this method. If this method shall be used for the analysis of coevolution in plants or mammals, larger databases will be needed. In this study we describe an improved version of the method, termed FunOrder 2 which solves the two mentioned limitations. For an automated detection of co-evolving genes, we determine the optimal number of gene groups in the FunOrder output and then use k-means clustering based on the first three principal components of a PCA. Further, we replace BLAST with the recently published and upgraded DIAMOND tool (14) to enable searching of larger databases and lay the foundation for future different applications of FunOrder 2.

RESULTS

Integration of the DIAMOND algorithm

The first major improvement of the FunOrder method was the integration of the DIAMOND algorithm (14, 15) for searching the proteome database instead of the previously used BLAST algorithm (13) (Fig 1). This change will allow the usage of larger databases in FunOrder 2, since DIAMOND is as sensitive as BLAST, but is faster and is adapted to larger databases (14). With DIAMOND the run time of the first step in the FunOrder pipeline was reduced significantly. For instance, the database search for the lovastatin BGC of *Aspergillus terreus* (lov) (16) took 1 m 25 sec using the original FunOrder method, and 45 sec real time using FunOrder 2. This difference will of course be more pronounced and obvious when a larger database is used.

To test, whether the integration of DIAMOND might have altered the ability of FunOrder to detect coevolution, we analysed the same control gene clusters (GCs) we had

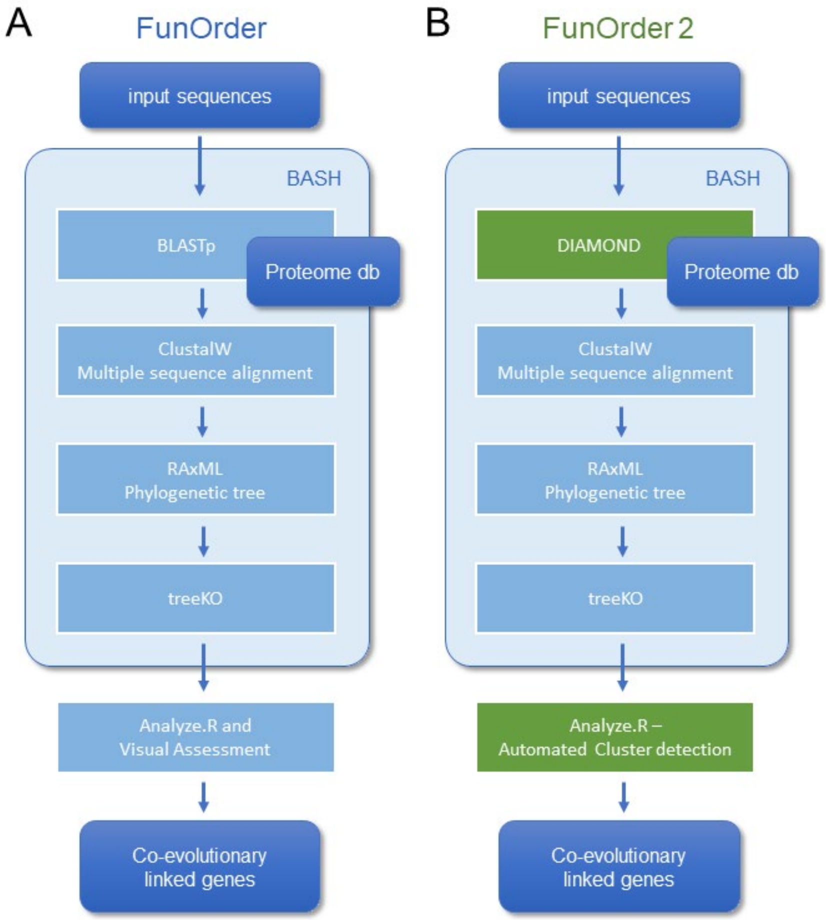


Fig 1. Comparison of the workflow of the original FunOrder method (A) and FunOrder 2 (B).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

TU WIEN Bibliothek Your knowledge hub

previously used to evaluate the original FunOrder method (1) and calculated the internal coevolution quotient (ICQ). The ICQ expresses how many genes in a gene cluster are detected as coevolutionary linked and is calculated subsequently to the treeKO comparison (Fig. 1). Since no other changes have been introduced until this point in the workflow, the ICQ values are a feasible way to compare BLAST and the DIAMOND software. We found only marginal differences between the original FunOrder method (using BLAST) and FunOrder 2 (using DIAMOND) (Table S1). For visualization, we compared the ICQ results in a kernel density plot (Fig 2). Therein, the curve for the ICQs of the positive control GCs (BioPath in Fig 2) slightly shifted to the left (higher internal coevolution) compared to the original method, while the curve for the negative control GCs (random GCs in Fig 2) slightly shifted to the right (lower internal coevolution). These results indicate that DIAMOND might be better suited than BLAST within the FunOrder method, as the usage of DIAMOND resulted in a better distinction of the positive and negative control GCs. The curve for the sequential GCs was flattened and broadened compared to the original curve (Fig 2), which can also be explained by the assumed better

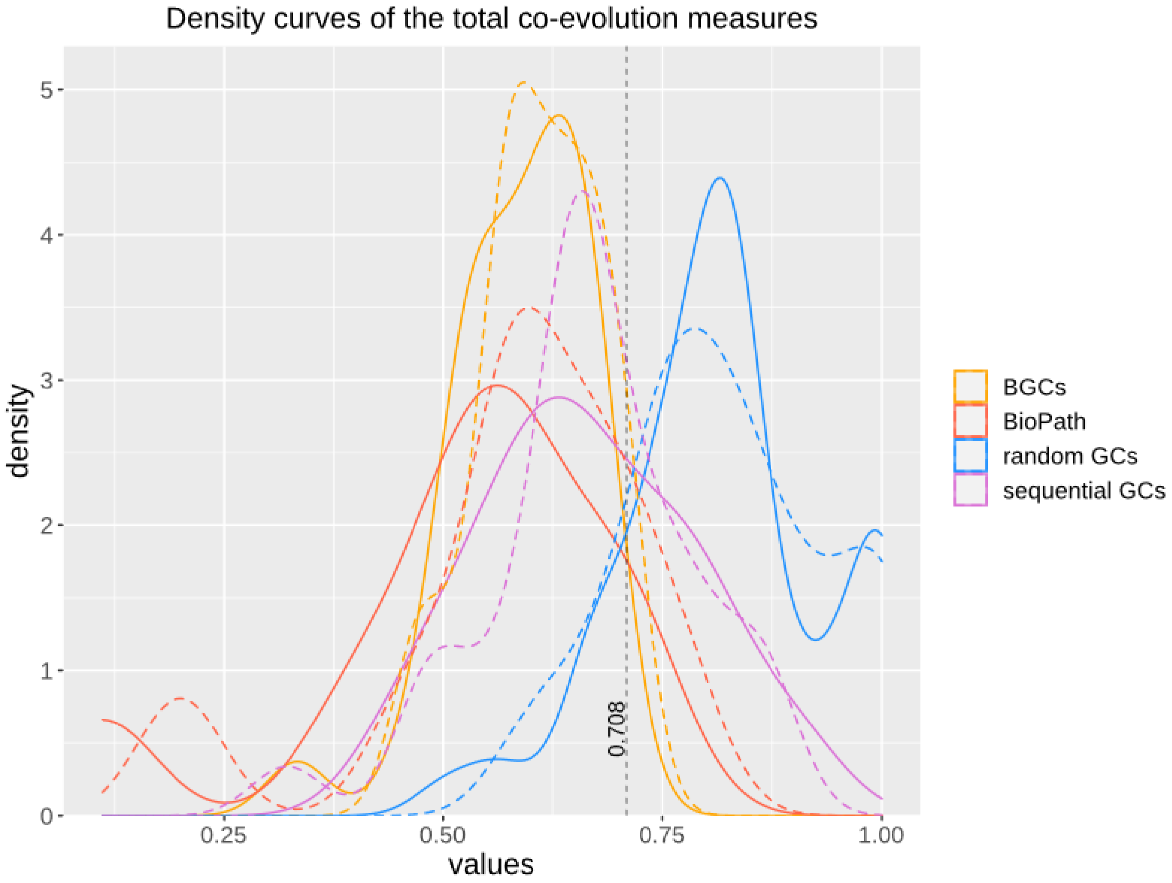


Fig 2. Kernel density plot of the ICQ values for co-evolutionary linked enzymes of different control sets comparing the original FunOrder method (dashed lines) and FunOrder 2 (solid lines). BGCs, previously empirically characterized fungal BGCs; BioPath, protein sets of conserved biosynthetic pathways of the primary metabolism; random GCs, randomly assembled protein sets from 134 fungal proteomes; sequential GCs, co-localized genes from random loci of different ascomycetes.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

performance of DIAMOND in this workflow. As the sequential GCs are random loci from different ascomycetes (1), they contain random numbers of co-evolved and independently evolved genes. Consequently, the usage of DIAMOND lowers the ICQ for GCs containing many co-evolved genes and raises the ICQ for GCs with many independently evolved genes compared to the original FunOrder method. This results in the detection of simultaneously more and less coevolution in all sequential GCs and therefore a flattening of the curve in Fig 2. For the benchmark BGCs, we could not observe a drastic change of the height or position of the curve, but a change of the shape with no significant differences of the variance and the mean (File S1). However, the changes of the curves of the random GCs and the BGCs, resulted in a new point of intersection (0.708), which should be considered in the final assessment of fungal BGCs, as described in our previous publication (1).

Automated Cluster definition

As mentioned, a major limitation of the original FunOrder method was the need for a manual assessment of the output, during which the proteins are grouped into clusters based on different data visualizations (Fig 1A). Please refer to our previous method for a detailed description of the procedure (1). To solve this problem, we integrated two R scripts for automatic definition of co-evolved protein groups (or clusters) (Fig 1B). The two R scripts use the first three principal components of the PCA of the strict and the combined distance matrices as input (Fig 1B) and group the proteins by k-means clustering. In the original FunOrder method, only the first two components were considered.

The first R-script for automated protein (or gene) clustering initially determines the optimal number of gene clusters within the first three principal components of the PCAs using the R Package NbClust (17). This package uses different indices and varies the number of clusters, distance measures and clustering methods to determine the optimal number of clusters in a data set based on the majority rule. If the prediction of the optimal number of clusters fails, the second (a back-up) script with a predefined number of clusters is called. The prediction of the optimal number of clusters might fail for instance if the majority rule cannot be applied. As we aim to distinguish biosynthetic, further essential, and gap genes in fungal BGCS, we predefined the number of clusters to 3. Regardless of the script used, the final output is an excel file (Table S2) and a color-coded visualization of the PCA (File S2).

To test how this automated cluster definition compares to the previously performed manual cluster definition, we analyzed the same 30 BGCs as in our previous study. To observe only the influence of the automated cluster definition, we kept the BLAST tool for the initial database search still in place (Fig. 1). Then, we

Table 1. Number of BGCs in which the automated cluster detection (in combination with BLAST or DIAMOND) delivered the same, better, worse, or different results for the given gene categories compared to the manual method.

BLAST / DIAMOND	same	better	worse	different genes
biosynthetic genes	16 / 17	4 / 4	10 / 9	- / -
further essential genes	11 / 11	3 / 3	5 / 5	- / -
gap genes	13 / 13	8 / 8	3 / 3	3 / 3
extra genes	16 / 16	5 / 5	2 / 3	1 / -

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

compared the obtained results to those of the previously performed manual analyses (1) (Table 1 and Table S3). In only 5 out of the tested 30 BGCs, the exact same results were obtained (Table S3). In 15 BGCs, the automated cluster definition missed at least one biosynthetic or further essential gene in comparison to the manual assessment, but it could detect more of these essential genes in 5 BGCs. Regarding the gap and extra genes, the automated cluster definition returned less false positives than the manual assessment in 12 BGCs but found more in 4 BGCs. In summary, the automated cluster detection appeared to be more stringent than the manual assessment method, which led to slightly reduced sensitivity but enhanced selectivity (see Table S3 for a detailed statistical analysis).

Next, we tested the simultaneous influence of DIAMOND and the automated clustering on the overall performance of FunOrder 2 during the analysis of fungal BGCs. To this end, we performed the same comparative analysis of the benchmark BGCs as described above. The results were very similar to the automated analysis using the BLAST analysis (Table 1 and Table S3). In a few cases, the usage of DIAMOND improved the automated cluster definition compared to BLAST, but it remained still more stringent than the manual assessment (Table 1 and Table S3). Fewer biosynthetic genes or further essential genes were detected in 13 of the 30 BGCs by FunOrder 2, but also fewer gap or extra genes in 12 BGCs (Table 1 and Table S3). Yet, FunOrder 2 clustered more genes together than the original method in other BGCs - to be precise, more essential genes were detected in in 5 BGCs and more gap or extra genes in 4 BGCs compared to the original method (Table S3). The overall enhanced stringency reduced the sensitivity slightly (Table 2) but also improved several statistic measures, including specificity, precision, and the normalized Matthew correlation coefficient (Table 2, in bold). To test if the observed differences have a

Table 2. Performance comparison of the original FunOrder (1) and FunOrder 2 for detecting relevant genes in fungal BGCs. Improved statistical measures are highlighted in bold.

	FunOrder essential genes	FunOrder 2 essential genes	FunOrder biosynthetic genes	FunOrder 2 biosynthetic genes
Sensitivity	0.6349	0.6266	0.6615	0.6564
Specificity	0.8112	**0.8541**	0.8112	**0.8541**
Precision	0.7766	**0.8162**	0.7457	**0.7901**
Negative Predictive Value	0.6823	**0.6886**	0.7412	**0.7481**
False Positive Rate	0.1888	**0.1459**	0.1888	**0.1459**
False Discovery Rate	0.2234	**0.1838**	0.2543	**0.2099**
False Negative Rate	0.3651	0.3734	0.3385	0.3436
Accuracy	0.7215	**0.7384**	0.7430	**0.764**
F1 Score	0.6986	**0.7089**	0.7011	**0.7171**
Matthews Correlation Coefficient	0.4524	**0.4926**	0.4797	**0.5242**
Normalized Matthews Correlation Coefficient	0.7262	**0.7463**	0.73985	**0.7621**
No-information error rate ni	0.5084	0.5084	0.5444	0.5444

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

significant impact on the overall applicability of FunOrder 2 in fungal BGCs, we further compared the percentages of correctly identified genes in each BGC between the original FunOrder and FunOrder 2 (Tables S3) in an ANOVA (File S1) and found no significant difference. Taken together, we conclude that the introduced changes allow the detection of coevolution between different proteins with an enhanced stringency and precision compared to the original method, and that FunOrder 2 can be used to identify essential genes in fungal BGCs.

The automation of the cluster detection in FunOrder 2 prevents a user bias and improves the overall speed. The analysis of the lovastatin BGC of *Aspergillus terreus* (lov) (16) with 17 genes, took 1 h 19 m 48 sec real time using 22 threads on an Ubuntu Linux system with 128 GB DDR4 RAM with the original FunOrder (excluding manual cluster definition) and 1 h 19 m 58 sec real time with FunOrder2. Notably, the runtime for FunOrder 2 includes already the automated detection and grouping of co-evolving genes, which takes an experienced user additional 30 - 45 minutes during the original method.

MATERIALS AND METHODS

Changes of the workflow

Within the previously developed work flow (1) (Fig. 1A), we replaced BLAST (13) with DIAMOND (14) for the database search (Fig 1B). The previous and subsequent steps up to the integrated R-scripts remained the same as described in the original FunOrder method (1). Notably, the BLAST algorithm was kept in the software bundle to extract the sequences from the local database and can be used for an optional remote search of the NCBI database (18). The distances measure obtained after the treeKO algorithm were compiled in matrices, which were used as input for three alternative R-scripts (Fig 1B). All R-scripts combine the strict distance matrix and the evolutionary distance matrix to a third distance, the combined distance matrix. These matrices were used for the calculation of the ICQ and as basis for the determination of co-evolved genes. The three scripts differ only in how exactly the co-evolved genes are determined. The first R-script is a revised version of the R-script used in the original FunOrder method (1). It was simplified by removing unnecessary Euclidean distance calculations and the order of the called functions was rearranged in a manner, that all calculations were performed on a single matrix and then on the next. The first matrix to be analyzed is the strict distance matrix followed by the combined distance matrix. Further, we rearranged the order of the visualizations in the output, which is saved as “FunOrder_Supplementary_Rplots.pdf”. This output is similar to the original FunOrder method and can be assessed manually as described previously (1).

The second and third R-scripts aim to determine the co-evolved genes automatically. To this end, a PCA is calculated for the strict and the combined distance matrices each. The first three principal components are then considered for defining the clusters by a k-means clustering approach. The difference between the second and the third R-script is the number of clusters used in the k-means clustering approach. In the second R-script, the optimal number of clusters in the first three principal components of the PCAs is determined by NbClust (17) using 28 indices (Table 3). We limited the maximum number of possibly definable clusters to 5 and chose the Ward´s minimum variance method based on the Euclidean distance for optimal cluster search within the NbClust function (19). The third R-script performs a k-means clustering with 3 clusters; it is only called as a back-up if the prediction of an optimal number of clusters in the second script fails. In both cases, the determined clusters are visualized in a color-coded plot of the first two principal components of the PCAs under “FunOrder_clustering_Rplots_pred.pdf” or “FunOrder_clustering_Rplots_defined.pdf” and as table under “cluster_definition_pred.xlsx” or “cluster_definition_3.xlsx”.

Table 3. Indices used to determine the optimal number of clusters.

Index	Reference
Bale index	(20)
Ball index	(21)
CCC index	(22)
CH index	(23)
C-index	(24)
DB index	(25)
Duda index	(26)
Dunn index	(27)
Frey index	(28)
Friedman index	(29)
Gamma index	(30)
Gap index	(31)
Gplus index	(32, 33)
Hartigan index	(34)
KL index	(35)
Marriot index	(36)
McClain index	(37)
Pseudot2 index	(26)
Ptbiserial index	(32, 38)
Ratkowsky index	(39)
Rubin index	(29)
Scott index	(40)
SD index	(41)
SDbw index	(42)
Silhouette index	(43)
Tau index	(32, 33)
Tracew index	(44)
Trcovw index	(44)

The software bundle is written in the BASH (Bourn Again Shell) environment and is deposited in the GitHub repository https://github.com/gvignolle/FunOrder (doi:10.5281/zenodo.5118984). Details on all included scripts can be found in the ReadMe file on the GitHub repository. FunOrder 2 requires some dependencies, for details and links to all dependencies please refer to the ReadMe file.

Control gene clusters

For the evaluation of FunOrder 2, we used the same control gene clusters (GC) as in the original study (1). As benchmark BGCs, we used 30 previously empirically defined BGCs. As negative controls, we used randomly assembled GCs. As positive control, we used enzymes of conserved metabolic pathways of the primary metabolism. The sequences of all test and control sets are deposited in the GitHub repository https://github.com/gvignolle/FunOrder.

Calculation of the internal coevolution quotient (ICQ)

“The internal coevolutionary quotient (ICQ) expresses how many genes in a GC or proteins in a protein set are co-evolved according to the previously defined threshold for strict and combined distances within the distance matrices of an analysed GC (or protein set).”(1) In accordance with the original method the ICQ values were calculated using Equation 1(1).

$$ICQ = 1 - \left\{\frac{g}{2 * [d * (d-1)]}\right\}$$

Equation 1. ICQ = internal coevolutionary quotient; g = number of strict distances < 0.7 and combined distances <= (0.6 * max value of the combined distance matrix) in all matrices; d = number of genes in the GC (1).

Performance evaluation

Similar as in the original method we analyzed 30 empirically characterized BGCs to evaluate the ability of FunOrder 2 to identify presumably co-evolved essential genes (as defined in Table S3) and to distinguish them from so-called gap genes and genes outside of the defined BGC borders. The genes clustering with the core enzyme(s) were considered as

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

"detected". As previously described "we counted the total number of (1a) detected essential genes or (1b) detected biosynthetic genes, (2a) not detected essential genes or (2b) not detected biosynthetic genes, (3) detected gap and extra genes, and (4) not detected gap or extra genes in all BGCs, and defined (1a or 1b) as true positives (TP), (2a or 2b) as false negatives (FN), (3) as false positives (FP), and (4) as true negatives (TN)" (1), which were finally used as input for a stringent statistical analysis (1).

DISCUSSION

The integration of the DIAMOND tool and the automated detection of co-evolved genes improved the run time and total analysis time, allowing high throughput analysis of protein sets and GCs without the risk of a user-bias. In general, the automated cluster definition appears more stringent than the manual assessment, which resulted in improved specificity and precision and a slightly reduced sensitivity during the analysis of fungal BGCs by FunOrder2 compared to the original method. In summary, we consider the integration of a fully automated cluster definition a major improvement, as the advantages (speed, reproducibility, precision) clearly outweigh the slightly reduced sensitivity.

As demonstrated, FunOrder 2 can be used to determine the essential genes in fungal BGCs, but this is not the only potential application of FunOrder 2. As protein coevolution can be used to predict protein-protein interactions and biosynthetically linked enzymes (1, 45), FunOrder 2 may be used to answer many different research questions. Given enough computational time, even complete fungal genomes might be assessed by our method. It is also exciting to speculate if and how FunOrder may be used in other clades of life. A limitation in this regard might be the maximum number of predicted clusters. NbClust limits the number of potential clusters to 15, we further lowered this number to 5 for the analysis of BGCs. This problem might be circumvented by an arbitrary definition of the number of clusters or by consecutive FunOrder analyses, in which a large output cluster is used as input for a new analysis. FunOrder 2 is provided with a database of ascomycete proteomes and can therefore be used for the detection of coevolution of proteins in this fungal division. If other divisions, classes, or even kingdoms shall be analyzed, a suitable new proteome database must be compiled and tested. As mentioned, the integration of the DIAMOND tool enables the integration of larger databases. However, at least 25 different proteomes must be used, because the phylogenetic trees are calculated with a maximum of 20 homologous sequences. Naturally, the proteomes should be of high quality (best RNASeq derived). The proteomes shall be equally distributed among the taxonomic rank to be analyzed but also take the size of the different subordinate ranks into consideration. Put differently, if a division contains 4 small classes, and two large classes, the database should contain proteomes of all six classes, but more from the larger classes than from the smaller classes. The database shall be representative sample of the phylogenetic group to be analyzed. This also means that highly divers phylogenetic groups need to be over-represented in comparison to evolutionary uninventive clades. Further, evolutionary outliers and special clades shall be considered in the database design. For instance, if a phylum contains a family that is the only member of its class, the user needs to decide whether that family shall be part of the proteome database at all, depending on the size and importance of the family. If the family shall be considered, several proteomes need to be included in the database, otherwise the evolutionary distances of the tested proteins might be too large to be successfully evaluated by FunOrder. Any new database must be tested thoroughly according to the procedure we described previously (1). This means, that suitable test gene clusters must be compiled and

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

that meaningful thresholds for the strict and combined distance should be defined. If possible, a test set of target gene clusters should be analyzed and compared to previous results. Please refer to our previous study on how we tested the ascomycete database, determined the thresholds, and tested the applicability of FunOrder for the detection of essential genes in BGCs in ascomycetes (1). A possible short-cut in this procedure might determining the thresholds of strict and combined distance via threshold optimizing (best obtained distinction of positive and negative control gene clusters). Please also refer to the technical guidelines for construction and integration of the database at the GitHub repository https://github.com/gvignolle/FunOrder.

ACKNOWLEDGEMENT

not applicable

REFERENCES

1. Vignolle GA, Schaffer D, Zehetner L, Mach RL, Mach-Aigner AR, Derntl C. FunOrder: A robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution. PLoS Comput Biol. 2021;17(9):e1009372.
2. Rafferty JP, Thompson JN. "coevolution". Encyclopedia Britannica. Accessed 12 December 2021(https://www.britannica.com/science/coevolution).
3. Fraser HB, Hirsh AE, Wall DP, Eisen MB. Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci U S A. 2004;101(24):9033-8.
4. Marcet-Houben M, Gabaldon T. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. Nucleic Acids Res. 2011;39(10):e66.
5. Osbourn A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends in Genetics. 2010;26(10):449-57.
6. Keller NP, Turner G, Bennett JW. Fungal secondary metabolism — from biochemistry to genomics. Nature Reviews Microbiology. 2005;3(12):937-47.
7. Alberti F, Foster GD, Bailey AM. Natural products from filamentous fungi and production by heterologous expression. Appl Microbiol Biotechnol. 2017;101(2):493-500.
8. Wang DN, Toyotome T, Muraosa Y, Watanabe A, Wuren T, Bunsupa S, et al. GliA in Aspergillus fumigatus is required for its tolerance to gliotoxin and affects the amount of extracellular and intracellular gliotoxin. Medical mycology. 2014;52(5):506-18.
9. Derntl C, Rassinger A, Srebotnik E, Mach RL, Mach-Aigner AR. Identification of the Main Regulator Responsible for Synthesis of the Typical Yellow Pigment Produced by *Trichoderma reesei*. Appl Environ Microbiol. 2016;82(20):6247-57.
10. Schrettl M, Carberry S, Kavanagh K, Haas H, Jones GW, O'Brien J, et al. Self-protection against gliotoxin--a component of the gliotoxin biosynthetic cluster, GliT, completely protects Aspergillus fumigatus against exogenous gliotoxin. PLoS Pathog. 2010;6(6):e1000952.
11. Tai Y, Liu C, Yu S, Yang H, Sun J, Guo C, et al. Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (Camellia sinensis). BMC Genomics. 2018;19(1):616.
12. Anyaogu DC, Mortensen UH. Heterologous production of fungal secondary metabolites in Aspergilli. Frontiers in Microbiology. 2015;6(77).
13. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421-.
14. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

using DIAMOND. Nature Methods. 2021;18(4):366-8.

15. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60.

16. Mulder KC, Mulinari F, Franco OL, Soares MS, Magalhaes BS, Parachin NS. Lovastatin production: From molecular basis to industrial process optimization. Biotechnol Adv. 2015;33(6 Pt 1):648-65.

17. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. 2014. 2014;61(6):36.

18. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.

19. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification. 2014;31(3):274-95.

20. Beale EML. Euclidean Cluster Analysis: Scientific Control Systems Ltd; 1969.

21. Ball G, Hall DJ. ISODATA: A Novel Method of Data Analysis and Pattern Classification. Stanford Research Institute, Menlo Park. 1965.

22. Sarle WS, Institute S. Cubic Clustering Criterion: SAS Institute; 1983.

23. Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics. 1974;3(1):1-27.

24. Hubert LJ, Levin JR. A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin. 1976;83(6):1072–80.

25. Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979;PAMI-1(2):224-7.

26. Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley; 1973.

27. Dunn JC. Well-Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybernetics. 1974;4(1):95-104.

28. Frey T, van Groenewoud H. A Cluster Analysis of the D-squared Matrix of White Spruce Stands in Saskatchewan Based on the Maximum-Minimum Principle. Journal of Ecology. 1972;60(3):873-86.

29. Friedman HP, Rubin J. On Some Invariant Criteria for Grouping Data. Journal of the American Statistical Association. 1967;62(320):1159-78.

30. Baker FB, Hubert LJ. Measuring the Power of Hierarchical Cluster Analysis. Journal of the American Statistical Association. 1975;70(349):31-8.

31. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(2):411-23.

32. Milligan GW. A monte carlo study of thirty internal criterion measures for cluster analysis. Psychometrika. 1981;46(2):187-99.

33. Rohlf FJ. Methods of Comparing Classifications. Annual Review of Ecology and Systematics. 1974;5(1):101-13.

34. Hartigan JA. Clustering Algorithms. John Wiley & Sons, New York. 1975.

35. Krzanowski WJ, Lai YT. A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. Biometrics. 1988;44(1):23-34.

36. Marriott FHC. Practical Problems in a Method of Cluster Analysis. Biometrics. 1971;27(3):501-14.

37. McClain JO, Rao VR. CLUSTISZ: A Program to Test for the Quality of Clustering of a Set of Objects. Journal of Marketing Research. 1975;12(4):456-60.

38. Milligan GW. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika. 1980;45(3):325-42.

39. Ratkowsky DA, Lance GN. Criterion for determining the number of groups in a classification. Australian Computer Journal. 1978(11): 115-7.
40. Scott AJ, Symons MJ. Clustering Methods Based on Likelihood Ratio Criteria. Biometrics. 1971;27(2):387-97.
41. Halkidi M, Vazirgiannis M, Batistakis Y, editors. Quality Scheme Assessment in the Clustering Process. Principles of Data Mining and Knowledge Discovery; 2000 2000//; Berlin, Heidelberg: Springer Berlin Heidelberg.
42. Halkidi M, Vazirgiannis M, editors. Clustering validity assessment: finding the optimal partitioning of a data set. Proceedings 2001 IEEE International Conference on Data Mining; 2001 29 Nov.-2 Dec. 2001.
43. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53-65.
44. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985;50(2):159-79.
45. Ochoa D, Pazos F. Practical aspects of protein co-evolution. Frontiers in Cell and Developmental Biology. 2014;2(14).

SUPPORTING INFORMATION

File S1. ANOVA for the percentage of correctly detected genes detected by FunOrder and FunOrder 2, respectively.

File S2. FunOrder 2 output of the Lovastatin BGC from *A. terreus* (lov).

Table S1. ICQ values of protein sets of conserved metabolic pathways of the primary metabolism (BioPath), sequential GCs and random GCs used in this study.

Table S2. FunOrder 2 output of the Lovastatin BGC from *A. terreus* (lov).

Table S3. Results for the analyses of benchmark BGCs using different versions of the FunOrder method.

DATA AVAILABILITY

The FunOrder tool, the relevant database, and the sequences and the FunOrder output of the negative control GCs and the positive control BGCs are available in the GitHub repository (https://github.com/gvignolle/FunOrder). We have also used Zenodo to assign a DOI to the repository: 10.5281/zenodo.5118984.

FUNDING

This study was supported by the Austrian Science Fund (FWF, https://www.fwf.ac.at/) [P 34036 to CD] and TU Wien (https://www.tuwien.at/) [PhD program TU Wien bioactive]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

GV: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation
RM: Resources, Writing – Review & Editing
AM: Resources, Writing – Review & Editing
CD: Conceptualization, Funding Acquisition, Methodology, Validation, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

S1 File

Statistical analysis of the relative discovery rates of essential or biosynthetic genes

All the statistical tests were performed in the R environment (1). The Shapiro-Wilk test used below was used to check for the normality of the percentages of detected essential or biosynthetic genes, as previously defined (2), between the FunOrder 1 and FunOrder 2 output (table 1). Normality assumptions underlie outlier detection hypothesis tests. If the p-value is above the set alpha significance value (0.01) then the null hypothesis is not discarded. In other words, it can be considered a normal distribution.

Table 1 Shapiro-Wilk normality tests.

data set	**p-value**
FunOrder 1 – % essential genes	0.2236
FunOrder 1 – % biosynthetic genes	0.007389
FunOrder 2 – % essential genes	0.1738
FunOrder 2 – % biosynthetic genes	0.05362

Table 2 Levene's Test for Homogeneity of Variance (center = median) performed on the percentages of detected essential or biosynthetic genes, as previously defined, between the FunOrder 1 and FunOrder 2 output.

	Df	**F value**	**Pr(>F)**
performance data sets	3	0.4007	0.7527

From the output in table 2, it can be seen that the p-value was not less than the significance level of 0.05. This means that there was no evidence to suggest that the variance is statistically significantly different for the data sets. Levene's test is an alternative to Bartlett's test when the data is not normally distributed.

Table 3 Computed one-way ANOVA test the analysis of variance performed on the percentages of detected essential or biosynthetic genes, as previously defined, between the FunOrder 1 and FunOrder 2 output.

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
performance data sets	3	1443	481.0	0.853	0.468
Residuals	116	65447	564.2		

The output in table 3 includes the columns F value and Pr(>F) corresponding to the p-value of the test. As the p-value is higher than the significance level 0.05, we could conclude that there are no significant differences between the percentages of relative discovery rate of essential or biosynthetic genes in the model summary between FunOrder 1 and FunOrder 2.

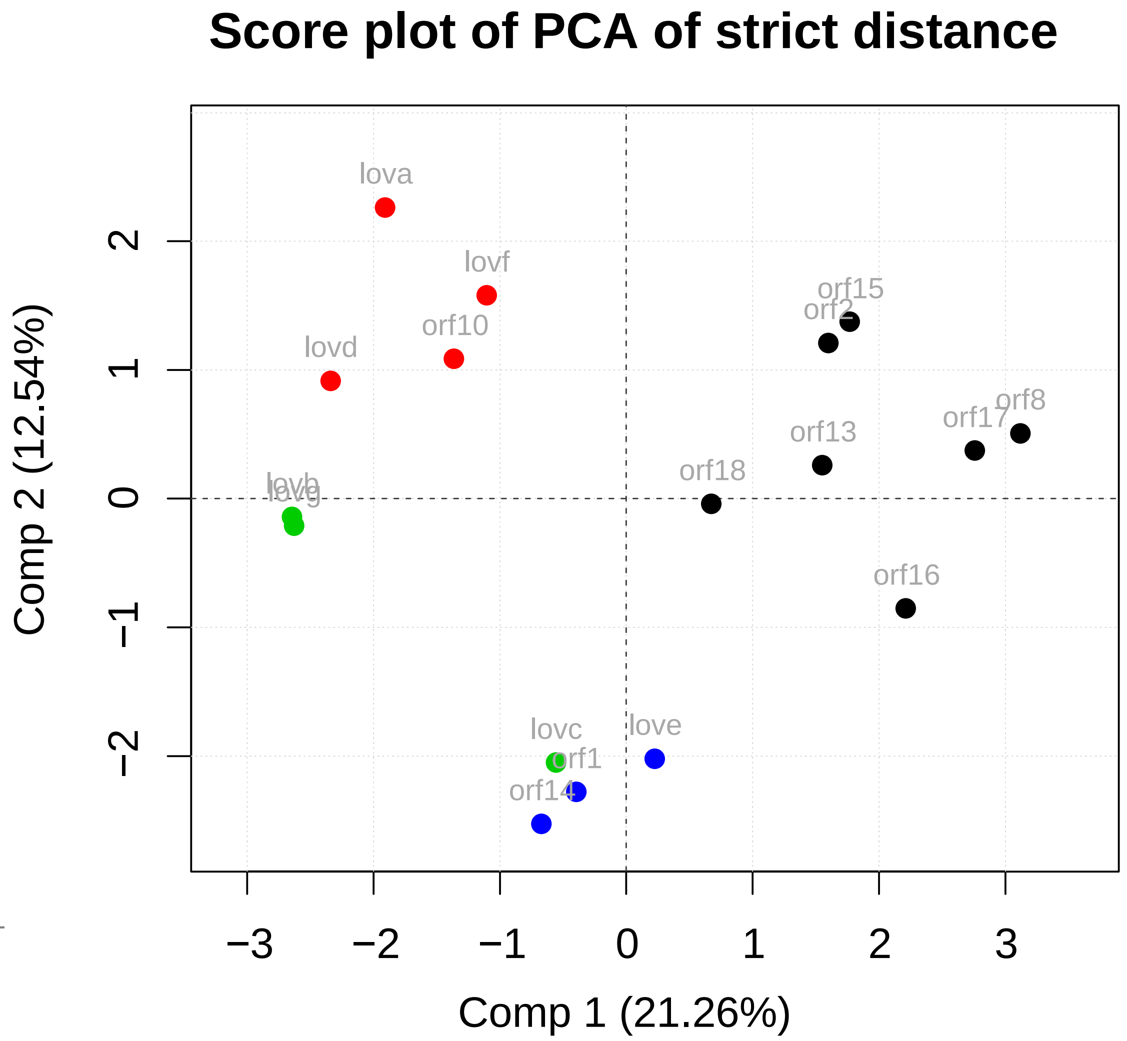
We further compared the internal co-evolution quotient (ICQ) of both the FunOrder 1 and FunOrder 2 output for the BGCs. A F-test resulted in F = 0.76644 with a p-value = 0.4783, since the p-value is above the significance level 0.05 we could conclude that there is no significant difference between the variances in the two sets of ICQs. We continued with a two sided two sample t-test to compare the means of the two datasets. The t-test had a p-value of 0.2682, since we obtained a p-value greater than 0.05 we can conclude that means of the two datasets have no significant difference and can be regarded as equal.

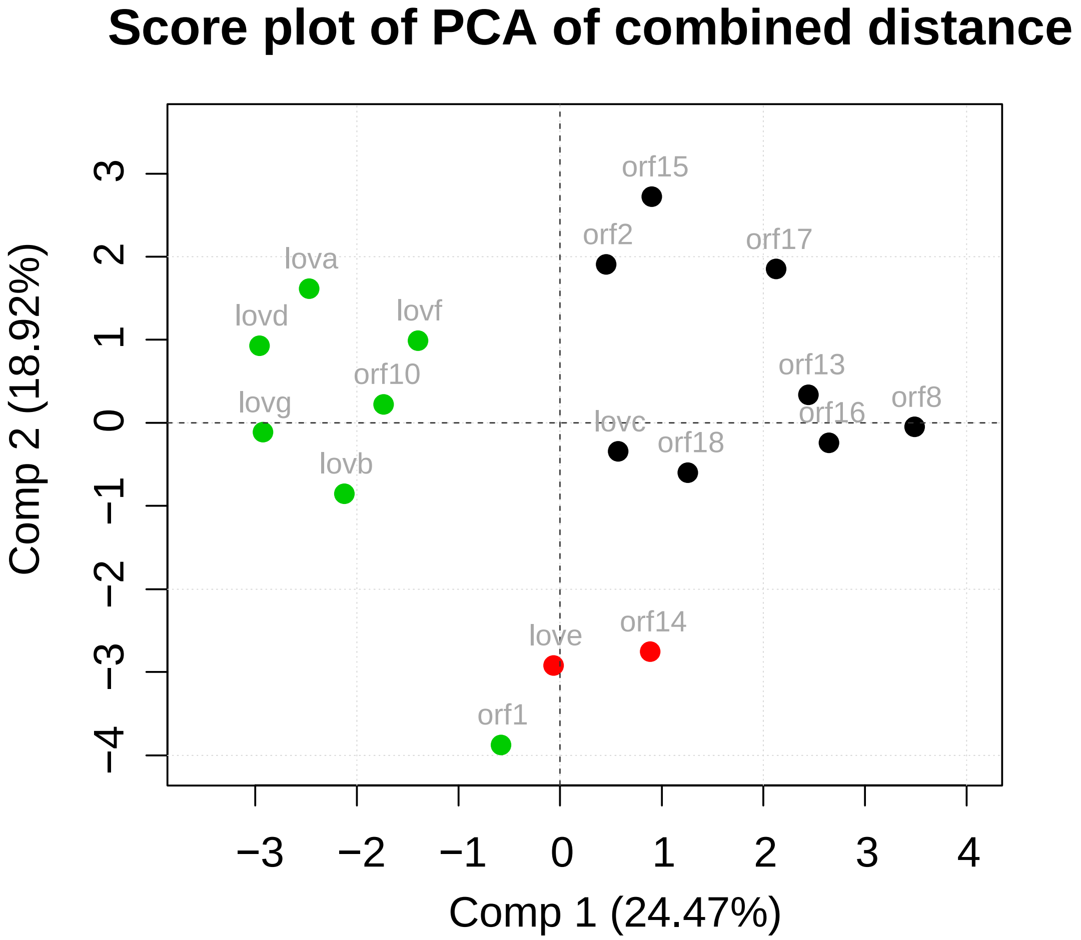
References:

1. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2019.
2. Vignolle GA, Schaffer D, Zehetner L, Mach RL, Mach-Aigner AR, Derntl C. FunOrder: A robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution. PLoS Comput Biol. 2021;17(9):e1009372.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

S2 File





Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

S1 Table

Biosynthetic pathways

Pathway	Species	FunOrder 2 - ICQ	org. FunOrder - ICQ
AAA_lysin_biosynthesis	Vanderwaltozyma_polyspora	0,3928571	0,4642857
citrate_cycle	Lachancea_thermotolerans	0,4727273	0,7727273
ergocalciferol_biosynthesis	Podospora_anserina	0,5934066	0,6593407
Glycolysis	Kluyveromyces_marxianus	0,5666667	0,5777778
Histidin_biosynthesis	Verticillium_alfalfae	0,547619	0,547619
IMP_biosynthesis	Naumovozyma_castellii	0,6363636	0,6
Neurosporaxanthin_biosynthesis	Baudoinia_panamericana	0,5128205	0,7179487
Pentosephosphate_pathway	Ashbya_gossypii	0,6964286	0,5892857
Pyrimidine_biosynthesis	Marssonina_brunnea	0,7307692	0,6758242
terpenoid_backbone_biosynthesis	Paracoccidioides_brasiliensis	0,1111111	0,2

Sequential GC

Species	FunOrder 2 - ICQ	org. FunOrder - ICQ
Alectoria_fallacina	0,5833333	0,6527778
Aplosporella_prunicola	0,5178571	0,3214286
Ascodesmis_nigricans	0,7321429	0,5892857
Ascoidea_rubescens	0,8	0,6
Aulographum_hederae	0,75	0,6666667
candida_albicans_L26	0,6666667	0,6190476
cephellophora_europea	0,6666667	0,4888889
Cryomyces_minteri	0,6166667	0,6666667
Dactylellina_haptotyla	0,75	0,55
Dactyrella_cylindrospora	0,5	0,8666667
Drechslerella_brochopaga	0,7857143	0,7142857
Eremomyces_bilateralis	0,4464286	0,7142857
Heterodermia_speciosa	0,6111111	0,5
Kalaharituber_pfeilii	NA	NA
Lasiodiplodia_theobromae	0,5972222	0,6527778
Neofusicoccum_parvum	0,6111111	0,6555556
Neolecta_irregularis	0,5555556	0,6444444
Orbilia_oligospora	0,9	0,8666667
Phialophora_americana	0,5238095	0,4761905
Piedraia_hortae	0,45	0,75
Polychaeton_citri	0,6888889	0,6888889
Pseudovirgaria_hyperparasitica	0,7666667	0,6666667
Pyronema_omphalodes	0,6607143	0,6428571
Rhinocladiella_mackenziei	0,8035714	0,6428571
Rhizodiscina_lignyota	0,6111111	0,7777778
Saccharata_proteae	0,5714286	0,6666667
Saitoella_complicata	0,8333333	0,7380952
Taphrina_deformans	0,9047619	0,7857143
Tirmania_nivea	0,6785714	0,8571429
Yamadazyma_tenuis	0,6666667	0,8

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Negative control GC

NC GC	Total number of trees calculated	FunOrder2 - ICQ	org. FunOrder - ICQ
1	5	0,95	0,8
2	4	0,6666667	0,66
3	7	0,8333333	0,64
4	4	0,8333333	0,83
5	4	0,5833333	0,66
6	3	1	1
7	4	0,6666667	0,66
8	3	1	1
9	7	0,7380952	0,66
10	3	1	1
11	7	0,8333333	0,81
12	5	0,85	0,75
13	5	0,95	0,6
14	6	0,9	0,66
15	3	1	0,66
16	3	1	1
17	4	0,8333333	0,83
18	7	0,7619048	0,81
19	6	0,8333333	0,73
20	5	0,8	0,7
21	3	1	1
22	5	0,95	0,95
23	4	0,75	0,92
24	7	0,8095238	0,62
25	8	0,7321429	0,68
26	7	0,7857143	0,86
27	5	0,7	0,65
28	6	0,8	0,77
29	7	0,7619048	0,74
30	3	0,5	0,5
31	8	0,6785714	0,61
32	3	0,8333333	0,83
33	5	0,8	0,8
34	5	1	0,75
35	5	0,8	0,75
36	5	0,85	0,75
37	4	1	0,83
38	4	0,8333333	0,66
39	5	0,85	0,9
40	6	0,7666667	0,76
41	5	0,9	0,55
42	3	0,8333333	0,83
43	4	0,6666667	0,58
44	7	0,8095238	0,76
45	7	0,7142857	0,76
46	8	0,6607143	0,59

47	5	0,8	0,75
48	5	0,8	0,85
49	5	0,55	0,5
50	4	0,8333333	0,83
51	6	0,8333333	0,76
52	6	0,8	0,83
53	5	0,8	0,7
54	5	0,9	0,8
55	5	0,75	0,65
56	7	0,8571429	0,81
57	8	0,7142857	0,73
58	4	0,75	0,75
59	5	0,75	0,7
60	2	1	1

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

S2 Table

Cluster definition based on the strict distance matrix

	cluster
orf1	4
orf2	1
lova	2
lovb	3
lovg	3
lovc	3
lovd	2
orf8	1
love	4
orf10	2
lovf	2
orf13	1
orf14	4
orf15	1
orf16	1
orf17	1
orf18	1

Cluster definition based on the combined distance matrix

	cluster
orf1	3
orf2	1
lova	3
lovb	3
lovg	3
lovc	1
lovd	3
orf8	1
love	2
orf10	3
lovf	3
orf13	1
orf14	2
orf15	1
orf16	1
orf17	1
orf18	1

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

S3 Table

The first part of this table (91 rows 28 columns) is too large to be displayed. The S3 Table can be found online at the publishing journal.

Statistical analysis

Original FunOrder method

	TP	**FN**	**TN**	**FP**
1) essential genes	153	88	189	44
2) biosynthetic genes	129	66	189	44

	essential genes	biosynthetic genes
Sensitivity	0,6349	0,6615
Specificity	0,8112	0,8112
Precision	0,7766	0,7457
Negative Predictive Value	0,6823	0,7412
False Positive Rate	0,1888	0,1888
False Discovery Rate	0,2234	0,2543
False Negative Rate	0,3651	0,3385
Accuracy	0,7215	0,743
F1 Score	0,6986	0,7011
Matthews Correlation Coefficient	0,4524	0,4797
Normalized Matthews Correlation Coefficient	0,7262	0,73985
No-information error rate ni	0,5084	0,5444

BLAST + Automated Cluster Detection

	TP	FN	TN	FP
1) essential genes	148	93	199	34
2) biosynthetic genes	125	70	199	34

	essential genes	biosynthetic genes
Sensitivity	0,6141	0,641
Specificity	0,8541	0,8541
Precision	0,8132	0,7862
Negative Predictive Value	0,6815	0,7398
False Positive Rate	0,1459	0,1459
False Discovery Rate	0,1868	0,2138
False Negative Rate	0,3859	0,359
Accuracy	0,7321	0,757
F1 Score	0,6998	0,7062
Matthews Correlation Coefficient	0,4813	0,5103
Normalized Matthews Correlation Coefficient	0,74065	0,75515
No-information error rate ni	0,5084	0,5444

DIAMOND + Automated Cluster Detection (FunOrder 2)

	TP	FN	TN	FP
1) essential genes	151	90	199	34
2) biosynthetic genes	128	67	199	34

	essential genes	biosynthetic genes
Sensitivity	0,6266	0,6564
Specificity	0,8541	0,8541
Precision	0,8162	0,7901
Negative Predictive Value	0,6886	0,7481
False Positive Rate	0,1459	0,1459
False Discovery Rate	0,1838	0,2099
False Negative Rate	0,3734	0,3436
Accuracy	0,7384	0,764
F1 Score	0,7089	0,7171
Matthews Correlation Coefficient	0,4926	0,5242
Normalized Matthews Correlation Coefficient	0,7463	0,7621
No-information error rate ni	0,5084	0,5444

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

AMERICAN SOCIETY FOR MICROBIOLOGY
Microbiology® Resource Announcements
GENOME SEQUENCES

Genome Sequence of the Black Yeast-Like Strain *Aureobasidium pullulans* var. *aubasidani* CBS 100524

Gabriel A. Vignolle,[a] **Robert L. Mach,**[a] **Astrid R. Mach-Aigner,**[a] **Christian Derntl**[a]

[a]Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Vienna, Austria

ABSTRACT In this work, we present the whole-genome sequence and the complete mitochondrial sequence of the black yeast-like strain *Aureobasidium pullulans* var. *aubasidani* CBS 100524, which produces the exopolysaccharide aubasidan and was previously isolated from *Betula* sp. slime flux from the Leningrad Region of Russia.

Aureobasidium pullulans is a yeast-like ascomycete with industrial relevance due to its extracellular polysaccharides (1). The main exopolysaccharide of *A. pullulans* var. *aubasidani* strain CBS 100524 is aubasidan rather than pullulan (2, 3). This strain was previously isolated from plant exudates of a *Betula* sp. from the Leningrad Region of Russia (2). Despite the difference in the secreted extracellular polysaccharides, *A. pullulans* var. *aubasidani* strain CBS 100524 is part of a main phylogenetic group (phylogenetic difference below 0.25 based on a multilocus alignment with a bootstrap value of 100) within the *A. pullulans* species complex. This group also includes the ex-neotype strain *A. pullulans var. pullulans* CBS 584.75 and the sequenced strain *A. pullulans* var. *pullulans* EXF-150 (3).

A. pullulans strain CBS 100524 was cultivated in malt extract medium (30 g/liter malt extract, 1 g/liter peptone) at 24°C and 220 rpm for 24 h. The biomass was filtered through Miracloth (EMD Millipore Corp., Burlington, MA, USA), lyophilized, and stored at −20°C. Genomic DNA was extracted as described in reference 4, sheared through sonication, purified using the GeneJET PCR purification kit (Thermo Fisher Scientific, Inc., Waltham, MA, USA), and then size selected for 800-bp fragments using NEBNext Ultra sample purification beads (New England Biolabs, Ipswich, MA, USA). The library was prepared using the NEBNext Ultra II DNA library kit with purification beads and NEBNext multiplex oligos for Illumina (index primer set 2) (both New England Biolabs) and sequenced on a MiSeq instrument using a v3 reagent kit (600 cycles, 2 × 300-bp paired-end reads) (both Illumina, Inc., San Diego, CA, USA).

The sequencing yielded 2,892,731 read pairs. First, a crude *de novo* assembly was performed using SPAdes v3.13.1 (5) with default parameters. From this initial assembly, mitochondrial sequences were identified by a BLAST analysis against the nonredundant nucleotide database (6). Next, these sequences were used as seed input for NOVOplasty v3.7 (7) for a *de novo* assembly of the mitochondrial genome sequence (one circular contig; size, 37,556 bp; coverage, 358×). Using the mitochondrial genome sequence as index built with Bowtie v1.2.2 (8), the mitochondrial reads were extracted from the raw reads. The mitochondrion-free reads were then re-paired using Fastq-pair (9), quality checked and trimmed using Trimmomatic (10), leaving 2,543,186 read pairs, and then mapped against the reference genome *A. pullulans* strain EXF-150 (GenBank accession no. GCA_000721785.1) with BWA (11) and combined and sorted using SAMtools v1.7 (12) and Picard (13). A first genome representation was extracted using ANGSD v0.925 (Analysis of Next Generation Sequencing Data) (14). The genome assembly was iteratively improved using SSPACE-Standard v3.0 (15), GapFiller v1-10 (16), and Pilon v1.21 (17). tRNA genes were detected using tRNAscan-SE v1.3.1 (18). Genes

Citation Vignolle GA, Mach RL, Mach-Aigner AR, Derntl C. 2021. Genome sequence of the black yeast-like strain *Aureobasidium pullulans* var. *aubasidani* CBS 100524. Microbiol Resour Announc 10:e01293-20. https://doi.org/10.1128/MRA.01293-20.

Editor Jason E. Stajich, University of California, Riverside

Copyright © 2021 Vignolle et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Christian Derntl, christian.derntl@tuwien.ac.at.

Received 10 November 2020
Accepted 23 February 2021
Published 25 March 2021

were predicted with AUGUSTUS v3.3.2 (19), trained with the reference genome *A. pullulans* strain EXF-150 according to reference 20. The assembly was masked using RepeatMasker v4.0.9 (21), based on the Dfam_3.0 database to identify repetitive elements. We used QUAST v5.0.2 (22, 23), including the fungal (fungi_odb9) Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2 (24), for the final evaluation.

The assembly consists of 83 scaffolds (total sequence length, 30,265,078 bp; N_{50}, 1,201,293 bp; GC content, 50.50%; mean coverage, 28×), and 10,978 genes (99.31% complete BUSCO genes found) and 353 tRNAs were predicted.

Data availability. The raw reads were uploaded to the Sequence Read Archive (SRA) under the accession no. SRR12830835. The complete genome sequence was deposited at DDBJ/ENA/GenBank under the accession no. JADGIM000000000. The version described in this paper is version JADGIM000000000.1. The complete mitochondrial genome sequence was deposited under GenBank accession no. MW148763.

ACKNOWLEDGMENTS

This study was supported by the Austrian Science Fund (FWF) (P29556) given to R.L.M. and by TU Wien (Ph.D. program TU Wien bioactive).

REFERENCES

1. Rekha MR, Sharma CP. 2007. Pullulan as a promising biomaterial for biomedical applications: a perspective. Trends Biomater Artif Organs 20:116–121.
2. Yurlova NA, de Hoog GS. 1997. A new variety of *Aureobasidium pullulans* characterized by exopolysaccharide structure, nutritional physiology and molecular features. Antonie Van Leeuwenhoek 72:141–147. https://doi.org/10.1023/a:1000212003810.
3. Zalar P, Gostinčar C, de Hoog GS, Uršič V, Sudhadham M, Gunde-Cimerman N. 2008. Redefinition of *Aureobasidium pullulans* and its varieties. Stud Mycol 61:21–38. https://doi.org/10.3114/sim.2008.61.02.
4. Paun O, Turner B, Trucchi E, Munzinger J, Chase MW, Samuel R. 2016. Processes driving the adaptive radiation of a tropical tree (Diospyros, Ebenaceae) in New Caledonia, a biodiversity hotspot. Syst Biol 65:212–227. https://doi.org/10.1093/sysbio/syv076.
5. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
6. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
7. Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. Nucleic Acids Res 45:e18. https://doi.org/10.1093/nar/gkw955.
8. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
9. Edwards JA, Edwards RA. 2019. Fastq-pair: efficient synchronization of paired-end fastq files. bioRxiv https://doi.org/10.1101/552885.
10. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
11. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
13. Broad Institute. 2019. Picard toolkit. http://broadinstitute.github.io/picard/.
14. Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15:356. https://doi.org/10.1186/s12859-014-0356-4.
15. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579. https://doi.org/10.1093/bioinformatics/btq683.
16. Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. Genome Biol 13:R56. https://doi.org/10.1186/gb-2012-13-6-r56.
17. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963.
18. Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.955.
19. Stanke M, Morgenstern B. 2005. AUGUSTUS: a Web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:W465–W467. https://doi.org/10.1093/nar/gki458.
20. Hoff KJ, Stanke M. 2019. Predicting genes in single genomes with AUGUSTUS. Curr Protoc Bioinformatics 65:e57. https://doi.org/10.1002/cpbi.57.
21. Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. http://repeatmasker.org.
22. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
23. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34:i142–i150. https://doi.org/10.1093/bioinformatics/bty266.
24. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Genome sequencing of *Wardomyces moseri*: a rare but cosmopolitan fungus with an outstanding secondary metabolite production potential

Gabriel A. Vignolle[1], Nadine Hochenegger[1], Jana M. U'Ren[2], Robert L. Mach[1], Astrid R. Mach-Aigner[1], Mohammad Javad Rahimi[1], Kamariah A. Salim[3], Chin Mei Chan[4], Linda B.L. Lim[4], Feng Cai[5], Irina S. Druzhinina[1,5], Christian Derntl[1§]

[1] Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Gumpendorfer Strasse 1a, 1060 Wien, Austria
[2] BIO5 Institute and Department of Biosystems Engineering, University of Arizona, Tucson, AZ, USA
[3] Environmental and Life Sciences, Faculty of Science, Universiti Brunei Darussalam, Jalan Tungku Gadong BE1410, Brunei Darussalam
[4] Chemical Sciences, Faculty of Science, Universiti Brunei Darussalam, Jalan Tungku Link, Brunei Darussalam
[5] Fingal Genomics Group, College of Resources and Environmental Sciences, Nanjing Agricultural University, Weigang NO. 1, Nanjing 210095, China
§Corresponding author

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

ABSTRACT

Background: Most secondary metabolites with industrial and biomedical importance are produced by only a small set of filamentous fungi of the Eurotiales (e.g. *Aspergillus*, *Penicillium*) and the Hypocreales, (e.g. *Tolypocladium*, *Fusarium*). Therefore, the analysis of filamentous fungi from other clades, promises the discovery of yet unknown substances with yet unknown properties. The ascomycete *Wardomyces moseri* was first isolated from a dead petiole of *Mauritia minor* in Colombia in 1980 and described by W. Gams in 1995. During a phylogenetic study in 2016, focusing on the taxonomy of the family *Microascaceae*, *W. moseri* was suggested to be phylogenetically misplaced and should therefore be re-evaluated.

Results: We analyse the slumbering metabolic potential of this historic fungus and re-evaluate its taxonomy, by sequencing the genomes of the ex-isotype strain *W. moseri* CBS 164.80 and two isolates from the opposite side of the world, *W. moseri* TUCIM 5827 and TUCIM 5799. We show how historic strains from already existing collections can be leveraged for the search of novel natural products.

Conclusion: We could demonstrate the vast and untapped secondary metabolite potential of the historic *W. moseri* strain CBS 164.80. Further, we identified numerous and diverse biosynthetic gene clusters (BGC), including a melanin cluster potentially responsible for the dark spore pigmentation. Many of these novel BGCs are not represented in the genomes of other compared fungi. Confirming the suggested slumbering potential in historic fungal strain collections. Furthermore, we leveraged the genome assemblies to re-evaluate a disputed taxonomic placement of the species and could indicate, that *Wardomyces moseri* is part of the family *Sporocadaceae* within the order of Xylariales (Dikarya, Ascomycota, Pezizomycotina, Sordariomycetes, Xylariomycetidae).

KEYWORDS

Fungi, *Wardomyces moseri*, genome mining, ascomycota, Xylariales, reclassification, secondary metabolism, comparative genomics

BACKGROUND

The ascomycete *Wardomyces moseri* was first isolated from a dead petiole of *Mauritia minor* in Colombia in 1980. Walter Gams described the fungus in 1995 and named it after his mentor Meinhard Moser (CBS 164.80) (4). This fungus forms sporodochium-like structures and aggregates conidia loosely in slimy masses. *W. moseri* was described already in 1995 as an unusual *Wardomyces* species, because of its easily liberated conidia. Later, Sandoval-Denis *et al.* showed that the large subunit (LSU) rRNA gene and the internally transcribed spacer (ITS) sequences of *W. moseri* clustered among the Xylariales but not with the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Wardomyces (5). *W. moseri* appears related to members of the *Amphisphaeriaceae* and *Clypeosphaeriaceae*. Based on these findings, *W. moseri* was suggested to be re-evaluated concerning its taxonomic placement. To date, there is only one more preprint mentioning this fungus indicating again the apparent misclassification (6).

The fungal order of Xylariales (Ascomycota) contains a large number of symbionts, saprotrophs, a variety of isolated endophytes, and plant pathogens (7-9). Many Xylariales are macromycetes, forming club or wart like fruiting bodies (stromata). Two prominent species of the Xylariales are *Eutypa lata* and *Pestalotiopsis fici*. *E. lata* is most commonly known as the vascular pathogen that causes Eutypa dieback in grapevines (10). Whereas *P. fici* is a commonly isolated endophyte from healthy plant tissue and, at the same time, a plant pathogen with a strong economic impact (1). The Xylariales are one of the largest clades of filamentous fungi and represent one of the most prolific lineages of secondary metabolite (SM) producers. Until now, several hundred SMs of unique carbon backbone structure were discovered from fungi of this order, including various drug lead compounds (7, 8).

In general, SMs are compounds with an abundance of diverse chemical structures and properties. They are found in each domain of life but are predominantly studied in microorganisms and plants. SMs are not essential for the survival and growth of an organism but can be beneficial under specific environmental conditions, e.g., antibiotics in competitive situations, pigments to withstand radiation, and toxins as either defensive or virulence factors (11, 12). SMs can be classified into different classes according to their biosynthetic pathways. In fungi, the two main classes are non-ribosomal peptides (*e.g.* the antibiotic penicillin (13) or the immunosuppressant cyclosporine (14)) and polyketides (e.g. the mycotoxin aflatoxin (15) or the cholesterol-lowering drug lovastatin (16)). Further SM classes are alkaloids, terpenes, melanins (17, 18), and ribosomally synthesized and post-translationally modified peptides (RiPPs) (19). The genes encoding the enzymes responsible for the production of SMs are spatially organized in biosynthetic gene clusters (BGCs) in many cases.

SMs from fungal sources have been used for medicinal purposes and to promote and maintain the human well-being already since ancient times (20-22). Fungal SM and chemically modified variants are widely used as antibiotics, immunomodulators and anti-cancer drugs (23). Interestingly, most SMs with industrial and biomedical importance are produced by only a small set of filamentous fungi of the Eurotiales (e.g. *Aspergillus*, *Penicillium*) and the Hypocreales, (e.g. *Tolypocladium*, *Fusarium*) (24-26). Therefore, the analysis of filamentous fungi from other clades, especially prolific SM producers such as the Xylariales promises the discovery of yet unknown substances with yet unknown properties (7-9, 27).

In this study, we isolate two new *W. moseri* strains (TUCIM 5827 and TUCIM 5799) and compare them to the type strain *W. moseri* CBS 164.80 in a comparative genetics approach. The genomes of all three strains were sequenced using an Illumina MiSeq platform, and their genes predicted and annotated. Based on high accuracy phylogenetic tree inference, we suggest replacing *W. moseri* in the family *Sporocadaceae* (order Xylariales).

RESULTS

Isolation of two new W. moseri strains

The epiphytic fungi TUCIM 5827 and TUCIM 5799 were isolated from the adaxial surface of the healthy high canopy leaf of *Shorea johorensis* (Dipterocarpaceae, Malvales; DNA BarCode maturase K (matK) deposited in NCBI GenBank MF993320.1,

Table 1 Average nucleotide identity (ANI) between the *W. moseri* strains.

Genomes compared	**ANI**
CBS 164.80 : TUCIM 5799	99.0276
CBS 164.80 : TUCIM 5827	99.0091
TUCIM 5799 : TUCIM 5827	99.1092

Laciny et al., 2008) (Borneo). The macro- and microscopic morphology of *W. moseri* CBS 164.80, TUCIM 5827 and TUCIM 5799 are shown in figure 1 (Fig. 1). For a detailed macroscopic and microscopic description of *W. moseri* CBS 164.80 we refer to the original publication by W. Gams (4). The ITS sequences of both isolates were highly similar to *W. moseri* CBS 164.80 (Additional file 1). Additionally, the average nucleotide identity (ANI) between the three *W. moseri* strains (Table 1) strongly suggested that these isolates belong to the same species. Furthermore, the high ANI suggests a stable genomic architecture and high relatedness especially considering the spatiotemporal distance of their isolation and origin.

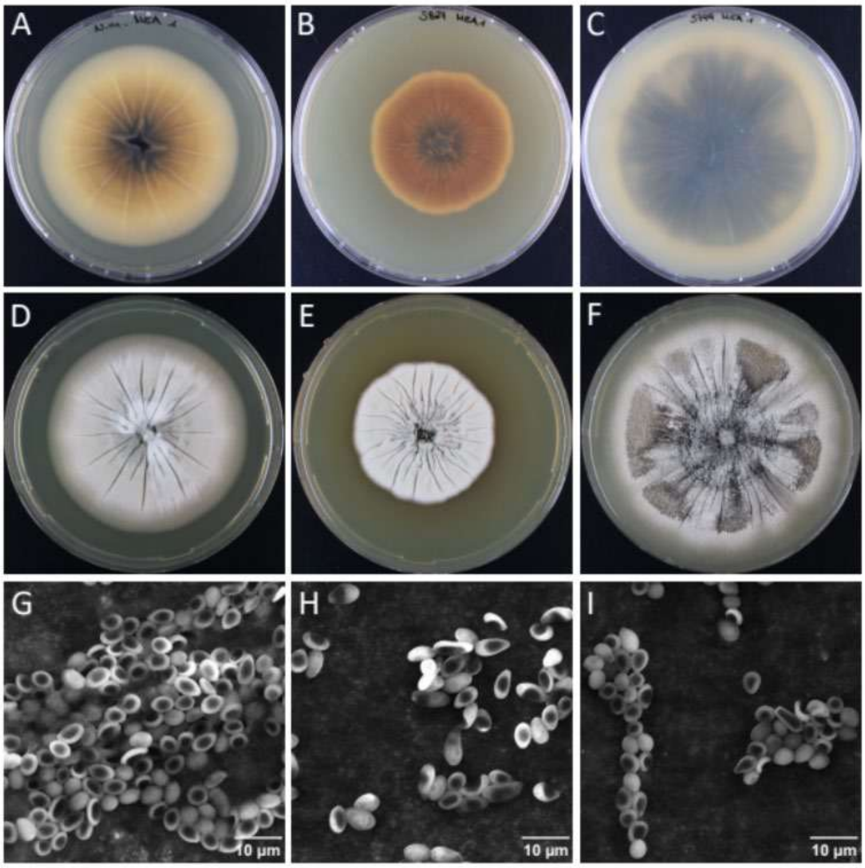


Fig. 1 Macro- and microscopic morphology of *W. moseri* CBS 164.80, TUCIM 5827 and TUCIM 5799. The three *W. moseri* strains were grown on malt extract agar plates at 28 °C in darkness and pictures were taken after 15 days of incubation. The first row displays the plates underside of *W. moseri* CBS 164.80 (A), TUCIM 5827 (B) and TUCIM 5799 (C). The second row shows pictures taken from above of *W. moseri* CBS 164.80 (D), TUCIM 5827 (E) and TUCIM 5799 (F). The third row visualizes the spores of *W. moseri* CBS 164.80 (G), TUCIM 5827 (H) and TUCIM 5799 (I) using scanning electron microscopy (SEM).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Phylogenetic placement

To determine the phylogenetic placement of *W. moseri*, we first performed a high accuracy orthogroup inference on the predicted proteomes of the three *W. moseri* strains, *Trichoderma reesei*, *Coccidioides immitis*, *Aspergillus flavus*, *P. fici*, *Sporothrix schenckii*, *E. lata*, *Penicillium roqueforti*, and *Fusarium fujikuroi*, applying 4213 single-locus gene trees, each based on a single copy orthologue (Fig. 2). This approach clearly places *W. moseri* among the Xylariales between *P. fici* and *E. lata* (Fig. 2), which had previously been suggested by Sandoval-Denis *et al.* (2). For a more precise phylogenetic placement, we performed a high accuracy orthogroup inference on the predicted proteomes of the three *W. moseri* strains, *E. lata*, *P. fici*, *Neopestalotiopsis clavispora*, *Truncatella*

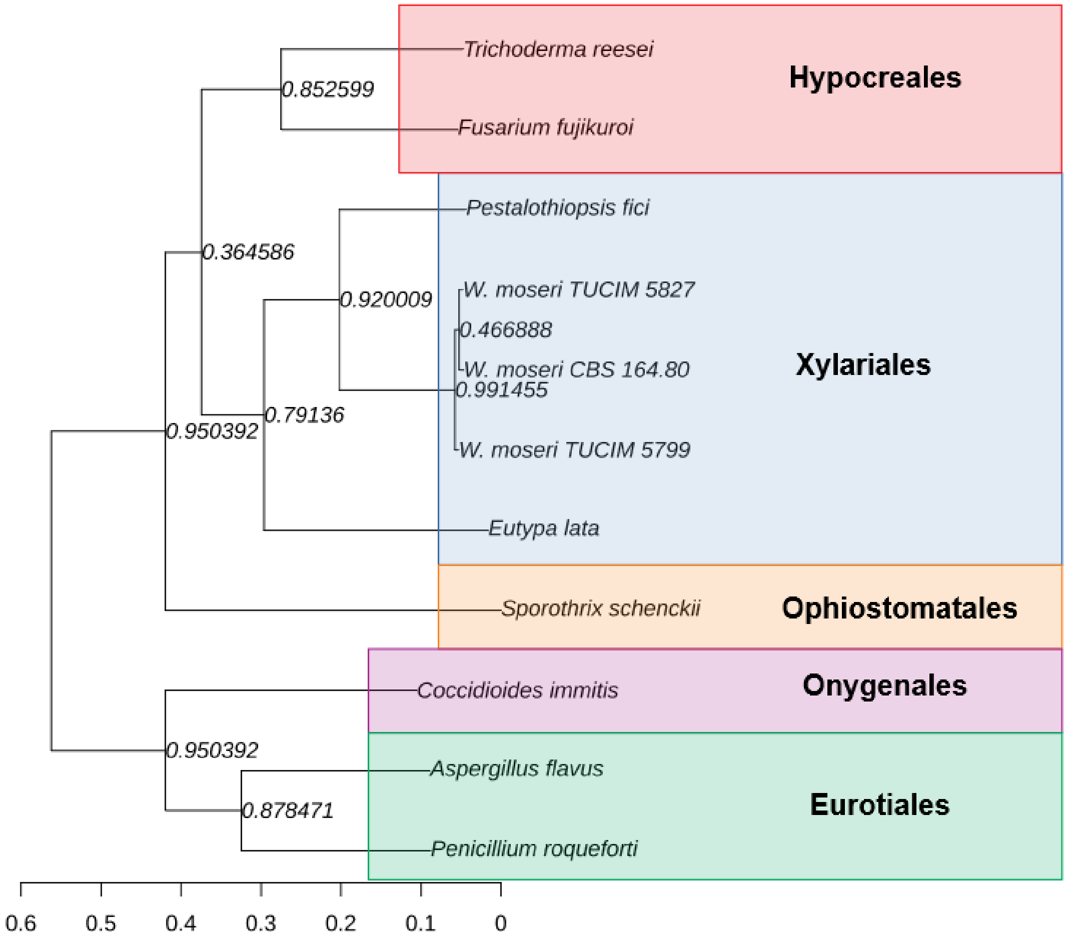


Fig. 2 Inferred rooted phylogenetic tree based on single-locus gene trees. Phylogenetic inference applying 4213 single-locus gene trees, each based on a single copy orthologue, from the predicted proteomes of *W. moseri* CBS 164.80*, W. moseri* TUCIM 5827*, W. moseri* TUCIM 5799*, T. reesei* QM6a, *C. immitis*, *A. flavus* NRRL3357, *P. fici*, *S. schenckii*, *E. lata* URCEL1, *P. roqueforti* LCP96 04111 and *F. fujikuroi* IMI 58589 (3). The species tree extrapolation was performed with STAG (Species Tree Inference from All Genes), which uses the fraction of species trees derived from single-locus gene trees supporting each bipartition as its degree of support for each node. The red colored box indicates the order of Hypocreales, the blue colored box the order of Xylariales, the yellow box the order of Ophiostomatales, the purple colored box the order of Onygenales and the green colored box the order of Eurotiales.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

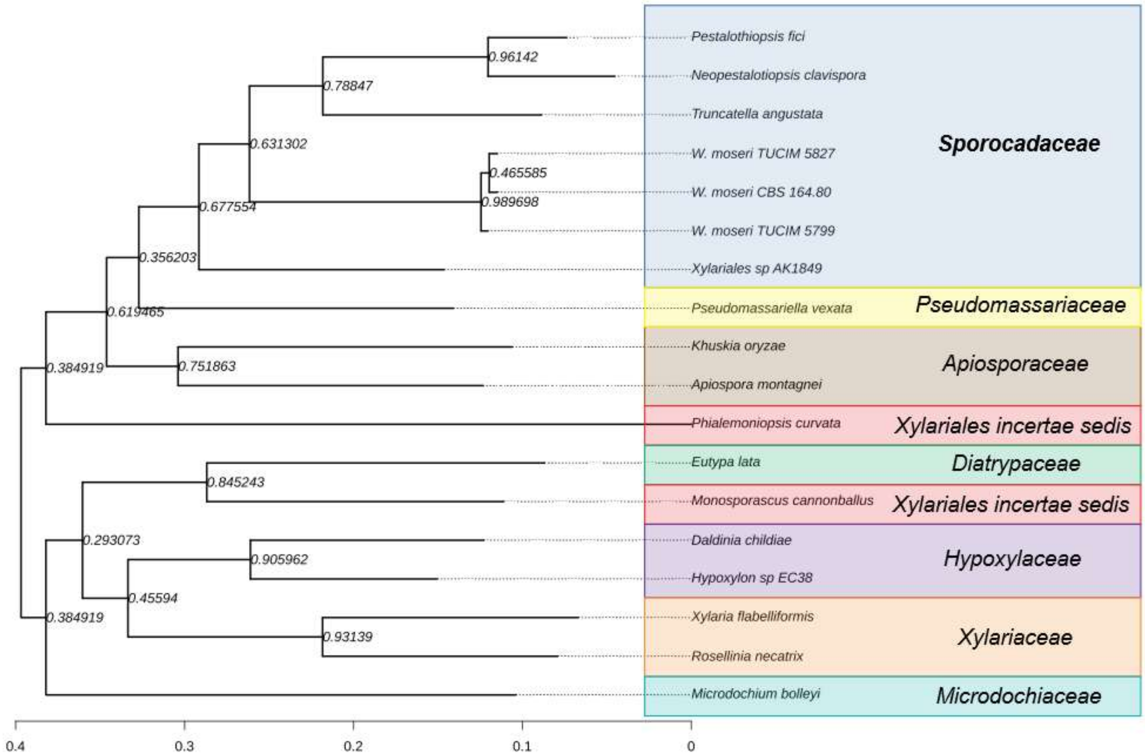


Fig. 3 Inferred rooted phylogenetic tree based on single-locus gene trees. Phylogenetic inference applying 3041 single-locus gene trees, each based on a single copy orthologue, from the predicted proteomes of *W. moseri* CBS 164.80*, W. moseri* TUCIM 5827*, W. moseri* TUCIM 5799*, P. fici*, *E. lata*, *Neopestalotiopsis clavispora*, *Truncatella angustata*, *Xylariales* sp. AK1849, *Pseudomassariella vexata*, *Khuskia oryzae*, *Apiospora montagnei*, *Phialemoniopsis curvata*, *Monosporascus cannonballus*, *Daldinia childiae*, *Hypoxylon* sp. EC38*, Xylaria flabelliformis*, *Rosellinia necatrix* and *Microdochium bolleyi* (3). The species tree extrapolation was performed with STAG (Species Tree Inference from All Genes), which uses the fraction of species trees derived from single-locus gene trees supporting each bipartition as its degree of support for each node. The blue colored box indicates members of the family of *Sporocadaceae*, the yellow box the family of *Pseudomassariaceae*, the brown box the family of Apiosporaceae, the red colored box indicates the family of *Xylariales incertae sedis*, the green colored box the family of *Diatrypaceae*, the purple colored box the family of *Hypoxylaceae*, the orange box the members of the family of *Xylariaceae* and the turquoise box the family *Microdochiaceae*.

angustata, *Xylariales* sp. AK1849, *Pseudomassariella vexata*, *Khuskia oryzae*, *Apiospora montagnei*, *Phialemoniopsis curvata*, *Monosporascus cannonballus*, *Daldinia childiae*, *Hypoxylon* sp. EC38*, Xylaria flabelliformis*, *Rosellinia necatrix* and *Microdochium bolleyi*, applying 3041 single-locus gene trees, each based on a single copy orthologue (Fig. 3). The phylogenetic tree places the *W. moseri* strains as sisters to a phylogenetic subtree containing the species *P. fici*, *N. clavispor*a and *T. angustata*. Further, *W. moseri* can be found between the above-mentioned subtree and *Xylariales* sp. AK1849, all of them have previously been taxonomically placed within the phylogenetic family *Sporocadaceae* (1, 28, 29) (Fig. 3).

The obtained results indicate that this fungus lineage is part of the family *Sporocadaceae* within the order of Xylariales (Dikarya, Ascomycota, Pezizomycotina, Sordariomycetes, Xylariomycetidae) and that it

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

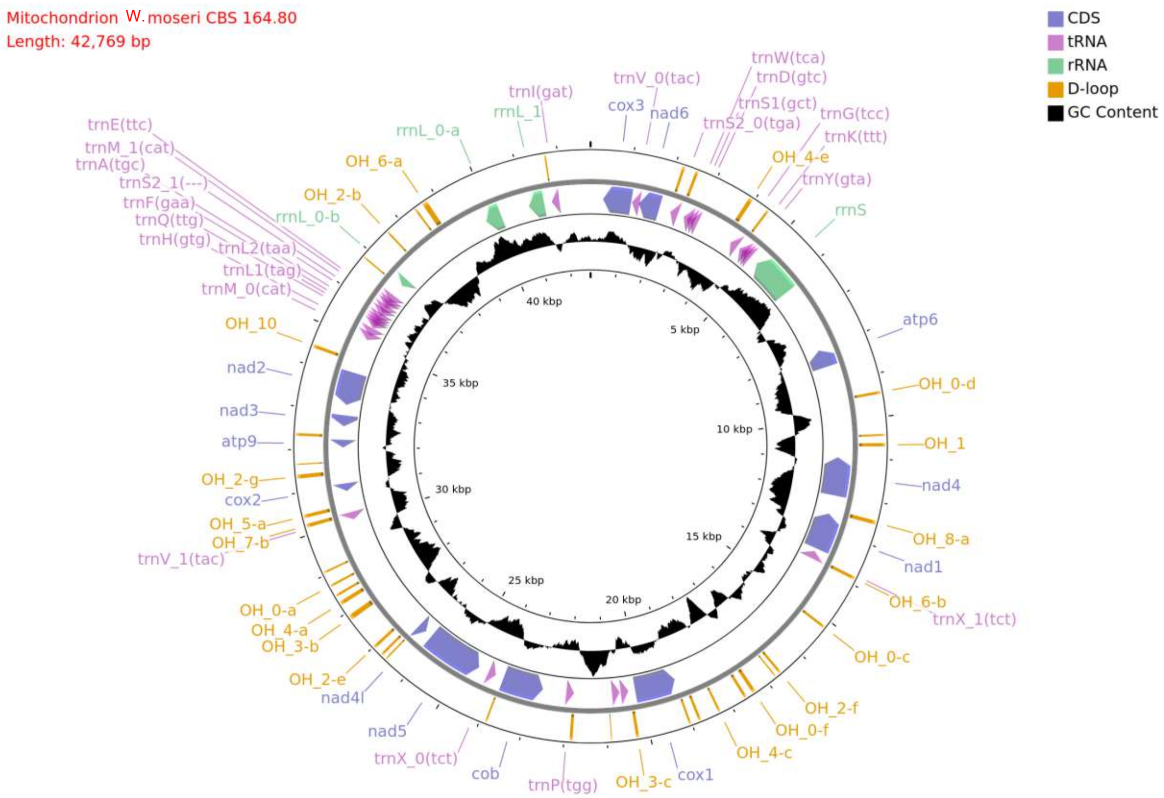


Fig. 4 Circular visualization of the whole circularized mitochondrial genome of *W. moseri* CBS 164.80. The purple boxes indicate coding sequences (CDS), the magenta boxes indicate tRNA genes, the mint boxes indicate rRNA genes and the orange boxes indicate D-loop control regions. The next inner circle represents a histogram of the GC content.

is not closely related to a yet described and sequenced genus in this clade. A recent preprint by Samarakoon et al. places *W. moseri* based on a multi locus ML tree (using ITS, LSU, rpb2, tub2 and tef1 genes) within the family *Amphisphaeriaceae* and next to *Beltraniaceae* (6). Based on our phylogenetic analysis, we propose to place *Wardomyces moseri* within the family *Sporocadaceae*.

Genome Sequencing

The total DNA of the strain *W. moseri* CBS 164.80, the strain *W. moseri* TUCIM 5827 and the strain *W. moseri* TUCIM 5799 was extracted, and two libraries were created with a DNA fragment length of 1293 ± 6 bp and 1136 ± 7 bp, the average DNA concentrations were 34.93 ± 0.25 ng/µl and 10.63 ± 0.23 ng/µl, resulting in a 40.899 nM and a 14.178 nM library respectively. Three sequencing runs on an Illumina MiSeq platform (two V3 Reagent Kit (600 cycles) and one V2 nano Reagent Kit (500 cycles)) resulted in a total of 67,670,936 pe-reads. The raw data were deposited at the Sequence Read Archive (SRA) under the accession SRR13570309, SRR13747339 and SRR13747338. Next, the mitochondrial reads were removed from the raw reads and then the mitochondrial and the nuclear genomes assembled individually.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Table 2 Genome assembly characteristics and found Benchmarking Universal Single-Copy Orthologues (BUSCO) genes of the assembled *W. moseri* strains CBS 164.80, TUCIM 5827 and TUCIM 5799.

Genome	**CBS 164.80**	**TUCIM 5827**	**TUCIM 5799**
Assembly size (bp)	43,702,215	46,154,457	44,394,130
G+C content (%)	52.77	52.65	52.66
Scaffolds (>= 0 bp)	230	2730	693
Scaffolds (>= 1000 bp)	193	609	221
Largest scaffold (bp)	2,337,669	1,719,970	2,329,648
N50 (bp)	506,940	462,712	764,765
L50 (scaffolds)	26	30	17
N's per 100 kbp	2.33	2.17	1.81
Complete BUSCO (%)	100.00	100.00	100.00
Partial BUSCO (%)	0.00	0.00	0.00

The mitochondrial genome

The extracted circularized mitochondrial genomes have a length of 42,769 bp, 42,769 bp, and 43,978 bp with GC contents of 27.52%, 27.53%, and 27.52% for the strain CBS 164.80, TUCIM 5827, and TUCIM 5799, respectively (Fig. 4). The respective average sequencing coverages were at 364x, 464x, and 8,939x. The mitochondrial genomes were deposited at GenBank with the accession no. MW554918, MW660809, and MW660808. Fungal mitochondrial genomes normally contain 14 protein-coding genes, i.e., three cytochrome c oxidase subunits (*cox1*, *cox2*, *cox3*), apocytochrome b (*cob*), seven NADH dehydrogenase subunits (*nad1*, *nad2*, *nad3*, *nad4*, *nad5*, *nad6*, *nad4L*), 3 ATP synthase F0 subunits (*atp6*, *atp8*, *atp9*) and 1 ribosomal protein S3 gene (*rps3*) (30, 31). This is also the case for the *W. moseri* strains, with one exception. The *atp8* gene (encoding for the ATP synthase F0 subunit 8) is missing in the mitochondrial genomes of the *W. moseri* strains. This gene was presumably transferred to the nuclear genome, as a single gene encoding for a putative ATP synthase subunit can be found in each genome of the three strains (JN550g13373 in *W. moseri* CBS 164.80; JX266g13823 in *W. moseri* TUCIM 5827; JX265g13592 in *W. moseri* TUCIM 5799).

Table 3 Masked repetitive elements found with RepeatMasker v4.0.9 and tRNA genes found by tRNAscan-SE v1.3.1, for the *W. moseri* strains CBS 164.80, TUCIM 5827 and TUCIM 5799 respectively. *Most repeats that were fragmented by insertions or deletions have been counted as one element.

Masked element	**Number of elements***			**Length occupied in bp**			**Percentage of sequence**		
Strain	**CBS**	**5827**	**5799**	**CBS**	**5827**	**5799**	**CBS**	**5827**	**5799**
SINEs	35	33	35	2,289	2,231	2,404	0.01%	-	0.01%
LINEs	223	220	222	16,838	16,757	17,399	0.04%	0.04%	0.04%
LTR elements	4	3	3	300	200	204	-	-	-
DNA elements	50	55	49	3,751	4,260	3,598	0.01%	0.01%	0.01%
Unclassified	1	1	1	142	72	142	-	-	-
Small RNA	86	74	78	12,181	11,815	12,157	0.05%	0.05%	0.03%
Simple repeats	7,572	7,532	7,432	306,538	297,695	294,925	0.70%	0.64%	0.66%
Low complexity	652	652	600	30,991	30,995	27,670	0.06%	0.07%	0.06%
tRNA	196	190	189	17,154	16,884	16,788	0.04%	0.04%	0.04%

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Nuclear genome assembly and annotation

The size of the nuclear genomes was between 43.7 Mbp and 46.1 Mbp and the average genome coverages were between 32x and 141x. The detailed results of the genomes and assembly characteristics (size, G+C content, characteristics for scaffold number and size, N50 and L50) are summarized in Table 2 and given in Additional files 2, 3, 4. To evaluate the completeness of the genome assembly, we performed a Benchmarking Universal Single-Copy Orthologues (BUSCO) analysis with the eukaryote dataset (32). 100% complete BUSCOs without duplicates were found in all three assemblies (Table 2, Additional files 2, 3, 4). Next, we masked the repetitive elements in the nuclear genome to reduce the number of false positives during the subsequent gene prediction. A total of 372,467 bp of the *W. moseri* strain CBS 164.80 genome was masked, this represents 0.85% of the total genome. Further 380,909 bp of the *W. moseri* strain TUCIM 5827 genome was masked and 375,287 bp of the *W. moseri* strain TUCIM 5799, this represents 0.83% and 0.84% of the total genomes respectively. (Table 3, Additional files 5, 6, 7). The used tool (RepeatMasker) predicts interspersed repeats, like short interspaced nuclear repeats (SINEs), transposable element like repeats, long interspaced nuclear repeats (LINEs) and long terminal repeats (LTR), small RNAs, simple repeats and low complexity repeats and also tRNA genes.

To verify and potentially complement the tRNA predictions, we also performed an tRNA prediction with tRNAscan-SE v1.3.1 (33) using the unmasked genome, because fungal specific SINEs are associated with tRNAs (34) and might therefore influence their detection. tRNAscan-SE predicted a total of 196, 190 and 189 tRNA genes, for each strain respectively (Additional files 8, 9, 10).

For the gene prediction, we used Augustus v3.3.2 (35), because no transcriptome data was available. Augustus was trained with the gene set of *P. fici*, because this was the closest related fungus with a published genome with high quality gene predictions (1, 10). The predicted gene sets were evaluated and annotated by blasting them against the UniProt database (Table 4). Further, we used the PANNZER2 web interface for a functional annotation of the *W. moseri* proteomes (Additional files 11, 12, 13). Genes that could not be annotated via the BLAST approach were primarily annotated through this approach.

CAZymes

A hallmark of fungal biology is their saprotrophic lifestyle. Fungi are thriving on plant biomass and other carbohydrate-rich materials by degrading complex and simple carbohydrates using so-called carbohydrate active enzymes (CAZymes) (36, 37). We used dbCAN2 (a meta-server for CAZyme annotation) and a HMMer (38) (Hidden Marcov model) search, a DIAMOND (39) search and a Hotpep (40) search to predict the CAZymes in the three *W. moseri* genomes (Fig. 5, Table 5). In total, 1,005, 1,018, and 1,011 CAZymes were predicted by all three methods in the CBS 164.80, the TUCIM 5827 and the TUCIM 5799 strains, respectively (Fig. 5, Additional files 14, 15, 16). 455, 460 and 455 of all predicted CAZymes genes were predicted by all three methods (Fig. 5).

Table 4 Gene predictions

Strain	**Predicted putative genes**	**genes without BLAST hits below E^{-5}**
CBS 164.80	13,929	4,797 (34.4%)
TUCIM 5827	14,595	5,352 (36.7%)
TUCIM 5799	14,160	4,964 (35.0%)

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 5 The carbohydrate active enzymes (CAZymes) found with dbCAN2 a meta-server for CAZyme annotation. Glycosyltransferases (GT); Glycoside Hydrolases (GH); carbohydrate esterases (CE); polysaccharide lyases (PL); Redox enzymes with auxiliary activities (AA).

Strain	**Total**	**GT**	**GH**	**CE**	**PL**	**AA**
CBS 164.80	1005	148	476	93	27	222
TUCIM 5827	1018	151	476	95	27	231
TUCIM 5799	1011	152	479	94	27	222
P. fici (1)	-	121	460	138	39	-

The dbCAN2 server also predicts certain subclasses of CAZyme. Glycosyltransferases catalyze glycosidic bond formation and inversion and are part of the posttranslational modification steps in different compound formation processes (41, 42). Glycoside hydrolases is a large family of enzymes which hydrolyses glycosidic bonds. Carbohydrate esterases catalyze de-N or de-O-acylation of ester bonds in saccharides like in pectin. Polysaccharide lyases cleave polysaccharide chains via β-elimination. Redox enzymes with auxiliary activities are involved in the breakdown processes of polysaccharides and lignin. The respective numbers of the predicted CAZymes subclasses are listed in Table 5. We could identify a relative high number of all groups, which is in accordance with the assumed plant-associated lifestyle of *W. moseri* and comparable to the number of

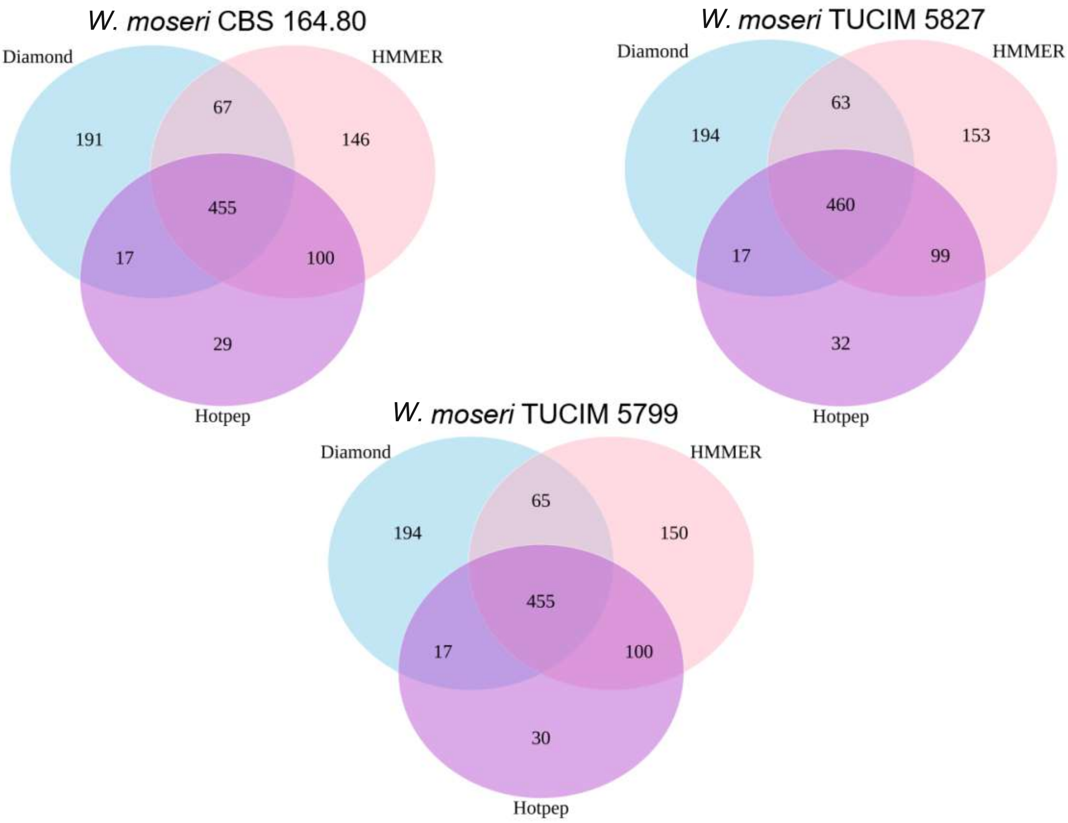


Fig. 5 Venn-plot among different search algorithms to finding CAZymes. Venn-plot showing the intersections of CAZymes predicted by different search algorithms Diamond (light blue), HMMER (pink) and Hotpep (orchid) for the three sequenced *W. moseri* strains.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

CAZymes previously reported for *P. fici* (1). Strikingly, *W. moseri* possess more glycosyltransferases than *P. fici*, but less carbohydrate esterases and polysaccharide lyases (Table 5). As a high amount of carbohydrate esterases and polysaccharide lyases have been suggested to play a role in pathogenicity (43) we speculate that *W. moseri* might have an overall lower plant pathogenicity potential than *P. fici*. Further the large arsenal of different CAZymes suggests *W. moseri* to exploit a diverse range of complex and simple carbon sources for growth (36) indicating a potential oligotrophic lifestyle (44) (Additional files 14, 15, 16).

Ligninolytic potential

In nature, plant biomass consists mainly of the two polysaccharide groups cellulose and different hemicelluloses and the polyaromatic lignin. While most fungi using the previous mentioned CAZymes can effectively degrade cellulose and the hemicelluloses, the depolymerization of liginin is only achieved by certain fungi. Enzymes with the potential to contribute to ligninolytic activities are lignin peroxidases (EC:1.11.1.14), manganese peroxidases (EC:1.11.1.13), specific laccases (EC:1.10.3.2) and versatile peroxidases (EC:1.11.1.16). A recent review by Kumar and Chandra indicates other enzymes such as aryl-alcohol oxidases (EC:1.1.3.7), lipases (EC:3.1.1.3), quinone reductases (EC:1.6.5.5), xylanase (EC:3.2.1.8), catechol 2,3-dioxygenases (EC:1.13.11.2) and feruloyl esterases (EC:3.1.1.73) to be indirect facilitators for the ligninolytic enzyme process (45). We used KofamKOALA (46) to search for these enzymes in the predicted proteomes of the *W. moseri* strains and could only find a small number of putative lignin-degrading enzymes (see Table 6, Additional files 17, 18, 19). Notably, no lignin peroxidases, specific laccases, or versatile peroxidases were found. These results strongly indicate that *W. moseri* cannot degrade lignin.

Proteases, Chitinases, Cutinases, Lipases

Next, we searched for putative proteases and peptidase inhibitors by aligning the predicted *W. moseri* proteomes against the MEROPS database (47) using blastp. Only genes with at least 20 hits aligned with an E-value less than E^{-5} were considered. We identified 604, 616, and 604 putative protease encoding genes in the strains CBS 164.80, TUCIM 5827 and TUCIM 5799, respectively,

Table 6 The enzymes potentially providing ligninolytic activities found using KofamKOALA. Lignin peroxidases (EC:1.11.1.14), manganese peroxidases (EC:1.11.1.13), specific laccases (EC:1.10.3.2), versatile peroxidases (EC:1.11.1.16), aryl-alcohol oxidases (EC:1.1.3.7), lipases (EC:3.1.1.3), quinone reductases (EC:1.6.5.5), xylanase (EC:3.2.1.8), catechol 2,3-dioxygenases (EC:1.13.11.2) and feruloyl esterases (EC:3.1.1.73). A thicker line separates the indirect facilitators from the four main enzymes (above the line) involved in the degradation of lignin.

Enzymes	**CBS 164.80**	**TUCIM 5827**	**TUCIM 5799**
Lignin peroxidases	0	0	0
manganese peroxidases	3	3	3
laccases	0	0	0
versatile peroxidases	0	0	0
aryl-alcohol oxidases	0	0	0
lipases	3	4	4
quinone reductases	6	6	6
xylanases	7	7	7
Catechol 2,3-dioxygenases	0	0	0
feruloyl esterases	12	12	12

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Table 7 The genes annotated as potential small, secreted cysteine rich proteins based on the PANNZER2 annotation.

	CBS 164.80	**TUCIM 5827**	**TUCIM 5799**
Trihydrophobin	JN550g3819; JN550g10131	JX266g12290; JX266g13180	JX265g4713; JX265g7211
Hydrophobin	JN550g12382	JX26612570	JX265g5813

Small, secreted protein	JN550g718;
	JN550g3349;
	JN550g5182;
	JN550g6140;
	JN550g7610:
	JN550g8062;
	JN550g10321
		JX266g3384;
		JX266g6643;
		JX266g9165;
		JX266g9719;
		JX266g10504;
		JX266g12338
			JX265g464;
			JX265g2970;
			JX265g5561;
			JX265g7101;
			JX265g7788;
			JX265g13552

Extracellular effector protein	JN550g5903	JX266g11925	JX265g12041

and 14 peptidase inhibitors in each of the three strains. Interestingly, 202 of the putative protease-coding genes from each strain are shared with *P. fici* with an E-value of 0.0 (Additional files 20, 21, 22), which indicates a high conservation of the proteases between the two genera. Furthermore, we found 11, 13 and 11 putative cutinases, 21, 23 and 21 potential chitinases and one secretory lipase (GO:0004806; EC:3.1.1.3; JN550g6955, JX266g10648, JX265g12244) each, based on the functional annotation by PANNZER2 (48) in the genomes of the *W. moseri* strains CBS 164.80, TUCIM 5827 and TUCIM 5799, respectively (Additional files 11, 12, 13).

Small Secreted Cysteine Rich proteins

Hydrophobins are small, secreted cysteine rich amphiphilic proteins self-assembling into insoluble polymerized amphipathic monolayers exclusively found in fungi. They were associated with pathogenicity, plat cell wall degradation, different developmental stages and proposed as potential supporters in plastic degradation (49). We detected two trihydrophobin genes and one hydrophobin gene in the CBS 164.80 strain; furthermore, we found seven genes predicted to be small secreted proteins and one gene predicted to be a extracellular effector protein. We detected two trihydrophobin genes and one hydrophobin gene in the TUCIM 5827 and TUCIM 5799 strains respectively; furthermore, we found six genes predicted to be small, secreted proteins and one extracellular effector protein in both strains (Table 7, Additional files 11, 12, 13).

Transcription factors

Transcription factors are essential for the regulation of gene expression at transcript level and are therefore of central importance for any biological system. We found 83 potential transcription factors in the predicted proteomes of the *W. moseri* strains, each. These 83 transcription factors belonged to 56 different types, according to a KEGG analysis (Additional file 23). As indicated in a recent review by Leiter *et al*. (50) the orthologues of *Schizosaccharomyces pombe Atf1* play a key role in the regulation of SM production as well as growth and development. We found the gene JN550g1888 in *W. moseri* CBS 164.80 to encode for an *Aft1*-like transcription factor. The gene was detected by a sequence similarity search with the *S. pombe Atf1* protein sequence and using our annotation (Additional file 11). Further, we searched for a conserved global regulator *veA* or velvet gene by a sequence similarity search with the *Fusarium verticillioides VE1* gene (51). Using this approach, we discovered the gene JN550g9662 in *W. moseri* CBS 164.80 to be a *veA*-like developmental and SM regulator. Identifying these two global regulators of on the one hand secondary metabolism and on the other growth and fungal development, will ease future studies in *W. moseri*.

Pathogenicity potential

Since no information about the ecological role of *W. moseri* except its association with the phyllopshere are available, we wanted to get an estimation about the pathogenicity potential of this fungus. To this end, we assessed the pathogenic potential of the *W. moseri* strains by comparing Host-pathogen protein–protein interactions (HPIs) that play an essential role in initiating infection to those of *A. flavus*, *C. immitis*, *S. schenkii*, *P. fici* and *T. reesei*. Possible HPIs of these fungi were predicted using the web version of Host Pathogen Interaction Database (HPIDB 3.0) (52). We found 470 putative gene products to have a pathogen host interaction in the predicted proteome of *W. moseri* CBS 164.80. Out of these 470 proteins, 460 proteins were predicted to have interactions with animals, 26 were associated with animals or plants, and 10 were predicted to have interactions with a plant host (Fig. 6, Additional files 24), the two TUCIM strains displayed a similarly low pathogenic potential (Fig. 6, Additional files 25, 26). These numbers are very low in comparison to the known human pathogenic species *C. immitis* and *S. schenckii*, the opportunistic pathogenic fungus *A. flavus*, the plant pathogenic fungus *P. fici* and the safely used *T. reesei* (generally regarded as safe - GRAS status) as a non-pathogenic strain (1, 53-56) (Fig. 6). Based on these comparative analyses we suggest a low overall pathogenic potential, if any, for the *W. moseri* strain CBS 164.80 and the two TUCIM strains.

Interestingly, two of the potential plant pathogenicity factors (JN550g516 and

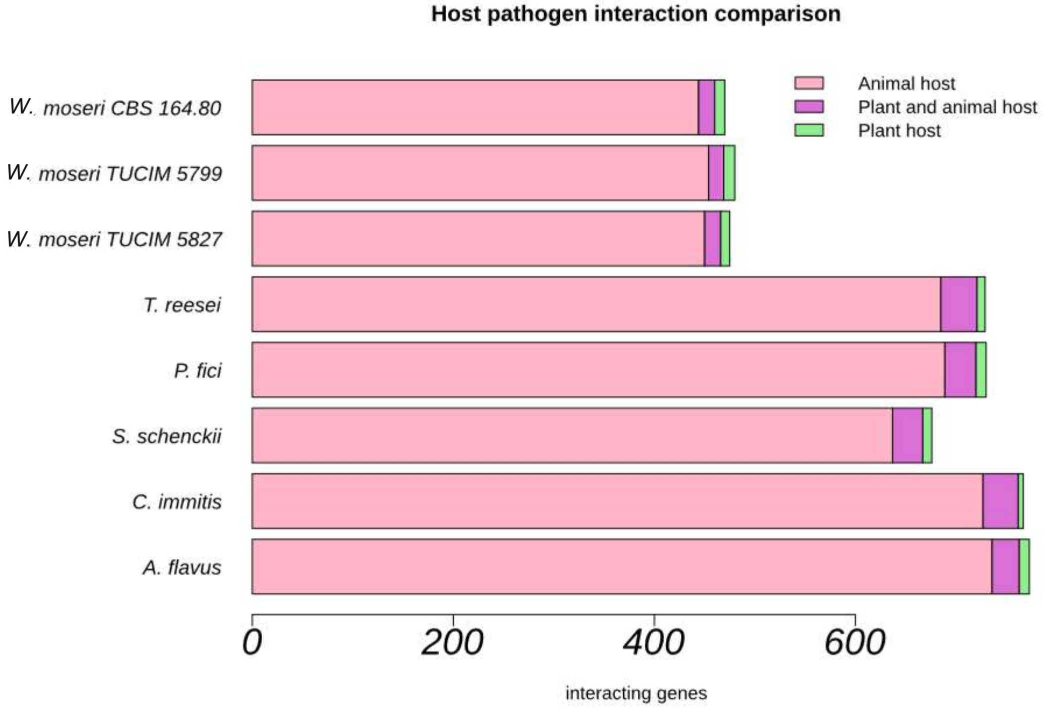


Fig. 6 Estimation of the pathogen host interaction potential. The stacked boxplot represents the numbers of putative gene products predicted to have associations with an animal and/or plant hosts for the proteomes of *W. moseri* CBS 164.80, *W. moseri* TUCIM 5827, *W. moseri* TUCIM 5799, *T. reesei*, *P. fici, S. schenkii*, *C. immitis* and *A. flavus*.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

JN550g9048) were annotated by PANNZER2 as an endo-1,4-beta-xylanase B and an endo-1,4-beta-xylanase G, respectively. They were predicted as part of the GH11 family by HMMER and Hotpep and to have a secretory signal according to a SignalP v5.0 prediction (57). A sequence similarity analysis using BLAST suggests that these two gene products might be undescribed xylanases since their sequence identity is in both cases below 67% for the best hit (JN550g516 66.66% with a xylanase from *Alternaria sp.* MG1 and JN550g9048 63.93% with a xylanase from *Verticillium dahliae*).

Further, we found a single potential TccC-type III insecticidal toxin in each strain (JN550g7831, JX266g13216, JX265g8960).

Organic acid production potential

To determine the potential for organic acid production in *W. moseri,* we searched for homologs of enzymes recognized to be involved in the organic acid production of filamentous fungi, especially in the organic acid producing *Aspergillus* spp. (58, 59). We were able to detect 27 of 31 enzyme types involved in organic acid production and three of six transporter types (Additional file 27). These enzymes are in theory sufficient for the production of fumaric acid, gluconic acid, succinic acid and malic acid (58, 59). We could not detect a glucokinase (EC:2.7.1.2), an oxaloacetase (EC:3.7.1.1), a trehalose phosphatase (EC:3.1.3.12), a β-fructofuranosidase (EC:3.2.1.26), a aconitate decarboxylase (EC:4.1.1.6), a citrate/malate antiporter, a citrate transporter, or a fructose transporter, which are necessary for the production and/or the secretion of trehalose, itaconic acid, oxalic acid, and citric acid.

Secondary Metabolism

To assess the SM production potential of *W. moseri*, we mined the genomes of the strains with antiSMASH 6 (60) and predicted a total of 63, 65, and 63 SM BGCs, in CBS 164.80, TUCIM 5827, and TUCIM 5799, respectively (Table 8). For comparison, we mined the genomes of fungi with a high SM production profile, i.e. *A. flavus*, *Penicillium chrysogenum*, *Tolypocladium inflatum*, *Fusarium oxysporum*, and *P. fici* and the industrial workhorse *T. reesei*, which has a relative small secondary metabolism (Table 8).

Most of the predicted BGCs of *W. moseri* did not have high similarities to previously described SM BGCs, with a few exceptions (Additional file 28). Region 4.3 of *W. moseri* CBS 164.80 (scaffold 4, JAFEVA010000004.1 317530 – 364110 nt; corresponding to TUCIM 5728 region 12.1, scaffold 12 JAFIMQ010000012.1 221463-

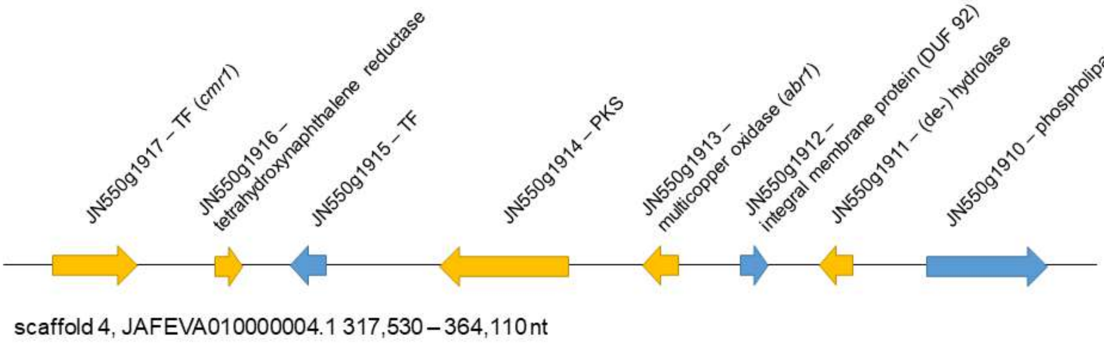

JN550g1917 – TF (cmr1)
JN550g1916 – tetrahydroxynaphthalene reductase
JN550g1915 – TF
JN550g1914 – PKS
JN550g1913 – multicopper oxidase (abr1)
JN550g1912 – integral membrane protein (DUF 92)
JN550g1911 – (de-) hydrolase
JN550g1910 – phospholipase D
scaffold 4, JAFEVA010000004.1 317,530 – 364,110 nt

Fig. 7 The potential melanin cluster of *W. moseri* CBS 164.80.
Yellow arrows indicate predicted core biosynthetic genes (by manual comparison to previously described DHN-melanin BGCs and biosynthetic pathways). TF, transcription factor; PKS, polyketide synthase.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

TU Bibliothek WIEN Your knowledge hub

268043 nt; and to TUCIM 5799 region 8.1, scaffold 8 JAFIMR010000008.1 317552-364132 nt) appears to be a potential dihydroxynaphthalene (DHN)-melanin BGC (Fig. 7, Additional file 28). Based on the knowledge about DHN-melanin biosynthesis in other fungi (61), we speculate that the first step in *W. moseri* is the formation of the 1,3,6,8-tetrahydroxynaphthalene by the non-reducing PKS (JN550g1914), followed by the formation of scytalone by the dehydratase JN550g1911 and the reduction by a tetrahydroxynaphthalene reductase (JN550g1916) to produce vermelone (62). This may be followed by a dehydration by the multicopper oxidase (JN550g1913) to produce 1,8-DHN the direct precursor of DHN-melanin, similar to the biosynthesis in *Aspergillus fumigatus* (63). 1,8-DHN needs to be polymerized by a laccase to yield DHN-melanin (61). There is no laccase present in direct proximity of the DHN-melanin BGC, but we found a laccase *abr2*-like enzyme (JN550g4679) elsewhere in the genome of *W. moseri*. Such partial or even total de-clustering of the enzymes involved in the biosynthesis of DHN-melanin is commonly occurring in fungi (61). The transcription factor JN550g1917 is similar to the melanin-specific transcriptional activator Cmr1 found in several fungi (64-66). The putative DHN-melanin cluster of *W. moseri* is highly likely to produce DHN-melanin and this could be responsible for the dark spore pigmentation (61), as melanins are usually polymerized and found in the fungal cell wall (67, 68). Notably, DHN-melanins may also be secreted (62). The observed brownish coloration of the agar plates (Fig 1) might be derivates or intermediates of the assumed DHN-melanin

Region 17.2 of *W. moseri* CBS 164.80 (scaffold 17 JAFEVA010000017.1 234418 – 293543 nt; TUCIM 5728: region 6.1, scaffold 6 JAFIMQ010000006.1 595090-654213 nt; TUCIM 5799: region 5.1, scaffold 5 JAFIMR010000005.1 1645008-1704131 nt) contains all biosynthetic genes for the production of a fusaric acid-like compound (Fig. 8). Nine of the 18 genes in this BGC share high homologies to the genes of the fusaric acid cluster of *F. verticillioides* strain 7600 (MIBiG: BGC0001190) (69). The two CYP P450 enzymes (JN550g5522 and JN550g5530) that do not share any homologies with the fusaric acid cluster of *F. verticillioides*, might be

Table 8 BGCs found with antiSMASH 6 in the genomes of the *W. moseri* strains and comparison genomes. Sid: Siderophore cluster; Ter: Terpene cluster; Ind: Indole cluster; Beta: Betalactone cluster; RiPP: fungal RiPP cluster

Genome	**Nrps**	**Nrps-like**	**T1pks**	**T3pks**	**Sid**	**Ter**	**Beta**	**Ind**	**mix**	**RiPP**
W. moseri CBS 164.80	9	8	21	1	1	11	1	3	7	1
W. moseri TUCIM 5827	10	8	21	1	1	11	1	4	7	1
W. moseri TUCIM 5799	9	8	19	1	1	11	1	3	9	1
P. fici	11	13	25	1	0	9	1	4	7	2
A. flavus	10	10	8	1	1	7	1	3	5	1
P. chrysogenum	8	10	15	0	0	3	2	0	4	0
T. inflatum	9	6	15	0	0	3	1	0	11	0
F. oxysporum	8	11	7	1	0	8	1	2	7	0
T. reesei	6	5	9	0	0	8	0	0	4	0

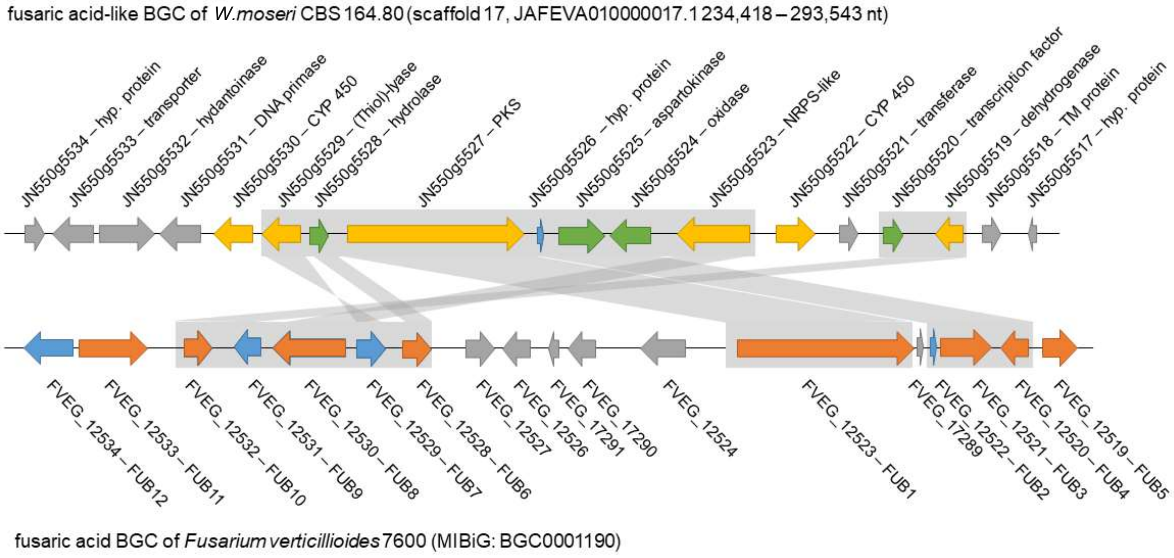


Fig. 8 Comparison of the fusaric acid-like BGC of *W. moseri* CBS 164.80 with the fusaric acid BGC of *Fusarium verticillioides*. The grey boxes and links indicate homologies between the two compared BGCs. Grey arrows represent genes that are considered as gap genes. Within the *W. moseri* BGC, yellow arrows indicate predicted core biosynthetic genes (by antiSMASH), green arrows indicated genes whose homologs are essential for fusaric acid production in at least one *Fusarium* species, and blue arrows indicate genes whose homologs are not essential but still conserved within different fusaric acid BGCs. Within the *F. verticillioides* BGC, orange arrows indicate genes essential for fusaric acid production, and blue boxes indicate conserved genes, which are not essential but may be important for full fusaric acid yield, according to (2). CYP 450, Cytochrome P450 monooxygenase; PKS, polyketide synthase; NRPS, non-ribosomal peptide synthetase.

involved in the detoxification of fusaric acid as indicated by Studt *et al*. (70) (Fig. 8).

The antiSMASH 6 analysis detected a single putative RiPP cluster in all three strains (*W. moseri* CBS 164.80: scaffold 76, JAFEVA010000076.1 94224 – 141768 nt; *W. moseri* TUCIM 5827: scaffold 37, JAFIMQ010000037.1 156492 – 204007 nt; *W. moseri* TUCIM 5799: scaffold 44, JAFIMR010000044.1 3419 – 50968 nt). The same cluster was identified as a potential fungal RiPP cluster by our manual method (71). The predicted RiPP BGC did not display significant similarities to known clusters from other fungi, but they were similar in gene composition and cluster structure among the three strains. The putative fungal RiPP cluster from *W. moseri* CBS 164.80 (scaffold 76, JAFEVA010000076.1 94224 – 141768 nt) is depicted in figure 9. The cluster contains 14 predicted genes, the gene JN550g12006 was predicted by our manual RiPP-precursor search as the only potential RiPP-precursor within the cluster (Additional file 28). The RiPP BGC prediction of antiSMASH 6.0 is based on a similarity to the ustiloxin BGC (60, 72). These BGCs normally contain at least one DUF3328 domain protein, (e.g. *ustYa*, *ustYb* in (73)), and the RiPP-precursor contain a repetitive core peptide structure. Interestingly, in the putative RiPP BGC of *W. moseri*, no DUF3328 protein was detected and the potential RiPP-precursor peptide does not share the repetitive core peptide structure. These findings indicate that *W. moseri* produces a potential novel fungal RiPP class.

We were able to show that our previously described method supports the

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

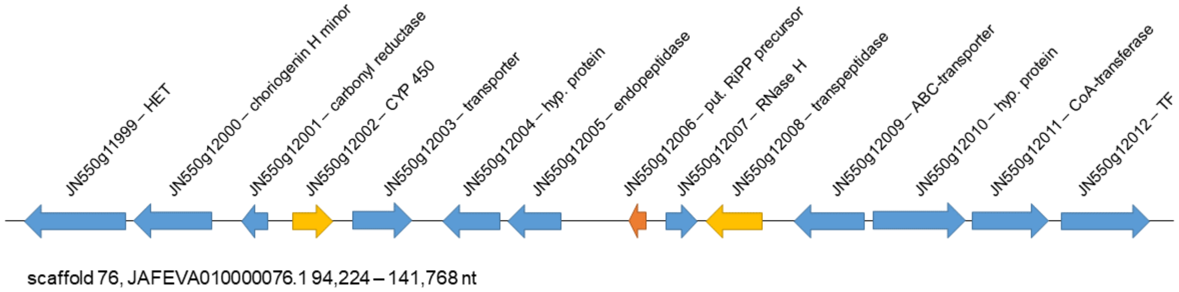

JN550g11999 – HET
JN550g12000 – choriogenin H minor
JN550g12001 – carbonyl reductase
JN550g12002 – CYP 450
JN550g12003 – transporter
JN550g12004 – hyp. protein
JN550g12005 – endopeptidase
JN550g12006 – put. RiPP precursor
JN550g12007 – RNase H
JN550g12008 – transpeptidase
JN550g12009 – ABC-transporter
JN550g12010 – hyp. protein
JN550g12011 – CoA-transferase
JN550g12012 – TF
scaffold 76, JAFEVA010000076.1 94,224 – 141,768 nt

Fig. 9 A putative RiPP BGC in *W. moseri* CBS 164.80. Yellow arrows indicate predicted core biosynthetic genes (by antiSMASH). The putative RiPP precursor is indicated in orange. HET, Heterokaryon incompatibility protein; CYP 450, cytochrome P450; TF, transcription factor.

detection of the potential RiPP-precursor peptide within novel predicted fungal RiPP clusters (71). Further, the high potential for novel SM discovery in *W. moseri* is reflected in the low amount of shared classical BGCs with the known SM producers. Furthermore, the similarities to *P. fici*, a high potential SM producer (74), makes this strain a likely candidate for novel SM discovery (1). This confirms the suggested slumbering SM potential of historic fungal strain collections (22, 75).

DISCUSSION

Plant-associated fungi are considered to be among the most prolific SM producers (22). Consequently, many new fungi have been isolated from the phyllosphere with the aim to find novel SM. In the recent years, the search area was broadened towards more extreme environments such as marine or arctic habitats (75). These efforts and the further sampling from host associated fungi have led to the discovery of manifold diverse species, which were described and classified, but remained understudied in respect of their secondary metabolism due to the sheer number of new isolates (22). *W. moseri* seems to be an example in this regard. Sandoval-Denis *et al.* showed that the ex-isotype culture of *W. moseri* clustered among the Xylariales and appeared to be related to members of the *Amphisphaeriaceae* and *Clypeosphaeriaceae* (5). A recent preprint by Samarakoon et al. places *W. moseri* based on a multi locus ML tree (using ITS, LSU, rpb2, tub2 and tef1 genes) within the family *Amphisphaeriaceae* and next to *Beltraniaceae* (6). Considering the limited number of available genomes from these phylogenetic families and based on our in depth phylogenetic analysis (3041 single-copy orthologues), we suggest placing this fungus in the family *Sporocadaceae* within the Xylariales. This topic will remain open for discussion.

The analysis of the primary carbon metabolism, characterized by the usage of specialized carbohydrate active enzymes (CAZymes) revealed *W. moseri* to exploit a diverse range of complex and simple carbon sources for growth (36) suggesting a oligotrophic lifestyle (44). The presence of a potential TccC-type III insecticidal toxin, paired with the high number of proteases and secreted chitinases suggests a potential insecticidal ability of *W. moseri* (76-78).

CONCLUSION

With this study, we present and discuss the high-quality genomes of three strains of *Wardomyces moseri*. This fungus had previously been isolated from Colombia and described as an unusual *Wardomyces* species (4). In this study, we could isolate two new strains from Brunei and demonstrate that the three strains are highly similar despite the temporal and geographical distances, indicating that *W. moseri* is a rare but cosmopolitan species. Based

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

on a detailed phylogenetic analysis, we suggest placing this fungus in the family *Sporocadaceae* within the Xylariales. The comparative genomic analysis revealed a rather average fungal genome in respect to size and gene composition with two outstanding features. *W. moseri* appears to possess a very low pathogenicity potential, while simultaneously a large secondary metabolite potential. Out of the large number of putative SM BGCs only a small proportion were similar to already known and described BGCs; we found a DHN-melanin BGC and a fusaric acid-like BGC. These findings suggest a great potential for the discovery of novel SMs in *W. moseri*.

METHODS

Sampling and strain purification

The epiphytic fungi TUCIM 5827 and TUCIM 5799 were isolated from the adaxial surface of the healthy leaf of *Shorea johorensis* (Dipterocarpaceae, Malvales; DNA BarCode maturase K (matK) deposited in NCBI GenBank MF993320.1, Laciny et al., 2008) sampled in the high canopy (40 – 60 m above ground) of the lowland tropical rain forest surrounding the Kuala Belalong Field Studies Center (KBFSC, 4°32'48.2"N 115°09'27.9"E) located in the Temburong District of Brunei Darussalam (Borneo). For this purpose, the adaxial surface of a freshly sampled leaf was scratched by the sterile electric toothbrush (2 min) in 25 ml of sterile water supplemented with Tween-20 (0.01%) in large sterile Petri plate (20 cm in diameter). The resulting suspension was collected in 50 ml falcons and centrifuged at 4ºC for 15 min at 14 000 rpm. The resulting pellet was resuspended in 4 ml of sterile water and used for serial dilution and plating on potato dextrose agar (PDA, Carl Roth) supplemented with 200 mg/l of chloramphenicol. Young single spore fungal colonies were detected with the use of a stereo microscope and aseptically transferred to fresh PDA plates and cultivated at 28°C in darkness. Agar plugs with pure mature cultures were preserved in 40% glycerol and stored at -80°C in TU Wien Collection of Industrial Microorganisms (TUCIM).

The ITS1 5.8S ITS2 regions of the TUCIM isolates were amplified by PCR using the primer pair ITS1F (5’-> 3’; CTTGGTCATTTAGAGGAAGTAA) and ITS4 (5’-> 3’; TCCTCCGCTTATTGATATGC) (79). The resulting ITS sequences from each strain were classified by performing a sequence similarity analysis using BLAST (non-redundant nucleotide database) (80).

To enable the growth on agar plates from spores as starting point, *W. moseri* CBS 164.80, TUCIM 5827, and TUCIM 5799 were cultivated and maintained on agar plates containing 30g/l oatmeal (S-Budget, SPAR Österreichische Warendhandels-AG; shredded to an approximate granular size of 0.25 mm). Spores were harvested from these oatmeal agar plates. For morphological comparison of the three strains, 5 µl of spore solution with OD_{600} 3 were applied to the middle of agar plates containing 20g/l malt extract. The malt extract agar plates were incubated at 28 °C for 15 days, after which pictures were taken.

The scanning electron microscopy (SEM) of the spores from *W. moseri* CBS 164.80, TUCIM 5827, and TUCIM 5799, respectively, was performed using COXEM EM-30AX PLUS with a SPT-20 Sputter. For sample preparation, spores of the respective strain were softly scratched off an overgrown oatmeal-plate with a cotton swab. The spores were carefully distributed over a silver stripe, which was adhered to the stage of the device. Further proceedings were done according to the manufacturer’s instructions. The pictures taken with the SEM were processed using the software Nanostation 3.0.4..

DNA extraction and library preparation

The type-strain *W. moseri* CBS 164.80 was ordered from the CBS Westerdijk Fungal Biodiversity Institute. The type strain and both TUCIM strains were cultivated in 250 ml Erlenmeyer flasks with 75 ml liquid malt extract (MEX) at 28°C and shaken at 180 rpm for 10 days in triplicates. The biomass was filtered through miracloth (EMD Millipore Corp., Burlington, MA, USA), pooled, placed in sterile 50 ml Cellstar tubes (Greiner Bio-One,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Kremsmünster, Austria), frozen in liquid nitrogen, lyophilized and stored at -20°C. For the DNA extraction, first the lyophilized biomass was disrupted using a Fast-Prep-24 (MP Biomedicals, Santa Ana/, CA, USA) with 0.37 g of small glass beads (0.1 mm diameter), 0.25 g of medium glass beads (1 mm diameter), and a single large glass bead (5 mm diameter) at 6 m/s for 30 sec. After the addition of 1 ml CTAB buffer (100 mM Tris.Cl, 20 mM EDTA, 1.4 M NaCl, 2 % (w/v) CTAB, pH = 8.0) and 4 µl β-mercaptoethanol, the samples were subjected to two further disruption treatments on the Fast-Prep-24 at 5 m/s for 30 sec and then incubated at 65°C for 20 min. The supernatant was extracted with phenol, chloroform, isoamylalcohol (25 : 24 : 1) followed by a chloroform extraction. The supernatant was treated with RNase A (Thermo Fisher Scientific, Inc., Waltham, MA, USA) according to the manufacturer's instructions. Finally, the DNA was precipitated with ethanol and dissolved in 10 mM Tris.Cl (pH = 8.0)

50μl of DNA, from each strain, were placed in 1.5ml TPX microtubes for Diagenode Bioruptor® Pico (Diagenode s.a., Liège, Belgium) and sonicated with the settings set to high and three cycles of 15 sec "on" and 60 sec "off". The sheared DNA was purified using "PCR purification kit #k0701 #k0702" (Thermo Fisher Scientific, Inc., Waltham, MA, USA) and then double side size selected with "NEBNext Ultra™ sample purification beads" (New England Biolabs, Ipswich, MA, USA) for 800 bp fragments. The library preparation was performed following the protocol of "NEBNext® Ultra™ II DNA Library Kit with Purification Beads" and "NEBNext® Multiplex Oligos for Illumina (Index Primer Set1 and Set2)" (New England Biolabs, Ipswich, MA, USA). The average size in bp of the library was measured with the fragment analyzer from Advanced Analytical Technologies using the Agilent dsDNA 915 Reagent Kit (35-5000bp) and analyzed with the PRO size software (Agilent Technologies, Santa Clara, California, USA). The exact DNA concentrations were measured with an "invitrogen™ Qubit™ fluorometer" in ng/μl (Thermo Fisher Scientific, Inc., Waltham, MA, USA) using a "Quant-iT™ dsDNA BR Assay" kit (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The libraries were diluted to the appropriate 4nM concentration for sequencing. The 4nM library was stored at -20°C.

Sequencing

The sequencing of the *W. moseri* library was performed on a Illumina MiSeq using two V3 Reagent Kit (600 cycles) and one V2 Nano Reagent Kit (500cycles) following the standard protocol of Illumina sequencing protocol without adding PhiX control to the runs (Illumina, San Diego, California, USA). The quality profiles and all further figures, if not specified otherwise, were visualized in R (81).

Extracting the mitochondrial genome and cleaning the raw reads

First a preliminary assembly was performed using SPAdes v3.13.1 (82) with default parameters for each strain separately. Mitochondrial sequences were identified in each strain by performing a sequence similarity analysis using BLAST (non-redundant nucleotide database) (80). Contigs ranging from 500 to 1000 bp were then used as seed input for NOVOplasty v3.7 (83) to extract the whole circularized mitochondrial genome of *W. moseri* CBS 164.80, TUCIM 5827 and TUCIM 5799. This was performed in an iterative manner. The mitochondrial genomes were visualized with CGViewer (84). The mitochondrial genomes were annotated with the automated MITOS2 web pipeline (85).

Using the mitochondrial genomes of the strains as input an index was built with bowtie v1.2.2 (86), respectively, and the mitochondrial flagged reads were extracted using --un option from each raw reads file. The clean raw reads were then re-paired with Fastq-pair (87) to use paired end read assemblers.

Whole - genome assembly

For each strain respectively, the raw cleaned paired end reads were quality trimmed using Trimmomatic (88) in the command line

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

and specifying PE for paired end reads and ILLUMINACLIP:Adapter-PE.fa:2:30:10:2:keepBothReads LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 to ensure high quality adapter-free reads. Then the cleaned raw reads were assembled using SPAdes v3.13.1 (82), for each strain separately. Furthermore, the high quality trimmed cleaned paired end reads were used for scaffolding with SSPACE-Standard v3.0 (89) in a iterative manner with following command line options -x 1 -m 50 -o 20 -k 8 -a 0.70 -n 30 -z 150 –b and –k 6. Ns introduced during the assemblies and the scaffolding, so called gaps, were closed with GapFiller v1-10 (90) using following commands -m 30 -o 6 -r 0.7 -n 10 -d 50 -t 10 -g 0 -i 5 -b.

The assemblies were further improved by using Pilon v1.21 (91) iteratively. We first indexed the assemblies with bwa (92), SAMtools v1.7 (93) and picard (94). The high quality trimmed cleaned paired end reads were mapped to the matching indexed assemblies of the individual *W. moseri* strains with bwa. The reads were mapped and combined in one step. Next, we sorted and created bam files from the sam files using SAMtools. Together with the paired sequencing reads, these were used as input for Pilon to iteratively improve each genome.

Gene prediction

Transfer RNA genes were detected using tRNAscan-SE v1.3.1 (33). Augustus v3.3.2 (35) was trained with the genome of *P. fici* (assembly PFICI; BioSample accession: SAMN02369365) following the protocol by Hoff & Stanke (95). The genomes were masked using RepeatMasker v4.0.9 (96) to identify repetitive elements.

Augustus was run with the species option set to pestalotiopsis_fici on the masked genome assemblies. The genomes and the gene sets were evaluated using Quast v5.0.2 (97, 98). Quast v5.0.2 includes a benchmarking with Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2, this was performed with the eukaryote dataset of 303 BUSCOs from 100 species (32). We further evaluated the gene predictions by aligning the amino acid sequences using Blastp v2.9.0+ (80) against the UniProt database(99).

DNA sequence-based phylogenetic placement

To get first indications of the potential phylogenetic placement of *W. moseri,* the complete genes of actin, calmodulin, the ITS1 5.8S ITS2 region, the SSU, the LSU, the RPB2 gene, *tef1* gene and the beta tubulin gene were extracted from the assemblies of the three *W. moseri* strains using the blastdbcmd software (4, 80, 100). To determine accurately the phylogenetic placement of *W.* moseri, we performed two high accuracy orthogroup inferences to provide phylogenetic inference using OrthoFinder (3) based on the predicted proteomes of type-strain *W. moseri* CBS 164.80, *W. moseri* TUCIM 5827, *W. moseri* TUCIM 5799, *T. reesei* QM6a (assembly v2.0 BioSample accession: SAMN02746107), *C. immitis* (assembly ASM14933v2, BioSample accession: SAMN02786853), *A. flavus* NRRL3357 (assembly JCVI-afl1-v2.0; BioSample accession: SAMN05591370), *P. fici* (assembly PFICI; BioSample accession: SAMN02369365), *S. schenckii* (assembly S_schenckii_v1; BioSample accession: SAMN07585147), *E. lata* URCEL1 (assembly URCEL1V03; BioSample accession: SAMN01906717), *P. roqueforti* LCP96 04111 (assembly ASM1553377v1; BioSample accession: SAMN14669941) and *F. fujikuroi* IMI 58589 (assembly Fusarium_fujikuroi_IMI58289_V2; BioSample accession: SAMEA3724789) and *Neopestalotiopsis clavispora* (assembly ASM1462143v1; BioSample accession: SAMN14260619), *Truncatella angustata* (assembly Truan1; BioSample accession: SAMN08150287), *Xylariales* sp. AK1849 (JGI assembly Xylariales sp. AK1849 v1.0), *Pseudomassariella vexata* CBS 129021 (assembly Pseve2; BioSample accession: SAMN05421895), *Khuskia oryzae* (JGI assembly Khuskia oryzae ATCC 28132 v1.0), *Apiospora montagnei* (JGI assembly Apiospora montagnei NRRL 25634 v1.0), *Phialemoniopsis curvata* (assembly

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

ASM435304v1; BioSample accession: SAMN11041535), *Monosporascus cannonballus* CBS 609.92 (assembly ASM415492v1; BioSample accession: SAMN09215312), *Daldinia childiae* (assembly Dalch_JS-1345; BioSample accession: SAMN12777271), *Hypoxylon* sp. EC38 (assembly HypEC38 v3.0; BioSample accession: SAMN01163462), *Xylaria flabelliformis* G536 (assembly ASM718279v1; BioSample accession: SAMN11912834), *Rosellinia necatrix* W97 (assembly Rnecatrix_2.0; BioSample accession: SAMD00023353) and *Microdochium bolleyi* J235TASD1 (assembly Microdochium bolleyi v1.0; BioSample accession: SAMN0486150). We verified the inferred phylogenetic trees using OrthoFinder by comparing them to the phylogenetic tree provided by the JGI (28). We compared the three *W. moseri* strains and calculated their ANI with fastANI (101).

Annotation

The gene sets were first annotated using Blastp against the UniProt protein database. Protein ANNotation with Z-scoRE (PANNZER2) (48) was used to provide both GO and free text DE producing an accurate functional annotation. CAZymes were annotated using the dbCAN2 (102) meta server by applying a HMMer (38) (Hidden Marcov model) search, a DIAMOND (39) search and a Hotpep (40) search and combining the three outputs. The dbCAN2 (102) server also includes a SignalP v5.0 prediction. (57) We searched the web version of HPIDB 3.0 (52) with the whole predicted proteome of the genome assemblies. Furthermore we performed a sequence similarity search against the MEROPS (47) database with Blastp v2.9.0+ (80). We performed a KEGG annotation for the complete predicted proteomes of the *W. moseri* strains using KofamKOALA (46).

Full genomes used for comparative analysis

The genomes of *A. flavus* NRRL3357 (assembly JCVI-afl1-v2.0; BioSample accession: SAMN05591370), *P. chrysogenum* (assembly ASM71027v1; BioSample accession: SAMN02742620), *T. inflatum* (assembly ASM394556v1; BioSample accession: SAMN08824660), *F. oxysporum* (assembly ASM14995v2; BioSample accession: SAMN02953675), *T. reesei* QM6a (assembly v2.0 BioSample accession: SAMN02746107) and *P. fici* (assembly PFICI; BioSample accession: SAMN02369365) were downloaded from the NCBI database and mined for secondary metabolite BGCs to be compared to the *W. moseri* genomes. To enable a comparative analysis of the pathogenic potential of *W. moseri* the proteomes of *A. flavus* NRRL3357, *P. fici*, *S. schenckii* (assembly S_schenckii_v1; BioSample accession: SAMN07585147), *T. reesei* QM6a (assembly v2.0 BioSample accession: SAMN02746107) and *C. immitis* (assembly ASM14933v2, BioSample accession: SAMN02786853) were downloaded from the NCBI database and evaluated using the web version of HPIDB 3.0. Further the genomes of *E. lata* URCEL1 (assembly URCEL1V03; BioSample accession: SAMN01906717), *P. roqueforti* LCP96 04111(assembly ASM1553377v1; BioSample accession: SAMN14669941) and *F. fujikuroi* IMI 58589 (assembly Fusarium_fujikuroi_IMI58289_V2; BioSample accession: SAMEA3724789) were downloaded to be included in the OrthoFinder analysis.

Genome mining

The command line version of antiSMASH v4.3.0 (103) and antiSMASH 6.0.0 web-version (60) (access 30.07.2021 and 08.10.2021) was used for genome mining for secondary metabolite biosynthetic gene clusters (BGC) with following specifications for the command line version: the taxon was specified with the option --taxon to be of fungal origin, --clusterblast, --smcogs, --full-hmmer, --asf, --subclusterblast and --knownclusterblast. Furthermore, the ClusterFinder algorithm was activated with the --inclusive option.

Further, we mined the genomes for putative ribosomally synthesized and post-translationally modified peptide (RiPP) using

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

the amino acid sequences of the genes classified as "other genes" in the BGCs as described by Vignolle *et al.* (71). The putative RiPP precursor peptides were further manually inspected using the PANNZER2 (48) annotation for all analyzed genomes (Additional files 29, 30, 31). Further, we performed a BiG-SCAPE (104) analysis with all BGCs predicted in the three *W. moseri* strains and the BGCs predicted for *P. fici*.

LIST OF ABBREVIATIONS

AA - redox enzymes with auxiliary activities
ANI – average nucleotide identity
BGC – biosynthetic gene cluster
CAZymes - Carbohydrate active enzymes
CE - carbohydrate esterases
CMA - cornmeal agar
CTAB – cetyl trimethylammonium bromid
DE - free text functional description
GH - Glycoside Hydrolases
GO - Gene ontology
GT - Glycosyltransferases
HPI - Host-pathogen protein–protein interaction
HMMER - Hidden Marcov model
LINE - long interspaced nuclear repeat
LTR – long terminal repeat
MEA - malt extract agar
MEX – malt extract
ML – maximum likelihood
NNI - Nearest-Neighbor-Interchange
NRPS – Non-ribosomal peptide synthetase
OA - oatmeal agar
PL - polysaccharide lyases
RiPPs – Ribosomally synthesized and post-translationally modified peptides
SEVAG – choloform/isoamylalcohol 24:1
SINE - short interspaced nuclear repeat
SM – secondary metabolite
SSCRP - small secreted cysteine rich proteins
T1pks – Type I polyketide synthase
T3pks – Type III polyketide synthase

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession PRJNA695409. The raw reads were uploaded to the Sequence Read Archive (SRA) under the accession SRR13570309, SRR13747339 and SRR13747338. The complete genomes were deposited at DDBJ/ENA/GenBank under the accessions JAFEVA000000000 (https://www.ncbi.nlm.nih.gov/Traces/wgs/JAFEVA01) for the CBS 164.80 strain, JAFIMQ000000000 (https://www.ncbi.nlm.nih.gov/Traces/wgs/JAFIMQ01) for the TUCIM 5827 strain and JAFIMR000000000 (https://www.ncbi.nlm.nih.gov/Traces/wgs/JAFIMR01) for the TUCIM 5799 strain. The versions described in this paper are version JAFEVA010000000.1, JAFIMQ010000000.1 and JAFIMR010000000.1. The complete mitochondrial genome of the strain CBS 164.80 was deposited with the GenBank accession no. MW554918 (https://www.ncbi.nlm.nih.gov/nuccore/MW554918). The complete mitochondrial genome of the strain TUCIM 5799 was deposited with the GenBank accession no. MW660809 (https://www.ncbi.nlm.nih.gov/nuccore/MW660809). The complete mitochondrial genome of the strain TUCIM 5827 was deposited with the GenBank accession no. MW660808 (https://www.ncbi.nlm.nih.gov/nuccore/MW660808). The datasets analyzed during the current study are publicly available in the NCBI National Center for Biotechnology Information repository (https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/). The datasets supporting the conclusions of this article are included within the article and its additional files.

Competing interests

The authors declare that they have no competing interests.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Funding
This study was supported by the Austrian Science Fund (FWF): P29556 given to RM, P34036 given to CD, and the PhD program TU Wien bioactive.

ACKNOWLEDGEMENTS

We kindly thank Gerd Mauschitz and Wolfgang Ipsmiller (both TU Wien) for providing access to the electron microscope.

ADDITIONAL FILES

Additional file 1. The results from the Sanger sequencing of the PCR amplified ITS1 5.8S ITS2 region. (.fasta)

Additional file 2. Genome assembly statistics report of *W. moseri* CBS 164.80. (.PDF)

Additional file 3. Genome assembly statistics report of *W. moseri* TUCIM 5827. (.PDF)

Additional file 4. Genome assembly statistics report of *W. moseri* TUCIM 5799. (.PDF)

Additional file 5. RepeatMasker analysis of repetitive elements in the genome of *W. moseri* CBS 164.80. (.txt)

Additional file 6. RepeatMasker analysis of repetitive elements in the genome of *W. moseri* TUCIM 5827. (.txt)

Additional file 7. RepeatMasker analysis of repetitive elements in the genome of *W. moseri* TUCIM 5799. (.txt)

Additional file 8. tRNA locations of *W. moseri* CBS 164.80. (.txt)

Additional file 9. tRNA locations of *W. moseri* TUCIM 5827. (.txt)

Additional file 10. tRNA locations of *W. moseri* TUCIM 5799. (.txt)

Additional file 11. Functional annotation analysis of the proteome of *W. moseri* CBS 164.80, performed with PANNZER2. (.txt)

Additional file 12. Functional annotation analysis of the proteome of *W. moseri* TUCIM 5827, performed with PANNZER2. (.txt)

Additional file 13. Functional annotation analysis of the proteome of *W. moseri* TUCIM 5799, performed with PANNZER2. (.txt)

Additional file 14. CAZyme analysis of the proteome of *W. moseri* CBS 164.80. (.xlsx)

Additional file 15. CAZyme analysis of the 1 proteome of *W. moseri* TUCIM 5827. (.xlsx)

Additional file 16. CAZyme analysis of the proteome of *W. moseri* TUCIM 5799. (.xlsx)

Additional file 17. KEGG annotation analysis of the proteome of *W. moseri* CBS 164.80, using KofamKOALA. (.txt)

Additional file 18. KEGG annotation analysis of the proteome of *W. moseri* TUCIM 5827, using KofamKOALA. (.txt)

Additional file 19. KEGG annotation analysis of the proteome of *W. moseri* TUCIM 5799, using KofamKOALA. (.txt)

Additional file 20. BLAST analysis of the proteome of *W. moseri* CBS 164.80 against the MEROPS database. (.xlsx)

Additional file 21. BLAST analysis of the proteome of *W. moseri* TUCIM 5827 against the MEROPS database. (.xlsx)

Additional file 22. BLAST analysis of the proteome of *W. moseri* TUCIM 5799 against the MEROPS database. (.xlsx)

Additional file 23. Transcription factors based on the KEGG annotation of *W. moseri* CBS 164.80. (.xlsx)

Additional file 24. Host pathogen interaction analysis results for the whole fungal proteome of *W. moseri* CBS 164.80. (.tsv)

Additional file 25. Host pathogen interaction analysis results for the whole fungal proteome of *W. moseri* TUCIM 5827. (.tsv)

Additional file 26. Host pathogen interaction analysis results for the whole fungal proteome of *W. moseri* TUCIM 5799. (.tsv)

Additional file 27 Genes possibly involved in organic acid production and secretion of *W. moseri* CBS 164.80, *W. moseri* TUCIM 5799 and *W. moseri* TUCIM 5827. (.xlsx)

Additional file 28. Similarities of predicted BGCs to described BGCs (MiBIG database), the annotations of the potential melanin BGC,

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

the potential fusaric acid BGC, and the potential RiPP BGC. (.xlsx)

Additional file 29. Table containing putative RiPP precursor annotation and manual refinement of *W. moseri* CBS 164.80. (.xlsx)

Additional file 30. Table containing putative RiPP precursor annotation and manual refinement of *W. moseri* TUCIM 5827. (.xlsx)

Additional file 31. Table containing putative RiPP precursor annotation and manual refinement of *W. moseri* TUCIM 5799. (.xlsx)

REFERENCES:

1. Wang X, Zhang X, Liu L, Xiang M, Wang W, Sun X, et al. Genomic and transcriptomic analysis of the endophytic fungus Pestalotiopsis fici reveals its lifestyle and high potential for synthesis of natural products. BMC Genomics. 2015;16:28.
2. Brown DW, Lee SH, Kim LH, Ryu JG, Lee S, Seo Y, et al. Identification of a 12-gene Fusaric Acid Biosynthetic Gene Cluster in Fusarium Species Through Comparative and Functional Genomics. Molecular plant-microbe interactions : MPMI. 2015;28(3):319-32.
3. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology. 2019;20(1):238.
4. Grams W. An unusual species of Wardomyces (Hyphomycetes). Beih Sydowia X. 1995:67-72.
5. Sandoval-Denis M, Guarro J, Cano-Lira JF, Sutton DA, Wiederhold NP, de Hoog GS, et al. Phylogeny and taxonomic revision of Microascaceae with emphasis on synnematous fungi. Stud Mycol. 2016;83:193-233.
6. Milan CS, Kevin DH, Sajeewa SNM, Marc S, Jones EBG, Itthayakorn P, et al. Taxonomy, Phylogeny, Molecular Dating and Ancestral State Reconstruction of Xylariomycetidae (Sordariomycetes). Fungal Diversity. 2021.
7. Franco MEE, Wisecaver JH, Arnold AE, Ju Y-M, Slot JC, Ahrendt S, et al. Secondary metabolism drives ecological breadth in the Xylariaceae. bioRxiv. 2021:2021.06.01.446356.
8. Becker K, Stadler M. Recent progress in biodiversity research on the Xylariales and their secondary metabolism. The Journal of Antibiotics. 2021;74(1):1-23.
9. Helaly SE, Thongbai B, Stadler M. Diversity of biologically active secondary metabolites from endophytic and saprotrophic fungi of the ascomycete order Xylariales. Natural product reports. 2018;35(9):992-1014.
10. Blanco-Ulate B, Rolshausen PE, Cantu D. Draft Genome Sequence of the Grapevine Dieback Fungus Eutypa lata UCR-EL1. Genome announcements. 2013;1(3).
11. Malik VS. Microbial secondary metabolism. Trends in Biochemical Sciences. 1980;5(3):68-72.
12. Keller NP, Turner G, Bennett JW. Fungal secondary metabolism — from biochemistry to genomics. Nature Reviews Microbiology. 2005;3(12):937-47.
13. van den Berg MA, Westerlaken I, Leeflang C, Kerkman R, Bovenberg RA. Functional characterization of the penicillin biosynthetic gene cluster of Penicillium chrysogenum Wisconsin54-1255. Fungal Genet Biol. 2007;44(9):830-44.
14. Weber G, Schörgendorfer K, Schneider-Scherzer E, Leitner E. The peptide synthetase catalyzing cyclosporine production in *Tolypocladium niveum* is encoded by a giant 45.8-kilobase open reading frame. Current Genetics. 1994;26:120-5.
15. Kensler TW, Roebuck BD, Wogan GN, Groopman JD. Aflatoxin: a 50-year odyssey of mechanistic and translational toxicology. Toxicol Sci. 2011;120 Suppl 1:S28-48.
16. Mulder KC, Mulinari F, Franco OL, Soares MS, Magalhaes BS, Parachin NS. Lovastatin production: From molecular basis to industrial process optimization. Biotechnol Adv. 2015;33(6 Pt 1):648-65.
17. Gomez BL, Nosanchuk JD. Melanin and fungi. Curr Opin Infect Dis. 2003;16(2):91-6.
18. Wheeler MH, Bell AA. Melanins and their importance in pathogenic fungi. Curr Top Med Mycol. 1988;2:338-87.
19. Luo S, Dong SH. Recent Advances in the Discovery and Biosynthetic Study of Eukaryotic RiPP Natural Products. Molecules. 2019;24(8).
20. Hoffmeister D, Keller NP. Natural products of filamentous fungi: enzymes, genes, and their regulation. Natural product reports. 2007;24(2):393-416.
21. Bassett EJ, Keith MS, Armelagos GJ, Martin DL, Villanueva AR. Tetracycline-labeled human bone from ancient Sudanese Nubia (A.D. 350). Science. 1980;209(4464):1532-4.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

22. Hyde KD, Xu J, Rapior S, Jeewon R, Lumyong S, Niego AGT, et al. The amazing potential of fungi: 50 ways we can exploit fungi industrially. Fungal Diversity. 2019;97(1):1-136.

23. Alberti F, Foster GD, Bailey AM. Natural products from filamentous fungi and production by heterologous expression. Applied microbiology and biotechnology. 2017;101(2):493-500.

24. van der Lee TAJ, Medema MH. Computational strategies for genome-based natural product discovery and engineering in fungi. Fungal Genet Biol. 2016;89:29-36.

25. Epstein SC, Charkoudian LK, Medema MH. A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences. Stand Genomic Sci. 2018;13:16.

26. de Vries RP, Riley R, Wiebenga A, Aguilar-Osorio G, Amillis S, Uchima CA, et al. Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus. Genome Biol. 2017;18(1):28.

27. Higginbotham SJ, Arnold AE, Ibañez A, Spadafora C, Coley PD, Kursar TA. Bioactivity of Fungal Endophytes as a Function of Endophyte Taxonomy and the Taxonomy and Distribution of Their Host Plants. PLOS ONE. 2013;8(9):e73192.

28. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 2014;42(Database issue):D26-31.

29. Franco MEE, Wisecaver JH, Arnold AE, Ju YM, Slot JC, Ahrendt S, et al. Ecological generalism drives hyperdiversity of secondary metabolite gene clusters in xylarialean endophytes. New Phytol. 2021.

30. Chen C, Li Q, Fu R, Wang J, Xiong C, Fan Z, et al. Characterization of the mitochondrial genome of the pathogenic fungus Scytalidium auriculariicola (Leotiomycetes) and insights into its phylogenetics. Scientific Reports. 2019;9(1):17447.

31. Kang X, Hu L, Shen P, Li R, Liu D. SMRT Sequencing Revealed Mitogenome Characteristics and Mitogenome-Wide DNA Modification Pattern in Ophiocordyceps sinensis. Front Microbiol. 2017;8:1422.

32. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2.

33. Lowe T, Eddy S. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research. 1997;25(5):955-64.

34. Kanhayuwa L, Coutts RH. Short Interspersed Nuclear Element (SINE) Sequences in the Genome of the Human Pathogenic Fungus Aspergillus fumigatus Af293. PLoS One. 2016;11(10):e0163215.

35. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33(Web Server issue):W465-7.

36. Khosravi C, Benocci T, Battaglia E, Benoit I, de Vries RP. Sugar catabolism in Aspergillus and other fungi related to the utilization of plant biomass. Adv Appl Microbiol. 2015;90:1-28.

37. de Vries RP, Visser J. *Aspergillus* enzymes involved in degradation of plant cell wall polysaccharides. Microbiol Mol Biol Rev. 2001;65(4):497-522.

38. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(Web Server issue):W29-37.

39. Buchfink B, Xie C, Huson D. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 2015;12(1):59-63.

40. Busk PK, Pilgaard B, Lezyk MJ, Meyer AS, Lange L. Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. BMC Bioinformatics. 2017;18(1):214.

41. Lairson LL, Henrissat B, Davies GJ, Withers SG. Glycosyltransferases: structures, functions, and mechanisms. Annu Rev Biochem. 2008;77:521-55.

42. Lee C, Kim S, Li W, Bang S, Lee H, Lee HJ, et al. Bioactive secondary metabolites produced by an endophytic fungus Gaeumannomyces sp. JS0464 from a maritime halophyte Phragmites communis. J Antibiot (Tokyo). 2017;70(6):737-42.

43. Zhao Z, Liu H, Wang C, Xu JR. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics. 2013;14:274.

44. Ho A, Di Lonardo DP, Bodelier PLE. Revisiting life strategy concepts in environmental microbial ecology. FEMS Microbiology Ecology. 2017;93(3).

45. Kumar A, Chandra R. Ligninolytic enzymes and its mechanisms for degradation of lignocellulosic waste in environment. Heliyon. 2020;6(2):e03170-e.

46. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020;36(7):2251-2.

47. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res. 2018;46(D1):D624-D32.

48. Toronen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids Res. 2018;46(W1):W84-W8.

49. Daly P, Cai F, Kubicek CP, Jiang S, Grujic M, Rahimi MJ, et al. From lignocellulose to plastics: Knowledge transfer on the degradation approaches by fungi. Biotechnol Adv. 2021;50:107770.

50. Leiter É, Emri T, Pákozdi K, Hornok L, Pócsi I. The impact of bZIP Atf1ortholog global regulators in fungi. Applied microbiology and biotechnology. 2021;105(14-15):5769-83.

51. Myung K, Zitomer NC, Duvall M, Glenn AE, Riley RT, Calvo AM. The conserved global regulator VeA is necessary for symptom production and mycotoxin synthesis in maize seedlings by Fusarium verticillioides. Plant Pathol. 2012;61(1):152-60.

52. Ammari MG, Gresham CR, McCarthy FM, Nanduri B. HPIDB 2.0: a curated database for host-pathogen interactions. Database (Oxford). 2016;2016.

53. Nevalainen H, Suominen P, Taimisto K. On the safety of Trichoderma reesei. Journal of Biotechnology. 1994;37:193-200.

54. Dixon DM. Coccidioides immitis as a Select Agent of bioterrorism. Journal of Applied Microbiology. 2001;91:602-5.

55. Barros MB, de Almeida Paes R, Schubach AO. Sporothrix schenckii and Sporotrichosis. Clin Microbiol Rev. 2011;24(4):633-54.

56. Hedayati MT, Pasqualotto AC, Warn PA, Bowyer P, Denning DW. Aspergillus flavus: human pathogen, allergen and mycotoxin producer. Microbiology. 2007;153(Pt 6):1677-92.

57. Almagro Armenteros JJ, Tsirigos KD, Sonderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37(4):420-3.

58. Kubicek CP, Punt P, Visser J. Production of Organic Acids by Filamentous Fungi. In: Hofrichter M, editor. Industrial Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 215-34.

59. Liaud N, Giniés C, Navarro D, Fabre N, Crapart S, Gimbert IH, et al. Exploring fungal biodiversity: organic acid production by 66 strains of filamentous fungi. Fungal Biology and Biotechnology. 2014;1(1):1.

60. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema Marnix H, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research. 2021;49(W1):W29-W35.

61. Eisenman HC, Casadevall A. Synthesis and assembly of fungal melanin. Applied microbiology and biotechnology. 2012;93(3):931-40.

62. Ebert MK, Spanner RE, de Jonge R, Smith DJ, Holthusen J, Secor GA, et al. Gene cluster conservation identifies melanin and perylenequinone biosynthesis pathways in multiple plant pathogenic fungi. Environmental microbiology. 2019;21(3):913-27.

63. Tsai H-F, Wheeler Michael H, Chang Yun C, Kwon-Chung KJ. A Developmentally Regulated Gene Cluster Involved in Conidial Pigment Biosynthesis in Aspergillus fumigatus. Journal of Bacteriology. 1999;181(20):6469-77.

64. Tsuji G, Kenmochi Y, Takano Y, Sweigard J, Farrall L, Furusawa I, et al. Novel fungal transcriptional activators, Cmr1p of Colletotrichum lagenarium and pig1p of Magnaporthe grisea, contain Cys2His2 zinc finger and Zn(II)2Cys6 binuclear cluster DNA-binding motifs and regulate transcription of melanin biosynthesis genes in a developmentally specific manner. Mol Microbiol. 2000;38(5):940-54.

65. Eliahu N, Igbaria A, Rose MS, Horwitz BA, Lev S. Melanin biosynthesis in the maize pathogen Cochliobolus heterostrophus depends on two mitogen-activated protein kinases, Chk1 and Mps1, and the transcription factor Cmr1. Eukaryot Cell. 2007;6(3):421-9.

66. Jiang H, Chi Z, Liu GL, Hu Z, Zhao SZ, Chi ZM. Melanin biosynthesis in the desert-derived Aureobasidium melanogenum XJ5-1 is controlled mainly by the CWI signal pathway via a transcriptional activator Cmr1. Curr Genet. 2020;66(1):173-85.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

67. Belozerskaya TA, Gessler NN, Aver'yanov AA, editors. Melanin Pigments of Fungi. In: Mérillon JM., Ramawat K. (eds) Fungal Metabolites. y. Chem: Springer; 2017.
68. Nosanchuk JD, Stark RE, Casadevall A. Fungal Melanin: What do We Know About Structure? Front Microbiol. 2015;6:1463.
69. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research. 2019;48(D1):D454-D8.
70. Studt L, Janevska S, Niehaus EM, Burkhardt I, Arndt B, Sieber CM, et al. Two separate key enzymes and two pathway-specific transcription factors are involved in fusaric acid biosynthesis in Fusarium fujikuroi. Environ Microbiol. 2016;18(3):936-56.
71. Vignolle GA, Mach RL, Mach-Aigner AR, Derntl C. Novel approach in whole genome mining and transcriptome analysis reveal conserved RiPPs in Trichoderma spp. BMC Genomics. 2020;21(1):258.
72. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47(W1):W81-w7.
73. Umemura M, Nagano N, Koike H, Kawano J, Ishii T, Miyamura Y, et al. Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in Aspergillus flavus. Fungal Genet Biol. 2014;68:23-30.
74. Wang X, Zhang X, Liu L, Xiang M, Wang W, Sun X, et al. Genomic and transcriptomic analysis of the endophytic fungus Pestalotiopsis fici reveals its lifestyle and high potential for synthesis of natural products. BMC Genomics. 2015;16(1):28.
75. Keller NP. Fungal secondary metabolism: regulation, function and drug discovery. Nat Rev Microbiol. 2019;17(3):167-80.
76. Charnley AK, St. Leger RJ. The Role of Cuticle-Degrading Enzymes in Fungal Pathogenesis in Insects. In: Cole GT, Hoch HC, editors. The Fungal Spore and Disease Initiation in Plants and Animals. Boston, MA: Springer US; 1991. p. 267-86.
77. Liu JR, Lin YD, Chang ST, Zeng YF, Wang SL. Molecular cloning and characterization of an insecticidal toxin from Pseudomonas taiwanensis. J Agric Food Chem. 2010;58(23):12343-9.
78. Yang G, Waterfield NR. The role of TcdB and TccC subunits in secretion of the Photorhabdus Tcd toxin complex. PLoS Pathog. 2013;9(10):e1003644-e.
79. Raja HA, Miller AN, Pearce CJ, Oberlies NH. Fungal Identification Using Molecular Tools: A Primer for the Natural Products Research Community. Journal of Natural Products. 2017;80(3):756-70.
80. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421-.
81. Team Rc. R: A language and environment for statistical computing. 2019.
82. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77.
83. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017;45(4):e18.
84. Grant JR, Stothard P. The CGView Server: a comparative genomics tool for circular genomes. Nucleic Acids Res. 2008;36(Web Server issue):W181-4.
85. Donath A, Juhling F, Al-Arab M, Bernhart SH, Reinhardt F, Stadler PF, et al. Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. Nucleic Acids Res. 2019;47(20):10543-52.
86. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.
87. Edwards JA, Edwards RA. fastq-pair: efficient synchronization of paired-end fastq files. bioRxiv. 2019.
88. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
89. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578-9.
90. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biology. 2012;13:R56.
91. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963.

92. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
93. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
94. Picard toolkit. Broad Institute, GitHub repository. 2019.
95. Hoff KJ, Stanke M. Predicting Genes in Single Genomes with AUGUSTUS. Curr Protoc Bioinformatics. 2019;65(1):e57.
96. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2015.
97. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.
98. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142-i50.
99. UniProt C. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506-D15.
100. Porras-Alfaro A, Liu KL, Kuske CR, Xie G. From genus to phylum: large-subunit and internal transcribed spacer rRNA operon regions show similar classification accuracies influenced by database composition. Appl Environ Microbiol. 2014;80(3):829-40.
101. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications. 2018;9(1):5114.
102. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46(W1):W95-W101.
103. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45(W1):W36-W41.
104. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 2020;16(1):60-8.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

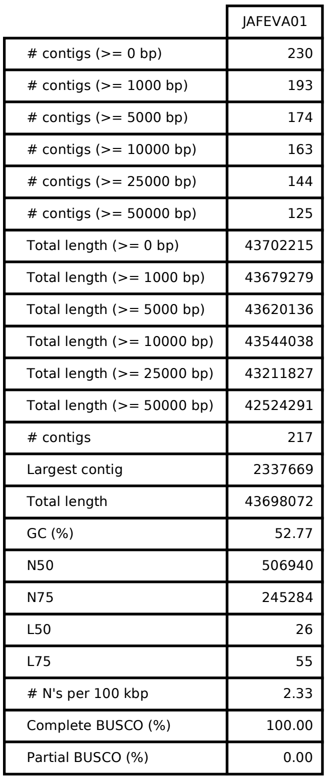
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 1

```
>ITS1_5.8S_ITS2_CBS
ATTATAGAGTTTATAAAACTCCCAAACCCATGTGAACTTACCATTGTTGCCTCGGCG
GAGCCTACCCTGTAGCTACCCTGTAAGGGCCTACCCTGTAGCGCACCCCGCCGGTGG
AATTTCAAACTCTTGTTATTTTTAAATGAATCTGAGCGTCTTATTTTAATAAGTCAAA
ACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGCGA
TAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAACGCACATTGCG
CCCATTAGTATTCTAGTGGGCATGCCTGTTCGAGCGTCATTTCAACCCTTAAGCCTAG
CTTAGTGTTGGGAATCTACTGTATTGTAGTTCCTGAAAAACAACGGCGGAACTATAG
TGTCCTCTGAGCGTAGTAATTTTTTATCTCGCTTTTGTCAGGTGCTGTAGCTCTTGCC
GCTAAACCCCCCAATTTTTAATGGTTGACCTCGGATCAGGTAGGAATACCCGCTGAA
CTTAAGCATATCAATAA
>ITS1_5.8S_ITS2_TUCIM_5799
ATTATAGAGTTTATAAAACTCCCAAACCCATGTGAACTTACCATTGTTGCCTCGGCG
GAGCCTACCCTGTAGCTACCCTGTAAGGGCCTACCCTGTAGCGCACCCCGCCGGTGG
AATTTCAAACTCTTGTTATTTTTAAATGAATCTGAGCGTCTTATTTTAATAAGTCAAA
ACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGCGA
TAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAACGCACATTGCG
CCCATTAGTATTCTAGTGGGCATGCCTGTTCGAGCGTCATTTCAACCCTTAAGCCTAG
CTTAGTGTTGGGAATCTACTGTATTGTAGTTCCTGAAAAACAACGGCGGAACTATAG
TGTCCTCTGAGCGTAGTAATTTTTTATCTCGCTTTTGTCAGGTGCTGTAGCTCTTGCC
GCTAAACCCCCAAATTTTTAATGGTTGACCTCGGATCAGGTAGGAATACCCGCTGAA
CTTAAGCATATCAATAAGCGGAGGAAA
>ITS1_5.8S_ITS2_TUCIM_5827
ATTATAGAGTTTATAAAACTCCCAAACCCATGTGAACTTACCATTGTTGCCTCGGCG
GAGCCTACCCTGTAGCTACCCTGTAAGGGCCTACCCNGTAGCGCACCCCGCCGGTGG
AATTTCAAACTCTTGTTATTTTTAAATGAATCTGAGCGTCTTATTTTAATAAGTCAAA
ACTTTCAACAACGGATCTCTTGGTTCTGGCATCGATGAAGAACGCAGCGAAATGCGA
TAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAACGCACATTGCG
CCCATTAGTATTCTAGTGGGCATGCCTGTTCGAGCGTCATTTCAACCCTTAAGCCTAG
CTTAGTGTTGGGAATCTACTGTATTGTAGTTCCTGAAAAACAACGGCGGAACTATAG
TGTCCTCTGAGCGTAGTAATTTTTTATCTCGCTTTTGTCAGGTGCTGTAGCTCTTGCC
GCTAAACCCCCCAATTTTTAATGGTTGACCTCGGATCAGGTAGGAATACCCGCTGAA
CTTAAGCATATCAATAAGCGGAGGAAAA
```

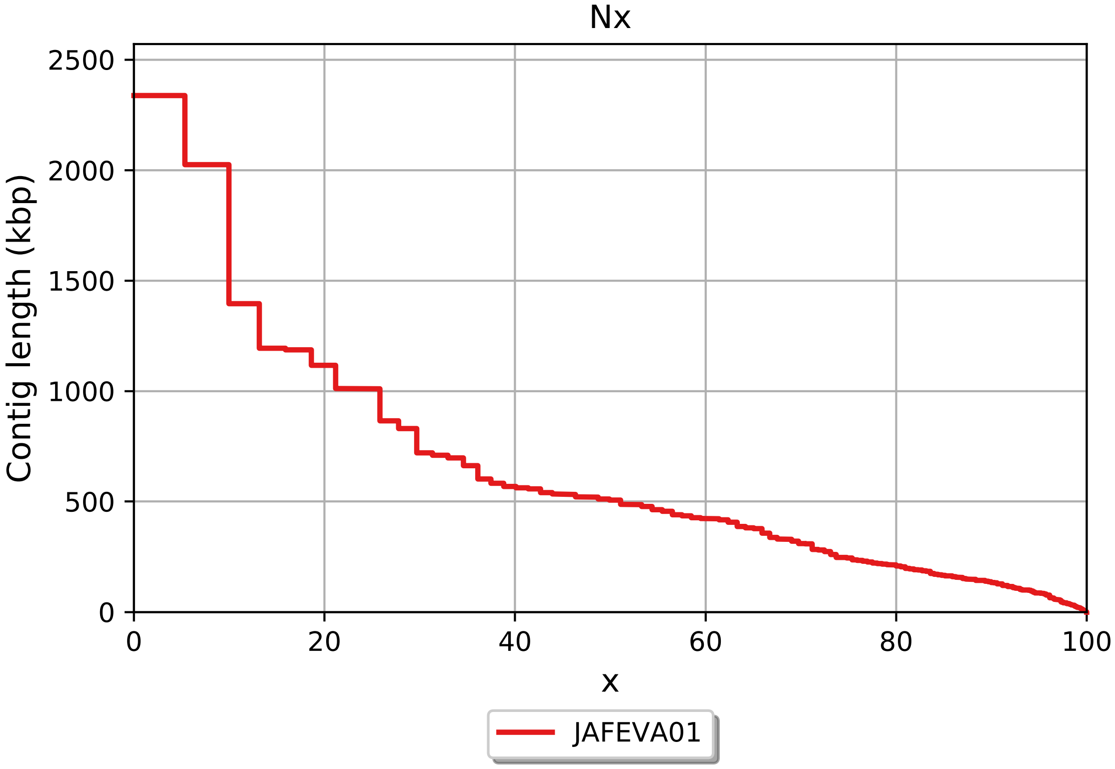
Additional File 2

Report

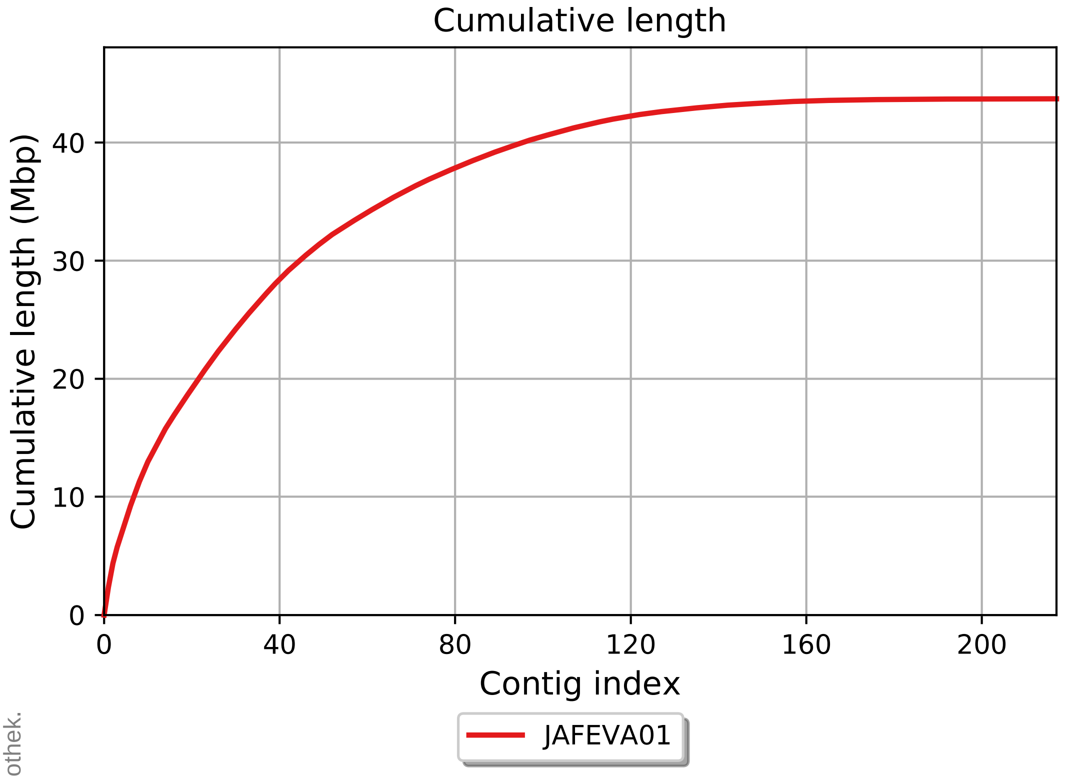


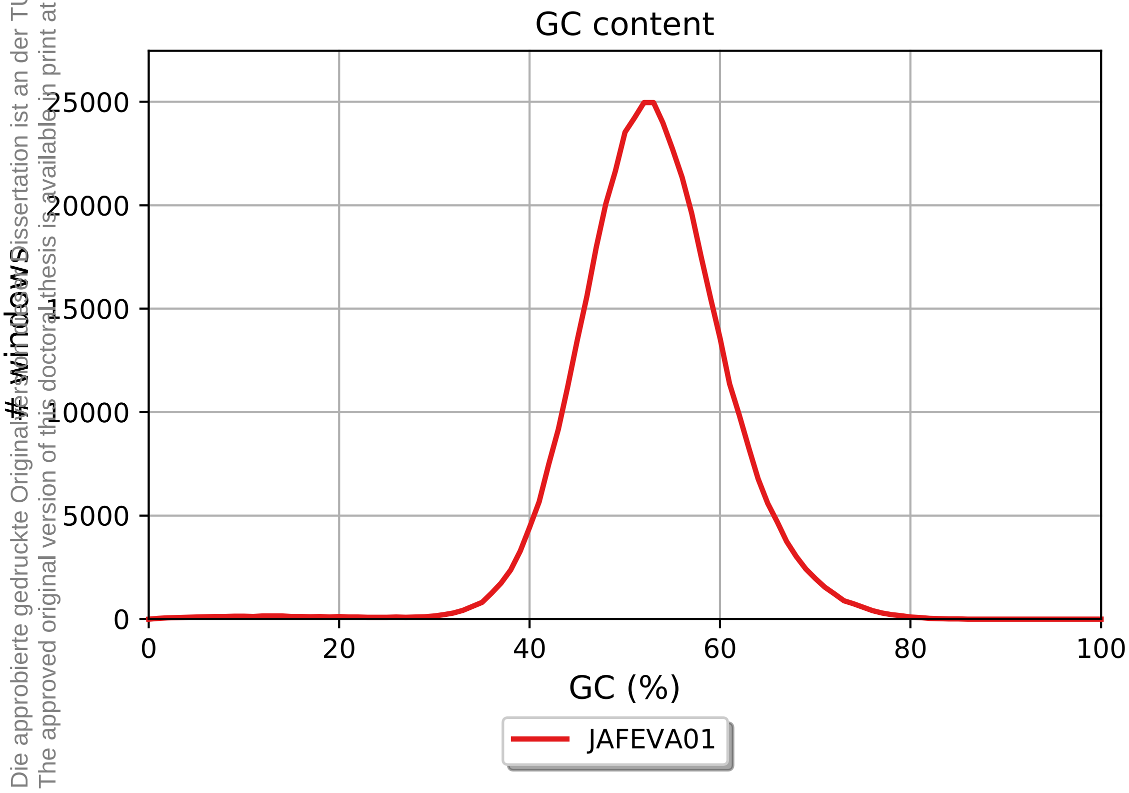
	JAFEVA01
# contigs (>= 0 bp)	230
# contigs (>= 1000 bp)	193
# contigs (>= 5000 bp)	174
# contigs (>= 10000 bp)	163
# contigs (>= 25000 bp)	144
# contigs (>= 50000 bp)	125
Total length (>= 0 bp)	43702215
Total length (>= 1000 bp)	43679279
Total length (>= 5000 bp)	43620136
Total length (>= 10000 bp)	43544038
Total length (>= 25000 bp)	43211827
Total length (>= 50000 bp)	42524291
# contigs	217
Largest contig	2337669
Total length	43698072
GC (%)	52.77
N50	506940
N75	245284
L50	26
L75	55
# N's per 100 kbp	2.33
Complete BUSCO (%)	100.00
Partial BUSCO (%)	0.00

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

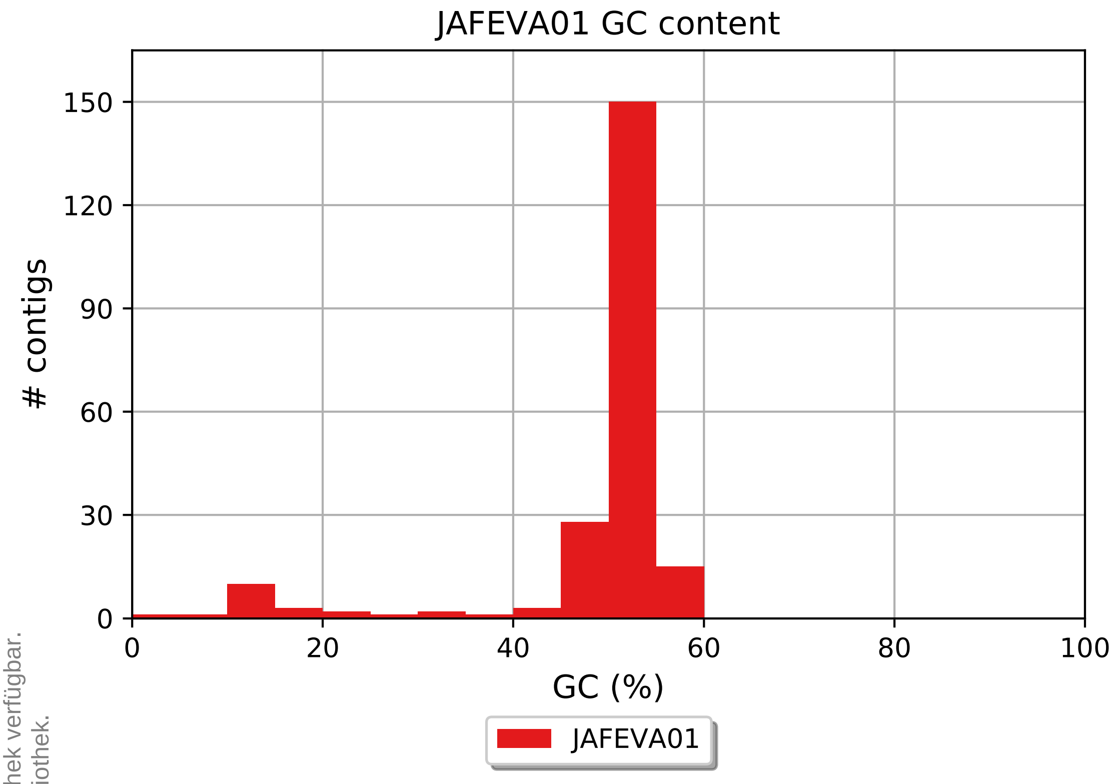


Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek
Your knowledge hub
WIEN





Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

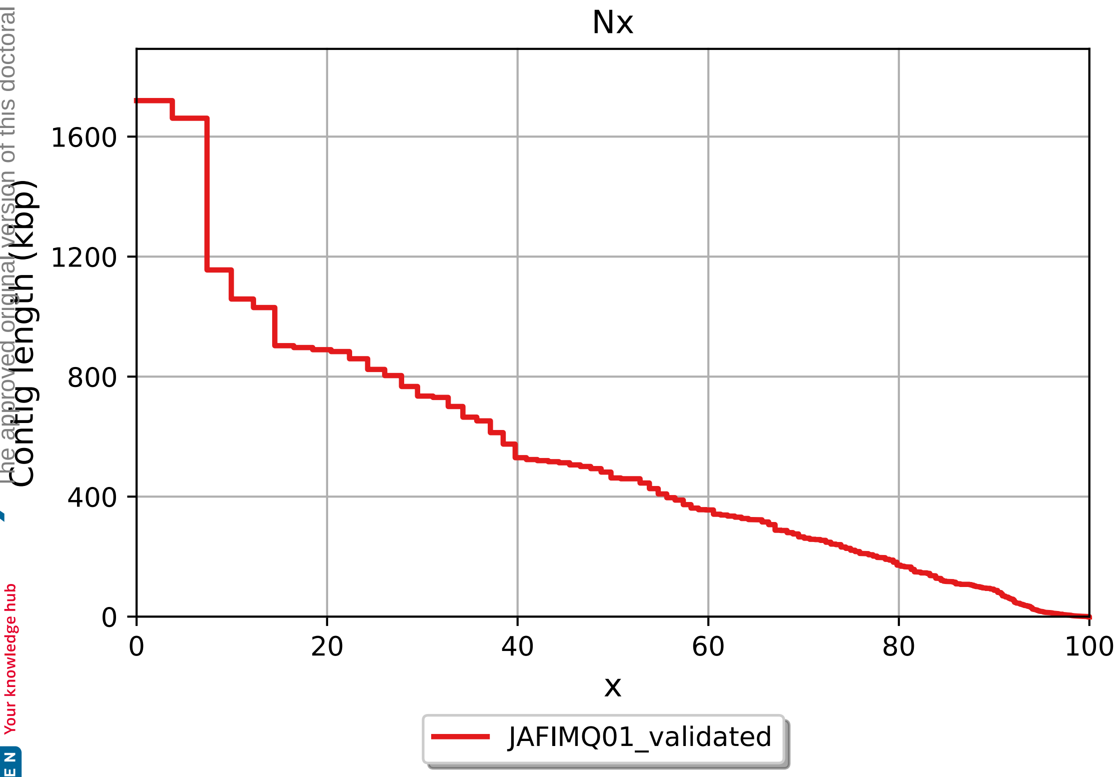
Additional File 3

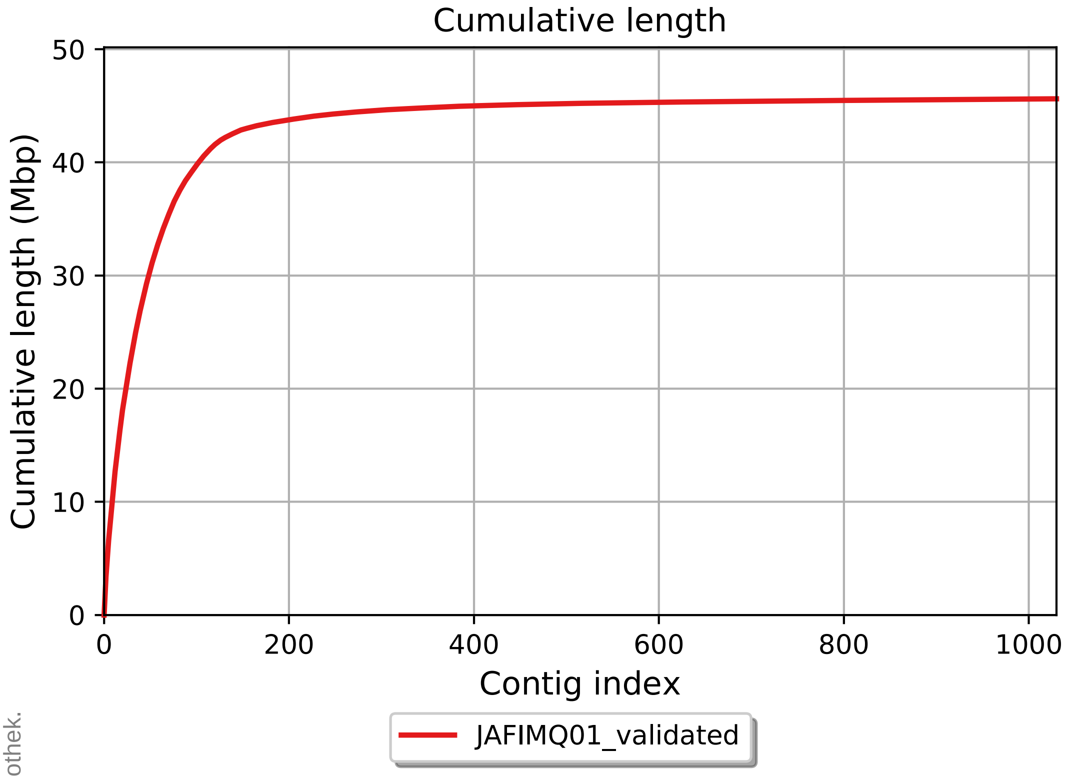
Report

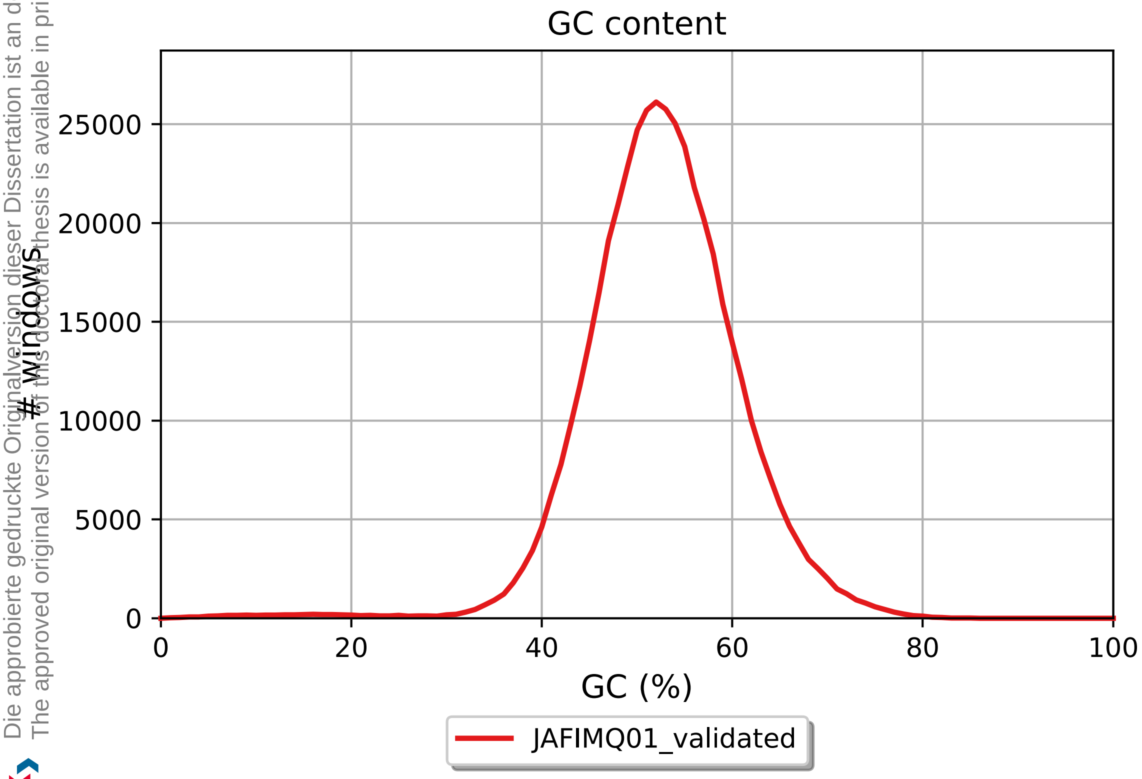
	JAFIMQ01_validated
# contigs (>= 0 bp)	2730
# contigs (>= 1000 bp)	609
# contigs (>= 5000 bp)	312
# contigs (>= 10000 bp)	230
# contigs (>= 25000 bp)	150
# contigs (>= 50000 bp)	127
Total length (>= 0 bp)	46154457
Total length (>= 1000 bp)	45329533
Total length (>= 5000 bp)	44679931
Total length (>= 10000 bp)	44119280
Total length (>= 25000 bp)	42908652
Total length (>= 50000 bp)	42023056
# contigs	1030
Largest contig	1719970
Total length	45619856
GC (%)	52.65
N50	462712
N75	221357
L50	30
L75	65
# N's per 100 kbp	2.17
Complete BUSCO (%)	100.00
Partial BUSCO (%)	0.00

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

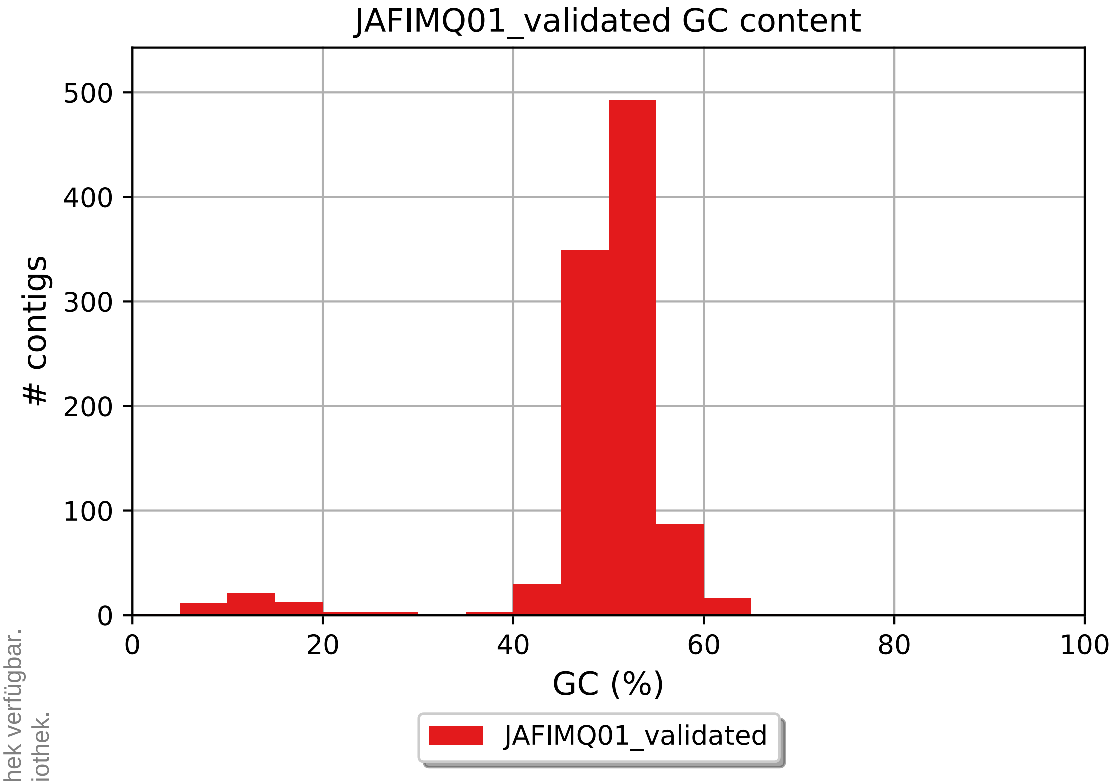
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.







Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



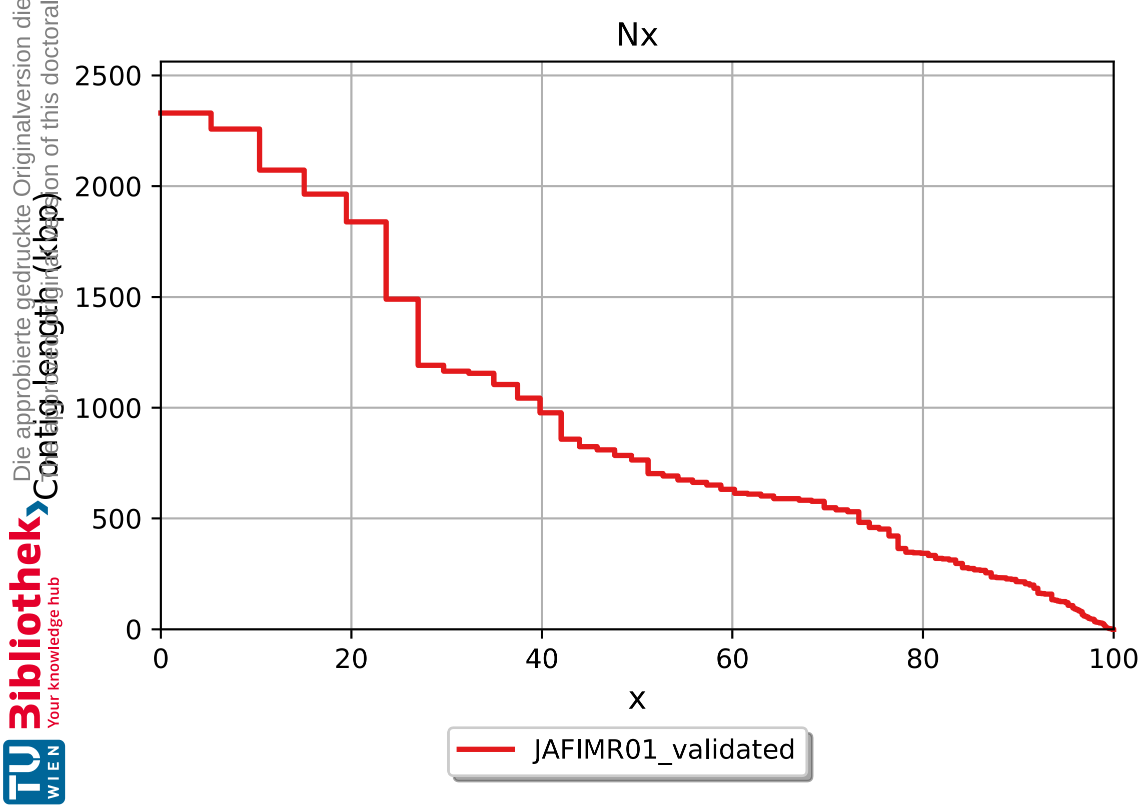
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 4

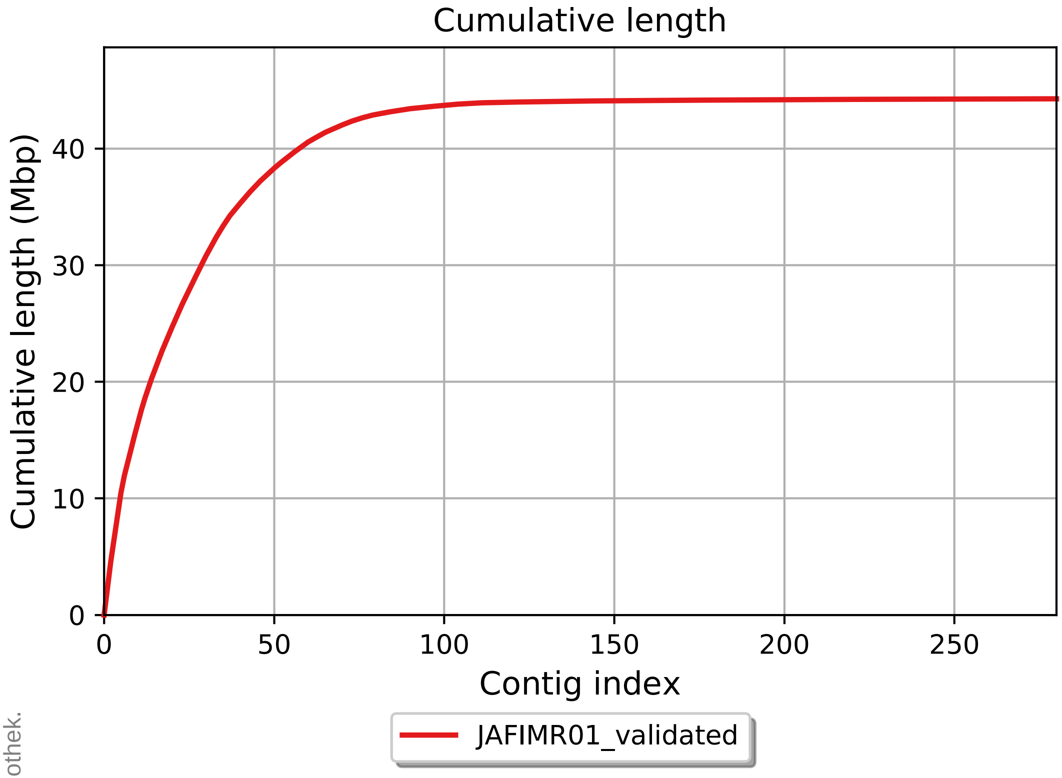
Report

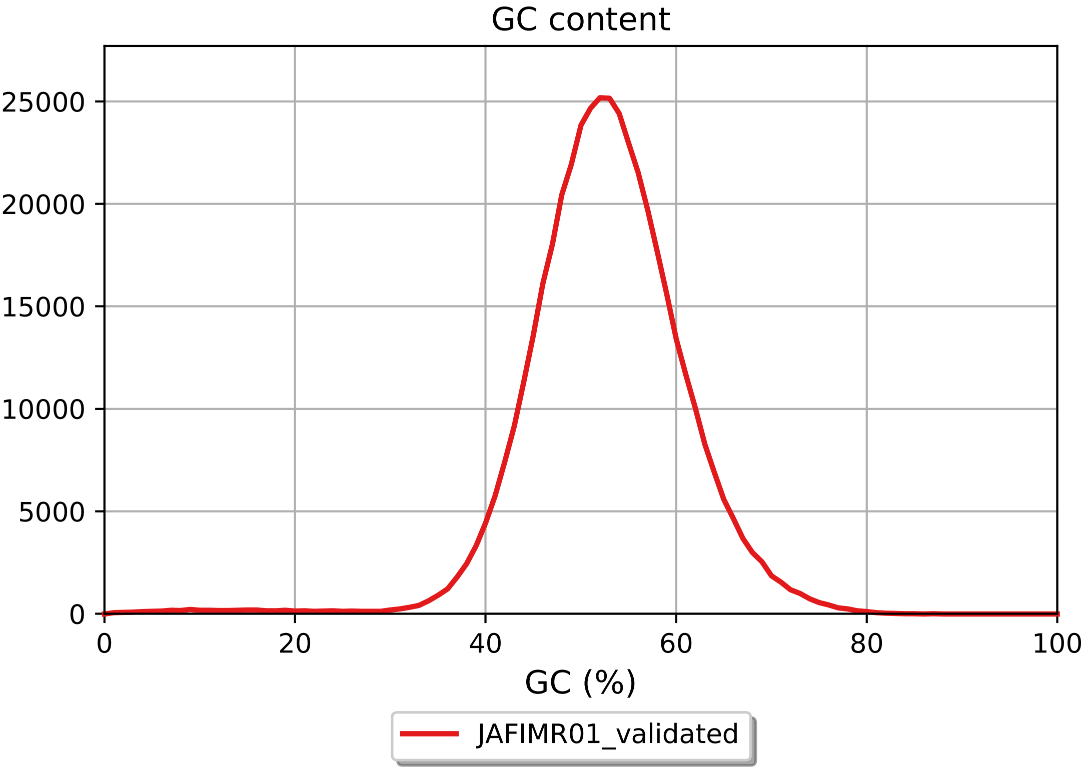
	JAFIMR01_validated
# contigs (>= 0 bp)	693
# contigs (>= 1000 bp)	221
# contigs (>= 5000 bp)	123
# contigs (>= 10000 bp)	110
# contigs (>= 25000 bp)	103
# contigs (>= 50000 bp)	83
Total length (>= 0 bp)	44394130
Total length (>= 1000 bp)	44223375
Total length (>= 5000 bp)	44001800
Total length (>= 10000 bp)	43909600
Total length (>= 25000 bp)	43797689
Total length (>= 50000 bp)	43099379
# contigs	280
Largest contig	2329648
Total length	44265620
GC (%)	52.66
N50	764765
N75	459401
L50	17
L75	35
# N's per 100 kbp	1.81
Complete BUSCO (%)	100.00
Partial BUSCO (%)	0.00

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

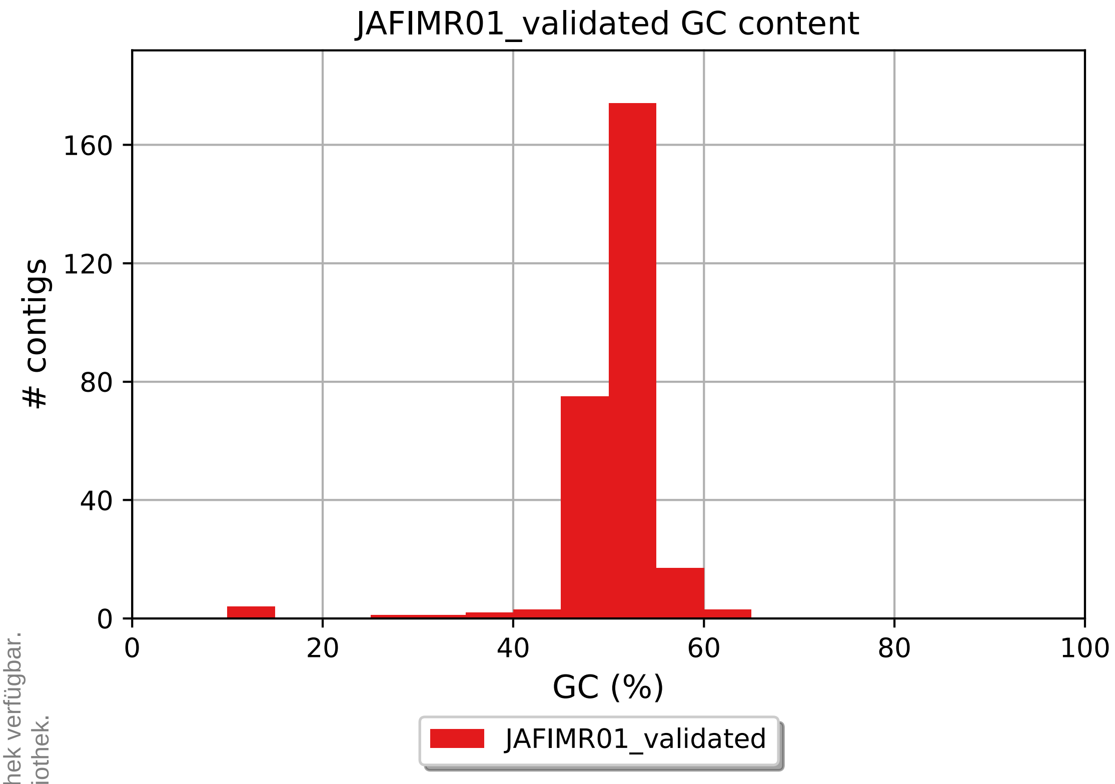


Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub





Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Additional File 5

```
==================================================
file name: W. moseri CBS
sequences:          230
total length:   43702215 bp  (43701207 bp excl N/X-runs)
GC level:         52.77 %
bases masked:     348341 bp ( 0.80 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:               36         2416 bp    0.01 %
      ALUs            0            0 bp    0.00 %
      MIRs           10          640 bp    0.00 %
LINEs:              221        16696 bp    0.04 %
      LINE1          11          891 bp    0.00 %
      LINE2          56         4199 bp    0.01 %
      L3/CR1         61         4243 bp    0.01 %
LTR elements:         5          336 bp    0.00 %
      ERVL            0            0 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI      4          258 bp    0.00 %
      ERV_classII     0            0 bp    0.00 %
DNA elements:        49         3664 bp    0.01 %
     hAT-Charlie      1           54 bp    0.00 %
     TcMar-Tigger     4          269 bp    0.00 %
Unclassified:         1          142 bp    0.00 %
Total interspersed repeats:    23254 bp    0.05 %
Small RNA:           74        10136 bp    0.02 %
Satellites:           0            0 bp    0.00 %
Simple repeats:    7309       287466 bp    0.66 %
Low complexity:     608        27917 bp    0.06 %
==================================================
* most repeats fragmented by insertions or deletions
  have been counted as one element

RepeatMasker Combined Database: Dfam_3.0
run with rmblastn version 2.9.0+
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Additional File 6

```
==================================================
file name: W. moseri TUCIM 5827
sequences:         2730
total length:   46154645 bp  (46153577 bp excl N/X-runs)
GC level:         52.63 %
bases masked:     363606 bp ( 0.79 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:               33         2231 bp    0.00 %
      ALUs            0            0 bp    0.00 %
      MIRs            9          651 bp    0.00 %
LINEs:              220        16757 bp    0.04 %
      LINE1          13         1058 bp    0.00 %
      LINE2          54         4203 bp    0.01 %
      L3/CR1         65         4473 bp    0.01 %
LTR elements:         3          200 bp    0.00 %
      ERVL            0            0 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI      3          200 bp    0.00 %
      ERV_classII     0            0 bp    0.00 %
DNA elements:        55         4260 bp    0.01 %
     hAT-Charlie      1           66 bp    0.00 %
     TcMar-Tigger     5          410 bp    0.00 %
Unclassified:         1           72 bp    0.00 %
Total interspersed repeats:    23520 bp    0.05 %
Small RNA:           74        11815 bp    0.03 %
Satellites:           0            0 bp    0.00 %
Simple repeats:    7532       297695 bp    0.64 %
Low complexity:     652        30995 bp    0.07 %
==================================================
* most repeats fragmented by insertions or deletions
  have been counted as one element

RepeatMasker Combined Database: Dfam_3.0
run with rmblastn version 2.9.0+
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 7

```
==================================================f
ile name: W. moseri TUCIM 5799
sequences:          693
total length:   44394166 bp  (44393399 bp excl N/X-runs)
GC level:         52.65 %
bases masked:     357820 bp ( 0.81 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:               35        2404 bp    0.01 %
      ALUs            0           0 bp    0.00 %
      MIRs            8         525 bp    0.00 %
LINEs:              222       17399 bp    0.04 %
      LINE1          12         837 bp    0.00 %
      LINE2          57        4623 bp    0.01 %
      L3/CR1         69        5111 bp    0.01 %
LTR elements:         3         204 bp    0.00 %
      ERVL            0           0 bp    0.00 %
      ERVL-MaLRs      0           0 bp    0.00 %
      ERV_classI      3         204 bp    0.00 %
      ERV_classII     0           0 bp    0.00 %
DNA elements:        49        3598 bp    0.01 %
     hAT-Charlie      1          66 bp    0.00 %
     TcMar-Tigger     5         324 bp    0.00 %
Unclassified:         1         142 bp    0.00 %
Total interspersed repeats:   23747 bp    0.05 %
Small RNA:           78       12157 bp    0.03 %
Satellites:           0           0 bp    0.00 %
Simple repeats:    7432      294925 bp    0.66 %
Low complexity:     600       27670 bp    0.06 %
==================================================
* most repeats fragmented by insertions or deletions
  have been counted as one element

RepeatMasker Combined Database: Dfam_3.0
run with rmblastn version 2.9.0+
```

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Additional File 8

Sequence Name	tRNA #	tRNA Begin	Bounds End	tRNA Type	Anti Codon	Intron Bounds Begin	End	Cove Score
--------	------	----	------	----	-----	-----	----	------
JAFEVA010000001.1	1	447552	447623	Thr	AGT	0	0	70.11
JAFEVA010000001.1	2	630694	630784	Phe	GAA	630731	630748	63.84
JAFEVA010000001.1	3	1363498	1363568	Gly	GCC	0	0	58.71
JAFEVA010000001.1	4	1606701	1606787	Arg	TCG	1606737	1606751	57.71
JAFEVA010000001.1	5	2231639	2231541	Leu	TAA	2231602	2231585	62.36
JAFEVA010000001.1	6	1811481	1811410	Asp	GTC	0	0	72.49
JAFEVA010000001.1	7	1550966	1550893	Val	AAC	0	0	75.41
JAFEVA010000001.1	8	766004	765899	Ile	AAT	765966	765935	61.75
JAFEVA010000002.1	1	218283	218372	Arg	ACG	218319	218336	53.20
JAFEVA010000002.1	2	421939	422048	Asn	GTT	421977	422012	54.65
JAFEVA010000002.1	3	1207820	1207924	Leu	AAG	1207859	1207878	61.86
JAFEVA010000002.1	4	1929213	1929294	Ala	AGC	1929249	1929258	66.20
JAFEVA010000002.1	5	1655982	1655887	Arg	CCT	1655946	1655923	60.41
JAFEVA010000002.1	6	947292	947204	Gln	CTG	947254	947239	59.00
JAFEVA010000002.1	7	785583	785512	Glu	CTC	0	0	66.10
JAFEVA010000002.1	8	720018	719945	Val	AAC	0	0	75.41
JAFEVA010000002.1	9	233601	233531	Gly	GCC	0	0	58.71
JAFEVA010000003.1	1	46019	46092	Val	AAC	0	0	78.00
JAFEVA010000003.1	2	386314	386384	Gly	GCC	0	0	58.71
JAFEVA010000003.1	3	563470	563542	Arg	CCG	0	0	64.28
JAFEVA010000003.1	4	1123235	1123164	Thr	AGT	0	0	68.37
JAFEVA010000003.1	5	798217	798128	Pro	AGG	798181	798164	67.69
JAFEVA010000003.1	6	292665	292594	Gly	TCC	0	0	66.41
JAFEVA010000003.1	7	219083	219013	Gly	GCC	0	0	58.71
JAFEVA010000004.1	1	200317	200407	Arg	TCG	200353	200371	54.51
JAFEVA010000004.1	2	926291	926377	Met	CAT	926327	926341	54.80
JAFEVA010000004.1	3	1108082	1107983	Ser	GCT	1108045	1108027	55.02
JAFEVA010000005.1	1	566982	567138	Arg	ACG	567018	567103	33.64
JAFEVA010000005.1	2	722011	721919	Phe	GAA	721974	721955	64.87
JAFEVA010000005.1	3	703334	703236	Leu	CAG	703296	703280	56.71
JAFEVA010000006.1	1	994957	995063	Ser	CGA	994995	995019	60.35
JAFEVA010000006.1	2	1086488	1086592	Met	CAT	1086526	1086556	76.54
JAFEVA010000006.1	3	1086901	1086816	Ala	TGC	1086865	1086852	53.44
JAFEVA010000006.1	4	417761	417690	Thr	AGT	0	0	68.37
JAFEVA010000006.1	5	36183	36085	Arg	TCT	36147	36121	65.58
JAFEVA010000007.1	1	369312	369399	Gln	CTG	369350	369364	60.27
JAFEVA010000007.1	2	655182	655092	Phe	GAA	655145	655128	64.47
JAFEVA010000007.1	3	502833	502741	Asp	GTC	502796	502776	64.71
JAFEVA010000008.1	1	857894	857806	Gln	TTG	857857	857841	59.11
JAFEVA010000008.1	2	440374	440270	Leu	AAG	440335	440316	61.86
JAFEVA010000009.1	1	186888	186976	Arg	CCG	186924	186940	48.51
JAFEVA010000009.1	2	230211	230282	Glu	TTC	0	0	63.63
JAFEVA010000009.1	3	470562	470641	Pseudo	???	0	0	22.51

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

JAFEVA010000009.1	4	738885	738957	Lys	CTT	0	0	75.42
JAFEVA010000009.1	5	801228	801311	Ala	AGC	801264	801275	65.70
JAFEVA010000009.1	6	865263	865178	Glu	TTC	865225	865213	57.30
JAFEVA010000009.1	7	635420	635321	Ile	AAT	635382	635357	61.84
JAFEVA010000009.1	8	397711	397639	Lys	CTT	0	0	75.42
JAFEVA010000009.1	9	326067	325954	Leu	CAA	326029	325998	56.84
JAFEVA010000009.1	10	128364	128266	Leu	CAG	128326	128310	54.67
JAFEVA010000010.1	1	646114	646035	His	GTG	646077	646070	51.84
JAFEVA010000010.1	2	99629	99543	Val	TAC	99592	99579	63.97
JAFEVA010000011.1	1	178640	178730	Asp	GTC	178677	178695	66.08
JAFEVA010000011.1	2	187248	187095	Leu	CAG	187211	187136	24.85
JAFEVA010000012.1	1	279052	279124	Lys	CTT	0	0	75.42
JAFEVA010000012.1	2	299262	299349	Arg	TCG	299298	299313	61.22
JAFEVA010000012.1	3	348269	348350	Glu	CTC	348306	348315	60.65
JAFEVA010000012.1	4	348716	348610	Met	CAT	348678	348646	76.17
JAFEVA010000013.1	1	501561	501662	Ala	AGC	501597	501626	63.92
JAFEVA010000013.1	2	616973	616883	Phe	GAA	616936	616919	64.47
JAFEVA010000013.1	3	302042	301964	His	GTG	302005	301999	45.90
JAFEVA010000014.1	1	196906	196999	Arg	CCT	196943	196963	58.83
JAFEVA010000014.1	2	234792	234873	Glu	TTC	234829	234838	56.92
JAFEVA010000015.1	1	95846	95919	Val	AAC	0	0	78.00
JAFEVA010000015.1	2	168499	168417	Gln	TTG	168462	168452	57.55
JAFEVA010000015.1	3	70398	70299	Ile	AAT	70360	70335	64.38
JAFEVA010000015.1	4	70003	69922	Ala	AGC	69967	69958	66.20
JAFEVA010000016.1	1	68287	68369	Glu	CTC	68324	68334	60.28
JAFEVA010000016.1	2	428802	428730	Lys	CTT	0	0	75.42
JAFEVA010000016.1	3	226155	226084	Gly	TCC	0	0	66.41
JAFEVA010000017.1	1	95600	95680	Asp	GTC	95637	95645	65.66
JAFEVA010000017.1	2	182295	182412	Ala	TGC	182333	182376	59.43
JAFEVA010000017.1	3	481039	480944	Pro	CGG	481003	480980	51.00
JAFEVA010000017.1	4	476617	476547	Gly	GCC	0	0	58.71
JAFEVA010000017.1	5	476195	476088	Ser	AGA	476157	476132	57.38
JAFEVA010000018.1	1	359015	359123	Ser	AGA	359053	359079	67.02
JAFEVA010000018.1	2	200499	200398	Ser	GCT	200462	200442	62.68
JAFEVA010000019.1	1	366350	366267	Ala	AGC	366314	366303	67.37
JAFEVA010000021.1	1	291634	291563	Thr	AGT	0	0	70.11
JAFEVA010000022.1	1	518727	518812	Thr	CGT	518763	518776	70.51
JAFEVA010000022.1	2	391570	391473	Ile	AAT	391532	391509	63.36
JAFEVA010000023.1	1	348009	348092	Ala	AGC	348045	348056	66.46
JAFEVA010000024.1	1	81189	81270	Gly	CCC	81225	81235	55.52
JAFEVA010000024.1	2	432370	432443	Val	AAC	0	0	76.92
JAFEVA010000024.1	3	432835	432916	Glu	CTC	432872	432881	60.29
JAFEVA010000024.1	4	465897	465998	Ser	CGA	465934	465954	60.11
JAFEVA010000025.1	1	63974	64127	Arg	GCG	64010	64092	40.02
JAFEVA010000026.1	1	177871	177943	Lys	CTT	0	0	75.42
JAFEVA010000026.1	2	387583	387666	Glu	CTC	387620	387631	59.91
JAFEVA010000026.1	3	161473	161371	Pro	TGG	161437	161407	60.19
JAFEVA010000027.1	1	146810	146706	Leu	AAG	146771	146752	61.86
JAFEVA010000028.1	1	92829	92684	Undet	???	0	0	34.69

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

JAFEVA010000028.1	2	59263	59172	Ser	AGA	59225	59216	69.74
JAFEVA010000028.1	3	59164	59065	Ile	AAT	59126	59101	64.53
JAFEVA010000029.1	1	61896	61987	Ser	AGA	61934	61943	69.13
JAFEVA010000029.1	2	66721	66631	Asp	GTC	66684	66666	66.22
JAFEVA010000030.1	1	210343	210442	Ile	AAT	210381	210406	64.39
JAFEVA010000030.1	2	303501	303622	Ser	CGA	303534	303587	20.93
JAFEVA010000030.1	3	400893	401025	Undet	???	0	0	22.80
JAFEVA010000030.1	4	401206	401286	Asp	GTC	401243	401251	63.74
JAFEVA010000030.1	5	271469	271388	Glu	CTC	271432	271423	59.87
JAFEVA010000032.1	1	58723	58793	Gly	GCC	0	0	58.71
JAFEVA010000032.1	2	219070	219155	Trp	CCA	219108	219119	49.28
JAFEVA010000032.1	3	98174	98103	Met	CAT	0	0	70.01
JAFEVA010000033.1	1	143361	143277	Tyr	GTA	143324	143313	63.18
JAFEVA010000034.1	1	208435	208229	Pro	CGG	208399	208262	32.57
JAFEVA010000035.1	1	284117	284029	Trp	CCA	284080	284065	61.35
JAFEVA010000036.1	1	239442	239516	Asn	GTT	0	0	70.83
JAFEVA010000037.1	1	413607	413509	Arg	TCT	413571	413545	63.55
JAFEVA010000037.1	2	151109	151016	Ile	TAT	151072	151052	59.25
JAFEVA010000038.1	1	172705	172777	Lys	CTT	0	0	75.42
JAFEVA010000038.1	2	230143	230247	Met	CAT	230181	230211	73.74
JAFEVA010000038.1	3	203378	203278	Ser	GCT	203341	203322	56.99
JAFEVA010000038.1	4	79498	79409	Met	CAT	79462	79445	56.34
JAFEVA010000040.1	1	68991	69108	Ala	TGC	69029	69072	56.81
JAFEVA010000040.1	2	254427	254354	Val	AAC	0	0	75.41
JAFEVA010000043.1	1	305568	305646	Glu	CTC	305605	305611	60.26
JAFEVA010000046.1	1	77394	77468	Asn	GTT	0	0	70.83
JAFEVA010000046.1	2	228291	228369	His	GTG	228328	228334	48.49
JAFEVA010000046.1	3	246660	246588	Val	CAC	0	0	76.81
JAFEVA010000049.1	1	183447	183583	Ser	CGA	183485	183548	30.51
JAFEVA010000053.1	1	109009	109092	Trp	CCA	109045	109056	50.62
JAFEVA010000055.1	1	55575	55686	Leu	CAA	55613	55642	58.72
JAFEVA010000055.1	2	43639	43567	Lys	CTT	0	0	75.42
JAFEVA010000056.1	1	125513	125628	Lys	TTT	125551	125592	61.14
JAFEVA010000056.1	2	118073	117969	Leu	AAG	118034	118015	61.86
JAFEVA010000060.1	1	135115	135043	Val	CAC	0	0	78.92
JAFEVA010000062.1	1	138363	138449	Cys	GCA	138399	138413	59.58
JAFEVA010000063.1	1	177449	177364	Ala	CGC	177413	177400	60.95
JAFEVA010000063.1	2	40394	40324	Pro	TGG	0	0	61.43
JAFEVA010000064.1	1	110286	110399	Pseudo	GTG	110327	110363	22.79
JAFEVA010000064.1	2	191269	191371	Tyr	GTA	191306	191335	60.09
JAFEVA010000064.1	3	149197	149109	Arg	ACG	149161	149145	51.78
JAFEVA010000065.1	1	88474	88544	Asn	GTT	0	0	58.07
JAFEVA010000065.1	2	88565	88636	Lys	TTT	0	0	41.28
JAFEVA010000065.1	3	175966	176073	Leu	TAG	176003	176031	49.38
JAFEVA010000066.1	1	15101	15174	Val	AAC	0	0	75.41
JAFEVA010000066.1	2	15587	15500	Gln	CTG	15549	15535	55.17
JAFEVA010000068.1	1	176381	176297	Ala	TGC	176345	176333	53.94
JAFEVA010000069.1	1	57512	57657	Undet	???	0	0	37.00
JAFEVA010000069.1	2	104589	104486	Lys	CTT	104552	104522	65.49

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

JAFEVA010000070.1	1	165514	165584	Gly	GCC	0	0	58.71
JAFEVA010000070.1	2	174857	174927	Gly	GCC	0	0	58.71
JAFEVA010000071.1	1	46591	46514	Leu	CAA	0	0	25.79
JAFEVA010000072.1	1	142942	142858	Ala	CGC	142906	142894	61.12
JAFEVA010000072.1	2	114562	114470	Ser	CGA	114525	114514	62.73
JAFEVA010000073.1	1	157017	156932	Met	CAT	156980	156968	54.36
JAFEVA010000073.1	2	11486	11333	Ser	TGA	11449	11378	23.91
JAFEVA010000078.1	1	162037	161955	Glu	CTC	162000	161990	59.64
JAFEVA010000079.1	1	132021	132119	Ile	AAT	132059	132083	63.48
JAFEVA010000079.1	2	52314	52244	Pro	TGG	0	0	61.01
JAFEVA010000080.1	1	128140	128218	His	GTG	128177	128183	48.49
JAFEVA010000082.1	1	137339	137422	Glu	CTC	137376	137387	61.19
JAFEVA010000083.1	1	47493	47563	Gly	GCC	0	0	58.71
JAFEVA010000084.1	1	31087	31000	Gln	CTG	31049	31035	61.05
JAFEVA010000087.1	1	19310	19397	Tyr	GTA	19349	19361	65.81
JAFEVA010000091.1	1	11730	11638	Gly	CCC	11692	11674	62.21
JAFEVA010000094.1	1	58551	58467	Tyr	GTA	58514	58503	62.41
JAFEVA010000094.1	2	42450	42379	Gly	TCC	0	0	66.80
JAFEVA010000095.1	1	342	413	Sup	TTA	0	0	32.21
JAFEVA010000096.1	1	81894	81976	Thr	TGT	81930	81940	66.18
JAFEVA010000100.1	1	61035	61126	Cys	GCA	61071	61090	63.44
JAFEVA010000102.1	1	4467	4396	Thr	AGT	0	0	68.37
JAFEVA010000107.1	1	6329	6413	Pro	TGG	6366	6377	48.17
JAFEVA010000127.1	1	23177	23283	Met	CAT	23215	23247	75.74
JAFEVA010000129.1	1	37068	36998	Asn	GTT	0	0	58.07
JAFEVA010000129.1	2	36977	36906	Lys	TTT	0	0	41.28
JAFEVA010000129.1	3	36674	36604	Gly	TCC	0	0	39.50
JAFEVA010000129.1	4	35408	35336	Pseudo	GTC	0	0	37.80
JAFEVA010000129.1	5	35202	35132	SeC	TCA	0	0	49.80
JAFEVA010000129.1	6	34913	34841	Pseudo	TGG	0	0	33.17
JAFEVA010000129.1	7	34820	34734	Pseudo	TGA	0	0	33.86
JAFEVA010000129.1	8	33722	33650	Val	TAC	0	0	53.91
JAFEVA010000129.1	9	31504	31433	Pseudo	GAT	0	0	29.76
JAFEVA010000129.1	10	26119	26049	Thr	TGT	0	0	34.60
JAFEVA010000129.1	11	25999	25928	Glu	TTC	0	0	54.15
JAFEVA010000129.1	12	25920	25849	Met	CAT	0	0	37.19
JAFEVA010000129.1	13	25779	25707	Pseudo	CAT	0	0	37.21
JAFEVA010000129.1	14	25678	25596	Pseudo	TAA	0	0	28.30
JAFEVA010000129.1	15	25521	25450	Ala	TGC	0	0	54.84
JAFEVA010000129.1	16	25426	25354	Phe	GAA	0	0	50.51
JAFEVA010000129.1	17	25268	25196	Pseudo	TTG	0	0	34.85
JAFEVA010000129.1	18	25052	24980	His	GTG	0	0	40.19
JAFEVA010000129.1	19	24909	24837	Pseudo	CAT	0	0	39.25
JAFEVA010000129.1	20	19850	19778	Val	TAC	0	0	49.28
JAFEVA010000129.1	21	13867	13797	Pseudo	TCT	0	0	45.82
JAFEVA010000129.1	22	10150	10080	Arg	ACG	0	0	50.29
JAFEVA010000129.1	23	3481	3411	Pseudo	TCT	0	0	40.52
JAFEVA010000135.1	1	2637	2746	Asn	GTT	2675	2710	60.39
JAFEVA010000138.1	1	29423	29511	Arg	ACG	29459	29475	51.64

JAFEVA010000141.1	1	9996	10085	Arg	ACG	10032	10049	50.14
JAFEVA010000142.1	1	20571	20501	Gly	GCC	0	0	58.71

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 9

Sequence		tRNA	Bounds	tRNA	Anti	Intron Bounds		Cove
Name	tRNA #	Begin	End	Type	Codon	Begin	End	Score
--------	-------	----	------	----	-----	-----	----	------
JAFIMQ010000001.1	1	449008	449095	Gln	CTG	449046	449060	59.64
JAFIMQ010000001.1	2	1207273	1207356	Glu	CTC	1207310	1207321	61.19
JAFIMQ010000001.1	3	1222373	1222460	Glu	TTC	1222411	1222425	57.05
JAFIMQ010000001.1	4	1324126	1324216	Asp	GTC	1324163	1324181	66.08
JAFIMQ010000001.1	5	1332737	1332583	Leu	CAA	1332700	1332624	26.18
JAFIMQ010000001.1	6	734101	734011	Phe	GAA	734064	734047	64.47
JAFIMQ010000001.1	7	582433	582341	Asp	GTC	582396	582376	63.93
JAFIMQ010000002.1	1	143647	143731	Tyr	GTA	143684	143695	62.41
JAFIMQ010000002.1	2	159799	159870	Gly	TCC	0	0	66.80
JAFIMQ010000002.1	3	752826	752897	Thr	AGT	0	0	70.11
JAFIMQ010000002.1	4	935946	936036	Phe	GAA	935983	936000	63.84
JAFIMQ010000002.1	5	1071111	1071006	Ile	AAT	1071073	1071042	61.75
JAFIMQ010000003.1	1	200610	200702	Phe	GAA	200647	200666	64.87
JAFIMQ010000003.1	2	219275	219373	Leu	CAG	219313	219329	56.71
JAFIMQ010000003.1	3	839959	840048	Pro	AGG	839995	840012	67.69
JAFIMQ010000003.1	4	1074311	1074239	Arg	CCG	0	0	64.28
JAFIMQ010000003.1	5	352375	352219	Arg	ACG	352339	352254	33.50
JAFIMQ010000004.1	1	890365	890287	Glu	CTC	890328	890322	60.26
JAFIMQ010000004.1	2	853562	853481	Gly	CCC	853526	853516	55.52
JAFIMQ010000004.1	3	505047	504974	Val	AAC	0	0	76.92
JAFIMQ010000004.1	4	504582	504501	Glu	CTC	504545	504536	60.29
JAFIMQ010000004.1	5	471603	471502	Ser	CGA	471566	471546	60.11
JAFIMQ010000004.1	6	413370	413272	Arg	TCT	413334	413308	62.78
JAFIMQ010000004.1	7	150977	150884	Ile	TAT	150940	150920	59.25
JAFIMQ010000005.1	1	9360	9469	Asn	GTT	9398	9433	54.65
JAFIMQ010000005.1	2	804596	804700	Leu	AAG	804635	804654	61.86
JAFIMQ010000005.1	3	544304	544216	Gln	CTG	544266	544251	59.00
JAFIMQ010000005.1	4	382939	382868	Glu	CTC	0	0	61.21
JAFIMQ010000005.1	5	317018	316945	Val	AAC	0	0	75.41
JAFIMQ010000006.1	1	408007	408102	Pro	CGG	408043	408066	51.00
JAFIMQ010000006.1	2	412264	412334	Gly	GCC	0	0	58.71
JAFIMQ010000006.1	3	412682	412789	Ser	AGA	412720	412745	56.74
JAFIMQ010000006.1	4	790855	790775	Asp	GTC	790818	790810	66.29
JAFIMQ010000006.1	5	706624	706505	Ala	TGC	706586	706541	59.19
JAFIMQ010000006.1	6	141049	140942	Leu	TAG	141012	140984	49.38
JAFIMQ010000007.1	1	219244	219450	Pro	CGG	219280	219417	33.07
JAFIMQ010000007.1	2	404669	404904	Gly	ACC	404706	404868	27.80
JAFIMQ010000007.1	3	654904	654977	Val	AAC	0	0	78.00
JAFIMQ010000007.1	4	727518	727436	Gln	TTG	727481	727471	57.55
JAFIMQ010000007.1	5	629446	629347	Ile	AAT	629408	629383	64.38
JAFIMQ010000007.1	6	629027	628946	Ala	AGC	628991	628982	66.20
JAFIMQ010000008.1	1	574492	574399	Arg	CCT	574455	574435	58.83
JAFIMQ010000008.1	2	536612	536531	Glu	TTC	536575	536566	56.92

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

JAFIMQ010000009.1	1	152503	152616	Pseudo	GTG	152544	152580	22.79
JAFIMQ010000009.1	2	233559	233661	Tyr	GTA	233596	233625	60.09
JAFIMQ010000009.1	3	191444	191356	Arg	ACG	191408	191392	51.78
JAFIMQ010000010.1	1	407341	407439	Leu	TAA	407378	407395	62.36
JAFIMQ010000010.1	2	827578	827649	Asp	GTC	0	0	72.49
JAFIMQ010000011.1	1	175249	175347	Ile	AAT	175287	175311	63.34
JAFIMQ010000011.1	2	256272	256342	Pro	TGG	0	0	61.01
JAFIMQ010000011.1	3	377571	377688	Ala	TGC	377609	377652	56.81
JAFIMQ010000011.1	4	524825	524909	Tyr	GTA	524862	524873	63.18
JAFIMQ010000012.1	1	384940	384850	Arg	TCG	384904	384886	54.51
JAFIMQ010000013.1	1	203065	203156	Ser	AGA	203103	203112	69.13
JAFIMQ010000013.1	2	694701	694780	His	GTG	694738	694745	51.84
JAFIMQ010000013.1	3	207869	207779	Asp	GTC	207832	207814	66.22
JAFIMQ010000014.1	1	144691	144761	Gly	GCC	0	0	58.71
JAFIMQ010000014.1	2	160006	159917	Arg	ACG	159970	159953	53.20
JAFIMQ010000015.1	1	449897	450003	Met	CAT	449935	449967	76.17
JAFIMQ010000015.1	2	520101	520029	Lys	CTT	0	0	75.42
JAFIMQ010000015.1	3	499895	499808	Arg	TCG	499859	499844	60.59
JAFIMQ010000015.1	4	450349	450268	Glu	CTC	450312	450303	60.65
JAFIMQ010000016.1	1	172234	172315	Ala	AGC	172270	172279	66.20
JAFIMQ010000016.1	2	406629	406715	Cys	GCA	406665	406679	60.35
JAFIMQ010000016.1	3	688097	688197	Ser	GCT	688134	688153	56.99
JAFIMQ010000016.1	4	661331	661227	Met	CAT	661293	661263	73.74
JAFIMQ010000017.1	1	423445	423530	Thr	CGT	423481	423494	70.51
JAFIMQ010000017.1	2	558618	558727	Asn	GTT	558656	558691	60.39
JAFIMQ010000017.1	3	494978	494887	Cys	GCA	494942	494923	63.44
JAFIMQ010000017.1	4	294215	294118	Ile	AAT	294177	294154	63.36
JAFIMQ010000018.1	1	92762	92617	Undet	???	0	0	34.69
JAFIMQ010000018.1	2	59245	59154	Ser	AGA	59207	59198	69.74
JAFIMQ010000018.1	3	59146	59047	Ile	AAT	59108	59083	64.66
JAFIMQ010000019.1	1	375643	375731	Trp	CCA	375680	375695	61.35
JAFIMQ010000019.1	2	193762	193678	Ala	CGC	193726	193714	61.12
JAFIMQ010000019.1	3	165392	165300	Ser	CGA	165355	165344	62.09
JAFIMQ010000020.1	1	566961	567033	Lys	CTT	0	0	75.42
JAFIMQ010000020.1	2	473837	473748	Met	CAT	473801	473784	56.34
JAFIMQ010000020.1	3	18522	18438	Pro	TGG	18485	18474	48.17
JAFIMQ010000021.1	1	3178	3277	Ser	GCT	3215	3233	55.66
JAFIMQ010000021.1	2	190077	189991	Met	CAT	190041	190027	54.80
JAFIMQ010000022.1	1	265272	265380	Ser	AGA	265310	265336	66.38
JAFIMQ010000022.1	2	107406	107305	Ser	GCT	107369	107349	62.68
JAFIMQ010000023.1	1	9980	10069	Arg	ACG	10016	10033	50.14
JAFIMQ010000023.1	2	380617	380534	Ala	AGC	380581	380570	67.10
JAFIMQ010000023.1	3	187840	187755	Met	CAT	187803	187791	54.36
JAFIMQ010000023.1	4	42616	42463	Ser	TGA	42579	42508	23.91
JAFIMQ010000024.1	1	372478	372580	Pro	TGG	372514	372544	60.19
JAFIMQ010000024.1	2	328781	328709	Lys	CTT	0	0	75.42
JAFIMQ010000024.1	3	119285	119202	Glu	CTC	119248	119237	59.91
JAFIMQ010000025.1	1	474516	474608	Gly	CCC	474554	474572	61.57
JAFIMQ010000026.1	1	143927	144010	Trp	CCA	143963	143974	48.76

JAFIMQ010000027.1	1	89307	89236	Thr	AGT	0	0	68.37
JAFIMQ010000028.1	1	114417	114489	Lys	CTT	0	0	75.42
JAFIMQ010000028.1	2	186057	186170	Leu	CAA	186095	186126	56.84
JAFIMQ010000028.1	3	383171	383269	Leu	CAG	383209	383225	54.67
JAFIMQ010000028.1	4	324476	324388	Arg	CCG	324440	324424	48.51
JAFIMQ010000028.1	5	281818	281747	Glu	TTC	0	0	63.63
JAFIMQ010000028.1	6	39732	39653	Pseudo	TGG	0	0	22.38
JAFIMQ010000029.1	1	10480	10564	Ala	TGC	10516	10528	53.94
JAFIMQ010000030.1	1	434706	434602	Leu	AAG	434667	434648	61.86
JAFIMQ010000032.1	1	352149	352064	Ala	CGC	352113	352100	60.95
JAFIMQ010000032.1	2	214884	214814	Pro	TGG	0	0	59.44
JAFIMQ010000033.1	1	108562	108652	Phe	GAA	108599	108616	63.84
JAFIMQ010000033.1	2	424031	424109	His	GTG	424068	424074	45.90
JAFIMQ010000033.1	3	224351	224250	Ala	AGC	224315	224286	63.29
JAFIMQ010000034.1	1	22743	22831	Arg	ACG	22779	22795	51.64
JAFIMQ010000034.1	2	227118	227190	Lys	CTT	0	0	75.42
JAFIMQ010000034.1	3	323698	323768	Gly	GCC	0	0	58.71
JAFIMQ010000034.1	4	293646	293540	Met	CAT	293608	293576	75.74
JAFIMQ010000034.1	5	215181	215070	Leu	CAA	215143	215114	58.72
JAFIMQ010000035.1	1	308331	308236	Arg	CCT	308295	308272	60.41
JAFIMQ010000036.1	1	136612	136683	Thr	AGT	0	0	70.11
JAFIMQ010000038.1	1	110159	110258	Ile	AAT	110197	110222	64.39
JAFIMQ010000038.1	2	203337	203458	Ser	CGA	203370	203423	21.56
JAFIMQ010000038.1	3	171300	171219	Glu	CTC	171263	171254	57.98
JAFIMQ010000041.1	1	221283	221356	Val	AAC	0	0	75.41
JAFIMQ010000041.1	2	221769	221682	Gln	CTG	221731	221717	55.17
JAFIMQ010000041.1	3	160169	160096	Val	AAC	0	0	78.00
JAFIMQ010000044.1	1	223984	223899	Trp	CCA	223946	223935	49.28
JAFIMQ010000045.1	1	4063	3977	Val	TAC	4026	4013	63.97
JAFIMQ010000046.1	1	165321	165403	Glu	CTC	165358	165368	59.64
JAFIMQ010000050.1	1	121187	121259	Val	CAC	0	0	78.92
JAFIMQ010000051.1	1	146335	146258	Leu	CAA	0	0	25.79
JAFIMQ010000054.1	1	189987	190058	Gly	TCC	0	0	66.41
JAFIMQ010000054.1	2	263669	263739	Gly	GCC	0	0	58.71
JAFIMQ010000054.1	3	96340	96270	Gly	GCC	0	0	58.71
JAFIMQ010000055.1	1	181122	181049	Val	AAC	0	0	75.41
JAFIMQ010000058.1	1	215980	215876	Leu	AAG	215941	215922	61.86
JAFIMQ010000060.1	1	235591	235665	Asn	GTT	0	0	70.83
JAFIMQ010000061.1	1	138346	138259	Gln	CTG	138308	138294	61.05
JAFIMQ010000062.1	1	22499	22581	Glu	CTC	22536	22546	60.92
JAFIMQ010000062.1	2	157097	157026	Gly	TCC	0	0	66.41
JAFIMQ010000063.1	1	131564	131481	Ala	AGC	131528	131517	66.59
JAFIMQ010000064.1	1	8168	8241	Val	AAC	0	0	75.41
JAFIMQ010000064.1	2	203102	203032	Gly	GCC	0	0	58.71
JAFIMQ010000066.1	1	169270	169184	Arg	TCG	169234	169220	57.71
JAFIMQ010000068.1	1	105529	105632	Lys	CTT	105566	105596	64.99
JAFIMQ010000068.1	2	152781	152636	Undet	???	0	0	37.29
JAFIMQ010000069.1	1	19283	19370	Tyr	GTA	19322	19334	65.81
JAFIMQ010000070.1	1	82754	82826	Lys	CTT	0	0	75.42

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

JAFIMQ010000071.1	1	185481	185559	His	GTG	185518	185524	48.49
JAFIMQ010000075.1	1	30678	30762	Ala	TGC	30714	30726	53.94
JAFIMQ010000075.1	2	122697	122591	Ser	CGA	122659	122635	60.35
JAFIMQ010000075.1	3	31090	30986	Met	CAT	31052	31022	76.54
JAFIMQ010000079.1	1	100339	100454	Lys	TTT	100377	100418	61.28
JAFIMQ010000079.1	2	92889	92785	Leu	AAG	92850	92831	61.86
JAFIMQ010000085.1	1	57487	57388	Ile	AAT	57449	57424	61.84
JAFIMQ010000092.1	1	101875	101805	Gly	GCC	0	0	58.71
JAFIMQ010000093.1	1	99677	99830	Arg	GCG	99713	99795	40.02
JAFIMQ010000094.1	1	44529	44661	Undet	???	0	0	22.80
JAFIMQ010000094.1	2	44837	44917	Asp	GTC	44874	44882	63.74
JAFIMQ010000100.1	1	70722	70648	Asn	GTT	0	0	70.83
JAFIMQ010000102.1	1	90725	90653	Lys	CTT	0	0	75.42
JAFIMQ010000102.1	2	28362	28279	Ala	AGC	28326	28315	66.46
JAFIMQ010000103.1	1	9139	9210	Met	CAT	0	0	70.01
JAFIMQ010000103.1	2	48504	48434	Gly	GCC	0	0	58.71
JAFIMQ010000105.1	1	2143	2045	Arg	TCT	2107	2081	65.58
JAFIMQ010000106.1	1	61188	61270	Thr	TGT	61224	61234	66.18
JAFIMQ010000108.1	1	66973	67043	Gly	GCC	0	0	58.71
JAFIMQ010000108.1	2	81266	81336	Gly	GCC	0	0	58.71
JAFIMQ010000113.1	1	52835	52747	Gln	TTG	52798	52782	59.11
JAFIMQ010000117.1	1	15042	15113	Thr	AGT	0	0	68.37
JAFIMQ010000118.1	1	3182	3254	Val	CAC	0	0	76.81
JAFIMQ010000118.1	2	21555	21477	His	GTG	21518	21512	48.49
JAFIMQ010000120.1	1	9317	9181	Ser	CGA	9279	9216	31.15
JAFIMQ010000133.1	1	3451	3521	Arg	ACG	0	0	50.29
JAFIMQ010000133.1	2	10120	10190	Pseudo	TCT	0	0	45.82
JAFIMQ010000133.1	3	19266	19336	Asn	GTT	0	0	58.07
JAFIMQ010000133.1	4	19357	19428	Lys	TTT	0	0	41.28
JAFIMQ010000133.1	5	19660	19730	Gly	TCC	0	0	39.50
JAFIMQ010000133.1	6	20925	20997	Pseudo	GTC	0	0	37.80
JAFIMQ010000133.1	7	21131	21201	SeC	TCA	0	0	49.80
JAFIMQ010000133.1	8	21420	21492	Pseudo	TGG	0	0	33.17
JAFIMQ010000133.1	9	21513	21599	Pseudo	TGA	0	0	33.86
JAFIMQ010000133.1	10	22616	22688	Val	TAC	0	0	53.91
JAFIMQ010000133.1	11	24834	24905	Pseudo	GAT	0	0	29.76
JAFIMQ010000133.1	12	30219	30289	Thr	TGT	0	0	34.60
JAFIMQ010000133.1	13	30339	30410	Glu	TTC	0	0	54.15
JAFIMQ010000133.1	14	30418	30489	Met	CAT	0	0	37.19
JAFIMQ010000133.1	15	30559	30631	Pseudo	CAT	0	0	37.21
JAFIMQ010000133.1	16	30660	30742	Pseudo	TAA	0	0	28.30
JAFIMQ010000133.1	17	30817	30888	Ala	TGC	0	0	54.84
JAFIMQ010000133.1	18	30912	30984	Phe	GAA	0	0	50.51
JAFIMQ010000133.1	19	31070	31142	Pseudo	TTG	0	0	34.85
JAFIMQ010000133.1	20	31286	31358	His	GTG	0	0	40.19
JAFIMQ010000133.1	21	31429	31501	Pseudo	CAT	0	0	39.25
JAFIMQ010000133.1	22	36487	36559	Val	TAC	0	0	49.28
JAFIMQ010000133.1	23	42472	42542	Pseudo	TCT	0	0	45.82
JAFIMQ010000180.1	1	14668	14739	Thr	AGT	0	0	68.37

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Additional File 10

Sequence		tRNA	Bounds	tRNA	Anti	Intron Bounds		Cove
Name	tRNA #	Begin	End	Type	Codon	Begin	End	Score
--------	------	----	------	----	-----	-----	----	------
JAFIMR010000001.1	1	100998	101093	Arg	CCT	101034	101057	60.41
JAFIMR010000001.1	2	802363	802451	Gln	CTG	802401	802416	59.00
JAFIMR010000001.1	3	1694027	1694098	Thr	AGT	0	0	68.37
JAFIMR010000001.1	4	2014143	2014232	Pro	AGG	2014179	2014196	67.69
JAFIMR010000001.1	5	2248615	2248543	Arg	CCG	0	0	64.28
JAFIMR010000001.1	6	1464658	1464560	Leu	TAA	1464621	1464604	62.36
JAFIMR010000001.1	7	1044381	1044310	Asp	GTC	0	0	72.49
JAFIMR010000001.1	8	541938	541834	Leu	AAG	541899	541880	61.86
JAFIMR010000002.1	1	402698	402800	Pro	TGG	402734	402764	60.19
JAFIMR010000002.1	2	655315	655405	Phe	GAA	655352	655369	63.84
JAFIMR010000002.1	3	970984	971062	His	GTG	971021	971027	45.90
JAFIMR010000002.1	4	2185597	2185676	His	GTG	2185634	2185641	51.84
JAFIMR010000002.1	5	1366990	1366845	Undet	???	0	0	34.69
JAFIMR010000002.1	6	1333418	1333327	Ser	AGA	1333380	1333371	69.74
JAFIMR010000002.1	7	1333319	1333220	Ile	AAT	1333281	1333256	64.53
JAFIMR010000002.1	8	771039	770938	Ala	AGC	771003	770974	63.29
JAFIMR010000002.1	9	386299	386227	Lys	CTT	0	0	75.42
JAFIMR010000002.1	10	171317	171234	Glu	CTC	171280	171269	59.91
JAFIMR010000003.1	1	794586	794668	Glu	CTC	794623	794633	60.92
JAFIMR010000003.1	2	1741276	1741384	Ser	AGA	1741314	1741340	66.38
JAFIMR010000003.1	3	2066453	2066523	Gly	GCC	0	0	58.71
JAFIMR010000003.1	4	1589902	1589801	Ser	GCT	1589865	1589845	62.68
JAFIMR010000003.1	5	1131358	1131286	Lys	CTT	0	0	75.42
JAFIMR010000003.1	6	929554	929483	Gly	TCC	0	0	66.41
JAFIMR010000003.1	7	324698	324612	Val	TAC	324661	324648	63.97
JAFIMR010000004.1	1	204776	204867	Ser	AGA	204814	204823	69.13
JAFIMR010000004.1	2	665097	665168	Thr	AGT	0	0	70.11
JAFIMR010000004.1	3	848439	848529	Phe	GAA	848476	848493	63.84
JAFIMR010000004.1	4	1593379	1593449	Gly	GCC	0	0	58.71
JAFIMR010000004.1	5	1844054	1844140	Arg	TCG	1844090	1844104	57.71
JAFIMR010000004.1	6	1788326	1788253	Val	AAC	0	0	75.41
JAFIMR010000004.1	7	983613	983508	Ile	AAT	983575	983544	61.75
JAFIMR010000004.1	8	209603	209513	Asp	GTC	209566	209548	66.22
JAFIMR010000005.1	1	1457984	1458079	Pro	CGG	1458020	1458043	51.00
JAFIMR010000005.1	2	1462423	1462493	Gly	GCC	0	0	58.71
JAFIMR010000005.1	3	1462844	1462951	Ser	AGA	1462882	1462907	57.38
JAFIMR010000005.1	4	1756388	1756271	Ala	TGC	1756350	1756307	59.43
JAFIMR010000005.1	5	1175002	1174895	Leu	TAG	1174965	1174937	49.38
JAFIMR010000005.1	6	563466	563382	Tyr	GTA	563429	563418	63.18
JAFIMR010000006.1	1	53987	54057	Gly	GCC	0	0	58.71
JAFIMR010000006.1	2	865675	865831	Arg	ACG	865711	865796	33.50
JAFIMR010000006.1	3	1017271	1017179	Phe	GAA	1017234	1017215	64.87
JAFIMR010000006.1	4	998611	998513	Leu	CAG	998573	998557	56.71

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

JAFIMR010000006.1	5	69293	69204	Arg	ACG	69257	69240	53.20
JAFIMR010000007.1	1	297736	297814	Glu	CTC	297773	297779	60.26
JAFIMR010000007.1	2	334555	334636	Gly	CCC	334591	334601	55.52
JAFIMR010000007.1	3	685694	685767	Val	AAC	0	0	76.92
JAFIMR010000007.1	4	686159	686240	Glu	CTC	686196	686205	60.29
JAFIMR010000007.1	5	719191	719292	Ser	CGA	719228	719248	60.11
JAFIMR010000007.1	6	777839	777937	Arg	TCT	777875	777901	62.78
JAFIMR010000007.1	7	1040206	1040299	Ile	TAT	1040243	1040263	59.25
JAFIMR010000008.1	1	481010	480920	Arg	TCG	480974	480956	54.51
JAFIMR010000009.1	1	137397	137480	Glu	CTC	137434	137445	61.19
JAFIMR010000009.1	2	152527	152614	Glu	TTC	152565	152579	57.05
JAFIMR010000009.1	3	295842	295914	Lys	CTT	0	0	75.42
JAFIMR010000009.1	4	381734	381833	Ile	AAT	381772	381797	61.84
JAFIMR010000009.1	5	629040	629112	Lys	CTT	0	0	75.42
JAFIMR010000009.1	6	701206	701319	Leu	CAA	701244	701275	60.03
JAFIMR010000009.1	7	898461	898559	Leu	CAG	898499	898515	54.67
JAFIMR010000009.1	8	839827	839739	Arg	CCG	839791	839775	48.51
JAFIMR010000009.1	9	797150	797079	Glu	TTC	0	0	63.63
JAFIMR010000009.1	10	546701	546622	Pseudo	???	0	0	22.51
JAFIMR010000009.1	11	216553	216470	Ala	AGC	216517	216506	66.46
JAFIMR010000010.1	1	628930	629017	Gln	CTG	628968	628982	61.05
JAFIMR010000010.1	2	873331	873430	Ile	AAT	873369	873394	64.39
JAFIMR010000010.1	3	965954	966075	Ser	CGA	965987	966040	21.56
JAFIMR010000010.1	4	933908	933827	Glu	CTC	933871	933862	59.87
JAFIMR010000011.1	1	62882	62953	Thr	AGT	0	0	70.11
JAFIMR010000011.1	2	731082	730990	Gly	CCC	731044	731026	61.57
JAFIMR010000012.1	1	514382	514452	Gly	GCC	0	0	58.71
JAFIMR010000012.1	2	756141	756218	Leu	CAA	0	0	25.79
JAFIMR010000012.1	3	916390	916291	Ser	GCT	916353	916335	55.02
JAFIMR010000012.1	4	280908	280820	Trp	CCA	280871	280856	61.35
JAFIMR010000013.1	1	192885	192975	Asp	GTC	192922	192940	66.08
JAFIMR010000013.1	2	201493	201340	Leu	CAA	201456	201381	25.56
JAFIMR010000014.1	1	509104	509310	Pro	CGG	509140	509277	32.57
JAFIMR010000014.1	2	694520	694755	Gly	ACC	694557	694719	21.33
JAFIMR010000015.1	1	75782	75872	Phe	GAA	75819	75836	64.47
JAFIMR010000015.1	2	227543	227635	Asp	GTC	227580	227600	63.93
JAFIMR010000015.1	3	361022	360935	Gln	CTG	360984	360970	59.64
JAFIMR010000016.1	1	314270	314341	Thr	AGT	0	0	68.37
JAFIMR010000016.1	2	692026	692124	Arg	TCT	692062	692088	65.58
JAFIMR010000017.1	1	153200	153285	Trp	CCA	153238	153249	49.28
JAFIMR010000017.1	2	574078	574214	Glu	CTC	574114	574181	28.36
JAFIMR010000017.1	3	34611	34540	Met	CAT	0	0	70.01
JAFIMR010000018.1	1	65822	65920	Ile	AAT	65860	65884	63.48
JAFIMR010000018.1	2	146686	146756	Pro	TGG	0	0	61.01
JAFIMR010000018.1	3	267989	268106	Ala	TGC	268027	268070	56.81
JAFIMR010000018.1	4	489411	489338	Val	AAC	0	0	75.41
JAFIMR010000019.1	1	549104	549020	Tyr	GTA	549067	549056	62.41
JAFIMR010000019.1	2	533079	533008	Gly	TCC	0	0	66.80
JAFIMR010000020.1	1	448491	448387	Leu	AAG	448452	448433	61.86

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

JAFIMR010000021.1	1	568711	568817	Met	CAT	568749	568781	76.17
JAFIMR010000021.1	2	638816	638744	Lys	CTT	0	0	75.42
JAFIMR010000021.1	3	618609	618522	Arg	TCG	618573	618558	60.59
JAFIMR010000021.1	4	569158	569077	Glu	CTC	569121	569112	60.65
JAFIMR010000022.1	1	182681	182774	Arg	CCT	182718	182738	58.83
JAFIMR010000022.1	2	220534	220615	Glu	TTC	220571	220580	56.92
JAFIMR010000023.1	1	134337	134424	Gln	CTG	134375	134389	55.17
JAFIMR010000023.1	2	196307	196380	Val	AAC	0	0	78.00
JAFIMR010000023.1	3	535980	536050	Gly	GCC	0	0	58.71
JAFIMR010000023.1	4	443174	443103	Gly	TCC	0	0	66.41
JAFIMR010000023.1	5	369552	369482	Gly	GCC	0	0	58.71
JAFIMR010000023.1	6	134823	134750	Val	AAC	0	0	75.41
JAFIMR010000024.1	1	164262	164348	Cys	GCA	164298	164312	60.35
JAFIMR010000024.1	2	416148	416257	Asn	GTT	416186	416221	60.39
JAFIMR010000024.1	3	352591	352500	Cys	GCA	352555	352536	63.44
JAFIMR010000025.1	1	111813	111902	Met	CAT	111849	111866	56.34
JAFIMR010000025.1	2	567263	567347	Pro	TGG	567300	567311	48.17
JAFIMR010000025.1	3	18618	18546	Lys	CTT	0	0	75.42
JAFIMR010000026.1	1	582435	582351	Ala	TGC	582399	582387	54.58
JAFIMR010000027.1	1	38323	38422	Arg	ACG	38359	38386	48.30
JAFIMR010000027.1	2	408324	408241	Ala	AGC	408288	408277	67.10
JAFIMR010000027.1	3	216138	216053	Met	CAT	216101	216089	54.36
JAFIMR010000027.1	4	70986	70833	Ser	TGA	70949	70878	23.91
JAFIMR010000029.1	1	538102	538194	Thr	TGT	538138	538158	70.61
JAFIMR010000030.1	1	462061	462139	His	GTG	462098	462104	48.49
JAFIMR010000030.1	2	150548	150476	Val	CAC	0	0	78.92
JAFIMR010000031.1	1	369316	369390	Asn	GTT	0	0	70.83
JAFIMR010000031.1	2	35301	35221	Asp	GTC	35264	35256	65.66
JAFIMR010000032.1	1	439504	439351	Arg	GCG	439468	439386	40.02
JAFIMR010000032.1	2	166224	166142	Glu	CTC	166187	166177	59.64
JAFIMR010000033.1	1	336882	336970	Gln	TTG	336919	336935	57.59
JAFIMR010000034.1	1	282184	282097	Tyr	GTA	282145	282133	65.81
JAFIMR010000035.1	1	350868	350783	Ala	CGC	350832	350819	60.95
JAFIMR010000035.1	2	214926	214856	Pro	TGG	0	0	61.43
JAFIMR010000036.1	1	146611	146507	Leu	AAG	146572	146553	61.86
JAFIMR010000037.1	1	117067	117171	Leu	AAG	117106	117125	61.86
JAFIMR010000037.1	2	109615	109500	Lys	TTT	109577	109536	61.28
JAFIMR010000039.1	1	105440	105513	Val	AAC	0	0	78.00
JAFIMR010000039.1	2	178086	178004	Gln	TTG	178049	178039	57.55
JAFIMR010000039.1	3	79975	79876	Ile	AAT	79937	79912	64.38
JAFIMR010000039.1	4	79580	79499	Ala	AGC	79544	79535	66.20
JAFIMR010000040.1	1	78605	78711	Met	CAT	78643	78675	79.84
JAFIMR010000040.1	2	156971	157082	Leu	CAA	157009	157038	58.72
JAFIMR010000040.1	3	145068	144996	Lys	CTT	0	0	75.42
JAFIMR010000040.1	4	45645	45575	Gly	GCC	0	0	58.71
JAFIMR010000042.1	1	326282	326211	Glu	CTC	0	0	61.21
JAFIMR010000042.1	2	260380	260307	Val	AAC	0	0	75.41
JAFIMR010000043.1	1	124030	124114	Ala	CGC	124066	124078	60.49
JAFIMR010000043.1	2	152374	152466	Ser	CGA	152411	152422	62.09

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

JAFIMR010000045.1	1	244909	244995	Met	CAT	244945	244959	54.80
JAFIMR010000046.1	1	158898	158981	Trp	CCA	158934	158945	50.62
JAFIMR010000048.1	1	180379	180296	Ala	AGC	180343	180332	67.23
JAFIMR010000049.1	1	95880	95799	Ala	AGC	95844	95835	66.20
JAFIMR010000052.1	1	44749	44837	Arg	ACG	44785	44801	51.78
JAFIMR010000052.1	2	2688	2586	Tyr	GTA	2651	2622	60.09
JAFIMR010000055.1	1	4451	4380	Thr	AGT	0	0	68.37
JAFIMR010000056.1	1	41901	42046	Undet	???	0	0	36.36
JAFIMR010000056.1	2	89268	89165	Lys	CTT	89231	89201	64.99
JAFIMR010000057.1	1	202819	202919	Ser	GCT	202856	202875	56.99
JAFIMR010000057.1	2	176060	175956	Met	CAT	176022	175992	73.74
JAFIMR010000058.1	1	187281	187366	Thr	CGT	187317	187330	70.51
JAFIMR010000058.1	2	56475	56378	Ile	AAT	56437	56414	63.36
JAFIMR010000059.1	1	14479	14551	Val	CAC	0	0	76.81
JAFIMR010000059.1	2	178839	178765	Asn	GTT	0	0	70.83
JAFIMR010000059.1	3	32826	32748	His	GTG	32789	32783	48.49
JAFIMR010000061.1	1	134397	134506	Asn	GTT	134435	134470	54.65
JAFIMR010000064.1	1	45215	45347	Undet	???	0	0	22.80
JAFIMR010000064.1	2	45528	45608	Asp	GTC	45565	45573	63.74
JAFIMR010000064.1	3	151758	151688	Gly	GCC	0	0	58.71
JAFIMR010000069.1	1	62240	62170	Gly	GCC	0	0	58.71
JAFIMR010000073.1	1	3605	3517	Arg	ACG	3569	3553	51.64
JAFIMR010000075.1	1	38228	38332	Met	CAT	38266	38296	75.90
JAFIMR010000075.1	2	38640	38556	Ala	TGC	38604	38592	53.94
JAFIMR010000077.1	1	58672	58778	Ser	CGA	58710	58734	60.35
JAFIMR010000089.1	1	2858	2928	Thr	TGT	0	0	34.60
JAFIMR010000089.1	2	2978	3049	Glu	TTC	0	0	54.15
JAFIMR010000089.1	3	3057	3128	Met	CAT	0	0	37.19
JAFIMR010000089.1	4	3198	3270	Pseudo	CAT	0	0	37.21
JAFIMR010000089.1	5	3299	3381	Pseudo	TAA	0	0	28.30
JAFIMR010000089.1	6	3456	3527	Ala	TGC	0	0	54.84
JAFIMR010000089.1	7	3551	3623	Phe	GAA	0	0	50.51
JAFIMR010000089.1	8	3709	3781	Pseudo	TTG	0	0	34.85
JAFIMR010000089.1	9	3925	3997	His	GTG	0	0	40.19
JAFIMR010000089.1	10	4068	4140	Pseudo	CAT	0	0	39.25
JAFIMR010000089.1	11	9127	9199	Val	TAC	0	0	49.28
JAFIMR010000089.1	12	15113	15183	Pseudo	TCT	0	0	45.82
JAFIMR010000089.1	13	18910	18980	Arg	ACG	0	0	50.29
JAFIMR010000089.1	14	25581	25651	Pseudo	TCT	0	0	45.82
JAFIMR010000089.1	15	34727	34797	Asn	GTT	0	0	58.07
JAFIMR010000089.1	16	34818	34889	Lys	TTT	0	0	41.28
JAFIMR010000089.1	17	35121	35191	Gly	TCC	0	0	39.50
JAFIMR010000089.1	18	36386	36458	Pseudo	GTC	0	0	37.80
JAFIMR010000089.1	19	36592	36662	SeC	TCA	0	0	49.80
JAFIMR010000089.1	20	36881	36953	Pseudo	TGG	0	0	33.17
JAFIMR010000089.1	21	36974	37060	Pseudo	TGA	0	0	33.86
JAFIMR010000089.1	22	38077	38149	Val	TAC	0	0	53.91
JAFIMR010000089.1	23	40280	40351	Pseudo	GAT	0	0	29.76

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 11

This File is too large to be displayed.

Additional File 12

This File is too large to be displayed.

Additional File 13

This File is too large to be displayed.

Additional File 14

This File is too large to be displayed.

Additional File 15

This File is too large to be displayed.

Additional File 16

This File is too large to be displayed.

Additional File 17

This File is too large to be displayed.

Additional File 18

This File is too large to be displayed.

Additional File 19

This File is too large to be displayed.

Additional File 20

This File is too large to be displayed.

Additional File 21

This File is too large to be displayed.

Additional File 22

This File is too large to be displayed.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Additional File 23

Gene ID	K-number	Type	Domain	Specific annotation
JN550g4767.t1	K09464	Eukaryotic	Basic leucine zipper (bZIP)	AP-1(-like) components, Fungal regulators
JN550g1939.t1	K09043	Eukaryotic	Basic leucine zipper (bZIP)	AP-1(-like) components, Fungal regulators
JN550g4788.t1	K09043	Eukaryotic	Basic leucine zipper (bZIP)	AP-1(-like) components, Fungal regulators
JN550g5010.t1	K09043	Eukaryotic	Basic leucine zipper (bZIP)	AP-1(-like) components, Fungal regulators
JN550g5088.t1	K09043	Eukaryotic	Basic leucine zipper (bZIP)	AP-1(-like) components, Fungal regulators
JN550g8375.t1	K09043	Eukaryotic	Basic leucine zipper (bZIP)	AP-1(-like) components, Fungal regulators
JN550g10328.t1	K09043	Eukaryotic	Basic leucine zipper (bZIP)	AP-1(-like) components, Fungal regulators
JN550g1888.t1	K09051	Eukaryotic	Basic leucine zipper (bZIP)	AP-1(-like) components, CRE-BP/ATF
JN550g9961.t1	K16230	Eukaryotic	Basic leucine zipper (bZIP)	CREB
JN550g11295.t1	K06648	Eukaryotic	Basic leucine zipper (bZIP)	ZIP only
JN550g8404.t1	K21642	Eukaryotic	Basic leucine zipper (bZIP)	ZIP only
JN550g1270.t1	K21451	Eukaryotic	Basic leucine zipper (bZIP)	ZIP only
JN550g6036.t1	K21452	Eukaryotic	Basic leucine zipper (bZIP)	ZIP only
JN550g4280.t1	K09102	Eukaryotic	Basic helix-loop-helix (bHLH)	INO
JN550g12841.t1	K22484	Eukaryotic	Basic helix-loop-helix (bHLH)	HLH domain only
JN550g8694.t1	K09175	Eukaryotic	Other basic domains	RF-X
JN550g1239.t1	K09184	Eukaryotic	Zinc finger	Cys4 GATA-factors
JN550g4890.t1	K09184	Eukaryotic	Zinc finger	Cys4 GATA-factors
JN550g5136.t1	K09184	Eukaryotic	Zinc finger	Cys4 GATA-factors
JN550g4563.t1	K09202	Eukaryotic	Zinc finger	Cys2His2 SP/KLF family and related proteins
JN550g1037.t1	K09467	Eukaryotic	Zinc finger	Cys2His2 metabolic regulators in fungi
JN550g126.t1	K09191	Eukaryotic	Zinc finger	Cys2His2 others
JN550g8198.t1	K07466	Eukaryotic	Zinc finger	Cys2His2 others
JN550g4159.t1	K19487	Eukaryotic	Zinc finger	Cys2His2 others
JN550g6632.t1	K21455	Eukaryotic	Zinc finger	Cys2His2 others
JN550g12709.t1	K21543	Eukaryotic	Zinc finger	Cys2His2 others
JN550g10772.t1	K21544	Eukaryotic	Zinc finger	Cys2His2 others
JN550g12722.t1	K21545	Eukaryotic	Zinc finger	Cys2His2 others
JN550g10621.t1	K11304	Eukaryotic	Zinc finger	Cys2HisCys zinc factors
JN550g7884.t1	K15263	Eukaryotic	Zinc finger	Cys2HisCys zinc factors
JN550g3264.t1	K14960	Eukaryotic	Zinc finger	CXXC CpG-binding proteins
JN550g1517.t1	K00558	Eukaryotic	Zinc finger	CXXC CpG-binding proteins
JN550g4596.t1	K00558	Eukaryotic	Zinc finger	CXXC CpG-binding proteins
JN550g12516.t1	K00558	Eukaryotic	Zinc finger	CXXC CpG-binding proteins
JN550g2137.t1	K09241	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g3158.t1	K09241	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g11523.t1	K09241	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g398.t1	K09242	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g3944.t1	K09242	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

JN550g2127.t1	K09246	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g6953.t1	K09246	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g8555.t1	K09246	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g2495.t1	K09248	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g4358.t1	K09248	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g6959.t1	K09248	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g9109.t1	K09248	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g10141.t1	K09248	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g11041.t1	K09248	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g3802.t1	K21547	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g4246.t1	K21547	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g4702.t1	K21547	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g9396.t1	K21547	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g6689.t1	K21632	Eukaryotic	Zinc finger	Cys6 metabolic regulators in fungi
JN550g1902.t1	K09250	Eukaryotic	Zinc finger	Other zinc fingers
JN550g10162.t1	K09250	Eukaryotic	Zinc finger	Other zinc fingers
JN550g13101.t1	K09250	Eukaryotic	Zinc finger	Other zinc fingers
JN550g8735.t1	K09313	Eukaryotic	Helix-turn-helix	Homeo domain CUT
JN550g231.t1	K09413	Eukaryotic	Helix-turn-helix	Fork head/winged helix other regulators
JN550g12198.t1	K24664	Eukaryotic	Helix-turn-helix	Fork head/winged helix other regulators
JN550g8881.t1	K09422	Eukaryotic	Helix-turn-helix	Tryptophan clusters Myb, Myb-factors
JN550g8881.t1	K09425	Eukaryotic	Helix-turn-helix	Tryptophan clusters Myb, Myb-like factors
JN550g10057.t1	K09448	Eukaryotic	Helix-turn-helix	TEA domain
JN550g1588.t1	K12412	Eukaryotic	beta-Scaffold factors with minor groove contacts	MADS-box regulators of differentiation, Yeast regulators
JN550g4838.t1	K09265	Eukaryotic	beta-Scaffold factors with minor groove contacts	MADS-box regulators of differentiation, Yeast regulators
JN550g4141.t1	K03120	Eukaryotic	beta-Scaffold factors with minor groove contacts	TATA-binding proteins
JN550g11055.t1	K09272	Eukaryotic	beta-Scaffold factors with minor groove contacts	HMG2-related
JN550g4264.t1	K22483	Eukaryotic	beta-Scaffold factors with minor groove contacts	Other HMG box factors
JN550g7759.t1	K09274	Eukaryotic	beta-Scaffold factors with minor groove contacts	Other HMG box factors
JN550g10477.t1	K08064	Eukaryotic	beta-Scaffold factors with minor groove contacts	Heteromeric CCAAT factors
JN550g7876.t1	K08065	Eukaryotic	beta-Scaffold factors with minor groove contacts	Heteromeric CCAAT factors
JN550g6664.t1	K08066	Eukaryotic	beta-Scaffold factors with minor groove contacts	Heteromeric CCAAT factors
JN550g4417.t1	K09275	Eukaryotic	beta-Scaffold factors with minor groove contacts	Grainyhead
JN550g1032.t1	K12769	Eukaryotic	beta-Scaffold factors with minor groove contacts	MYRF
JN550g4026.t1	K22758	Eukaryotic	Other transcription factors	Others
JN550g11059.t1	K11215	Eukaryotic	Other transcription factors	Others
JN550g604.t1	K21631	Eukaryotic	Other transcription factors	Others
JN550g13866.t1	K21631	Eukaryotic	Other transcription factors	Others
JN550g13957.t1	K21631	Eukaryotic	Other transcription factors	Others

JN550g3547.t1	K12763	Eukaryotic	Other transcription factors	Others
JN550g8201.t1	K12763	Eukaryotic	Other transcription factors	Others
JN550g9201.t1	K05527	Prokaryotic	Other transcription factors	Others
JN550g6644.t1	K03707	Prokaryotic	Other transcription factors	Others
JN550g2349.t1	K07734	Prokaryotic	Other transcription factors	Others

Gene ID	K-number	Type	Specific annotation
JN550g7831.t1	K11021	Type III toxins: Intracellular toxins	TccC-type insecticidal toxin

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 24

This File is too large to be displayed.

Additional File 25

This File is too large to be displayed.

Additional File 26

This File is too large to be displayed.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 27

Protein	Accession	*W. moseri* CBS Gene ID	*W. moseri* TUCIM 5827 Gene ID	*W. moseri* TUCIM 5799 Gene ID
Hexokinase	EC:2.7.1.1	JN550g1159.t1, JN550g3215.t1, JN550g3392.t1, JN550g4061.t1, JN550g4763.t1, JN550g5430.t1	JX266g2067.t1, JX266g7681.t1, JX266g7905.t1, JX266g10351.t1, JX266g10546.t1, JX266g12239.t1	JX265g148.t1, JX265g988.t1, JX265g1828.t1, JX265g7829.t1, JX265g10550.t1, JX265g12577.t1
Glucokinase	EC:2.7.1.2	0	0	0
Glucose-6-P isomerase	EC:5.3.1.9	JN550g2462.t1	JX266g1113.t1	JX265g3719.t1
6-Phosphofructokinase	EC:2.7.1.11	JN550g1435.t1	JX266g8822.t1	JX265g8460.t1
Fructose-bisP aldolase	EC:4.1.2.13	JN550g4314.t1, JN550g7570.t1, JN550g8229.t1, JN550g11550.t1	JX266g538.t1, JX266g4192.t1, JX266g7191.t1, JX266g13488.t1	JX265g2230.t1, JX265g4508.t1, JX265g6100.t1, JX265g9144.t1
Triosephosphate isomerase	EC:5.3.1.1	JN550g5424.t1, JN550g7567.t1	JX266g4189.t1, JX266g10345.t1	JX265g1822.t1, JX265g2227.t1
Glyceraldehyde-3-P DH	EC:1.2.1.12	JN550g3126.t1	JX266g206.t1	JX265g6541.t1
Phosphoglycerate kinase	EC:2.7.2.3	JN550g11486.t1	JX266g7256.t1	JX265g9080.t1
Phosphoglycerate mutase	EC:5.4.2.12	JN550g11656.t1	JX266g13133.t1	JX265g852.t1
PEP hydratase	EC:4.2.1.11	JN550g10481.t1	JX266g9525.t1	JX265g10229.t1
Pyruvate kinase	EC:2.7.1.40	JN550g12018.t1	JX266g8377.t1	JX265g12477.t1
Pyruvate DH complex	EC:1.2.4.1	JN550g10570.t1, JN550g739.t1	JX266g8236.t1, JX266g13747.t1	JX265g483.t1, JX265g11260.t1
Citrate synthase	EC:4.1.3.7	JN550g4898.t1, JN550g6168.t1, JN550g11399.t1, JN550g12418.t1, JN550g12419.t1,	JX266g2878.t1, JX266g8903.t1, JX266g9691.t1, JX266g10830.t1, JX266g10831.t1	JX265g2942.t1, JX265g8211.t1, JX265g8378.t1, JX265g5149.t1, JX265g5150.t1
Pyruvate carboxylase	EC:6.4.1.1	JN550g6224.t1	JX266g8304.t1	1
PEP carboxykinase	EC:4.1.1.49	JN550g10785.t1	JX266g5390.t1	JX265g1280.t1
Malate dehydrogenase	EC:1.1.1.37	JN550g2922.t1, JN550g7143.t1, JN550g8734.t1, JN550g11372.t1, JN550g12911.t1, JN550g13842.t1	JX266g1531.t1, JX266g7948.t1, JX266g9018.t1, JX266g11793.t1, JX266g6408.t1, JX266g8876.t1	JX265g4138.t1, JX265g9631.t1, JX265g13587.t1, JX265g13990.t1, JX265g788.t1, JX265g8406.t1
Oxaloacetase	EC:3.7.1.1	0	0	0
Glucose oxidase	EC:1.1.3.4	JN550g13020.t1, JN550g13653.t1	JX266g12607.t1, JX266g13506.t1	JX265g737.t1, JX265g10203.t1
Gluconolactonase	EC:3.1.1.17	JN550g373.t1	JX266g1036.t1	JX265g2564.t1
Trehalose P synthase	EC:2.4.1.15	JN550g10750.t1, JN550g3330.t1, JN550g11833.t1	JX266g8942.t1	JX265g9552.t1,
Trehalose phosphatase	EC:3.1.3.12	0	0	0

6-Phosphofructo-2-kinase	EC:2.7.1.105	JN550g9253.t1, JN550g7727.t1	JX266g11121.t1, JX266g1690.t1	JX265g5089.t1, JX265g3978.t1
Aconitate hydratase	EC:4.2.1.36	JN550g2797.t1, JN550g11799.t1	JX266g6213.t1, JX266g6963.t1	JX265g9169.t1, JX265g6769.t1
Aconitate decarboxylase	EC:4.1.1.6	0	0	0
b-Fructofuranosidase	EC:3.2.1.26	0	0	0
Mannitol-1P DH	EC:1.1.1.17	JN550g6958.t1, JN550g7916.t1, JN550g12515.t1, JN550g13023.t1	JX266g4385.t1	JX265g9044.t1, JX265g9352.t1, JX265g12247.t1
Isocitrate DH (NADP)	EC:1.1.1.42	JN550g1757.t1, JN550g11639.t1	JX266g12694.t1	JX265g500.t1, JX265g2076.t1
Oxoglutarate DH	EC:1.2.4.2	JN550g12602.t1	JX266g8554.t1	JX265g5303.t1
Succinate DH (NADP)	EC:1.3.5.1	JN550g1305.t1	JX266g4798.t1	JX265g12209.t1
Fumarate hydratase	EC:4.2.1.2	JN550g1594.t1	JX266g1399.t1	JX265g665.t1
Glc transporter		JN550g1335.t1, JN550g1743.t1, JN550g1264.t1, JN550g2950.t1, JN550g7540.t1, JN550g9198.t1, JN550g9845.t1, JN550g10198.t1, JN550g11140.t1, JN550g11835.t1, JN550g12501.t1	JX266g8176.t1, JX266g3546.t1, JX266g4161.t1, JX266g4828.t1, JX266g6249.t1, JX266g7807.t1, JX266g8739.t1, JX266g10324.t1, JX266g11640.t1, JX266g11766.t1, JX266g12217.t1, JX266g12708.t1	JX265g42.t1, JX265g514.t1, JX265g1471.t1, JX265g2199.t1, JX265g2771.t1, JX265g4949.t1, JX265g9205.t1, JX265g7459.t1, JX265g9338.t1, JX265g10595.t1, JX265g10674.t1, JX265g12179.t1, JX265g13657.t1
Citrate/malate antiporter		0	0	0
succinate-fumarate transporter		JN550g8827.t1	JX266g5799.t1	JX265g8779.t1
Citrate transporter		0	0	0
Sucrose transporter		JN550g67.t1, JN550g6186.t1, JN550g6296.t1	JX266g728.t1, JX266g8232.t1, JX266g8579.t1	JX265g5408.t1, JX265g7551.t1, JX265g10522.t1
Fructose transporter		0	0	0
Alternative oxidase		JN550g976.t1, JN550g6275.t1, JN550g3325.t1, JN550g9818.t1, JN550g12703.t1	JX266g298.t1, JX266g6686.t1, JX266g13408.t1, JX266g1881.t1, JX266g8253.t1	JX265g5518.t1, JX265g7765.t1, JX265g12910.t1, JX265g5319.t1, JX265g11504.t1

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Additional File 28

antiSMASH 4 results

Similar known cluster	MIBiG ID	Type	Cluster No in CBS	Location in CBS	similarity CBS to MIBIG	Cluster No in TUCIM 5827	Location in TUCIM 5827	similarity TUCIM 5827 to MIBIG	Cluster No in TUCIM 5799	Location in TUCIM 5799	similarity TUCIM 5799 to MIBIG
Depudecin	BGC0000046	T1pks	Cluster 8	JAFEVA010000001; 1474369-1521786 nt	50%	Cluster 139	JAFIMQ010000064; 37595-85010 nt	50%	Cluster 31	JAFIMR010000004; 1711506-1758920 nt	50%
Pestheic acid	BGC0000121	T1pks	Cluster 28	JAFEVA010000005; 496849-542699 nt	25%	Cluster 16	JAFIMQ010000003; 376620-422470 nt	25%	Cluster 47	JAFIMR010000006; 795568-841418 nt	25%
Nivalenol	BGC000127	Terpene-Nrps	Cluster 45	JAFEVA010000008; 892047-954376 nt	8%	Cluster 128	JAFIMQ010000056; 63663-125990 nt	8%	Cluster 19	JAFIMR010000003; 1280798-1343122 nt	8%
Fumonisin	BGC0000062	putative	Cluster 55	JAFEVA010000011; 265042-376911 nt	8%	Cluster 6	JAFIMQ010000001; 1410505-1504667 nt	8%	Cluster 77	JAFIMR010000013; 279232-373099 nt	8%
Citrinin	BGC0001338	T1pks-Nrps	Cluster 58	JAFEVA010000012; 23838-108416 nt	18%	Cluster 35	JAFIMQ010000010; 301372-349261 nt	18%	Cluster 164	JAFIMR010000051; 146144-230732 nt	18%
Fusaric acid	BGC0001190	T1pks	Cluster 74	JAFEVA010000017; 234418-293543 nt	40%	Cluster 21	JAFIMQ010000006; 595090-654213 nt	40%	Cluster 40	JAFIMR010000005; 1645008-1704131 nt	40%
Huperzine A	BGC0000812	putative	Cluster 124	JAFEVA010000041; 832-41099 nt	7%	Cluster 142	JAFIMQ010000065; 182553-219895 nt	7%	Cluster 66	JAFIMR010000010; 470151-519359 nt	7%
Azanigerone	BGC0001143	putative	Cluster 135	JAFEVA010000050; 1-20038 nt	13%	Cluster 106	JAFIMQ010000040; 43943-61749 nt	13%	Cluster 35	JAFIMR010000005; 201783-221058 nt	13%
Pseurotin A	BGC0001037	T1pks-Nrps	Cluster 142	JAFEVA010000057; 40715-93063 nt	40%	Cluster 111	JAFIMQ010000042; 122903-175251 nt	40%	Cluster 124	JAFIMR010000029; 264106-316457 nt	40%
Stipitatic acid	BGC0000154	T1pks	Cluster 149	JAFEVA010000069; 139069-187236 nt	14%	Cluster 144	JAFIMQ010000068; 1-71005 nt	14%	Cluster 167	JAFIMR010000056; 123790-211383 nt	14%
PR toxin	BGC0000667	Terpene	Cluster 166	JAFEVA010000084; 93982-115213 nt	50%	Cluster 137	JAFIMQ010000061; 200253-221484 nt	50%	Cluster 67	JAFIMR010000010; 543818-565049 nt	50%
Ferrichrome	BGC0000901	Nrps	-	-	-	Cluster 116	JAFIMQ010000047; 295666-323880 nt	66%	-	-	-
Tryptoquialanine	BGC0001142	putative	-	-	-	-	-	-	Cluster 39	JAFIMR010000005; 1515656-1561789 nt	9%
Brassicicene C	BGC0000685	Indole	-	-	-	-	-	-	Cluster 89	JAFIMR010000017; 656036-677413 nt	22%

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

antiSMASH 6 results

Similar known cluster	**MIBiG ID**	**Type**	**Region in CBS**	**Location in CBS**	**similarity CBS to MIBIG**	**Region in TUCIM 5827**	**Location in TUCIM 5827**	**similarity TUCIM 5827 to MIBIG**	**Region in TUCIM 5799**	**Location in TUCIM 5799**	**similarity TUCIM 5799 to MIBIG**
Depudecin	BGC0000046	T1pks	Region 1.3	JAFEVA010000001; 1474369-1521786 nt	50%	Region 64.1	JAFIMQ010000064; 37595-85010 nt	50%	Region 4.2	JAFIMR010000004; 1711506-1758920 nt	50%
Naphthalene	BGC0001906	T1pks	Region 4.3	JAFEVA010000004; 317530-364110 nt	33%	Region 12.1	JAFIMQ010000012; 221463-268043 nt	33%	Region 8.1	JAFIMR010000008; 317552-364132 nt	33%
Gibberellin	BGC0001604	Terpene	Region 5.1	JAFEVA010000005; 454771-477803 nt	57%	Region 3.2	JAFIMQ010000003; 443642-462167 nt	57%	Region 6.1	JAFIMR010000006; 753385-776417 nt	57%
Neosartorin	BGC0001988	T1pks	Region 5.2	JAFEVA010000005; 496849-542699 nt	21%	Region 3.1	JAFIMQ010000003; 377894-419682 nt	21%	Region 6.2	JAFIMR010000006; 795568-841418 nt	21%
Nivalenol	BGC000127	Terpene-Nrps	Region 8.4	JAFEVA010000008; 892047-954376 nt	8%	Region 56.1	JAFIMQ010000056; 72834-119502 nt	8%	Region 3.2	JAFIMR010000003; 1289954-1336637 nt	8%
Epipyriculol	BGC0001749	T1pks	Region 9.2	JAFEVA010000009; 832293-865631 nt	29%	Region 1.3	JAFIMQ010000001; 1207690-1255337 nt	29%	Region 9.2	JAFIMR010000009; 137840-185488 nt	29%
Citrinin	BGC0001338	T1pks-Nrps	Region 12.1	JAFEVA010000012; 23838-108416 nt	18%	Region 10.2	JAFIMQ010000010; 301372-349261 nt	18%	Region 51.1	JAFIMR010000051; 146144-230732 nt	18%
Fusaric acid	BGC0001190	T1pks	Region 17.2	JAFEVA010000017; 234418-293543 nt	54%	Region 6.1	JAFIMQ010000006; 595090-654213 nt	54%	Region 5.1	JAFIMR010000005; 1645008-1704131 nt	54%
Squalestatin S1	BGC0001839	Terpene	Region 22.2	JAFEVA010000022; 123022-144551 nt	40%	Region 17.1	JAFIMQ010000017; 89705-111234 nt	40%	Region 63.1	JAFIMR010000063; 12526-34055 nt	40%
Neurosporin A	BGC0001697	T1pks-Nrps	Region 27.1	JAFEVA010000027; 199140-260418 nt	26%	Region 9.3	JAFIMQ010000009; 317685-378946 nt	26%	Region 36.1	JAFIMR010000036; 199044-260481 nt	26%

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Pseurotin A	BGC0001037	T1pks-Nrps	Region 57.1	JAFEVA010000057; 40715-93063 nt	40%	Region 42.1	JAFIMQ010000042; 122903-175251 nt	40%	Region 29.1	JAFIMR010000029; 264106-316457 nt	40%
Eupenifeldin	BGC0001976	T1pks	Region 69.1	JAFEVA010000069; 139069-187236 nt	63%	Region 68.1	JAFIMQ010000068; 3404-71005 nt	72%	Region 56.1	JAFIMR010000056; 124266-211383 nt	72%
(-)-Mellein	BGC0001244	T1pks	Region 74.1	JAFEVA010000074; 32980-78564 nt	100%	Region 142.1	JAFIMQ010000142; 1-37021 nt	100%	Region 28.1	JAFIMR010000028; 187968-226811 nt	100%
PR toxin	BGC0000667	Terpene	Region 84.1	JAFEVA010000084; 93982-115213 nt	50%	Region 61.1	JAFIMQ010000061; 200253-221484 nt	50%	Region 10.3	JAFIMR010000010; 543818-565049 nt	50%
Koraiol	BGC0001642	Terpene	Region 103.1	JAFEVA010000103; 50306-71982 nt	100%	Region 98.1	JAFIMQ010000098; 36363-58039 nt	100%	Region 44.2	JAFIMR010000044; 275212-296888 nt	100%
Dimethylcoprogen	BGC0001249	Nrps	Region 110.1	JAFEVA010000110; 983-46328 nt	100%	Region 9.2	JAFIMQ010000009; 257354-300373 nt	100%	Region 73.1	JAFIMR010000073; 71619-106735 nt	100%
Swainsonine	BGC0001794	Nrps-like-T1pks	-	-	-	Region 108.1	JAFIMQ010000108; 25624-73159 nt	66%	Region 3.3	JAFIMR010000003; 2025104-2072639 nt	66%
Brassicicene C	BGC0000685	Indole	-	-	-	-	-	-	Region 17.1	JAFIMR010000017; 656036-677413 nt	22%

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Potential melanin BGC annotation

CBS 164.80 gene ID	**TUCIM 5872 gene ID**	**TUCIM 5799 gene ID**	**Annotation**
JN550g1910	JX266g3884	JX265g4319	phospholipase D
JN550g1911	JX266g3883	JX265g4318	putative alpha/beta-hydrolase or haloalkane dehalogenase
JN550g1912	JX266g3882	JX265g4317	integral membrane protein
JN550g1913	JX266g3881	JX265g4316	(conidial pigment biosynthesis) multicopper oxidase *abr1*
JN550g1914	JX266g3880	JX265g4315	non-reducing polyketidesynthase
JN550g1915	JX266g3879	JX265g4314	RNA II specific transcription factor
JN550g1916	JX266g3878	JX265g4313	tetrahydroxynaphthalene reductase
JN550g1917	JX266g3877	JX265g4312	transcription factor *Cmr1*

Potential fusaric acid BGC annotation

CBS 164.80 gene ID	**TUCIM 5872 gene ID**	**TUCIM 5799 gene ID**	**Annotation**
JN550g5517	JX266g2331	JX265g3282	hypothetical protein
JN550g5518	JX266g2330	JX265g3281	transmembrane protein
JN550g5519	JX266g2329	JX265g3280	dehydrogenase
JN550g5520	JX266g2328	JX265g3279	transcription factor (similar to Fusaric acid cluster transcription factor FUB10)
JN550g5521	JX266g2327	JX265g3278	N-methyltransferase
JN550g5522	JX266g2326	JX265g3277	Cytochrome P450 monooxygenase
JN550g5523	JX266g2325	JX265g3276	Non-canonical non-ribosomal peptide synthetase
JN550g5524	JX266g2324	JX265g3275	FMN-dependent alpha-hydroxy acid oxidase
JN550g5525	JX266g2323	JX265g3274	Aspartokinase
JN550g5526	JX266g2322	JX265g3273	Fusaric acid biosynthesis protein 2 / hypothetical protein
JN550g5527	JX266g2321	JX265g3272	reducing polyketide synthase
JN550g5528	JX266g2320	JX265g3271	hydrolase
JN550g5529	JX266g2319	JX265g3270	O-acetylhomoserine (Thiol)-lyase
JN550g5530	JX266g2318	JX265g3269	Cytochrome P450 monooxygenase
JN550g5531	JX266g2317	JX265g3268	DNA primase large-subunit

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

JN550g5532	JX266g2316	JX265g3267	putative D-/L-hydantoinase subunit
JN550g5533	JX266g2315	JX265g3266	putative nucleoside transporter
JN550g5534	JX266g2314	JX265g3265	hypothetical protein

Potential fungal-RiPP BGC annotation

CBS 164.80 gene ID	**TUCIM 5872 gene ID**	**TUCIM 5799 gene ID**	**Annotation**
JN550g11999	JX266g8396	JX265g11666	HET domain-containing protein
JN550g12000	JX266g8395	JX265g11665	potential choriogenin H minor
JN550g12001	JX266g8394	JX265g11664	carbonyl reductase
JN550g12002	JX266g8393	JX265g11663	cytochrome p450
JN550g12003	JX266g8392	JX265g11662	major facilitator transporter
JN550g12004	JX266g8391	JX265g11661	hypothetical protein
JN550g12005	JX266g8390	JX265g11660	aspartyl endopeptidase
JN550g12006	JX266g8389	JX265g11659	potential RiPP-precursor
JN550g12007	JX266g8388	JX265g11658	RNase H
JN550g12008	JX266g8387	JX265g11657	gamma-glutamyltranspeptidase
JN550g12009	JX266g8386	JX265g11656	ABC-transporter
JN550g12010	JX266g8385	JX265g11655	hypothetical protein
JN550g12011	JX266g8384	JX265g11654	CoA-transferase
JN550g12012	JX266g8383	JX265g11653	fungal-specific transcription factor

Additional File 29

This File is too large to be displayed.

Additional File 30

This File is too large to be displayed.

Additional File 31

This File is too large to be displayed.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Appendix

The appendix presents published interdisciplinary studies fruited from collaborations with different working groups. The supplementary files of these studies were not included, but can be viewed online and were deposited by their respective corresponding author with the publishing journal.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Ellena *et al. BMC Genomics* (2021) 22:679
https://doi.org/10.1186/s12864-021-07990-8

BMC Genomics

RESEARCH Open Access

Genome sequencing of the neotype strain CBS 554.65 reveals the MAT1–2 locus of *Aspergillus niger*

Valeria Ellena[1,2], Sjoerd J. Seekles[3,4], Gabriel A. Vignolle[2], Arthur F. J. Ram[3,4] and Matthias G. Steiger[1,2*]

Abstract

Background: *Aspergillus niger* is a ubiquitous filamentous fungus widely employed as a cell factory thanks to its abilities to produce a wide range of organic acids and enzymes. Its genome was one of the first *Aspergillus* genomes to be sequenced in 2007, due to its economic importance and its role as model organism to study fungal fermentation. Nowadays, the genome sequences of more than 20 *A. niger* strains are available. These, however, do not include the neotype strain CBS 554.65.

Results: The genome of CBS 554.65 was sequenced with PacBio. A high-quality nuclear genome sequence consisting of 17 contigs with a N50 value of 4.07 Mbp was obtained. The assembly covered all the 8 centromeric regions of the chromosomes. In addition, a complete circular mitochondrial DNA assembly was obtained. Bioinformatic analyses revealed the presence of a MAT1-2-1 gene in this genome, contrary to the most commonly used *A. niger* strains, such as ATCC 1015 and CBS 513.88, which contain a MAT1-1-1 gene. A nucleotide alignment showed a different orientation of the MAT1–1 locus of ATCC 1015 compared to the MAT1–2 locus of CBS 554.65, relative to conserved genes flanking the MAT locus. Within 24 newly sequenced isolates of *A. niger* half of them had a MAT1–1 locus and the other half a MAT1–2 locus. The genomic organization of the MAT1–2 locus in CBS 554.65 is similar to other *Aspergillus* species. In contrast, the region comprising the MAT1–1 locus is flipped in all sequenced strains of *A. niger*.

Conclusions: This study, besides providing a high-quality genome sequence of an important *A. niger* strain, suggests the occurrence of genetic flipping or switching events at the MAT1–1 locus of *A. niger*. These results provide new insights in the mating system of *A. niger* and could contribute to the investigation and potential discovery of sexuality in this species long thought to be asexual.

Keywords: Sexual development, Mating-type locus, Mitochondrial DNA, Centromere, ATCC 16888, NRRL 326

---

* Correspondence: matthias.steiger@tuwien.ac.at
[1]Austrian Centre of Industrial Biotechnology (ACIB GmbH), Muthgasse, 18 Vienna, Austria
[2]Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Gumpendorfer Straße 1a BH, 1060 Vienna, Austria
Full list of author information is available at the end of the article

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Background

Aspergillus niger is a filamentous fungus classified in the section *Nigri* of the genus *Aspergillus.* Its versatile metabolism allows it to grow in a wide variety of environments [1]. Since the early twentieth century it has become a major industrial producer of organic acids, such as citric and gluconic acid, and enzymes, including amylases and phytases [2, 3]. The United States Food and Drug Administration has given it GRAS (Generally Regarded As Safe) status because of its long history of industrial use [3].

First genome sequencing projects were focused on industrial relevant strains. In 2007, the genome sequence of the enzyme-producing strain CBS 513.88 was published [4], followed by the sequencing of the citric acid-producing strain ATCC 1015 in 2011 [5]. At the moment, the genome sequences of 23 *A. niger* strains are available in GenBank. Surprisingly, the *A. niger* strain CBS 554.65 has not yet been sequenced although it is the official neotype strain of this species [6]. This strain was isolated from a tannic-gallic acid fermentation in Connecticut (USA) and it is listed as the (neo-)type strain by international strain collections, such as the Westerdijk Institute (CBS 554.65), the American Type Culture Collection (ATCC 16888) and the ARS Culture Collection (NRRL 326). According to the International Code of Nomenclature for algae, fungi and plants (Shenzhen Code) a neotype is “a specimen or illustration selected to serve as nomenclatural type if no original material exists, or as long as it is missing” [7]. The importance of strain CBS 554.65 lies in its use as biological model and reference strain for morphological observations and taxonomical studies. *A. niger* was previously shown to be able to form sclerotia [8–11], which are an important prerequisite for the sexual development in closely related species. In 2016 the presence of a MAT1–2 locus in the genome of CBS 554.65 was mentioned in a study [12], making this strain an interesting candidate for investigating sexuality in *A. niger.*

The MAT loci are regions of the genome which contain one or more open reading frames of which at least one encodes a transcription factor [13, 14]. Conventionally, the MAT locus containing a transcription factor with an α1 domain similar to the MATα1 of *S. cerevisiae* is called MAT1–1, while the MAT locus containing a transcription factor with a high mobility group (HMG) domain is called MAT1–2 [13]. The corresponding genes are usually called MAT1-1-1 and MAT1-2-1 [13]. The first number indicates that the two sequences are found in the same locus. Due to their sequence dissimilarities they are not termed alleles but idiomorphs [15]. MAT1-1-1 and MAT1-2-1 are major players in the sexual cycle of fungi. They contain DNA binding motifs and were shown to control the expression of pheromone and pheromone-receptor genes during the mating process [16–18]. In heterothallic species, which are self-incompatible, only one of the two MAT genes is found and mating can occur only between strains of opposite mating-type [13]. In homothallic species, which are self-fertile, both MAT genes are present, either linked or unlinked, in the same genome [19]. In the ascomycetes, the sequences flanking the MAT loci are highly conserved [13, 20, 21]. In the aspergilli, as well as in other fungi, including yeasts, the MAT idiomorphs are usually flanked by the genes *slaB*, encoding for a cytoskeleton assembly control factor, and the DNA lyase *apnB*. An anaphase promoting complex gene (*apcE*) is also sometimes present [21].

Although present in previously sequenced genomes, the second mating-type locus of *A. niger* has not been described in detail. In this study, we present the full genome sequence of a MAT1–2 *A. niger* strain and compare its MAT locus to the one of strain ATCC 1015 and those of 24 de novo sequenced *A. niger* isolates containing both MAT1–1 and MAT1–2 loci.

Materials and methods

Strains

The genetic organization of the MAT locus present in *A. niger* CBS 554.65 (ATCC 16888, NRRL 326) was analyzed and compared to the MAT locus of *A. niger* ATCC 1015 and 24 *A. niger* isolates obtained from the Westerdijk Fungal Biodiversity Institute (Uppsalalaan 8, Utrecht, the Netherlands). The isolates analyzed are listed in Table S1 (Additional file 1).

Media

The morphology of strain CBS 554.65 was inspected on minimal medium [22] and malt extract agar (30 g/L malt extract (AppliChem, Darmstadt, Germany) and 5 g/L peptone from casein (Merck KGaA, Darmstadt, Germany)). The strain was 4-point inoculated and incubated at 30 °C for one week.

Genome sequencing and annotation

The genome of the *A. niger* neotype strain CBS 554.65 was sequenced with the PacBio® technology using the PacBio SEQUEL system (Sequencing Chemistry S/P2-C2/5.0) by the Vienna Biocenter Core Facilities (VBCF). The genome was assembled with the default HGAP4 pipeline in PacBio SMRTlink version 5.1.0.26412. The mitochondrial DNA was assembled using CLC Genomic Workbench 12.0 (QIAGEN). The genome annotation of CBS 554.65 was performed with Augustus [23], by training the tool on the genome annotation of the strain ATCC 1015 as reference.

PCRs were performed on the genomic DNA of CBS 554.65 to confirm sequencing and assembly results.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Primer pairs chr5_left_fwd/chr5_left_rev and chr5_right_fwd_1/chr5_right_rev_1 were used to amplify 1756 bp and 1638 bp respectively in the left and in the right region of chr5_00008F. Primers B150 and B151 were used to amplify 1644 bp in the MAT1–1 locus of ATCC 1015. Primers B151 and B152 were used to amplify 2009 bp in the MAT1–2 locus of CBS 554.65. PCR products were sequenced by Microsynth AG.

The MAT locus sequences of 24 *A. niger* isolates were extracted from the complete genome sequences obtained with the Illumina technology and assembled using SPADes [24] (data not published). Homologues of the MAT genes in these isolates were determined based on local Blastn searches using genes obtained from CBS 554.65 and ATCC 1015 as query. In 18 out of the 24 *A. niger* isolates the MAT locus was distributed over multiple scaffolds. In order to verify the location of the MAT genes and their orientation in these strains, diagnostic PCRs and subsequent sequencing were performed to fill in silico gaps within the MAT locus. Primers used in this study are listed in Table S2 (Additional file 2).

Bioinformatic analyses

The genome and the gene set of CBS 554.65 were evaluated using Quast v5.0.2 [25, 26], which includes a benchmarking with Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2. This was performed with the fungal dataset of 290 BUSCOs from 85 fungal species [27]. The genome was masked using RepeatMasker v4.0.9 to identify repetitive elements [28]. Transfer RNA genes were detected using tRNAscan-SE v1.3.1 [29].

The unprocessed reads were mapped to the assembly with the Burrows-Wheeler Alignment Tool (bwa) [30, 31] and the mapping was sorted with SAMtools [32]. The average coverage based on the sorted mapping was calculated in the R environment [33]. The mappings for each individual scaffold were plotted in R and coverage graphs for each scaffold obtained.

The proteomes of the strains CBS 554.65 and NRRL3 were aligned using DIAMOND blastp [34, 35] with an E-value of e^{-10}. The output, consisting of the unique proteins of CBS 554.65 compared to NRRL3, was filtered with a blastx analysis to remove unannotated proteins and analyzed with pannzer2 [36]. The same analysis was performed on the complete proteome of strain CBS 554.65. A singular enrichment analysis (SEA) was performed on the GO term set of unique proteins of CBS 554.65 referenced to the entire GO term set of CBS 554.65 with agriGO [37, 38].

The genome sequences of strains ATCC 1015, NRRL3 and CBS 513.88 were retrieved from JGI [39]. Analyses of the position of the MAT genes within the MAT locus for *A. niger* strains were performed either on BLAST, by searching in the whole-genome shotgun contig database (wgs) of *A. niger*, or on CLC Main Workbench 20.0.2 (QIAGEN). The same analysis was performed for *A. welwitschiae* strains on BLAST against the whole-genome shotgun contig database (wgs) limited by organism (*Aspergillus niger*) and with FungiDB for the other *Aspergillus* species [40]. Sequence analyses and alignments were performed with CLC Main Workbench 20.0.2 (QIAGEN).

Results and discussion

Morphology of strain CBS 554.65

The strain CBS 554.65 is the *A. niger* neotype, a reference strain for morphological and taxonomical analyses. The morphology of this strain grown on minimal medium and malt extract agar can be observed in Fig. 1. On both media CBS 554.65 forms abundant conidia, black on minimal medium and dark brown on malt extract agar.

Genome sequence and analysis

The genome sequencing of the neotype strain CBS 554.65 yielded 5.3 Gbp in 287,000 subreads. The mean length was 18.4 Kbp for the longest subreads and half of the data was in reads longer than 29 Kbp. The assembly consisted of 17 contigs with a total of 40.4 Mbp and a 127-fold coverage. Half of the size of the genome is comprised in 4 scaffolds (L50) of which the smallest has a length of 4.07 Mbp (N50). The GC content is 49.57%. 100% complete BUSCOs (Benchmarking Universal Single-Copy Orthologs) with 2 duplicated and no fragmented BUSCOs were found. The repetitive regions were identified with RepeatMasker v4.0.9 [28]. Using this approach, we were able to recognize interspersed repeats, such as long interspaced nuclear repeats (LINEs) and long terminal repeats (LTR), short interspaced nuclear repeats (SINEs), transposable element like repeats as well as small RNAs, tRNA genes, simple repeats and low complexity repeats. A total of 669,638 bp of the genome was flagged as repetitive, this represents 1.66% of the total genome. In addition, a tRNA prediction with

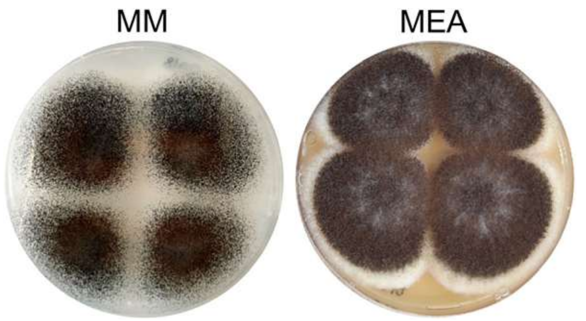


Fig. 1 Morphology of the neotype strain CBS 554.65 on minimal medium (MM) and malt extract agar (MEA)

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

tRNAscan-SE v1.3.1 was performed using the unmasked genome, because fungal specific SINEs were associated with tRNAs. Complete genome characteristics are reported in Tables S3 and S4 of Additional file 3.

The nuclear genome was annotated with Augustus, using the genome of strain ATCC 1015 as reference. Based on this automated annotation 12,240 protein coding genes were predicted. Table 1 shows some basic characteristics of the CBS 554.65 nuclear genome, calculated with Quast, in comparison to the characteristics of other three sequenced *A. niger* strains, CBS 513.88, ATCC 1015 and NRRL3, obtained from JGI.

The CBS 554.65 genome assembly has an increased quality compared to the assemblies of the other strains, with a higher coverage, a higher N50 value and a lower L50 value. CBS 554.65 has a larger genome, while the GC content is similar in the 4 strains. For each of the 8 chromosomes, a putative centromeric region between 88 and 100 kb was identified, which is highlighted in Fig. 2 with vertical black lines. These regions have a GC content between 17.1 and 18.4%, significantly lower than the GC content characterizing the total genome (49.57%) and do not contain any predicted ORF. The only exception is a single ORF of 219 nucleotides in the centromere of chromosome 1. This is found in a 7 kb region of the centromere with a higher GC content compared to the GC content of the entire centromere, suggesting the presence of a mobile element. A conserved domain search [43] on this sequence gave as hits CHROMO and chromo shadow domains (accession: cd00024), ribonuclease H-like superfamily domain (accession: cl14782), integrase zinc binding domain (accession: pfam17921), reverse transcriptase domain (accession: cd01647), RNase H-like domain found in reverse transcriptase (accession: pfam17919) and a retropepsin-like domain (accession: cd00303). The presence of the last four domains suggests that the analyzed sequence has a retroviral or a retrotransposon origin. Similar sequences with domains for reverse transcriptase were also found in the centromeres of chromosomes 5, 6 and 7. Transposons and retrotransposons have been identified in the centromeres of other eukaryotes, including fungi [44, 45]. Blast analyses of the single chromosomes of strain CBS 554.65 against the complete genome of strain NRRL3 and of strains CBS 513.88 showed that the putative centromeres are almost completely lacking from the genome assembly of NRRL3 (Fig. 2, grey areas in the blast graph) and CBS 513.88 (Fig. S1, Additional file 4). Although difficult to identify, centromeric regions in filamentous fungi are composed of complex and heterogeneous AT rich sequences which can stretch up to 450 kb [45, 46]. Due to the likely presence of near-identical long repeats, centromeres are difficult to sequence and assemble [46] which explains why they are lacking in strain NRRL3. The blast analyses against NRRL3 and CBS 513.88 showed that other large regions of the genome of CBS 554.65 do not find homology in NRRL3 or in CBS 513.88. To confirm that these unique regions are not artifacts, the sequencing reads of CBS 554.65 were remapped to the genome. 298,301 reads (90.38% of the total reads) were remapped to the nuclear genome yielding an average coverage calculated on scaffold level of 127x. Figure S2 in the additional file 5 shows the coverage plots for each of the 17 contigs constituting the nuclear genome sequence. Continous coverage was also obtained for the CBS 554.65 regions not found in NRRL3 such as those present in chromosome 2 (chr2_00000F), chromosome 4 (chr4_000001F) and chromosome 5 (chr5_000008F) (Fig. S2, Additional file 5). Moreover, two analytic PCR reactions were successfully performed on the non-homologous region on chromosome 5 (chr5_000008F, Fig. 2). Sequencing of the PCR products confirmed the sequence obtained by genome assembly. The long reads and the high coverage characterizing this genome project allow to assemble sequences which are missing from previous genome assemblies obtained with other sequencing technologies. The number of protein-coding genes in CBS 554.65 is in line with what was found in ATCC 1015 and NRRL3. The large difference in the protein-coding genes in strain CBS 513.88 is likely caused by overpredictions, as previously suggested [5]. A comparison of the proteome of CBS 554.65 and NRRL3 by a blastp

Table 1 Comparison of the basic characteristics of the nuclear genomes of 4 different *A. niger* strains

	CBS 554.65 (This study)	**CBS 513.88 [4, 5]**	**ATCC 1015 [5]**	**NRRL3 [41, 42]**
Genome size (Mb)	40.42	33.98	34.85	35.25
Coverage	127x	7.5x	8.9x	10x
Number of contigs	17	471	24	15
Number of scaffolds	17	19	24	15
Scaffold N50 (Mbp)	4.07	2.53	1.94	2.81
Scaffold L50	4	6	6	5
GC content (%)	49.57	50.4	50.3	49.92
Protein-coding genes	12,240	14,097	11,910	11,846

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

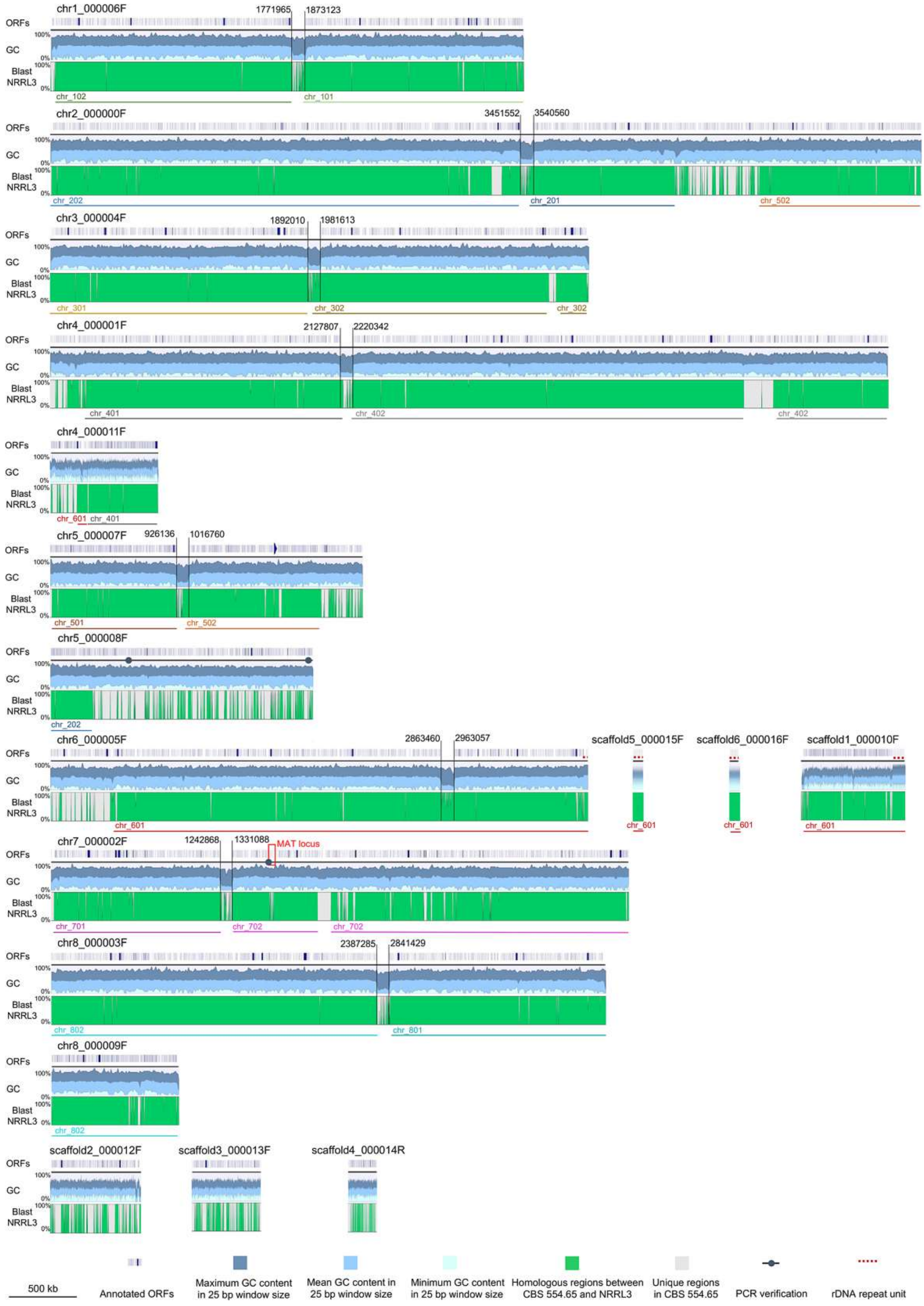


Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 Assembly of the genome sequence of CBS 554.65 consisting of 17 contigs (in scale). For each contig (black horizontal lines) the annotated ORFs (first row), the GC content (second row) and the conservation compared to NRRL3 (third row) are schematically represented. The annotation was obtained with Augustus. The GC content was calculated using a window size of 25 bp. The upper and darker graph represents the maximum GC content value observed in that region, the middle graph represents the mean GC value and the lower graph represents the minimum GC value. The conservation graph (last row) was obtained by blasting each contig of CBS 554.65 against the whole genome of strain NRRL3. The results shown here were additionally confirmed using Mauve [47] by performing progressive alignments of each CBS 554.65 scaffold with the complete genome sequence of NRRL3 (data not shown). Green areas indicate genomic regions conserved between the two strains, grey areas indicate regions only found in CBS 554.65 and not in NRRL3. Below the conservation graph lines representing the chromosomes of strain NRRL3 are reported, as a result of the blast analysis. Chr6_00005F, scaffold1_000010F, scaffold5_000015F and scaffold6_000016F contain the highly repetitive ribosomal DNA (rDNA) gene unit, indicated with a dashed line on top of the scaffolds. Notably, for each of the 8 identified chromosomes, a centromeric region of at least 80 kb could be identified where ORFs are not annotated (indicated with two parallel and vertical lines; the first and the last nucleotide after and before the annotated ORFs, respectively, are indicated). These regions correspond to a decrease in the GC content (as indicated in the GC graph) and are only partially present in the genome of strain NRRL3 (grey areas in the blast graph). Dots on chr5_000008F and on chr7_000002F indicate the region where the PCRs were performed. The MAT locus analyzed in the following paragraphs is indicated by a red box on chromosome 7. Fig. S1 in the additional file 4 reports the comparison of the CBS 554.65 genome to the one of strain CBS 513.88. Additional information on the length of the contigs and the coordinates of the alignments are reported in Table S8 of Additional file 7

analysis showed that there are 694 unique protein sequences in the proteome of CBS 554.65 compared to NRRL3 (additional file 6, Table S6) and 209 unique protein sequences in the proteome of NRRL3 compared to CBS 554.65 (additional file 6, Table S7). GO terms were assigned to proteins and a GO term enrichment analysis was performed with agriGO [37, 38]. 39 GO terms were significantly enriched in the set of unique CBS 554.65 GO terms when referenced to the entire CBS 554.65 GO term set (additional file 6, Table S5, Figs. S3 and S4). Interestingly, GO terms related to thiamine, cholesterol metabolic processes as well as RNA processing are enriched. Overall, this demonstrates that in this genome sequence novel protein sequences were detected, which are absent from previous reference genome projects and might yield novel insights into the biology of this fungus.

Mitochondrial DNA

The mitochondrial DNA is often neglected in genome projects, which tend to focus on the nuclear genome. In *A. niger* only one mitochondrial DNA (mtDNA) assembly was reported, for the strain N909 [48]. In this study, the mtDNA of strain CBS 554.65 was de novo assembled from PacBio reads as a circular DNA with a length of 31,363 bp. MtDNA is abundant in whole genome sequencing projects and the read coverage of the assembly (average: 1220 x, min: 328 x, max: 1674 x) is thus higher than that for the nuclear genome. In total 18 ORFs, 26 tRNA and 2 rRNA sequences were annotated (Fig. 3). All 15 core mitochondrial genes reported for *Aspergillus* species were identified with a similar gene organization [49]. In addition, three accessory genes *orf1L*, *orf3* and *endo1* were annotated. The gene *endo1* is located in the intron of *cox1* and encodes a putative homing endonuclease gene belonging to the LAGLIDADG family frequently found in the *cox1* intron of other filamentous fungi [49]. The gene *orf3* encodes a hypothetical protein of 191 residues, which is also present in the mtDNA of strain N909 but was not annotated there. Surprisingly this unknown protein has a good hit against an unknown protein of *Staphylococcus aureus* (99% identity, WP_117225298.1), however not against other proteins of *Aspergillus* species. In *A. niger* strain N909 two other unknown proteins are encoded in *orf1* and *orf2*. These two open reading frames are connected to *orf1L* in *A. niger* CBS 554.65 yielding a potential protein product

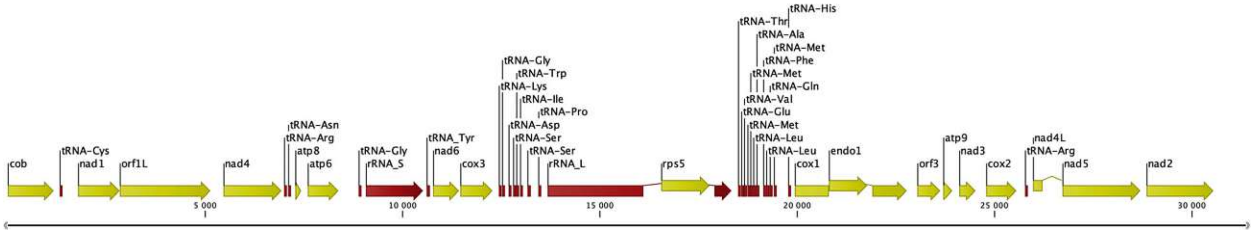


Fig. 3 Annotation of the 31 kbp circular mtDNA sequence (displayed in a linear projection): ORF (yellow), rRNA, tRNA (red)

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

with 739 amino acid residues. This is similar to an open reading frame located at the same position between *nad1* and *nad4* in the mtDNA of *A. flavus* NRRL 3357 (AFLA_m0040), with a size of 667 amino acid residues. In the N-terminal region of both putative proteins, transmembrane spanning regions can be predicted supposing a location in a mitochondrial membrane. However the C-terminal regions are not conserved between *A. niger* and *A. flavus* proteins. We suggest to use the mitochondrial assembly of CBS 554.65 as a reference sequence for *A. niger* mitochondria because it is known that strain N909 is resistant to oligomycin [50]. This resistance is typically linked to mutations in the mtDNA, either in *atp6* [51] or *atp9* [52], and indeed two mutations are found in *atp6* of strain N909 (L26W and S173L).

Discovery and sequencing of a MAT1–2 *A. niger* strain

The genome sequencing and analysis of strain CBS 554.65 allowed to determine the mating-type of this strain. The sequence of the putative MAT1-2-1 gene (g9041) was searched in the standard nucleotide collection database (nr/nt) using Blastn. This gave as hits the mating-type HMG-box protein MAT1-2-1 of other aspergilli, including *A. neoniger* (with an identity of 93.25%) and *A. tubingensis* (with an identity of 93.07%). As such, we consider gene g9041 to be homologous to the MAT1-2-1 gene of other *Aspergillus* species.

This is in line with a previous study that showed the presence of a MAT1-2-1 sequence in the CBS 554.65 strain through a PCR approach [12]. Here we report the complete genome sequence of an *A. niger* strain having a MAT1-2-1 gene. The availability of this genome sequence represents an important tool for further studies investigating the sexual potential of *A. niger*. The presence of both opposite mating-type genes in different strains belonging to the same species represents a strong hint of a sexual lifestyle [14].

MAT1–2 locus analysis and comparison to MAT1–1

The locus of strains CBS 554.65 containing the MAT1-2-1 gene was compared in silico to the locus of strain ATCC 1015 containing the MAT1-1-1 gene. This was done to determine whether the genes flanking the MAT1-1-1 gene are also present in the genome of the MAT1–2 strain and vice versa. A region of 40,517 bp, spanning from gene Aspni7|39467 (genomic position 2,504,615 in the v7 of the ATCC 1015 genome) to gene Aspni7|1128148 (genomic position 2,545,131) was aligned to the corresponding region of strain CBS 554.65 (Fig. 4). In CBS 554.65 the two genes homologous to Aspni7|39467 (g9051) and Aspni7|1128148 (g9036) are comprised in a sequence of 43,891 bp, almost 4 kb longer than in ATCC 1015. The identifiers of the genes included in these regions are indicated in Fig. 4 and additionally reported in Table 2, with their predicted function retrieved from FungiDB or blast analysis. The alignment shows that the MAT genes occupy the same genomic location at chromosome 7. The genes comprised in the analyzed loci are mostly conserved between the two strains, with the exception of genes Aspni7|1178859 (MAT1-1-1), Aspni7|1128137 and Aspni7|1160288, unique for ATCC 1015, and g9046, g9041 (MAT1-2-1) and g9040–2 (MAT1-2-4), unique for CBS 554.65. Aspni7|1128137 has predicted metal ion transport activity and it is found in other *Aspergillus* species, either heterothallic with a MAT1-1-1 or a MAT1-2-1 gene or homothallic. It is not found near the MAT gene, with the exception of *A. brasiliensis* and *A. ochraceoroseus*. Aspni7|1160288 has a domain with

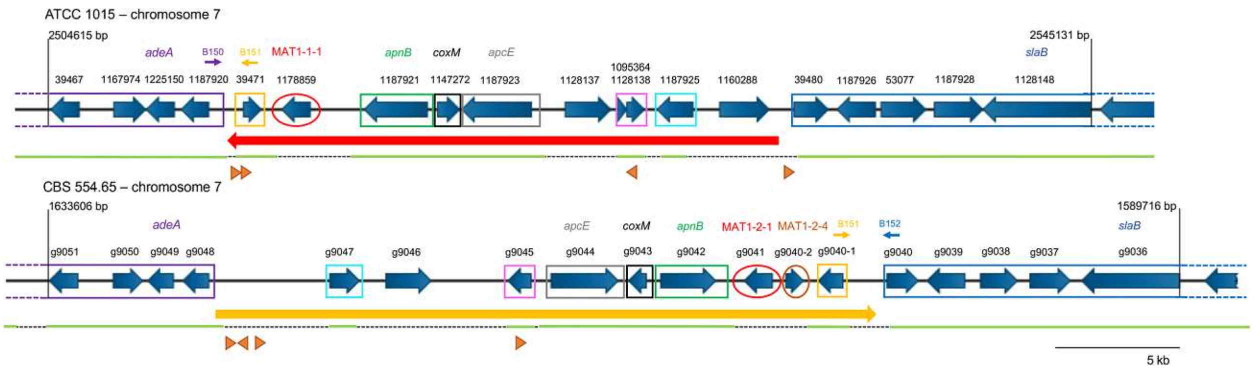


Fig. 4 Nucleotide alignment between the same genomic region of ATCC 1015 (MAT1–1) and CBS 554.65 (MAT1–2). Genes found in both strains are indicated with a box of the same color, MAT genes are indicated with a circle and unmarked genes are unique in each strain. Below each genomic region, green lines indicate regions homologous in the two strains and dotted lines regions unique for each strain. A red arrow indicates the genomic region of ATCC 1015 which contains the MAT1-1-1 gene and appears flipped compared to the corresponding region in CBS 554.65 (yellow arrow). Small arrows with numbers B150, B151 and B152 indicate primers used for PCRs. Orange triangles indicate the presence of a 7 bp motif (5′-TTACACT)

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Table 2 List of genes included in the genomic region comprising the MAT genes

ATCC 1015	CBS 554.65	Predicted function retrieved from FungiDB or blast
Aspni7|39467	g9051	Hypothetical protein
Aspni7|1167974	g9050	CIA30-domain containing protein – Ortholog(s) have role in mitochondrial respiratory chain complex I assembly
Aspni7|1225150	g9049	SAICAR synthetase (*adeA*)
Aspni7|1187920	g9048	Homolog in CBS 513.88 has domain(s) with predicted catalytic activity, metal ion binding, phosphoric diester hydrolase activity
Aspni7|39471	g9040–1	Hypothetical protein
Aspni7|1178859	–	Mating-type protein MAT1-1-1
Aspni7|1187921	g9042	DNA lyase Apn2|Hypothetical protein
Aspni7|1147272	g9043	Hypothetical cytochrome C oxidase|Mitochondrial cytochrome c oxidase subunit VIa
Aspni7|1187923	g9044	Ortholog(s) are anaphase-promoting complex proteins
Aspni7|1128137	–	Homolog in CBS 513.88 has domain(s) with predicted metal ion transmembrane transporter activity, role in metal ion transport, transmembrane transport and membrane localization
Aspni7|1095364	g9045	HAD-like protein; Homolog in CBS 513.88 has domain(s) with predicted hydrolase activity
Aspni7|1128138	g9045	HAD-like protein; Homolog in CBS 513.88 has domain(s) with predicted hydrolase activity
Aspni7|1187925	g9047	Glycosyltransferase Family 8 protein - Ortholog(s) have acetylglucosaminyltransferase activity, role in protein N-linked glycosylation and Golgi medial cisterna localization
Aspni7|1160288	–	Aspartic protease|Hypothetical aspartic protease
Aspni7|39480	g9040	WD40 repeat-like protein
Aspni7|1187926	g9039	Aldehyde dehydrogenase
Aspni7|53077	g9038	CoA-transferase family III
Aspni7|1187928	g9037	Salicylate hydroxylase
Aspni7|1128148	g9036	Cytoskeleton assembly control protein Sla2
–	g9046	Hypothetical protein
–	g9041	Mating-type HMG-box protein MAT1-2-1
–	g9040–2	Hypothetical protein – Putative homologue of MAT1–2-4 of *A. fumigatus*

predicted role in proteolysis and its homolog in other aspergilli is present at another genomic locus, not in proximity to the MAT gene. A homolog of gene g9046 was found by Blastn search in *Aspergillus vadensis*, in a different location of the genome than the MAT locus. These results suggest that these unique genes are unlikely to be part of the "core" MAT locus. The gene g9040–2 is a putative homolog of the MAT1-2-4 gene in *A. fumigatus*, an additional mating-type gene required for mating and cleistothecia formation [53]. Another difference between ATCC 1015 and CBS 554.65 is represented by the gene putatively encoding for a HAD-like protein. While this gene is complete in CBS 554.65 (g9045), it appears disrupted in ATCC 1015 and, therefore, doubly annotated in this strain (Aspni7|1095364 and Aspni7|1128138). The other genes present in the selected genomic region show a high level of conservation, with a higher synteny further away from the MAT genes (genes in the purple and blue boxes). Moreover, genes encoding for the DNA lyase a*pnB*, the cytoskeleton control assembly factor *slaB* and the anaphase promoting complex *apcE* are present in both MAT loci. These genes are normally found in the MAT loci of other fungi, including yeast [21]. Their presence in the MAT loci of *A. niger* further confirms the high level of conservation characterizing this locus. In heterothallic ascomycetes the MAT genes are commonly included between the genes *apnB* and *slaB* [21]. From the alignment in Fig. 4 the relative position of the MAT genes to *apnB* and *slaB* can be analyzed. In CBS 554.65 the MAT1-2-1 gene (g9041) is flanked by *apnB* and *slaB* respectively upstream and seven genes downstream. In contrast, in the MAT1–1 locus of strain ATCC 1015 the MAT gene is flanked downstream by *apnB* and upstream by a conserved sequence including *adeA*, while *slaB* is found on the same side of *apnB*. The entire genomic locus, containing the MAT1-1-1 gene and eight other genes (23 kbp indicated by the red arrow in Fig. 4), shows a flipped orientation compared to the corresponding locus in CBS 554.65 containing the MAT1-2-1 gene (indicated by an orange arrow in Fig. 4). The ORF direction of the conserved genes *apnB*, *coxM* and *apcE* additionally confirms the different orientation of this locus in the two strains. In addition, PCRs performed with primers B150, B151

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

and B152 (Fig. 4) yielded expected bands, confirming the orientation of the MAT loci of both ATCC 1015 and CBS 554.65. By sequence analysis, a repetitive 7 bp DNA motif (5′-TTACACT) was found in the MAT1–1 locus (orange triangles in Fig. 4), where the homology between the MAT1–1 and MAT1–2 loci breaks (in proximity to *adeA* and *slaB*). An additional site of this motif was found in the gene encoding a HAD-like hydrolase (Aspni7|1128138). This motif is present at similar positions in at least two other sequenced MAT1–1 strains of *A. niger* (N402, CBS 513.88). Differently, the MAT1–2 strain presents this motif only at the site close to the *adeA* gene and in the putative HAD-like hydrolase gene (g9045), but not at the site close to the *slaB* gene.

Methods to identify the opposite mating-type in strains isolated from natural sources often rely on the use of primers designed to bind to *apnB* and *slaB*, since these are the genes that commonly flank the MAT gene itself [54, 55]. In both mating-type *A. niger* strains, *slaB* is found more than 12 kbp away from the MAT gene. In addition, the relative orientation of *apnB* to *slaB* is different in strains having opposite mating types. This might explain why the MAT1–2 locus was only mentioned by one previous study [12] but never described in detail so far.

Both the particular orientation of the MAT locus and the presence of a repetitive motif in the MAT loci suggest that a genetic switch or a flipping event might have occurred or is still ongoing in *A. niger*, which might affect the expression of the MAT genes. Genetic switching events at the MAT locus are known for other ascomycetes, particularly yeasts. For instance, in *S. cerevisiae* a switching mechanism involving an endonuclease and two inactive but intact copies of the MAT genes allows to switch the MAT type of the cell [56]. Expression of the MAT gene is instead regulated in the methylotrophic yeasts *Komagataella phaffii* and *Ogataea polymorpha* via a flip/flop mechanism [57, 58]. In these species, a 19 kbp sequence including both mating type genes is flipped so that a MAT gene will be close to the centromere (5 kbp from the centromere) and, therefore, silenced while the other will be transcribed. In CBS 554.65 the region comprising the MAT1-2-1 gene is present at around 280 kbp downstream of the putative centromere, which is much further away of what observed for *K. phaffi* and *O. polymorpha*. However, in certain basidiomycetes, such as *Microbotryum saponariae* and *Microbotryum lagerheimii*, the mating-type locus HD (containing the homeodomain genes) is around 150 kbp distant from the centromere and linked to it [59]. It was proposed that the proximity to the centromere in these species might be enough to reduce recombination events [59]. The effect of the distance between the centromere and the MAT genes in *A. niger* merits further attention, especially in view of a potential sexual cycle characterizing this species.

Inversion at the MAT locus have been described for certain homothallic filamentous fungi such as *Sclerotinia sclerotiorum* and *Sclerotinia minor* [60, 61]. Field analysis of a large number of isolates showed that strains belonging to these species can either present a non-inverted or an inverted MAT locus. In the inverted orientation two of the four MAT genes at the locus have the opposite orientation and one gene is truncated. In the case of *S. sclerotiorum*, differences in the gene expression were observed between inverted and non-inverted strains. This inversion, induced by crossing-over between two identical inverted repeat present in the locus, likely happens during the sexual cycle before meiosis [60]. The analysis of a larger number of *A. niger* isolates is required to investigate whether opposite orientations of both MAT loci exist for this species as well and what the implications of such inversions might be. Chromosomal inversions are considered to prevent recombination between sex determining genes in higher eukaryotes, such as animals and plants [62]. Further studies are required to investigate whether *A. niger* possesses a genetic switching mechanism controlling its sexual development.

Genetic comparison of MAT loci in different aspergilli and additional *A. niger* strains

This study revealed a particular configuration for the MAT1–1 locus of strain ATCC 1015. For this reason, the orientation of the MAT locus of additional *Aspergillus* species for which a genome sequence is available was analyzed. Firstly, the genes *adeA* and *slaB* were retrieved as they are conserved and often found at the right and left flank of the MAT gene, respectively (Fig. 4). Subsequently, the position of the MAT gene was compared to the three conserved genes *apnB*, *coxM* and *apcE*. The MAT gene could be either included between *adeA* and *apnB*, like in ATCC 1015 (flipped position), or between *apnB and slaB*, like in CBS 554.65 (conserved position). The results of this analysis are reported in Table 3. A complete table with the identifiers of all genes analyzed is reported in the Additional file 8.

Table 4 MAT genes which are found between *apnB* and *slaB* are considered to have a “conserved” position, while MAT genes identified between *adeA* and *apnB* are considered as “flipped”. *Aspergillus* species are grouped in sections based on the most updated classification [71]. For each species it is indicated if a sexual cycle has been reported in the literature.

In the analyzed *Aspergillus* sequences the MAT gene (either MAT1-1-1 or MAT1-2-1) was mostly found between the genes *apnB* and *slaB*, such as in CBS 554.65 (conserved). The only exceptions, showing a

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Table 3 MAT gene identifiers of the analyzed *Aspergillus* strains and their position in the MAT locus

Section	Species	Strain	Mating-type gene - MAT	Mating-type	MAT position	Sexual cycle described for the species
Nigri	*A. welwitschiae*	CBS 139.54	172,181	MAT1–1	flipped	No
	A. kawachii (A. luchuensis)	IFO 4308	AKAW_03832	MAT1–2	conserved	No
	A. luchuensis	106.47	ASPFODRAFT_180958	MAT1–1	conserved	No
	A. tubingensis	G131	Not annotated	MAT1–2	conserved	Yes [63]
		CBS 134.48	ASPTUDRAFT_124452	MAT1–1	conserved	
	A. niger	CBS 554.65	g9041	MAT1–2	conserved	No
		ATCC 1015	ASPNIDRAFT2_1178859	MAT1–1	flipped	
	A. brasiliensis	CBS 101740	ASPBRDRAFT_167991	MAT1–2	flipped	No
	A. carbonarius	ITEM 5010	ASPCADRAFT_1991	MAT1–2	conserved	No
	A. aculeatus	ATCC 16872	ASPACDRAFT_1867751	MAT1–2	conserved	No
Nidulantes	*A. versicolor*	CBS 583.65	ASPVEDRAFT_82222	MAT1–2	conserved	No
	A. sydowii	CBS 593.65	ASPSYDRAFT_87884	MAT1–2	conserved	No
Ochraceorosei	*A. ochraceoroseus*	IBT 24754	P175DRAFT_0477739	MAT1–1	conserved	No
Flavi	*A. flavus*	NRRL 3357	AFLA_103210	MAT1–1	conserved	Yes [64]
	A. oryzae	BCC7051	OAory_01101300	MAT1–2	conserved	No
		RIB40	AO090020000089	MAT1–1	conserved	
Circumdati	*A. steynii*	IBT 23096	P170DRAFT_349471	MAT1–2	conserved	No
Candidi	*A. campestris*	IBT 28561	P168DRAFT_313902	MAT1–1	conserved	No
			P168DRAFT_285957	MAT1–2	conserved	
Terrei	*A. terreus*	NIH2624	ATEG_08812	MAT1–1	conserved	Yes [65]
Fumigati	*A. novofumigatus*	IBT 16806	P174DRAFT_462167	MAT1–2	conserved	No
	A. fischeri	NRRL 181	NFIA_071100	MAT1–1	conserved	Yes [66]
			NFIA_024390	MAT1–2	conserved	
	A. fumigatus	Af293	Afu3g06170	MAT1–2	conserved	Yes [67]
		A1163	AFUB_042900	MAT1–1	conserved	
			AFUB_042890	MAT1–2	conserved	
Clavati	*A. clavatus*	NRRL1	ACLA_034110	MAT1–1	conserved	Yes [68]
			ACLA_034120	MAT1–2	conserved	
Aspergillus	*A. glaucus*	CBS 516.65	ASPGLDRAFT_89185	MAT1–1	n.a.[1]	Yes [69, 70]
Cremei	*A. wentii*	DTO 134E9	ASPWEDRAFT_184745	MAT1–2	conserved	No

[1] Conserved genes not in the MAT locus

configuration similar to the MAT1–1 locus of ATCC 1015, were the MAT1-1-1 gene of *A. welwitschiae* and the MAT1-2-1 gene of *A. brasiliensis*. This analysis could not be performed on the MAT1–2 locus of *A. welwitschiae* nor on the MAT1–1 locus of *A. brasiliensis*, due to the unavailability of sequences for strains of the opposite mating type. Seven of the analyzed species, including the closely related *A. tubingensis*, were reported to have a sexual cycle. A conserved position of the MAT gene was observed for all of these species with the exception for *A. glaucus*, whose conserved genes were not found in the vicinity of the MAT gene. These observations suggest that the position of the MAT gene and the orientation of the locus might have an impact on the sexual development of the respective fungus.

Since the orientation observed for the MAT1–1 locus of ATCC 1015 might be peculiar for this *A. niger* strain only, additional genome sequences were analyzed to determine the orientation of the MAT locus of other sequenced strains of *A. niger* (Table 4). 18 out of 23 *A. niger* strain sequences deposited in GenBank contain a MAT1-1-1 gene and they all show the same orientation of the MAT locus as observed in ATCC 1015. The other 5 strains contain a MAT1–2 locus and they all show the

TU Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table 4 Mating-type and MAT gene position of the analyzed *A. niger* strains. 48 *A. niger* strains have been analyzed in respect to their MAT locus configuration. Newly sequenced *A. niger* isolates and CBS 554.65 are reported in rows above the dashed line. Among these, 12 have a MAT1–1 and 13 a MAT1–2 locus. Previously sequenced *A. niger* strains are reported in rows below the dashed line. Among these, a bias towards MAT1–1 strains is present. All the MAT1–1 strain have a flipped orientation of the MAT locus and all the MAT1–2 strains a conserved one. *MAT locus distributed over multiple scaffolds which could not be combined

MAT1–1				**MAT1–2**			
***A. niger* strain**	**Isolation source**	**MAT position**	**GenBank accession**	***A. niger* strain**	**Isolation source**	**MAT position**	**GenBank accession**
CBS 112.32	Unknown, Japan	flipped	MW809488	CBS 554.65	Tannin-gallic acid fermentation, USA	conserved	PRJNA715116
CBS 147371	Green coffee bean, India	flipped	MW809493	CBS 113.50	Leather, unknown	conserved	MW809487
CBS 147320	Grape, Australia	flipped	MW809494	CBS 124.48	Unknown	conserved	MW809489
CBS 147345	Unknown, USA	flipped	MW809501	CBS 118.52	Unknown	conserved	Incomplete coverage*
CBS 147347	Petridish, soft drink factory, The Netherlands	flipped	MW809503	CBS 147321	Arctic soil, Norway	conserved	MW809495
CBS 769.97	Leather, Unknown	flipped	MW809504	CBS 147322	Coffee, Brazil	conserved	MW809496
CBS 115989	Unknown	flipped	MW809505	CBS 147323	Raisin, Turkey	conserved	MW809497
CBS 147352	Air next to bottle blower, Mexico	flipped	MW809506	CBS 147324	Unknown	conserved	MW809498
CBS 147353	Food factory of Sanquinetto, Italy	flipped	MW809507	CBS 147482	Surface water, Portugal	conserved	Incomplete coverage*
CBS 115988	Unknown	flipped	MW809491	CBS 147344	Coffee beans, Thailand	conserved	MW809499
CBS 131.52	Leather, unknown	flipped	MW809490	CBS 133816	Black pepper, Denmark	conserved	MW809500
CBS 147343	Coffee bean, Thailand	flipped	MW809508	CBS 147346	CF patient material, The Netherlands	conserved	MW809502
H915–1	Soil, China	flipped	PRJNA288269	CBS 630.78	Army equipment, South Pacific Islands	conserved	MW809492
L2	Soil, China	flipped	PRJNA288269	RAF106	Pu'er tea, China	conserved	PRJNA503751
LDM3	Industrial production, China	flipped	PRJNA562509	3.316	Laboratory, China	conserved	PRJNA597564
FDAARGOS_311	USA	flipped	PRJNA231221	An76	Soil, China	conserved	PRJDB4313
N402 (ATCC 64974)	Laboratory, The Netherlands	flipped	PRJEB21769	JSC-093350089	ISS environmental surface, USA	conserved	PRJNA355122
ATCC 10864	Soil, Peru	flipped	PRJNA300350	MOD1-FUNGI2	Red seedless grapes, USA	Genes in different scaffolds	PRJNA482816
F3_1F3_F	ISS environmental surface, USA	flipped	PRJNA667181				
F3_4F2_F	ISS environmental surface, USA	flipped	PRJNA667181				
F3_4F1_F	ISS environmental surface, USA	flipped	PRJNA667181				
DSM 1957	Soil, France	flipped	PRJNA566102				
FGSC A1279	Laboratory, The Netherlands	flipped	PRJNA255851				

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table 4 Mating-type and MAT gene position of the analyzed *A. niger* strains. 48 *A. niger* strains have been analyzed in respect to their MAT locus configuration. Newly sequenced *A. niger* isolates and CBS 554.65 are reported in rows above the dashed line. Among these, 12 have a MAT1–1 and 13 a MAT1–2 locus. Previously sequenced *A. niger* strains are reported in rows below the dashed line. Among these, a bias towards MAT1–1 strains is present. All the MAT1–1 strain have a flipped orientation of the MAT locus and all the MAT1–2 strains a conserved one. *MAT locus distributed over multiple scaffolds which could not be combined *(Continued)*

MAT1–1				**MAT1–2**			
***A. niger* strain**	**Isolation source**	**MAT position**	**GenBank accession**	***A. niger* strain**	**Isolation source**	**MAT position**	**GenBank accession**
A1	Soil, China	flipped	PRJNA288269				
ATCC 1015	USA	flipped	PRJNA15785				
ATCC 13496	Soil, USA	flipped	PRJNA209543				
CBS 101883 (*A. lacticoffeatus*)	Coffee beans, Sumatra	flipped	PRJNA479910				
CBS 513.88	Unknown	flipped	PRJNA19275				
SH-2	Soil, China	flipped	PRJNA196564				
ATCC 13157 (*A. phoenicis*)	Whole shelled corn	flipped	PRJNA209548				

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

same conserved orientation as observed in the strain CBS 554.65. The orientation could not be determined for one MAT1–2 strain, MOD1FUNGI2, since the different analyzed genes are present in different scaffolds in the available genome sequence. Overall, 80% of the sequenced strains contain a MAT1–1 locus. The selection procedure of strains for whole-genome sequencing might be biased by their industrial relevance and might not resemble the mating-type distribution in the environment. Therefore, 24 randomly picked isolates of *A. niger* were sequenced and the MAT loci analyzed: 12 contain the MAT1–1 locus and 12 the MAT1–2 locus (Table 4).

The MAT locus configuration of these strains is similar to the configuration of strain ATCC 1015, in the case of the MAT1–1 strains, and to CBS 554.65, in the case of at least 10 out of 12 MAT1–2 strains. In the two remaining MAT1–2 strains (CBS 118.52 and CBS 147482) a gap between two genomic scaffolds could not be closed by PCR. This is likely due to the presence of a region with multiple G repeats. However, when the two separate scaffolds of these isolates were aligned to the MAT1–2 locus of CBS 554.65, they appeared to have the same locus configuration as the other 10 MAT1–2 isolates. Similarly to what was observed for ATCC 1015 and CBS 554.65, the HAD-like protein encoding gene appears disrupted in all the MAT1–1 isolates and complete in all the MAT1–2 isolates. Further studies are required to investigate whether the disruption of this gene in the MAT1–1 strains plays a role in the context of fungal development. Overall, the MAT1–1 configuration described in Fig. 4 is a peculiar feature of *A. niger* and its close relative *A. welwitschiae*. Despite the unusual orientation, the presence of a 50:50 ratio of MAT1–1:MAT1–2 among 24 randomly selected *A. niger* isolates is remarkable and suggests that sexual reproduction is occurring in this species. Interestingly, MAT1–1 occurs at higher frequency in commonly used industrial and laboratory strains. This could be pure coincidence, but it could also indicate a phenotypic difference between strains with opposite matingtypes.

Conclusions

The *A. niger* neotype strain CBS 554.65 has now a high quality genome sequence, which covers all the 8 centromeres and includes a complete mtDNA sequence. This sequence represents an important tool for further studies. The analysis of this genome revealed the presence of a second mating-type locus (MAT1–2) in this strain, making it a suitable reference strain to investigate fungal development in *A. niger*. The position and the orientation of the MAT1-2-1 gene of all the 15 MAT1–2 *A. niger* strains analyzed was found to be similar to that of other aspergilli, with the MAT gene included between the genes *apnB* and *slaB*. The unusual orientation of the MAT1-1-1 locus found in the already sequenced *A. niger* strains and in other 12 newly sequenced isolates indicates that flipping or switching events have occurred at the MAT locus. Further research is required to investigate whether this difference in the position of the MAT genes in the opposite mating-type strains could have an effect on the expression of the genes included in this genomic region. These flipping events might have a direct impact on the sexual development in *A. niger*.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07990-8.

Additional file 1: Table S1. *Aspergillus niger* strains used in this study.

Additional file 2: Table S2. List of primers used in this study.

Additional file 3: Table S3. Genome characteristics and found Benchmarking Universal Single-Copy Orthologues (BUSCO) genes of the assembled *Aspergillus niger* strain CBS 554.65. **Table S4.** Masked repetitive elements found with RepeatMasker v4.0.9 and tRNA genes found by tRNAscan-SE v1.3.1.

Additional file 4: Fig. S1. Assembly of the genome sequence of CBS 554.65 consisting of 17 contigs (in scale). For each contig (black horizontal lines) the annotated ORFs (first row), the GC content (second row) and the conservation compared to CBS 513.88 (third row) are schematically represented.

Additional file 5: Fig. S2. Coverage plots of the scaffolds obtained by remapping the reads to the CBS 554.65 genome assembly.

Additional file 6: Table S5. GO term enrichment analysis of the unique GO term set of CBS 554.65 referenced to the entire GO term set of CBS 554.65. The unique CBS 554.65 proteins compared to NRRL3 are 694, of which 176 had at least one GO term assigned. **Fig. S3.** GO term enrichment analysis of the unique GO term set of CBS 554.65 referenced to the entire GO term set of CBS 554.65, assigned to the biological process ontology. **Fig. S4.** GO term enrichment analysis of the unique GO term set of CBS 554.65 referenced to the entire GO term set of CBS 554.65, assigned to the molecular function ontology. **Table S6.** Unique protein sequences in the proteome of CBS 554.65 compared to NRRL3 by a blastp analysis. **Table S7.** Unique protein sequences in the proteome of NRRL3 compared to the entire proteome of CBS 554.65 by a blastp analysis.

Additional file 7: Table S8. Lenght of the contigs of CBS 554.65 and coordinates of the NRRL3 and CBS 513.88 contig alignments to CBS 554.65.

Additional file 8: Table S9. Gene identifiers of the analyzed *Aspergillus* strains and their position in the MAT locus.

Acknowledgements
The authors want to thank Heinz Ekker from the Vienna Bioscience Core Facility for running the assembly and providing bioinformatic support.

Authors' contributions

MGS conceived the study. VE and MGS designed the experiments and the bioinformatic analyses. VE performed the experiments and the bioinformatic analyses of CBS 554.65 and ATCC 1015 with the help of GAV. SJS and AFJR provided the sequence data of the MAT loci of the isolates. SJS performed the experiments and the bioinformatic analyses of the isolates. VE prepared the manuscript with input from MGS and all authors provided critical feedback. All authors read and approved the final manuscript.

Funding
The COMET center: acib: Next Generation Bioproduction is funded by BMK, BMDW, SFG, Standortagentur Tirol, Government of Lower Austria und Vienna

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Business Agency in the framework of COMET - Competence Centers for Excellent Technologies. The COMET-Funding Program is managed by the Austrian Research Promotion Agency FFG. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Program.
Part of this study was funded by the TIFN project "Heterogeneity in spores of food spoilage fungi" (AF-15507).

Availability of data and materials

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the bioproject PRJNA715116 (accession JAGRPH000000000) [https://www.ebi.ac.uk/ena/browser/view/PRJNA715116]. The version described in this paper is version JAGRPH010000000. The genome reads of strain CBS 554.65 are available in the European Nucleotide Archive (ENA) at EMBL-EBI under accession numbers PRJEB42544 [https://www.ebi.ac.uk/ena/browser/view/PRJEB42544]. The mitochondrial genome of strains CBS 554.65 has been deposited at GenBank under the accession MW816869 [https://www.ncbi.nlm.nih.gov/nuccore/MW816869.1]. The MAT loci sequences of the *A. niger* isolates have been deposited at GenBank under the accessions: MW809487-MW809508. [https://www.ncbi.nlm.nih.gov/nuccore/MW809487, https://www.ncbi.nlm.nih.gov/nuccore/MW809488, https://www.ncbi.nlm.nih.gov/nuccore/MW809489, https://www.ncbi.nlm.nih.gov/nuccore/MW809490, https://www.ncbi.nlm.nih.gov/nuccore/MW809491, https://www.ncbi.nlm.nih.gov/nuccore/MW809492, https://www.ncbi.nlm.nih.gov/nuccore/MW809493, https://www.ncbi.nlm.nih.gov/nuccore/MW809494, https://www.ncbi.nlm.nih.gov/nuccore/MW809495, https://www.ncbi.nlm.nih.gov/nuccore/MW809496, https://www.ncbi.nlm.nih.gov/nuccore/MW809497, https://www.ncbi.nlm.nih.gov/nuccore/MW809498, https://www.ncbi.nlm.nih.gov/nuccore/MW809499, https://www.ncbi.nlm.nih.gov/nuccore/MW809500, https://www.ncbi.nlm.nih.gov/nuccore/MW809501, https://www.ncbi.nlm.nih.gov/nuccore/MW809502, https://www.ncbi.nlm.nih.gov/nuccore/MW8094503, https://www.ncbi.nlm.nih.gov/nuccore/MW8094504, https://www.ncbi.nlm.nih.gov/nuccore/MW809505, https://www.ncbi.nlm.nih.gov/nuccore/MW809506, https://www.ncbi.nlm.nih.gov/nuccore/MW809507, https://www.ncbi.nlm.nih.gov/nuccore/MW809508].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Austrian Centre of Industrial Biotechnology (ACIB GmbH), Muthgasse, 18 Vienna, Austria. [2]Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Gumpendorfer Straße 1a BH, 1060 Vienna, Austria. [3]TiFN, P.O. Box 557, 6700, AN, Wageningen, The Netherlands. [4]Leiden University, Institute of Biology Leiden, Molecular Microbiology and Biotechnology, Sylviusweg 72, 2333, BE, Leiden, The Netherlands.

Received: 18 February 2021 Accepted: 3 September 2021
Published online: 21 September 2021

References

1. Schuster E, Dunn-Coleman N, Frisvad J, van Dijck P. On the safety of *Aspergillus niger* - a review. Appl Microbiol Biotechnol. 2002;59(4-5):426–35. https://doi.org/10.1007/s00253-002-1032-6.
2. Currie JN. The citric acid fermentation of *Aspergillus niger*. J Biol Chem. 1917; 31(1):15–37. https://doi.org/10.1016/S0021-9258(18)86708-4.
3. Baker SE, Bennett J. The Aspergilli. In: Goldman GH, Osmani SA, editors. CRC Press; 2007. https://doi.org/10.1201/9781420008517.
4. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, et al. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nat Biotechnol. 2007;25(2):221–31. https://doi.org/10.1038/nbt1282.
5. Andersen MR, Salazar MP, Schaap PJ, van de Vondervoort PJI, Culley D, Thykaer J, et al. Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. Genome Res. 2011; 21(6):885–97. https://doi.org/10.1101/gr.112169.110.
6. Kozakiewicz Z, Frisvad JC, Hawksworth DL, Pitt JI, Samson RA, Stolk AC. Proposal for nomina specifica conservanda and rejicienda in *Aspergillus* and *Penicillium* (Fungi). Taxon. 1992;41(1):109. https://doi.org/10.2307/1222500. https://www.jstor.org/stable/1222500.
7. Turland NJ, Wiersema JH, Barrie FR, Greuter W, Hawksworth DL, Herendeen PS, et al. International Code of Nomenclature for algae, fungi and plants. vol. 159. Koeltz Botanical Books; 2018. https://doi.org/10.12705/Code.2018.
8. Jørgensen TR, Burggraaf A-M, Arentshorst M, Schutze T, Lamers G, Niu J, et al. Identification of SclB, a Zn (II)2Cys6 transcription factor involved in sclerotium formation in *Aspergillus niger*. Fungal Genet Biol. 2020;139: 103377. https://doi.org/10.1016/j.fgb.2020.103377.
9. Frisvad JC, Petersen LM, Lyhne EK, Larsen TO. Formation of sclerotia and production of indoloterpenes by *Aspergillus niger* and other species in section *Nigri*. PLoS One. 2014;9(4):e94857. https://doi.org/10.1371/journal.pone.0094857.
10. Ellena V, Bucchieri D, Arcalis E, Sauer M, Steiger MG. Sclerotia formed by citric acid producing strains of *Aspergillus niger*: induction and morphological analysis. Fungal Biol. 2021;125(6):485–94. https://doi.org/10.1016/j.funbio.2021.01.008.
11. Jørgensen TR, Nielsen KF, Arentshorst M, Park J, van den Hondel CA, Frisvad JC, et al. Submerged conidiation and product formation by *Aspergillus niger* at low specific growth rates are affected in aerial developmental mutants. Appl Environ Microbiol. 2011;77(15):5270–7. https://doi.org/10.1128/AEM.00118-11.
12. Mageswari A, Kim J, Cheon K-H, Kwon S-W, Yamada O, Hong S-B. Analysis of the MAT1-1 and MAT1-2 gene ratio in black koji molds isolated from meju. Mycobiology. 2016;44(4):269–76. https://doi.org/10.5941/MYCO.2016.44.4.269.
13. Debuchy R, Turgeon BG. Mating-type structure, evolution, and function in Euascomycetes. Growth, Differ. Sex. Growth, Di, Berlin/Heidelberg: Springer-Verlag; 2006, p. 293–323. https://doi.org/10.1007/3-540-28135-5_15.
14. Dyer PS, Kück U. Sex and the imperfect fungi. The Fungal Kingdom. 2017; 5(3):193–214. https://doi.org/10.1128/microbiolspec.funk-0043-2017.
15. Metzenberg RL, Glass NL. Mating type and mating strategies in *Neurospora*. BioEssays. 1990;12(2):53–9. https://doi.org/10.1002/bies.950120202.
16. Lee SC, Ni M, Li W, Shertz C, Heitman J. The evolution of sex: a perspective from the fungal kingdom. Microbiol Mol Biol Rev. 2010;74(2):298–340. https://doi.org/10.1128/MMBR.00005-10.
17. Coppin E, Debuchy R, Arnaise S, Picard M. Mating types and sexual development in filamentous ascomycetes. Microbiol Mol Biol Rev. 1997; 61(1):411–28. https://doi.org/10.5424/sjar/2014121-4340.
18. Kück U, Böhm J. Mating type genes and cryptic sexuality as tools for genetically manipulating industrial molds. Appl Microbiol Biotechnol. 2013; 97(22):9609–20. https://doi.org/10.1007/s00253-013-5268-0.
19. Pöggeler S. Mating-type genes for classical strain improvements of ascomycetes. Appl Microbiol Biotechnol. 2001;56(5-6):589–601. https://doi.org/10.1007/s002530100721.
20. Galagan JE, Hynes M, Pain A, Machida M, Purcell S, Peñalva MÁ, et al. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature. 2005;438(7071):1105–15. https://doi.org/10.1038/nature04341.
21. Dyer PS. Sexual reproduction and significance of MAT in the *Aspergilli*. In: Heitman J, Kronstad J, Taylor JCL, editors. Sex Fungi. ASM Press, American Society of Microbiology; 2007, p. 123–42. https://doi.org/10.1128/9781555815837.ch7.
22. Arentshorst M, Ram AFJ, Meyer V. Using non-homologous end-joining-deficient strains for functional gene analyses in filamentous fungi. In: Bolton MD, Thomma BPHJ, editors. Plant fungal Pathog. Methods Protoc, vol. 835. Humana Pre, Totowa, NJ: Humana Press; 2012. p. 133–50. https://doi.org/10.1007/978-1-61779-501-5_9.
23. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005; 33(Web Server):W465–7. https://doi.org/10.1093/nar/gki458.
24. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77. https://doi.org/10.1089/cmb.2012.0021.

25. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5. https://doi.org/10.1093/bioinformatics/btt086.
26. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142–50. https://doi.org/10.1093/bioinformatics/bty266.
27. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351.
28. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. http://www.repeatmasker.org.
29. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–64. https://doi.org/10.1093/nar/25.5.955.
30. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
31. Li H, Durbin R. Fast and accurate long-read alignment with burrows–wheeler transform. Bioinformatics. 2010;26(5):589–95. https://doi.org/10.1093/bioinformatics/btp698.
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
33. Team RC. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2019.
34. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60. https://doi.org/10.1038/nmeth.3176.
35. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421. https://doi.org/10.1186/1471-2105-10-421.
36. Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids Res. 2018;46(W1):W84–8. https://doi.org/10.1093/nar/gky350.
37. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010;38(suppl_2):W64–70. https://doi.org/10.1093/nar/gkq310.
38. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017;45(W1):W122–9. https://doi.org/10.1093/nar/gkx382.
39. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 2014;42(D1):D26–31. https://doi.org/10.1093/nar/gkt1069.
40. Basenko E, Pulman J, Shanmugasundram A, Harb O, Crouch K, Starns D, et al. FungiDB: an integrated Bioinformatic resource for Fungi and oomycetes. J Fungi. 2018;4(1):39. https://doi.org/10.3390/jof4010039.
41. Aguilar-Pontes MV, Brandl J, McDonnell E, Strasser K, Nguyen TTM, Riley R, et al. The gold-standard genome of *Aspergillus niger* NRRL 3 enables a detailed view of the diversity of sugar catabolism in fungi. Stud Mycol. 2018;91:61–78. https://doi.org/10.1016/j.simyco.2018.10.001.
42. Vesth TC, Nybo JL, Theobald S, Frisvad JC, Larsen TO, Nielsen KF, et al. Investigation of inter- and intraspecies variation through genome sequencing of *Aspergillus* section *Nigri*. Nat Genet. 2018;50(12):1688–95. https://doi.org/10.1038/s41588-018-0246-1.
43. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):D265–8. https://doi.org/10.1093/nar/gkz991.
44. Talbert PB, Henikoff S. What makes a centromere? Exp Cell Res. 2020;389(2):111895. https://doi.org/10.1016/j.yexcr.2020.111895.
45. Smith KM, Galazka JM, Phatale PA, Connolly LR, Freitag M. Centromeres of filamentous fungi. Chromosom Res. 2012;20(5):635–56. https://doi.org/10.1007/s10577-012-9290-3.
46. Friedman S, Freitag M. Centrochromatin of Fungi. Prog Mol Subcell Biol. 2017;56:85–109. https://doi.org/10.1007/978-3-319-58592-5_4.
47. Darling ACE, Mau B, Blattner FR, N.T. P. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 2004;14:1394–1403. https://doi.org/10.1101/gr.2289704, 7.
48. Juhász Á, Pfeiffer I, Keszthelyi A, Kucsera J, Vágvölgyi C, Hamari Z. Comparative analysis of the complete mitochondrial genomes of *Aspergillus niger* mtDNA type 1a and *Aspergillus tubingensis* mtDNA type 2b. FEMS Microbiol Lett. 2008;281(1):51–7. https://doi.org/10.1111/j.1574-6968.2008.01077.x.
49. Joardar V, Abrams NF, Hostetler J, Paukstelis PJ, Pakala S, Pakala SB, et al. Sequencing of mitochondrial genomes of nine *Aspergillus* and *Penicillium* species identifies mobile introns and accessory genes as main sources of genome size variability. BMC Genomics. 2012;13(1):698. https://doi.org/10.1186/1471-2164-13-698.
50. Juhász Á, Láday M, Gácser A, Kucsera J, Pfeiffer I, Kevei F, et al. Mitochondrial DNA organisation of the mtDNA type 2b of *Aspergillus tubingensis* compared to the *Aspergillus niger* mtDNA type 1a. FEMS Microbiol Lett. 2004;241(1):119–26. https://doi.org/10.1016/j.femsle.2004.10.025.
51. Niedzwiecka K, Tisi R, Penna S, Lichocka M, Plochocka D, Kucharczyk R. Two mutations in mitochondrial ATP6 gene of ATP synthase, related to human cancer, affect ROS, calcium homeostasis and mitochondrial permeability transition in yeast. Biochim Biophys Acta - Mol Cell Res. 1865;2018(1):117–31. https://doi.org/10.1016/j.bbamcr.2017.10.003.
52. Ward M, Wilkinson B, Turner G. Transformation of *Aspergillus nidulans* with a cloned, oligomycin-resistant ATP synthase subunit 9 gene. Mol Gen Genet MGG. 1986;202(2):265–70. https://doi.org/10.1007/BF00331648.
53. Yu Y, Amich J, Will C, Eagle CE, Dyer PS, Krappmann S. The novel *Aspergillus fumigatus* MAT1-2-4 mating-type gene is required for mating and cleistothecia formation. Fungal Genet Biol. 2017;108:1–12. https://doi.org/10.1016/j.fgb.2017.09.001.
54. Ramirez-Prado JH, Moore GG, Horn BW, Carbone I. Characterization and population analysis of the mating-type genes in *Aspergillus flavus* and *Aspergillus parasiticus*. Fungal Genet Biol. 2008;45(9):1292–9. https://doi.org/10.1016/j.fgb.2008.06.007.
55. Houbraken J, Dyer PS. Induction of the sexual cycle in filamentous ascomycetes. In: van den Berg MA, Maruthachalam K, editors. Genet. Transform. Syst. Fungi, Vol. 2, vol. 2, Cham: Springer International Publishing; 2015, p. 23–46. https://doi.org/10.1007/978-3-319-10503-1_2.
56. Haber JE. Mating-type genes and MAT switching in *Saccharomyces cerevisiae*. Genetics. 2012;191(1):33–64. https://doi.org/10.1534/genetics.111.134577.
57. Hanson SJ, Byrne KP, Wolfe KH. Flip/flop mating-type switching in the methylotrophic yeast *Ogataea polymorpha* is regulated by an Efg1-Rme1-Ste12 pathway. PLoS Genet. 2017;13(11):1–26. https://doi.org/10.1371/journal.pgen.1007092.
58. Hanson SJ, Byrne KP, Wolfe KH. Mating-type switching by chromosomal inversion in methylotrophic yeasts suggests an origin for the three-locus *Saccharomyces cerevisiae* system. Proc Natl Acad Sci U S A. 2014;111(45):E4851–8. https://doi.org/10.1073/pnas.1416014111.
59. Carpentier F, Rodríguez De La Vega RC, Branco S, Snirc A, Coelho MA, Hood ME, et al. Convergent recombination cessation between mating-type genes and centromeres in selfing anther-smut fungi. Genome Res. 2019;29(6):944–53. https://doi.org/10.1101/gr.242578.118.
60. Chitrampalam P, Inderbitzin P, Maruthachalam K, Wu BM, Subbarao K V. The *Sclerotinia sclerotiorum* Mating Type Locus (MAT) Contains a 3.6-kb Region That Is Inverted in Every Meiotic Generation. PLoS One 2013;8. https://doi.org/10.1371/journal.pone.0056895, 8, 2.
61. Chitrampalam P, Pryor BM. Characterization of mating type (MAT) alleles differentiated by a natural inversion in *Sclerotinia minor*. Plant Pathol. 2015;64(4):911–20. https://doi.org/10.1111/ppa.12305.
62. Wright AE, Dean R, Zimmer F, Mank JE. How to make a sex chromosome. Nat Commun. 2016;7(1):12087. https://doi.org/10.1038/ncomms12087.
63. Horn BW, Olarte RA, Peterson SW, Carbone I. Sexual reproduction in *aspergillus tubingensis* from section *Nigri*. Mycologia. 2013;105(5):1153–63. https://doi.org/10.3852/13-101.
64. Horn BW, Moore GG, Carbone I. Sexual reproduction in *Aspergillus flavus*. Mycologia. 2009;101(3):423–9. https://doi.org/10.3852/09-011.
65. Arabatzis M. Sexual reproduction in the opportunistic human pathogen *Aspergillus terreus* 2013;105:71–79. https://doi.org/10.3852/11-426,
66. Raper KB, Fennell DI. The genus *Aspergillus*; 1965.
67. O'Gorman CM, Fuller HT, Dyer PS. Discovery of a sexual cycle in the opportunistic fungal pathogen *Aspergillus fumigatus*. Nature. 2009;457(7228):471–4. https://doi.org/10.1038/nature07528.
68. Ojeda-López M, Chen W, Eagle CE, Gutiérrez G, Jia WL, Swilaiman SS, et al. Evolution of asexual and sexual reproduction in the Aspergilli. Stud Mycol. 2018;91:37–59. https://doi.org/10.1016/j.simyco.2018.10.002.

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

69. Link HF. Observationes in ordines plantarum naturales. Dissertatio I Mag Ges Naturf Freunde Berlin. 1809;3:3–42.
70. Chen AJ, Hubka V, Frisvad JC, Visagie CM, Houbraken J, Meijer M, et al. Polyphasic taxonomy of *Aspergillus* section *Aspergillus* (formerly *Eurotium*), and its occurrence in indoor environments and food. Stud Mycol. 2017;88: 37–135. https://doi.org/10.1016/j.simyco.2017.07.001.
71. Houbraken J, Kocsubé S, Visagie CM, Yilmaz N, Wang XC, Meijer M, et al. Classification of *Aspergillus*, *Penicillium*, *Talaromyces* and related genera (*Eurotiales*): an overview of families, genera, subgenera, sections, series and species. Stud Mycol. 2020;95:5–169. https://doi.org/10.1016/j.simyco.2020.05.002.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

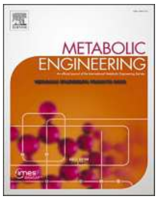
Learn more biomedcentral.com/submissions

Metabolic Engineering 68 (2021) 68–85

Contents lists available at ScienceDirect

Metabolic Engineering

journal homepage: www.elsevier.com/locate/meteng



Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

A quantitative metabolic analysis reveals *Acetobacterium woodii* as a flexible and robust host for formate-based bioproduction

Christian Simon Neuendorf[a], Gabriel A. Vignolle[a], Christian Derntl[a], Tamara Tomin[b], Katharina Novak[a], Robert L. Mach[a], Ruth Birner-Grünberger[b,c], Stefan Pflügl[a,*]

[a] *Technische Universität Wien, Institute for Chemical, Environmental and Bioscience Engineering, Research Area Biochemical Engineering, Gumpendorfer Straße 1a, 1060, Vienna, Austria*

[b] *Technische Universität Wien, Institute for Chemical Technologies and Analytics, Research Group Bioanalytics, Getreidemarkt 9, 1060, Vienna, Austria*

[c] *Medical University of Graz, Diagnostic and Research Institute of Pathology, Center for Medical Research, Stiftingtalstrasse 24, 8036, Graz, Austria*

ARTICLE INFO

Keywords:
Formate bioeconomy
Acetogens
Substrate co-utilization
Mixed gas fermentation
Continuous fermentation
Metabolic modelling and -omics analysis

ABSTRACT

Cheap and renewable feedstocks such as the one-carbon substrate formate are emerging for sustainable production in a growing chemical industry. We investigated the acetogen *Acetobacterium woodii* as a potential host for bioproduction from formate alone and together with autotrophic and heterotrophic co-substrates by quantitatively analyzing physiology, transcriptome, and proteome in chemostat cultivations in combination with computational analyses. Continuous cultivations with a specific growth rate of 0.05 h^{-1} on formate showed high specific substrate uptake rates (47 mmol g^{-1} h^{-1}). Co-utilization of formate with H_2, CO, CO_2 or fructose was achieved without catabolite repression and with acetate as the sole metabolic product. A transcriptomic comparison of all growth conditions revealed a distinct adaptation of *A. woodii* to growth on formate as 570 genes were changed in their transcript level. Transcriptome and proteome showed higher expression of the Wood-Ljungdahl pathway during growth on formate and gaseous substrates, underlining its function during utilization of one-carbon substrates. Flux balance analysis showed varying flux levels for the WLP (0.7–16.4 mmol g^{-1} h^{-1}) and major differences in redox and energy metabolism. Growth on formate, H_2/CO_2, and formate + H_2/CO_2 resulted in low energy availability (0.20–0.22 ATP/acetate) which was increased during co-utilization with CO or fructose (0.31 ATP/acetate for formate + $H_2/CO/CO_2$, 0.75 ATP/acetate for formate + fructose). Unitrophic and mixotrophic conversion of all substrates was further characterized by high energetic efficiencies. *In silico* analysis of bioproduction of ethanol and lactate from formate and autotrophic and heterotrophic co-substrates showed promising energetic efficiencies (70–92%). Collectively, our findings reveal *A. woodii* as a promising host for flexible and simultaneous bioconversion of multiple substrates, underline the potential of substrate co-utilization to improve the energy availability of acetogens and encourage metabolic engineering of acetogenic bacteria for the efficient synthesis of bulk chemicals and fuels from sustainable one carbon substrates.

1. Introduction

En route to a circular bioeconomy, industrial biotechnology becomes a key technology to achieve the United Nations sustainable development goals (Arora and Mishra, 2019) and reduce human-made CO_2 emissions (Köpke and Simpson, 2020). However, to meet the rising global demand for chemicals and fuels (Panich et al., 2021), cheap and sustainable feedstocks are needed for industrial bioproduction.

One-carbon substrates and H_2 are emerging as promising alternatives to traditional, agro-based biotechnological feedstocks. Currently, gaseous carbon and energy sources (CO_2, CO, H_2) are available from large point sources (e.g. steel mills) (Köpke and Simpson, 2020; Novak et al., 2021) and can be obtained via gasification of solid municipal waste or residual biomass (Liew et al., 2016). Moreover, bioproduction of ethanol from CO via gas fermentation has already been commercialized (Vees et al., 2020).

---

* Corresponding author.

E-mail addresses: christian.neuendorf@tuwien.ac.at (C.S. Neuendorf), gabriel.vignolle@tuwien.ac.at (G.A. Vignolle), christian.derntl@tuwien.ac.at (C. Derntl), tamara.tomin@tuwien.ac.at (T. Tomin), katharinanovak.kn@gmail.com (K. Novak), robert.mach@tuwien.ac.at (R.L. Mach), ruth.birner-gruenberger@tuwien.ac.at (R. Birner-Grünberger), stefan.pfluegl@tuwien.ac.at (S. Pflügl).

https://doi.org/10.1016/j.ymben.2021.09.004
Received 14 June 2021; Received in revised form 30 August 2021; Accepted 15 September 2021
Available online 16 September 2021

1096-7176/© 2021 The Authors. Published by Elsevier Inc. on behalf of International Metabolic Engineering Society. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

In the future, circular carbon economies are anticipated to be based on feedstocks obtained from renewable energy (e.g. wind, solar) and CO_2 as abundantly available carbon source (Claassens et al, 2018, 2019; Cotton et al., 2020; Yishai et al., 2016). At this point, CO_2 might be directly captured and concentrated from air (Chatterjee and Huang, 2020; Fasihi et al., 2019; Realmonte et al., 2019).

CO, H_2 and formate have been described as suitable microbial electron donors and are therefore promising mediators between chemical feedstock production and utilization for bioproduction (Boone et al., 1989; Diender et al., 2015). The small one-carbon compound formate is particularly interesting in this context as it can be efficiently produced from CO_2 via electrochemical reduction, hydrogenation and photoreduction (Yishai et al., 2016). Consequently, formate may serve as a chemical energy storage system for excess electricity in the future. While CO and H_2 can also be produced electrochemically or via water hydrolysis with promising efficiencies (Haas et al., 2018; Hardt et al., 2021), formate has several advantages as a substrate compared to the direct utilization of gaseous feedstocks. In contrast to gas fermentations which are typically limited by the gas-liquid mass transfer (Van Hecke et al., 2019), formate is completely miscible with water and can be directly added to the cultivation medium. In addition, the transport and storage of gaseous substrates such as H_2 and CO is challenging due to their high reactivity and toxicity (Cotton et al., 2020; Karmann et al., 2017). Interestingly, the interconversion of H_2, CO and CO_2 to formate by acetogenic bacteria could additionally provide a solution for storage of H_2 (Müller, 2019; Schuchmann and Müller, 2013; Schwarz et al., 2020; Schwarz and Müller, 2020). In the future, formate may be produced in Power-to-X (P2X) approaches. Therefore, formate supply and prices may correlate with the availability of electricity (Li et al., 2012; Yishai et al., 2016).

Currently, several natural and metabolically engineered formatotrophs are investigated for their applicability in formate-based bioproduction. An important criterion for potential microbial hosts is the amount of energy of the substrate that is retained in the product (i.e. energy efficiency) (Claassens et al., 2019). Natural formatotrophs such as *Pseudomonas* species and *Cupriavidus necator* suffer from low energy efficiency on formate, which in turn limits product yields (Claassens et al., 2020; Goldberg et al., 1976). Hence, the metabolic engineering of *Escherichia coli* and *Saccharomyces cerevisiae* for growth on formate focused on efficient assimilation routes such as the reductive glycine pathway (rGLY) (Gonzalez de la Cruz et al., 2019; Kim et al., 2020). Out of all engineered and natural formatotrophs, acetogens show the highest energy efficiency for formate assimilation (Cotton et al., 2020). Acetogens are strictly anaerobic bacteria that utilize the Wood-Ljungdahl Pathway (WLP) and an interlinked redox balancing system for the growth on a variety of one-carbon substrates (Schuchmann and Müller, 2014). The model acetogen *Acetobacterium woodii* utilizes the four one-carbon sources CO, CO_2, formate and methanol (Balch et al., 1977; Bertsch and Müller, 2015a; Kremp et al., 2018; Moon et al., 2021) and is considered for industrial production of the platform chemical acetate from gaseous substrates (Demler and Weuster-Botz, 2011; Kantzow et al., 2015; Novak et al., 2021). Its suitable substrate spectrum and energy-efficient metabolism make *A. woodii* a promising microbial platform organism for sustainable bioprocesses.

In the future, formate may either serve as the main carbon source or as a supplementary substrate in flexible bioprocesses. The co-utilization of formate with other carbon and energy sources such as carbohydrates might offer advantages compared to the use of one-carbon substrates. In addition to higher carbon conversion efficiencies, mixotrophic substrate utilization might be used as a strategy to improve bioenergetics in acetogens single substrates (Jones et al., 2016; Maru et al., 2018; Molitor et al., 2017). Notably, mixotrophic utilization of carbohydrates and gaseous substrates is easily achieved by some acetogens such as *A. woodii* (Braun and Gottschalk, 1981). In addition, the future bioeconomy needs to react flexibly to fluctuating energy and substrate availabilities (Blank et al., 2020; Liew et al., 2016; Wendisch et al., 2016; Yishai et al., 2016). Therefore, co-utilization of substrates and robust process performance with varying substrate supply are desirable.

In this study, we aimed to obtain a quantitative understanding of unitrophic and mixotrophic formate utilization by *A. woodii* to evaluate its potential for formate-based bioproduction. To that end, chemostat cultivations were used to study single substrate (formate, H_2/CO_2 and fructose) utilization on a physiological, transcriptomic and proteomic level. Additionally, we tested whether *A. woodii* can co-utilize formate with gaseous (H_2/CO_2 and $H_2/CO/CO_2$) and heterotrophic (fructose) substrates. Transcriptome and proteome data together with metabolic modelling revealed a high flexibility and robustness of *A. woodii* to utilize multiple substrates simultaneously. Metabolic modelling further highlighted how intracellular energy availability can be controlled by substrate co-utilization. Finally, we discuss the energetic efficiency and strategies for formate-based bioproduction of novel products with *A. woodii*.

2. Material and methods

2.1. Bacterial strain

Acetobacterium woodii DSM1030 was used in all experiments. For cryo-preservation, cell suspensions supplemented with a final sucrose concentration of 125 g L^{-1} were stored at −80 °C.

2.2. Growth medium

For shaken cultivation in serum bottles, cells were grown on a phosphate-buffered medium as previously described (Novak et al., 2021). The medium contained per liter: 1 g NH_4Cl, 2 g yeast extract, 3.47 g NaCl, 0.1 g $MgSO_4{\cdot}7\ H_2O$, 1.76 g KH_2PO_4, 8.44 g K_2HPO_4, 0.5 g cysteine-HCl·H_2O, 0.25 mL sodium resazurin (0.2% w/v), 20 mL adapted trace element solution DSMZ141 and 10 mL vitamin solution DSMZ 141. The vitamin solution from medium recipe DSMZ 141 contained per liter: 2 mg Biotin, 2 mg folic acid, 10 mg pyridoxine-HCl, 5 mg thiamine-HCl, 5 mg riboflavin, 5 mg nicotinic acid, 5 mg D-Ca-panthothenate, 0.1 mg vitamin B_{12}, 5 mg p-Aminobenzoic acid and 5 mg Lipoic acid. The adapted trace element solution based on DSMZ141 contained per liter: 1.5 g nitrilotriacetic acid, 3 g $MgSO_4{\cdot}7\ H_2O$, 0.5 g $MnSO_4{\cdot}\ H_2O$, 1 g NaCl, 0.1 g $FeSO_4{\cdot}7\ H_2O$, 0.152 g $Co(II)Cl_2{\cdot}6\ H_2O$, 0.1 g $CaCl_2{\cdot}2\ H_2O$, 0.18 g $ZnSO_4{\cdot}7\ H_2O$, 0.01 g $CuSO_4{\cdot}5\ H_2O$, 0.02 g $KAl(SO_4)_2{\cdot}12\ H_2O$, 0.01 g boric acid, 0.01 g $Na_2MoO_4{\cdot}2\ H_2O$, 0.033 g $Ni(II)SO_4{\cdot}6\ H_2O$, 0.3 mg $Na_2SeO_3{\cdot}5\ H_2O$ and 0.4 mg $Na_2WO_4{\cdot}2\ H_2O$. Formate or fructose were added from anaerobic stocks with concentrations of 230 g $^{L-1}$ or 250 g L^{-1}, respectively. The pH of the medium for serum bottle cultivation was adjusted to 7.2 with 5 M KOH unless stated otherwise. The medium composition was adapted for bioreactor cultivations: There, the amount of vitamin and trace element solution were doubled, Ca-pantothenate was added to a final concentration of 1 mg L^{-1} (Godley et al., 1990) and $FeSO_4{\cdot}7\ H_2O$ to a final concentration of 26.9 mg L^{-1} (Demler and Weuster-Botz, 2011). The phosphate salt concentrations were reduced to 0.33 g L^{-1} KH_2PO_4 and 0.45 g L^{-1} K_2HPO_4 and the pH of the medium was adjusted to 7.0 with 5 M KOH. Antifoam Struktol SB2020 (Schill und Seilacher, Hamburg, Germany) was added to the medium in a ratio of 1:5,000 (v/v).

2.3. Growth conditions

For growth of pre-cultures and small-scale batch cultivations, cells were grown in 125 mL serum bottles using 50 mL medium. All serum bottle cultures were incubated at 30 °C and 200 rpm in a rotary shaker (Infors AG, Bottmingen, Switzerland). The headspace of the serum bottles was flushed for 1 min with the same gas mixture also used for the respective cultivation. For growth of pre-cultures, 28 mM fructose was used as the carbon source with a N_2 atmosphere in the serum bottle. For autotrophic and mixotrophic experiments, a pre-mixed gas mixture of

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

80/20% (v/v) H_2/CO_2 (Air Liquide Austria GmbH, Schwechat, Austria) was used at a total pressure of 2.5 bar. The headspace was replaced daily with H_2/CO_2. During serum bottle cultivations, 2 mL samples were routinely withdrawn for OD_{600} determination and HPLC analysis.

Continuous cultivations were conducted either in a DASbox® Mini Bioreactor system (Eppendorf AG, Jülich, Germany) or in a DASGIP® Multibioreactor system (Eppendorf AG, Jülich, Germany). A filling volume of 200 mL and an agitation rate of 500 rpm were used for DASbox® cultivations and a filling volume of 1000 mL and an agitation rate of 400 rpm were used for cultivations with the DASGIP® system. For all cultivations, a temperature of 30 °C was used. The reaction volume was maintained at a constant volume using a dip tube and a peristaltic pump (Ismatec SA, Glattburg, Germany). An aeration rate of 0.25 vvm was applied. For the growth conditions H_2/CO_2, formate + H_2/CO_2 and formate + $H_2/CO/CO_2$, pre-mixed gas mixtures with CO (60:9.5:10.6:19% $H_2/CO/CO_2/N_2$) and without CO (60:9.5:29.6% $H_2/CO_2/N_2$) (Air Liquide Austria GmbH, Schwechat, Austria) were used. For growth on formate + fructose, nitrogen gas (purity in % >99.999) (Messer Austria GmbH, Gumpoldskirchen, Austria) was utilized. For growth on fructose alone, nitrogen and carbon dioxide (purity in % >99.995) (Air Liquide Austria GmbH, Schwechat, Austria) were mixed in a ratio of 80:20% N_2/CO_2 by the DASGIP® MX4/1 Gas Mixing Module (Eppendorf AG, Jülich, Germany).

The pH was maintained at 7.0 using 5 M KOH and 2 M phosphoric acid. The medium was sparged continuously with the indicated gases at a rate of 0.25 vvm. Offgas from the DASbox® Mini Bioreactor system was analyzed continuously with a gas chromatograph (Trace GC Ultra, Thermo Fisher Scientific, Waltham/MA, USA). Offgas from the DASGIP Multibioreactor system was analyzed continuously with a DASGIP® GA Exhaust Analyzing Module (Eppendorf AG, Jülich, Germany).

2.4. Biomass concentration determination

The cell dry weight was determined at steady state conditions as follows: 5 mL of freshly sampled culture broth were transferred into dried and pre-weighed glass tubes. The tubes were centrifuged for 10 min at 4 °C and 4,800 rpm (2,396 g), washed with 2.5 mL distilled water and centrifuged again. The samples were dried at 105 °C for 24 h, subsequently cooled in a desiccator for at least 1 h and finally weighed. For cultures grown on formate, a sample volume of 25 mL and a washing volume of 10 mL were used instead. Biomass determination was performed in triplicates. A correlation coefficient between OD_{600} and cell dry weight (biomass = $0.38{*}OD_{600}$) was determined and used to estimate the biomass concentrations at all other points.

2.5. Bioreactor off-gas analysis

A Trace GC Ultra gas chromatograph (Thermo Fisher Scientific, Waltham/MA, USA) was used to analyze the reactor off-gas for H_2, CO, CO_2 and N_2. The gas chromatograph was equipped with a ShinCarbon ST 100/120 packed column (Restek Corporation, Bellefonte/PA, USA) and a thermal conductivity detector operated in constant temperature mode with 200 °C transfer temperature, 240 °C block temperature and 370 °C filament temperature. Argon 5.0 (Messer Austria GmbH, Gumpoldskirchen, Austria) was used as the carrier gas at a constant flow rate of 2.0 mL/min. Samples with a volume of 100 μL were injected with a split ratio of 20. After the injection, the oven temperature was kept constant at 30 °C for 6.5 min, then increased to a temperature of 250 °C with a 16 °C/min ramp and finally kept at 250 °C for 0.75 min. An electrical valve system allowed the automatic off-gas analysis of each of the four bioreactors of the DASbox system in 2 h intervals. The chromatograms were recorded and evaluated using Chromeleon 7.2.10 Chromatography Data System (Thermo Scientific, Waltham/MA, USA). Calibration was performed with premixed defined gas mixtures containing H_2, CO, CO_2 and N_2.

Off-gas analysis with the DASGIP® GA Exhaust Analyzing Module (Eppendorf AG, Jülich, Germany) was performed after calibrating the module with pressurized air and premixed calibration gas. The module was used to analyze the exhaust gas for CO_2.

2.6. Organic acid, sugar, and amino acid analysis

All organic acid, sugar and amino acid analysis were carried out with an Ultimate 3000 High Performance Liquid Chromatograph (HPLC) (Thermo Scientific, Waltham/MA, USA). Control, monitoring and evaluation of the analysis was performed with Chromeleon 7.2.6 Chromatography Data System (Thermo Fisher Scientific, Waltham/MA, USA).

Fructose, formate, and acetate quantification in sample supernatants were achieved with an Aminex HPX-87H column (300 × 7.8 mm, Bio Rad, Hercules/CA, USA). The mobile phase was 4 mM H_2SO_4, and the column was operated at a velocity of 0.6 mL/min, 60 °C for 30 min. Detection was performed with a refractive index (Refractomax 520, Thermo Fisher Scientific, Waltham/MA, USA) and a diode array detector (Ultimate 3000, Thermo Fisher Scientific, Waltham/MA, USA). Prior to analysis, 450 μL of culture supernatant were mixed with 50 μL of 40 mM H_2SO_4 and centrifuged for 5 min at 14,000 rpm (21,913 g) and 4 °C. 10 μL of this samples was injected for analysis (Erian et al., 2018). Standards at defined concentrations of fructose, formate, acetate, and ethanol were treated the same way.

Amino acids were analyzed with a reversed phase column (Agilent Eclipse AAA, 3 × 150 mm, 3.5 μm) with a guard column (Agilent Eclipse AAA, 4.6 × 12.5 mm, 5 μm) and a gradient of eluent (A) 40 mM NaH_2PO_4 monohydrate pH 7.8 and eluent (B) MeOH/ACN/MQ (45/45/10 (v/v/v)). At a flowrate of 1.2 mL/min and a column temperature of 40 °C, samples were analyzed with an injection volume of 10 μL. In-needle derivatization was performed with *ortho*-phtaldialdehyde (OPA) containing 1% 3-MPA and 9-Fluormethylencarbonylchlorid (FMOC). Samples and standards were spiked with norvaline and sarcosine as internal standards. Detection was carried out with a fluorescence detector (FLD-3400RS), detecting secondary amines and sarcosine at Ex 266 nm/Em 305 nm and primary amines and norvaline at Ex 340 nm/Em 450 nm (Hofer et al., 2018).

2.7. Rate calculations and elemental balancing

For determination of the volumetric acetate formation rate (r_{ace}) and biomass formation rate (r_X), the dilution rate D was multiplied with the average acetate and biomass concentration from at least two data points from steady state conditions. Volumetric fructose and formate consumption rates were calculated by multiplying the feed concentration with the dilution rate.

The calculation of volumetric gas uptake rates XUR [mmol L^{-1} h^{-1}] from GC data was performed as follows: the molar fraction of N_2, CO, CO_2 and H_2 were determined in the reactor exhaust gas. Mass balances were established assuming that no N_2 is consumed (NTR = 0). The reactor gas inflow rate was measured and balancing of N_2 allowed calculation of the reactor exhaust gas flow. All gas transfer rates [mmol L^{-1} h^{-1}] were calculated from the volumetric gas inflow rate q_{in} [L h^{-1}], the molar fraction of the respective gas in the inlet gas ($y_{x,in}$) and exhaust gas ($y_{x,out}$) and the calculated volumetric exhaust gas flow q_{out} [L h^{-1}] as follows:

$$XUR = \frac{q_{in} \cdot y_{x,in} - q_{out} \cdot y_{x,out}}{V_{molar} \cdot V_{reactor}}$$

where V_{molar} [L mol^{-1}] is the molar gas volume at 20 °C and 1.013 bar, and $V_{reactor}$ [L] is the filling volume of the bioreactor. No corrections for the dissolved gas in the harvest flow or the CO_2/HCO_3^- equilibrium were applied.

To perform elemental balancing, a carbon content of 45% (w/w) was used for *A. woodii* biomass (Godley et al., 1990). A degree of reduction (DoR) of 4.15 mol electrons per mol of carbon was assumed for biomass

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

(Rittmann et al., 2012). Yeast extract was neglected for the calculation of the C- and DoR-balance.

2.8. *Transcriptome and proteome analysis*

2.8.1. *Sampling*

After a culture reached steady state conditions, a 5 mL sample was withdrawn, divided into 1 mL aliquots, and centrifuged for 1 min at 11,000 g and −4 °C. After removing the supernatant, the pellet was snap-frozen in liquid nitrogen. The samples were stored at −80 °C until further processing.

2.8.2. *RNA extraction and RNAseq*

Cell pellets were resuspended in 1 ml Invitrogen TRIzol Reagent (ThermoFisher Scientific, Waltham/MA, USA) and lyzed using a Fast-Prep-24 (MP Biomedicals, Santa Ana/CA, USA) with 0.37 g of glass beads (0.1 mm diameter) at 6 m/s for 40 s. Samples were incubated at room temperature for 5 min and then centrifuged at 12,000 g for 5 min. 750 µl of the supernatant were mixed with 750 µl ethanol and RNA isolated using the Direct-zol RNA Miniprep Kit (Zymo Research, Irvine/CA, USA) according to the manufacturer's instructions. This Kit includes a DNAse treatment step. Integrity, Quality, and Quantity of the isolated RNA was checked on a 5200 Fragment Analyzer System (Agilent, Santa Clara/CA, USA) and a NanoDrop One UV–Vis Spectrophotometer (ThemoFisher Scientific, Waltham/, MA, USA).

Preparation of RNA libraries and Sequencing on an Illumina Next-Seq, v2.5, 1 × 75bp with a target of 5 million reads per sample was performed by Microsynth (Microsynth AG, Balgach, Switzerland). Transcriptomic data were uploaded to the SRA database (accession number PRJNA737050).

2.8.3. *Transcriptome analysis*

The obtained reads were inspected using FastQC v0.11.5, analyzed and quality trimmed using Trimmomatic (Bolger et al., 2014). A reference transcriptome was extracted from the reference genome of *A. woodii* DSM 1030 (Poehlein et al., 2012) and the corresponding gff file using gffread v0.12.7 (Pertea and Pertea, 2020). A salmon index was created by using salmon 1.4.0 (Patro et al., 2017) on the reference transcriptome and the samples were quantified, including the –gcBias flag to account for the effects of sample specific biases such as fragment-level GC bias. The quantification results were imported into the R environment and analyzed with the DESeq2 (Love et al., 2014) package and the packages tximport, ggplo2, vsn, pheatmap, RColorBrewer and limma (R Core Team, 2013; Soneson et al., 2016; Zhu et al., 2019).

2.8.4. *Sample preparation for proteome analysis*

For the proteomics analysis, samples were lysed in 100 µl of the lysis buffer (100 mM Tris pH 8.6, 1% sodium dodecyl-sulphate (SDS), 40 mM chloroacetamide and 10 mM (tris(2-carboxyethyl)phosphine) (TCEP)) followed by three cycles of sonication (15 s per cycle, 20% amplitude). Lysates were then spun down for 5 min at 14,000 *g* and 100 µg of protein (after protein estimation) were precipitated overnight using acetone. The following day, protein pellets were re-solubilized in 50 µl of 25% trifluoroethanol (TFE) in 100 mM Tris (pH = 8.6), after which solution was diluted to 10% TFE with 100 mM ammonium-bicarbonate and subjected to overnight digestion with trypsin (1:67 ratio protein to trypsin). Resulting peptide mixture was offline desalted, then chromatographically separated using an Ultimate 3000 RCS Nano Dionex system equipped with an Ionopticks Aurora Series UHPLC C18 column (250 mm × 75 µm, 1.6 µm) (Ionopticks, Australia). Solvent A was 0.1% formic acid in water and solvent B acetonitrile containing 0.1% formic acid. Total run per sample was 136.5 min with the following gradient: 0–5.5 min: 2% B; 5.5–65.5 min: 2–17% B; 65.5–95.5 min: 25–37% B, 105.5–115.5 min: 37–95% B, 115.5–125.5 min: 95% B; 125.5–126.5 min: 95-2% B; 126.5–136.5 min: 2% B at a flow rate of 400 nl/min and 50 °C. Peptides were measured on the timsTOF mass spectrometer (Bruker Daltonics, Germany) that was operated in positive mode with enabled trapped Ion Mobility Spectrometry (TIMS) at 100% duty cycle (100 ms cycle time). Scan mode was set to parallel accumulation–serial fragmentation (PASEF) for the scan range of 100–1700 m/z. Source capillary voltage was set to 1500 V and dry gas flow to 3 L/min at 180 °C.

2.8.5. *Statistical analysis of proteome data*

Raw data processing was carried out using MaxQuant (v1.6.17.0) (Cox and Mann, 2008; Tyanova et al., 2016a). Database matching was performed against a genome predicted publicly available *A. woodii* protein database (GCF_000247605.1_ASM24760v1; downloaded on February 24, 2021, 3546 entries). For peptide as well as protein matching, false discovery rate was set to 1%, minimum peptide length was set to six and up to two mis-cleavages were allowed. Oxidation of methionine was set as variable and carbamidomethylation as fixed modification. Match between run feature was enabled for the match window of 1 min and alignment window of 20 min.

Resulting table of protein "Intensities" was then imported to Perseus (v 1.6.14.0) (Tyanova et al., 2016b), where data was transformed, normalized (mean subtraction per column) and grouped. Matrix was then filtered to keep only those proteins with reported values in at least three replicates in at least one of the groups. Missing values were consequently imputed from normal distribution (downshift 1.8, width 3) and pairwise Student's t-tests were carried out between the groups with multi-testinging correction (permutation-based FDR <5%). All the raw proteomics data including the search parameters, database used as well as results output was deposited to the ProteomeXchange Consortium via the PRIDE partner repository (Perez-Riverol et al., 2019) with the dataset identifier PXD026569.

2.9. *Metabolic modelling and FBA*

A previously published *A. woodii* core model (Koch et al., 2019) with 118 reactions was used to perform flux balance analysis (FBA) and model intracellular fluxes. Energy conservation and redox balancing were considered by the model as previously described for *A. woodii* (Schuchmann and Müller, 2014). A biomass composition similar to *Clostridium autoethanogenum* was assumed (Valgepea et al., 2017). The *CellNetAnalyzer* toolbox (Klamt et al., 2007; von Kamp et al., 2017) was used for flux balance analysis (FBA). The experimentally determined specific rates for biomass formation, substrate uptake (formate, fructose, CO, CO_2, H_2) and metabolite formation were used to constrain the model. The determined rates beared redundancies with respect to carbon and redox balances in the metabolic model. Consequently, fluxes were corrected prior to the FBA calculations to obtain a consistent system. This was achieved by minimizing the relative changes in the measured rates needed to yield a consistent flux scenario ("Check feasibility" function in *CellNetAnalyzer*). The growth rate was kept constant under all conditions. As the objective function, the pseudo reaction that quantifies the non-growth associated ATP maintenance (NGAM) demand was maximized. Thereby, intracellular flux distributions and an upper bound of ATP available for NGAM processes could be described. To estimate variations in fluxes, flux variability analysis was performed with and without NGAM as the constraint.

3. Results and discussion

3.1. *Chemostat cultivations to investigate physiology and systems level response of A. woodii*

Even though formate has previously been used as a growth substrate for *A. woodii*, there is no quantitative data set describing the physiological behavior during growth on this one-carbon compound. Therefore, steady state cultivation data were obtained by establishing

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

chemostat cultivations of *A. woodii* at a dilution rate of 0.05 h^{-1}. This growth rate previously proved to be the half-maximum growth rate of *A. woodii* for the substrate conditions investigated here (Novak et al., 2021). A total of six different conditions were tested, including formate, H_2/CO_2, fructose, formate + H_2/CO_2, formate + $H_2/CO_2/CO$ and formate + fructose (Table 1). For each steady state condition, the physiological behavior was investigated and complemented by transcriptomics (RNAseq) and proteomics analyses.

3.1.1. *A. woodii efficiently utilizes formate for growth and acetate production in chemostat cultures*

In a first step, we evaluated whether chemostat cultivations of *A. woodii* with formate as the sole carbon and energy source can be established. To that end, batch cultures were transferred to continuous mode by supplying a feed containing 100 mM formate at a rate of 0.05 h^{-1}. Indeed, cells completely consumed formate and stable steady state formation of biomass and acetate production was observed. However, carbon-limited cultures under these conditions showed extremely low biomass concentrations of 0.14 g L^{-1} (see Table 1). To evaluate whether the biomass concentration and the volumetric formate turnover could be boosted, the formate concentration in the feed was increased to 200 mM. As a result, the volumetric formate uptake rate roughly increased by 2-fold (Table 2) and formate was fully consumed. Although the formate concentration was doubled, steady state concentrations for biomass and acetate only increased by 57 and 71% to 0.22 g L^{-1} and 3.17 g L^{-1}, respectively (Table 1). A possible explanation is the yeast extract that was used in the same concentration for all cultivations. Yeast extract has been shown to increase biomass and acetate yields of *A. woodii* cultures (Tschech and Pfennig, 1984), thus leading to an overestimation of acetate yields on formate.

During growth on formate, 4 mol formate are required to form 1 mol acetate (Bertsch and Müller, 2015a). The remaining carbon is oxidized to 2 mol CO_2 to provide enough reduction power for carbon fixation in the WLP (Fig. 4). The reaction stoichiometry therefore shows a carbon efficiency of only 50%. Moreover, the ATP yield for stoichiometric formate conversion is only 0.3 ATP per mol acetate (Müller, 2019). Using 200 mM formate, near stoichiometric conversion of formate to acetate and CO_2 was observed. The yields for acetate and CO_2 were 0.26 mol mol^{-1} and 0.47 mol mol^{-1}, respectively. As the carbon and DoE balance were closed (Table 1), the influence of yeast extract seemed negligible for growth on 200 mM formate. The high amount of carbon liberated as CO_2 combined with a low ATP yield make formate a challenging anaerobic substrate and provide a possible explanation for the low biomass yields observed during growth of *A. woodii* on formate (Table 2).

Generally, specific rates can be used to extract information on the physiological behavior and boundaries of a microbial cell factory utilizing a given substrate. Additionally, sound physiological data are crucial to obtain useful results from metabolic modelling (section 3.3). Therefore, we next analyzed cell specific formate uptake and acetate formation rates of *A. woodii* during growth on formate. Despite the low biomass yields, a specific formate uptake rate of 47 mmol g^{-1} h^{-1} corresponding to ~1 g g^{-1} h^{-1} was observed for the 200 mM chemostat. Moreover, the specific production rate for acetate was ~12 mmol g^{-1} h^{-1}. Combined with the favorable acetate yields, these values indicate that *A. woodii* can convert formate to acetate at high rates and efficiency. *A. woodii* could therefore be an interesting organism for anaerobic formate-based bioproduction. Additionally, the data obtained here provide a reference data set under well-defined conditions which can be used for comparison of *A. woodii* physiology during formate utilization to other substrates.

3.1.2. *Quantitative comparison shows similarities of formate and autotrophic H_2/CO_2 utilization but not with heterotrophic fructose utilization*

To obtain a picture of the physiological behavior of *A. woodii* during growth on formate, H_2/CO_2 and fructose were studied as reference substrates for autotrophic and heterotrophic fermentation. Fermentation data for H_2/CO_2 (Kantzow et al., 2015; Novak et al., 2021) and fructose (Godley et al., 1990) have already been reported. However, to ensure comparability and to obtain samples for the transcriptomic and proteomic analyses we decided to generate the reference data for both substrates using the same cultivation conditions and media as for the formate cultivations. Changing only the respective carbon and energy sources, chemostat cultivations for both substrates were successfully established. Gas-limited cultures on H_2/CO_2 with a gas containing 60% H_2 and 9.5% CO_2 showed 4-fold higher biomass concentrations compared to the 200 mM formate culture (Table 1). This increase is also reflected in the biomass yield which was ~50% higher for H_2/CO_2 compared to formate (g mol^{-1} basis, Table 1). A higher biomass yield is

Table 1
Yield coefficients for biomass formation and acetate production during growth of *A. woodii* on single and mixed substrates. Mean values and standard deviations were calculated from biological triplicates.

Growth condition	Product concentration [g L^{-1}]		Acetate yields [mol mol^{-1} substrate]					Biomass yields [g mol^{-1} substrate]					Balances [%]	
	Acetate	Biomass	$Y_{Ace/For}$	$Y_{Ace/Fru}$	$Y_{Ace/CO2}$	$Y_{Ace/H2}$	$Y_{Ace/sumC}$	$Y_{X/For}$	$Y_{X/Fru}$	$Y_{X/CO2}$	$Y_{X/H2}$	$Y_{X/sumC}$	C	DoR
Formate (102 mM)	1.85 ± 0.07	0.135 ± 0.011	0.312 ± 0.011	–	–	–	0.312 ± 0.011	1.32 ± 0.11	–	–	–	1.32 ± 0.11	128 ± 7	154 ± 7
Formate (200 mM)	3.17 ± 0.05	0.22 ± 0.01	0.263 ± 0.005	–	–	–	0.263 ± 0.005	1.09 ± 0.03	–	–	–	1.09 ± 0.03	104 ± 1	113 ± 2
$H_2:CO_2$ (60:9.5)	15.3 ± 1.0	0.93 ± 0.10	–	–	0.449 ± 0.021	0.228 ± 0.012	0.449 ± 0.021	–	–	1.62 ± 0.14	0.82 ± 0.08	1.62 ± 0.14	96 ± 5	99 ± 5
Fructose (34.1 ± 0.5 mM)	4.77 ± 0.09	1.76 ± 0.07	–	2.33 ± 0.01	–	–	2.33 ± 0.01	–	51.5 ± 0.9	–	–	51.5 ± 0.9	108 ± 4	111 ± 2
Formate (100 mM) $H_2:CO_2$ (60:9.5)	16.3 ± 1.1	0.98 ± 0.06	2.71 ± 0.18	–	0.529 ± 0.015	0.244 ± 0.012	0.442 ± 0.011	9.8 ± 0.6	–	1.92 ± 0.12	0.89 ± 0.03	1.59 ± 0.07	94 ± 3	98 ± 1
Formate (100 mM) $H_2:CO_2:CO$ (60:9.5:10.6)	16.2 ± 1.2	1.28 ± 0.01	2.70 ± 0.20	–	0.676 ± 0.003	0.285 ± 0.003	0.445 ± 0.001	12.8 ± 0.1	–	3.21 ± 0.24	1.35 ± 0.08	2.11 ± 0.15	97 ± 1	102 ± 1
Formate (202 ± 2 mM) Fructose (35.3 ± 1.1 mML)	7.88 ± 0.16	1.94 ± 0.03	0.649 ± 0.008	3.72 ± 0.04	–	–	0.553 ± 0.005	9.60 ± 0.05	55.0 ± 1.0	–	–	8.17 ± 0.02	106 ± 1	108 ± 1
$H_2:CO_2:CO$ (60:9.5:10.6) [a]	17.8 ± 1.5	1.54 ± 0.12	–	–	0.43 ± 0.01	0.23 ± 0.01	–	–	–	–	–	–	93 ± 1	87 ± 1

[a] Data from (Novak et al., 2021).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table 2
Volumetric and specific substrate uptake and acetate formation rates from *A. woodii* chemostat cultivations on single and mixed substrates. Mean values and standard deviations were calculated from biological triplicates. Specific rates are uptake rates (negative values for qCO_2 indicate production) for formate, fructose and H_2, CO_2 and CO, and production rates for acetate. Volumetricc rates are uptake rates (negative values for CO_2UR indicate production) for formate, fructose and H_2, CO_2 and CO, and production rates for acetate.

Growth condition	Dilution rate [h^{-1}]	Specific rates [mmol g^{-1} h^{-1}]						Volumetric rates [mmol L^{-1} h^{-1}]					
		qAce	qFor	qFru	qCO_2	qH_2	qCO	rAce	rFru	rFor	CO2UR	HUR	COUR
Formate (102 mM)	0.052 ± 0.002	11.89 ± 1.34	39.4 ± 4.8	–	−28.5 ± 1.3	−9.0 ± 2.7	–	1.59 ± 0.09	–	5.28 ± 0.21	−3.33 ± 0.26	−1.21 ± 0.38	–
Formate (200 mM)	0.051 ± 0.001	12.2 ± 0.3	47.0 ± 1.2	–	−22.5 ± 0.7	–	–	2.67 ± 0.06	–	10.3 ± 0.1	−4.92 ± 0.03	–	–
$H_2:CO_2$ (60:9.5)	0.054 ± 0.002	15.1 ± 1.1	–	–	33.8 ± 3.2	66.9 ± 7.5	–	14.0 ± 0.7	–	–	30.9 ± 0.9	60.9 ± 4.6	–
Fructose (34.1 ± 0.5 mM)	0.049 ± 0.001	2.20 ± 0.04	–	0.95 ± 0.02	0.07 ± 0.22	–	–	3.87 ± 0.06	1.66 ± 0.02	–	0.13 ± 0.40	–	–
Formate (100 mM) $H_2:CO_2$ (60:9.5)	0.055 ± 0.002	15.1 ± 0.3	5.6 ± 0.5	–	33.8 ± 3.2	66.9 ± 7.5	–	15.0 ± 0.9	–	5.47 ± 0.16	30.9 ± 0.9	60.8 ± 4.5	–
Formate (100 mM) $H_2:CO_2:CO$ (60:9.5:10.6)	0.054 ± 0.002	11.6 ± 0.4	4.3 ± 0.2	–	17.2 ± 0.7	40.9 ± 2.0	4.6 ± 0.4	15.2 ± 0.6	–	5.51 ± 0.20	22.0 ± 1.0	52.3 ± 1.5	5.9 ± 0.6
Formate (202 ± 2 mM) Fructose (35.3 ± 1.1 mM)	0.050 ± 0.001	3.40 ± 0.04	5.24.0 ± 0.01	0.92 ± 0.02	−2.7 ± 0.1	–	–	6.60 ± 0.17	1.77 ± 0.06	10.2 ± 0.2	−5.17 ± 0.15	–	–
$H_2:CO_2:CO$ (60:9.5:10.6) [a]	0.05	10.8 ± 0.4	–	–	17.1 ± 1.2	46.6 ± 2.2	8.7 ± 0.2	16.6 ± 0.7	–	–	26.2 ± 0.2	71.6 ± 2.3	13.4 ± 1.4

[a] Data from (Novak et al., 2021).

consistent with the higher carbon efficiency observed during growth on H_2/CO_2. Analogously to biomass, the acetate titer and yield were 4.8-fold and 70% higher for H_2/CO_2. However, the specific acetate productivity on formate was almost equal to the value of H_2/CO_2 (q_{Ace} 80% for formate compared to H_2/CO_2). Likewise, the specific uptake rates for all three carbon and energy sources were within the same range i.e., ~47 mmol g^{-1} h^{-1} for formate compared to ~34 mmol g^{-1} h^{-1} and 67 mmol g^{-1} h^{-1} for CO_2 and H_2, respectively. Hence, despite the drastic differences in titers and volumetric rates for the two conditions, a comparable physiological behavior could be observed. Regardless of the distinct differences to autotrophic growth, formate utilization of *A. woodii* via the WLP shares significant similarities with H_2/CO_2 utilization.

Next, fructose-grown chemostat cultures were compared to growth on formate. To ensure comparability, an equimolar amount of carbon (33.3 mM fructose) was used. Because fructose contains significantly more energy than formate (combustion energies of 2,930 kJ/mol and 245 kJ/mol, respectively), heterotrophic cultures have significantly higher ATP gains compared to formate cultures (see also 3.3.2). Consequently, the steady state biomass concentration of the fructose fermentation was 8-fold higher than for formate. As a result, the molar (g mol^{-1}) and C-molar (g C-mol^{-1}) biomass yields increased 50- and 8-fold, respectively. In contrast, the acetate titer was only increased by 50% (Table 1). The observed acetate yield of 2.33 mol mol^{-1} is in good agreement with previously reported values but it is only 78% of the theoretical maximum for homoacetogenic acetate production (Beck et al., 2019; Braun and Gottschalk, 1981; Wiechmann et al., 2020). Theoretically, acetogens could convert 1 mol of a hexose into 3 mol acetate by using CO_2 and reduction equivalents produced during sugar catabolism in the WLP for carbon fixation. However, the theoretical value does not consider that growth requires significant portions of cellular resources. As previously observed, heterotrophic cultures required CO_2 to fully consume fructose (Godley et al., 1990). The reason for this behavior is rooted in the function of the WLP as an electron sink. In the absence of sufficient amounts of CO_2, *A. woodii* cannot re-oxidize electron carriers.

A comparison of the specific rates showed that due to the high biomass concentrations of the heterotrophic cultures, substrate uptake rates for formate were ~50-fold higher compared to fructose. Similarly, the biomass specific acetate formation rate was 5.6-fold higher for formate. Combined, these observations could indicate that cells use high specific substrate turnover of low energy substrates to provide enough ATP for growth and maintenance of biomass, especially under anaerobic conditions (Rintala et al., 2008).

In conclusion, the physiological behavior of *A. woodii* during growth on formate and fructose differed significantly. These observations are in line with the different properties of the two substrates and the metabolic pathways involved in their utilization.

3.1.3. *Metabolic flexibility and robustness of A. woodii is revealed by efficient co-utilization formate with gaseous or heterotrophic substrates*

In a future bioeconomy, flexible substrate co-utilization is anticipated to become an important feature of microbial production hosts. Consequently, we investigated the ability of *A. woodii* to utilize formate together with H_2/CO_2 or fructose. Furthermore, a gas containing additional CO was tested for co-utilization with formate.

For gaseous co-substrates, the same gas-limited conditions as for the H_2/CO_2 condition described above were used to establish steady state continuous cultures. The liquid feed was supplied at $D = 0.05\ h^{-1}$ and initially contained formate (100 mM). Both conditions, formate + H_2/CO_2 and formate + $H_2/CO_2/CO$ could successfully be established in carbon-limited chemostats. Generally, a stabilizing effect of formate on fermentation of gaseous substrates was noticed. In our previous study, autotrophic cultures were sensitive to perturbations (e.g. antifoam pump failure) that caused product titers and gas uptake rates to fluctuate and prohibited cultures to maintain steady state conditions (Novak et al., 2021).

A comparison to the 100 mM formate and the H_2/CO_2 culture showed that for formate + H_2/CO_2 the steady state acetate concentration was only 5% lower compared to the sum of acetate for the individual substrates (16.3 and 17.2 g L^{-1}, respectively) (Table 1). Similarly, the volumetric acetate productivity for formate + H_2/CO_2 was comparable to the sum of the induvial substrates. The biomass concentration for formate + H_2/CO_2 increased 5% compared to H_2/CO_2 but was 10% lower compared to the sum of the individual substrates. Compared to H_2/CO_2, formate addition to cultures did not affect the acetate yield (0.44 mol mol^{-1} for formate + H_2/CO_2, Table 1). This observation, however, might be a result of the relatively small contribution of formate to the final acetate titer as underlined by the specific substrate utilization rates. Although both cultures (formate and formate + $H_2/$

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

CO_2) were provided with the same volumetric formate feeding rate, the higher biomass concentration for formate + H_2/CO_2 decreased the specific formate uptake rate to only 14% of the value observed for unitrophic formate utilization (Table 2). Moreover, the presence of formate reduced the specific uptake rates for H_2 and CO_2 by 7 and 16%, respectively. These shifts in utilization of gaseous substrates indicate that formate partially replaced H_2 and CO_2 as energy and carbon sources under limiting chemostat conditions. Despite substrate co-utilization, the total specific acetate productivity for formate + H_2/CO_2 did not change compared to the H_2/CO_2 culture. This physiological behavior is in line with the observation that q_{Ace} was comparable when formate and H_2/CO_2 were used individually. Overall, the flexible adjustments of substrate utilization are quite remarkable given that co-utilization of formate and H_2/CO_2 requires the hydrogen-dependent CO_2 reductase (HDCR) of *A. woodii* to react to changing concentrations of educts and products of the reaction. For serum bottle cultures grown on increasing concentrations of formate and a H_2/CO_2 gas phase an initial lag phase was found. The length of the lag phase depended on the initial formate concentration (Fig. S1), indicating kinetic and thermodynamic regulation as the key determinant of flow at the HDCR. Regardless of potential initial inhibitions, all batch cultures eventually fully consumed formate and H_2/CO_2 from the gas phase and produced biomass and acetate.

Next, we expanded the investigation of co-utilization of formate and gaseous substrates to a gas stream which contained CO in addition to H_2/CO_2. In our previous study, we had shown that batch cultures containing formate + $H_2/CO_2/CO$ first co-utilized formate and CO, and upon limitation of CO in the liquid culture, also H_2/CO_2 and CO. However, no information on the ability of *A. woodii* to co-utilize all four carbon and energy sources was gained. Chemostat cultures using 100 mM formate and $H_2/CO_2/CO$ in the gas stream showed a 30% higher biomass concentration compared to formate + H_2/CO_2 (Table 1). CO is an intermediate of the WLP obtained by reducing CO_2 with the low potential reduction equivalent ferredoxin. When supplying limiting amounts of CO with CO_2/H_2, less ferredoxin is oxidized for CO_2 reduction (Novak et al., 2021). Hence, more ferredoxin can be allocated to the energy conserving reaction of the Rnf complex (section 3.3.2) (Schuchmann and Müller, 2014), improving overall bioenergetics, hence enabling an increase in biomass concentration (Bertsch and Müller, 2015a, 2015b; Novak et al., 2021).

In contrast to the biomass concentration, acetate titer, productivity and yield did not change for formate + $H_2/CO_2/CO$ compared to formate + H_2/CO_2. Consequently, q_{Ace} decreased by 23% because of the higher biomass concentration (Table 2). A similar decrease was observed for the specific uptake rates for formate, H_2 and CO_2. For q_{H2} and q_{CO2}, the decrease was 34% and 40%, respectively. These changes indicate that in addition to the higher biomass concentration, cell specific rates were decreased by the presence of CO. Analogous to formate addition to H_2/CO_2 cultures, CO replaced H_2 and CO_2 as carbon and energy sources in the formate + $H_2/CO_2/CO$ culture (see section 3.3 below for details on intracellular flux distributions). Collectively, *A. woodii* proved to be extremely flexible in utilizing up to four different carbon and energy sources simultaneously, including three gaseous substrates. Future work towards bioprocess development could further explore this important metabolic feature by varying formate and gas utilization and by expanding the system to other gas compositions. Combined, these measures will allow controlling specific uptake rates for individual substrates, which can be used as a strategy to control metabolism and intracellular fluxes (section 3.3.2 and 3.4).

Another substrate that could help to improve bioenergetics is the utilization of hexose sugars in combination with formate. *A. woodii* is known to efficiently utilize H_2/CO_2 and fructose but utilization together with formate has so far not been described. To that end, we aimed to establish continuous cultures fed with equimolar amounts of carbon from formate (200 mM = 200 mM carbon) and fructose (33.3 mM = 200 mM carbon). Carbon-limited steady states could be achieved which showed biomass and acetate concentrations of 1.9 and 7.9 g L^{-1}, respectively. These values are in both cases close to the sum of the cultures for formate and fructose utilization alone (Table 1). As formate was completely consumed, carbon catabolite repression (CCR) could not be observed even with the relatively high fructose concentrations. CCR was previously found to prevent co-consumption of methanol and glucose in *Eubacterium limosum* (Loubiere et al., 1992) and to cause poor H_2/CO_2 consumption by *Clostridium aceticum* (Braun and Gottschalk, 1981) and *Moorella thermoacetica* (Huang et al., 2012) in the presence of fructose or glucose, respectively. In contrast, in other acetogens including *Clostridium ljungdahlii*, gas consumption was not inhibited by fructose (Jones et al., 2016).

Furthermore, formate was able to replace the need for CO_2 to establish fructose-limited steady state conditions. As expected for formate utilization, co-utilization with fructose resulted in CO_2 production from formate (Table 2), indicating that formate served both as carbon and energy source in addition to fructose. The cell-specific fructose uptake rate was comparable for formate + fructose to fructose alone but due to the higher biomass concentration q_{For} was only 11% of the value for unitrophic formate utilization. Nevertheless, the presence of formate increased q_{Ace} by 50% compared to fructose utilization alone. This observation shows how co-utilization of a high and low energy substrate can improve physiological performance data beyond what is possible for unitrophic substrate utilization. While formate utilization improved cell-specific acetate productivity, fructose improved the overall bioenergetics. Consequently, addition of fructose or glucose could be used to improve the bioenergetics of formate-utilizing *A. woodii* in the future and enable shifting carbon flux away from acetate in metabolically engineered strains (section 3.4). In summary, the physiological data presented here demonstrate *A. woodii* as a robust host for formate-based bioconversion which can efficiently co-utilize autotrophic and heterotrophic carbon and energy sources in combination with formate.

3.2. A transcriptomic and proteomic analysis highlights substrate-specific regulation of pathways

The physiological study highlighted the ability of *A. woodii* to utilize different carbon and energy sources simultaneously. The individual substrates investigated are assimilated via separate pathways, provide different amounts of energy, and donate electrons with different potentials. However, biomass and acetate were the only products detected in the culture broth. The question arises how the cell flexibly adapts to various substrates while maintaining the same product spectrum.

Both formate and fructose were supplied to the culture by a liquid feed, possibly requiring the expression of genes for the uptake of these carbon sources from the medium. On the other hand, the one-carbon substrates formate, CO_2, and CO are all assimilated via the WLP, suggesting a similar overall metabolism and gene expression for autotrophic and formatotrophic growth. As formate is an intermediate of the methyl-branch of the WLP, simultaneous oxidation of formate to CO_2 and H_2 and activation of formate to formyl-THF might require fine-tuning of enzyme expression in the WLP. On top of that, co-utilization of formate with other substrates might require the activation of additional gene clusters e.g., for CO oxidation or fructose uptake. To thoroughly understand the utilization of different substrates and to examine the adaptation of the gene expression of *A. woodii*, we analyzed the transcriptome and proteome under different growth conditions.

RNA-seq allowed the detection of 3662 transcripts, covering the whole 3546 protein-coding ORF of the genome of *A. woodii*. Additionally, our investigation of the proteome is the first published LC-MS/MS-based proteome study for *A. woodii* and enabled the detection and quantification of 1881 polypeptides from all samples altogether. We performed a differential expression analysis of all six growth conditions using formate as reference condition (Fig. 1). The transcriptome of different growth conditions was investigated for similarities by identifying changes in the transcription of common genes.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

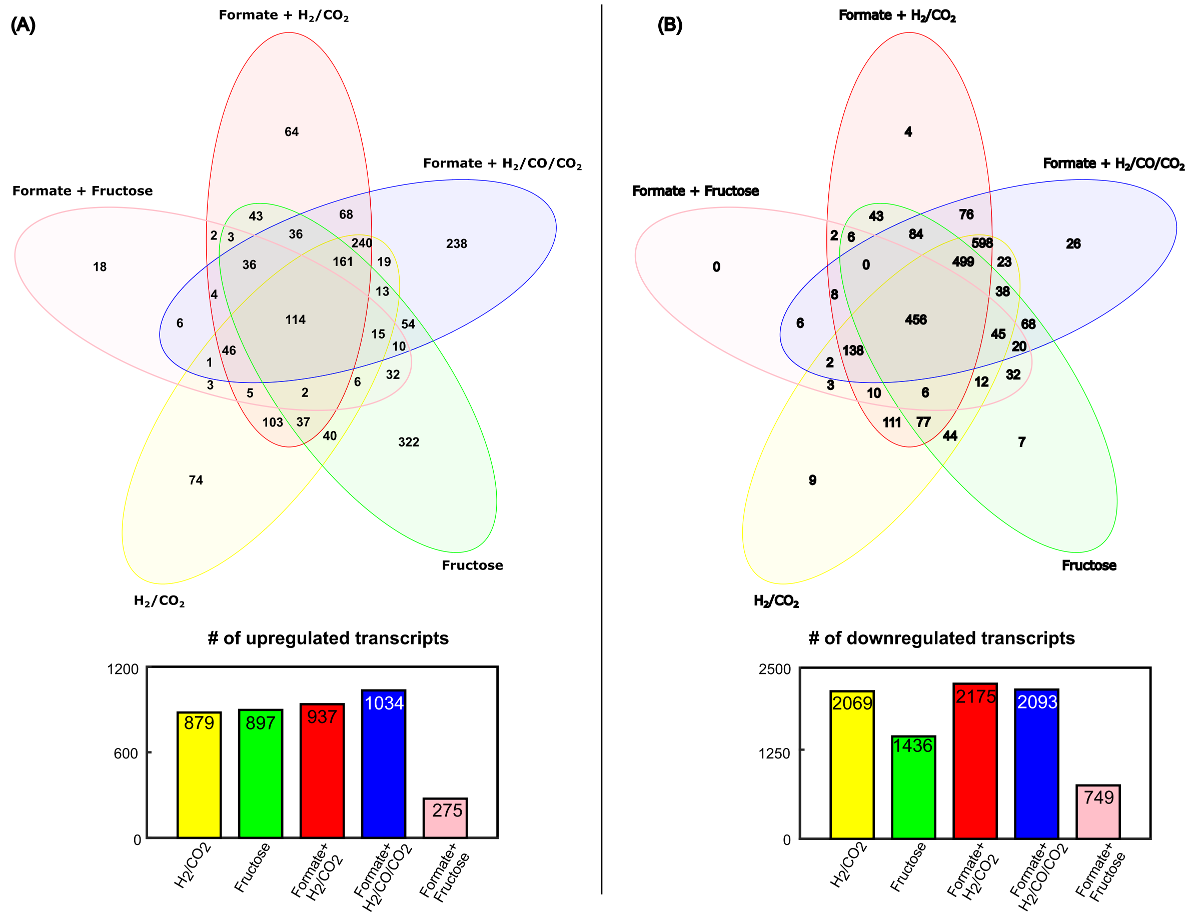


Fig. 1. Differential gene transcription analysis of *A. woodii* for growth on the six different substrate conditions tested. Figures indicate the number of differentially expressed genes as compared to growth on formate. (A) Number of upregulated genes; (B) number of downregulated genes.

456 genes were down-regulated and 114 genes were upregulated on a transcription level on all other growth conditions compared to formate, indicating an adaptation of the cell to the utilization of formate with high specific rates (47 mmol g^{-1} h^{-1}, section 3.1.1). For the other three growth conditions on one-carbon substrates (H_2/CO_2, formate + H_2/CO_2 and formate + $H_2/CO/CO_2$), 240 common genes were up-regulated and 598 common genes down-regulated as compared to growth on only formate. Under all three conditions, gaseous substrates were utilized and elevated acetate concentrations of ~15–17 g L^{-1} were reached (Table 1) which might trigger changes in the expression of common genes. With 264 and 329 exclusive changes in transcript levels, respectively, the growth conditions formate + $H_2/CO/CO_2$ and fructose indicated the most distinct adaptation to the respective substrates. In contrast, the growth on formate + fructose revealed only 275 up-regulated genes and 749 down-regulated genes as compared to growth on formate, indicating a similar transcriptome for these conditions.

This first differential analysis of the transcriptome suggests that *A. woodii* adapts to the supply of different substrates on a global level. In a previous proteome analysis of *A. woodii,* enzymes linked to glycolysis and the WLP were found to be expressed differently on fructose and H_2/CO_2 (Poehlein et al., 2012). In contrast, previous -omics studies of the acetogens *Clostridium ljungdahlii* and *Clostridium autoethanogenum* suggested a stable expression of genes under various substrate uptake and product formation rates, indicating a robust expression as the basis for metabolic flexibility of acetogens (Richter et al., 2016; Valgepea et al., 2017).

We next aimed to understand central adaptations in the expression of genes and proteins that are involved in acetate and biomass formation. To that end, we focused on the two central pathways that lead to acetyl-CoA synthesis (WLP, glycolysis + pyruvate decarboxylation), on enzymes involved in the supply of reduced reduction equivalents (electron bifurcating hydrogenase HydABCD, HDCR, Rnf complex) and on proteins that catalyze the conservation of energy (ATPase, Pyruvate kinase, Phosphoglycerate kinase).

3.2.1. *The WLP is highly expressed during growth on one-carbon substrates*

The WLP is responsible for the assimilation of the one-carbon substrates formate, CO_2, and CO. All one-carbon substrates were found to be taken up by *A. woodii* with high specific rates (section 3.1.2), indicating a highly active WLP.

Indeed, gene clusters of the methyl-branch of the WLP and the carbon monoxide dehydrogenase/acetyl-CoA synthetase (CODH/ACS) (Poehlein et al., 2012) were among the 20 genes that showed the highest intermediate normalized mean read count across all growth conditions (Fig. S2). A third highly transcribed gene cluster is the electron-bifurcating hydrogenase which is responsible for the oxidation of H_2 and supplying the WLP with reduced ferredoxin (Fd^{2-}) and NADH. To compare the expression of the WLP between growth conditions and to highlight up- and downregulation, differential transcriptome and proteome analyses were performed (Fig. 2).

Comparing the gene expression during growth on formate and H_2/CO_2 revealed comparable levels of expression for most genes of the WLP. However, the monofunctional CODH CooS and the neighboring iron-sulfur protein (Awo_c19060) and the CODH/Acetyl-CoA synthase β subunit AcsB2 were upregulated under autotrophic growth, both on a transcriptome and proteome level. This upregulation of genes of the WLP may be linked to the higher specific acetate formation rate (25% increased) during autotrophic growth compared to formatotrophic

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

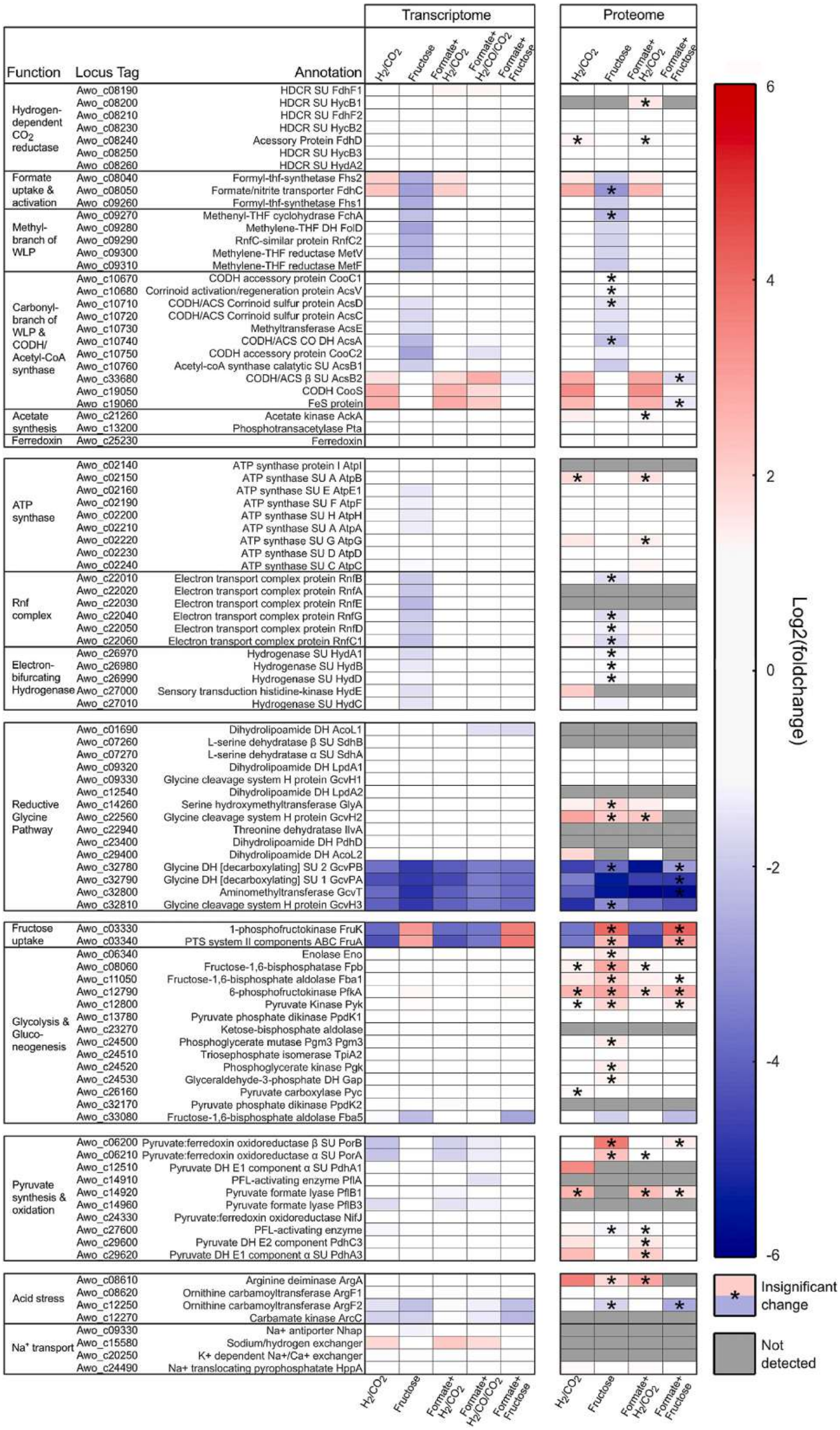


Fig. 2. Differential transcriptomic and proteomic analysis for growth of *A. woodii* on single and mixed substrates. Formate was chosen as the reference growth condition. SU = subunit, DH = dehydrogenase, PFL = pyruvate formate lyase.

growth (Table 2). A comparison of the proteome also revealed 2- to 4-fold higher levels for the ATPase subunits C and G, the electron transport complex protein RnfC1 and the Hydrogenase associated proteins B and E for autotrophic growth.

Heterotrophic growth on fructose showed several adaptations of the expression of genes of the WLP compared to growth on formate: transcript and protein levels of genes involved in the methyl-branch and genes of the CODH/Acetyl-CoA synthase were found in significantly lower levels (3- to 7-fold lower transcript levels, 3- to 5-fold lower protein levels) (Fig. 2). Transcriptome analysis additionally revealed 2-

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

to 5-fold lower transcript levels of several ATPase, Rnf complex and Hydrogenase subunit genes. The low specific acetate formation during growth on fructose requires only a small contribution of the WLP to acetate formation (section 3.3). As the WLP genes are among the highest transcribed genes in *A. woodii*, cutting back their expression potentially allows the cell to save energy. In contrast, the HDCR was not regulated on a transcript and protein level. H_2 serves as a substrate of the HDCR and needs to be provided via oxidation of NADH and Fd^{2-} during growth on fructose (Wiechmann et al., 2020). Providing a high HDCR activity might therefore be necessary to capture intracellular H_2 and funnel it towards the WLP.

The comparison of gene expression for growth on formate and formate + H_2/CO_2 revealed changes similar to the comparison of gene expression for growth on formate and autotrophic growth on H_2/CO_2. In detail, the same genes of the WLP showed higher transcript and protein levels (Fig. 2). The similar specific H_2 uptake rates for autotrophic growth and growth on formate + H_2/CO_2 (Table 2) may dictate the regulation of the WLP. Moreover, the co-utilization of formate + $H_2/CO/CO_2$ revealed similar transcriptional changes of the WLP genes as observed for the other growth conditions with gaseous substrates with a few notable exceptions. Compared to growth on H_2/CO_2 and formate + H_2/CO_2, formate + $H_2/CO/CO_2$ showed a weaker up-regulation of the monofunctional CODH CooS, a lower transcript number of the CODH accessory protein Cooc2 and a higher transcript level of the CODH/ACS β subunit AcsB2. These findings indicate an adaptation to the external supply of CO. By reducing the level of transcripts for CODH functions, an unnecessary assignment of Fd^{2-} for the reduction of CO_2 to CO might be avoided (section 3.3.1).

Compared to growth on formate, higher transcript and protein levels of the formyl-THF-synthetase Fhs2 and the putative formate transporter FdhC were found for H_2/CO_2 and formate + H_2/CO_2 but not for formate + $H_2/CO/CO_2$. In addition to a similar expression of Fhs2 and FdhC, the conditions formate and formate + $H_2/CO/CO_2$ share a specific acetate formation rate of ~12 mmol g^{-1} h^{-1} which is 20% lower than for growth on H_2/CO_2 and formate + H_2/CO_2 (Table 2). Potentially, a faster formation of acetate causes the intracellular pH to drop. A lowered intracellular pH at a constant external pH impairs the diffusive uptake of formic acid from the medium which is supposedly facilitated by FdhC (Moon et al., 2021). At low intracellular pH values, stronger expression of FdhC might therefore enable faster equilibration of internal and external formate pools. In contrast, stronger expression of Fhs2 might allow faster formate activation to keep the intracellular formate pool low and to avoid formate efflux into the medium.

A. woodii grown on formate + fructose showed almost no adaptations of the expression of the WLP compared to formate-grown cultures. This finding agrees well with the observation that there were few overall changes in the transcriptome between the two growth conditions (Fig. 1). A high level of WLP enzymes might be required to fully convert the additionally supplied formate.

The expression of the WLP adapts to the supplied carbon sources despite carrying the strongest transcribed genes throughout all growth conditions on single carbon sources. Surprisingly, lower transcript and protein levels of AcsB2 and CooS were found for growth on formate compared to growth on gaseous substrates, indicating a potentially lower activity of the acetyl-CoA generating step of the WLP. However, the expression of genes of the methyl-branch was not changed during growth on formate. The methyl-branch of the WLP can also fuel the reductive glycine pathway (rGLY), another carbon fixation pathway recently described in the acetogen *Clostridium drakei* (Song et al., 2020). Therefore, we next examined the expression level of genes of the rGLY.

3.2.2. *The glycine cleavage system is upregulated during growth on formate*

Compared to cultures grown under all other conditions, formate-grown cultures showed a strong expression of the genes GcvPA, GcvPB, GcvT and GcvH3 with an 11- to 32-fold increase on a transcript level and a 7- to 55-fold increase on a protein level. These genes are neighboring in the genome of *A. woodii* (see Fig. 3) and are part of the glycine cleavage system (GCS) that allows both glycine synthesis from one-carbon substrates and the degradation of glycine. The GCS forms a functional subunit of the reductive glycine pathway (rGLY) (Bar-Even et al., 2013).

An upregulation of the GCS could indicate a potential flux of one carbon metabolites to glycine. Nevertheless, the remaining genes of the rGLY that allow pyruvate synthesis from glycine via serine (Fig. 3) were not upregulated during growth on formate (Fig. 2). It is therefore hard to estimate the actual activity of the rGLY. To investigate the function of the GCS, we examined if glycine was accumulating in the medium or taken up. During growth on formate and fructose the small amounts of glycine that were provided with the feed (via yeast extract) were almost completely consumed (Table 3). The highest specific uptake rate was determined for growth on formate, being 8-fold higher than for growth on fructose. The transcript changes observed for the GCS genes *gcvPA, gcvPB, gcvT* and *gcvH3* were drastically higher than in a recent study that compared the transcriptome of *A. woodii* batch cultivations on formate, fructose and H_2/CO_2 (Moon et al., 2021). Our carbon-limited continuous cultivation on formate with a high ratio of glycine to biomass potentially triggered the increased expression of the GCS to enable the uptake of glycine as an additional carbon and energy source. For growth on fructose, a bigger share of glycine might have been directly incorporated into biomass, rendering an upregulation of the GCS obsolete.

Despite the role of the GCS in glycine degradation, *A. woodii* could theoretically use the rGLY for assimilation of single carbon sources. The acetogen *C. drakei* was shown to possess all genes of the rGLY as well. For *C. drakei*, metabolic modelling suggested an almost negligible flux through the rGLY and CO_2 fixation mainly via the WLP and the glycine synthase-reductase pathway (GSRP) (Song et al., 2020). *A. woodii* is lacking a glycine synthase-reductase and can therefore only fix CO_2 via the rGLY and the WLP. To further analyze the importance of glycine-forming pathways in the one carbon assimilation of *A. woodii* and other acetogens, further investigations are necessary.

3.2.3. *Pyruvate synthesis is not regulated on a gene expression level*

During growth on one-carbon substrates such as formate, CO and CO_2, the synthesis of pyruvate is based on the carboxylation of acetyl-CoA. Pyruvate is an important metabolite that links CO_2 fixation to major biomass-forming reactions. The genome of *A. woodii* encodes two pyruvate:ferredoxin oxidoreductases (PFOR), three pyruvate formate lyases (PFL), and three pyruvate dehydrogenases (PDH) that could be responsible for the carboxylation of acetyl-CoA. The Fd^{2-}-consuming PFOR reaction is considered the active pyruvate synthesis route in acetogens (Furdui and Ragsdale, 2000). However, the PFL reaction would allow acetogens to synthesize pyruvate from formate, thereby saving valuable Fd^{2-} for other reactions. To investigate the role of different pyruvate-forming enzymes under different growth conditions, we examined their expression level on a transcript and protein level.

We calculated the transcript level for each gene by multipying the intermediate normalized mean read count with the read length (75 bp) and dividing it by the length of the respective gene. The PFOR gene *nifJ* was identified as the highest transcribed gene for pyruvate synthesis: the transcript level of 255 for *nifJ* was ~5-fold higher than the transcript levels of the highest transcribed PDH genes *pdhC3, pdhB3* and *pdhA3,* ~7-fold higher than the PFOR subunit genes *porA* and *porB,* and ~90-fold higher than the highest transcribed PFL gene *pflB3*. NifJ was also measured in all proteome samples and showed the highest intensity of all pyruvate-forming enzymes.

A comparison of transcript and protein levels revealed stable expression of NifJ under all growth conditions. Stable and strong expression of *nifJ* indicates that pyruvate synthesis via acetyl-CoA might indeed rely on the PFOR reaction. To verify the importance of NifJ or other pathways for pyruvate synthesis in more detail, activities from PFOR, PDH and PFL enzymes in crude extracts could be analyzed for different conditions using *in vitro* assays. Additionally, the generation of

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

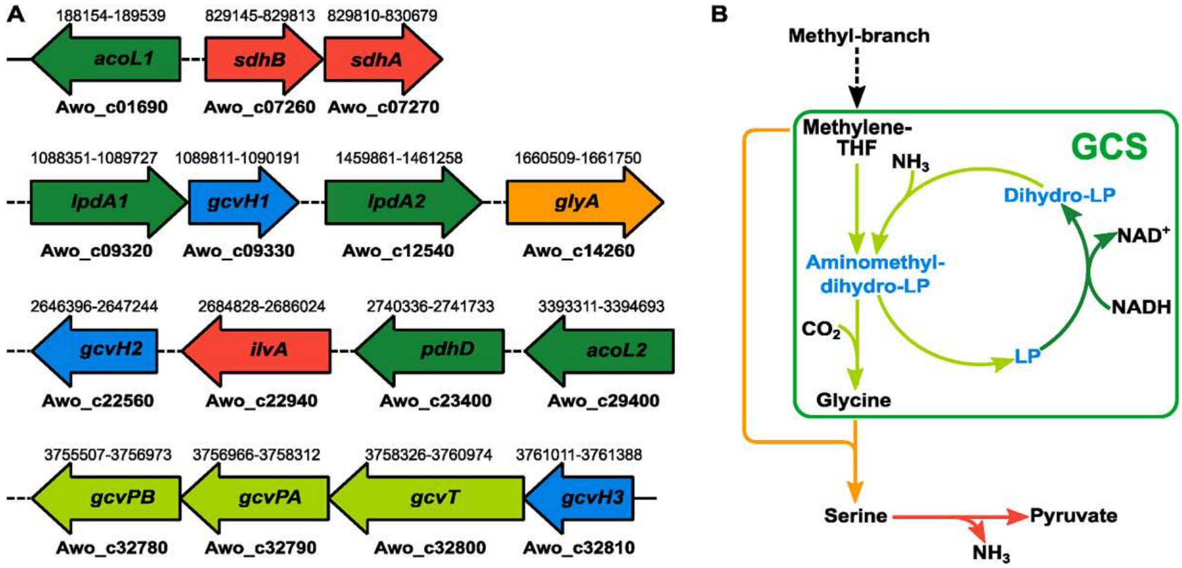


Fig. 3. Arrangement and function of the genes associated to the reductive glycine pathway of *A. woodii*. A: Genomic organization of rGLY genes. acoL1, acoL2, lpdA1, lpdA2 and pdhD: Dihydrolipoamide dehydrogenase; sdhB and sdhA: L-serine dehydratase subunits; gcvH1, gcvH2 and gcvH3: Glycine cleavage system H-Protein; glyA: Serine hydroxymethyltransferase; ilvA: Threonine dehydratase; gcvPB and gcvPA: Glycine dehydrogenase subunits; gcvT: Aminomethyltransferase; B: Metabolic map of the rGLY. LP: Lipoprotein.

Table 3
Glycine uptake rates for A. woodii chemostat cultivations on single substrates. A glycine concentration of 307 ± 12 μM was determined for the feed. r_{Gly}, volumetric glycine uptake rate, g_{Gly}, specific glycine uptake rate. Dilution rates and biomass concentrations from Tables 1 and 2

Condition	Formate	H_2/CO_2	Fructose
r_{Gly} [μmol L^{-1} h^{-1}]	14.1	4.0	14.7
q_{Gly} [μmol g^{-1} h^{-1}]	67	4.3	8.5
Relative glycine consumption [%]	89	24	96

a *nifJ* deletion mutant might be interesting for future studies as it would enable pyruvate formation via the PFL reaction. The PFL route via formate and acetyl-CoA was already shown to be feasible *in vivo* in anaerobically-grown *E. coli* (Zelcbuch et al., 2016).

3.2.4. *Fructose supply activates uptake via the phosphotransferase system*

In contrast to growth on one-carbon substrates, the direct supply of fructose allows the synthesis of pyruvate via glycolysis. Fructose utilization is initiated by substrate uptake via the phosphotransferase system (PTS). During growth on fructose and formate + fructose, a 2.5- to 4-fold increase in the transcript levels of *fruA* and *fruK* was noted compared to growth on formate (Fig. 2). Therefore, the expression of the fructose uptake system seems to be linked to the presence of fructose in the growth medium.

3.2.5. *Site product formation is regulated during growth on different carbon sources*

A. woodii is equipped with the genetic information to produce several fermentation products including lactate and ethanol. The genes for lactate utilization and potential lactate formation are organized in the operon lctCDEF. In our study, the highest transcript levels of the lctCDEF operon were found in formate-grown cells with 13- to 20-fold higher transcription compared to growth on fructose (Fig. S3). A down-regulation of this operon was previously described for growth on fructose, H_2/CO_2, methanol, and ethylene glycol. The operon was shown to be activated by the presence of D- and L-lactate, leading to a ~300-fold increase in the transcription level (Schoelmerich et al., 2018). Hence, the operon was not fully activated for growth on formate and the low transcription underlines why no lactate was formed under any of the growth conditions.

Ethanol formation from fructose has been described for phosphate-limited cultures of *A. woodii* (Buschhorn et al., 1989). Current studies highlighted the importance of the bi-functional alcohol dehydrogenase AdhE for ethanol formation and consumption of *A. woodii* (Trifunović et al., 2020). Interestingly, ~4-fold higher *adhE* transcript levels were found in fructose-grown cells as compared to formate-grown cells. The AdhE protein was detected in all analyzed fructose samples (Fig. S3). Ethanol formation from acetyl-CoA could serve as an alternative electron sink to the WLP. Interestingly, ethanol was neither detected in the culture supernatant of our study nor in studies where the re-oxidation of reduction equivalents via the WLP was blocked (Godley et al., 1990; Wiechmann et al., 2020). Further research is needed to understand the relevance of AdhE during growth of *A. woodii* on fructose.

3.2.6. *A single ferredoxin is dominantly expressed*

Ferredoxin serves as a carrier for electrons with a low reduction potential. Fd^{2-} is critical for the reduction of CO_2 in the carbonyl-branch of the WLP but also plays a crucial role in building up the sodium gradient at the Rnf complex which drives ATP synthesis. The genome of *A. woodii* encodes eleven potential ferredoxins. Among those, Awo_c25230 is transcribed with the highest intermediate normalized mean read count and without changes between different growth conditions (Fig. 2). Awo_c25230 was also detected in all proteome measurements, underlining the abundance and importance of this ferredoxin.

3.3. *Metabolic modelling highlights major differences of intracellular flux levels and directionality*

The data obtained from chemostat cultivations were used to perform flux balance analysis with the stoichiometric core model of *A. woodii*. Metabolic modelling enabled us to investigate which pathways are involved in the utilization and co-utilization of substrates, to highlight reactions that build the fundament for the high metabolic flexibility of *A. woodii,* and to access the turnover of reduction equivalents and the available energy. Maximizing non-growth associated ATP maintenance (NGAM) was used as an objective function. To check flux variations, flux variability analysis (FVA) was additionally performed (File S2).

Modelling the growth of *A. woodii* on formate suggested a high flux of

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

13 mmol L^{-1} h^{-1} through the WLP (Fig. 4). Formate was partly degraded by the HDCR to supply the cell with CO_2 and H_2. 94% of the acetyl-CoA formed was converted to acetate to gain ATP via substrate level phosphorylation. The remaining share of acetyl-CoA was fueling anabolic reactions of the cell. During growth on formate, the overall supply of electrons was insufficient to reduce all CO_2 formed from formate degradation, leading to a net release of CO_2. When the uptake of glycine from the medium was neglected, the flux from methylene-THF to glycine via the GCS was 100-fold smaller than the flux through the WLP, indicating a minor role of the GCS as carbon fixation pathway under these conditions. Including a glycine uptake rate of 67 µmol g^{-1} h^{-1} (Table 3) as an additional constraint for FBA did not increase degradation of glycine

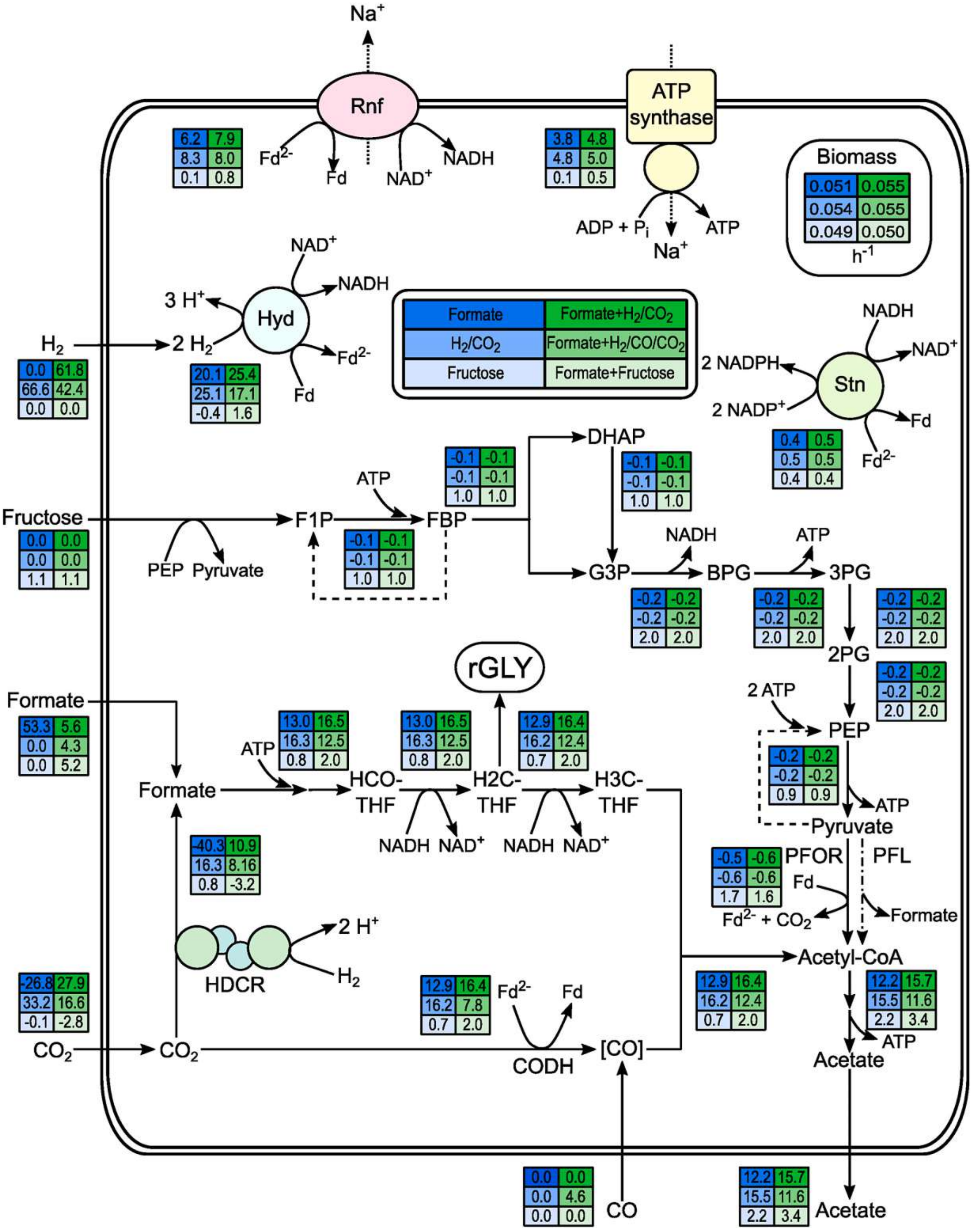


Fig. 4. Metabolic flux map of *A. woodii* for growth on different substrates. Boxed values show flux levels in mmol g^{-1} h^{-1} for six different growth conditions. rGLY = reductive glycine pathway, Rnf = Rnf complex, Hyd = electron-bifurcating hydrogenase, Stn = *Sporomusa* type Nfn (Kremp et al., 2020), HDCR = hydrogen-dependent carbon dioxide reductase, PFOR = pyruvate:ferredoxin oxidoreductase, PFL = Pyruvate formate lyase, F1P = fructose-1-phosphate, FBP = fructose bisphosphate, DHAP = dihydroxyacetone phosphate, G3P = glyceraldehyde-3-phosphate, BPG = bis-phosphoglycerate, 3 PG = 3-phosphoglycerate, 2 PG = 2-phosphoglycerate, PEP = phosphoenolpyruvate.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

via the GCS (data not shown), which is in stark contrast to the distinct upregulation of GCS gene expression (section 3.2.2). However, the glycine uptake rate was ~800-fold lower compared to the formate uptake rate, providing a potential explanation for the negligible influence on intracellular flux distributions.

Growth on H_2/CO_2 led to a 25% higher flux through the WLP as compared to growth on formate, agreeing well with the higher expression of WLP genes (section 3.2.1). Formate was formed from H_2 and CO_2 by the HDCR instead of being lysed. Apart from the HDCR reaction, the metabolic fluxes were similar to formatotrophic growth.

Heterotrophic growth on fructose varied significantly from the formatotrophic and autotrophic condition. The degradation of fructose via glycolysis provided energy via substrate level phosphorylation and central carbon metabolites for anabolic reactions, making costly gluconeogenesis obsolete. The direction of the hydrogenase reaction was inverted to form H_2 from NADH and Fd^{2-}. The flux through the WLP was only 66% of the initial fructose uptake rate, underlining the role of glycolysis as the central energy-providing pathway. Despite a lower expression of WLP gene clusters (section 3.2.1), the WLP allowed full reoxidation of reduction equivalents obtained from fructose degradation.

The growth on formate + H_2/CO_2 revealed flexible adaptation of fluxes to the available substrates. External formate fueled the internal metabolite pool and reduced the reaction rate of the HDCR by 33% compared to growth on H_2/CO_2. However, the overall flux to acetyl-CoA remained unchanged. For growth on formate + $H_2/CO/CO_2$, formate and CO fueled the carbonyl-branch and the methyl-branch of the WLP.

Growth of *A. woodii* on formate + fructose represented a metabolic mixture of the growth on the isolated carbon sources. Fructose was degraded via glycolysis to provide energy and metabolites for the anabolism. Simultaneously, formate was used as a substrate of the WLP, providing additional energy. The flux through the WLP to acetyl-CoA was ~280% of the flux during growth on fructose, thereby contributing significantly to acetate production and energy generation. This increased flux was supported by a stronger expression of WLP genes as compared to growth on fructose (section 3.2.1). However, the activity of the WLP was still 6.5-fold lower than for growth on formate. When co-utilizing formate and fructose, 39% of formate was activated by the formyl-THF-synthetase while the remaining part was converted to CO_2 and H_2. Hence, the reduction equivalents obtained from glycolysis allowed to utilize more formate in the methyl-branch as compared to growth on formate. Formate was shown before to serve as an electron acceptor in the methyl-branch of the WLP (Wolin et al., 2003; Wiechmann et al., 2020). However, the CO_2 released from formate and fructose degradation equaled the amount that was released for growth on formate (section 3.1.3).

Comparing modelling results for different growth conditions highlights reactions and pathway functionalities and explains the metabolic flexibility of *A. woodii*: the direction of the hydrogenase reaction is adapted to allow either oxidation of H_2 for the supply of reduction equivalents or generation of H_2 for formate synthesis via the HDCR. Excess formate is lysed by the HDCR to release H_2 to provide additional reduction power. While the expression level of the HDCR is not adapted (section 3.2.1), the fluxes from H_2/CO_2 to formate vary greatly in direction and overall level for the different growth conditions. The stable expression of the HDCR might enable complete and fast utilization of the electron donors formate and H_2.

3.3.1. *A. woodii utilizes the WLP for energy conservation and as a redox sink*

Generally, the WLP serves as an electron sink (Schuchmann and Müller, 2014). In *A. woodii*, electrons are provided by oxidation of H_2, fructose, CO or formate-derived H_2. We investigated the provision and consumption of reduction equivalents by eight key reactions of the central carbon metabolism to underline differences in their fate and the contribution of the WLP to their reoxidation (Fig. 5).

During growth on formate and H_2/CO_2, all Fd^{2-} is supplied by the oxidation of H_2 through the hydrogenase HydABCD. 75% of NADH is obtained by H_2 oxidation and the remaining part by oxidation of Fd^{2-}. Nearly all NADH and the remaining Fd^{2-} are consumed in the WLP while the gluconeogenetic reactions (PFOR and G3P DH) and NADPH

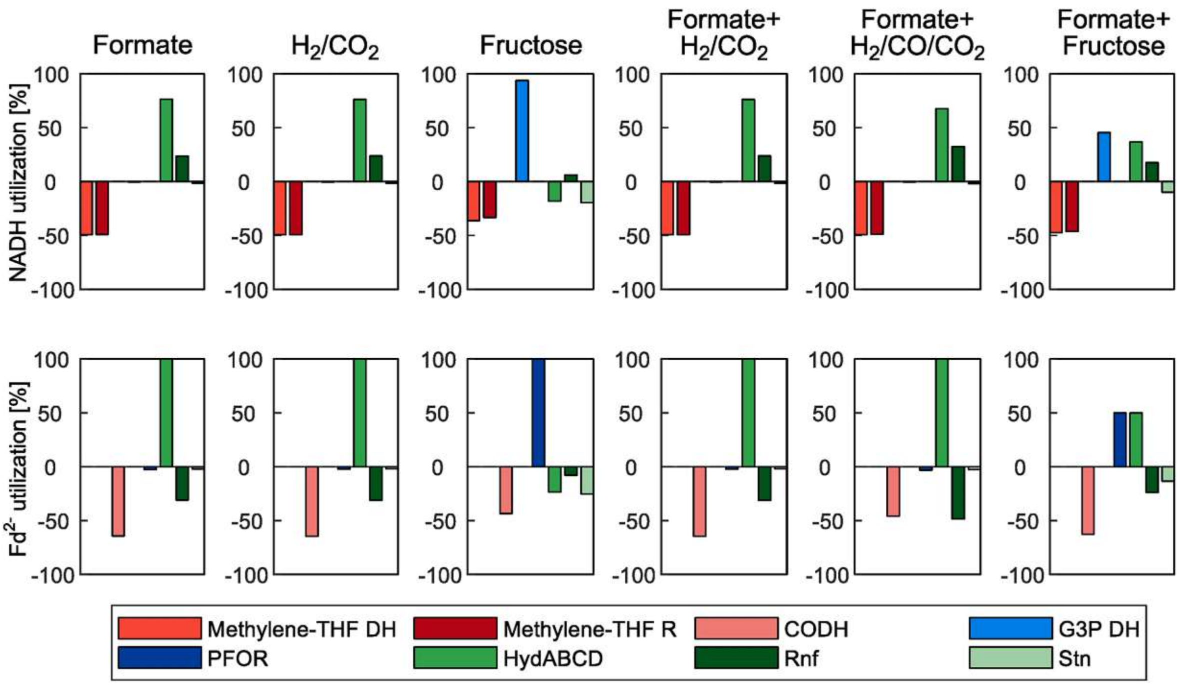


Fig. 5. Relative contribution of central redox reactions to the NADH and Fd^{2-} pool under different substrate conditions. Negative rates indicate oxidation of the respective reduction equivalent, positive rates reduction. Reaction rates were normalized by the total rate of reduction of each reduction equivalent by all eight considered reactions. DH = dehydrogenase, R = reductase, G3P = 3-phosphoglycerate, PFOR = pyruvate:ferredoxin oxidoreductase, HydABCD = electron bifurcating hydrogenase, Rnf = Rnf complex, Stn = *Sporomusa* type Nfn. The rates for the generation and consumption of NADH and Fd^{2-} were derived from the metabolic modelling results (Fig. 4).

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

forming reaction (Stn) contribute to a negligible amount. The overall contribution of reactions to the supply and oxidation of reduction equivalents is identical for formatotrophic growth and autotrophic growth on H_2/CO_2.

In contrast, PFOR is the sole source of Fd^{2-} during heterotrophic growth on fructose. The electron-bifurcating hydrogenase operates in reverse direction and consumes NADH and Fd^{2-} to supply H_2 for formate formation by the HDCR. Only 6% of NADH are generated from Fd^{2-} via the Rnf complex, correlating with the lower expression of Rnf complex genes (section 3.2.1). Compared to growth on formate, a bigger share of NADH (19%) and Fd^{2-} (25%) is consumed via Stn, guiding reduction power towards NADPH-consuming anabolic reactions (Kremp et al., 2020). Nevertheless, the WLP still functions as an electron sink during growth on fructose. Due to the minor contribution of the Rnf complex in redox balancing, the function of the WLP and Rnf complex in energy conservation becomes subordinate.

Co-utilization of formate + H_2/CO_2 relies on the oxidation of H_2 to provide electrons, similar to the respective unitrophic growth conditions, resulting in an identical share of redox reactions in the conversion of reduction equivalents. When formate + $H_2/CO/CO_2$ are co-utilized, Fd^{2-} is also exclusively generated from oxidation of H_2. However, less Fd^{2-} is consumed by the CODH and 48% of Fd^{2-} is oxidized via the Rnf complex.

During growth on formate + fructose, electrons are equally provided from formate-derived H_2 and glycolytic redox reactions. Half of the Fd^{2-} is provided by the electron-bifurcating hydrogenase and the other half by the PFOR. 18% of NADH are generated by the reaction of the Rnf complex, indicating a stronger role of the WLP in energy conservation compared to growth on fructose. A notable share of electrons is transferred to NADPH via Stn to fuel anabolic reactions.

A. woodii can flexibly adapt to the electrons supplied by substrate oxidation. While re-oxidation of reduction equivalents by the WLP is crucial during heterotrophic growth, energy conservation via the Rnf complex was found to play only a minor role. All electrons and carbon sources are used to produce acetyl-CoA via the glycolysis and the WLP, yielding acetate as the only product in addition to biomass.

Establishing *A. woodii* as a platform organism for formate-based bioproduction requires an extension of the product spectrum to industrially relevant bulk and commodity chemicals. To understand the formation of acetate as the sole product and to determine potential limitations for the synthesis of other products, we examined the ATP availability of *A. woodii* for growth on different substrates.

3.3.2. *Co-utilization of substrates allows modulation of ATP availability*

The synthesis of ATP from acetyl-CoA plays an important role in the energy household of *A. woodii*: during autotrophic and formatotrophic growth. During growth on fructose, acetate formation from pyruvate-derived acetyl-CoA enables synthesis of additional ATP. Acetyl-CoA is also an intermediary metabolite for the synthesis of industrially relevant products (Vees et al., 2020). However, withdrawing acetyl-CoA for the synthesis of metabolites other than acetate is only possible if net energy conservation of the cell is ensured. Consequently, product yields are constrained by ATP availability (Bertsch and Müller, 2015b).

To investigate the available energy of *A. woodii*, the ATP gain per acetate was calculated (Table 4). The metabolic model of *A. woodii* allowed consideration of energetic costs for gluconeogenesis which are necessary to evaluate growth-coupled production of metabolites. A second approach to access the ATP availability of the cell is to determe the non-growth associated ATP maintenance (NGAM) which reflects the surplus ATP that cannot be associated to growth.

During growth on formate, the lowest ATP/acetate ratio of 0.2 was determined, being 91% of the value for growth on H_2/CO_2 (Table 5). Both for formatotrophic and autotrophic growth, a high specific acetate formation rate was required to supply the cell with sufficient energy. Consequently, little energy could be invested in energy-negative production pathways. The NGAM value for growth on formate was 56% lower than for autotrophic growth on H_2/CO_2. The lower ATP maintenance costs might be linked to the low acetate concentration of 3.1 g L^{-1} for the formatotrophic culture as compared to the high acetate concentration of 15.3 g L^{-1} for the autotrophic cultivation on H_2/CO_2. We found an inhibitory effect of high acetate concentrations on growth of *A. woodii* in our previous study (Novak et al., 2021) and a link between ATP maintenance costs and acetate concentrations has already been postulated for other acetogens (Valgepea et al., 2017). Integrating glycine uptake into the FBA for growth on formate increased the NGAM value by ~5%, indicating that glycine uptake may have increased ATP availability when formate was the carbon source.

Growth of *A. woodii* on fructose allowed the highest ATP generation per formed acetate, being 5.2-fold higher than for growth on formate. The NGAM value for growth on fructose was 2.5-fold higher than for growth on formate and was comparable to the value for autotrophic growth on H_2/CO_2 (Table 4). As the acetate concentration of 4.8 g L^{-1} for heterotrophic growth was comparable to the concentration for growth on formate, product inhibition is unlikely to be responsible for the increased ATP maintenance costs. Analyzing samples from fructose-limited chemostat cultivations under the microscope showed *A. woodii* cells to be noticeably motile in contrast to cells from formatotrophic and autotrophic cultures. Thus, the higher ATP availability during growth on fructose might have enabled energy investment into inefficient cellular functions such as movement, thereby increasing maintenance costs.

For co-utilization of formate + H_2/CO_2, the same ATP gain per acetate was determined as for autotrophic growth on H_2/CO_2. The NGAM values for both conditions were also comparable. As the supply route of reduction equivalents is the same (section 3.3.1), assuming a similar energy state of the cell seems plausible. For growth on formate + $H_2/CO_2/CO$, a 41% higher ATP gain was observed as compared to formate + H_2/CO_2. Improved bioenergetics through supply of CO is in line with previous reports for *A. woodii* and other acetogens (Hermann et al., 2020; Novak et al., 2021). Supplying CO directly to the carbonyl-branch of the WLP enabled lower specific flux through the WLP while maintaining the same specific flux through the Rnf complex (section 3.3.1).

When growing on formate + fructose, a 3.8-fold higher ATP gain per acetate was observed compared to growth on formate. By providing fructose in a molar concentration six times lower compared to formate, the energetic availability of the cell could be drastically increased. The computed NGAM value was 2.9-fold higher than for growth on formate, indicating that the addition of fructose increased the amount of ATP wasted. Indeed, cells grown on formate + fructose were also motile when inspected under the microscope.

In conclusion, the mixing of low energy substrates, e.g., formate and H_2, with energy-rich substrates, e.g., CO and fructose, allows improving the bioenergetics of *A. woodii*. This additional energy could ultimately

Table 4
ATP yields per formed acetate and non-growth associated ATP maintenance (NGAM) for growth of *A. woodii* on different substrate mixtures. For calculation of the yields, the reactions of the following enzymes were considered: Formyl-THF synthetase, ATPase, acetate kinase, PTS fructose, 6-phosphofructokinase, phosphoglycerate kinase, pyruvate kinase.

Condition	Formate	H_2/CO_2	Fructose	Formate + H_2/CO_2	Formate + $H_2/CO/CO_2$	Formate + Fructose
ATP Yield (mol ATP/mol Acetate)	0.20	0.22	1.04	0.22	0.31	0.75
Non-growth associated ATP maintenance (NGAM)	0.6473[a]	1.4549	1.6445	1.4460	1.6225	1.8528

[a] The NGAM value was 0.681 mmol g^{-1} h^{-1} when a glycine uptake rate of 0.067 mmol g^{-1} h^{-1} was used as an additional constraint in FBA.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table 5
Comparison of the energetic efficiency of different acetogens and microorganisms during growth and product formation on one carbon substrates and sugar substrates. Energetic efficiency was calculated according to (Claassens et al., 2019). rGLY (eng.) refers to engineered, synthetic formatotrophy.

Organism	Substrate	Active assimilation pathway	Product(s)	Energetic efficiency [%]	Reference
A. woodii	Formate	WLP	Acetate	84.2	Tschech and Pfennig (1984)
A. woodii DSM1030		WLP	Acetate	93.7	This study
A. woodii DSM1030		WLP	Acetate	80.7	Moon et al. (2021)
E. coli		rGLY (eng.)	Biomass	18.8	Kim et al. (2020)
C. necator		rGLY (eng.)	Biomass	21.2	Claassens et al. (2020)
C. necator		Calvin cycle	Biomass	23.7	Claassens et al. (2020)
Pseudomonas 1		Serine cycle	Biomass	52.2	Goldberg et al. (1976)
Methylotroph strain M2		Serine cycle	Biomass	60.4	Kelly et al. (1994)
A. woodii DSM1030	Formate + H_2/CO_2	WLP	Acetate	75.2	This study
A. woodii DSM1030	Formate + $H_2/CO_2/CO$	WLP	Acetate	78.2	This study
A. woodii DSM1030	H_2/CO_2	WLP	Acetate	76.3	This study
C. ljungdahlii		WLP	Acetate, Ethanol, 2,3BDO (check)	80.8	Hermann et al. (2020)
C. autoethanogenum	$H_2/CO/CO_2$	WLP	Acetate, Ethanol, 2,3BDO	79.6	Valgepea et al. (2017)
C. ljungdahlii		WLP	Acetate, Ethanol, 2,3BDO	76.3	Hermann et al. (2020)
A. woodii DSM1030		WLP	Acetate	75.1	Novak et al. (2021)
C. ljungdahlii	CO	WLP	Acetate, Ethanol, 2,3BDO (check)	72.1	Hermann et al. (2020)
C. autoethanogenum		WLP	Acetate, Ethanol, 2,3BDO	69.6	Valgepea et al. (2018)
P. pastoris PC4002	Methanol	DHA cycle	Biomass	36.1	Shay et al. (1987)
P. pastoris CBS 704		DHA cycle	Biomass	36.9	Hazeu and Donker (1983)
Pseudomonas 1		Serine cycle	Biomass	34.1	Goldberg et al. (1976)
Pseudomonas C		Serine cycle	Biomass	48.7	Battat et al. (1974)
B. methanolicus MGA3		RuMP cycle	Biomass	43.3	Schendel et al. (1990)
B. methanolicus MGA3		RuMP cycle	Biomass	45.1	Pluschkell and Flickinger (2002)
E. coli		rGLY	Biomass	11.8	Kim et al. (2020)
A. woodii DSM1030+		WLP	Acetate	82.7	Tschech and Pfennig (1984)
Acetobacterium sp.+		WLP	Acetate	87.0	Bainotti et al. (1998)
Acetobacterium sp.	Methanol + Formate	WLP	Biomass	74.4	Bainotti and Nishio (2000)
C. acetobutylicum CAB1060	Glucose	–	Butanol, Ethanol	65.4	Nguyen et al. (2018)
S. cerevisiae		–	Ethanol, Glycerol	81.6	Nissen et al. (1997)
A. woodii DSM1030	Fructose	WLP	Acetate	65.0	Godley et al. (1990)
A. woodii DSM1030		WLP	Acetate	69.2	This study
A. woodii DSM1030	Formate + Fructose	WLP	Acetate	74.3	This study

be used to synthesize relevant bulk chemicals from sustainable carbon and energy sources such as H_2, formate, CO, and CO_2. Genetic tools for the plasmid-based overexpression of pathways and for the deletion of genes in *A. woodii* are available (Beck et al., 2019; Hoffmeister et al., 2016; Wiechmann et al., 2020), enabling to broaden the product spectrum in the future. As *A. woodii* naturally directs all excess carbon and reduction equivalents towards the formation of acetate, additional genetic modifications might be needed to improve heterologous product synthesis.

3.4. *Formate-based bioproduction achieves excellent energy efficiencies*

One-carbon sources such as CO_2, CO, formate and methanol are considered as promising platform feedstocks of the future bioeconomy (Bar-Even et al., 2013; Claassens et al., 2019; Cotton et al., 2020). Table 5 shows the energetic efficiencies obtained for *A. woodii* and different substrates used in this study and compares them to values reported for acetogens and other common microbial hosts. Overall, acetogens show superior energetic efficiency on all substrates analyzed, with the highest values for one carbon substrates. Compared to gaseous substrates, formate as a miscible one carbon substrate showed even higher energetic efficiencies. The high efficiency make formate a promising substrate for bioproduction of chemicals and fuels. However, acetogens such as *A. woodii* and *E. limosum* (Litty and Müller, 2021) form acetate as the exclusive product during growth on formate. Metabolic engineering of *A. woodii* might allow to implement strategies for production of other metabolites. Production of ethanol and lactate from the substrates considered in this study were analyzed using flux balance analysis (Table 6). To that end, NGAM values and specific substrate uptake rates from experimental results (Fig. 4 and Table 4) were used as model inputs together with a specific growth rate of 0.02 h^{-1} and maximizing the ethanol or lactate yield was used as objective function. Metabolic modelling showed that by smart co-feeding of substrates flexible production scenarios for formate upgrading with high energy efficiencies can be devised. Supplementation of relatively minor quantities of CO and fructose increases the energy availability (section 3.3.2), and thus enables exclusive formation of ethanol or lactate without co-production of acetate. Co-utilization of H_2 allows complete fixation of CO_2, improving the carbon efficiency of the process or even facilitating net CO_2 uptake. The superior energy efficiency and straight forward substrate co-utilization make *A. woodii* an excellent candidate for formate-based bioproduction.

4. Conclusion

The quantitative physiological, transcriptomic, proteomic, and computational analysis of this study revealed *A. woodii* metabolism to be highly flexible in terms of substrate co-utilization. The -omics analysis together with metabolic modelling provided insights into the adaptations of acetogen metabolism to utilization of different substrates. Utilization of formate, autotrophic and heterotrophic substrates was characterized by high energetic efficiencies, a crucial aspect for economic viability of bioprocesses for chemicals and fuels production from one carbon substrates. *In silico* analysis underlined the potential of substrate co-utilization for improving the bioenergetics which could facilitate the implementation of metabolic engineering strategies for formate-based production of ethanol and lactate. Collectively, the results of this study highlight *A. woodii* as a promising host for bioprocesses rooted in sustainable substrates.

Conflict of interests

The authors declare no competing interests.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Table 6
In silico predictions of the efficiency of formate-based bioproduction of ethanol and lactate with *A. woodii*. Reaction stoichiometries were obtained from FBA simulations ($q_{For} = 50$ mmol g^{-1} h^{-1}, $\mu = 0.02$ h^{-1}, experimentally determined NGAM values for the individual conditions used from Table 4). AdhE = bifunctional alcohol dehydrogenase, Aor = aldehyde oxidoreductase, Ldh = lactate dehydrogenase, Co-Ldh = electron-confurcating lactate dehydrogenase.

Product	Pathway	Co-substrate (s)	Reaction Stoichiometry (normalized)	Energetic efficiency [%]
Ethanol	AdhE	none	100 Formate = 14.2 Acetate + 6.6 Ethanol + 56.8 CO_2	87.1
	Aor		100 Formate = 16.1 Ethanol + 66.2 CO_2	89.4
	AdhE	Fructose	100 Formate + 1.5 Fructose = 19.2 Ethanol + 69.3 CO_2	90.0
	AdhE	H_2/CO_2	100 Formate + 66 CO_2 + 332.4 H_2 = 32.1 Acetate + 50.1 Ethanol	86.6
	AdhE	$H_2/CO_2/CO$	100 Formate + 66 CO_2 + 9.2 CO + 361.1 H_2 = 27.0 Acetate + 45.8 Ethanol	70.9
	AdhE	CO	100 Formate + 24.7 CO = 20.2 Ethanol + 82.6 CO_2	88.1
	AdhE	H_2/CO	100 Formate + 38.8 CO + 276.1 H_2 = 68.6 Ethanol	87.3
Lactate	Ldh	none	100 Formate = 19.4 Acetate + 3.1 Lactate + 50.1 CO_2	86.1
	Co-Ldh		100 Formate = 14.2 Acetate + 6.6 Lactate + 50.1 CO_2	87.1
	Ldh	Fructose	100 Formate + 8.6 Fructose = 33.2 Lactate + 50.1 CO_2	90.8
	Co-Ldh		100 Formate + 2.3 Fructose = 20.7 Lactate + 50.1 CO_2	90.1
	Ldh	H_2/CO_2	100 Formate + 66 CO_2 + 232.2 H_2 = 56.7 Acetate + 17 Lactate	85.4
	Co-Ldh		100 Formate + 66 CO_2 + 232.2 H_2 = 54.8 Lactate	87.8
	Ldh	$H_2/CO_2/CO$	100 Formate + 66 CO_2 + 9.2 CO + 221.4 H_2 = 45.9 Acetate + 27.2 Lactate	91.0
	Co-Ldh		100 Formate + 66 CO_2 + 9.2 CO + 221.4 H_2 = 57.8 Lactate	93.0
	Ldh	CO	100 Formate + 90.6 CO = 31.2 Lactate + 95.4 CO_2	86.7
	Co-Ldh		100 Formate + 24.7 CO = 20.2 Lactate + 62.4 CO_2	88.1
	Ldh	H_2/CO	100 Formate + 403.1 CO + 503.3 H_2 = 167.1 Lactate	86.0
	Co-Ldh		100 Formate + 31.6 CO + 131.8 H_2 = 43.3 Lactate	87.5

CRediT authorship contribution statement

Christian Simon Neuendorf: Investigation, Writing – original draft, Visualization. **Gabriel A. Vignolle:** Formal analysis. **Christian Derntl:** Investigation, Formal analysis. **Tamara Tomin:** Investigation, Formal analysis. **Katharina Novak:** Investigation. **Robert L. Mach:** Resources. **Ruth Birner-Grünberger:** Resources, Methodology. **Stefan Pflügl:** Conceptualization, Writing – original draft, Resources, Supervision, Project administration, Funding acquisition.

Acknowledgements

The authors are indebted to Samuele Verra for excellent technical assistance with the amino acid analysis. voestalpine Stahl GmbH, OMV Downstream GmbH and Austrian Airlines AG are gratefully acknowledged for financial support and the Austrian Research Promotion Agency (FFG) for funding. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymben.2021.09.004.

Funding

CN (#874503), KN (#858702), and SP (#858702, #874503) received funding from the Austrian Research Promotion Agency (FFG). TT and RBG received funding from the Austrian Science fund (FWF) (Doctoral school “DK Metabolic and Cardiovascular disease” (W1226) and SFB “Lipid hydrolysis” (F73)), Medical University of Graz and TU Wien.

References

Arora, N.K., Mishra, I., 2019. United Nations sustainable development goals 2030 and environmental sustainability: race against time. Environmental Sustainability 2, 339–342. https://doi.org/10.1007/s42398-019-00092-y.

Bainotti, A.E., Nishio, N., 2000. Growth kinetics of Acetobacterium sp. on methanol-formate in continuous culture. J. Appl. Microbiol. 88, 191–201. https://doi.org/10.1046/j.1365-2672.2000.00854.x.

Bainotti, A.E., Yamaguchi, K., Nakashimada, Y., Nishio, N., 1998. Kinetics and energetics of Acetobacterium sp. in chemostat culture on methanol-CO2. J. Ferment. Bioeng. 85, 223–229. https://doi.org/10.1016/S0922-338X(97)86772-4.

Balch, W.E., Schoberth, S., Tanner, R.S., Wolfe, R.S., 1977. Acetobacterium, a new Genus of hydrogen-oxidizing, carbon dioxide-reducing, anaerobic bacteria. Int. J. Syst. Bacteriol. 27, 355–361. https://doi.org/10.1099/00207713-27-4-355.

Bar-Even, A., Noor, E., Flamholz, A., Milo, R., 2013. Design and analysis of metabolic pathways supporting formatotrophic growth for electricity-dependent cultivation of microbes. Biochim. Biophys. Acta Bioenerg. 1827, 1039–1047. https://doi.org/10.1016/j.bbabio.2012.10.013.

Battat, E., Goldberg, I., Mateles, R.I., 1974. Growth of Pseudomonas C on C1 compounds: continuous culture. Appl. Microbiol. 28, 6.

Beck, M.H., Flaiz, M., Bengelsdorf, F.R., Dürre, P., 2019. Induced heterologous expression of the arginine deiminase pathway promotes growth advantages in the strict anaerobe Acetobacterium woodii. Appl. Microbiol. Biotechnol. https://doi.org/10.1007/s00253-019-10248-9.

Bertsch, J., Müller, V., 2015a. CO metabolism in the acetogen Acetobacterium woodii. Appl. Environ. Microbiol. 81, 5949–5956. https://doi.org/10.1128/AEM.01772-15.

Bertsch, J., Müller, V., 2015b. Bioenergetic constraints for conversion of syngas to biofuels in acetogenic bacteria. Biotechnol. Biofuels 8, 210. https://doi.org/10.1186/s13068-015-0393-x.

Blank, L.M., Narancic, T., Mampel, J., Tiso, T., O’Connor, K., 2020. Biotechnological upcycling of plastic waste and other non-conventional feedstocks in a circular economy. Curr. Opin. Biotechnol. 62, 212–219. https://doi.org/10.1016/j.copbio.2019.11.011.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Boone, D.R., Johnson, R.L., Liu, Y., 1989. Diffusion of the interspecies electron carriers H_2 and formate in methanogenic ecosystems and its implications in the measurement of K_m for H_2 or formate uptake. Appl. Environ. Microbiol. 55, 1735–1741. https://doi.org/10.1128/aem.55.7.1735-1741.1989.

Braun, K., Gottschalk, G., 1981. Effect of molecular hydrogen and carbon dioxide on chemo-organotrophic growth of Acetobacterium woodii and Clostridium aceticum. Arch. Microbiol. 128, 294–298. https://doi.org/10.1007/BF00422533.

Buschhorn, H., Dürre, P., Gottschalk, G., 1989. Production and utilization of ethanol by the homoacetogen Acetobacterium woodii. Appl. Environ. Microbiol. 55, 1835–1840. https://doi.org/10.1128/AEM.55.7.1835-1840.1989.

Chatterjee, S., Huang, K.-W., 2020. Unrealistic energy and materials requirement for direct air capture in deep mitigation pathways. Nat. Commun. 11, 3287. https://doi.org/10.1038/s41467-020-17203-7.

Claassens, N.J., Bordanaba-Florit, G., Cotton, C.A.R., De Maria, A., Finger-Bou, M., Friedeheim, L., Giner-Laguarda, N., Munar-Palmer, M., Newell, W., Scarinci, G., Verbunt, J., de Vries, S.T., Yilmaz, S., Bar-Even, A., 2020. Replacing the Calvin cycle with the reductive glycine pathway in Cupriavidus necator. Metab. Eng. 62, 30–41. https://doi.org/10.1016/j.ymben.2020.08.004.

Claassens, N.J., Cotton, C.A.R., Kopljar, D., Bar-Even, A., 2019. Making quantitative sense of electromicrobial production. Nat Catal 2, 437–447. https://doi.org/10.1038/s41929-019-0272-0.

Claassens, N.J., Sánchez-Andrea, I., Sousa, D.Z., Bar-Even, A., 2018. Towards sustainable feedstocks: a guide to electron donors for microbial carbon fixation. Curr. Opin. Biotechnol. 50, 195–205. https://doi.org/10.1016/j.copbio.2018.01.019.

Cotton, C.A., Claassens, N.J., Benito-Vaquerizo, S., Bar-Even, A., 2020. Renewable methanol and formate as microbial feedstocks. Curr. Opin. Biotechnol. 62, 168–180. https://doi.org/10.1016/j.copbio.2019.10.002.

Cox, J., Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372. https://doi.org/10.1038/nbt.1511.

Demler, M., Weuster-Botz, D., 2011. Reaction engineering analysis of hydrogenotrophic production of acetic acid by Acetobacterium woodii. Biotechnol. Bioeng. 108, 470–474. https://doi.org/10.1002/bit.22935.

Diender, M., Stams, A.J.M., Sousa, D.Z., 2015. Pathways and bioenergetics of anaerobic carbon monoxide fermentation. Front. Microbiol. 6 https://doi.org/10.3389/fmicb.2015.01275.

Erian, A.M., Gibisch, M., Pflügl, S., 2018. Engineered *E. coli* W enables efficient 2,3-butanediol production from glucose and sugar beet molasses using defined minimal medium as economic basis. Microb. Cell Factories 17, 190. https://doi.org/10.1186/s12934-018-1038-0.

Fasihi, M., Efimova, O., Breyer, C., 2019. Techno-economic assessment of CO2 direct air capture plants. J. Clean. Prod. 224, 957–980. https://doi.org/10.1016/j.jclepro.2019.03.086.

Furdui, C., Ragsdale, S.W., 2000. The role of pyruvate ferredoxin oxidoreductase in pyruvate synthesis during autotrophic growth by the wood-ljungdahl pathway. J. Biol. Chem. 275, 28494–28499. https://doi.org/10.1074/jbc.M003291200.

Godley, Andrew R., Linnett, Paul E., Robinson, John P., 1990. The effect of carbon dioxide on the growth kinetics of fructose-limited chemostat cultures of Acetobacterium woodii DSM 1030. Arch. Microbiol. 154 https://doi.org/10.1007/BF00249170.

Goldberg, I., Rock, J.S., Ben-Bassat, A., Mateles, R.I., 1976. Bacterial yields on methanol, methylamine, formaldehyde, and formate. Biotechnol. Bioeng. 18, 1657–1668. https://doi.org/10.1002/bit.260181202.

Gonzalez de la Cruz, J., Machens, F., Messerschmidt, K., Bar-Even, A., 2019. Core catalysis of the reductive Glycine pathway demonstrated in yeast. ACS Synth. Biol. 8, 911–917. https://doi.org/10.1021/acssynbio.8b00464.

Haas, T., Krause, R., Weber, R., Demler, M., Schmid, G., 2018. Technical photosynthesis involving CO2 electrolysis and fermentation. Nat Catal 1, 32–39. https://doi.org/10.1038/s41929-017-0005-1.

Hardt, S., Stapf, S., Filmon, D.T., Birrell, J.A., Rüdiger, O., Fourmond, V., Léger, C., Plumeré, N., 2021. Reversible H2 oxidation and evolution by hydrogenase embedded in a redox polymer film. Nat Catal 4, 251–258. https://doi.org/10.1038/s41929-021-00586-1.

Hazeu, W., Donker, R.A., 1983. A continuous culture study of methanol and formate utilization by the yeast Pichia pastoris. Biotechnol. Lett. 5, 399–404. https://doi.org/10.1007/BF00131280.

Hermann, M., Teleki, A., Weitz, S., Niess, A., Freund, A., Bengelsdorf, F.R., Takors, R., 2020. Electron availability in CO_2 , CO and H_2 mixtures constrains flux distribution, energy management and product formation in *Clostridium ljungdahlii.* Microb. Biotechnol. 13, 1831–1846. https://doi.org/10.1111/1751-7915.13625.

Hofer, A., Kamravamanesh, D., Bona-Lovasz, J., Limbeck, A., Lendl, B., Herwig, C., Fricke, J., 2018. Prediction of filamentous process performance attributes by CSL quality assessment using mid-infrared spectroscopy and chemometrics. J. Biotechnol. 265, 93–100. https://doi.org/10.1016/j.jbiotec.2017.11.010.

Hoffmeister, S., Gerdom, M., Bengelsdorf, F.R., Linder, S., Flüchter, S., Öztürk, H., Blümke, W., May, A., Fischer, R.-J., Bahl, H., Dürre, P., 2016. Acetone production with metabolically engineered strains of Acetobacterium woodii. Metab. Eng. 36, 37–47. https://doi.org/10.1016/j.ymben.2016.03.001.

Huang, H., Wang, S., Moll, J., Thauer, R.K., 2012. Electron bifurcation involved in the energy metabolism of the acetogenic bacterium Moorella thermoacetica growing on glucose or H2 plus CO2. J. Bacteriol. 194, 3689–3699. https://doi.org/10.1128/JB.00385-12.

Jones, S.W., Fast, A.G., Carlson, E.D., Wiedel, C.A., Au, J., Antoniewicz, M.R., Papoutsakis, E.T., Tracy, B.P., 2016. CO2 fixation by anaerobic non-photosynthetic mixotrophy for improved carbon conversion. Nat. Commun. 7, 12800. https://doi.org/10.1038/ncomms12800.

Kantzow, C., Mayer, A., Weuster-Botz, D., 2015. Continuous gas fermentation by Acetobacterium woodii in a submerged membrane reactor with full cell retention. J. Biotechnol. 212, 11–18. https://doi.org/10.1016/j.jbiotec.2015.07.020.

Karmann, S., Follonier, S., Egger, D., Hebel, D., Panke, S., Zinn, M., 2017. Tailor-made PAT platform for safe syngas fermentations in batch, fed-batch and chemostat mode with Rhodospirillum rubrum. Microb. Biotechnol. 10, 1365–1375. https://doi.org/10.1111/1751-7915.12727.

Kelly, D.P., Baker, S.C., Trickett, J., Davey, M., Murrell, J.C., 1994. Methanesulphonate utilization by a novel methylotrophic bacterium involves an unusual monooxygenase. Microbiology 140, 1419–1426. https://doi.org/10.1099/00221287-140-6-1419.

Kim, S., Lindner, S.N., Aslan, S., Yishai, O., Wenk, S., Schann, K., Bar-Even, A., 2020. Growth of E. coli on formate and methanol via the reductive glycine pathway. Nat. Chem. Biol. 16, 538–545. https://doi.org/10.1038/s41589-020-0473-5.

Klamt, S., Saez-Rodriguez, J., Gilles, E.D., 2007. Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst. Biol. 1, 2. https://doi.org/10.1186/1752-0509-1-2.

Koch, S., Kohrs, F., Lahmann, P., Bissinger, T., Wendschuh, S., Benndorf, D., Reichl, U., Klamt, S., 2019. RedCom: a strategy for reduced metabolic modeling of complex microbial communities and its application for analyzing experimental datasets from anaerobic digestion. PLoS Comput. Biol. 15, e1006759 https://doi.org/10.1371/journal.pcbi.1006759.

Köpke, M., Simpson, S.D., 2020. Pollution to products: recycling of 'above ground' carbon by gas fermentation. Curr. Opin. Biotechnol. 65, 180–189. https://doi.org/10.1016/j.copbio.2020.02.017.

Kremp, F., Poehlein, A., Daniel, R., Müller, V., 2018. Methanol metabolism in the acetogenic bacterium Acetobacterium woodii. Environ. Microbiol. 20, 4369–4384. https://doi.org/10.1111/1462-2920.14356.

Kremp, F., Roth, J., Müller, V., 2020. The *Sporomusa* type Nfn is a novel type of electron-bifurcating transhydrogenase that links the redox pools in acetogenic bacteria. Sci. Rep. 10, 14872. https://doi.org/10.1038/s41598-020-71038-2.

Li, H., Opgenorth, P.H., Wernick, D.G., Rogers, S., Wu, T.-Y., Higashide, W., Malati, P., Huo, Y.-X., Cho, K.M., Liao, J.C., 2012. Integrated electromicrobial conversion of CO2 to higher alcohols. Science 335, 1596. https://doi.org/10.1126/science.1217643, 1596.

Liew, F., Martin, M.E., Tappel, R.C., Heijstra, B.D., Mihalcea, C., Köpke, M., 2016. Gas fermentation—a flexible platform for commercial scale production of low-carbon-fuels and chemicals from waste and renewable feedstocks. Front. Microbiol. 7 https://doi.org/10.3389/fmicb.2016.00694.

Litty, D., Müller, V., 2021. Butyrate production in the acetogen *Eubacterium limosum* is dependent on the carbon and energy source. Microb. Biotechnol. 1751–7915, 13779. https://doi.org/10.1111/1751-7915.13779.

Loubiere, P., Gros, E., Paquet, V., Lindley, N.D., 1992. Kinetics and physiological implications of the growth behaviour of Eubacterium limosum on glucose/methanol mixtures. J. Gen. Microbiol. 138, 979–985. https://doi.org/10.1099/00221287-138-5-979.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8.

Maru, B.T., Munasinghe, P.C., Gilary, H., Jones, S.W., Tracy, B.P., 2018. Fixation of CO2 and CO on a diverse range of carbohydrates using anaerobic, non-photosynthetic mixotrophy. FEMS (Fed. Eur. Microbiol. Soc.) Microbiol. Lett. 365 https://doi.org/10.1093/femsle/fny039.

Molitor, B., Marcellin, E., Angenent, L.T., 2017. Overcoming the energetic limitations of syngas fermentation. Curr. Opin. Chem. Biol. 41, 84–92. https://doi.org/10.1016/j.cbpa.2017.10.003.

Moon, J., Dönig, J., Kramer, S., Poehlein, A., Daniel, R., Müller, V., 2021. Formate metabolism in the acetogenic bacterium *Acetobacterium woodii.* Environ. Microbiol. 1462–2920, 15598. https://doi.org/10.1111/1462-2920.15598.

Müller, V., 2019. New horizons in acetogenic conversion of one-carbon substrates and biological hydrogen storage. Trends Biotechnol. https://doi.org/10.1016/j.tibtech.2019.05.008.

Nguyen, N.-P.-T., Raynaud, C., Meynial-Salles, I., Soucaille, P., 2018. Reviving the Weizmann process for commercial n-butanol production. Nat. Commun. 9. https://doi.org/10.1038/s41467-018-05661-z.

Nissen, T.L., Schulze, U., Nielsen, J., Villadsen, J., 1997. Flux distributions in anaerobic, glucose-limited continuous cultures of Saccharomyces cerevisiae. Microbiology 143, 203–218. https://doi.org/10.1099/00221287-143-1-203.

Novak, K., Neuendorf, C.S., Kofler, I., Kieberger, N., Klamt, S., Pflügl, S., 2021. Blending industrial blast furnace gas with H2 enables Acetobacterium woodii to efficiently co-utilize CO, CO2 and H2. Bioresour. Technol. 323, 124573. https://doi.org/10.1016/j.biortech.2020.124573.

Panich, J., Fong, B., Singer, S.W., 2021. Metabolic engineering of Cupriavidus necator H16 for sustainable biofuels from CO2. Trends Biotechnol. 39, 412–424. https://doi.org/10.1016/j.tibtech.2021.01.001.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419. https://doi.org/10.1038/nmeth.4197.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yılmaz, Ş., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jarnuczak, A.F., Ternent, T., Brazma, A., Vizcaíno, J.A., 2019. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 47, D442–D450. https://doi.org/10.1093/nar/gky1106.

Pertea, G., Pertea, M., 2020. GFF utilities: GffRead and GffCompare. F1000Res 9, 304. https://doi.org/10.12688/f1000research.23297.2.

Pluschkell, S.B., Flickinger, M.C., 2002. Dissimilation of [13C]methanol by continuous cultures of Bacillus methanolicus MGA3 at 50 °C studied by 13C NMR and isotope-ratio mass spectrometry. Microbiology 148, 3223–3233. https://doi.org/10.1099/00221287-148-10-3223.

Poehlein, A., Schmidt, S., Kaster, A.-K., Goenrich, M., Vollmers, J., Thürmer, A., Bertsch, J., Schuchmann, K., Voigt, B., Hecker, M., Daniel, R., Thauer, R.K., Gottschalk, G., Müller, V., 2012. An ancient pathway combining carbon dioxide fixation with the generation and utilization of a sodium Ion gradient for ATP synthesis. PloS One 7, e33439. https://doi.org/10.1371/journal.pone.0033439.

R Core Team, 2013. R: A Language and Environment for Statistical Computing.

Realmonte, G., Drouet, L., Gambhir, A., Glynn, J., Hawkes, A., Köberle, A.C., Tavoni, M., 2019. An inter-model assessment of the role of direct air capture in deep mitigation pathways. Nat. Commun. 10, 3277. https://doi.org/10.1038/s41467-019-10842-5.

Richter, H., Molitor, B., Wei, H., Chen, W., Aristilde, L., Angenent, L.T., 2016. Ethanol Production in Syngas-Fermenting Clostridium Ljungdahlii Is Controlled by Thermodynamics rather than by Enzyme Expression 9, pp. 2392–2399. https://doi.org/10.1039/c6ee01108j.

Rintala, E., Wiebe, M.G., Tamminen, A., Ruohonen, L., Penttilä, M., 2008. Transcription of hexose transporters of Saccharomyces cerevisiae is affected by change in oxygen provision. BMC Microbiol. 8, 53. https://doi.org/10.1186/1471-2180-8-53.

Rittmann, S., Seifert, A., Herwig, C., 2012. Quantitative analysis of media dilution rate effects on Methanothermobacter marburgensis grown in continuous culture on H2 and CO2. Biomass Bioenergy 36, 293–301. https://doi.org/10.1016/j.biombioe.2011.10.038.

Schendel, F.J., Bremmon, C.E., Flickinger, M.C., Hanson, R.S., 1990. L-lysine production at 50°C by mutants of a newly isolated and characterized methylotrophic Bacillus sp. Appl. Environ. Microbiol. 56, 8.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Schoelmerich, M.C., Katsyv, A., Sung, W., Mijic, V., Wiechmann, A., Kottenhahn, P., Baker, J., Minton, N.P., Müller, V., 2018. Regulation of lactate metabolism in the acetogenic bacterium *Acetobacterium woodii*. Environ. Microbiol. 20, 4587–4595. https://doi.org/10.1111/1462-2920.14412.

Schuchmann, K., Müller, V., 2013. Direct and reversible hydrogenation of CO2 to formate by a bacterial carbon dioxide reductase. Science 342, 1382–1385. https://doi.org/10.1126/science.1244758.

Schuchmann, K., Müller, V., 2014. Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. Nat. Rev. Microbiol. 12, 809–821. https://doi.org/10.1038/nrmicro3365.

Schwarz, F.M., Ciurus, S., Jain, S., Baum, C., Wiechmann, A., Basen, M., Müller, V., 2020. Revealing formate production from carbon monoxide in wild type and mutants of Rnf- and Ech-containing acetogens, *Acetobacterium woodii* and *Thermoanaerobacter kivui*. Microb. Biotechnol. 13, 2044–2056. https://doi.org/10.1111/1751-7915.13663.

Schwarz, F.M., Müller, V., 2020. Whole-cell biocatalysis for hydrogen storage and syngas conversion to formate using a thermophilic acetogen. Biotechnol. Biofuels 13, 32. https://doi.org/10.1186/s13068-020-1670-x.

Shay, L.K., Hunt, H.R., Wegner, G.H., 1987. High-productivity fermentation process for cultivating industrial microorganisms. J. Ind. Microbiol. 2, 79–85. https://doi.org/10.1007/BF01569506.

Soneson, C., Love, M.I., Robinson, M.D., 2016. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res 4, 1521. https://doi.org/10.12688/f1000research.7563.2.

Song, Y., Lee, J.S., Shin, J., Lee, G.M., Jin, S., Kang, S., Lee, J.-K., Kim, D.R., Lee, E.Y., Kim, S.C., Cho, S., Kim, D., Cho, B.-K., 2020. Functional cooperation of the glycine synthase-reductase and Wood–Ljungdahl pathways for autotrophic growth of Clostridium drakei. Proc. Natl. Acad. Sci. Unit. States Am. https://doi.org/10.1073/pnas.1912289117, 201912289.

Trifunović, D., Berghaus, N., Müller, V., 2020. Growth of the acetogenic bacterium Acetobacterium woodii by dismutation of acetaldehyde to acetate and ethanol. Environmental Microbiology Reports 12, 58–62. https://doi.org/10.1111/1758-2229.12811.

Tschech, A., Pfennig, N., 1984. Growth yield increase linked to caffeate reduction in Acetobacterium woodii. Arch. Microbiol. 137, 163–167. https://doi.org/10.1007/BF00414460.

Tyanova, S., Temu, T., Cox, J., 2016a. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat. Protoc. 11, 2301–2319. https://doi.org/10.1038/nprot.2016.136.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., Cox, J., 2016b. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods 13, 731–740. https://doi.org/10.1038/nmeth.3901.

Valgepea, K., de Souza Pinto Lemgruber, R., Abdalla, T., Binos, S., Takemori, N., Takemori, A., Tanaka, Y., Tappel, R., Köpke, M., Simpson, S.D., Nielsen, L.K., Marcellin, E., 2018. H2 drives metabolic rearrangements in gas-fermenting Clostridium autoethanogenum. Biotechnol. Biofuels 11, 55. https://doi.org/10.1186/s13068-018-1052-9.

Valgepea, K., de Souza Pinto Lemgruber, R., Meaghan, K., Palfreyman, R.W., Abdalla, T., Heijstra, B.D., Behrendorff, J.B., Tappel, R., Köpke, M., Simpson, S.D., Nielsen, L.K., Marcellin, E., 2017. Maintenance of ATP homeostasis triggers metabolic shifts in gas-fermenting acetogens. Cell Systems 4, 505–515. https://doi.org/10.1016/j.cels.2017.04.008.

Van Hecke, W., Bockrath, R., De Wever, H., 2019. Effects of moderately elevated pressure on gas fermentation processes. Bioresour. Technol. 293, 122129. https://doi.org/10.1016/j.biortech.2019.122129.

Vees, C.A., Neuendorf, C.S., Pflügl, S., 2020. Towards continuous industrial bioprocessing with solventogenic and acetogenic clostridia: challenges, progress and perspectives. J. Ind. Microbiol. Biotechnol. 47, 753–787. https://doi.org/10.1007/s10295-020-02296-2.

von Kamp, A., Thiele, S., Hädicke, O., Klamt, S., 2017. Use of CellNetAnalyzer in biotechnology and metabolic engineering. J. Biotechnol. 261, 221–228. https://doi.org/10.1016/j.jbiotec.2017.05.001.

Wendisch, V.F., Brito, L.F., Gil Lopez, M., Hennig, G., Pfeifenschneider, J., Sgobba, E., Veldmann, K.H., 2016. The flexible feedstock concept in Industrial Biotechnology: metabolic engineering of Escherichia coli, Corynebacterium glutamicum, Pseudomonas, Bacillus and yeast strains for access to alternative carbon sources. J. Biotechnol. 234, 139–157. https://doi.org/10.1016/j.jbiotec.2016.07.022.

Wiechmann, A., Ciurus, S., Oswald, F., Seiler, V.N., Müller, V., 2020. It does not always take two to tango: "Syntrophy" via hydrogen cycling in one bacterial cell. ISME J. https://doi.org/10.1038/s41396-020-0627-1.

Wolin, M.J., Miller, T.L., Collins, M.D., Lawson, P.A., 2003. Formate-Dependent Growth and Homoacetogenic Fermentation by a Bacterium from Human Feces: Description of Bryantella formatexigens gen. nov., sp. nov. AEM 69, 6321–6326. https://doi.org/10.1128/AEM.69.10.6321-6326.2003.

Yishai, O., Lindner, S.N., Gonzalez de la Cruz, J., Tenenboim, H., Bar-Even, A., 2016. The formate bio-economy. Curr. Opin. Chem. Biol. 35, 1–9. https://doi.org/10.1016/j.cbpa.2016.07.005.

Zelcbuch, L., Lindner, S.N., Zegman, Y., Vainberg Slutskin, I., Antonovsky, N., Gleizer, S., Milo, R., Bar-Even, A., 2016. Pyruvate formate-lyase enables efficient growth of *Escherichia coli* on acetate and formate. Biochemistry 55, 2423–2426. https://doi.org/10.1021/acs.biochem.6b00184.

Zhu, A., Ibrahim, J.G., Love, M.I., 2019. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics 35, 2084–2092. https://doi.org/10.1093/bioinformatics/bty895.

Acknowledgements

Foremost, I would like to thank my supervisor Univ. Prof. Robert Mach for his reliable support, scientific advice, and granting me the opportunity to do my doctoral thesis under his supervision. Furthermore, I would like to pay my special regards to my co-supervisor Dr. Christian Derntl. He was always available whenever I had a question and his clear supervision inspired me the entire time. Further, I would like to extend my special thanks to Prof. Astrid Mach-Aigner for her scientific advice and the possibility to do my doctoral thesis in their working group.

I would like to extend my gratitude toward the PhD program TU Wien bioactive for giving me the chance to grow beyond my expectations and work in such a highly competitive program. I want to thank all my colleagues from the bioactive doctoral college and the associated colleagues for interesting discussions and insights in a diverse range of different scientific topics.

Last, but not least, my thanks go to family and friends, especially my mother Cornelia Vignolle, my father Giorgio Vignolle, my grandparents Emilia and Bruno Vignolle, and my grandparents Ursula and Prof. Gerhart Schüring, the former being an inspiration to pursue a doctoral degree. Primarily I thank my future wife Mag.pharm. Anna Stich for the scientific discussions, the laughs, her unyielding support and the strength she gives me every single day.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Curriculum Vitae

Personal Data

Name	Mag.pharm. Gabriel Alexander Vignolle
Contact	Geigergasse 11/6, 1050 Wien
Phone (TU)	+43 1 58801-166566
Mobile	+43 6763708879
E-mail	gabriel.vignolle@tuwien.ac.at
ORCID	0000-0002-1369-5150
Date and place of birth	29.08.1988 in Luxemburg, Luxembourg
Nationality	Luxemburg

Research relevant employments

February 2019 – to date	**University assistant (PhD thesis)** TU Wien Institute of Chemical, Environmental and Biological Engineering Vienna, Austria **Lecturer (166.231 Applied Bioinformatics)** TU Wien Institute of Chemical, Environmental and Biological Engineering Vienna, Austria
July 2018 – December 2018	**Project Co-worker** University of Vienna Department of Botany and Biodiversity Research Vienna, Austria
2017 – 2018	**Project Co-worker** University of Vienna Department of Botany and Biodiversity Research Vienna, Austria

Education

February 2019 – to date	**PhD thesis** TU Wien Institute of Chemical, Environmental and Biological Engineering Vienna, Austria Supervisors: Univ.Prof. Mag. Dr.rer.nat. Robert Mach Mag.rer.nat. Dr.rer.nat. Christian Derntl Topic: “Modeling novel bioinformatics approaches to investigate bioactive substance production based on genomics and transcriptomics”
September 2017 – February 2019	**Diploma thesis** University of Vienna Department of Pharmaceutical Sciences Division of Pharmacognosy Vienna, Austria Supervisors: Univ.Prof. Dr. Sergey B. Zotchev Dr. Andrea Kodym Topic: “Effect of Cryopreservation on the Microbiome of Plants”
March 2012 – February 2019	**Studies in Pharmacy (Diploma)** University of Vienna Department of Pharmaceutical Sciences Vienna, Austria Graduation 2019 (Mag.pharm)
1994 - 2006	**School education and graduation** European school of Luxemburg, Luxembourg

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.

Skills and trainings

Advanced trainings:

- Ecological and evolutionary genomics seminar (May 2019)
- Biostatistics (June 2019)
- Bioinformatics of nucleic acids (July 2019)
- Device officer of Illumina MiSeq (September 2019 – to date)
- Management and leadership classes (January 2020 - January 2021)
- Statistics (June 2020)
- Machine Learning class (September 2020)
- Big Data on the Vienna scientific cluster course (VSC) (January 2021)
- Technology Assessment and Sustainable Development (July 2021)
- Several seminars and classes connected to the bioactive doctoral college (2019 – 2021)

Languages:

- German: first language
- Italian: first language
- English: business fluent
- French: basic knowledge
- Luxemburgish: basic knowledge

Scientific contributions

Publications:

Vignolle GA, Mach RL, Mach-Aigner AR, Derntl C. Novel approach in whole genome mining and transcriptome analysis reveal conserved RiPPs in *Trichoderma* spp. *BMC Genomics* 21:258 (2020)

Vignolle GA, Mach RL, Mach-Aigner AR, Derntl C. Genome Sequence of the Black Yeast-Like Strain *Aureobasidium pullulans* var. *aubasidani* CBS 100524. *Microbiology Resource Announcement*. (2021)

Vignolle GA, Schaffer D, Zehetner L, Mach RL, Mach-Aigner AR, Derntl C. FunOrder: A robust and semi-automated method for the identification of essential biosynthetic genes through computational molecular co-evolution. *PLOS Computational Biology* 17(9): e1009372. (2021)

Vignolle GA, Hochenegger NJ, U'Ren JM, Mach RL, Mach-Aigner AR, Rahimi MJ, Salim KA, Chan CM, Lim LBL, Cai F, Druzhinina1 IS, Derntl C. "Genome sequencing of *Wardomyces moseri*: a rare but cosmopolitan fungus with an outstanding secondary metabolite production potential." Submitted to *BMC Genomics*

Vignolle GA, Mach RL, Mach-Aigner AR, Derntl C. "FunOrder 2.0 – a fully automated method for the identification of co-evolved genes." Submitted to *PLOS Computational Biology*

Neuendorf CS, Vignolle GA, Derntl C, Tomin T, Novak K, Mach RL, Birner-Grünberger R, Pflügl S. A quantitative metabolic analysis reveals *Acetobacterium woodii* as a flexible and robust host for formate-based bioproduction, *Metabolic Engineering*, ISSN 1096-7176 (2021)

Ellena V, Seekles SJ, Vignolle GA, Ram AFJ, Steiger MG. Genome sequencing of the neotype strain CBS 554.65 reveals the MAT1–2 locus of *Aspergillus niger*. *BMC Genomics* 22, 679 (2021)

Oberhofer M, Malfent F, Zehl M, Urban E, Wackerlig J, Reznicek G, Vignolle GA, Rückert C, Busche T, Wibberg D, Zotchev SB. "Biosynthetic potential of the endophytic fungus *Chalara* sp. BL73 revealed via compound identification and genome mining" Submitted to *Applied and Environmental Microbiology*

Talks:

GÖCH, GDCh; 18th Austrian Chemistry Days; Linz, Austria; September 24-27, 2019
"Novel approaches in genome mining reveal preserved RiPPs in *Trichoderma reesei*"
Vignolle GA, Seidl BK, Mach RL, Schumacher R, Mach-Aigner AR, Derntl C.

Funding ID/grants:

(Co-author, peer reviewed)
FWF-Project Nr. P 34036 "Identification and Characterization of Novel Fungal RiPPs"

Posters:

54th Popgroup; online, University of Liverpool; January 4-6, 2021
"Population genomics shows last European stand of *Artemisia laciniata* is diverse despite population size" Hedderich C, Vignolle GA, Hatfaludi T, Tkach N, Hoffmann MH, Korobkov AA, Kodym A, Paun O.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

ECFG15; Rome, Italy; February 18th, 2020
"Novel method in genome mining and transcriptome analysis reveal undiscovered RiPPs in *Trichoderma* spp." Vignolle GA, Mach RL, Mach-Aigner AR, Derntl C.

ECFG15; Rome, Italy; February 18th, 2020
"Bioprospecting a newly identified fungus from the Borneo rain forest regarding its bioactive properties" Hochenegger NJ, Vignolle GA, Mach RL, Druzhinina IS, Mach-Aigner AR, Derntl C.

ConsGen18; Vienna, Austria; February 27th, 2018
"Population genomics shows last European stand of *Artemisia laciniata* (Asteraceae) as fairly diverse despite extreme population size" Hatfaludi T, Kodym A, Vignolle GA, Tkach N, Hoffmann MH, Korobkov AA, Paun O.

Die approbierte gedruckte Originalversion dieser Dissertation ist an der TU Wien Bibliothek verfügbar.
The approved original version of this doctoral thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub