Diplomarbeit

Evaluation and testing of fungal cluster border prediction based on computational molecular evolution (FunOrder) program

Ausgeführt am Institut für

Verfahrenstechnik, Umwelttechnik und Technische Biowissenschaften

der Technischen Universität Wien

unter der Anleitung von Univ.Prof. Mag. Dr.rer.nat. Robert Mach, sowie Univ.Ass. Mag.pharm. Gabriel Alexander Vignolle und Univ.Ass. Mag.rer.nat. Dr.rer.nat Christian Derntl als verantwortlich mitwirkenden Universitätsassistenten

durch

Denise Schaffer, BSc

Wien, 23.12.2020 Denise Schaffer, BSc

Author

Denise Schaffer, BSc

Institution: Technical University of Vienna

Study Programme: Master programme Technical Chemistry

Track: Biotechnology and Bioanalytics

Reviewer and Supervisor

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

Institution: Technical University of Vienna

Department: Technical Chemistry

Institute: Chemical, Environmental and Bioscience Engineering

Co-Supervisors

Univ.Ass. Mag.pharm. Gabriel Alexander Vignolle

Institution: Technical University of Vienna

Department: Technical Chemistry

Institute: Chemical, Environmental and Bioscience Engineering

Univ.Ass. Mag.rer.nat Dr.rer.nat. Christian Derntl, BSc

Institution: Technical University of Vienna

Department: Technical Chemistry

Institute: Chemical, Environmental and Bioscience Engineering

Eidesstaatliche Erklärung

Hiermit versichere ich eidesstaatlich, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst habe.

Wien, 23.12.2020 Denise Schaffer, BSc

Acknowledgements

Foremost, I would like to thank my supervisor, Univ. Prof. Robert Mach, head of the Institute of Chemical, Environmental and Bioscience Engineering at the Technical University of Vienna. His door was always open whenever I had a question, also during the Covid19 pandemic. I especially want to thank him for his honest and great support prior this project, which motivated me after throwbacks, and encouraged me to new possibilities.

Besides my supervisor, I would like to pay my special regards to my co-supervisor and the programmer of FunOrder, Mag. pharm. Gabriel Alexander Vignolle. Despite the Covid19 pandemic, he was always available whenever I had a question about the program, evaluation, or writing. His clear and favorably supervision motivated me during the entire time. I am also extremely grateful for his feedback after proofreading this thesis. His great advice pushed me further and allowed me to go for the limit. I whole-heartedly want to thank him for that.

I wish to show my sincere gratitude to my colleagues from my collaborating group for Synthetic Biology and Molecular Biotechnology within the Institute of Chemical, Environmental and Bioscience Engineering for the warm and welcoming atmosphere. In this connection, I would like to particularly single out and thank my second co-supervisor, Dr.rer.nat. Christian Derntl, and the head of the research group, Assistant Prof. Astrid Mach-Aigner.

My deepest gratitude also goes to my colleagues during my studies at technical University of Vienna, especially my colleagues from Fachschaft Technische Chemie, who helped me throughout my studies and made me enjoy them even more.

In addition, I would like to thank my parents and my siblings for never stopping believing in me and for always making me feel confident in my abilities. I would like to particularly single out and thank my father, Dieter Schaffer, and my sister, Mag. Natalia Corrales-Diez, for their support during the last months.

Finally, I could not have completed this thesis without the constant emotional support and motivation of my beloved partner, Kevin Takacs. I would like to thank him with all my heart for his amazing support.

Last, but not least, I would like to thank all fungal species out there, without which I would not have been able to write this thesis and which challenge us researcher's day after day to go for the limit.

Table of contents

1. List of abbreviation

A	Adenylation
ACP	Acyl carrier protein
aMEM	average manual evaluation measure
AMP	Adenosine-5'-monophosphate
AT	Acetyl transferase
BGC	Biosynthetic gene cluster
BLAST	Basic Local Alignment Search Tools
BLOSUM	Blocks substitution matrix
C	Condensation
CMet	C-methyl transferase
CsA	Cyclosporine A
DH	Dehydratase
DKC	Dieckmann cyclase
DMAPP	Dimethylallyl diphosphate
Dmb	2-Pyridon-Desmethylbassianin
E	Epimerisation
EMBOSS	European Molecular Biology Open Software Suite
ER	Enoyl reductase
FN	False negative
FP	False positive
GMP	Guanosine 5'-monophosphate
IMP	Inosin 5'-monophosphate
IPP	Isopentenyl diphosphate
kb	Kilobasepair
KR	Ketoreductase, or Ketoacyl reductase
KS	Ketosynthase
Mb	Megabasepair
MCC	Matthews Correlation Coefficient
MEM	Manual evaluation measure
MIBIG	Minimum Information about a Biosynthetic Gene cluster

MIxS	Minimum Information about any Sequence
ML	Maximum likelihood
MP	Maximum parsimony
MPA	Mycophenolic acid
MSA	Multiple Sequence Alignment
MT	Methyltransferase
NCBI	National Center for Biotechnology Information
NJ	Neighbour joining
NRPS	Non-ribosomal Peptide Synthetase
OOB	Out-of-bag
ORF	Open reading frame
P450	Cytochrome P450 oxidase
PAM	Percent accepted mutation
PCA	Principal Component Analysis
PCP	Peptidyl Carrier Protein
PKS	Polyketide synthase
PLS DA	Partial Least Discriminant Analysis
RAxML	Randomized accelerated maximum likelihood
RF	Robinson-Foulds
RiPPs	Ribosomally synthesized and post-translationally modified peptides
ROC	Receiver operating characteristics
SM	Secondary metabolite
T	Thiolation
TC	Terpene cyclase
TE	Thioesterase
Ten	Tenellin
TN	True negative
TP	True positive
UPGMA	Unweighted Pair Group Method with Arithmetic mean
XMP	Xanthosine-5'-monophosphate

2. Abstract

English Version

Fungal cluster border prediction based on computational molecular co-evolution (=FunOrder) is a genome mining program which finds evolutionary connections between fungal genes and hence, find genes involved in the biosynthesis of a compound. During this project, the program was evaluated by manually examination of phylogenetic trees based on the genes within experimentally validated fungal biosynthetic gene clusters (BGC) as positive controls and genes from randomly generated clusters as negative controls. The evaluation data was then used to define the borders of co-evolution between protein families within BGCs in fungi. The aim of the project was to verify if FunOrder has the ability to predict correct fungal cluster borders and therefore can contribute to the research for novel secondary metabolites.

German Version

FunOrder (Fungal cluster border prediction based on computational molecular co-evolution) ist ein Genome Mining Programm, welches die evolutionäre Verbindung zwischen Genen und Gencluster aus Pilzen untersucht. Während des Projekts sollte das Programm durch die manuelle Auswertung phylogenetischer Bäume von Genen aus experimentell ermittelten pilzlichen biosynthetischen Gencluster (BGC) als Positiv- und aus zufällig generierten Cluster als Negativkontrollen evaluiert werden. Die daraus erzielten Daten wurden anschließend verwendet, um die koevolutionären Grenzen von Proteinfamilien in Gencluster von Pilzen zu definieren. Ziel des Projektes war die Untersuchung und Verifikation, ob FunOrder korrekte Clustergrenzen in Pilzen vorhersagen kann und damit als neues Tool zur Entdeckung von neuen Sekundärmetaboliten geeignet ist.

3. Introduction

3.1 Fungal secondary metabolites (SM)

Secondary metabolites are low-molecular-weight, often bioactive compounds produced by a cell.[1] The probably most famous secondary metabolite is the broad-spectrum antibiotic penicillin which was discovered in 1929.[2] Since then, an effective control of infections was possible for the first time in history and the research for new antibiotics began.[3,4] While the discovery of Penicillin happened by accident, 16 years later in 1945, Giuseppe Brotzu searched especially for microorganisms that inhibited bacterial development in sea water and therefore discovered penicillium acremonium that produces cephalosporins, another β-lactam antibiotic like penicillin.[5] In 2011, Brakhage assumed that most of these metabolites are produced for the competition and communication with other organisms.[6] In fact, secondary metabolites provide their producers to interact and compete with other organisms, unlike primary metabolites which are required to ensure growth of the organisms that produce them.[7,8] Hence, secondary metabolites not only include beneficial products like antibiotics and other pharmaceuticals (e.g. anti-cancer agents and immunosuppressant), but also toxins, hormones, insecticides and carcinogens.[1,6,9,10] Mycotoxins are indeed the subject of major concern because they lead to severe mycoses and mycotoxicosis.[11] Both beneficial and harmful properties of these metabolites lead to a greater importance to study them and their producers.

Today it is well-known that the genetic information of the specific organism forms the basis to produce these secondary metabolites. For the discovery of novel secondary metabolites, the genomic information from microorganisms is leveraged (see 3.3 Genome Mining).[12] Hence, microbial pharmaceuticals are now indispensable in the treatment of various clinical disorders. Furthermore, secondary metabolites are also used as herbicides, insecticides, and fungicides.[13]

But as a Darwinian consequence, the usage of microbial, antibiotic products led to an increased number of resistances to antibiotics and therefore concerns about these antibiotic resistances has risen in the early beginning of the 21st century leading to a more careful usage of antibiotics and pesticides, but also to an urgent need to discover new microbial products.[13,14] Furthermore, the access to novel natural products is limited, as most loci for secondary metabolites are disabled, also referred as “silent”, under specific, in most cases unknown circumstances, because of the absence of stimuli (e.g., nutrient sources).[15,16] Additionally, the main part of experimental evidence for secondary metabolites derive from bacteria, as they are fast and cost-efficient producers. In fact, it is estimated that the total number of bacteria species is around 100 million, whereas approximately 5 million fungal species exist but only about 99 000 of them were identified.[11,17] Hence, fungi might encode a vast potential for the discovery of novel natural products. Genome analysis indeed revealed that some fungal species possesses up to 80 BGCs in their genomes. This and the knowledge about the silent genes postulate that there are still much more compounds to find.[16,17,18]

3.2 Biosynthetic Gene Cluster (BGC)

The genome of an organism possesses the blueprints for producing secondary metabolites and is described by a unique nucleotide sequence. This sequence is built up by the four nucleotides: adenine (A), guanine (G), cytosine (C) and thymine (T). The unique composition of these nucleotides forms the genomic sequence. For fungi, the genome sequence lengths vary from 8,97 Mb to 177,57 Mb.[19] Segments of these genomic sequences that can be transcribed into a biological active mRNA are called genes. These mRNA segments are then translated into amino acid sequences, which are folded into a 3D structure forming the finished protein. The whole process constructing the product is called (gene)expression. However, secondary metabolites are mostly not produced by a single gene only, but by a collaboration of two or more genes that together encode the specific biosynthetic pathway to produce that specific metabolite.[20] As secondary metabolites are important for microbial

competing and communication, regulation is often done by environmental stimuli, including light, pH, carbon source, nitrogen source, reactive oxygen species and temperature.[21] Further investigations also revealed that they are often regulated by a cluster-specific transcription factor.[10 22]

Back in the early 21st century the first complete microbial genome sequences revealed two further important aspects; Firstly, these genes are typically located side by side, clustered into so-called **biosynthetic gene clusters (BGC)** with modest sizes containing up to 20 genes in fungal species.[13 22 23] These clusters contain all genes for the biosynthesis of all compounds needed in the biosynthetic pathway, such as precursors, modificatory, resistance, and regulation genes, to produce the final product(s).[23] Secondly, BGCs mostly contain one or more core biosynthesis genes encoding multimodular enzymes, like polyketide synthases (PKS) or non-ribosomal peptide synthetases (NRPS).[23] PKS and NRPS are multi-functional megasynthases that utilizes a biosynthetic mechanism similar to the fatty acid biosynthesis. In fact, PKS are homologous to fatty acid synthases.[6 10 13 24] Furthermore, it became apparent that the sequences of PKS and NRPS enzymes are very conserved, meaning that they rarely change during evolutionary processes, and have therefore highly predictive quality.[4 23] Thus, their typical structure can be visualized (see Table 3.1)

Table 3.1: Typical structures of biosynthetic gene clusters (BGC) encoding polyketides and non-ribosomal peptides.[20 25 26 27 28]

Products	Core enzyme	Typical domains	Typical additional domains
Polyketides	Polyketide synthase	Acetyltransferase (AT)	Ketoreductase (KR)
		Acyl carrier protein (ACP)	Dehydratase (DH)
			Enoylreductase (ER)
		Ketosynthase (KS)	Methyltransferase (MT)
Non-ribosomal peptides	Non-ribosomal peptide synthetase	Adenylation (A)	Thioesterase (TE)
		Condensation (C)	Reductase
		Peptidyl Carrier protein (PCP) = Thiolation (T)	Epimerisation (E)
			Methyltransferase (MT)

As shown in Table 3.1, a typical NRPS module minimally consists of three domains: an Adenylation domain (A) that activates the amino acids, a Condensation domain and Peptidyl carrier protein domain, also known as thiolation domain, that serves as an anchor for the growing peptide chain. Whereas a PKS module minimally consists of an acyltransferase (AT) domain for unit selection and transfer, an acyl carrier protein for unit loading and a ketoacyl synthase domain, that condensates decarboxylatively the unit.[6 25 26 27] Not all clusters produce their compounds using only one core enzyme. Some BGCs synthesize their products from two synthases and/or synthetases forming a hybrid BGC. Examples for this are the fumagillin cluster, a PKS-TC hybrid, or echinocandin, a PKS-NRPS hybrid.[16] Although, it became apparent that not all BGC comprise similar modularity like polyketides and non-ribosomal peptides, e.g., terpene BGCs. All fungal terpenoids derive from the common five-carbon isoprenyl diphosphate intermediates, isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), in short isoprenes. They are the start material for the biosynthesis of terpenoids and isoprenoids by either terpene cyclases (TC) or dimethylallyl-tryptophan synthases (DMATs).[29 30] Over the past view years, another fungal biosynthetic class was identified: ribosomally synthesized and post-translationally modified peptides (RiPPs).[31] They are synthesized by short precursor peptides comprising a leader peptide and a core peptide. After synthesizing the precursor in the ribosome, the core peptide is post-translationally modified by tailoring enzymes and then usually cleaved from the leader peptide, yielding the final product. Fungal

RiPPs include six classes: amatoxins/phallotoxins[32], and borosins[33], which both are found in the basidiomycetes. In Ascomycetes, RiPPs are produced by dikaritins[34], which are classified as ustiloxins, asperipins, phomopsins, and epichloëcyclins.[35 36 37] The knowledge about the different enzyme classes were used to discover novel BGCs (see 3.3 Genome Mining). To date, a huge amount of genomic data has been deposited in publicly accessible databases (see 3.4 Databases).[38] The availability of genome sequences and the research to identify the containing genes resulted in a large number of secondary metabolite gene clusters per organism, most of them unknown and therefore referred as "orphan" or "cryptic".[10 16]

However, another important aspect regarding the BGCs is that they are often concentrated at the telomeres (sub-telomeric regions) and at the centromeres.[10 16] It is well known that sub-telomerically located genes, like BGC genes, are often repressed, which is called telomere position effect.[16 21] Also, chromatin modifiers which control silencing and transcription effects have a high impact due to their location in sub-telomeric region.[10 16 21] This knowledge hardens the perception that most of the BGCs are silent under standard conditions.[10 15 21 23] In addition, the inability to cultivate some potential producers aggravates the research for novel secondary metabolites.[10 15] These findings emphasized the research for new approaches.

3.3 Genome Mining

At the time it came to known that microbes possessed an unexplored potential for producing secondary metabolites, the idea of genome mining was born, which involves the prediction and isolation of microbial products by using the genetic data of biosynthetic gene clusters (BGC).[4] In fact, the conserved characteristics of the core enzymes in biosynthetic gene clusters (BGC) are therefore exploited for genome mining purposes.[4 23] Over the time, two ways of establishing a link between a BGC and a secondary metabolite (SM) were established. Either the SM is identified for a specific BGC by elucidating its biosynthetic pathway, in the following referred as first strategy, or a BGC is identified for a specific SM by homology search, retro-biosynthesis, or comparative genomics, in the following referred as second strategy.[4 39]

For the first strategy the starting point is a putative BGC that is investigated. These strategies were summarized as molecular and epigenetics-based methods and methods that attempt to predict natural condition that led to activation.[6] A flow diagram of key questions determining the exact strategy was proposed by Inge Kjaerbolling *et* al. in 2019, starting with the question whether the cluster is in a cultivable host and if this fungus is engineerable.[39] If not, heterologous expression strategies are used, otherwise a homologous expression is possible, meaning that it can be expressed in the native host. In the latter case it is relevant whether the cluster is silent or not. Silenced clusters need to be activated first, e.g., by varying the environmental stimuli or overexpressing cluster specific transcription factors. Already activated clusters can be directly used for deletion strategies to elucidate the function of the cluster genes. This is usually done by sequentially deleting or disrupting genes followed by metabolite profiling. Thus, it is possible to identify metabolites missing in the strain and its intermediates.[6 39]

The second strategy is the association of a BGC with a specific SM it is producing (Figure 3.1).[39] The goal is the annotation of the identified cluster, which means the functional classification of the genes. This strategy is mainly based on bioinformatic prediction also referred to as in silico mining.[4 6] It describes the usage of genomic information for the discovery of new products, but also of new processes and targets using computing technologies.[4 12 38]
The starting point for this strategy is the selected secondary metabolite or, alternatively, SMs with already identified BGCs which are used for the search of similar BGCs. It concerns three in-silico methods namely homology search, retro-biosynthesis, and comparative genomics.[39]

During the first method (Figure 3.1, subpart strategy 1), homology search, a secondary metabolite with identified BGC is used for elucidating the BGC of a similar and known secondary metabolite by applying alignment tools, such as Basic Local Alignment Search Tools (BLAST) (see 3.5.1.1 BLAST). The second method (Figure 3.1, subpart strategy 2) starts with a chemically characterized secondary metabolite using an in-silico retro-biosynthesis to identify intermediates and compounds that are needed to produce the specific metabolite. For the third method (Figure 3.1, subpart strategy 3) a set of producing organisms is known and compared to each other in order to find and identify homologous gene clusters that produce the candidate metabolite. Such approaches use core genes, tailoring enzymes, or even phylogeny-based mining methods to find homologous sequences. A common example is antiSMASH (see 3.3.1 Cluster border definition and prediction algorithms). An example for phylogeny-based genome mining is EvoMining, which searches for homologues of housekeeping genes in secondary metabolite clusters.[4 6 38 39]

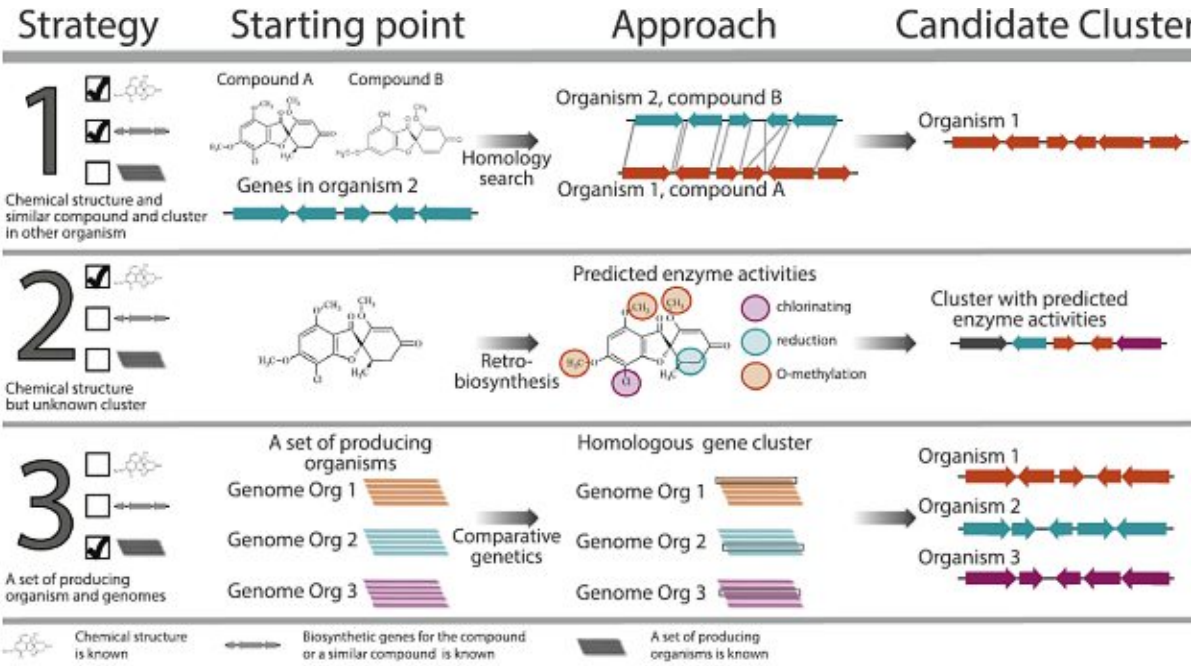


Figure 3.1: According to Kjaerbolling et al. in 2019, there are three bioinformatic methods (in the figure referred as "strategy") for linking secondary metabolites with a biosynthetic gene cluster (BGC). 1) Homology search defines a method at which a compound A and its BGC is already characterized and compared to a compound B and its genome sequence to elucidate the BGC of compound B. 2) Retro-biosynthesis is the prediction of enzymes needed to produce a chemically classified compound. 3) Comparative genomics enables the identification of homologous gene clusters for a candidate compound by comparing a known set of producing organisms and genomes.[39]

The second strategy (the linkage of a BGC with a secondary metabolite) ends in the identification of a putative BGC that can be verified using the first strategy by expressing the cluster homologously or heterologously.

In this thesis, the focus lies on the second strategy: By using comparative genomics based on phylogeny, the bioinformatic prediction enables the linkage of a BGC with a putative secondary metabolite.

3.3.1 Cluster border definition and prediction algorithms

As written in 3.2 Biosynthetic Gene Cluster (BGC) and 3.3 Genome Mining, the core genes (PKS and NRPS) are highly predictable due to the strikingly conserved biosynthetic principles and biosynthetic machineries, even though secondary metabolites are highly diverse.[4 23] This enables the bioinformatic prediction of BGCs.

The first reported tool for automated bioinformatic gene prediction was the proprietary DECIPHER® search engine and database published in 2003.[23] Six years later, the first open-source pipelines CLUSEAN and NP.searcher (both for bacteria) were released.[4 23] Finally, in 2011, the first version of antiSMASH, which, amongst others, incorporates the CLUSEAN pipeline, was released and was since then steadily extended. AntiSMASH is an open-source genome mining platform that enables large-scale genome mining studies.[23] Next to antiSMASH, other noteworthy tools have been developed, like PRISM, which have a focus on predicting chemical structures of the biosynthetic pathway, and SMURF, a tool for mining fungal PKS.[23] These cluster prediction tools rely on collections of protein domains found in known clusters.[22] Such domains are then searched in new sequences to find homologies and to that effect new clusters.[22 23] AntiSMASH identifies core enzymes like PKS or NRPS according to a collection of biosynthesis enzymes.

Hence, the prediction of secondary metabolites starts with identifying conserved biosynthetic genes and subsequently analysing them about their putative biosynthetic function. For this, gene annotations must be available on the genome of interest, which can be found in databases like GenBank® built by the National Center for Biotechnology Information (see 3.4.1 GenBank®).[23 40] If there are no annotations available, it is possible to use a gene finding tool, e.g., antiSMASH uses Prodigal for bacteria and GlimmerHMM for plants and fungi.[9 23] The next step is the identification of core enzymes, like PKS or NRPS, to identify BGCs. As these enzymes frequently share common patterns of amino acids, profile-based methods like Hidden Markov models (HMMs) are used to identify these patterns.[9 23] Once the core enzyme is identified, co-located genes are compared. In this connection, secondary metabolite genome mining tools, like AntiSMASH, PRISM and SMURF, use manually curated BGC rules following Boolean logic to decide whether these adjacent genes are part of the cluster or not.[23] An example is that BGCs encoding nonribosomally synthesized peptides typically contain at least one Condensation, Adenylation, and Peptidyl Carrier Protein domains, next to its core enzyme, NRPS.[23 41]

Today, antiSMASH incorporates two approaches for bacterial BGC (CLUSEAN and ClusterFinder) and one approach for both, bacterial and fungal BGC (NRPSpredictor).[23] The latter one, NRPSpredictor, is a machine learning approach that uses a high amount of core enzyme sequences to predict substrate specificity.[9 23] However, while all the above-mentioned tools work well for similar clusters, these tools have difficulties when searching for novel ones, especially for novel RiPPs and terpenoids clusters as they are not as highly conserved as PK and NRP clusters.[22 29 42]

An alternative is CASSIS/SMIPS, a toolkit that uses the biological principle that BGCs contain a higher density of shared transcription factors and common regulatory patterns to predict core genes.[9 23 43] Secondary Metabolites by InterProScan (SMIPS) is a genome wide detector for core genes, like PKS, NRPS or dimethylallyl tryptophan synthases (DMATS).[22] It can be used separately or together with Cluster Assignment by Islands of Sites (CASSIS).[22] CASSIS is a method for BGC prediction in eukaryotic genomes searching for cluster-specific motifs in the vicinity of the detected core genes.[22] It is a further development of the method Motif Density Method (MDM), assuming the density of binding motifs for cluster specific transcription factors (csTF) must be higher within the cluster and lower outside.[22] But, this method, too, uses sequence similarity searches and it is therefore not likely to find completely new BGCs.[44]

Hence, EvoMining, a tool based on phylogenetics (see 3.5 Phylogenetic trees), was created for bacteria and archaea species, which provide a genome mining approach using the evolutionary insights to discover novel biosynthetic gene clusters.[44] However, this approach refers to bacteria and archaea lineages, only.[44] Therefore, an approach based on fungal phylogenetics is still missing. During this thesis a suchlike, novel genome mining tool for fungal species is presented.

Nevertheless, Genome mining is a bioinformatic, in-silico method. Therefore, experimental verification of obtained putative BGC genes is inevitable.

3.4 Databases

The elucidation of genomes led to an enormous amount of data, which had to be organized and accessed easily. Hence, databases were built-up. There is a high number of databases used in genomic studies. One of the main databases for nucleotide sequences is GenBank®.[45] For biosynthetic gene clusters, one of the most important databases is Minimum Information about a Biosynthetic Gene cluster (MIBiG).[46]

3.4.1 GenBank®

GenBank® is a database that contains publicly available nucleotide sequences obtained primarily by submission supporting bibliographic and biological annotation.[45] Uploaded sequences receive an unique accession number for better retrieval. Also, GenBank® provides several other ways to retrieve data, e.g., BLAST tool to search and align sequences from GenBank® to a query sequences or Entrez Nucleotide to search for identifiers and annotations.[40] Data can be downloaded in GenBank® or Fasta format, both text based. GenBank® files have endings with .gb or .gbk and contain sequences, annotations, information about the organism and sometimes translated amino acid sequences, while Fasta files with endings .fasta, .fna, .ffn, .ffa or .frn contain amino acid or nucleotide sequences only. GenBank® is built by the National Center for Biotechnology Information (NCBI) and is extended every two months.[45] [40]

However, as the sequences and information depend on the uploads of researchers worldwide, the quality of the genomic data varies highly.

3.4.2 MIBiG

Minimum Information about a Biosynthetic Gene cluster (MIBiG) is a database that offers an improved access to information about secondary metabolites gene clusters. Building on Minimum Information about any Sequence (MIxS) framework, which is a standard for describing sequence data, MIBiG covers four general group parameters that each BGC must fulfil. These parameters include associated publications, description of the genomic loci, the chemical product and its features, and experimental verified genes.[47] [20] Furthermore, class-specific checklists for gene clusters must be fulfilled, e.g., there must be acyltransferase domain and starter units for polyketide BGCs or precursor peptides and peptide modifications for RiPP pathway clusters. Both, MIBiG and MIxS framework provide a characterization of a biosynthetic pathway.[20]

3.5 Phylogenetic trees

A phylogenetic tree is a diagrammatic representation comprising leaves, nodes and branches showing the evolutionary linkage among various taxa based on their molecular evolution at the level of nucleic acids or proteins.[48] [49] Hence, phylogeny is referred to as the evolutionary history of organisms or species.[48] [49] [50] [51] It forms the basis for comparative genomics, a research field for the comparison of genomic sequences showing the relation between them.[48] [49] [52]

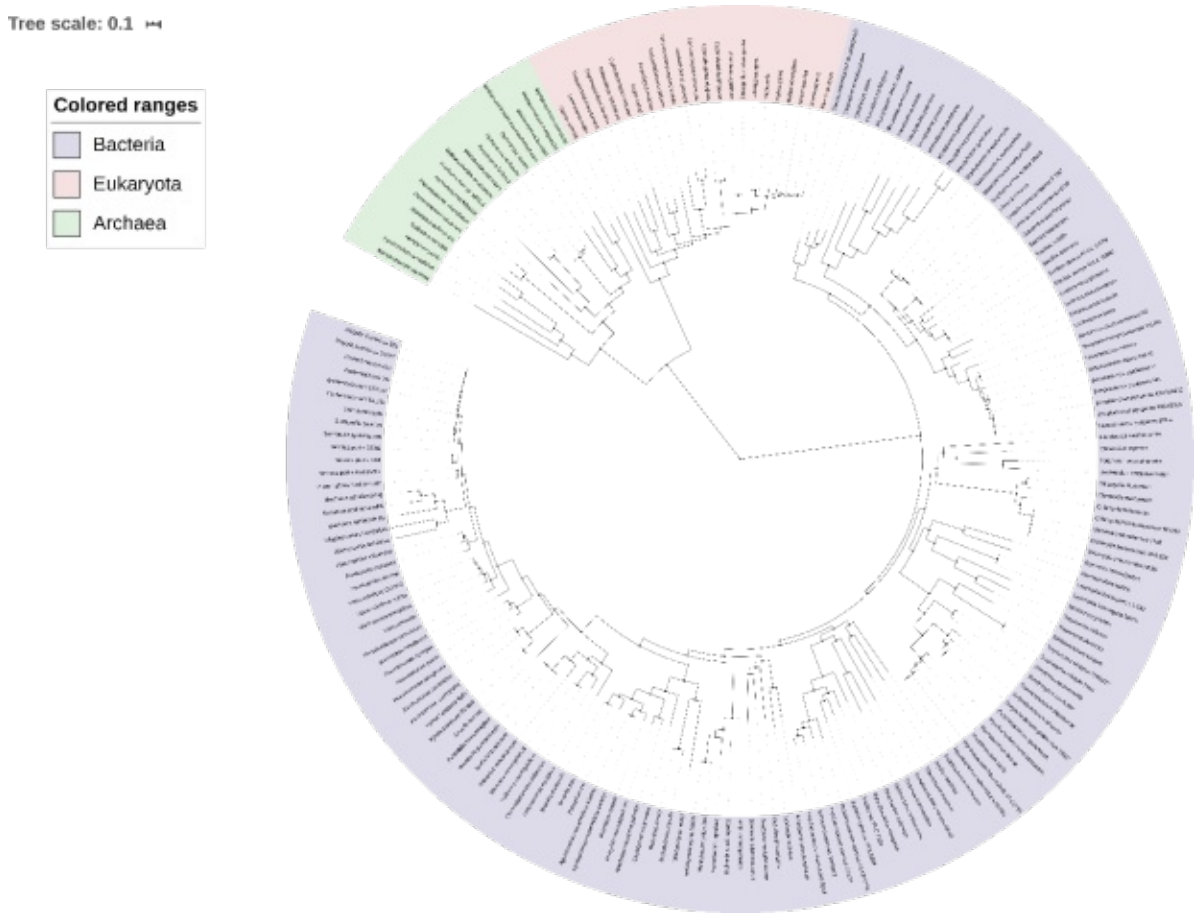


Figure 3.2: Overview over the Tree of life obtained from the Interactive Tree Of Life.[53] The three domains, Bacteria, Eukarya, and Archaea are linked together through the last universal common ancestor (LUCA), that is represented in the middle. As represented in blue, the most known living species belong to bacteria.[53] [54]

Evolutionary changes are caused by an accumulation of alterations of the genetic composition, varied by mutation, gene flow, genetic drift, or natural selection.[48] Hence, evolution leads to the formation of new species. The evolutionary divergences of the DNA or protein sequences between the species can be then shown in phylogenetic trees.[48] [55] Accordingly, in evolutionary history there must have been one theoretical primordial ancestral form, from which life on earth evolved 3.6 billion years ago, called the last universal common ancestor (LUCA).[48] [50] [51] All its descendants constitute the tree of life.[48] Such trees starting with the ancestor from which the rest of the tree diverges are called rooted trees.[48] [49] The contrary are unrooted trees, that are inferred more often.

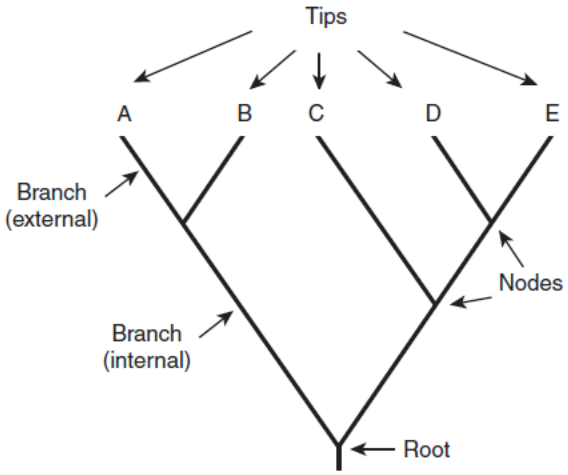


Figure 3.3: Components of a phylogenetic tree obtained from the book Encyclopedia of Evolutionary Biology, 2016.[56] Single species, individuals, or even sequences are represented by leaves (tips) that are evolutionary connected through branches. Two branches run into nodes that represent the last common ancestor. Rooted trees also exhibit an overall common ancestor (root).[56]

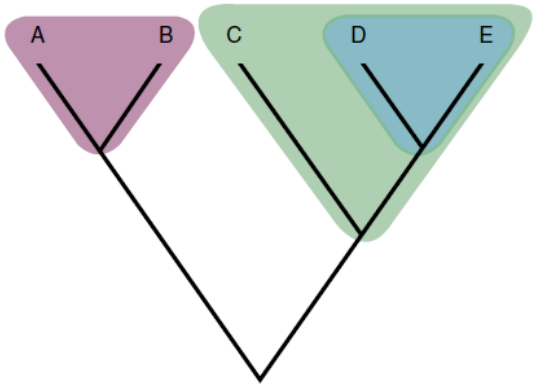


Figure 3.4: Highlighted groups are called clades, which include a node and all its descended lineages. Figure obtained from the book Encyclopedia of Evolutionary biology, 2016.[56]

Like a tree, a phylogenetic tree consists of branches and leaves. The leaves represent populations, species, individuals or even genes that are connected by branches.[56] A group of connected leaves with only one node are called clades. The branching points between two branches are called nodes, which represents the last common ancestors of the related leaves.[48] [56] At this point, the ancestor was split into two different descendant lineages, whose individuals did not exchange genes anymore. Although, a lineage splitting event occurred, it is not necessarily based on trait divergence, the accumulation of mutation can lead to a trait evolution in the descendant lineages.[56]

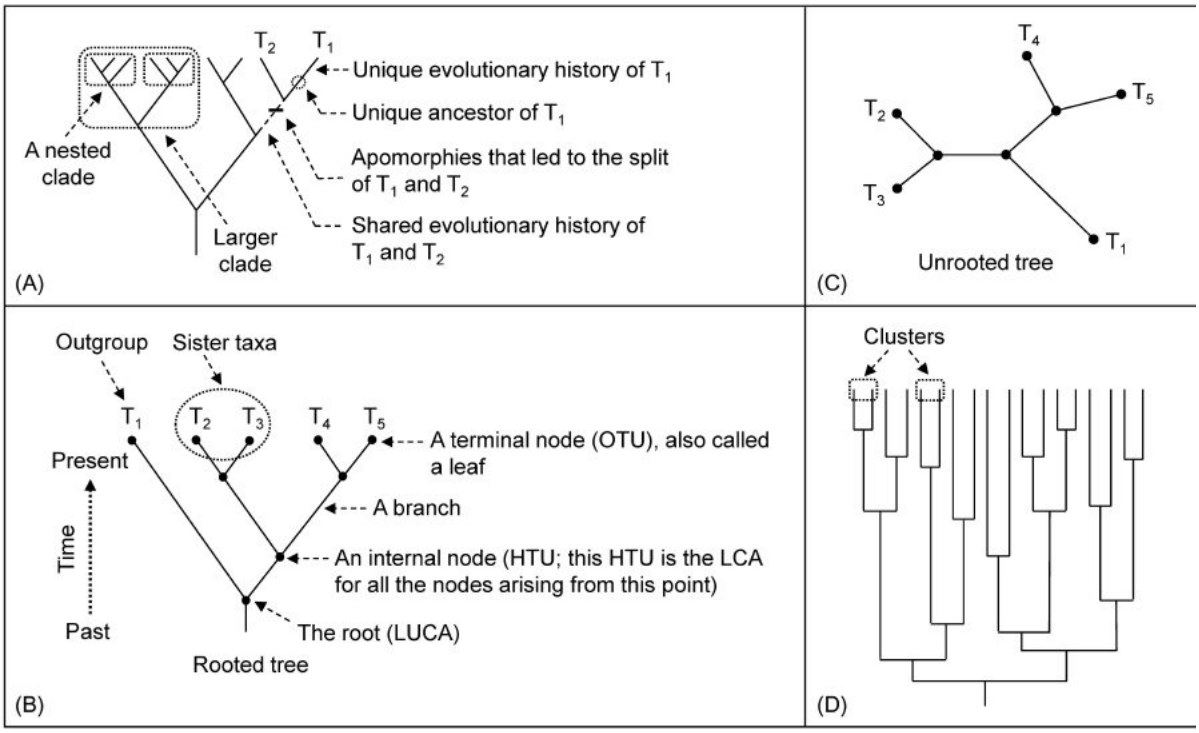


Figure 3.5: Various types of phylogenetic trees. The figure was obtained from Bioinformatics for Beginners, 2016.[49] A, B and D show rooted trees, while C represents an unrooted one. While A shows an unscaled tree (cladogram) and B a scaled one (phylogram), the tree in D represents a dendrogram derived from hierarchical clustering.[49]

Phylogenetic trees can be unrooted or rooted, but also scaled or unscaled.[49] In scaled trees, branch lengths are proportional to evolutionary divergence, comprising e.g., the amount of nucleotide substitutions.[49] There are therefore various types of phylogenetic trees; Cladograms represent a hierarchical but unscaled tree topology showing the relationship of clades, while Phylograms are scaled trees, which means the longer the branch the more changes occurred.[49] A dendrogram is an arrangement of clusters showing their relationship and is not solely used in phylogeny, but also outside bioinformatics.[49 56]

The construction of phylogenetic trees starts by first aligning two or multiple sequences according to their evolutionary relationship. In this connection, mostly amino acid sequences are used.[49 55] In the second step an evolutionary model is determined. Based on that and the alignment the distances between the sequences is estimated by either distance-based or discrete methods.[49] Using these methods, the phylogenetic tree is constructed.[49]

The study of phylogenetic trees gives insight into the evolution of genes, genomes, and species. For this, two or more phylogenetic trees are compared, which is discussed in more detail in 3.5.4 Inferring Co-evolution.[57]

3.5.1 Multiple sequence alignment

As written in the previous chapter, the inference of phylogenetic trees starts with the alignment of two or more sequences, which is the most important step in the construction as a good alignment yields a reliable tree.[49] This multiple sequence alignment (MSA) comprises two main assumptions; Firstly, the sequences are homologous and secondly, point mutations evolved independently.[49] Homologous segments in genomic data are defined as sharing a common ancestor.[49 58] However, MSA is an important method, on which, next to phylogenetic reconstruction, many other in silico analysis depend, e.g., domain analysis and motif finding. The goal is to infer the evolutionary, functional, or structural relationship of the sequences. For this, homologous sequences are aligned, and gaps are inserted, if needed.[59] For the construction of phylogenetic trees, the evolutionary

relationship is represented, and the gaps therefore stand for insertions and deletions (indels) within the genome. Such changes in the genome are hypothesized as evolvements from a common ancestor.[58] [59]

As previously explained, MSAs consist of two or more pairwise alignments between the given sequences, which have similar lengths. Pairwise alignments are used to find the best match between two query sequences. This can be done globally or locally. Global alignments assume that the two sequences are basically similar over the entire length, trying to align every residue of the sequences, while local alignments only search for segments that match well.[60] [61]

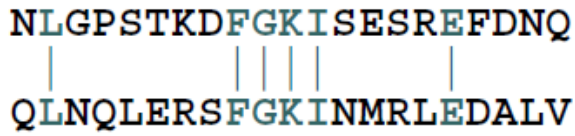


Figure 3. 6: Example for a global alignment.[60]

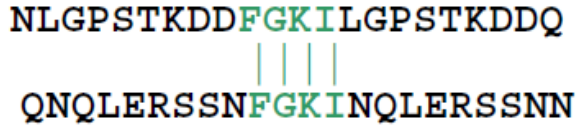


Figure 3. 7: Example for a local alignment.[60]

A popular approach for global pairwise alignment is the Needle Wunsch algorithm, which is using two-dimensional arrays representing every possible comparison between two amino acid or nucleotide sequences by pathways through the array, yielding the best similarity score by backtracking.[62]

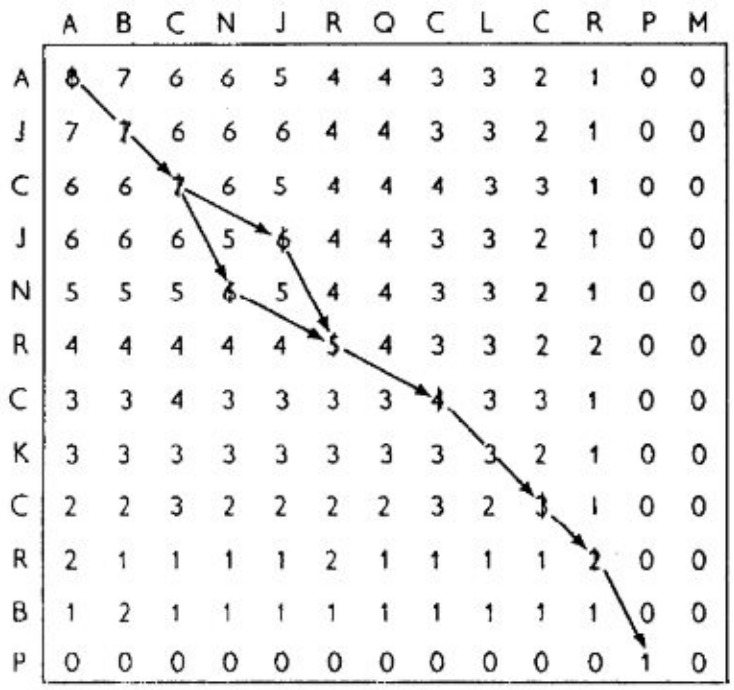


Figure 3.8: Completed similarity matrix using Needleman & Wunsch algorithm showing the best pathway through the array obtained from Needleman et al., 1970.[62] *Numbers in each cell represent the largest number of identical pairs that can be found if that cell is the origin of the pathway. Here, for each identical pair, the value one was given, non-identical pairs were given the value zero, scoring the maximum match, terminating at the largest number in the first row or column, 8 in this case, which is the similarity score.*[62]

In contrast, Smith-Waterman algorithm is a general local alignment approach. Such local alignment tools are generally preferred for database searches.[63]

	0	C	A	G	C	C	U	C	G	C	U	U	A	G
0	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
A	*0*	0	5	0	0	0	0	0	0	0	0	0	5	0
A	*0*	0	5	2	0	0	0	0	0	0	0	0	5	2
U	*0*	0	0	2	0	0	5	0	0	0	5	5	0	2
G	*0*	0	0	***5***	0	0	0	2	5	0	0	2	2	5
C	*0*	5	0	0	***10***	5	0	5	0	10	1	0	0	0
C	*0*	5	2	0	5	***15***	6	5	4	5	7	1	0	0
A	*0*	0	10	1	0	***6***	12	3	2	1	2	4	6	0
U	*0*	0	1	7	0	5	***11***	9	1	0	6	7	1	3
U	*0*	0	0	0	4	4	10	***8***	6	0	5	11	4	1
G	*0*	0	0	5	0	3	1	7	***13***	4	3	2	8	9
C	*0*	5	0	0	10	5	0	6	4	***18***	9	8	7	6
C	*0*	5	2	0	5	15	6	5	4	9	15	6	5	4
G	*0*	0	2	7	0	6	12	3	10	8	6	12	3	10
G	*0*	0	0	7	4	5	3	9	8	7	5	3	9	8

Figure 3.9: Completed similarity matrix for local alignment of two small sequences using Smith-Waterman algorithm, showing the best matches by high numbers obtained from Khajeh-Saeed et al., 2010.[64] The best local alignment is highlighted in blue, yielding the following aligned sequences: GCC-UCGC and GCCAUUGC.[64]

Usually, protein sequences are used for the alignment for phylogenetic tree inference, as they have more characters (20) than nucleotides (4), amino acid matrices are more sophisticated, and there is no codon bias for the same amino acid in different species.[49] [55] But multiple sequence alignment can also be applied to DNA and RNA sequences.

3.5.1.1 BLAST

Another heuristic, local alignment algorithm is the Basic Local Alignment Search Tool (BLAST).[63] [65] It is an important and robust tool for database searching, especially in GenBank®, aligning the query sequence with the sequences deposited in the database searching for the highest similarity scores.[65] As the functions and properties of a gene underlie their structure and composition, a BLAST search gives hints about that.[66]

Depending on the given and the target sequence, there are various BLAST programs, e.g., blastp provides an alignment of a protein sequence with protein sequences from the database, while blastn uses a given nucleotide sequence to search in a nucleotide database. There is also the possibility to receive translated sequences, e.g., with blastx.(Information, 2008 #649) [65]

To measure the local similarity a maximal segment pair (MSP) score was introduced by Altschul *et* al. and implemented in BLAST.[63] The MSP score is getting higher, the longer identical segments of two sequences are. Also, BLAST allow the implementation of substitution matrices, like PAM, and can be mathematically analysed (see 3.5.2. Evolutionary distances).[63] This makes BLAST a valuable tool for genome mining.[63]

3.5.2. Evolutionary distances

A vital prerequisite for creating an accurate MSA is representing evolutionary relationship, which is done by the determination of evolutionary distances.[49]

At this juncture, the simplest way to calculate evolutionary distance between two sequences is using the uncorrected p-distance, quantifying the number of substitutions:

$$p = \frac{D}{L} (Bawono, 2014\ \#240)$$

L… Number of positions in the sequence alignment, excluding gaps
D… Number of positions that contain different residues

However, this is not very accurate, as a substitution may have evolved from another substituted nucleotide leading to an underestimation of true evolutionary distance.[50] Hence, an evolutionary model must be determined, which is another important step for the reconstruction of phylogenetic trees, as they are used for calculating the evolutionary distances correctly, based on biological principles.[49] [50]

In 1965, the molecular clock hypothesis was proposed by Zuckerkandl and Pauling, which assumes that all species evolved with constant rate.[50] Shortly after, Jukes-Cantor one-parameter model was proposed by Jukes and Cantor in 1969 assuming all nucleotides occur in equal frequencies.[49] [67] However, both have been found to be oversimplified, because nucleotides and amino acids do not occur in equal frequencies and studies have shown that this is a result of different evolutionary pressures at different times.[50] [67] Furthermore, there are hot spots for mutations in the genome, depending on the chromosomal position, G+C content and the efficiency of their repairing system.[67] It is also known that the nucleotide replacement by a similar base facilitates further misincorporation errors.[67] This is why transition mutation (A ↔ G and C ↔ T) are more common than transversion mutations (A ↔ T, G ↔ C).[49] [67] Later on, in 1980, Kimura's two-parameter model for nucleotide sequences was proposed, that incorporates the different mutation rates.[49] More complex algorithms are the Felsenstein model and the Hasegawa-Kishono-Yano model, which assume both, the different frequencies of nucleotides and the different rates of mutation type.[49]

For amino acid sequences there are separate substitution models. The simplest model is the Bishop-Friday model that does not incorporate different amino acid frequencies and different substitution rates.[49] More complex models are Dayhoff's model, leading to PAM (Percent accepted mutation) and Henikoff's model, leading to BLOSUM (Blocks substitution matrix).[49] [68] [69] Both are substitution matrices used for similarity measurements during alignment of protein sequences that represent all possible exchanges of amino acids. Next to the mutation rates, they incorporate the relative frequency of amino acid pairs estimating the evolutionary likelihoods of conservations and mutations of residues in amino acid sequences.[49] However, BLOSUM are mostly used for database searches, as the approach perform better in alignments and homology searches than PAM, while PAM is usually used for phylogenetic relations.[69]

3.5.3. Tree Construction

The final phylogenetic tree is inferred based on the multiple sequence alignment. There are two methods applicable: The distance-based method and character-based (discrete) method.

The first class, distance-based methods, are based on cluster algorithms producing distance matrices to estimate the evolutionary distance. As an adjustment of the received distances, the previous determined evolutionary model is used. The most common examples here are neighbour joining (NJ), Unweighted Pair Group Method with Arithmetic mean (UPGMA), the least squares, and the minimum evolution methods. Among them, UPGMA and NJ method are the most common ones. Both works quite similar; they are clustering procedures, constructing clusters (UPGMA) or distance matrices by Hamming distance (NJ) at each step yielding the final phylogenetic tree. However, as UPGMA is based on the molecular clock hypothesis, it is not often used today.[70] Furthermore, despite their advantages to be very fast, applicable for close entities, and computationally efficient, these methods also undermine evolutionary relationship of distant sequences, e.g., NJ lacks information when sequences are converted into distances.[66] [71] [72]

The second class for tree construction, the discrete method, involves maximum parsimony (MP) and maximum likelihood (ML) methods utilizing the sequences themselves rather than their pairwise distance. Unlike distance-based methods, character-based methods use evolutionary models at all stages during the tree-building process. Maximum parsimony assumes that the simplest tree is the

most plausible one, using the smallest number of evolutionary changes needed by one sequence to convert to another. It therefore assumes, that change over evolutionary times is improbable, which is considered as wrong.[73] Also, maximum parsimony can perform quite poorly when the branch lengths vary.[66] Furthermore, the more taxa for reconstructing the phylogenetic tree are used, the more tree topologies become possible.[74] To overcome these disadvantages, Maximum likelihood was developed, which is a statistical method using the given data to determine the probability of substitutions to construct the phylogenetic tree with the most probability that the selected evolutionary model is predicting.[73] It is a method, that can be evaluated, and it is able to give insights into sequence evolution. However, it is computationally demanding and can be therefore quite slow.[49 71 72] A further development is Randomized accelerated maximum likelihood (RAxML), a fast, maximum likelihood tree search algorithm (see 3.5.3.1 RAxML).[75] Yet these approaches have one disadvantage; they can miss the global optimal tree.[66]

One of the first multiple sequence alignment and tree construction tool was Clustal, creating independently pairwise alignments calculating all distances between sequence pairs.[58] While, in the original Clustal programs, initial guide trees were used to establish the multiple alignment using UPGMA, ClustalW ("W" for "weights"), developed in 1994, use neighbour-joining methods and comprise weight matrices which depend on the estimated evolutionary sequence history.[76 77] Current Clustal programs derive from that 1994' ClustalW and have been amended and added to many times.[77] Since then, other more improved alignment tools such as MUSCLE, t-Coffee and MAFFT were developed, which perform iterated alignment steps to assure its accuracy.[58] Today, three types of Clustal programs exist: the classical tools ClustalX and ClustalW, and an optimized, faster and scaleable one, ClustalΩ.[78]

As the Clustal programs are not freely available, the European Molecular Biology Open Software Suite (EMBOSS) was established, providing free open-source software analysis packages, covering applications for e.g., sequence alignments, identification of protein motifs or rapid database searching with sequence patterns. It also provides multiple sequence alignment by ClustalW using the EMBOSS application EMMA, which is a property of the program that was evaluated during this thesis.[79 80]

3.5.3.1 RAxML

Maximum likelihood methods have one big disadvantage: they come with high computational costs, depending on the thoroughness of the search. Hence, randomized accelerated maximum likelihood (RAxML) a significantly faster search algorithm based on maximum likelihood for high performance computing was developed. Its focus is on the computation of large phylogenetic trees with over 1000 taxa, starting with building an initial parsimony tree using stepwise addition for tree building.[74 75] Stepwise addition, a method attaching new sequences on three starting sequences yielding an optimum at each step, has two main advantages; firstly, it is fast, and secondly, later steps allow the reversion of earlier pairing decisions.[66] Disadvantages are that it solely yields one tree, which often has no global optimum, and it is not as fast as neighbour-joining.[66] The phylogenetic tree is optimized by iterated subtree rearrangements, which are repeated until no better topology is found. Today's RAxML supports not only DNA and protein data, but also RNA and binary, multi-state morphological data. It also offers four different bootstrapping methods and several post-analysis functions, yielding a robust tree.[75]

3.5.3.2 Bootstrap

Bootstrapping is a nonparametric, resampling method for inferring uncertainty for phylogenetic trees which was first introduced by Felsenstein in 1983.[72 81] In this connection, the original data are randomly sampled and replaced, receiving bootstrapped values, which are then compared to the

original estimates. If a bootstrapped tree is different to the original tree, it can be assumed that the underlying data have weak evolutionary signals, and vice versa.[72] Though, it is not a test on how reliable the tree is, but how stabile it is.[81]

Other resampling methods are cross-validation, jackknifing, which is also called "leave n out" procedure, and Bayesian simulation.[72] [82]

3.5.4 Inferring Co-evolution

In 1959, Demerec and Hartman postulated that gene clusters will not be separated by natural selection as they confer an evolutionary advantage for their host organism.[83] Hence, the comparison of two or more evolutionary trees is an important aspect for inferring co-evolution among organisms and deciphering geographical areas. To achieve this, different metrics can be distinguished computing the dissimilarity between two phylogenetic trees.[57] They can be divided into two groups, firstly, counting the minimum number of operations required to transform one phylogenetic tree into another tree, e.g., nearest neighbour interchange (NNI) or subtree-prune-regraft (SPR) distances. The second group include distances in which the information of two phylogenetic trees is split into sets. Those sets are then compared, and their similarity is measured. Such methods include the Robinson-Foulds and the Quartet distances.[57] [84] However, all these algorithms share one main disadvantage, namely leaves are compared to only one leaf of the other tree, and they, hence, are unable to deal with trees that contain gene duplication or gene loss.[85] This means that only 11 to 37% of the fungal genomes are accounted with these methods.[85]

An alternative is TreeKO algorithm, which measures the number of inferred duplications and losses events by using backward selection (also called pruning, or decomposition).[85] The similarity between pruned trees is calculated using the RF distance formula, yielding a so-called strict distance.[85] Hence, the strict distance is a weighted RF penalizing gene duplications and gene losses, which is most appropriate when searching for protein families with a similar evolutionary history.[85]

The evaluated program in this thesis uses a TreeKO algorithm to automatically calculate the distances of the computed phylogenetic trees. One goal of this thesis was to range these distances by using the data from the manual evaluation.

3.6 Multivariate Statistics

When using basic statistics, usually univariate data is applied, utilizing only one dependent variable. A dependent variable is the input variable which depends on other, independent variables. For example, when investigating the dose effect of a substance on the frequency of symptoms, the dose is the dependent and the frequency of symptoms is the independent variable.[86]

However, natural behaviour is usually not described by only one input variables, but by a multitude of variables. Such methods and algorithms are called multivariate statistics. A special case is bivariate statistics comprising two variables, yielding the relationship between them.[87]

In the following sub chapters, the methods used in this thesis are described.

3.6.1 PCA

Principal component analysis (PCA) is one of the most popular multivariate statistical approach allowing the examination of the relationship among variables explaining their variability as much as possible by linearly combining them. The goal of a PCA is to represent similarity patterns by extracting the most important information. In a pre-processing step, the data are centred and either a covariance matrix by dividing all elements with the root of the observations or a correlation matrix by standardizing the data is established. During the analysis, the variables are linearly combined to new orthogonal variables called principal components based on singular value decomposition (SVD).

SVD allows an optimal approximation of a matrix by reducing the dimensionality yielding a second matrix comprising the principal components. On a geometrically view, a linear combination of a two-dimensional data point comprising a x and a y value represents its new definition by other lines at any angle from any direction receiving new values x' and y'. For a set of data points, the variability is summarized by the standard deviation and then maximized by increasing the standard deviation to a maximum. The whole process is called projecting a data point. The new values, called factor scores, are therefore projections of the original observations.[82] [88]
But PCA are not only used for descriptive but also for predictive purposes, where the values of the novel observations shall be estimated by the PCA model. For such cases, an evaluation is needed by using resampling techniques (e.g., bootstrapping, or cross-validation).[82]

One main disadvantage regarding PCA, but also other established methods, like partial least squares (PLS), is the need of complete datasets, which is seldom the case in research. Hence, missing data must be eliminated, or estimated by various methods. This is a crucial step, as outliers have a high impact on PCA results.[88] [89]
A popular technique is the NIPALS algorithm based on linear regression, which can be applied on random missing data patterns.[90]
Another method is the non-parametric MissForest algorithm, a random forest approach, which can handle any type of data, even mixed data. While the original random forest algorithm requires dependent variables without missing data for training, MissForest uses the given observations to directly predict the missing values.[89]

In this thesis, a MissForest algorithm was used to perform PCA.

3.6.2 PLS DA

As the name implies, Partial least squares-discriminant analysis (PLS-DA) is an algorithm used for classification of multivariate data, combining dimensionality reduction and discriminant analysis. It involves several mathematical steps facilitating predictive and descriptive modelling as well as discriminative variable selection.[91]
To classify datasets, categorical variables of the training set must be recoded into continuous variables first. In this junction, two algorithms, PLS1-DA and PLS2-DA, can be distinguished: The former is used for binary modelling, while the latter is applied on multi-class problems.
Next, the covariance between input and output variables is maximized and subsequently scores and loadings are determined. These data are used to estimate the regression coefficient receiving the first PLS component. This procedure is repeated as often as PLS components are required for the desired model using the last calculated residuals as new input and output variables. The outcome is a regression coefficient matrix as well as all needed PLS components that can be applied on a test set. The classification of the test set allows its description or prediction.[91]

3.7 FunOrder

FunOrder, short for fungal cluster border prediction based on computational molecular evolution, is a program written in python and bash by Mag.pharm. Gabriel A. Vignolle. It works with amino acid sequences from GenBank® files creating phylogenetic trees using a manually created fungal database. Its goal is to predict the genes involved in the biosynthesis of a fungal secondary metabolite and by this facilitate the molecular research for novel biosynthetic gene clusters.

3.7.1 Subprograms

FunOrder is a program for detection of the genes necessary for the biosynthesis of a specific compound within a BGC and further for cluster border discrimination using RAxML and EMBOSS.

In the first step FunOrder extracts the amino acid sequences and split them from a GenBank® file into single FASTA files. Next, a sequence alignment for each sequence is done using blastp on a manually

created fungal database. This database contains 134 proteomes over the Ascomycota tree of life and 2 basidiomycetes. The latter are used as an outgroup. A remote BLAST search can be performed if desired. The top 20 hits for each sequence are listed and a multiple sequence alignment by means of EMMA, a ClustalW wrapper from EMBOSS, is done creating dendrograms. Subsequently, the phylogenetic trees are constructed using RAxML. Furthermore, strict distances and speciation distances between all generated trees are calculated via TreeKO algorithm. However, these distances were not used during evaluation.

3.7.2 Aim

The aim of this project was to evaluate the program fungal cluster border prediction based on computational molecular evolution (=FunOrder) and analyse its calculated phylogenetic manually, based on validated methods. During the analysis, the goal was to answer the following questions:

1. Has FunOrder the ability to predict positive genes correctly?
    a. If Yes, can it differentiate between core genes and adjacent genes?
2. Can FunOrder differentiate between positive and negative controls?

The evaluation data were then used to define the borders of the co-evolutionary distances calculated by the TreeKO algorithm of FunOrder to be able to use the program. The project was therefore a possibility to define the borders of co-evolution between protein families within BGCs in fungi. As the distances calculated by FunOrder were not handed over to the writer, the evaluation was executed as a blind study.

As FunOrder is a program that predicts genes involved in the biosynthesis of secondary metabolites like pharmaceuticals, FunOrder can be used in the drug discovery as well as in other research topics regarding secondary metabolites from fungi.

4. Materials and Methods

4.1 FunOrder

Fungal cluster b**order** prediction based on computational molecular co-evolution (FunOrder) is a program written by Mag.pharm. Gabriel A. Vignolle in the programming languages python, perl regular expressions, and bash. Based on a GenBank® file (.gbk or .gb) as input, the amino acid sequences are extracted and converted into fasta files using the package biopython and the program "convert genbank to fasta" by Cedar McKay and Gabrielle Rocap, University of Washington.[92] The converted fasta files are then split into single fasta files, that are called queries hereinafter, by the emboss function seqretsplit containing one sequence each.[80 93] A sequence similarity search using blastp of the package BLAST[65] for each fasta file is carried out and the top 20 hits are outputted. The used database for the search contained 134 proteomes over the Ascomycota tree of life and 2 basidiomycetes and was manually created by Mag.pharm. Gabriel A. Vignolle. Duplicated sequences in this database were removed prior using. A further remote sequence similarity search using BLAST in the database from NCBI[65 94] remains optional. Sequence duplications originating from using a query sequence that also occurs in the database were removed using a custom perl script. The obtained hits are then aligned against the query sequence using the multiple sequence alignment function emma[95] from the emboss package[80 93] yielding dendrograms and a multiple sequence alignment. To create the final phylogenetic tree, a rapid bootstrap analysis and a search for the best-scoring maximum likelihood tree using the RAxML package[74 75] is executed. Finally, the phylogenetic trees are compared using the free treeKO software by Marina Marcet-Houben and Toni Gabaldon[85] written in python, where strict distances and speciation distances between all generated trees are calculated.

4.2. Positive controls

A set of 29 experimentally verified positive controls were obtained from literature research. These controls were restricted to biosynthetic gene clusters (BGC) with defined cluster borders, tested by *in vitro* methods such as gene inactivation or heterologous recombination. All sequences except of four derived from Minimum Information about a Biosynthetic Gene cluster (MIBiG) database. Two of the remaining four sequences were defined by their locus tags based on their respective literature, one sequence was downloaded directly from GenBank®. The last sequence was received using a sequence similarity search in the proteome of the fungus and each gene sequences were downloaded separately from GenBank®. FunOrder was then fed with the positive controls and tree comparisons were made according to chapter 4.4. Tree comparison. Because of the high number of genes and species-to-query-similarities within two trees, not all genes within a cluster could be compared. Therefore, nearly all positive clusters had missing data. To receive a meaningful evaluation, genes necessary for the biosynthetic pathway according to the respective literature were specified. Phylogenetic trees of those genes were compared. Furthermore, core genes like polyketide synthases (PKS) and non-ribosomal peptide synthetases (NRPS) were compared to all genes in their respective cluster.

4.3. Negative controls

Two groups of negative controls were used for evaluation: 42 synthetic BGCs and 60 random BGCs, which are defined in the following subsections.

4.3.1. Random BGCs

To establish the random BGCs, various amino acid sequences of random sizes were randomly sampled from fungal proteomes of the database and then randomly concatenated. FunOrder calculated multiple phylogenetic trees for each cluster which were used as negative controls for

further evaluation. Tree comparison was made according to chapter 4.4. Tree comparison. Within each random cluster all genes were compared with each other.

4.3.2. Synthetic BGCs

Synthetic BGCs were completely randomly generated by creating ATGC strings and put together to generate a DNA sequence. In a further step, those in-silico nucleotide sequences were translated into an amino acid sequence. FunOrder analysed those synthetic BGCs, but no phylogenetic trees from synthetic controls could be established.

4.4. Tree comparison

The phylogenetic trees computed by FunOrder, representing each a gene within a cluster, were examined, and compared using the online tool phylo.io.[96] As the coevolutionary linkage was of interest, similar leaves between two trees were determined. No similar leaves indicated that the respective genes did not share any evolutionary traits. To receive a meaningful evaluation, the following parameters were examined: branch length differences, node differences, and branch colours between the leaves and query, and the overall topologies of both trees. The branch lengths were measured with a ruler, converted into the distances using the caption of phylo.io[96] and the differences were then calculated. The nodes between a species and the query were counted and subtracted with the number of nodes of the second phylogenetic tree, yielding the node differences. The branch colour was a tool from phylo.io[96] that described the similarity to the most common node. Other than the nodes and branch lengths, the branch colours were defined as a categorical class were 0 to 40% of similarity was represented as “yellow”, 40 – 66,6% was defined as “green” and the rest was regarded as “blue”. The topologies of two phylogenetic trees were specified using phylo.io[96] and compared in five different ways, that can be found in Table 4.1.

Table 4.1: Definition of the parameter “topology”. To specify the coevolutionary linkage between two genes, the topologies of their phylogenetic trees were compared using the tool phylo.io[96], next to the branch lengths, nodes, and branch colours. During the evaluation, five different ways to define the topology comparison were used: same, very similar, similar, somewhat similar, and different.

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	Δdistance < 2, colour min. 50%
somewhat similar	either 1 or 2 similar species with Δdistances < 0.5 and Δnodes <3, or more species but only little similarities
different	no similarities or only 1 similar species

Using the conversion explained in Table 4.2, the parameters were converted into numerical data to enable a statistical evaluation.

Table 4.2: Caption for the conversion of the parameters to receive numerical data only. The measures were used to calculate the manual evaluation measures (MEMs) and average manual evaluation measures (aMEMs). ΔDistance, branch length distances between a leave and the query compared to another phylogenetic tree; ΔNodes, node differences between a leave and the query compared to another phylogenetic tree; Colour, branch colours according phylo.io[96] between a leave and the query compared to another phylogenetic tree; topology, comparison of the topologies of two phylogenetic trees

ΔDistance	ΔNodes	Colour	Topology	Measure
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

Average pairwise distances of all four measures were determined. If the trees contained more than two similar leaves, another average pairwise would be calculated, called manual evaluation measure (MEM). These MEMs were put together in matrices to evaluate the coevolutionary linkages between genes and to decide, whether FunOrder could predict positive genes correctly and whether it could differentiate between core genes and adjacent genes. The matrices consisting of the MEMs were therefore used to calculate heatmaps, dendrograms and principal component analysis (PCA).

To evaluate whether FunOrder can differentiate between positive and negative controls and predict biosynthetic gene clusters correctly, further average pairwise distances per cluster of the previously obtained MEMs were determined, called average manual evaluation measures (aMEM). They were used for building up a confusion matrix and a receiver operating characteristics (ROC) curve.

4.5. Statistical Evaluation

Statistical evaluation of the manual evaluation measures (MEMs) was made in RStudio, Version 4.0.2. The R script used for evaluation is attached in the supplement (Supplement 9.1 and Supplement 9.2.) The data was imported from .csv sheets. The column names were defined, and matrices applied numerically using the R function matrix. Because the rownames were removed during the import, the columnnames where set as rownames. From these prepared matrices, heatmaps were constructed applying the R functions heatmap.2 from the package gplots.[97] To eliminate missing data in the positive controls for further plots, a random Forest approach (MissForest package) was used. Then, datasets were scaled using the function scale, and next Euclidean distance applying the dist function from the package stats were computed on both, scaled and unscaled data. Subsequently, a Ward clustering was executed using the hclust function of stats yielding the final dendrograms. Within the hclust function, two different methods were applied; ward.D which did not implement Ward's clustering criterion from 1963, and ward.D2 which implemented that criterion. With the latter one the dissimilarities were squared before cluster updating.[98] In the end, four different dendrograms with the parameters scaled or unscaled data, and ward.D or ward.D2 approach were established for negative controls. Positive controls further included dendrograms with missing data or with approximated data. The principal component analysis was computed using the datasets approximated by MissForest approach and the pca function from the package mdatools.[99]

To decide whether FunOrder can differentiate between positive and negative controls, the average manual evaluation measures were used to establish a confusion matrix. Because of the high amount of missing data in the positive controls, two different thresholds were used: 1.5 for negative controls and 2.0 for positive controls. According to the thresholds, true negatives (TN), and true positives (TP), as well as false negatives (FN), and false positives (FP) were determined to establish the confusion matrix and to calculate the performance metrics. The formulas were listed in Table 4.3.

Table 4.3: Formulas of the respective performance metrics based on the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics were used to analyze the performance of the classification.[100]

Performance metrics	Formula
True positive rate (Sensitivity)	$Sensitivity\ [\%] = \frac{\sum TP}{\sum TP + \sum FN} * 100$
True negative rate (Selectivity)	$Selectivity\ [\%] = \frac{\sum TN}{\sum TN + \sum FP} * 100$
False positive rate (FPR)	$FPR\ [\%] = \frac{\sum FP}{\sum TP + \sum FN} * 100$
False negative rate (FNR)	$FNR\ [\%] = \frac{\sum FN}{\sum TN + \sum FP} * 100$
Accuracy	$Accuracy\ [\%] = \frac{\sum TP + \sum TN}{\sum TP + \sum FP + \sum TN + \sum FN} * 100$
Positive predicted value (Precision)	$Precision\ [\%] = \frac{\sum TP}{\sum TP + \sum FP} * 100$
Negative predicted value (NPV)	$NPV\ [\%] = \frac{\sum TN}{\sum TN + \sum FN} * 100$
F_1 score	$F_1[\%] = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} * 100$

A further performance metrics was calculated to evaluate the quality of the classification, namely the Matthews correlation coefficient (MCC):

$$MCC = \frac{\sum TP * \sum TN - \sum FP * \sum FN}{\sqrt{(\sum TP + \sum FP)(\sum TP + \sum FN)\,(\sum TN + \sum FP)(\sum TN + \sum FN)}} \quad [100]$$

The ROC curve was established in Microsoft Excel for Microsoft 365 MSO 64-bit using the confusion matrix and the previously calculated sensitivity and false positive rate (FPR).

To verify the results, a Partial least discriminant analysis (PLS-DA) was established in DataLab[101], Version 4.0, based on the raw data for branch lengths and node differences, topology, and branch colours. As the topology and branch colours belonged to a categorical class, they were converted to numerical data first (see Table 4.2).

5. Results

5.1 Positive controls

The 30 positive BGCs were restricted to clusters verified experimentally, only. Most of them were obtained from MIBiG. The following four positive control BGCs were obtained differently:

- Sorbicillin BGC, *T. reesei*
- Xanthocillin BGC, *A. fumigatus*
- Fumagillin BGC, *A. fumigatus*
- Lovastatin BGC, *A. terreus*

Sorbicillin BGC (Positive Control 27) was downloaded from GenBank®, Accession number GL985056.

Xanthocillin and fumagillin BGC were defined by their locus tags in literature.[102] [103] These locustags were searched in NCBI and the obtained results used for evaluating FunOrder. Both clusters were also found in the MIBiG repository.

A lovastatin BGC missing one of the core enzymes (*lovB*) could be found in GenBank®, Accession No. AH007774.[104] A sequence similarity search using BLAST was done for searching the whole cluster including *lovB* in the proteome of *A.terreus* strain NIH2426, for which the genome was the only one that was publicly available.[105] All cluster genes could be found in GenBank® separately. Another sequence similarity search using the single genes verified the correctness of the acquired BGC. The query coverage for almost all found genes resulted in 99-100%, except of ORF12, ORF8 and ORF16. ORF12 could not be found in the proteome of *A. terreus* strain NIH2426. The sequence similarity searches of ORF8 and ORF16 yielded a coverage of 95% and 83%, respectively. As the obtained BGC from *A. terreus* strain NIH2426 included the core enzyme *lovB*, it was used for evaluation.

The cyclosporine cluster from *Tolypocladium inflatum* (MIBiG repository BGC0000334)[106] found in the MIBiG database was redefined in the following as MIBiG cluster. This BGC had no congruence with the one described in the literature (Figure 5.2).[107]

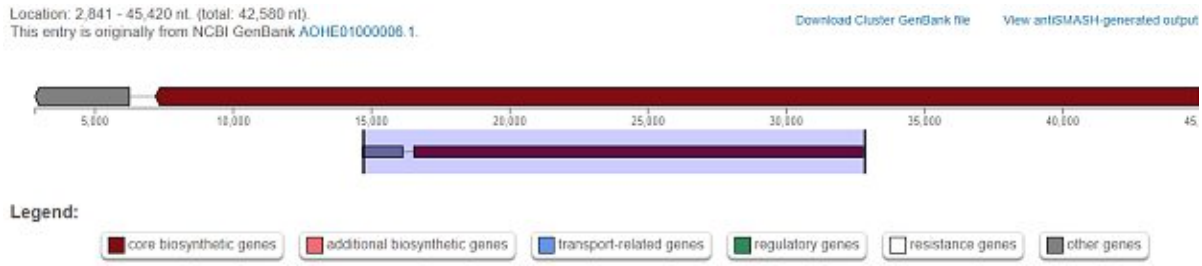


Figure 5.1: Cyclosporine cluster from Tolypocladium inflatum downloaded from the MIBiG repository BGC0000334[106], which was redefined as MIBiG cluster. The cluster contained only two sequences: one small gene in grey followed by a bigger one in red. The location defines the geographical coordinates of the cluster. The cluster is therefore located in the genome between 2841 and 45420 nucleotides (nt) and has a total length of 42560 nt.

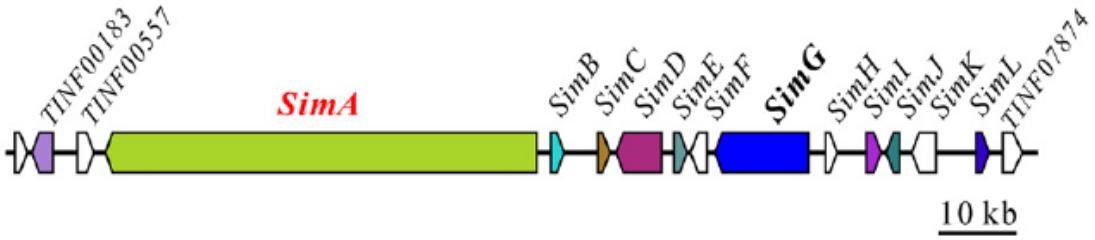


Figure 5.2: The structure of the cyclosporine biosynthetic cluster according to literature. The core enzyme is a NRPS defined as simA gene. All followed genes are renamed according to the core gene by Xiuqing Yang et al. According to literature the gene ID for simA is TINF00159, which could not be found in GenBank®.[107]

As shown in Figure 5.2 the cyclosporine cluster contained at least 12 genes (*simA* – *simL*) and 3 additional ones with the gene IDs TINF00183, TINF00557, TINF07874.[107] The MIBiG cluster however included only two genes (Figure 5.1). Next, the gene IDs stated in the publication of Xiuqing Yang, 2018, were searched in GenBank® but could not be found there.[107] Subsequently, the core gene *simA* was found in Uniprot[108] and its nucleotide sequence was downloaded from NCBI (Accession number AOHE01000194). Notably, the species *Tolypocladium inflatum* was called differently in various papers and databases (*Tolypocladium niveum*, *Beauveria nivea*, *Trichoderma polysporum*).

To review if the MIBiG cluster (Figure 5.1) was part of the cyclosporine cluster, the received gene *simA* was globally aligned with the MIBiG cluster via EMBOSS Stretcher online.The result suggested that the two genes from the MIBiG cluster were only a part of the *simA* gene that was found in Uniprot and therefore do not include the whole BGC. Consequently, the MIBiG cluster was not used for evaluation.

A second cyclosporine cluster from *Beauveria felina* was found and downloaded from the MIBiG repository (Figure 5.3). Notably, the number of genes in this BGC were in line with the research of Xiuyang Yang *et* al.[107]

Figure 5.3: Cyclosporine biosynthetic gene cluster from Beauveria felina. The cluster was obtained from MIBiG (BGC0001565) and contained 12 genes. Although Xiuyang Yang et al. investigated the cyclosporine cluster in Tolypocladium inflatum (also called Beauveria nivea), the number of genes in the BGC from Beauveria felina were in line with the literature.[109] *The location defines the geographical coordinates of the cluster. The cluster is therefore located in the genome between 1 and 90396 nucleotides (nt) and has a total length of 90396 nt.*

The new cyclosporine cluster from *B. felina* (Figure 5.3) was globally aligned using EMBOSS Stretcher online with *simA* gene from *T. inflatum* and with the MIBiG cluster from *T. inflatum*, respectively. The results verified that *simA* was part of the BGC from *B. felina*.

The BGC from *B. felina* was subsequently used for a sequence similarity search in the genome of *T. inflatum* yielding a cyclosporine cluster sequence, which was redefined as GenBank cluster. A global alignment via EMBOSS Stretcher online suggested that the *simA* gene was part of the BGC from *T. inflatum* as well. However, since this sequence was not annotated and the genes could not be obtained separately, the GenBank cluster could not be used for evaluation.

The previously stated results led to the usage of the cyclosporine BGC from *B. felina* for evaluation, whereas the clusters from *T. inflatum* were dismissed.

The finally utilized positive control BGCs and their source entries are listed in Table 5.1.

Table 5.1: Source entries of all used positive controls. "No" referred to the characterized number of the respective biosynthetic gene cluster (BGC) used for this thesis. For each BGC an abbreviation was further defined listed in "BGC abbr.". Source entries either corresponded to the accession number in the MIBiG repository (starting with "BGC") or in the GenBank® database. The Lovastatin BGC was received by a BLAST analysis and therefore no source entry is available.

No	BGC	BGC abbr.	Source entry
1	Tetramic acid	tas	KP8352.02
2	Mycophenolic acid	mpa	BGC0000104
3	Mycophenolic acid	mpa2	BGC0001360
4	Mycophenolic acid	mpa3	BGC0001677
5	Botrydial	bot	BGC0000631
6	Leporin B	lep	BGC0001445
7	Fumitremorgin	ftm	BGC0000356
8	Tenellin	ten	BGC0001049
9	Ilicicolin H	ili	BGC0002035
10	2-Pyridon-Desmethylbassianin	dmb	BGC0001136
11	Xanthocillin	xan	BGC0001990
12	Fumagillin	fma	BGC0001067
13	Terrein	ter	BGC0000161
14	Pneumocandin	pne	BGC0001035
15	Fumonisin	fum1	BGC0000063
16	Destruxin	dtxs	BGC0000337
17	Paxillin	pax	BGC0001082
18	Aflatoxin	afl	BGC0000008
19	Pestheic acid	pes	BGC0000121
20	Cephalosporin	cep	BGC0000317
21	Penicillin	pen1	BGC0000404
22	Penicillin	pen2	EF601124.1
23	Sorbicillin	sor1	BGC0001404
24	Ustiloxin B	ust	NW_002477245
25	Lovastatin	lov	-
26	Compactin	com	BGC0000039
27	Sorbicillin	sor2	GL985056
28	Fumonisin	fum2	BGC0000062
29	Cyclosporine C	cyc	BGC0001565

Each positive control BGC was examined and genes necessary for the biosynthetic pathway according to literature were specified and in the following redefined as necessary genes.[34 84 102-105 107 110-173] All used positive controls with their necessary genes can be found in Table 5.2.

.

Table 5.2: List of all used positive controls, including their organisms, core enzymes, represented as multimodular enzymes and enzympe type. Furthermore, empirically verified genes necessary for the secondary metabolite production, represented as "necessary genes", are listed.

No	BGC	Organism	Multimodular Enzyme	Enzyme type	Necessary Genes
1	Tetramic acid	*Hapsidospora irregularis*	tasS	PKS/NRPS	tasS, tasC, tasA
2	Mycophenolic acid	*Penicillium brevicompactum*	mpaC	NRPKS	mpaC, mpaD, mpaE, mpaA, mpaB, mpaG, mpaH, mpaF
3	Mycophenolic acid	*Penicillium roqueforti*	mpaC	NRPKS	mpaC, mpaDE, mpaA, mpaB, mpaG, mpaH, mpaF
4	Mycophenolic acid	*Penicillium roqueforti*	mpaC	NRPKS	mpaC, mpaDE, mpaA, mpaB, mpaG, mpaH, mpaF
5	Botrydial	*Botrytis cinera*	bot2	Terpen	bot1, bot2, bot6
6	Leporin B	*Aspergillus flavus*	lepA	PKS/NRPS	lepA, lepD, lepE, lepF, lepG, lepH
7	Fumitremorgin	*Aspergillus fumigatus*	ftmA	NRPS	ftmA, ftmB, ftmC, ftmD, ftmE, ftmF, ftmG, ftmH
8	Tenellin	*Beauveria bassiana*	tenS	PKS/NRPS	tenS, tenC, tenA, tenB
9	Ilicicolin H	*Neonectria sp. DH2*	iliA	PKS/NRPS	iliA, iliB, iliC, iliD
10	2 Pyridon Desmethyl-bassianin	*Beauveria bassiana*	dmbS	PKS/NRPS	dmbS, dmbC
11	Xanthocillin	*Aspergillus fumigatus*	xanB	ICS	xanA, xanB, xanC, xanE, xanF, xanG
12	Fumagillin	*Aspergillus fumigatus*	370	PKS	Af370, Af380, Af390, Af520
13	Terrein	*Aspergillus terreus*	terA	NRPKS	terA, terB, terC, terD, terE, terF, terR
14	Pneumocandin	*Glarea iozoyensis*	GLPKS4	PKS	GL10032, GL10034, GL10035, GL10043, GL10044
			GLNRPS4	NRPS	
15	Fumonisin	*Fusarium oxysporum*	fum1	PKS	Fum1, Fum21, Fum6, Fum7, Fum8, Fum3, Fum10, Fum11, Fum12, Fum13, Fum14
			fum10	ACS	
			fum14	NRPS	
			fum16	ACS	
16	Destruxin	*Metarhizium robertsii*	dtxS1	NRPS	DtxS1, DtxS2, DtxS3, DtxS4
17	Paxillin	*Penicillium paxilli*	paxG	Terpen	PaxB, PaxC, PaxD, PaxD, PaxM, PaxP, PaxQ
			paxD	Terpen	

18	Aflatoxin	*Aspergillus flavus*	pksA	PKS	AflR, Fas1, Fas2, PksA, CypA, NorB, Nor1, AvnA, NorA, AdhA, CypX, AvfA, EstA, VBS, VerB, Ver1, MoxY, OmtB, OmtA, OrdA, OrdB
19	Pestheic acid	*Pestalotiopsis fici*	ptaA	NRPKS	PtaA, PtaB, PtaC, PtaE, PtaF, PtaH, PtaI, PtaJ, PtaM
20	Cephalosporine	*Acremonium chrysogenum*	pcbAB	ACVS	pcbAB, pcbC, cefD1, cefD2, cefEF, cefG
			cefD1	IPNS	
21	Penicillin	*Penicillium chrysogenum*	pcbAB	ACVS	pcbAB, penDE, pcbC
22	Penicillin	*Penicillium chrysogenum*	pcbAB	ACVS	pcbAB, penDE, pcbC
23	Sorbicillin	*Penicillium rubens*	sorA, sorB	NRPKS, NRPKS	sorA, sorB, sorC, sorD, sorR1, sorR2, sorT
24	Ustiloxin B	*Aspergillus flavus*	ustC	RiPPs pathway	ustO, ustF1, ustC, ustU, ustA, ustYa, ustP1, ustP2, ustYb, ustH, ustD, ustF2, ustQ, ustT, ustR1, ustR2, ustM
			ustYa		
			ustP2		
			ustH		
25	Lovastatin	*Aspergillus terreus*	lovB	LNKS	lovF, lovD, lovG
			lovF	LDKS	
26	Compactin	*Penicillium citrinum*	mlcA	NKS	mlcD, mlcG, mlcH
			mlcB	DKS	
27	Sorbicillin	*Trichoderma reesei*	sor1	PKS	sor1, sor2, sor3, sor4, ypr1
			sor2	PKS	
28	Fumonisin	*Fusarium verticillioides*	fum1	PKS	fum1, fum21, fum6, fum7, fum8, fum3, fum10, fum11, fum12, fum13, fum14
			fum10	ACS	
			fum14	NRPS	
			fum16	ACS	
29	Cyclosporine C	*Beauveria felina*	simA	CsA	simA, simB, simC, simD, simG, simI, simJ, simL

Seven controls contained more genes and therefore more trees were acquired than expected according to their respective literature.[34 120 121 122 123 125 132 133 139 148 153 157 168 174] In four of the analysed clusters, core genes were missing (Table 5.3).

FunOrder was fed with all positive control BGCs and for each gene within a cluster a phylogenetic tree was inferred. Four of the analysed clusters were missing phylogenetic trees after running FunOrder (Table 5.3).

Table 5.3: Additional and missing genes in gbk-files compared to literature and missing trees after running FunOrder. Genes marked with an asterisk (*) are core genes according to their respective literature. "BGC No" refers to the characterized number of the respective biosynthetic gene cluster.[34 107 110 111 113 119 120 121 122 123 125 126 127 128 132 133 134 139 148 153 154 157 160 165 168 170 174]

BGC No	Additional genes	Missing genes in file	Missing trees
5	bot6, bot7	-	-
6	gen1, gen2	-	gen2
8	ten1	-	-
9	gene1	-	-
13	-	-	terI
14	GL100 20 - GL10029, GL10046 - GL10050	-	GL10025, GL10046, GL10048
15	-	fum20	-
18	-	cypA*, norB*	-
20	-	cefEF*	-
21	-	orf1	-
22	-	orf20c	-
24	AFLA94900, AFLA94910, AFLA94920	-	ustU, ustP1*
26	-	orf12	-
27	orf118 - orf124, orf126, orf131	sor3*	orf126, orf131
28	-	orf20	-
29	-	simC*, simL*	-

In three cases (BGC No 6, 14 and 27 in Table 5.3) the missing phylogenetic trees resulted due to additional genes, e.g., *gen2* for the Leporin B BGC (BGC No 6 in Table 5.3). Furthermore, in the Ustiloxin BGC (BGC No 25 in Table 5.3) the genes *ustU* and *ustP1* did not result in phylogenetic trees. Notably, both genes have only very small sequences compared to all other genes:

ustU: MRWRGRMEFKTRGATVWRDGPLTLALRRLAMTSSVVICSHWPRVTCELKINLAPVWEDSCCLLLCAELLMEGRLLGAQFNSASSQTCLYLIDG

ustP1: MGFSWYGVLLFVQLISSTIVYASDPCAQIDHYVAWGKKQGRNKISGIPGHLAYDVSSMPFRSDLAVKL

After receiving the phylogenetic trees, for each positive control the core genes and multimodular enzymes were defined and their trees were compared with the online visualization tool phylo.io.[96] The tree comparisons were performed as described in chapter 4.4. Tree comparison. The four parameters (branch length differences, node differences, topology, and colour) were used to calculate an average value of the comparison of two genes within a cluster, called manual evaluation measure (MEM). As the clusters contained more than two genes, the MEMs within a cluster were averaged yielding an average manual evaluation measure (aMEM). Because of the high number of genes and species-to-query-similarities within two trees, not all genes within a cluster could be compared. To receive a meaningful evaluation the previously defined necessary genes needed for secondary metabolite production were compared with each other. To further reduce the number of comparisons, another declaration was made: If a biosynthetic pathway was suggested or defined in

literature, received trees of associated genes were compared to trees of these genes that were followed by them according to the pathway, only. However, backbone producing enzymes, like PKS and NRPS, were compared to all genes within a cluster.

This procedure resulted in a high amount of missing data in nearly all clusters (Table 5.4), despite the great number of comparisons made (910 of 2588).

Table 5.4: Obtained tree numbers, done comparisons compared to total number of potential comparisons, average manual evaluation measure for positive genes due to literature (aMEM core) and for all compared genes (aMEM total) in the positive controls, respectively. "No" refers to the characterized number of the respective biosynthetic gene cluster (BGC) used for this thesis. In the column labelled "BGC" the abbreviations of the clusters according to Figure 5.1 were used.

Positive Controls

No	BGC	Trees	Possible comparisons	Done Comparisons	Missing Comparisons	aMEM core	aMEM total
1	tas	8	28	9	19	2.52	2.47
2	mpa	8	28	19	9	2.15	2.16
3	mpa2	7	21	16	5	2.00	2.02
4	mpa3	7	21	16	5	2.35	2.25
5	bot	7	21	11	10	2.28	2.36
6	lep	10	45	16	29	2.35	2.28
7	ftm	9	36	24	12	1.63	1.78
8	ten	5	10	6	4	2.46	2.11
9	ili	6	15	7	8	2.38	2.25
10	dmb	4	6	6	0	2.40	2.35
11	xan	7	21	14	7	1.91	1.90
12	fma	15	105	30	75	1.66	1.90
13	ter	10	45	24	21	1.62	1.71
14	pne	28	378	102	276	1.57	1.51
15	fum1	17	136	65	71	2.25	2.30
16	dtxs	21	210	57	153	2.70	2.03
17	pax	8	28	22	6	2.24	1.90
18	afl	24	276	60	216	2.06	2.09
19	pes	18	153	31	122	1.91	1.87
20	cep	7	21	12	9	2.27	2.24
21	pen1	15	105	15	90	2.53	2.13
22	pen2	15	105	15	90	2.46	2.14
23	sor1	7	21	14	7	2.35	2.28
24	ust	19	171	99	72	2.21	1.90
25	lov	17	136	33	103	2.28	1.80
26	com	9	36	26	10	1.86	2.00
27	sor2	13	78	35	43	2.12	2.26
28	fum2	23	253	89	164	2.21	2.20
29	cyc	13	78	35	43	1.77	1.51

The manual evaluation measures (MEMs) are listed in the supplement (Supplement 9.3 - Supplement 9.31). As previously defined, the MEMs were averaged for each cluster, yielding the so-called average manual evaluation measure (aMEM, see Table 5.4). The MEMs and aMEMs were applied to further statistical evaluation, like heatmaps and dendrograms, as well as a manually created confusion matrix and a receiver operating characteristic (ROC) curve. The latter two were both compared to a partial least squares discriminant analysis (PLS DA) done using the raw data. Because of the missing

data a higher threshold of the aMEM for the positive controls (2.0) than for the negative controls (1.5) was used for the confusion matrix and the ROC curve.

5.2 Negative controls

The 42 synthetic BGCs analysed with FunOrder did not output any phylogenetic trees. They were therefore not included in the statistical evaluation.

Phylogenetic trees for each of the 60 random BGC were obtained, which were therefore used as negative controls for further analysis. The comparisons of the phylogenetic trees were performed as described in chapter 4.4. Tree comparison. The MEMs are listed in the supplement (Supplement 9.32). All negative control BGCs with their respective number of phylogenetic trees, completed comparisons and aMEMs are shown in Table 5.5.

Table 5.5: Obtained tree numbers, done comparisons, and calculated average manual evaluation measure (aMEM) for each random control. "No" refers to the characterized number of the respective biosynthetic gene cluster used for this thesis.

Random Cluster			
No	**Trees**	**Comparisons**	**aMEM**
1	5	10	0.4
2	4	6	1.2
3	7	21	1.3
4	4	6	0.5
5	4	6	0.9
6	3	3	1.2
7	4	6	0.8
8	3	3	1.4
9	7	21	1.2
10	3	3	0.8
11	7	21	1.0
12	5	10	1.0
13	5	10	0.4
14	6	15	0.4
15	3	3	0.5
16	3	3	0.8
17	4	6	0.9
18	7	21	0.9
19	6	15	0.9
20	5	10	0.7
21	3	3	1.1
22	5	10	1.1
23	4	6	1.2
24	7	21	0.9
25	8	28	1.0
26	7	21	0.6
27	5	10	1.0
28	6	15	0.8
29	7	21	1.0
30	3	3	2.1
31	8	28	0.8
32	3	3	0.5
33	5	10	0.7

34	5	10	0.5
35	5	10	0.7
36	5	10	1.1
37	4	6	0.5
38	4	6	1.0
39	5	10	1.1
40	6	15	0.4
41	5	10	1.2
42	3	3	0.5
43	4	6	1.0
44	7	21	0.6
45	7	21	0.7
46	8	28	1.1
47	5	10	0.6
48	5	10	0.9
49	5	10	1.3
50	4	6	1.5
51	6	15	0.7
52	6	15	0.7
53	5	10	0.7
54	5	10	0.7
55	5	10	0.7
56	7	21	0.9
57	8	28	0.9
58	4	6	1.3
59	5	10	0.7
60	2	1	1.5

As shown in Table 5.5, most of the aMEMs were lower than 1.5. A threshold of 1.5 was therefore used for the confusion matrix and the receiver operating characteristic (ROC) curve.

5.4 Statistical Evaluation

5.4.1 Manual evaluation measure (MEM)

5.4.1.2 Positive controls

The manual evaluation measures (MEMs) for all positive controls were calculated according to chapter 4.4. Tree comparison and with them matrices were assembled (Supplement 9.3 - Supplement 9.31). These matrices were used to visualize the data with heatmaps, dendrograms and for establishing a principal component analysis (PCA). As the MEMs represented the similarity between two phylogenetic trees of genes within a biosynthetic gene cluster (BGC), the evaluation was done to analyse whether there was a similar coevolution between cluster genes. Due to the missing comparisons, a Random forest algorithm was used to infer the missing data points in the matrices from the analysed comparisons to enable the calculation of PCAs. The approximated data were also used for the computation of the dendrograms and heatmaps. Before analysis, the data was scaled, and two different wards approaches were applied yielding four divergent datasets. The dendrograms were then calculated for all parameters (without and with missing data, scaled and unscaled data as well as both wards approaches). In total, 58 heatmaps, 232 dendrograms and 29 PCAs of the positive controls were calculated.

The differently computed dendrograms were compared with each other and no differences between the two ward's approaches could be determined. Furthermore, no differences in the clustering between scaled and unscaled dendrograms were apparent. Discrepancies were only visible for dendrograms that were approximated by the random forest algorithm when comparing with the ones with missing data (Supplement 9.33 - Supplement 9.61). However, as 28 of 29 matrices included more than 20% missing data, the PCA results were in question. Thus, the PCA results were critically discussed and a higher focus on the heatmaps and dendrograms without approximated data were set during evaluation.

In the following subchapters the MEMs and their statistical analysis of eleven of the 29 positive controls were introduced in more detail. Contrary to the estimation, eight positive controls exhibited low average manual evaluation measures (aMEM). They were therefore also presented in the following. The plots of the remaining controls were listed in the supplement (Supplement 9.33 - Supplement 9.61).

The manual evaluation of FunOrder was used to range the distances calculated by FunOrder using the TreeKO algorithm. To review if the manual evaluation made sense, an additional BGC (fusaric acid BGC from *Fusarium fujikoroi*) was added and automatically analysed by FunOrder itself. This time the distances calculated by FunOrder were used for the evaluation of the biosynthetic gene cluster. The results were attached in the supplement (Supplement 9.63).

1. **Mycophenolic acid BGC from *Penicillium roqueforti* and *Penicillium brevicompactum* (mpa)**

Mycophenolic acid is a natural antibiotic, which is also used as immunosuppressive drug for organ transplantations and autoimmune diseases.[173] It inhibits the production of inosine-5'-monophosphate dehydrogenase (IMPDH), which is a rate-controlling enzyme in the guanosine-5'-monophosphate (GMP) biosynthesis that converts inosine-5'-monophosphate (IMP) to xanthosine-5'-monophosphate (XMP). This is an important reaction in almost all living organisms. Hence, the producing organism will need a resistance gene against mycophenolic acid, which was found to be produced by the gene *mpaF*.[156]

Three different Mycophenolic acid BGCs were analysed. The first two BGCs were derived from the species *Penicillium roqueforti* (Positive control 03 and positive control 04) and the third from *Penicillium brevicompactum* (Positive control 02). These three BGCs were downloaded from the MIBiG repository and served as a first exemplary positive control analysis. The main difference in the clusters were that *P. brevicompactum* had two single genes for *mpaD* and *mpaE*, while in *P. roqueforti* the genes for *mpaD* und *mpaE* came together as *mpaDE*.

The illustrated biosynthetic pathway for *P. brevicompactum* in Figure 5.4 was used as a basis for the analysis of the three Mycophenolic acid BGCs.

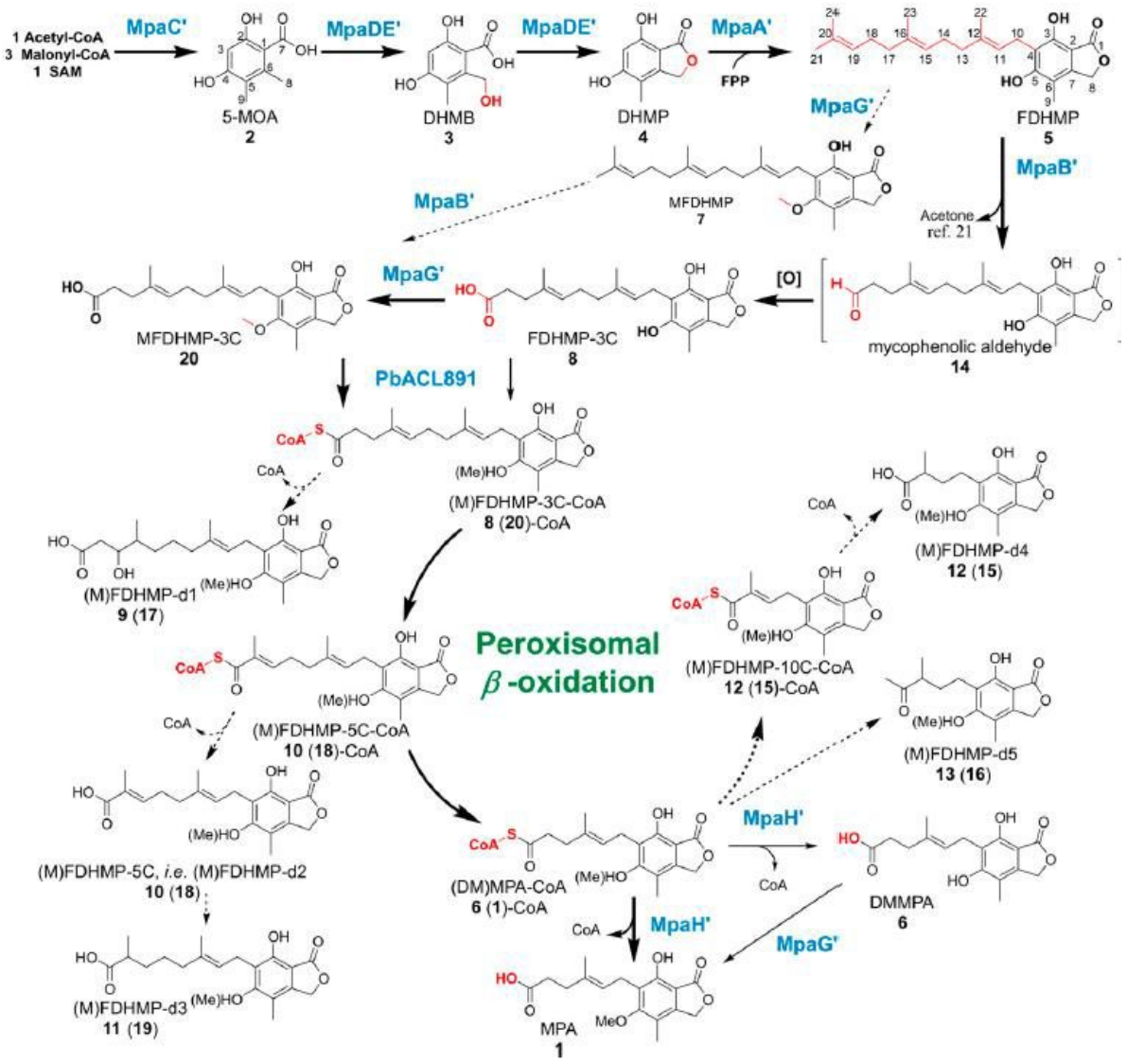


Figure 5.4: Biosynthetic pathway of the production of mycophenolic acid (MPA, 1) in P. brevicompactum according to Zhang et al., 2019.[173] The gene mpaC uses Acetyl-CoA, Malonyl-CoA and S-adenosyl-L-methionine (SAM) to produce the first intermediate 5-methylorsellinic acid (5-MOA, 2) which is then converted to 3,5-dihydroxy-7-(hydroxymethyl)-6-methylbenzoic acid (DHMB, 3) by the C8 hydroxylation activity of mpaD. Subsequently, 3,5-dihydroxy-6-methylphtalide (DHMP) is produced by mpaE. According to literature the products of mpaD (a cytochrome P450 domain) and mpaE (a hydrolase domain) are fused together and represented as mpaDE'. The next steps in the pathway are proposed only. DHMP is farnesylated using farnesyl pyrophosphate (FPP) and the gene mpaA to 4-farnesyl-3,5-dihydroxy-6-methylphtalide (FDHMP, 5). According to Zhang et al. (2019), all further steps until the production of demethylmcyophenolic acid (DMMPA, 6) are speculated, e.g., the production of 5-O-methyl-FDHMP (MFDHMP, 7). The last step includes the O-methylation of DMMPA by mpaG yielding the final product (MPA).[173]

Genes, which follow the previous one in the pathway, were compared to each other. For example, the gene *mpaA* was checked against *mpaDE (or mpaD and mpaE, respectively)*, *mpaG* and *mpaB*. As *mpaF* was identified to be a self-resistance gene[135], it was compared to all genes in the cluster. The core gene *mpaC* produces a polyketide synthase (PKS).[135] Like *mpaF*, *mpaC* was also compared to all genes in the cluster.

The resulting MEMs were assembled to matrices and heatmaps with and without missing data were calculated to show the evolutionary linkage between biosynthetic cluster genes. (Figure 5.5)

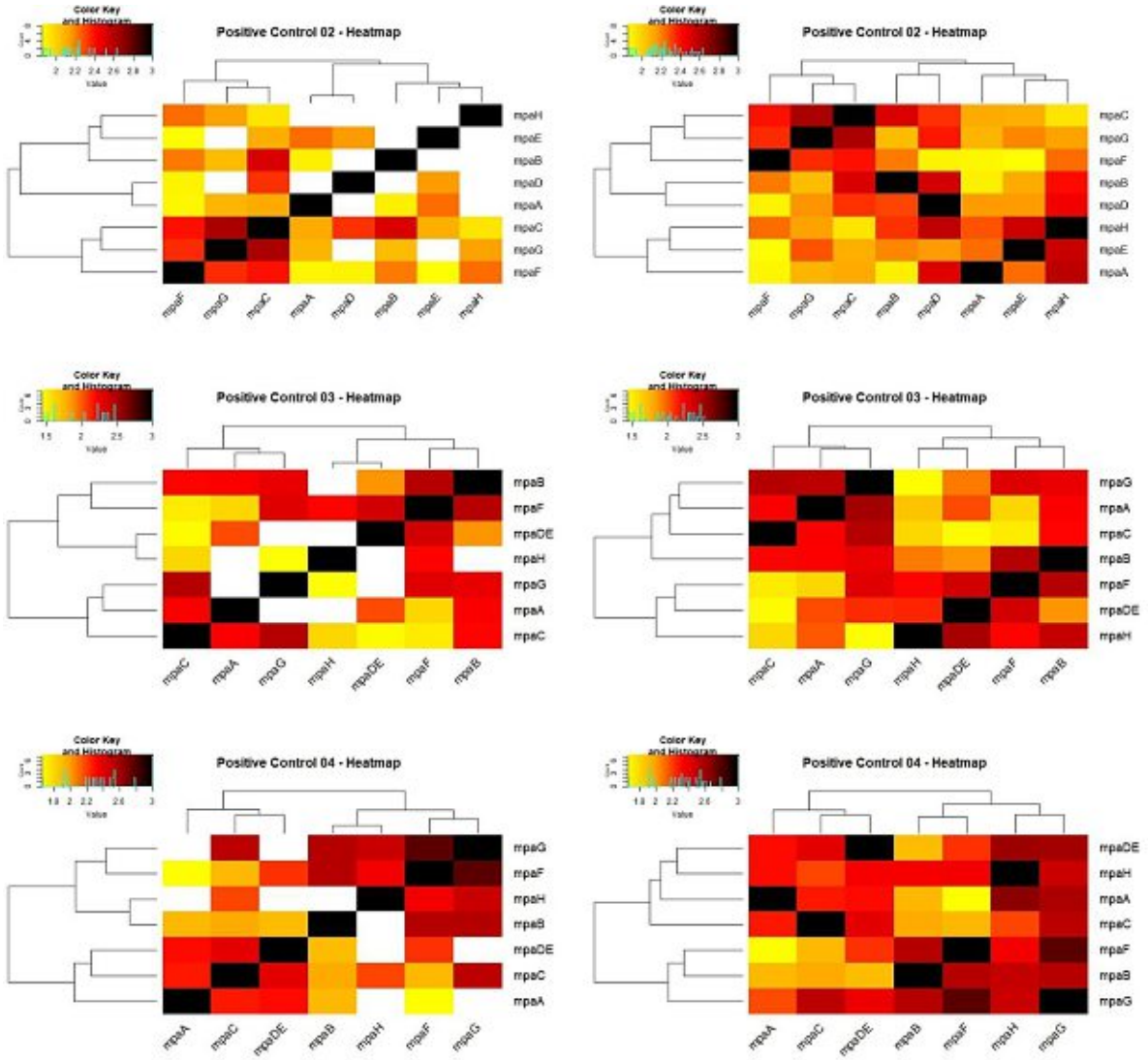


Figure 5.5: Heatmaps of the manual evaluation measures (MEMs) with missing data (left) and without (right) for control 02 (top), control 03 (middle) and control 04 (bottom). The missing data were approximated by a random forest approach (MissForest package in RStudio), yielding complete matrices and therefore complete heatmaps. The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes. Positive control 02 referred to the cluster from P. brevicompactum (MIBiG BGC0000104), control 03 and 04 derived from P. roqueforti. While control 03 referred to MIBiG BGC0001360, control 04 referred to MIBiG BGC0001677.

According to literature, the gene *mpaG*, a S-adenosyl-L-methionine-dependent O-methyltransferase, catalyzed the last step of the synthesis of the final product, the mycophenolic acid (MPA).[172] As previously stated, mycophenolic acid is an antibiotic, for which the organism needed the resistance gene *mpaF*.[156] A co-evolutionary linkage between *mpaG* and *mpaF* therefore appeared to be appropriate. In fact, the calculated heatmaps (Figure 5.5) showed that the genes *mpaG* and *mpaF* cluster together, especially for the controls 02 and 04, indicating that the genes share evolutionary traits. In the organism *P. brevicompactum* the core enzyme *mpaC* clustered together with *mpaG*, too, suggesting a co-evolutionary linkage and supporting the previously stated importance of *mpaG* in the biosynthesis.

Additionally to the heatmaps, dendrograms were computed using two different Ward's minimum variance approaches: ward.D2, which implemented the Ward's clustering criterion from 1963, and

ward.D, which did not implement that criterion. Comparing the dendrograms generated with ward.D and ward.D2, no differences were apparent.

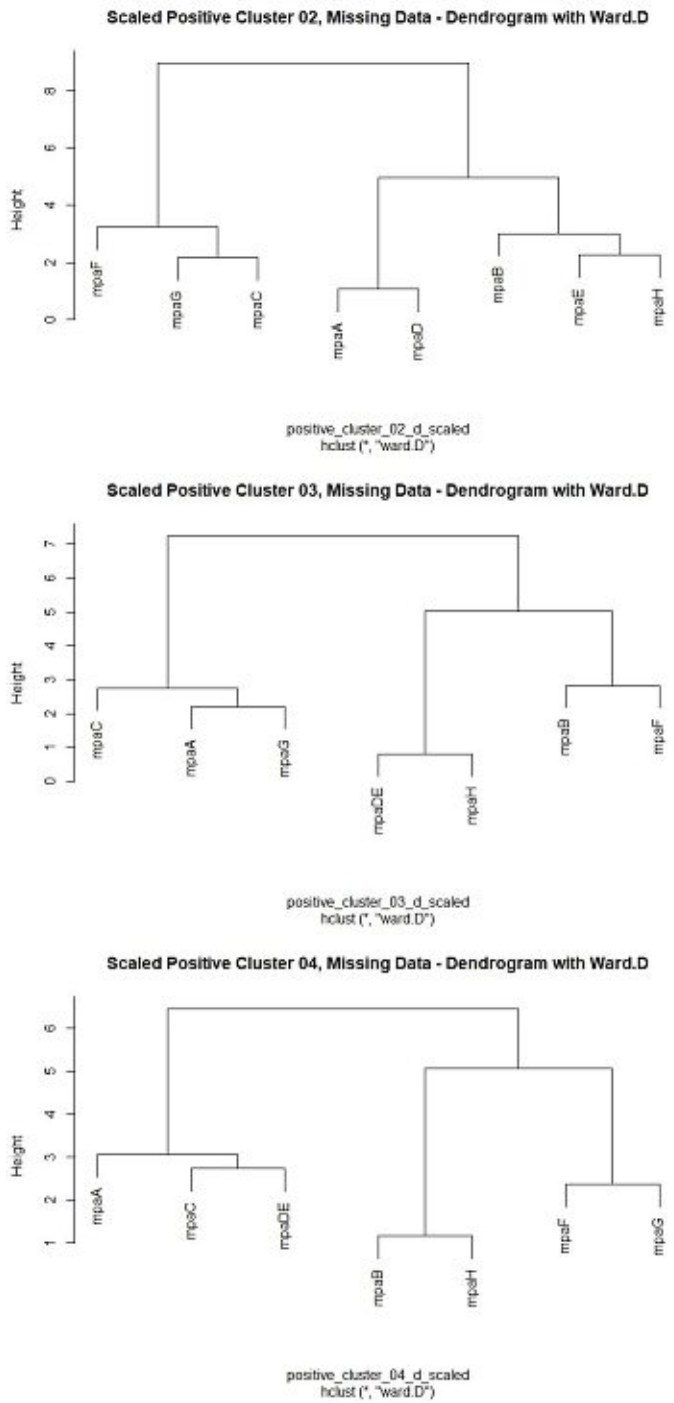


Figure 5.6: Scaled Dendrogram with missing data of positive controls 02, 03 and 04 (mpa BGC). The former conducted matrices were scaled and a distance matrix were returned using the functions scale and dist in RStudio. The illustrated dendrograms were computed using the hclust function and the ward clustering method ward.D in RStudio. The clustering of mpaF and mpaG suggested a coevolutionary linkage between them, which was expected, because mpaG catalyzes the production of mycophenolic acid (MPA) and mpaF is the selfresistance gene against MPA.

The dendrograms in Figure 5.6 were computed by Ward's minimum variance (ward.D) using scaled data without approximation by MissForest algorithm. Similar to the heatmaps, the genes *mpaF* and *mpaG* cluster together for the controls 02 and 04, supporting the previously assumed evolutionary linkage between them. However, *mpaG* did not cluster with *mpaF*, but with *mpaC*, the core enzyme, and *mpaA* for the control 03, whereas the self-resistance gene *mpaF* clustered with *mpaB*. According

to literature, the gene *mpaB* is a membrane-associated oxygenase mediating the cleavage of the farnesyl side chain yielding mycophenolic aldehyde.[173] A coevolutionary linkage between *mpaB* and *mpaF* could not be shown in the other two controls though. Notably, control 02 and control 04 derived from two different organisms, whereas control 03 and control 04 both originated from *P. roqueforti*. However, the presumption of resemblances in the clustering patterns between the controls 03 and 04 could not be confirmed, whereas there were similar patterns between the controls 02 and 04.

The scores of the principal component analysis (PCA) of all three mpa clusters were illustrated in Figure 5.7. To overcome the missing values in the matrices, the data were approximated by a random forest approach (see Figure 5.5) and these new data sets were used to perform PCA. As shown in Figure 5.5 subpart control 02, the genes *mpaC*, *mpaF* and *mpaG* clustered together, confirming the previous statements, and indicating a coevolutionary linkage. However, in the subparts control 03 and control 04 a likewise pattern was not apparent. The core gene *mpaC* clustered with *mpaDE* (control 04) or with *mpaA* and *mpaG* (control 03). The self-resistance gene *mpaF* diverged (control 03) or clustered with *mpaG* (control 04), similar to control 02. A discrepancy between control 03 and 04 could be already represented in the previously illustrated heatmaps and dendrograms. The BGCs derived from the same species (*P. roqueforti*) but different loci. Hence, the results indicated that the two BGCs (control 03 and control 4) diverged differently.

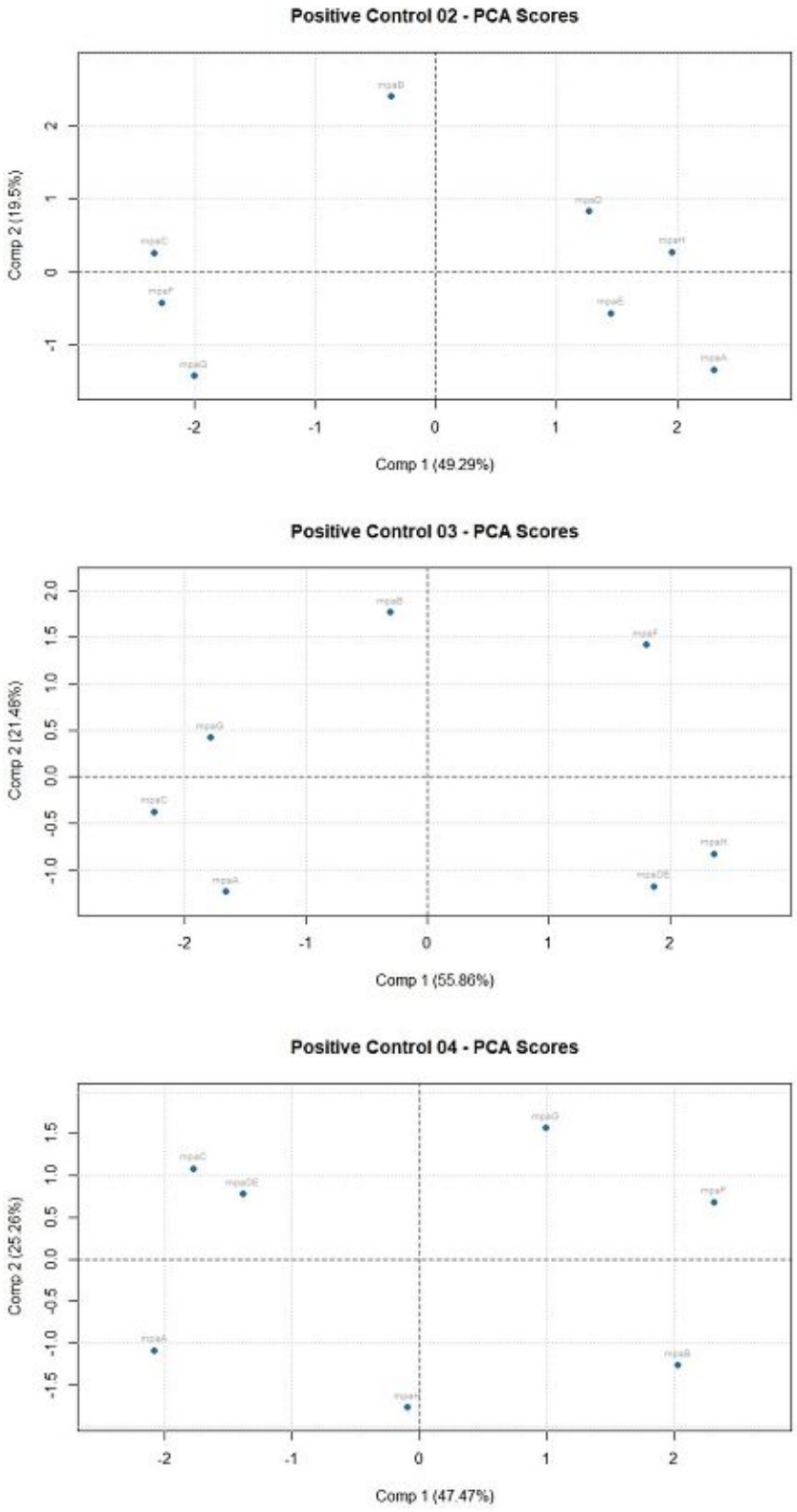


Figure 5.7: Scores of the calculated principal component analysis (PCA) of the mycophenolic acid BGCs. A PCA calculates linear combination of the original variables in multivariate data sets yielding principal components, whose scores were plotted. The clustering of the scores indicates a coevolutionary linkage. Hence, the plots suggested that the genes mpaF mpaC and mpaA in control 03, mpaC, mpaA and mpaG in control 03, and mpaC and mpaDE in control 4 shared evolutionary traits.

2. Tenellin BGC from *Beauveria bassiana* (ten) and 2-Pyridon-Desmethylbassianin BGC from *Beauveria bassiana* (dmb)

Tenellin (ten) and 2-Pyridon-Desmethylbassianin (dmb) are both yellow pigments found in and produced by *Beauveria bassiana*.[162] Heneghan *et al*. compared these two clusters and revealed that the structure of the BGCs is identical having a 90% identity value, and that the core genes *tenS* and *dmbS* are homologous (Figure 5.8).[140] Due to these similarities, the clusters ten and dmb were visualized and described together in the following.

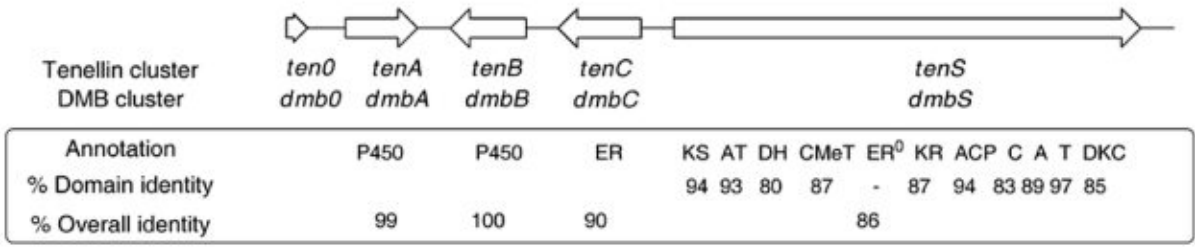


Figure 5.8: Comparison of tenellin biosynthetic gene cluster (BGC) and desmethylbassianin BGC, which produced orthologues enzymes according to Heneghan et al., 2011.[140] The arrows on the top illustrated the structure of the BGCs according to literature. The figure was extracted from Heneghan et al. (2011), figure 2, in which the domain sequences were compared and showed similar identities. Abbrevations: KS, β-ketoacyl synthase; AT, acyl transferase; DH, dehydratase; ER, enoyl reductase; ER^0, defective enoyl reductase; CMeT, C-methyl transferase; KR, β-ketoacyl reductase; ACP, acyl carrier protein; C, condensation; A, adenylation; T, thiolation; DKC, Dieckmann cyclase; P450, cytochrome P450 oxidase.[140]

Tenellin BGC (Positive control 08) contains a hybrid PKS-NRPS (*tenS*), the core enzyme, two cytochrome P450 oxidases (*tenA*, *tenB*) and a trans-acting enoyl reductase (*tenC*). The dmb BGC is producing orthologous proteins. Like *tenA* and *tenB*, *dmbA* and *dmbB* encode for cytochrome P450 oxidases. The orthologue gene of *tenC* is *dmbC* encoding a trans-acting enoyl reductase. The core enzyme of 2-Pyridon-Desmethylbassianin (Positive control 10), a hybrid PKS-NRPS, is produced by *dmbS*.[140] According to the proposed biosynthetic pathway of Tenellin BGC, *tenC* and the core enzyme *tenS* are producing the backbone pretenellin-A that is further used for synthesizing the final product by the genes *tenA* and *tenB*.[133 139 140] The production of 2-Pyridon-Desmethylbassianin is following a similar pathway.[129]

During analysis, the dmb BGC was completely evaluated, while the ten BGC included an additional gene (*ten1*), for which the only references was the figure of the comparison with the dmb BGC, marked as *ten0* (Figure 5.9). However, no further details about *ten0* were discussed in the respective literature.[140] Other papers assumed that the Tenellin BGC did not include this additional gene.[133 129] Furthermore, the extracted dmb BGC did not include a *dmb0* gene, either (Figure 5.9). The gene *ten1* was therefore assumed to be an additional gene dispensable for the biosynthetic pathway and was compared to the core enzyme, only.

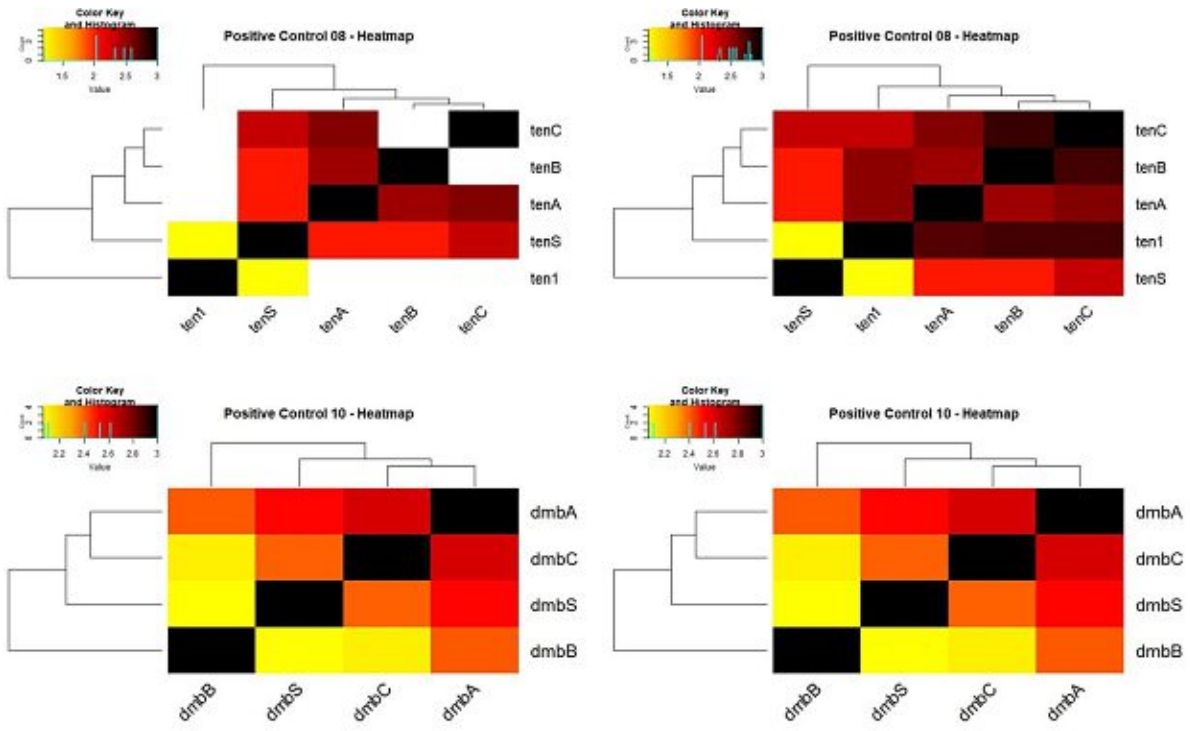


Figure 5.9: Heatmaps of the manual evaluation measures (MEMs) with missing data (left) and without (right) for control 08 (top) and control 10 (bottom). A random forest approach (MissForest package in RStudio), yielding complete matrices and therefore complete heatmaps approximated the missing data. The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes. The BGCs for tenellin (control 08) and desmethylbassianin (control 10) were extracted from MIBiG (BGC0001049 & BGC0001136). While control 10 was completely evaluated and no approximation was needed, control 08 included an additional gene (ten1) which was compared to the core gene tenS.

As expected, the MEM between the genes *ten1* and *tenS* was low, assuming that *ten1* did not share any evolutionary traits with the core enzyme *tenS* (Figure 5.9), while the rest of the evaluated genes of control 08 clustered together indicating a strong co-evolutionary linkage. Comparing the two heatmaps of control 08, the approximation of the missing data apparently overestimated the values of *ten1* indicating that *ten1* shared more evolutionary traits with the rest of the cluster than the core enzyme. As the gene *ten1* was compared to the core gene only, the approximation of its remaining values was arguable, particularly because *ten1* was assumed to be an additional gene. The examination of the heatmap containing the missing data showed that the genes *tenA* and *tenC* had a strong evolutionary linkage, similar to the genes *dmbA* and *dmbC* in the dmb BGC. Positive control 10 was completely evaluated and therefore no approximation was needed. The results for the dmb BGC could be therefore seen as reliable. Interestingly, the gene *dmbB,* encoding a cytochrome P450 oxidase like *dmbA*, hardly cluster with the rest of the dmb genes, implying that *dmbB* had less evolutionary linkages with the other genes.[140]

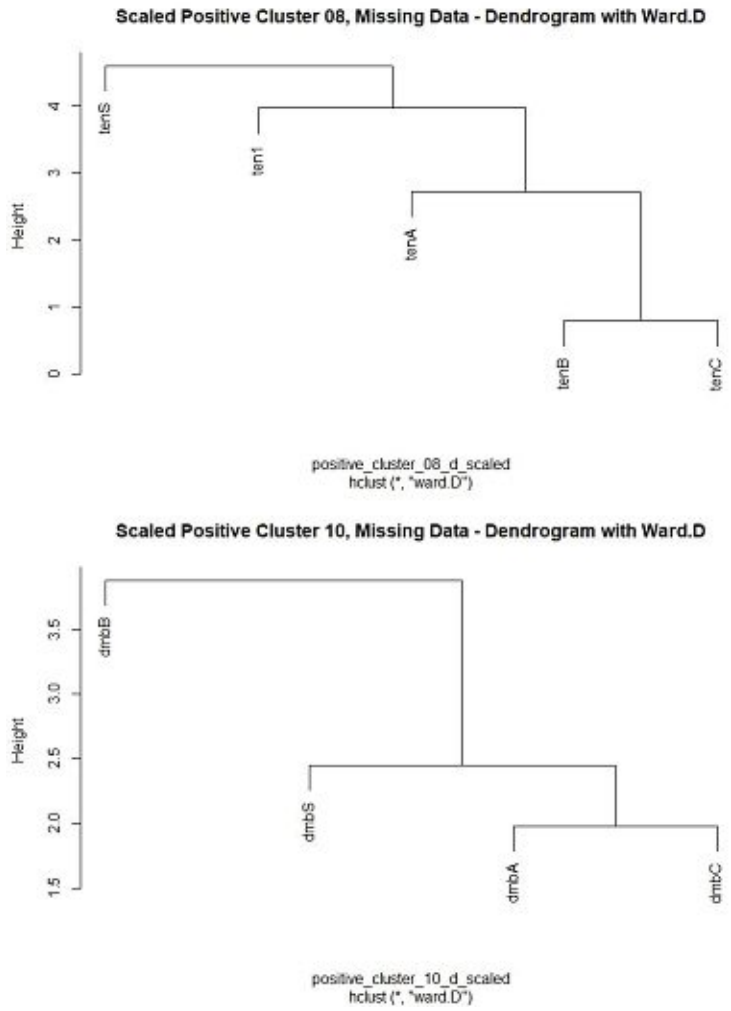


Figure 5.10: Scaled Dendrogram with missing data of positive controls 08 (tenellin BGC), and 10 (2-pyridon-desmethylbassianin BGC). The former conducted matrices were scaled and a distance matrix were returned using the functions scale and dist in RStudio. The illustrated dendrograms were computed using the hclust function and the ward clustering method ward.D in RStudio. The dendrogram suggested a coevolutionary linkage between tenB and tenC in control 08, and between dmbA and dmbC, confirming the results from the heatmaps in Figure 5.9.

The dendrogram of the Tenellin BGC showed similar results as the approximated heatmap, assuming that the additional gene *ten1* clustered more with the rest of the BGC than the core enzyme and that *tenB* and *tenC* were strongly coevolutionary linked (Figure 5.10). The dendrogram of control 10 confirmed the previously stated results from the heatmaps, indicating that *dmbA* and *dmbC* shared more evolutionary traits than the other genes.

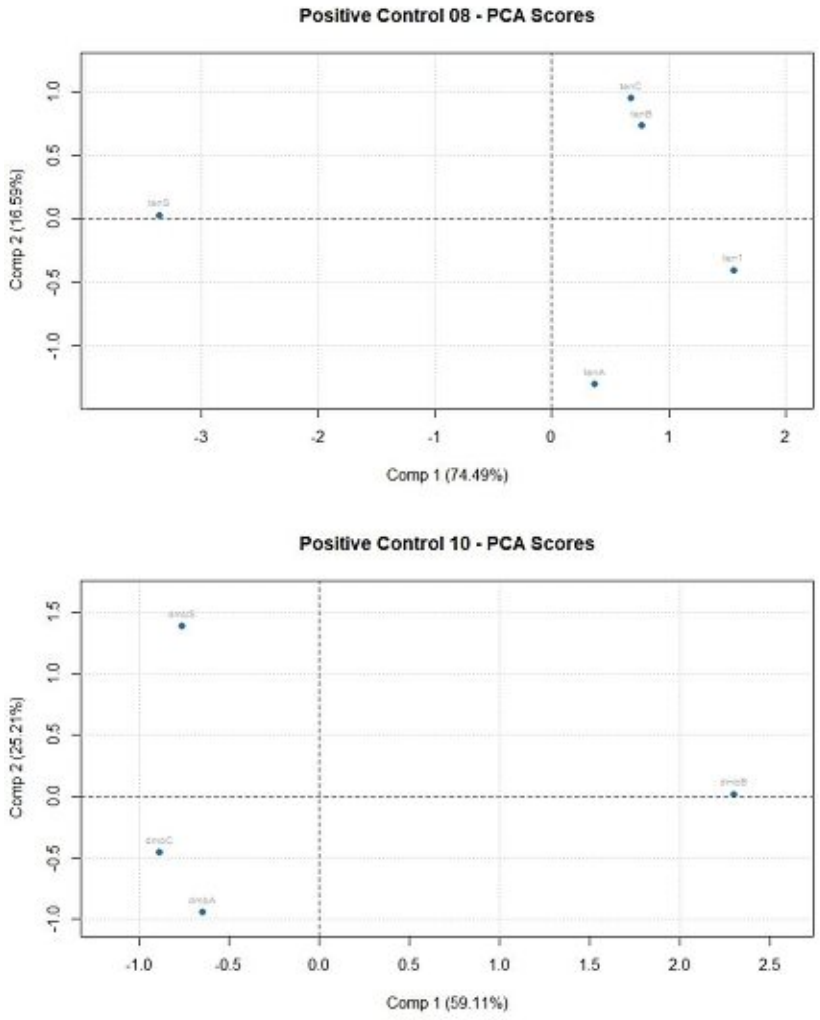


Figure 5.11: Scores of the calculated principal component analysis (PCA) of the tenellin BGC (control 08) and 2-pyridon-desmethylbassianin (control 10). A PCA calculates linear combination of the original variables in multivariate data sets yielding principal components, whose scores were plotted. The clustering of the scores indicates a coevolutionary linkage. Hence, the plots suggested that the genes tenC and tenB in control 08, dmbC and dmbA shared evolutionary traits.

The principal component analysis (PCA) was computed using the approximated data and its scores were illustrated in Figure 5.11. Like the heatmaps and the dendrograms they show an aggregation of *tenC* and *tenB* (Control 08) as well as of *dmbC* and *dmbA* (Control 10), while *tenS* and *dmbB* were shifting to different directions. As PCA used the approximated data, the analysis for the Tenellin BGC was in question. However, the scores confirmed the previous statements, especially for the dmb BGC, for which all MEMs were analysed and calculated.

3. Fumonisin BGC from *Fusarium oxysporum* and *Fusarium verticillioides* (fum)

Fumonisins belong to the sphinganine-analog mycotoxins (SAMT) and are common contaminants of corn and maize produced by the genus *Fusarium*.[126] Furthermore, they cause several diseases in animals which are associated to human oesophageal cancer and neural tube defects.[126] During this thesis, two Fumonisin BGCs were analysed. Positive control 28 derived from *Fusarium verticillioides* and positive control 15 derived from *Fusarium oxysporum*. According to Proctor *et* al. (2008) and their analysis the overall identity of the coding regions of the two BGCs lies in the range of 88-92% identity.[154]

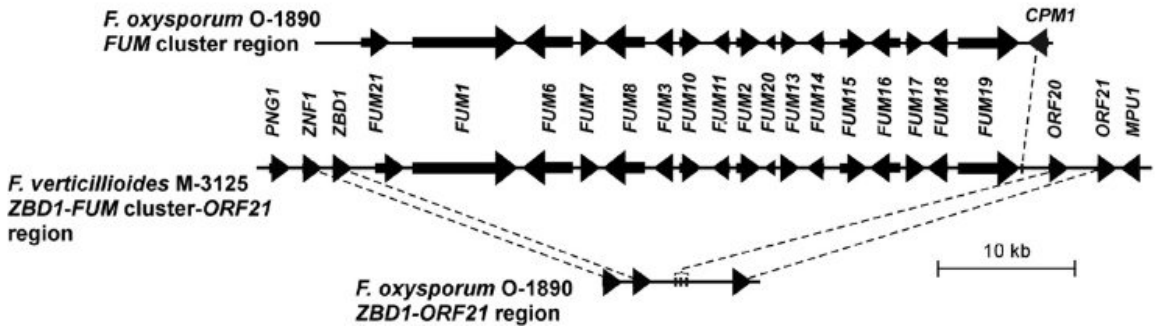


Figure 5.12: Comparison of the fumonisin (fum) BGC in F. oxysporum and in F. verticillioides obtained from Proctor et al, 2008.[154] The arrows illustrated the structure of the BGCs according to literature. According to Proctor et al. (2008), the fumonisin BGC in the M-3125 strain of F. verticilloides is located between the genes ZBD1 and ORF21. Hence, the genes before and after did not belong to the fum cluster.[154]

The genes *fum13*, *fum14*, *fum1*, and *fum10* were found to be core enzymes in the Fumonisin BGC and were therefore compared to all other genes in the cluster (Figure 5.12). Further gene comparisons were made according to the proposed biosynthetic pathway.[126] [119] As it had high similarity to acyl-CoA synthetases, *fum16* was analysed and compared with all other genes of the BGC as well. However, according to deletion studies in Butchko *et* al. (2006) the knockout of *fum16* did not apparently alter the fumonisin production.[119] *Fum10*, like *fum16*, had also a high degree of similarity to acyl-CoA synthetases, while the prediction of *fum14* was the encoding of a protein with high similarity to a NRPS.[119] However, the deletion of *fum10* and *fum14* resulted in an accumulation of two fumonisin derivatives.[119] Following the disruption of *fum1*, the production of fumonisin was reduced by 99%. *Fum1* was formerly referred to as fum5 and encodes a type I PKS.[155] The knockout of *fum8* resulted in no production of any fumonisin.[154] However, *fum8* was not compared to all genes in the cluster. Another gene that significantly reduced fumonisin production when knocked out was *fum13*, a hypothetical short chain dehydrogenase/reductase.[117]

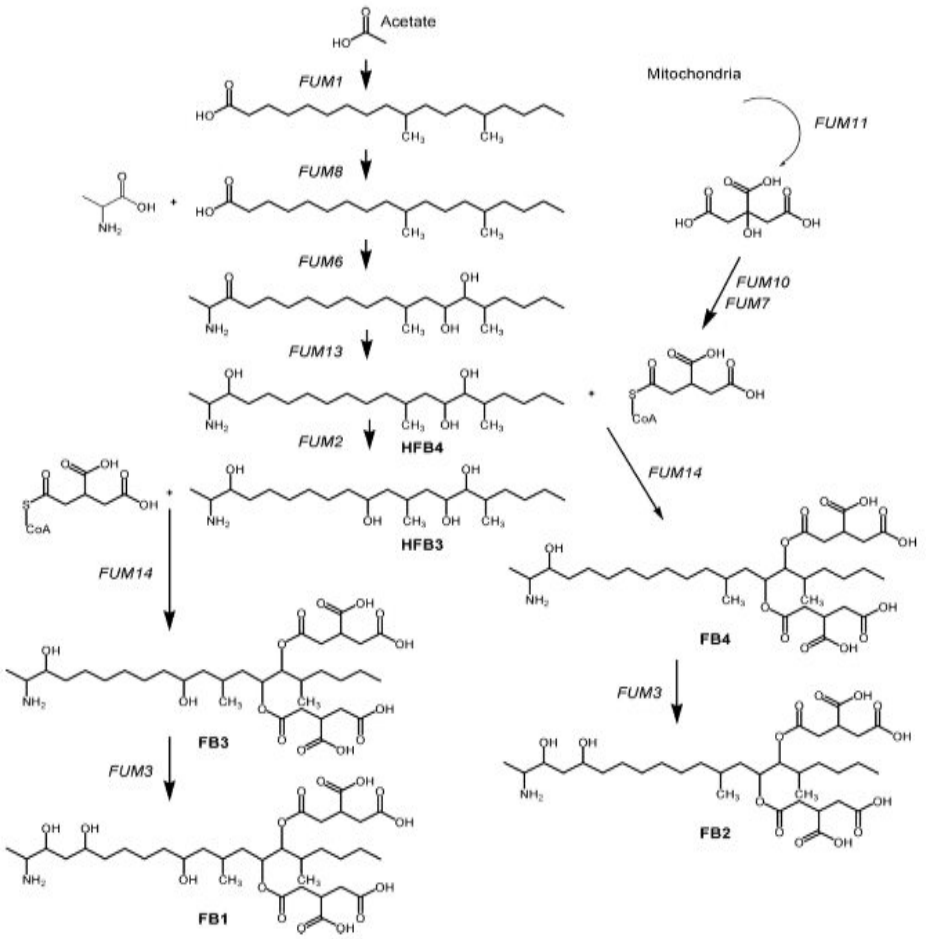


Figure 5.13: Proposed biosynthetic pathway of fumonisin BGC. According to Butchko et al. (2006), the first intermediate is produced by the gene fum1 using acetate as precursor molecule.[119] The intermediate is then converted to HFB4 (hydro-fumonisin B4) using the genes fum8, fum6, fum13 and fum2. The final products (Fumonisin B1, B2, B3 and B4 abbreviated as FB1, FB2, FB3 and FB4) are synthesized out of HFB4 and its derivative HFB3. To obtain FB2 and FB4, the gene fum11 makes a precursor available for biosynthesis which is then converted by fum10 and fum7.[119]

As previously explained, the analysis of the Fumonisin BGCs, too, were done using the biosynthetic pathway (Figure 5.13). In total, 10 out of 18 (Control 15) and 10 out of 23 (Control 28), respectively, were set as empirically verified genes necessary for the secondary metabolite production.

Heatmaps of both clusters, with and without missing data, are shown in Figure 5.14. As shown in Figure 5.14 subpart control 15, the genes *fum6* and *fum13,* as well as *fum1* and *fum8* clustered together, confirming their coevolutionary linkage according to their biosynthetic pathway. However, this could not be shown in the heatmap of control 28. According to the colour key of the histogram, the comparisons of *fum10* with the rest of the cluster genes yielded in high values, assuming the share of evolutionary traits between *fum10* and the rest of the BGC. Contrary, *cpm1* in control 15 yielded comparatively low values. In fact, according to Figure 5.14. *cpm1* appeared in control 15, only, and was not a part of the BGC in control 28.

When comparing the two heatmaps, resemblances in the clustering patterns of the genes *fum1*, *fum14* and *fum16* could be found, supporting a coevolutionary linkage between them. Further resemblances could be found between the genes *fum6*, *fum10* and *fum19*.

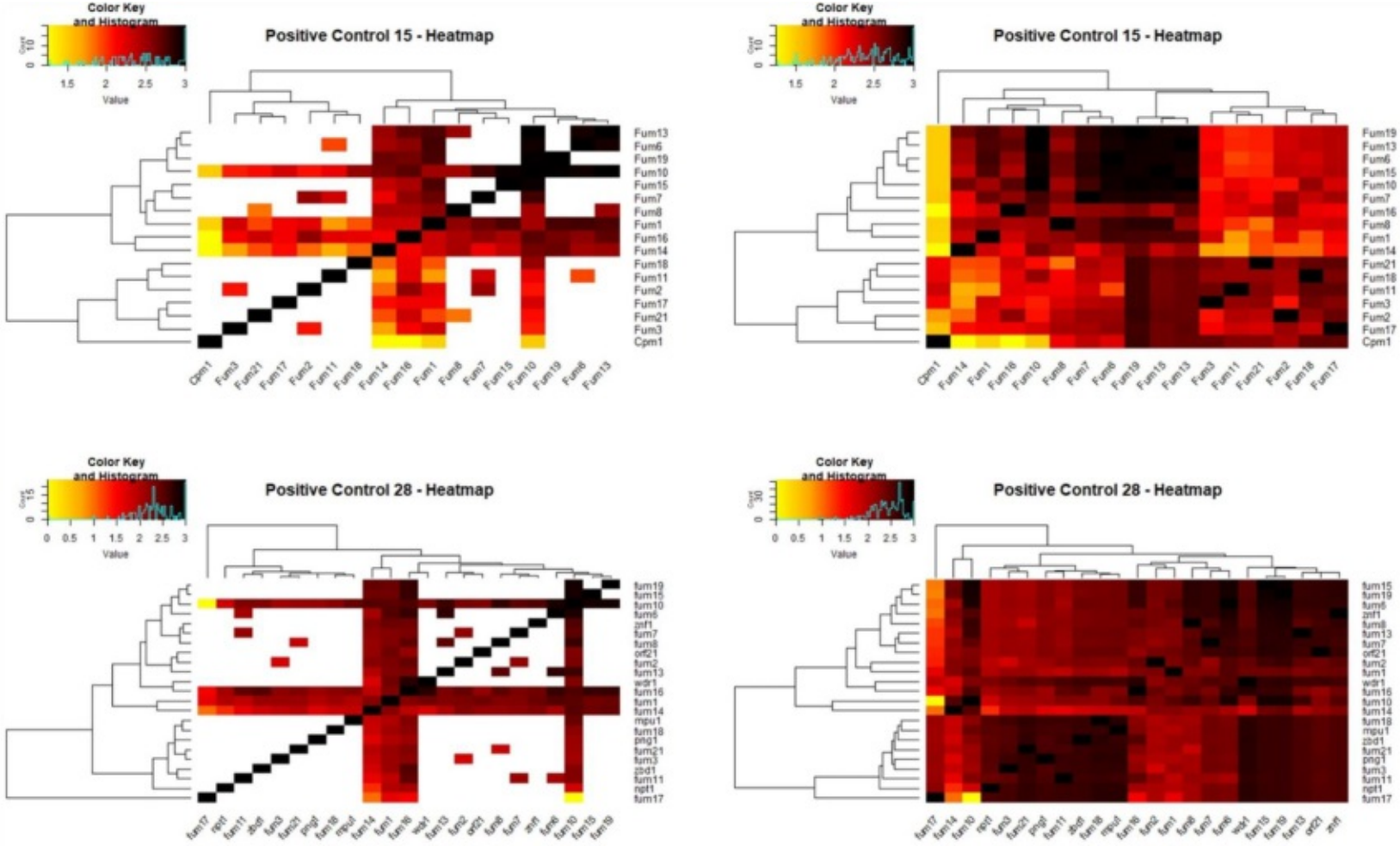


Figure 5.14: Heatmaps of the manual evaluation measures (MEMs) with missing data (left) and without (right) for control 15 (top) and control 28 (bottom). The missing data was approximated by a random forest approach (MissForest package in RStudio), yielding complete matrices and therefore complete heatmaps. The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes. Positive control 15 refers to fumonisin BGC in F. oxysporum, while positive control 28 represents the cluster of F. verticilloides.

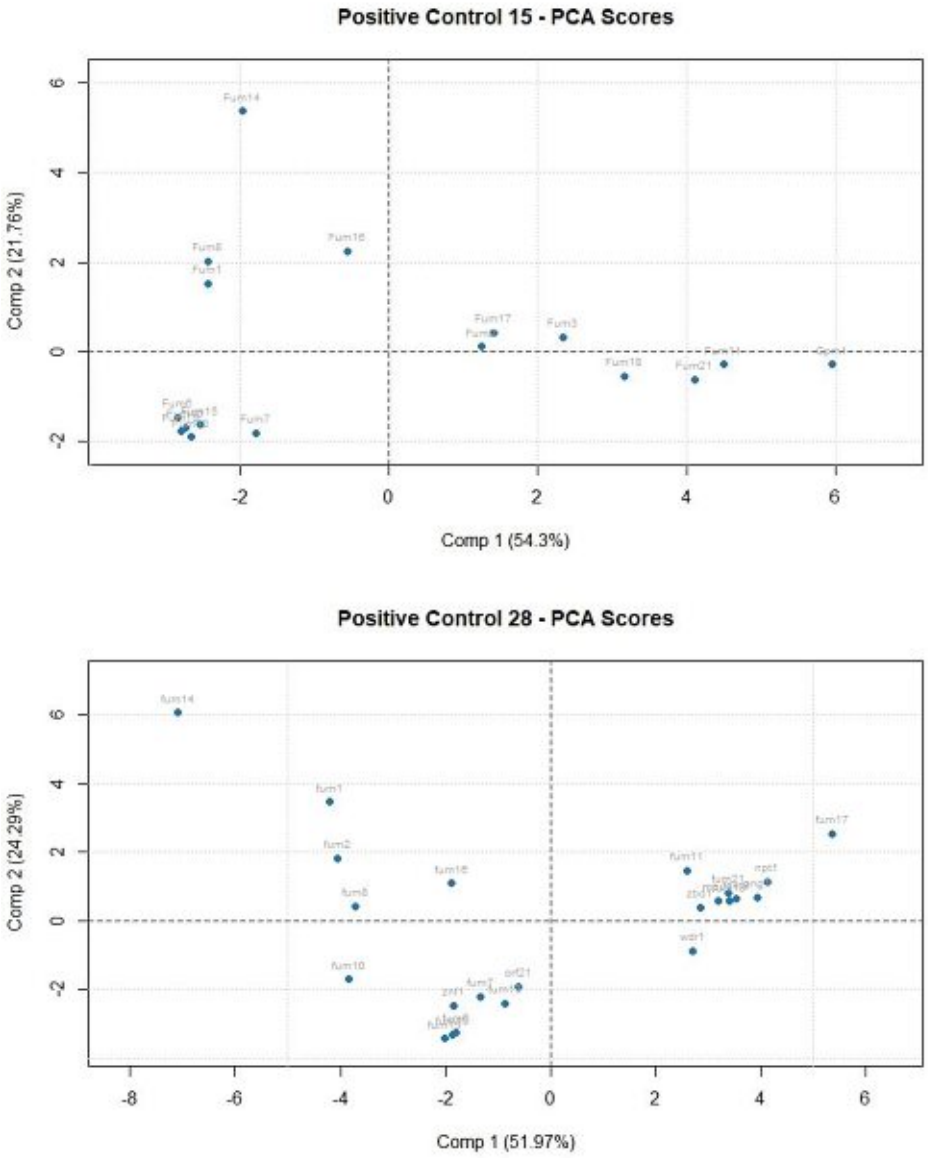


Figure 5.15: Scores of the calculated principal component analysis (PCA) of the fumonisin BGCs. A PCA calculates linear combination of the original variables in multivariate data sets yielding principal components, whose scores were plotted. The clustering of the scores indicates a coevolutionary linkage. Hence, the plots suggested that most genes shared evolutionary linkages. As the BGCs contained a high number of missing data, a clustering of those genes was expected and can be seen as an artifact. Regarding the evaluated genes, the clustering patterns in both clusters indicated a coevolutionary linkage between fum1, fum8 and fum16.

As shown in Figure 5.15, both BGCs comprised a lot of missing data which were approximated to perform PCA. The core genes (*fum1*, *fum10*, *fum14*, and *fum16*) were the only ones that were compared to all other cluster genes during evaluation. Focusing on the examination of these four genes indicated a similar clustering pattern between the control 15 and 28, especially for the genes *fum1*, *fum8* and *fum16*. As the other MEMs were approximated and the heatmaps in Figure 5.15 indicated an overestimation of their values, the clustering of these genes in the scores plot was therefore reasonable.

4. Penicillin BGC from *Penicillium chrysogenum* (pen)

Penicillin, a β-lactam drug, is the probably most famous antibiotic and was discovered in 1929. Improved industrial strains of *Penicillium chrysogenum* are still used for the production of penicillin,

exceeding 45.000 tons annually.[160] Two penicillin BGCs from different strains of *P. chrysogenum* were used for evaluation (Positive control 21 and 22).

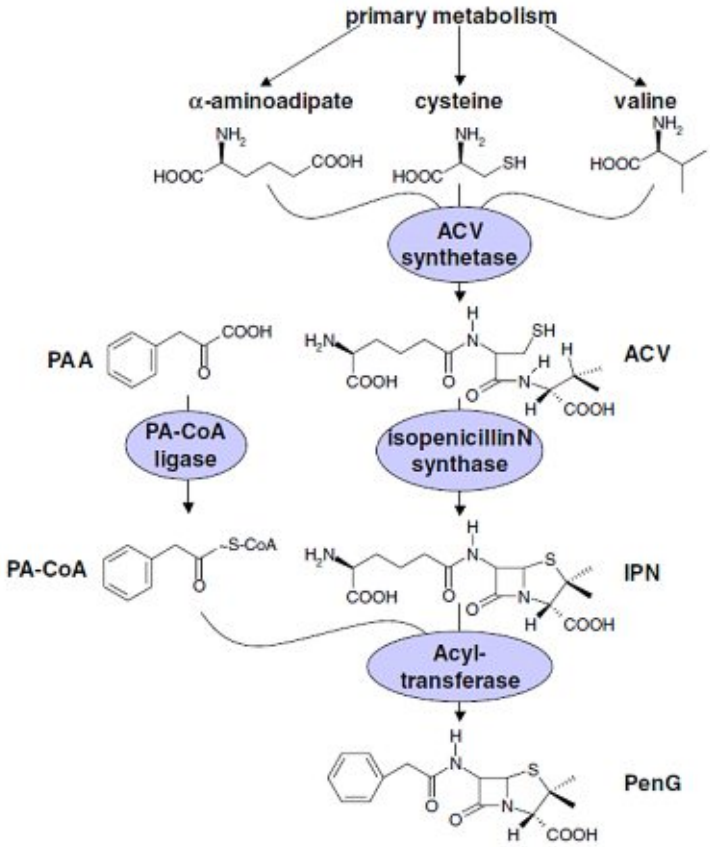


Figure 5.16: Biosynthetic pathway for Penicillin production according to van den Berg et al., 2007.[160] The gene pcbAB encodes for ACV synthetase, penDE encodes for isopenicillin N-acyltransferase and pcbC is encoding isopenicillin synthase. Together they form a tri-partite biosynthetic gene cluster (BGC) to build up Penicillin out of the precursors α-aminoadipate, cysteine and valine. ACV, L-α-aminoadipoyl-L-cysteinyl-D-valine; IPN, isopenicillin N, PenG, penicillin G; PAA, phenylacetic acid; PAA-CoA, phenylacetyl-coenzyme A.[160]

The Penicillin BGC contains three biosynthetic genes (*pcbAB*, *pcbC*, *penDE*) and 12 other Open Reading Frames (ORF) in *P. chrysogenum*. However, previous studies have shown that the three enzymes alone are sufficient to restore full β-lactam synthesis in a mutant lacking the complete region, not needing transporters or further genes that catalyse direct or indirect steps in the pathway.[160] Hence, comparison was made only between those genes, as they were form a tri-partite BGC according to literature (Figure 5.16).[160] The core gene *pcbAB*, encoding α-aminoadipoyl-D-cysteinyl-D-valine (ACV) synthetase, was additionally compared to all the ORFs.

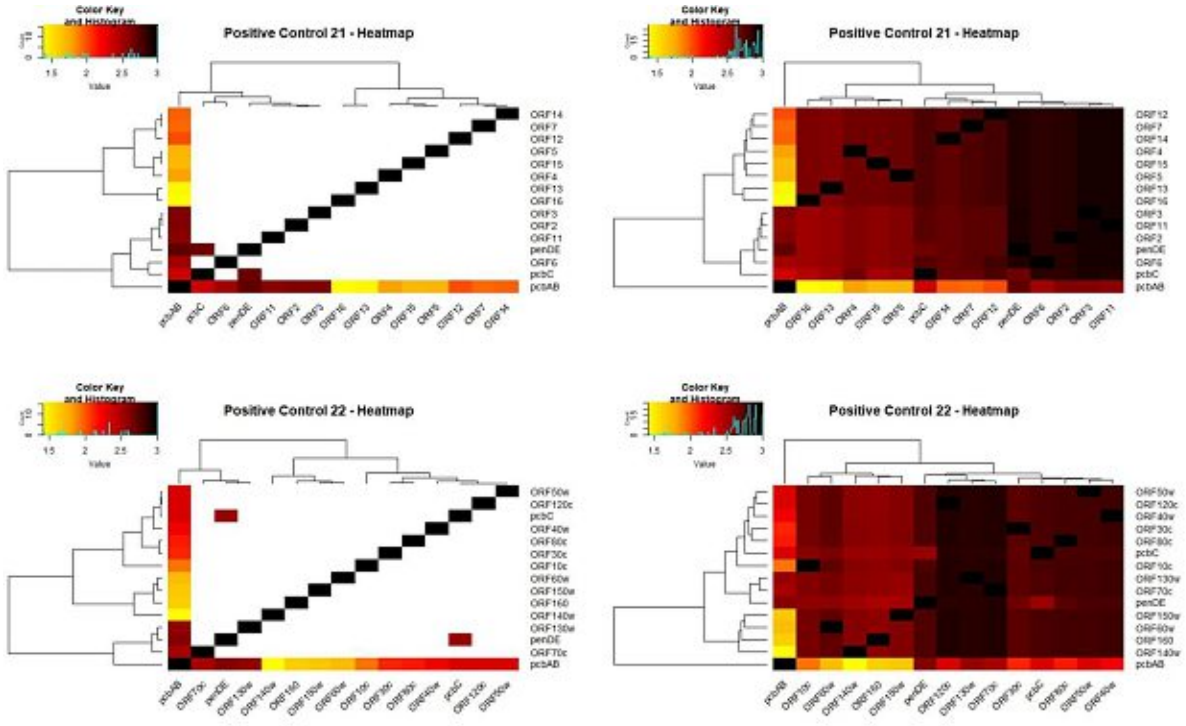


Figure 5.17: Heatmaps of the manual evaluation measures (MEMs) with missing data (left) and without (right) for control 21 (top) and control 22 (bottom). The missing data were approximated by a random forest approach (MissForest package in RStudio), yielding complete matrices and therefore complete heatmaps. The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes. Both clusters were Penicillin BGCs from two different strains (AS-P-78 and Wisconsin 54-1255) of the organism P. chrysogenum. Positive control 21 referred to MIBiG entry BGC0000404, whereas positive control 22 were extracted from NCBI GenBank®, accession number EF601124.1.

In Figure 5.17 the heatmaps of both BGC is illustrated, showing a high clustering pattern between the genes *pcbAB*, *penDE, pcbC* and further ORF genes. In control 21, the additional genes *ORF2*, *ORF3*, *ORF6* and *ORF11* clustered together with the tri partite BGC, whereas in control 22 the genes *ORF70c* and *ORF130w* were clustering with *pcbAB* and *penDE*. According to the approximated heatmaps, the ORF genes shifted to a strong coevolutionary linkage. As the ORF genes were compared to the core gene, only, the clustering pattern was in question, though. The examination of the tri partite BGC in control 21 indicated that all three genes were highly clustered, supporting their coevolutionary linkage. However, in control 22 only two parts of the previously introduced tri-partite BGC clustered together (*pbcAB* and *penDE*).[160]

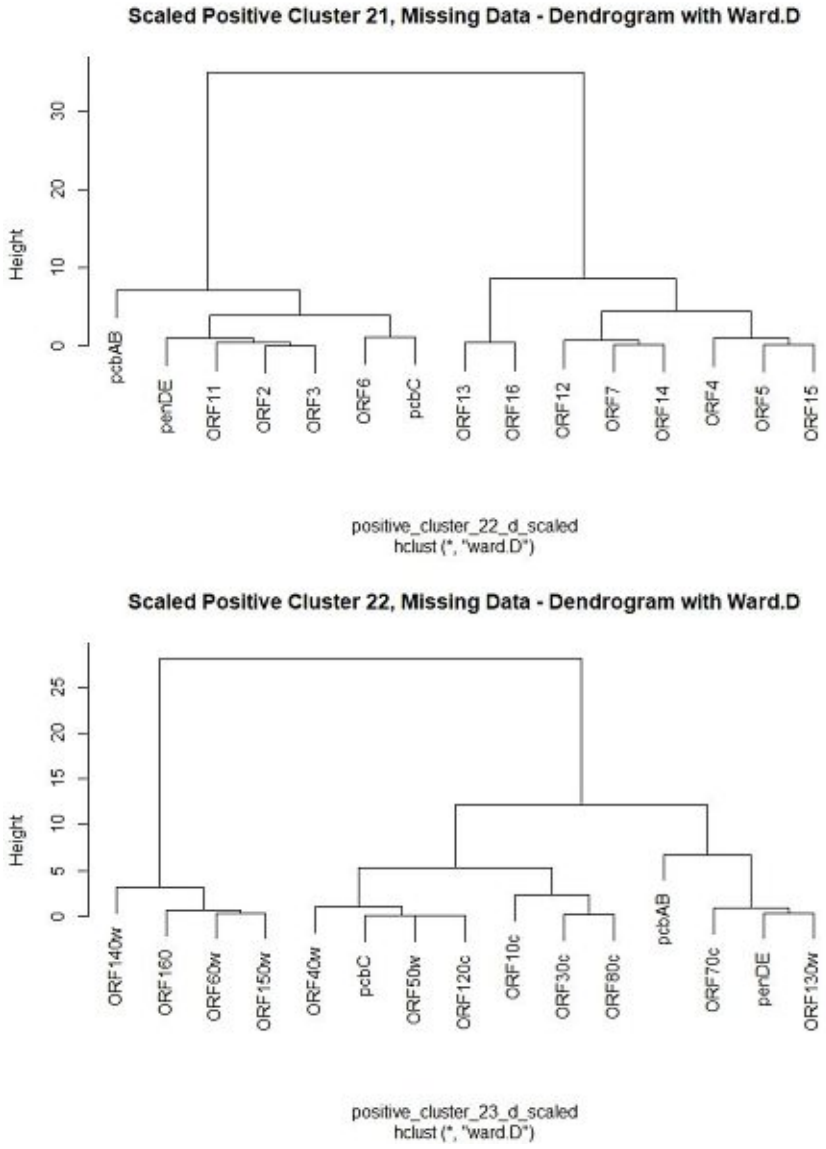


Figure 5.18: Scaled Dendrogram with missing data of positive controls 21, and 22 (two penicillin BGC from P. chrysogenum). The former conducted matrices were scaled and a distance matrix were returned using the functions scale and dist in RStudio. The illustrated dendrograms were computed using the hclust function and the ward clustering method ward.D in RStudio. Even though the penicillin BGCs included more genes, the product was produced by a tri partite BGC, consisting of the genes pcbAB, penDE, and pcbC.[160] These genes were clustering together in one clade indicating a coevolutionary linkage, as expected.

As illustrated in Figure 5.18 the genes *pcbAB*, *penDE* and *pcbC* were clustered in one clade in both controls. In control 22 the gene *pcbC* could be found in the same superclade, unlike the heatmap. Hence, more additional genes could be found there. As the ORF genes were singly compared to the core enzyme *pcbAB*, the splitting of the dendrogram was arguable. However, as the tri partite BGC genes clustered together in one clade, a coevolutionary linkage was reasonable.

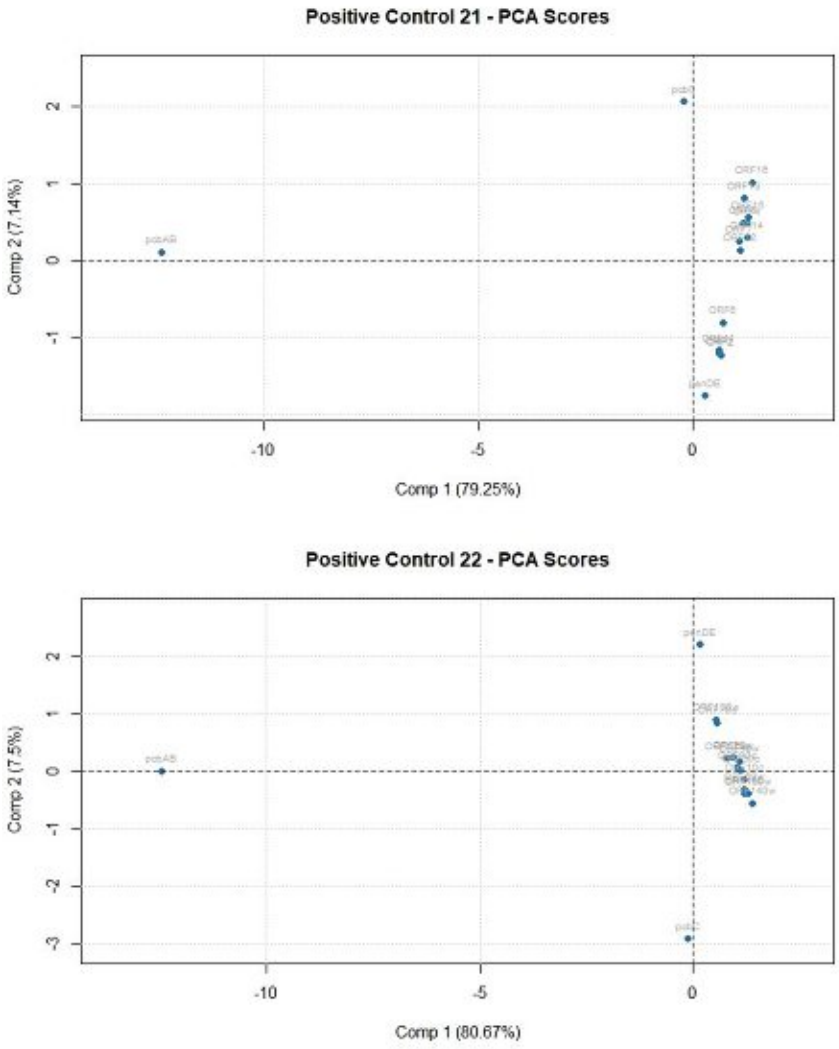


Figure 5.19: Scores of the calculated principal component analysis (PCA) of the penicillin BGCs. A PCA calculates linear combination of the original variables in multivariate data sets yielding principal components, whose scores were plotted. The clustering of the scores indicates a coevolutionary linkage. Hence, the plots suggested that most genes shared evolutionary linkages. As the BGCs contained a high number of missing data, a clustering of those genes was expected. Regarding the evaluated genes, penDE, pcbAB, and pcbC, the gene scores diverged, indicating no coevolutionary linkage, other than assumed.

The PCA were computed using the approximated data. As the approximation overestimated the values (see Figure 5.17), the ORF genes clustered together in the scores plot, while the three main genes diverged (Figure 5.19). Further plots based on the MEM matrices of the two Penicillin BGCs can be found in the supplement (Supplement 9.53 and Supplement 9.54)

5. Sorbicillin BGC from *Penicillium rubens* and *Trichoderma reesei* (sor)

Sorbicillinoids are yellow pigments produced by various fungi having antiviral, anti-inflammatory, and anti-microbial activities.[157] [175] To ensure a wider perspective during the evaluation the BGC from *Penicillium rubens* (Positive control 23) and the BGC from *Trichoderma reesei* (Positive control 27) were used.

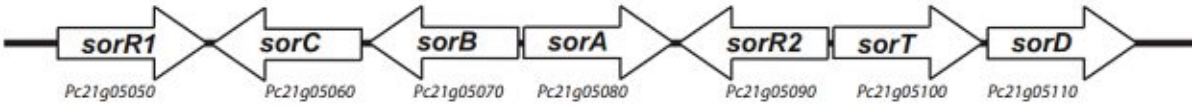


Figure 5.20: Sorbicillin cluster from Penicillium chrysogenum (Positive control 23). The arrows illustrated the structure of the BGC according to literature. The figure was extracted from Guzman-Chavez et al. (2017), figure 1. Pc21 g05050 (sorR1; transcriptional factor), Pc21 g05060 (sorC; monooxygenase), Pc21 g05070 (sorB; non-reduced polyketide synthase), Pc21 g05080 (sorA; highly reduced polyketide synthase), Pc21 g05090 (sorR2; transcriptional factor), Pc21 g05100 (sorT; MFS transporter) and Pc21 g05110 (sorD; oxidase).[132]

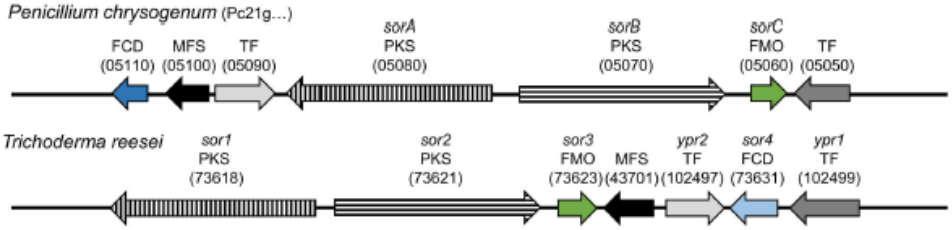


Figure 5.21: Comparison between Sorbicillin BGC from Trichoderma reesei (Positive control 27) and from P. chrysogenum. The arrows illustrated the structure of the BGC according to literature and arrows with the same filling were homologs. The figure was extracted from Derntl et al. (2017), figure 2. FCD, FAD/FMN-containing dehydrogenase; MFS, transporter of the multifacilitator superfamily; TF, transcription factor; PKS, polyketide synthase; FMO, FAD-dependent monooxygenase; HYD, hydrolase.[125]

As shown in Figure 5.21 the BGC structures of the two clusters varied. The extracted cluster from *T. reesei* (Positive control 27) lacked gene *sor3*, but included 7 additional genes (*118*, *119*, *120*, *121*, *122*, *123*, *124*). Gene 128 was annotated as a major facilitating transporter and therefore assumed to be the gene between *sor3* and *ypr2* in Figure 5.21, marked as “MFS”.

The genes *sor1, sor3*, *sor4* and *ypr1* in *T. reesei* were knocked out in gene deletion studies, resulting in no pigment for Δ*ypr1*, Δ*sor1* and Δ*sor3*, while Δ*sor4* mutants produced reduced amounts of sorbicillinoids.[125] Therefore, the proposed biosynthetic pathway was defined as shown in Figure 5.23. As described, the key intermediates are produced by the genes *sor1* and *sor2*. A Knoevenagel cyclization yield in the production of sorbicillin and dihydrosorbicillin, which both can be converted into further derivatives (sorbicillinol and dihydrosorbicillinol) by the gene products of *sor3* and *sor4*.[125] [132] In *P. chrysogenum*, gene deletion studies showed that *ΔsorA* and *ΔsorB* were core genes, generating the intermediates sorbicillin and dihydrosorbicillin (Figure 5.22).[132] [157]

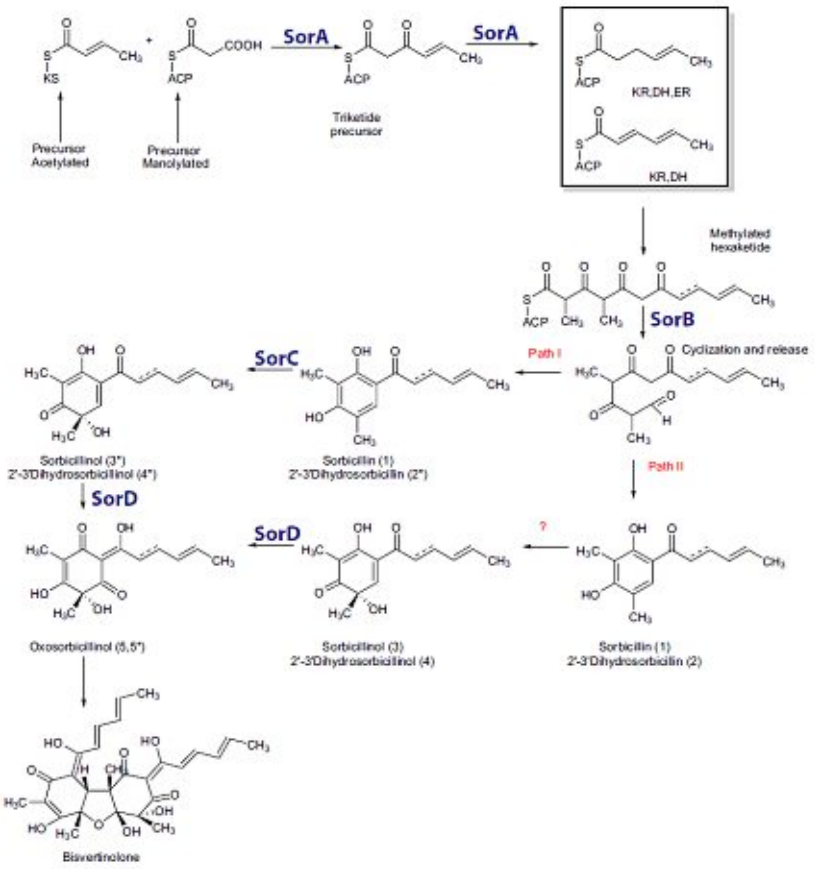


Figure 5.22: Proposed biosynthetic pathway of the sorbicillin cluster in P. chrysogenum according to Guzman-Chavez et al., 2007.[132] Starting with two precursors, the PKS sorA is forming the backbone with its domains KR (ketoreductase), DH (dehydratase) and ER (enoylreductase). The intermediate is then cyclized and released by the PKS sorB, yielding sorbicillin and dihydrosorbicillin. Further sorbicillin derivatives can be achieved using the genes sorC and sorD to convert the previously generated products.[132]

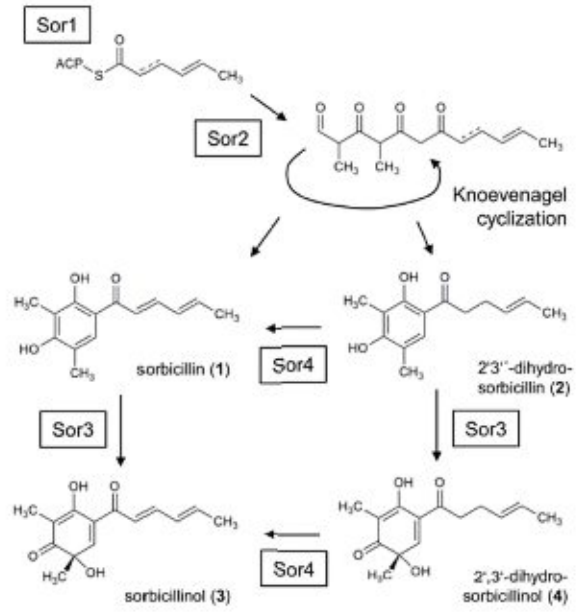


Figure 5.23: Proposed biosynthetic pathway of the sorbicillin cluster in T. reesei according to Derntl et al., 2017.[125] The key intermediates are produced by the core genes sor1 and sor2. A Knoevenagel cyclization forms the products sorbicillin and dihydrosorbicillin, which can be further converted by the genes sor3 and sor4, yielding sorbicillinol and dihydrosorbicillinol.[125]

According to literature, the gene *sor1* is the homolog of *sorA* in *P. chrysogenum*, which is the first PKS of the biosynthetic pathway (Figure 5.23). Analogical, *sor2* is homologous to *sorB*, which is a PKS as well.[125] Both genes were therefore operated as core enzymes and compared to all other genes in the cluster. In positive control 28 a further gene was annotated as a PKS/NRPS like protein (*124*). Hence, it was compared to all other genes as well.

According to the biosynthetic pathway of *T. reesei* and *P. chrysogenum*, the genes *sor3* and *sor4* in *T. reesei* were homologous to the genes *sorC* and *sorD* in *P. chrysogenum* (Figure 5.22, and Figure 5.23)

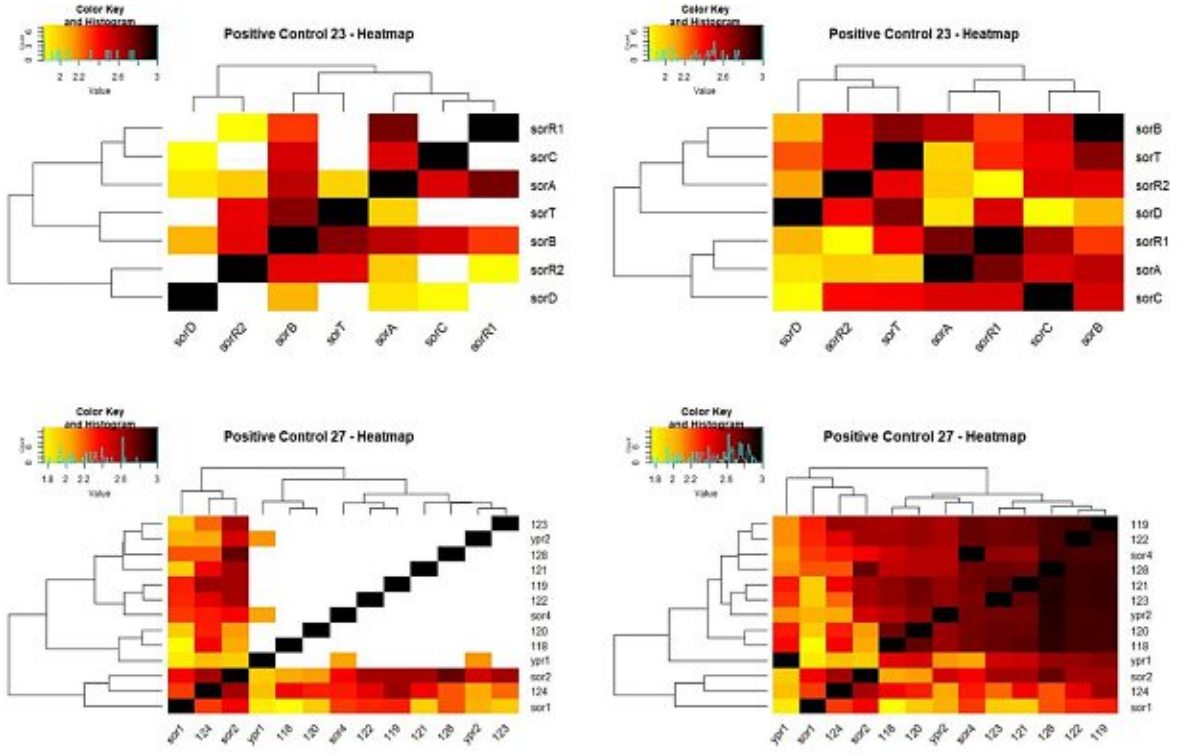


Figure 5.24: Heatmaps of the manual evaluation measures (MEMs) with missing data (left) and without (right) for control 21 (top) and control 22 (bottom). The missing data were approximated by a random forest approach (MissForest package in RStudio), yielding complete matrices and therefore complete heatmaps. The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes. Sorbicillin cluster from P. chrysogenum (Positive Control 23) derived from MIBiG (BGC0001404), while the cluster from T. reesei (Positive Control 27) was extracted from GenBank® with the accession number GL985056

The calculated heatmaps were illustrated in Figure 5.24. The genes *sor1*, *sor2* and *124* clustered in the heatmap of positive control 27 indicating a coevolutionary relationship. As this cluster was missing the *sor3* gene, no comparison to *sor4* could be achieved. In *P. chrysogenum* (control 23) the clustering patterns of *sorA* and *sorC* as well as *sorB* and *sorT* indicated that these two pairs share evolutionary traits. Interestingly, the comparison of the genes *sorD* and *sorC* yielded in comparably low MEM, assuming that they were not coevolutionary linked.

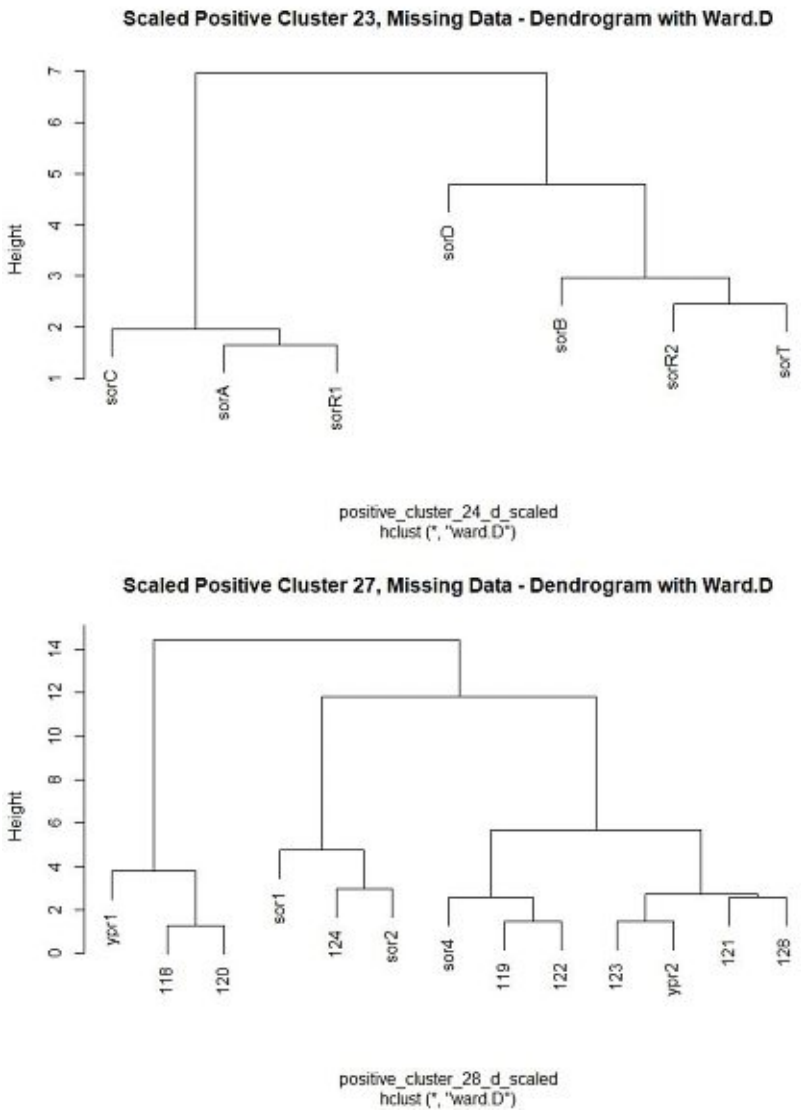


Figure 5.25: Scaled Dendrogram with missing data of positive controls 23, and 27 (two sorbicillin BGC from P. chrysogenum and T. reesei, respectively). The former conducted matrices were scaled and a distance matrix were returned using the functions scale and dist in RStudio. The illustrated dendrograms were computed using the hclust function and the ward clustering method ward.D in RStudio. The genes sorC, sorR1 and sorA were clustering together in control 23, while the genes sor1, 124 and sor2 shared the same clade in control 27, both indicating a coevolutionary linkage.

The scaled dendrograms in Figure 5.25 included the missing data. A similar result compared to the heatmap was therefore apparent. In control 23 (*P. chrysogenum*) the genes *sorC* and *sorA* clustered together with *sorR1*, while *sorB* and *sorT* share the same clade with *sorR2*. According to literature, the genes *sorR1* and *sorR2* expressed transcriptional factors.[132] Like the heatmaps, the *T. reesei* genes *sor1*, *124*, and *sor2* shared the same clade in the dendrogram (control 27 in Figure 5.25).

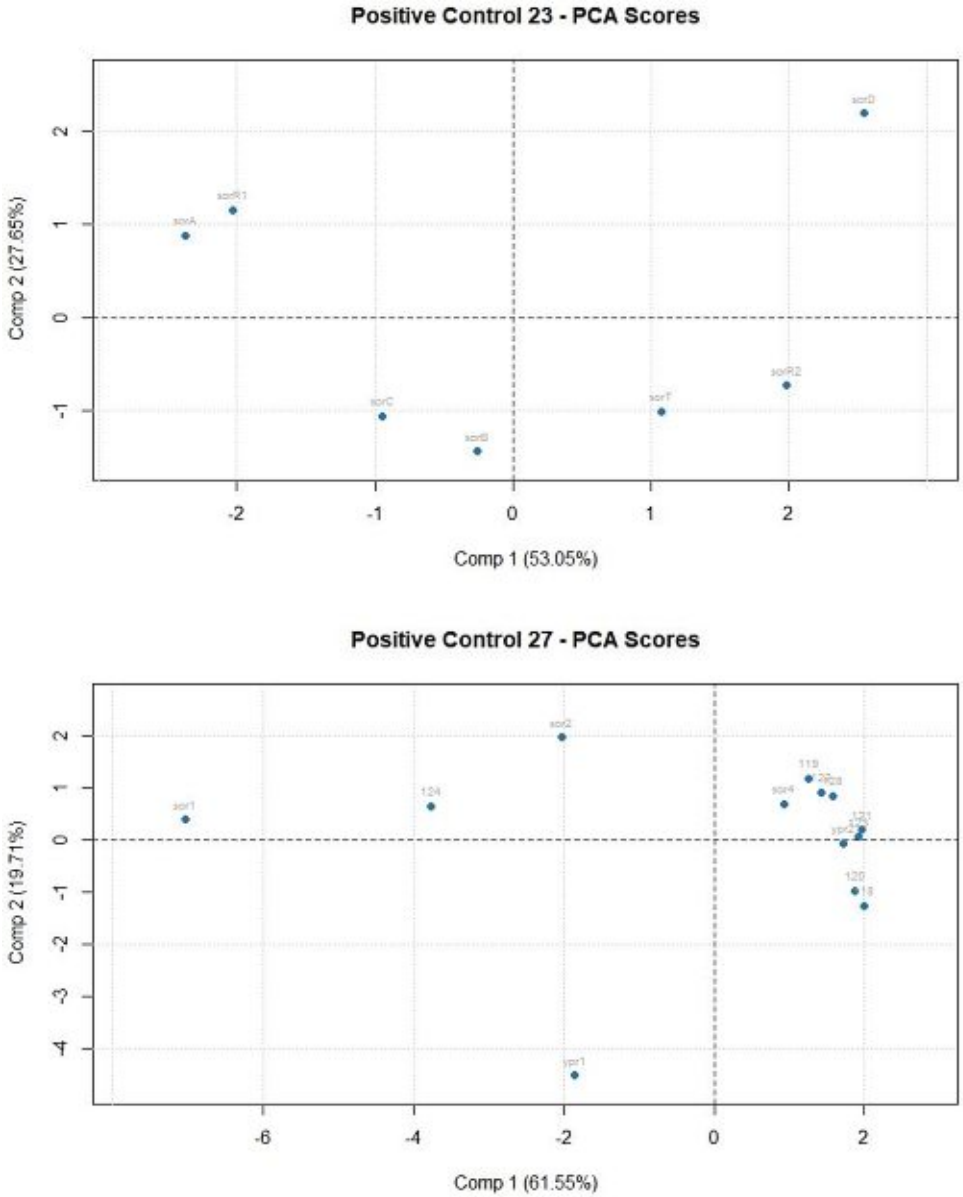


Figure 5.26: Scores of the calculated principal component analysis (PCA) of the sorbicillin BGCs. A PCA calculates linear combination of the original variables in multivariate data sets yielding principal components, whose scores were plotted. The clustering of the scores indicates a coevolutionary linkage. Hence, the plots suggested that the genes sorA and sorR1 in control 23, and the approximated genes in control 27 shared evolutionary linkages. As the BGCs contained a high number of missing data, a clustering of those genes in control 27 was expected. Regarding the evaluated core genes, the scores diverged, indicating no coevolutionary linkage, other than assumed.

Examining the PCA scores in Figure 5.26 subpart control 23 indicated that the genes *sorA* and *sorR1* clustered together, while *sorC* diverged from them clustering with *sorB*, other than illustrated in the previous plots. The subpart control 27 showed a clustering of all genes, except of *ypr1*, *sor1*, *124* and *sor2* which diverged. As the PCA scores were calculated using the approximated data sets, a discrepancy was expected. However, the core enzymes (*sor1*, *sor2*, *124*) were completely analysed and were still located in the same quadrant of the plot indicating a slight coevolutionary linkage. As control 23 lacked less data compared to control 27, the PCA scores of control 23 could be therefore regarded as more reliable. However, the core enzymes diverged instead of clustering together indicating no evolutionary linkage, other than illustrated in the heatmap and dendrogram.

6. Positive control BGCs with low aMEM

Eight out of 29 positive controls had only low average manual evaluation measures (aMEM), despite high MEM values or because of low MEM values. As the aMEMs were calculated averages of the genes that were empirically verified as necessary for secondary metabolite production, high aMEMs were expected for all BGCs. Hence, a deeper insight into the diversity among these positive controls was illustrated in the following.

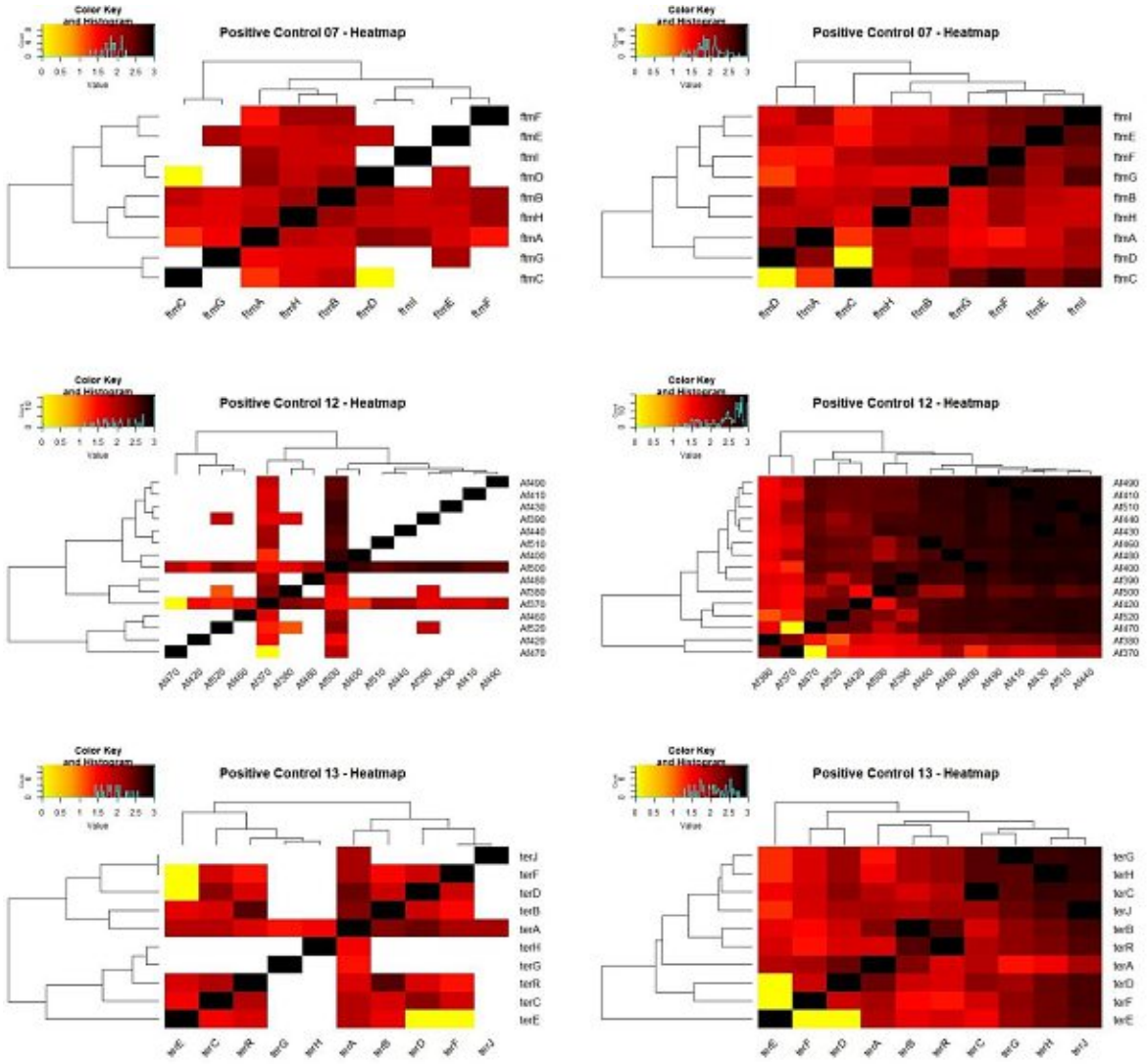


Figure 5.27: Heatmaps of the manual evaluation measures (MEMs) with missing data (left) and without (right) for control 07 (top), control 12 (middle), and control 13 (bottom). Despite high MEM values the average manual evaluation measures (aMEM) were very low for these clusters (1.63, 1.66, and 1.62, respectively). The missing data were approximated by a random forest approach (MissForest package in RStudio), yielding complete matrices and therefore complete heatmaps. The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes. Positive control 07 referred to the fumitremorgin BGC from Aspergillus fumigatus (MIBiG BGC0000356). The fumagillin BGC (control 12, MIBiG BGC0001067) derived from A. fumigatus as well, whereas control 13 referred to the terrain BGC from A. terreus (MIBiG BGC0000161).

Despite high MEM values, the aMEMs of fumitremorgin BGC (Positive control 07), fumagillin BGC (Positive control 12), and terrein BGC (Positive control 13) were very low (1.63, 1.66, and 1.62, respectively).

The first intermediate of fumitremorgin B, brevianamide F, is produced out of the precursors L-tryptophane and L-proline using the ftmA product, a NRPS. A further conversion by FtmB, a dimethylallyltryptophan synthase, yields the intermediate tryprostatin B.144 149 In another study, it was pointed out that the gene ftmH, a tryptophane dimethylallyltransferase, was involved in the last step of the biosynthetic pathway forming fumitremorgin B.131 144 These three genes were therefore assigned as core enzymes. Further gene deletion studies indicated that ftmC, ftmE and ftmG, which all showed similarities to fungal cytochrome P450, were important for hydroxylation, C-N bond formation and dihydroxylation, while the disruption of ftmF, a verruculogen synthase, suggested that it is not involved in the biosynthesis of fumitremorgin B, but catalyzes the conversion of fumitremorgin B to verruculogen.144 145 According to the illustrated heatmap in Figure 5.27 subpart control 07, the core genes clustered together indicating an evolutionary relationship, as expected. However, when examining the MEM values, it became obvious that they fell below the threshold of 2.0 for almost all gene comparisons which were identified as necessary for the secondary metabolite production (see 5.4.2 Average manual evaluation measure (aMEM)) (see Supplement 9.3 - Supplement 9.31). Hence, the BGC yielded a low aMEM.

Fumagillin, an anti-angiogenic secondary metabolite, is produced by the core gene *Af370*, a PKS, in *A. fumigatus*. The fumagillin BGC (control 12) included 15 genes, but only 3 genes (*Af370*, *Af380*, *Af520*) were assumed to be necessary for the biosynthetic pathway.[103] However, as *Af500* was annotated as a NRPS-like protein, it was also treated like a core gene. According to Lin *et* al. (2013) the genes *Af370* and *Af380* encode a highly reducing polyketide synthase (HR-PKS) and a α/β hydrolase, respectively, while *Af520* was assumed to encode a terpene cyclase and be responsible for the production of the intermediate β-trans-bergamotene.[103] According to the heatmap in Figure 5.27 subpart control 12, the genes *Af370* and *Af380* clustered together as expected. Interestingly, the gene *Af500* showed high MEMs for all other genes, which were assigned to be not necessary for the production of fumagillin, indicating that they were coevolutionary linked forming another, unknown BGC.

Positive control 13 referred to the terrain BGC from *A. terreus*. Terrein has antimicrobial, antiproliferative, and antioxidative activities and has a phytotoxic effect on plant growth.[170] Gene deletion studies showed that the genes *terA*, the core gene producing a NRPS, *terR*, an essential coregulator, and *terF* were indispensable for the biosynthesis. Furthermore, the disruption of the genes *terB*, *terC*, *terD*, and *terF*, respectively, showed no terrain production, indicating that they contribute to the biosynthetic pathway. Contrariwise, the genes *terH* and *terI* were assumed to be dispensable.[170] Examining the biosynthesis, *terA* produces the backbone intermediate which is further converted by *terB* yielding 6-hydroxymellein. The final product is then formed by several steps using the genes *terC*, *terD*, *terE* and *terF*.[170] FunOrder did not produced any tree for gene *terI*, which indicated that the database did not have homologs to this gene.[170] The heatmap in Figure 5.27 subpart control 13 illustrated that *terA* and *terB* cluster together, which went with the biosynthetic pathway. Notably, the gene *terE* had the least clustering patterns with other genes. Moreover, the comparison of the phylogenetic trees with *terF* and *terD* showed no congruence and exhibited low MEMs, other than expected. However, the MEMs of those genes (*terF*, *terD*) compared with the core gene *terA* were high and additionally, they shared the same clade in the heatmap, supporting a coevolutionary linkage.

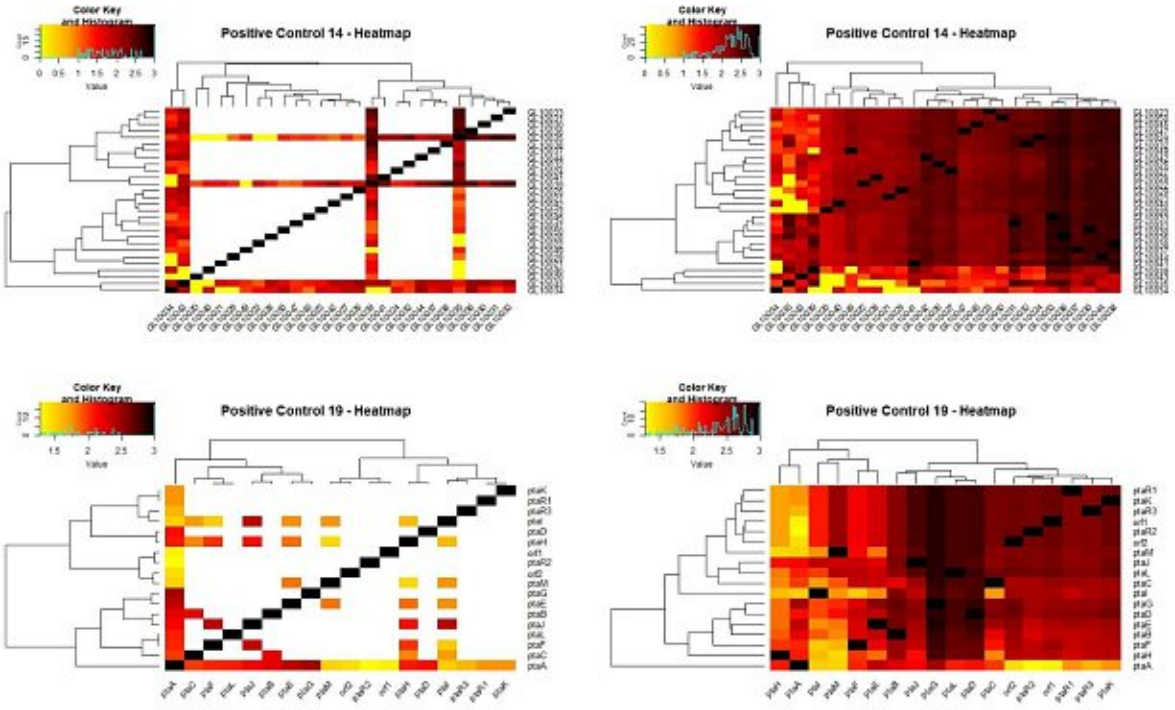


Figure 5.28: Heatmaps of the manual evaluation measures (MEMs) with missing data (left) and without (right) for control 14 (top), and control 19 (bottom). These clusters scored low average manual evaluation measures (aMEM) and had a suspiciously big number of missing data. The missing data were approximated by a random forest approach (MissForest package in RStudio), yielding complete matrices and therefore complete heatmaps. However, as illustrated the approximation were overestimated by the algorithm. The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes. Positive control 14 referred to the pneumocandin BGC from Glaera lozoyensis (MIBiG BGC0001035), while control 19 referred to the pestheic acid BGC (MIBiG BGC0000121) from Pestalotiopsis fici.

The clusters illustrated in Figure 5.28 had a suspiciously big number of missing data, because of the size of their clusters. As shown in the right part of Figure 5.28, the two BGCs constituted as good examples for the overestimation of the approximation of the data sets, which were calculated by the Miss Forest algorithm. The plots calculated with the approximated data were therefore observed with suspicion.

Control 14 referred to the pneumocandin BGC from the fungus *Glarea lozoyensis* yielding an aMEM of 1.57. Pneumocandins are lipohexapeptides with antifungal activities.[121] [123] The cluster contained two core genes, a NRPS (*GL10035*) and a PKS (*GL10034*), that were disrupted in gene deletion studies, which interrupted the production of pneumocandin.[123] Further studies demonstrated that *GL10043*, an acyl adenonsine-5'-monophosphate (AMP) -ligase, was important for the shuttling of the PKS to the core hexapeptide initiating the lipoinitation step. Furthermore, the optimal function of the PKS was assumed to be maintained by a second gene (*GL10032*).[121] The core genes (*GL10034* and *GL10035*), as well as the gene producing the ligase (*GL10032*) were compared to all other genes. Furthermore, as the gene *GL10039* was annotated as a synthase, the analysis was made similarly to the core genes. According to literature, the genes between *GL10037* and *GL10041* belong to a different BGC, the L-homotyrosine BGC.[123] Hence, a coevolutionary linkage with the other genes in the pneumocandin BGC was not estimated. In fact, the majority of the genes between *GL10037* and *GL10041* shared the same super clade in the heatmap (Figure 5.28, subpart control 14). Notably, the second core gene *GL10035* did also cluster with the L-homotyrosine BGC. However, the PKS gene (*GL10034*) and the gene *GL10043* were clustering together in the heatmap, as estimated, indicating a coevolutionary linkage. The low aMEM resulted due to low contributing MEMs between *GL10034*

and the genes *GL10032* and *GL10035*. This pointed to the direction that either the two gene pairs apparently did not share evolutionary traits, or there was too much missing data for a clear clustering pattern.

As shown in the heatmap, the pestheic acid BGC (Positive control 19) showed generally low MEM values. A low aMEM (1.91) was therefore expectable. Pestheic acid is a diphenyl ether produced by a NRPKS (*ptaA*) in the endophytic fungus *Pestalotiopsis fici*.[167] According to literature, ptaA forms the backbone, which is further converted by *ptaB* and *ptaC* to the intermediate endocrocin. Further conversions by the genes *ptaH*, *ptaI*, *ptaJ*, *ptaF*, *ptaM* and *ptaE* result in the final product, pestheic acid.[167] Gene deletion studies further showed that the disruption of *ptaA* and *ptaE*, respectively, resulted in no product, indicating an importance for the biosynthesis of pestheic acid. The gene *ptaA* was used as a core gene and compared to all other cluster genes. Additionally, the received trees of associated genes were compared to trees of those genes that were followed by them due to the pathway. According to analysis, the *ptaA* showed high MEM values when compared with *ptaB* and *ptaC*, but also compared to *ptaE*. Furthermore, they shared the same superclade in the heatmap. However, according to literature a further exploration of the genetic and biochemical mechanisms was still outstanding.[167] Hence, the contributions of the genes in biosynthetic pathway and therefore their contribution to the calculation of the aMEMs were questionable. In fact, most of these gene comparison yielded low MEMs, supporting the suggestion of a further exploration of this cluster.

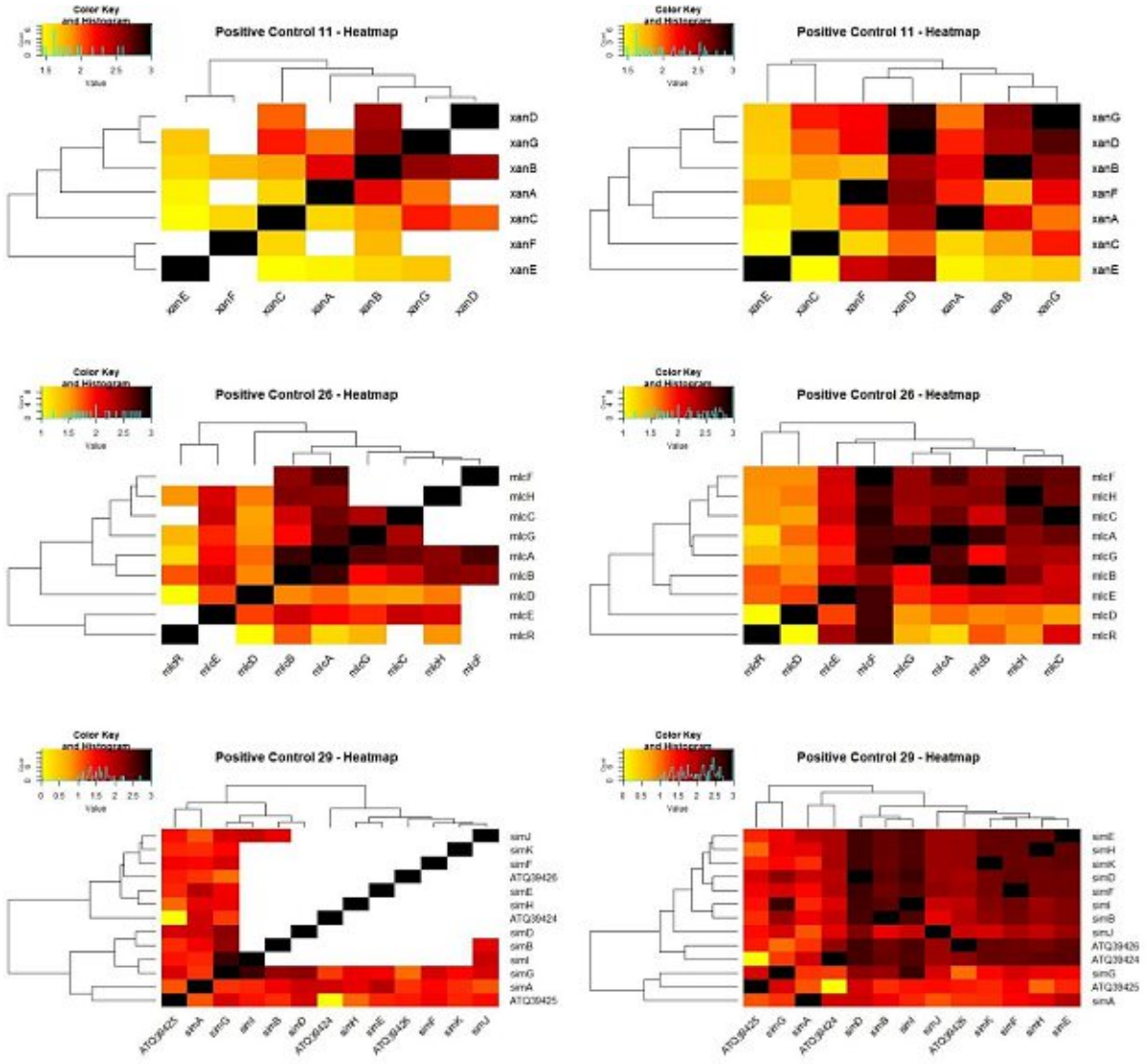


Figure 5.29: Heatmaps of the manual evaluation measures (MEMs) with missing data (left) and without (right) for control 11 (top), control 26 (middle), and control 29 (bottom). Because of low MEM values the average manual evaluation measures (aMEM) were also comparably low for these clusters (1.91, 1.86, and 1.77, respectively). The missing data were approximated by a random forest approach (MissForest package in RStudio), yielding complete matrices and therefore complete heatmaps. The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes. Positive control 11 referred to the xanthocillin BGC from Aspergillus fumigatus (MIBiG BGC0001990). The compactin BGC (control 26, MIBiG BGC0000039) derived from Penicillium citrinum, whereas control 29 referred to the cyclosporine BGC from Beauveria felina (MIBiG BGC0001565).

The three positive controls illustrated in Figure 5.29 generally demonstrated low MEM values and were therefore further discussed in the following.

Xanthocillin is an isocyanide produced by various fungi. Isocyanides are assumed to be chalkophores, meaning that they are involved in the copper-uptake of the organism. In fact, in *Aspergillus fumigatus*, xanthocillin is produced under copper-depleted conditions.[102] During the analysis, a xanthocillin BGC from *A. fumigatus* (positive control 11) comprising 7 genes was used. Gene deletion studies demonstrated that *xanB* was the core enzyme producing an isocyanide synthase-dioxygenase, and *xanC* was a regulation factor that upregulated all other genes, except of *xanD*.[102] Hence, these two genes were compared to all other cluster genes. As described by Lim *et* al. (2018), *xanB* produces the intermediate which is further converted by *xanG* to yield xanthocillin. A

subsequent conversion by *xanA* and/or *xanE* result in the production of various xanthocillin derivatives.[102] When inspecting the heatmap in Figure 5.29, subpart control 11, the genes *xanB*, *xanG* and *xanD* clustered together, suggesting that they share evolutionary traits. In fact, the comparison between *xanB* and the genes *xanG* and *xanD*, respectively, yielded in high MEMs, whereas the rest of the comparisons scored low values. Notably, the regulation factor *xanC* did not scored high MEMs, indicating no coevolutionary linkage with the rest of the genes. Hence, a low aMEM value (1.91) was expectable.

Control 26 referred to the compactin BGC. Compactin is an inhibitor of 3-hydroxy-3-methylglutaryl-coenzynme A (HMG-CoA) reductase and used as a substrate to produce the pharmaceutical drug, pravastatin sodium.[111] The compactin cluster contained 10 genes (*mlcA* – *mlcI* and *mlcR*), including two backbone producing enzymes (*mlcA* and *mlcB*), which both were PKSs. According to the gene disruption study, the deletion of *mlcA* yielding no product demonstrated its importance for the biosynthesis. The disruption of *mlcB* meanwhile resulted in the production of a compactin-precursor, indicating that *mlcB* is involved in the biosynthetic pathway.[111] Hence, a strong evolutionary linkage between those two genes was estimated. As illustrated in the heatmap in Figure 5.29 subpart control 26, the core genes cluster together, supporting the previously estimation. The aMEM value was calculated based on the proposed biosynthetic pathway. Abe *et* al. (2001) proposed that *mlcE* and *mlcD* were resistant genes and in 2002, the functional analysis of *mlcR* revealed that it was a regulatory gene.[110] [111] A coevolutionary linkage of all three genes (*mlcE*, *mlcD*, and *mlcR*) to the pathway genes was therefore estimated. However, those gene comparison yielded only low MEM values which contributed to the low aMEM score.

The cyclosporine BGC (control 29) from *Beauveria felina* (also known as *Amphichorda felina*) comprised 14 genes. According to literature, the core gene was *simA*, a NRPS, yielding cyclosporine C (CsC), a similar compound of the characterized immunosuppressant drug cyclosporine A (CsA).[165] In fact, the gene clusters for CsA and CsC shared a high sequence similarity according to literature.[165] However, as the cyclosporine BGC from *Tolypocladium inflatum* was not available and the only available BGC derived from *Beauveria felina*, which lacked a proposed biosynthetic pathway, the evaluation was done using the biosynthetic pathway of CsA. A low aMEM (1.77) was therefore not surprising. According to that pathway, an important intermediate for cyclosporine production was (4*R*)-4-([(*E*)-2-butenyl]-4-methyl-L-threonine (BMT), which was produced by the PKS *simG*, and the genes *simI* and *simJ*. The gene *simB* converts L-Alanin into D-Alanin, which, together with BMT, is subsequently used by *simA* as starting material to produce cyclosporine A.[107] The genes *simA* and *simG* were therefore marked as core genes and compared to all other genes. Examinating the heatmap in Figure 5.29 subpart control 29, the genes *simG*, *simI*, *simB* and *simD* clustered together and shared a superclade with *simA*, which went along with the biosynthetic pathway indicating a slight coevolutionary linkage. However, the MEMs of *simA* were low compared to the MEMs of *simG*, which pointed to the direction that *simG* shared more evolutionary traits with the other genes than *simA*.

5.4.1.1 Negative controls

The analysis of the negative controls was carried out as described in the chapter 4.4. Tree comparison. Heatmaps, dendrograms and PCAs were calculated for 59 out of the 60 random controls. The remaining control (negative control 60) included only two phylogenetic trees yielding only one MEM. To establish an array for the statistical analysis, at least three phylogenetic trees were needed. Hence control 60 was excluded for the further analysis. However, the MEM value of negative control 60, which equalled the aMEM value, was 1.5 (see Table 5.5).

In the following, nine random controls were picked to show the diversity of the analysis of the negative control BGCs.

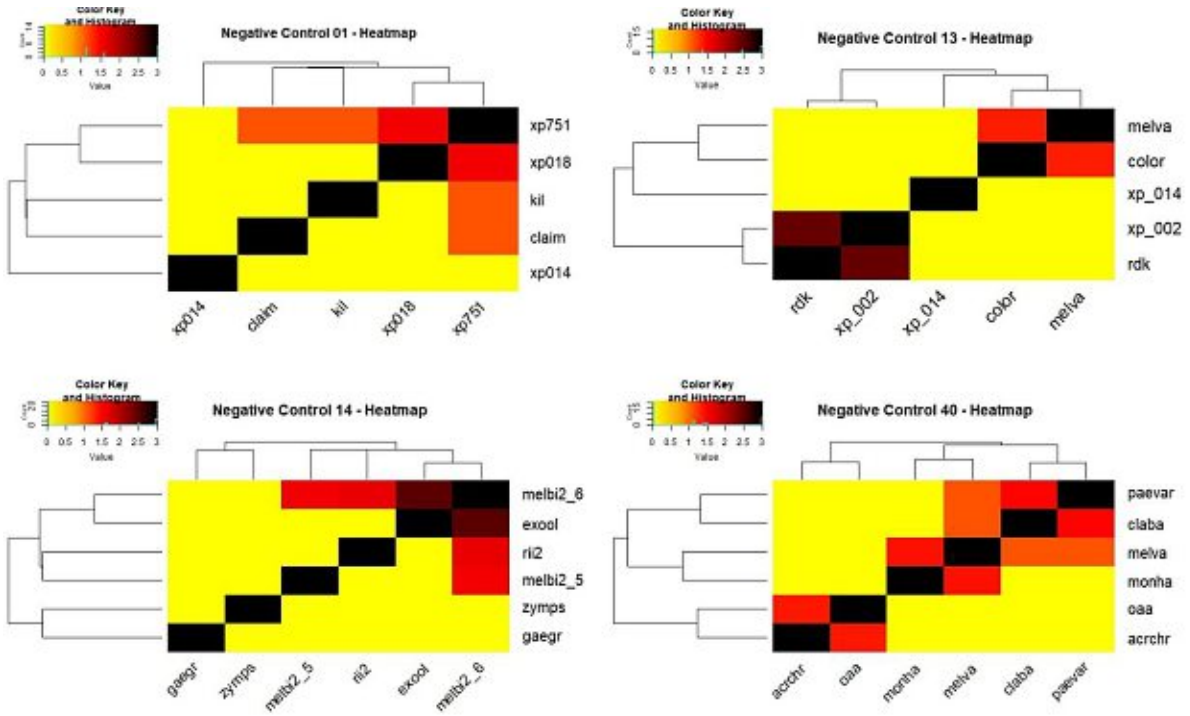


Figure 5.30: Heatmaps of the manual evaluation measures (MEMs) of the negative controls that all scored the lowest average manual evaluation measure (aMEM) of 0.4: negative control 01 (top left), control 13 (top right), control 14 (bottom left), and control 40 (bottom right). The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes.

The heatmaps with the lowest and the highest score for aMEM were shown in Figure 5.30 and Figure 5.31. Random control no. 14 had the lowest aMEM score (0.4), alongside with the random controls number 1, 13 and 40, indicating that FunOrder could differentiate between randomly generated genes and real biosynthetic gene clusters. However, some MEMs were comparably high, e.g., the comparison of exool and melbi2_6 in negative control 14 or rdk and xp_002 in negative control 13, indicating a coevolutionary linkage between them. Though, the rest of the phylogenetic trees did not share similar topologies. This pointed to the direction, that the MEM values were only valid, when the aMEM of the cluster was high.

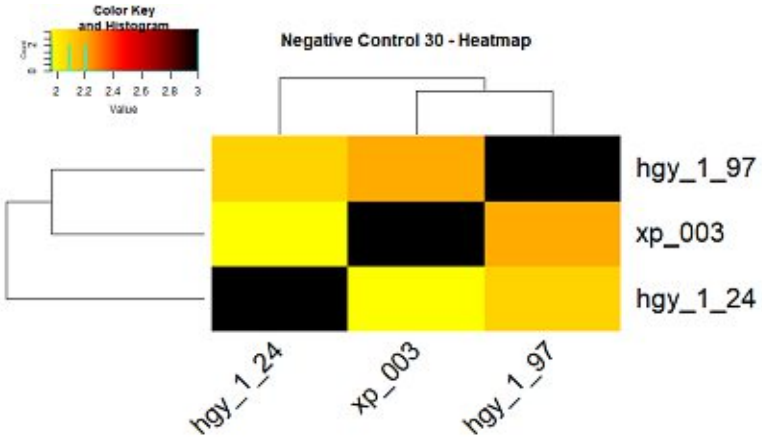


Figure 5.31: Heatmaps of the manual evaluation measures (MEMs) of negative control 30, which scored the highest average manual evaluation measure (aMEM) of 2.1 indicating an evolutionary linkage. As shown, the MEMs were

comparably high (2.09, 2.20, and 1.96), yielding a high aMEM. This BGC was regarded as false positive. The colour key on the upper left side of each heatmap illustrated the MEM value range between ~1.9 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes.

Negative control 30 resulted in the highest aMEM (2.1), followed by control 60 with an aMEM of 1.5. While control 30 contained only 3 phylogenetic trees (Figure 5.31), negative control 60 included only two and could therefore not be further analysed. According to that, control 60 could not be regarded as truly false positive. As illustrated in Figure 5.31 negative control 30 yielded comparably high MEM values (2.09, 2.20, and 1.96), indicating a coevolutionary linkage of the randomly generated genes. As FunOrder calculated only a few phylogenetic trees out of control 30 indicated that the identification of BGCs of only three genes must be regarded with suspicion. However, when comparing the number of used random controls the likelihood of receiving a false positive was rather low (1:60 or rather 1:30).

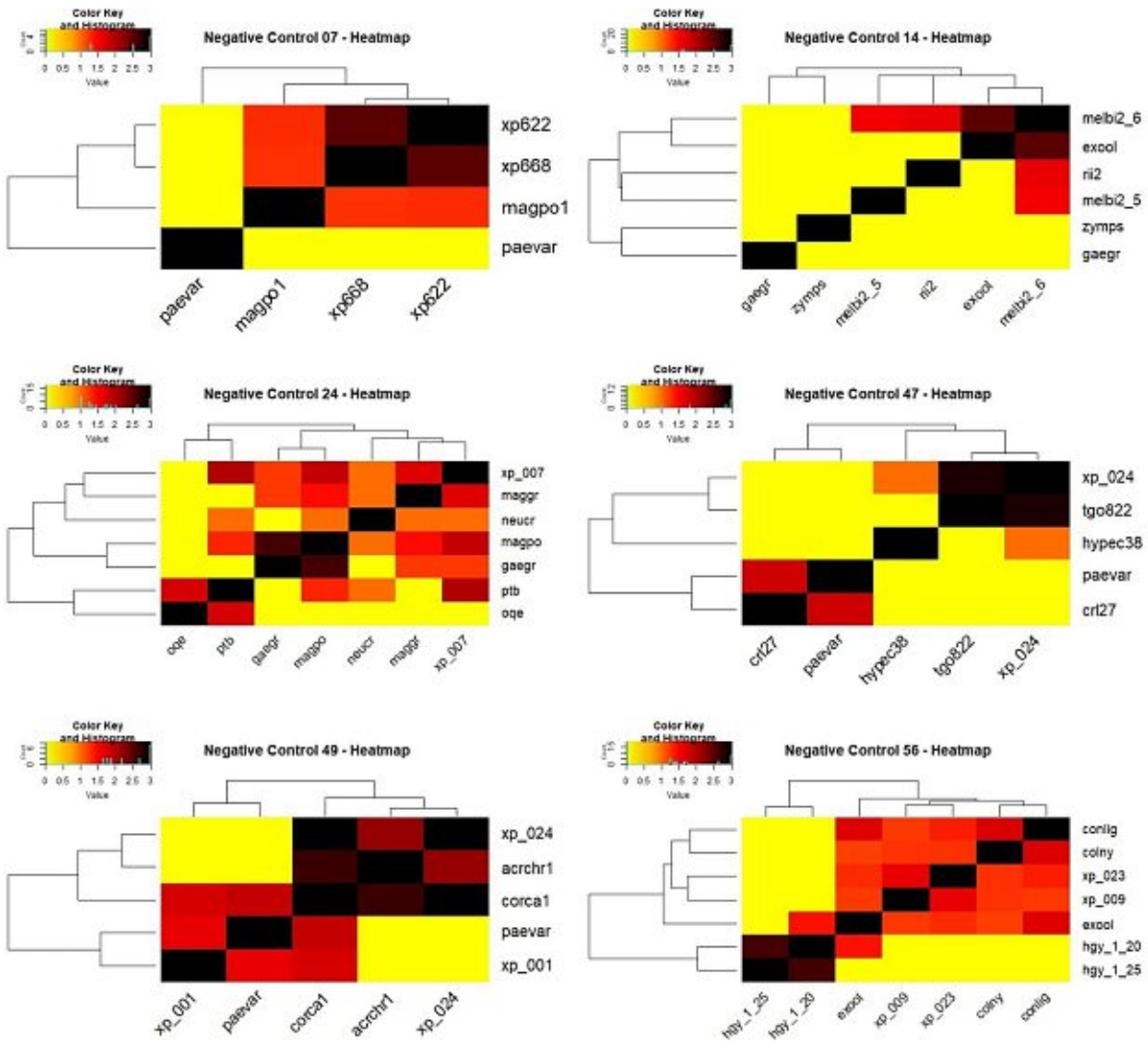


Figure 5.32: Heatmaps of the manual evaluation measures (MEMs) of the negative control 07, 14, 24, 47, 49, and 56, which indeed had moderate average manual evaluation measures (aMEM), but at least one high MEM score equalling or exceeding 2.5. Among them, negative control 49 scored the highest aMEM (1.3), because of two high MEM scores (xp_024:corca1 á 2.67, acrchr1:corca1 á 2.69). The colour key on the upper left side of each heatmap illustrated the MEM value range between 0 and 3, where 3 implied a coevolutionary linkage between two genes and 0 did not. According to the MEM values, dendrograms were assigned automatically to the heatmaps, clustering the genes.

Six random controls resulted in low aMEM scores but had single high MEM scores that equalled or exceeded a value of 2.5. Out of these six controls, negative control 49 yielded the highest aMEM (1.3), because two tree comparisons had a high MEM score of 2.67 and 2.69 (Figure 5.32). All other controls scored aMEMs lower than 1.0. Additionally, control 14 belonged to the controls that scored the lowest aMEM of all random controls (0.4). All that supported the previously assumption, that MEM values were only valid if the aMEM of the cluster was high.

5.4.2 Average manual evaluation measure (aMEM)

One goal of this thesis was to establish whether FunOrder estimates biosynthetic gene clusters correctly. Hence, the previously measured manual evaluation measures (MEMs) were combined to average manual evaluation measures (aMEM). They were used for establishing a receiver operating curve (ROC) and a confusion matrix.

Table 5.6: Overview of the average manual evaluation measures (aMEM) of all controls (negative controls on the left, positives on the right). The threshold for the negative controls were set to 1.5, while for the positive controls it was set to 2.0. False positives and false negatives were highlighted red, whereas true positives and true negatives were highlighted green. "No" referred to the characterized number of the respective biosynthetic gene cluster (BGC) in case of the positive controls, or rather the respective random cluster in the case of the negative controls, which were used for this thesis. The tables were split for a better overview.

Negative Controls					
No	aMEM	No	aMEM	No	aMEM
1	0.41	21	1.15	41	1.15
2	1.15	22	1.07	42	0.49
3	1.33	23	1.21	43	1.02
4	0.55	24	0.92	44	0.61
5	0.86	25	1.03	45	0.70
6	1.17	26	0.57	46	1.11
7	0.85	27	1.05	47	0.56
8	1.38	28	0.84	48	0.90
9	1.21	29	0.99	49	1.31
10	0.83	30	2.08	50	1.49
11	0.96	31	0.84	51	0.68
12	1.00	32	0.46	52	0.73
13	0.38	33	0.68	53	0.69
14	0.38	34	0.50	54	0.67
15	0.51	35	0.71	55	0.74
16	0.84	36	1.09	56	0.86
17	0.90	37	0.53	57	0.88
18	0.94	38	0.98	58	1.34
19	0.87	39	1.13	59	0.70
20	0.68	40	0.44	60	1.50

Positive Controls			
No	aMEM	No	aMEM
1	2.52	21	2.53
2	2.15	22	2.46
3	2.00	23	2.35
4	2.35	24	2.21
5	2.28	25	2.28
6	2.35	26	1.86
7	1.63	27	2.12
8	2.46	28	2.21
9	2.38	29	1.77
10	2.40		
11	1.91		
12	1.66		
13	1.62		
14	1.57		
15	2.25		
16	2.70		
17	2.24		
18	2.06		
19	1.91		
20	2.27		

The threshold for the negative controls were set to 1.5. Because of the high amount of missing data in the positive controls, a higher threshold (2.0) was used for them. All aMEMs were illustrated in Table 5.6. and true negatives (TN) and true positives (TP) were highlighted in green, while the false positives (FP) and false negatives (FN) were highlighted in red.

A boxplot was established to visualize the differences between the aMEMs of the positive and negative controls (Figure 5.33). Only one negative control cluster (negative control 30) overlapped

with the positive controls. Hence, a good distinction between the controls was possible, indicating that FunOrder could discriminate between randomly generated clusters and truly BGCs.

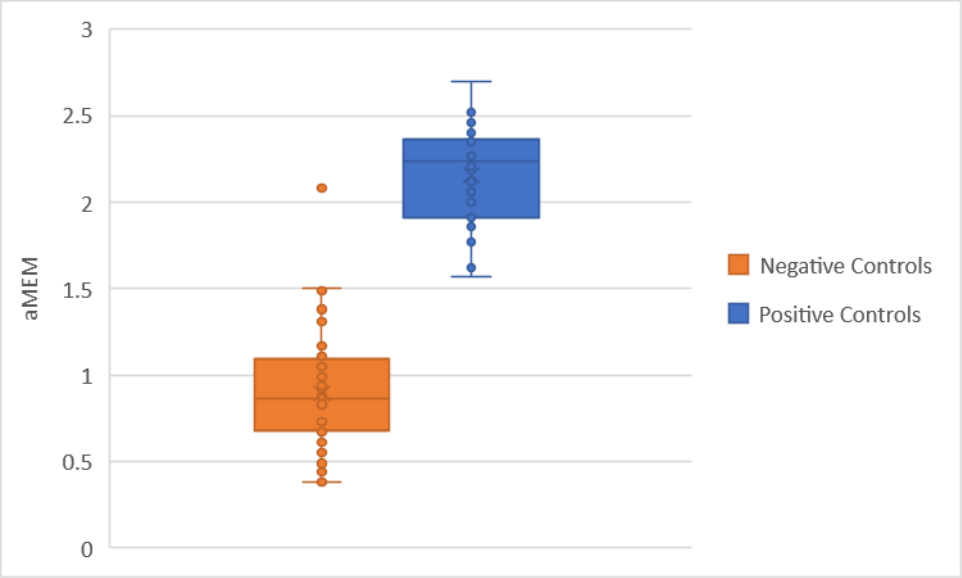


Figure 5.33: Boxplot of all average manual evaluation measures (aMEM) of positive controls (blue) and negative controls (orange) visualizing the distinction between them. True negative controls scored aMEMs below 1.5, while true positive controls had aMEM values above 2.0. The only outlier was negative control 30 whose aMEM overlapped with those of the positive controls.

A confusion matrix was established (Table 5.7) and all calculations concerning sensitivity, selectivity, accuracy, and precision were done (see 4.5. Statistical Evaluation) (Table 5.8).

Table 5.7: Confusion matrix of all average manual evaluation measures (aMEMs). In total 89 controls were used (29 positive and 60 negative controls). The aMEMs of these controls were examined upon their thresholds and defined whether the controls could be declared as positive or negative. In total 21 of 29 positive controls were stated as true positives and 58 of 60 negative controls were stated as true negatives. Meanwhile, 2 of 60 negative clusters were declared as false positives and 8 of 29 positive clusters were declared as false negatives.

			True Condition		Total Population
			Positive	Negative	
	Predicted Condition	Total Population	29	60	89
Predicted Condition	Positive	23	21	2	
	Negative	66	8	58	
	Total Population	89			

As shown in Table 5.8, 21 out of 29 positive BGCs were stated as true positives and 58 out of 60 random clusters were stated as false positives. Based on these values, metrics such as sensitivity, selectivity, accuracy, and precision were calculated to analyse the performance of the evaluation (Table 5.8). The true positive rate, also called sensitivity, was measured according to chapter 4.5. Statistical Evaluation, and scored 72.41%. The true negative rate, also called specificity or selectivity, scored 96,67%, whereas the accuracy, the ratio between the number of correctly classified controls and the total number of controls[176], scored 88,76%. The predictive power of the model was represented by those three metrics: selectivity, sensitivity, and accuracy.[177] The higher these metrics, the better the classification. Furthermore, the precision, also called positive predicted value, was calculated, which specified the proportion of correct positive predictions.[177] A precision score of 91,30% gave therefore good reasons to confide the 23 positive predictions, which was supported by

the F_1 score (80.77%), the mean of precision and sensitivity. The negative predicted value which specified the proportion of correct negative predictions was slightly lower than the precision and scored 87,88%. According to Chicco *et* al., 2020, the accuracy and the F_1 score lacked the ratio between positive (TP, FP) and negative variables (TN, FN). Hence, they can yield in misleading results.[176] As both, negative and positive controls were of interest, the Matthews correlation coefficient was calculated which represented the correlation between true and predicted class (see 4.5. Statistical Evaluation). The higher the Matthews correlation coefficient, also called phi-coefficient, the better the classification.[178] As the coefficient returned a number between -1 and +1, the score of 0.74 indicated a reliable classification.

Hence, the performance metrics pointed to the direction, that FunOrder predicted positive genes correctly and that it was able to differentiate between positive and negative controls.

Table 5.8: Metrics based on the values derived from confusion matrix representing the performance of the classification (performance metrics). The score and its unit of each metric were illustrated. The predictive power of the model was represented by the sensitivity, selectivity, and the accuracy. The higher these values, the better the classification.

Performance metrics	Score	Unit
True positive rate (Sensitivity)	72.41	%
True negative rate (Selectivity)	96.67	%
False Positive rate	3.33	%
False Negative rate	2.76	%
Accuracy	88.76	%
Positive Predicted Value (Precision)	91.30	%
Negative Predicted Value	87.88	%
F_1 Score	80.77	%
Matthew Correlation Coefficient	0.74	-

A manual receiver operating characteristics (ROC) curve illustrated the plotting of the true positive rate over the false positive rate. It was established using Microsoft Excel, receiving an area under the curve (AUC) of 0.857 for two thresholds and an AUC of 0.895 for one threshold (Figure 5.34). Next to the accuracy, the AUC represented the predictive power of the model. The more AUC reached 1.00 the better was the classification.[177] A higher AUC for only one threshold was expectable, because the two thresholds should involve the imprecision of the approximated values. However, the AUC for two thresholds were still high, indicating that FunOrder differentiated between positive and negative controls and furthermore, that it predicted positive genes correctly.

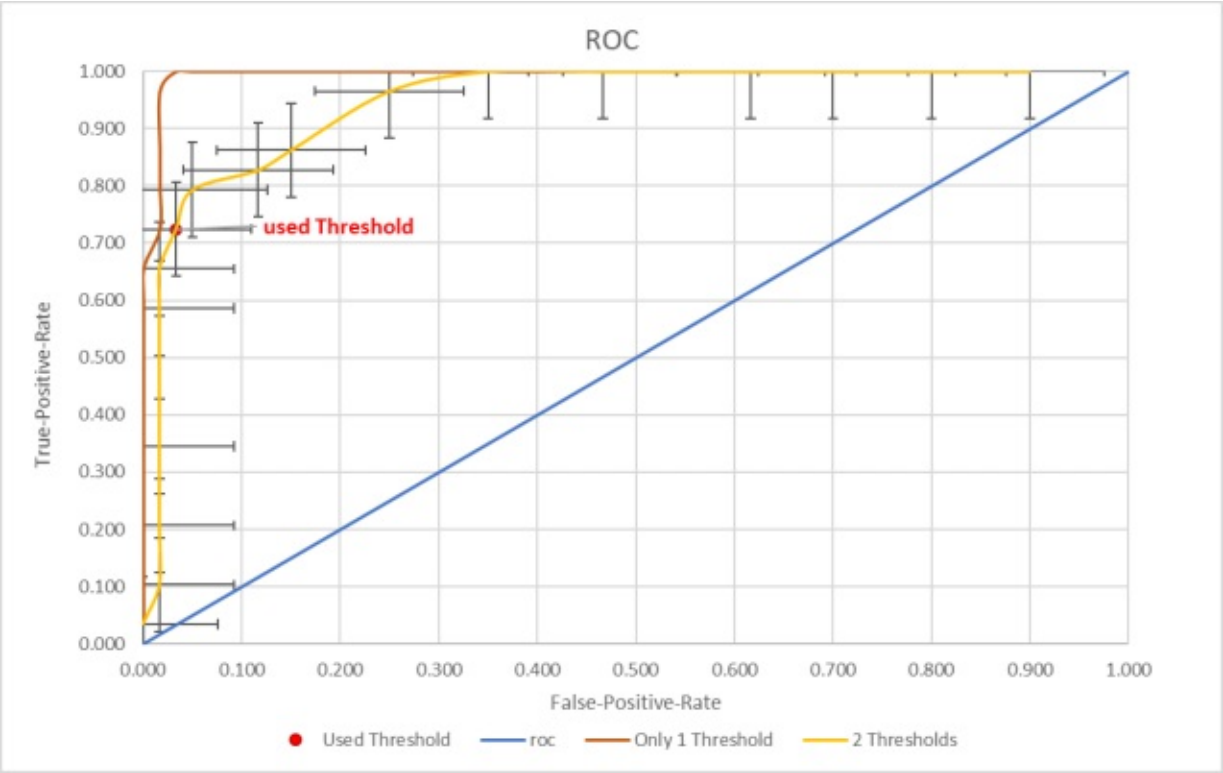


Figure 5.34: Receiver operating curve (ROC) of the average manual evaluation measures (aMEM). The red dot represented the used threshold of 1.5 for negative und 2.0 for positive controls. The area under the curve (AUC) was 0.86 for the yellow, and 0.90 for the red curve. The more AUC reached 1.00 the better was the classification.

5.4.3 PLS DA of raw data

To compare the previous results a partial least discriminant analysis of the raw data was established using DataLab, Version 4.0. In this connection, the branch length differences, node differences, branch colours and topologies were used for the classification. The result of the PLS DA was illustrated in Figure 5.35, resulting in an AUC of 0.83, supporting the analysis presented in the previous chapters. The sensitivity was even higher than in the manually calculated confusion matrix, but the False Positive Rate (FPR or FP) was higher as well.

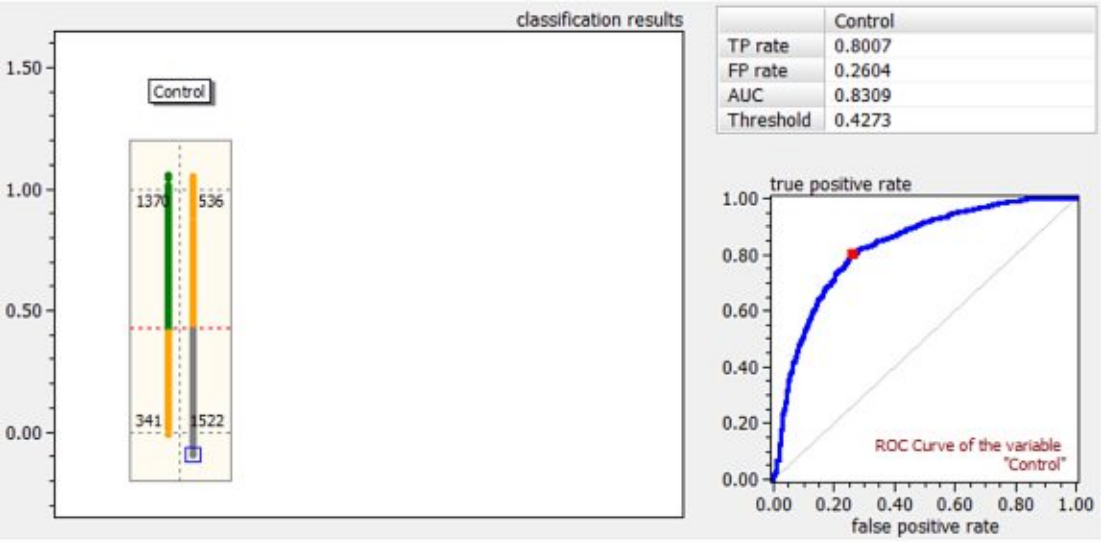


Figure 5.35: Partial least square discriminant analysis (PLS-DA) results of raw data (branch length difference, node difference, branch colour, topology), performed in DataLab, Version 4.0. As shown, the area under the curve (AUC) in the PLS-DA scored 0.83, and the true positive (TP) rate and the false positive rate (FP) were both higher than in the manually conducted confusion matrix.

The loadings of the predictors were shown in Figure 5.36, illustrating their input on the classification. According to that, the values of the predictors colour and topology contributed most to the PLS DA and the classification, while the distances and nodes did not. As the colour was a representation of the tree topology based on the description of the online tree visualization tool phylo.io[96] (see 4.4. Tree comparison), the main predictor for inferring coevolutionary linkage was therefore the topology of the phylogenetic trees.

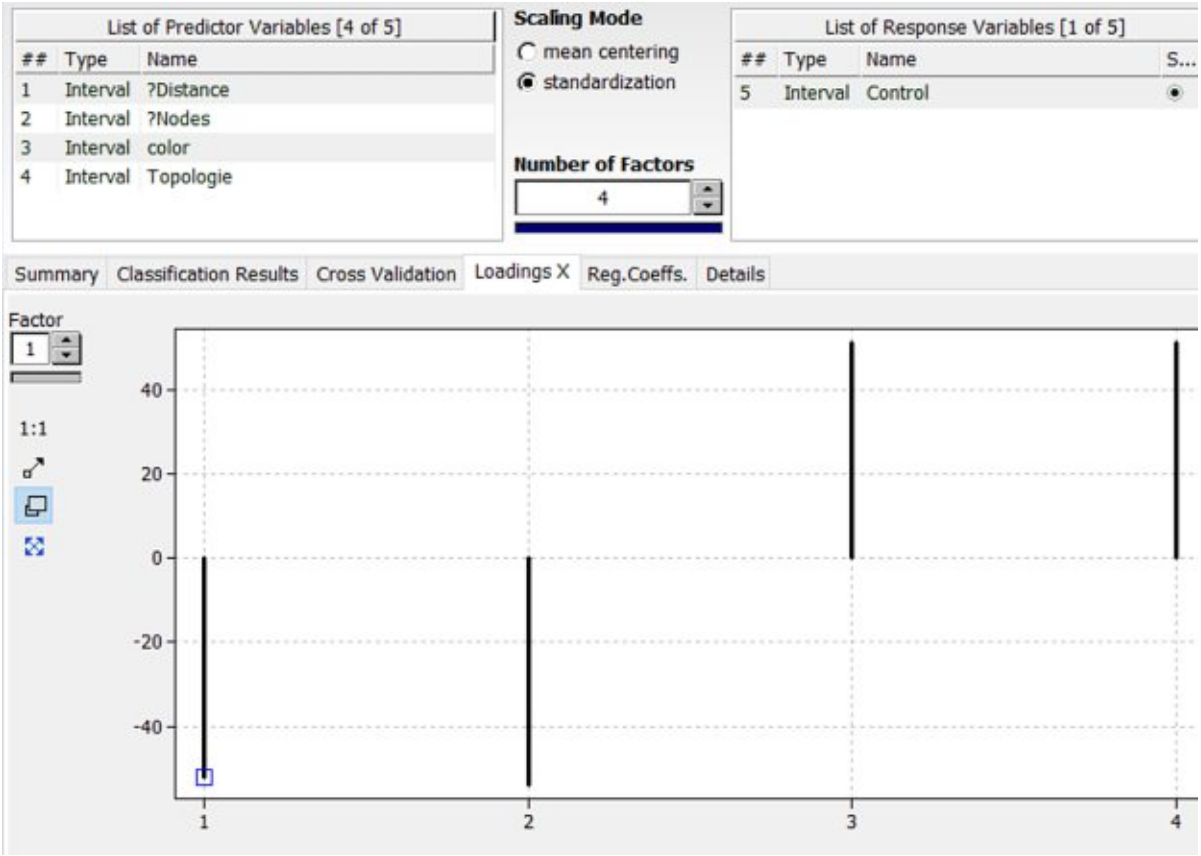


Figure 5.36: Loadings of the predictors in partial least discriminant analysis (PLS DA), showing which predictor has which input on the classification. Hence, the predictor 3 (colour) and 4 (topology) contributed most to the PLS DA.

6. Discussion

The main goal during this thesis was to infer, whether FunOrder had the ability to predict genes involved in the biosynthesis of a compound and hence, can be used as a genome mining tool for novel biosynthetic gene clusters (BGC). To answer this question, a set of 102 negative, and 29 positive controls were used. 42 out of the 102 negative controls consisted of completely in-silico created clusters, called synthetic BGC, while the remaining 60 clusters were created by concatenating random fungal gene sequences, called random BGCs. Regarding the positive controls, the selected BGCs showed a great diversity among the fungal taxa (see Table 5.2). The products of these clusters represented a high variety, too. They included antibiotics (e.g., penicillin and mycophenolic acid), other pharmaceuticals (e.g., lovastatin and fumagillin), pigments (e.g., tenellin and sorbicillin), and mycotoxins (e.g., aflatoxin and fumonisin), among others. This provided a broad compilation of the 29 different, positive controls.

The controls were analysed by FunOrder, the 42 synthetic BGCs yielded no phylogenetic tree, while for each of the random controls trees were obtained. Therefore, it was concluded that FunOrder can distinguish between completely randomized negative controls (synthetic BGCs) and controls based on real sequences. Regarding the random clusters, the number of the established phylogenetic trees ranged between two trees per cluster (Negative control 60) and eight trees per cluster (Negative controls 25, 31, 46, and 57). The calculated heatmaps of the negative controls showed generally low MEMs indicating no coevolutionary linkages, as expected. Examining the positive controls revealed the absence of 9 phylogenetic trees in five controls (see Table 5.3). Except of one gene (*ustP1* in positive control 24), these genes were not involved in the biosynthetic pathway, which supported the postulation of Demerec and Hartman (1959) that cluster genes were coevolutionary linked and genes outside the clusters were not.[83] However, while all negative controls were compared, genes necessary for the biosynthetic production, only, were analysed in the positive controls. 910 phylogenetic trees of in total 2588 were compared during the thesis. This biased the further statistical analysis, preventing a reliable statement upon the question whether FunOrder can distinguish between core genes and adjacent genes. The positive controls showed therefore a great amount of missing data (> 20%) and thus were approximated it by a random forest approach. The used MissForest algorithm calculated the missing values based on the available ones. A certain analysis upon their evolutionary linkages was therefore not provided. Furthermore, the comparisons between heatmaps with missing and with approximated data illustrated that the random forest approach overestimated these values, as all of the predicted values were higher than the original ones. In fact, this approach considers only a randomly drawn subset of variables to predict a missing one, called out-of-bag (OOB) observations.[179] These OOBs can be used to estimate the prediction error of the random forest approach, namely the OOB-error, which overestimates the true-prediction error in two-class-problems, especially with few observations like in the used positive controls.[179] This led to the conclusion, that the approximated MEMs of the positive controls must be observed with suspicion. Nevertheless, the coevolutionary linkages of the genes were analysed using the established heatmaps, dendrograms, and principal component analysis (PCA). During the evaluation of the heatmaps and dendrograms, the examination of the overestimated data was de-emphasized and instead the plots with missing data were reviewed. The PCA results, which were calculated using the approximated data, showed that the earlier missing genes were clustering together, while the evaluated genes mostly diverged, other than actually assumed (e.g., *penDE*, *pcbAB* and *pcbC* in the penicillin BGC, Figure 5.19). This derived from the overestimation by the random forest approach and can be seen as an artifact. On that basis, evaluated genes that in fact clustered together in the PCA scores must be regarded as highly coevolutionary related (e.g., *fum1* and *fum16* in the fumonisin BGC, Figure 5.15). All the results (see 5.4.1 Manual evaluation measure) pointed to the direction,

that FunOrder has the ability to predict positive genes correctly. Still, the previously explained bias could not be eliminated. To overcome that bias, a complete evaluation and the comparison of 1678 more phylogenetic trees would have been needed.

Another question to answer was whether FunOrder can distinguish between positive and negative controls. Hence, the MEMs were averaged, yielding the so-called average manual evaluation measures (aMEM). To ensure a strict differentiation, the bias was de-emphasized by using two thresholds for aMEMs, which was set to 1.5 and lower for negative controls, and 2.0 and above for positive controls. The discrimination was shown in a boxplot (Figure 5.33) indicating a good distinction. However, two negative controls yielded high aMEMs (see Table 5.6) and were considered as false positives. When examining the amount of obtained phylogenetic trees, both controls contained only two (negative control 60) or rather three (negative control 30) trees, indicating that the correct identification of small BGCs by FunOrder had a likelihood of 30:1. After establishing a confusion matrix based on the aMEMs (Table 5.7) the performance metrics were calculated. Best classification results tend to 100% regarding sensitivity, selectivity, accuracy, and precision, respectively, while the false positive rate and the false negative rate approach 0 %. Their values (see Table 5.8) gave therefore reason to confide the classification. However, as these metrics did not include the ratio between positive and negative variables yielding probably misleading results[176], the Matthews correlation coefficient was determined (see Table 5.8). The coefficient scored 0.74 indicating a good classification of the negative and positive controls. Based on the confusion matrix, a ROC curve was established (see Figure 5.34) and the AUC was calculated, which score 1.00 for best models. In this thesis the AUC was calculated for two thresholds to overcome the bias which derived from the comparison of the positive controls. Despite two thresholds, the AUC scored 0.86. All these results indicated that FunOrder can distinguish between positive and negative controls. Furthermore, the classification results point to the direction that the discrimination is robust.

To confirm these results, a partial least discriminant analysis (PLS DA) was established using the parameters of the raw data (branch length differences, node differences, branch colours, and topologies). According to Figure 5.35, the PLS DA was comparable to the manually established classification results. In fact, PLS DA scored an even higher sensitivity. Thus, the analysis therefore confirmed the previously illustrated results. Furthermore, PLS DA clarified that the topologies and the branch colours, which were representations of the topologies, had the most input to the classification, indicating they were sufficient for inferring coevolutionary linkages (Figure 5.36).

Finally, the evaluation results were used to range the distances calculated by FunOrder using the TreeKO algorithm. An additional BGC (fusaric acid BGC from *Fusarium fujikoroi*) was automatically analysed by FunOrder itself and this time the distances calculated by TreeKO algorithm were used for the evaluation of the BGC. The obtained results verified the evolutionary connection between the cluster genes (see Supplement 9.63), for which examples are introduced in the following. According to the respective literature[159] one of the core genes of the BGC (*fub8*) encoding a non-ribosomal peptide synthetase should have high coevolutionary linkages with the gene *fub6*, a NAD(P)-dependent dehydrogenase. Based on the biosynthetic pathway, *fub8* binds O-acetyl-L homoserine yielding an NRPS-bound intermediate which is further converted by *fub6*.[159] The second core gene (*fub1*) encoding a polyketide synthase produces trans-2-hexenal, which is released by *fub4*, a homoserine O-acetyltransferase. This indicates that they share evolutionary traits. These assumptions were confirmed by the results (see Supplement 9.63). Studt *et* al. proposed that fub1, fub8, and fub4 form an enzyme complex.[159] In fact, the distances between *fub1* and *fub8*, as well as *fub4* and *fub8* indicated some evolutionary linkages. Further important genes are *fub12* which is involved in the production of two fusaric acid derivatives, and *fub11*, which exports fusaric acid out of the cell.[159] Their evolutionary linkages were confirmed by the results as well (see Supplement

9.63). According to these results, FunOrder is able to provide information upon the coevolutionary linkage of the investigated genes. This indicates that the program can contribute to the discovery of novel BGCs.

Taken together, the data and all the results point to the direction that FunOrder has the ability to predict fungal cluster genes correctly and that it differentiates between real biosynthetic gene clusters (BGC) and randomly assembled, negative controls. Furthermore, they indicate that the differentiation done by FunOrder is robust. However, it remains unclear whether FunOrder can distinguish between biosynthetic genes needed for the biosynthesis of a secondary metabolite and adjacent ones, as not all established phylogenetic trees of the positive controls were evaluated.

7. Conclusion

Fungal cluster b**order** prediction based on computational coevolution (FunOrder) is able to predict genes involved in the biosynthesis of a secondary metabolite by means of coevolutionary linkages. The approach works with common Genbank® files as input providing a simple handling. The program is based on sequence similarity searches using BLAST, multiple sequence alignments by means of emma, an EMBOSS approach, RAxML, and TreeKO algorithm yielding evaluation and strict distances. FunOrder offers the possibility of inferring novel fungal biosynthetic gene clusters (BGC) based on their coevolutionary linkage, similar to EvoMining, a genome mining tool for bacteria and archaea species.[44] Thereby, FunOrder is a promising genome mining tool for unknown fungal BGCs based on phylogenetics and – to the writer's knowledge – the first suchlike approach for fungal species. Furthermore, it is an auspicious approach for the ongoing research of novel fungal secondary metabolites that can be used as pharmaceuticals.

8. References

1 Keller, N. P., Turner, G. & Bennett, J. W. Fungal secondary metabolism - from biochemistry to genomics. *Nat Rev Microbiol* **3**, 937-947, doi:10.1038/nrmicro1286 (2005).

2 Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *British journal of experimental pathology* **10**, 226-236 (1929).

3 Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336-343, doi:10.1038/nature17042 (2016).

4 Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes - a review. *Nat Prod Rep* **33**, 988-1005, doi:10.1039/c6np00025h (2016).

5 Bo, G. Giuseppe Brotzu and the discovery of cephalosporins. *Clin Microbiol Infect* **6**, 6-9, doi:10.1111/j.1469-0691.2000.tb02032.x (2000).

6 Brakhage, A. A. & Schroeckh, V. Fungal secondary metabolites - strategies to activate silent gene clusters. *Fungal Genet Biol* **48**, 15-22, doi:10.1016/j.fgb.2010.04.004 (2011).

7 Ramawat, K. G. & Goyal, S. in *Co-Evolution of Secondary Metabolites* (eds Jean-Michel Mérillon & Kishan Gopal Ramawat) 3-17 (Springer International Publishing, 2020).

8 Künzler, M. How fungi defend themselves against microbial competitors and animal predators. *PLOS Pathogens* **14**, e1007184, doi:10.1371/journal.ppat.1007184 (2018).

9 Chavali, A. K. & Rhee, S. Y. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. *Brief Bioinform* **19**, 1022-1034, doi:10.1093/bib/bbx020 (2018).

10 Brakhage, A. A. Regulation of fungal secondary metabolism. *Nat Rev Microbiol* **11**, 21-32, doi:10.1038/nrmicro2916 (2013).

11 Hautbergue, T., Jamin, E. L., Debrauwer, L., Puel, O. & Oswald, I. P. From genomics to metabolomics, moving toward an integrated strategy for the discovery of fungal secondary metabolites. *Nat Prod Rep* **35**, 147-173, doi:10.1039/c7np00032d (2018).

12 Challis, G. L. Genome Mining for Novel Natural Product Discovery. *J Med Chem* **51**, 2618-2628, doi:10.1021/jm700948z (2008).

13 Rutledge, P. J. & Challis, G. L. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat Rev Microbiol* **13**, 509-523, doi:10.1038/nrmicro3496 (2015).

14 Livermore, D. M. The need for new antibiotics. *Clin Microbiol Infect* **10**, 1-9 (2004).

15 Hertweck, C. Hidden biosynthetic treasures brought to light. *Nat Chem Biol* **5**, 450-452, doi:10.1038/nchembio0709-450 (2009).

16 Pfannenstiel, B. T. & Keller, N. P. On top of biosynthetic gene clusters: How epigenetic machinery influences secondary metabolism in fungi. *Biotechnol Adv* **37**, 107345, doi:10.1016/j.biotechadv.2019.02.001 (2019).

17 Starke, R., Capek, P., Morais, D., Callister, S. J. & Jehmlich, N. The total microbiome functions in bacteria and fungi. *J Proteomics* **213**, 103623, doi:10.1016/j.jprot.2019.103623 (2020).

18 Tsukada, K. *et al.* Synthetic biology based construction of biological activity-related library of fungal decalin-containing diterpenoid pyrones. *Nat Commun* **11**, 1830, doi:10.1038/s41467-020-15664-4 (2020).

19 Mohanta, T. K. & Bae, H. The diversity of fungal genome. *Biol Proced Online* **17**, 8, doi:10.1186/s12575-015-0020-z (2015).

20 Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol* **11**, 625-631, doi:10.1038/nchembio.1890 (2015).

21 Palmer, J. M. & Keller, N. P. Secondary metabolism in fungi: does chromosomal location matter? *Curr Opin Microbiol* **13**, 431-436, doi:10.1016/j.mib.2010.04.008 (2010).

22 Wolf, T., Shelest, V., Nath, N. & Shelest, E. CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics* **32**, 1138-1143, doi:10.1093/bioinformatics/btv713 (2016).

23 Blin, K., Kim, H. U., Medema, M. H. & Weber, T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief Bioinform* **20**, 1103-1113, doi:10.1093/bib/bbx146 (2019).

24 Swadha, A. & Debasisa, M. in *Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications* (eds Liu Limin Angela, Wei Dongqing, Li Yixue, & Lei Huimin) 380-405 (IGI Global, 2011).

25 Cummings, M., Breitling, R. & Takano, E. Steps towards the synthetic biology of polyketide biosynthesis. *FEMS Microbiol Lett* **351**, 116-125, doi:10.1111/1574-6968.12365 (2014).

26 Nivina, A., Yuet, K. P., Hsu, J. & Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases. *Chem Rev* **119**, 12524-12547, doi:10.1021/acs.chemrev.9b00525 (2019).

27 Challis, G. L. & Naismith, J. H. Structural aspects of non-ribosomal peptide biosynthesis. *Curr Opin Struct Biol* **14**, 748-756, doi:10.1016/j.sbi.2004.10.005 (2004).

28 EMBL-EBI. *Aromatic prenyltransferase DMATS-type, fungi*, <https://www.ebi.ac.uk/interpro/entry/InterPro/IPR012148/> (2017).

29 Schmidt-Dannert, C. Biosynthesis of terpenoid natural products in fungi. *Adv Biochem Eng Biotechnol* **148**, 19-61, doi:10.1007/10_2014_283 (2015).

30 Helfrich, E. J. N., Lin, G. M., Voigt, C. A. & Clardy, J. Bacterial terpene biosynthesis: challenges and opportunities for pathway engineering. *Beilstein J Org Chem* **15**, 2889-2906, doi:10.3762/bjoc.15.283 (2019).

31 van der Lee, T. A. J. & Medema, M. H. Computational strategies for genome-based natural product discovery and engineering in fungi. *Fungal Genet Biol* **89**, 29-36, doi:10.1016/j.fgb.2016.01.006 (2016).

32 Hallen, H. E., Luo, H., Scott-Craig, J. S. & Walton, J. D. Gene family encoding the major toxins of lethal Amanita mushrooms. *PNAS* **104**, 19097-19101 (2007).

33 van der Velden, N. S. *et al.* Autocatalytic backbone N-methylation in a family of ribosomal peptide natural products. *Nat Chem Biol* **13**, doi:10.1038/nchembio.2393 (2017).

34 Umemura, M. *et al.* Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in Aspergillus flavus. *Fungal Genet Biol* **68**, 23-30, doi:10.1016/j.fgb.2014.04.011 (2014).

35 Johnson, R. D. *et al.* A novel family of cyclic oligopeptides derived from ribosomal peptide synthesis of an in planta-induced gene, gigA, in Epichloe endophytes of grasses. *Fungal Genet Biol* **85**, 14-24, doi:10.1016/j.fgb.2015.10.005 (2015).

36 Vogt, E. & Kunzler, M. Discovery of novel fungal RiPP biosynthetic pathways and their application for the development of peptide therapeutics. *Appl Microbiol Biotechnol* **103**, 5567-5581, doi:10.1007/s00253-019-09893-x (2019).

37 Vignolle, G. A., Mach, R. L., Mach-Aigner, A. R. & Derntl, C. Novel approach in whole genome mining and transcriptome analysis reveal conserved RiPPs in Trichoderma spp. *BMC Genomics* **21**, 258, doi:10.1186/s12864-020-6653-6 (2020).

38 Corre, C. & Challis, G. L. in *Comprehensive Natural Products II: Chemistry and Biology* Vol. 2 (eds Lewis N. Mander & Hung-wen Liu) 429-453 (Elsevier, 2010).

39 Kjaerbolling, I., Mortensen, U. H., Vesth, T. & Andersen, M. R. Strategies to establish the link between biosynthetic gene clusters and secondary metabolites. *Fungal Genet Biol* **130**, 107-121, doi:10.1016/j.fgb.2019.06.001 (2019).

40 (NCBI), N. C. f. B. I. *GenBank Overview*, <https://www.ncbi.nlm.nih.gov/genbank/> (2013).

41 Blin, K. *et al.* antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* **45**, W36-W41, doi:10.1093/nar/gkx319 (2017).

42 Russell, A. H. & Truman, A. W. Genome mining strategies for ribosomally synthesised and post-translationally modified peptides. *Comput Struct Biotechnol J* **18**, 1838-1851, doi:10.1016/j.csbj.2020.06.032 (2020).

43 Weber, T. & Kim, H. U. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synth Syst Biotechnol* **1**, 69-79 (2016). <http://www.secondarymetabolites.org/mining/>.

44 Selem-Mojica, N., Aguilar, C., Gutierrez-Garcia, K., Martinez-Guerrero, C. E. & Barona-Gomez, F. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microb Genom* **5**, doi:10.1099/mgen.0.000260 (2019).

45 Benson, D. A. *et al.* GenBank. *Nucleic Acids Res* **41**, D36-42, doi:10.1093/nar/gks1195 (2013).

46 Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res* **48**, D454-D458, doi:10.1093/nar/gkz882 (2020).

47 Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* **29**, 415-420, doi:10.1038/nbt.1823 (2011).

48 Choudhuri, S. in *Bioinformatics for Beginners* 27-53 (2014).

49 Choudhuri, S. in *Bioinformatics for Beginners* 209-218 (2014).

50 Bawono, P. & Heringa, J. in *Comprehensive Biomedical Physics* 93-110 (2014).

51 Velasco, J. Universal common ancestry, LUCA, and the Tree of Life: three distinct hypotheses about the evolution of life. *Biology & Philosophy* **33**, 31, doi:10.1007/s10539-018-9641-3 (2018).

52 Touchman, J. Comparative Genomics. *Nature Education Knowledge* **3** (2010).

53 Ciccarelli, F. D. *et al. Interactive Tree Of Life*, <https://itol.embl.de> (2006).

54 Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283-1287, doi:10.1126/science.1123061 (2006).

55 Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* **21**, 428-444, doi:10.1038/s41576-020-0233-0 (2020).

56 Scott, A. D. & Baum, D. A. in *Encyclopedia of Evolutionary Biology* 270-276 (2016).

57 Hayati, M. & Chindelevitch, L. Computing the distribution of the Robinson-Foulds distance. *Comput Biol Chem* **87**, 107284, doi:10.1016/j.compbiolchem.2020.107284 (2020).

58 Stavrinides, J. & Ochman, H. in *Encyclopedia of Microbiology* (ed Moselio Schaechter) (2009).

59 Chatzou, M. *et al.* Multiple sequence alignment modeling: methods and applications. *Brief Bioinform* **17**, 1009-1023, doi:10.1093/bib/bbv099 (2016).

60 Naveed, T., Siddiqui, I. S. & Ahmed, S.

61 Contributors, W. *Sequence Alignment*, <https://en.wikipedia.org/wiki/Sequence_alignment#Alignment_methods> (2020).

62 Needleman, S. B. & Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J Mol Bio* **48**, 443-453 (1970).

63 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J Mol Bio* **215**, 403-410 (1990).

64 Khajeh-Saeed, A., Poole, S. & Blair Perot, J. Acceleration of the Smith–Waterman algorithm using single and multiple graphics processors. *Journal of Computational Physics* **229**, 4247-4258, doi:10.1016/j.jcp.2010.02.009 (2010).

65 McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* **32**, W20-25, doi:10.1093/nar/gkh435 (2004).

66 Holder, M. & Lewis, P. O. Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet* **4**, 275-284, doi:10.1038/nrg1044 (2003).

67 Lio, P. & Goldman, N. Models of Molecular Evolution and Phylogeny. *Genome Res* **8**, 1233-1244, doi:10.1101/gr.8.12.1233 (1998).

68 Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. in *Atlas of Protein Sequence and Structure* Vol. 5 345-352 (Dayhoff, M. O. , 1978).

69 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *PNAS* **89**, 10915-10919 (1992).

70 Rizzo, J. & Rouchka, E. Review of Phylogenetic Tree Construction. (2007).

71 Godini, R. & Fallahi, H. A brief overview of the concepts, methods and computational tools used in phylogenetic tree construction and gene prediction. *Meta Gene* **21**, doi:10.1016/j.mgene.2019.100586 (2019).

72 Akhtar, M. S. & Alaraidh, I. A. *Essentials of Bioinformatics, Volume III*. Vol. III (Springer Nature Switzerland AG, 2019).
73 Felsenstein, J. Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *J Mol Evo* **17**, 363-376 (1981).
74 Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456-463, doi:10.1093/bioinformatics/bti191 (2005).
75 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313, doi:10.1093/bioinformatics/btu033 (2014).
76 Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gab penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680 (1994).
77 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).
78 Higgins, D. G. & Sharp, P. M. **2020** (1988).
79 Faller, M. *emma*, <http://www.sacs.ucsf.edu/Documentation/emboss/emma.html > (1999).
80 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277, doi:10.1016/s0168-9525(00)02024-2 (2000).
81 Holmes, S. in *Statistical Science* Vol. 18 241-255 (Institute of Mathematical Statistics, 2003).
82 Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 433-459, doi:10.1002/wics.101 (2010).
83 Demerec, M. & Hartman, P. E. Complex Loci in Microorganisms. *Annual Review of Microbiology* **13**, 377-406, doi:10.1146/annurev.mi.13.100159.002113 (1959).
84 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. *Curr Genet* **26**, 120-125 (1994).
85 Marcet-Houben, M. & Gabaldon, T. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Research* **39**, e66, doi:10.1093/nar/gkr087 (2011).
86 contributors, W. *Dependent and independent variables*, <https://en.wikipedia.org/wiki/Dependent_and_independent_variables> (2020).
87 Lohninger, H. *Grundlagen der Statistik*, <http://www.statistics4u.info/fundstat_germ/cc_multivar_stat.html> (2005).
88 Belle, G. v., Fisher, L. D., Heagerty, P. J. & Lumley, T. *Biostatistics: A Methodology for the Health Sciences*. 2 edn, (John Wiley & Sons, Inc., 2004).
89 Stekhoven, D. J. & Buhlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112-118, doi:10.1093/bioinformatics/btr597 (2012).
90 Nelson, P. R. C., Taylor, P. A. & MacGregor, J. F. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems* **35**, 45-65 (1996).
91 Lee, L. C., Liong, C. Y. & Jemain, A. A. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst* **143**, 3526-3539, doi:10.1039/c8an00599k (2018).
92 McKay, C. & Rocap, G. *Convert Genbank or EMBL files to FASTA*, <https://rocaplab.ocean.washington.edu/tools/genbank_to_fasta/> (
93 Rice, P., Bleasby, A., Ison, J., Mullan, L. & Bottu, G. *EMBOSS Users Guide*, <http://emboss.open-bio.org/html/use/index.html> (2009).
94 Camacho, C., Madden, T. L., Tao, T., Agarwala, R. & Morgulis, A. *BLAST Command Line Applications User Manual*, <https://www.ncbi.nlm.nih.gov/books/NBK537770/> (2008).
95 Faller, M. *Programm emma*, <http://www.csd.hku.hk/bruhk/emboss/emma.html> (1999).

96 Robinson, O., Dylus, D. & Dessimoz, C. Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. *Mol Biol Evol* **33**, 2163-2166, doi:10.1093/molbev/msw080 (2016).
97 Warnes, G. R. *et al. Various R Programming Tools for Plotting Data*, <https://cran.r-project.org/web/packages/gplots/gplots.pdf > (2020).
98 Murtagh, F. & Legendre, P. Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. *Journal of Classification* **31**, 274-295, doi:10.1007/s00357-014-9161-z (2014).
99 Kucheryavskiy, S. mdatools – R package for chemometrics. *Chemometrics and Intelligent Laboratory Systems* **198**, 103937, doi:https://doi.org/10.1016/j.chemolab.2020.103937 (2020).
100 Powers, D. M. W. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* **2**, 37-63 (2011).
101 Lohninger, H. *DataLab - A program for statistical analysis*, <http://datalab.epina.at/de_home.html > (2000).
102 Lim, F. Y. *et al.* Fungal Isocyanide Synthases and Xanthocillin Biosynthesis in Aspergillus fumigatus. *mBio* **9**, doi:10.1128/mBio.00785-18 (2018).
103 Lin, H. C. *et al.* The fumagillin biosynthetic gene cluster in Aspergillus fumigatus encodes a cryptic terpene cyclase involved in the formation of beta-trans-bergamotene. *J Am Chem Soc* **135**, 4616-4619, doi:10.1021/ja312503y (2013).
104 Savitha, J., Bhargavi, S. D. & Praveen, V. K. Complete Genome Sequence of Soil Fungus Aspergillus terreus (KM017963), a Potent Lovastatin Producer. *Genome Announc* **4**, doi:10.1128/genomeA.00491-16 (2016).
105 Guo, C. J. & Wang, C. C. Recent advances in genome mining of secondary metabolites in Aspergillus terreus. *Front Microbiol* **5**, 717, doi:10.3389/fmicb.2014.00717 (2014).
106 Cluster, M. I. a. a. B. G. *BGC0000334: cyclosporine biosynthetic gene cluster from Tolypocladium inflatum NRRL8044*, <https://mibig.secondarymetabolites.org/repository/BGC0000334/index.html#r1c1 > (2020).
107 Yang, X. *et al.* Cyclosporine Biosynthesis in Tolypocladium inflatum Benefits Fungal Adaptation to the Environment. *ASM* **9**, doi:10.1128/mBio (2018).
108 Resource, T. U. P. *UniProtKB - Q09164_TOLIN: Cyclosporine synthetase*, <https://www.uniprot.org/uniprot/Q09164> (1996).
109 cluster, M. I. a. a. B. G. *BGC0001565: cyclosporin C biosynthetic gene cluster from Beauveria felina*, <https://mibig.secondarymetabolites.org/repository/BGC0001565/index.html#r1c1 > (2020).
110 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* **268**, 352-361, doi:10.1007/s00438-002-0755-5 (2002).
111 Abe, Y. *et al.* Molecular cloning and characterization of an ML-236B (compactin) biosynthetic gene cluster in Penicillium citrinum. *Mol Genet Genomics* **267**, 636-646, doi:10.1007/s00438-002-0697-y (2002).
112 Barrios-Gonzalez, J., Banos, J. G., Covarrubias, A. A. & Garay-Arroyo, A. Lovastatin biosynthetic genes of Aspergillus terreus are expressed differentially in solid-state and in liquid submerged fermentation. *Appl Microbiol Biotechnol* **79**, 179-186, doi:10.1007/s00253-008-1409-2 (2008).
113 Bhatnagar, D., Cary, J. W., Ehrlich, K., Yu, J. & Cleveland, T. E. Understanding the genetics of regulation of aflatoxin production and Aspergillus flavus development. *Mycopathologia* **162**, 155-166, doi:10.1007/s11046-006-0050-9 (2006).
114 Bhatnagar, D., Ehrlich, K. C. & Cleveland, T. E. Molecular genetic analysis and regulation of aflatoxin biosynthesis. *Appl Microbiol Biotechnol* **61**, 83-93, doi:10.1007/s00253-002-1199-x (2003).

115 Brown, D. W., Butchko, R. A., Busman, M. & Proctor, R. H. The Fusarium verticillioides FUM gene cluster encodes a Zn(II)2Cys6 protein that affects FUM gene expression and fumonisin production. *Eukaryot Cell* **6**, 1210-1218, doi:10.1128/EC.00400-06 (2007).
116 Bushley, K. E. *et al.* The genome of tolypocladium inflatum: evolution, organization, and expression of the cyclosporin biosynthetic gene cluster. *PLoS Genet* **9**, e1003496, doi:10.1371/journal.pgen.1003496 (2013).
117 Butchko, R. A., Plattner, R. D. & Proctor, R. H. FUM13 Encodes a Short Chain Dehydrogenase/Reductase Required for C-3 Carbonyl Reduction during Fumonisin Biosynthesis in Gibberella moniliformis. *J Agric Food Chem* **51**, 3000-3006 (2003).
118 Butchko, R. A., Plattner, R. D. & Proctor, R. H. FUM9 is required for C-5 hydroxylation of fumonisins and complements the meitotically defined Fum3 locus in Gibberella moniliformis. *Appl Environ Microbiol* **69**, 6935-6937, doi:10.1128/aem.69.11.6935-6937.2003 (2003).
119 Butchko, R. A., Plattner, R. D. & Proctor, R. H. Deletion Analysis of FUM Genes Involved in Tricarballylic Ester Formation during Fumonisin Biosynthesis. *J Agric Food Chem* **54**, 9398-9404 (2006).
120 Cary, J. W. *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* **81**, 88-97, doi:10.1016/j.fgb.2015.05.010 (2015).
121 Chen, L. *et al.* Engineering of New Pneumocandin Side-Chain Analogues from Glarea lozoyensis by Mutasynthesis and Evaluation of Their Antifungal Activity. *ACS Chem Biol* **11**, 2724-2733, doi:10.1021/acschembio.6b00604 (2016).
122 Chen, L. *et al.* Engineering of Glarea lozoyensis for exclusive production of the pneumocandin B0 precursor of the antifungal drug caspofungin acetate. *Appl Environ Microbiol* **81**, 1550-1558, doi:10.1128/AEM.03256-14 (2015).
123 Chen, L. *et al.* Genomics-driven discovery of the pneumocandin biosynthetic gene cluster in the fungus Glarea lozoyensis. *BMC Genomics* **14**, 339, doi:10.1186/1471-2164-14-339 (2013).
124 Del-Cid, A. *et al.* Identification and Functional Analysis of the Mycophenolic Acid Gene Cluster of Penicillium roqueforti. *PLoS One* **11**, e0147047, doi:10.1371/journal.pone.0147047 (2016).
125 Derntl, C. *et al.* In Vivo Study of the Sorbicillinoid Gene Cluster in Trichoderma reesei. *Front Microbiol* **8**, 2037, doi:10.3389/fmicb.2017.02037 (2017).
126 Du, L. *et al.* Biosynthesis of sphinganine-analog mycotoxins. *J Ind Microbiol Biotechnol* **35**, 455-464, doi:10.1007/s10295-008-0316-y (2008).
127 Ehrlich, K. C., Chang, P. K., Yu, J. & Cotty, P. J. Aflatoxin biosynthesis cluster gene cypA is required for G aflatoxin formation. *Appl Environ Microbiol* **70**, 6518-6524, doi:10.1128/AEM.70.11.6518-6524.2004 (2004).
128 Fierro, F. *et al.* 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* **43**, 618-629, doi:10.1016/j.fgb.2006.03.001 (2006).
129 Fisch, K. M. *et al.* Rational domain swaps decipher programming in fungal highly reducing polyketide synthases and resurrect an extinct metabolite. *J Am Chem Soc* **133**, 16635-16641, doi:10.1021/ja206914q (2011).
130 Gillot, G. *et al.* Genetic basis for mycophenolic acid production and strain-dependent production variability in Penicillium roqueforti. *Food Microbiol* **62**, 239-250, doi:10.1016/j.fm.2016.10.013 (2017).
131 Grundmann, A., Kuznetsova, T., Afiyatullov, S. & Li, S. M. FtmPT2, an N-prenyltransferase from Aspergillus fumigatus, catalyses the last step in the biosynthesis of fumitremorgin B. *Chembiochem* **9**, 2059-2063, doi:10.1002/cbic.200800240 (2008).
132 Guzman-Chavez, F. *et al.* Mechanism and regulation of sorbicillin biosynthesis by Penicillium chrysogenum. *Microb Biotechnol* **10**, 958-968, doi:10.1111/1751-7915.12736 (2017).
133 Halo, L. M. *et al.* Late Stage Oxidations during the Biosynthesis of the 2-Pyridone Tenellin in the Entomopathogenic Fungus Beauveria bassiana. *J Am Chem Soc* **130**, 17988-17996 (2008).

134 Hamed, R. B. *et al.* The enzymes of beta-lactam biosynthesis. *Nat Prod Rep* **30**, 21-107, doi:10.1039/c2np20065a (2013).
135 Hansen, B. G. *et al.* A new class of IMP dehydrogenase with a role in self-resistance of mycophenolic acid producing fungi. *BMC Microbiol* **11**, 202, doi:10.1186/1471-2180-11-202 (2011).
136 Hansen, B. G. *et al.* Involvement of a natural fusion of a cytochrome P450 and a hydrolase in mycophenolic acid biosynthesis. *Appl Environ Microbiol* **78**, 4908-4913, doi:10.1128/AEM.07955-11 (2012).
137 Hansen, B. G. *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* **77**, 3044-3051, doi:10.1128/AEM.01768-10 (2011).
138 Hendrickson, L. *et al.* Lovastatin biosynthesis in Aspergillus terreus: characterization of blocked mutants, enzyme activities and a multifunctional polyketide synthase gene. *Chemistry & Biology* **6** (1999).
139 Heneghan, M. N. *et al.* First heterologous reconstruction of a complete functional fungal biosynthetic multigene cluster. *Chembiochem* **11**, 1508-1512, doi:10.1002/cbic.201000259 (2010).
140 Heneghan, M. N. *et al.* The programming role of trans-acting enoyl reductases during the biosynthesis of highly reduced fungal polyketides. *Chemical Science* **2**, doi:10.1039/c1sc00023c (2011).
141 Hoffmann, K., Schneider-Scherzer, E., Kleinkauf, H. & Zocher, R. Purification and Characterizationo of Eucaryotic Alanine Racemase Acting as Key Enzyme in Cyclosporin Biosynthesis. *J Biol Chem* **269**, 12710-12714 (1994).
142 Kakule, T. B., Zhang, S., Zhan, J. & Schmidt, E. W. Biosynthesis of the tetramic acids Sch210971 and Sch210972. *Org Lett* **17**, 2295-2297, doi:10.1021/acs.orglett.5b00715 (2015).
143 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* **77**, 1061-1067, doi:10.1271/bbb.130026 (2013).
144 Kato, N. *et al.* Identification of cytochrome P450s required for fumitremorgin biosynthesis in Aspergillus fumigatus. *Chembiochem* **10**, 920-928, doi:10.1002/cbic.200800787 (2009).
145 Kato, N. *et al.* Gene disruption and biochemical characterization of verruculogen synthase of Aspergillus fumigatus. *Chembiochem* **12**, 711-714, doi:10.1002/cbic.201000562 (2011).
146 Kennedy, J. *et al.* Modulation of Polyketide Synthase Activity by Accessory Proteins During Lovastatin Biosynthesis. (1999).
147 Lia, Y. *et al.* Tricarballylic ester formation during biosynthesis of fumonisin mycotoxins in Fusarium verticillioides. *Mycology* **4**, 179-186, doi:10.1080/21501203.2013.874540 (2013).
148 Lin, X. *et al.* Heterologous Expression of Ilicicolin H Biosynthetic Gene Cluster and Production of a New Potent Antifungal Reagent, Ilicicolin J. *Molecules* **24**, doi:10.3390/molecules24122267 (2019).
149 Maiya, S., Grundmann, A., Li, S. M. & Turner, G. Improved tryprostatin B production by heterologous gene expression in Aspergillus nidulans. *Fungal Genet Biol* **46**, 436-440, doi:10.1016/j.fgb.2009.01.003 (2009).
150 Manzoni, M. & Rollini, M. Biosynthesis and biotechnological production of statins by filamentous fungi and application of these cholesterol-lowering drugs. *Appl Microbiol Biotechnol* **58**, 555-564, doi:10.1007/s00253-002-0932-9 (2002).
151 Mulder, K. C. *et al.* Lovastatin production: From molecular basis to industrial process optimization. *Biotechnol Adv* **33**, 648-665, doi:10.1016/j.biotechadv.2015.04.001 (2015).
152 Pinedo, C. *et al.* Sesquiterpene synthase from the botrydial biosynthetic gene cluster of the phytopathogen Botrytis cinerea. *ACS Chem Biol* **3**, 791-801, doi:10.1021/cb800225v (2008).
153 Porquier, 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* **96**, 33-46, doi:10.1016/j.fgb.2016.10.003 (2016).
154 Proctor, R. H., Busman, M., Seo, J. A., Lee, Y. W. & Plattner, R. D. A fumonisin biosynthetic gene cluster in Fusarium oxysporum strain O-1890 and the genetic basis for B versus C fumonisin production. *Fungal Genet Biol* **45**, 1016-1026, doi:10.1016/j.fgb.2008.02.004 (2008).
155 Proctor, R. H., Desjardins, A. E., Plattner, R. D. & Hohn, T. M. A Polyketide Synthase Gene Required for Biosynthesis of Fumonisin Mycotoxins in Gibberella fujikuroi Mating Population A. *Fungal Genet Biol* **27**, 100-112 (1999).
156 Regueira, T. B. *et al.* Molecular basis for mycophenolic acid biosynthesis in Penicillium brevicompactum. *Appl Environ Microbiol* **77**, 3035-3043, doi:10.1128/AEM.03015-10 (2011).
157 Salo, O. *et al.* Identification of a Polyketide Synthase Involved in Sorbicillin Biosynthesis by Penicillium chrysogenum. *Appl Environ Microbiol* **82**, 3971-3978, doi:10.1128/AEM.00350-16 (2016).
158 Scott, B. *et al.* Deletion and gene expression analyses define the paxilline biosynthetic gene cluster in Penicillium paxilli. *Toxins (Basel)* **5**, 1422-1446, doi:10.3390/toxins5081422 (2013).
159 Studt, L. *et al.* Two separate key enzymes and two pathway-specific transcription factors are involved in fusaric acid biosynthesis in Fusarium fujikuroi. *Environ Microbiol* **18**, 936-956, doi:10.1111/1462-2920.13150 (2016).
160 van den Berg, M. A., Westerlaken, I., Leeflang, C., Kerkman, R. & Bovenberg, R. A. Functional characterization of the penicillin biosynthetic gene cluster of Penicillium chrysogenum Wisconsin54-1255. *Fungal Genet Biol* **44**, 830-844, doi:10.1016/j.fgb.2007.03.008 (2007).
161 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* **109**, 1287-1292, doi:10.1073/pnas.1115983109 (2012).
162 Wat, C.-K., McInnes, G., Smith, D. G., Wright, J. L. C. & Vining, L. C. The yellow pigments of Beauveria species. Structures of tenellin and bassianin. *Can J Chem* **55**, 4090-4098, doi:10.1139/v77-580 (1977).
163 Weber, G. & Leitner, E. Disruption of the cyclosporin synthetase gene of Tolypocladium niveum. *Curr Genet* **26**, 461-467 (1994).
164 Xie, X., Watanabe, K., Wojcicki, W. A., Wang, C. C. & Tang, Y. Biosynthesis of lovastatin analogs with a broadly specific acyltransferase. *Chem Biol* **13**, 1161-1169, doi:10.1016/j.chembiol.2006.09.008 (2006).
165 Xu, L. *et al.* Identification of cyclosporin C from Amphichorda felina using a Cryptococcus neoformans differential temperature sensitivity assay. *Appl Microbiol Biotechnol* **102**, 2337-2350, doi:10.1007/s00253-018-8792-0 (2018).
166 Xu, W. *et al.* LovG: the thioesterase required for dihydromonacolin L release and lovastatin nonaketide synthase turnover in lovastatin biosynthesis. *Angew Chem Int Ed Engl* **52**, 6472-6475, doi:10.1002/anie.201302406 (2013).
167 Xu, X. *et al.* Identification of the first diphenyl ether gene cluster for pestheic acid biosynthesis in plant endophyte Pestalotiopsis fici. *Chembiochem* **15**, 284-292, doi:10.1002/cbic.201300626 (2014).
168 Yakasai, A. A. *et al.* Nongenetic reprogramming of a fungal highly reducing polyketide synthase. *J Am Chem Soc* **133**, 10990-10998, doi:10.1021/ja204200x (2011).
169 Yu, F. *et al.* Structure and biosynthesis of heat-stable antifungal factor (HSAF), a broad-spectrum antimycotic with a novel mode of action. *Antimicrob Agents Chemother* **51**, 64-72, doi:10.1128/AAC.00931-06 (2007).
170 Zaehle, C. *et al.* Terrein biosynthesis in Aspergillus terreus and its impact on phytotoxicity. *Chem Biol* **21**, 719-731, doi:10.1016/j.chembiol.2014.03.010 (2014).
171 Zaleta-Rivera, K. *et al.* A Bidomain Nonribosomal Peptide Synthetase Encoded by FUM14 Catalyzes the Formation of Tricarballylic Esters in the Biosynthesis of Fumonisins. *Biochemistry* **45**, 2561-2569 (2006).

172 Zhang, W. *et al.* Functional characterization of MpaG', the O-methyltransferase involved in the biosynthesis of mycophenolic acid. *Chembiochem* **16**, 565-569, doi:10.1002/cbic.201402600 (2015).
173 Zhang, W. *et al.* Compartmentalized biosynthesis of mycophenolic acid. *Proc Natl Acad Sci U S A* **116**, 13305-13310, doi:10.1073/pnas.1821932116 (2019).
174 Pinedo, C. *et al.* Sesquiterpene Synthase from the Botrydial Biosynthetic Gene Cluster of the Phytopathogen Botrytis cinerea. *ACS Chemical Biology* **3**, doi:10.1021/cb800225v (2008).
175 Lalchhandama, K. Reappraising Fleming's snot and mould. *Science Vision* **20**, 29-42, doi:10.33493/scivis.20.01.03 (2020).
176 Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6, doi:10.1186/s12864-019-6413-7 (2020).
177 Nisbet, R., Miner, G. & Yale, K. in *Handbook of Statistical Analysis and Data Mining Applications* 215-233 (2018).
178 Shmueli, B. *Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard of*, <https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a > (2019).
179 Janitza, S. & Hornung, R. On the overestimation of random forest's out-of-bag error. *PLoS One* **13**, e0201904, doi:10.1371/journal.pone.0201904 (2018).

9. Supplement

Supplement 9.1: R script used for the statistical analysis of the positive controls, yielding heatmaps, dendrograms, scatterplots and PCA results. Prior analysis, the data matrix was prepared to obtain numeric matrices and the data was then approximated with a random forest approach (MissForest). The calculation of heatmaps and dendrograms were executed using data with and without missing data.

```
##############################################
##                                          ##
##                  BGC-Analysis            ##
##                Positive control          ##
##                                          ##
##############################################

#############################################################
#                 written by Gabriel Alexander Vignolle     #
#                     modified by Denise Schaffer           #
#############################################################

library("readr"); packageVersion("readr")
library("stats"); packageVersion("stats")
library("outliers"); packageVersion("outliers")
library("dplyr"); packageVersion("dplyr")
library("ggplot2"); packageVersion("ggplot2")
library("gplots"); packageVersion("gplots")
library("car"); packageVersion("car")
library("moonBook"); packageVersion("moonBook")
library("e1071"); packageVersion("e1071")
library("lmtest"); packageVersion("lmtest")
library("metaheuristicOpt"); packageVersion("metaheuristicOpt")
library("factoextra"); packageVersion("factoextra")
library("pls"); packageVersion("pls")
library("idx2r"); packageVersion("idx2r")
library("mdatools"); packageVersion("mdatools")
library("missForest"); packageVersion("missForest")

#### Matrix Preparation ####
positive_cluster_01 <- read_delim ("~/Pos_control/positive_cluster_01/distance_matrix/positive_cluster_01.csv",
                    ";", escape_double = FALSE, col_types = cols(Gen = col_skip()), trim_ws = TRUE)
positive_cluster_01_numeric <- as.matrix(sapply(positive_cluster_01, as.numeric))
rownames(positive_cluster_01_numeric) <- colnames(positive_cluster_01)
pc01.imp <- missForest(positive_cluster_01_numeric)

#### Heatmap computation ####
#With approximated data
heatmap.2(pc01.imp$ximp, trace="none", margins = c(8, 8), srtRow=0, srtCol=45, keysize = 1.5,
      col= colorRampPalette(c("yellow", "orange", "red", "darkred", "black"))(100), main = "")
title(main = "Positive Control 01 - Heatmap")
#With Missing Data
heatmap.2(positive_cluster_01_numeric, trace="none", margins = c(8, 8), srtRow=0, srtCol=45, keysize = 1.5,
      col= colorRampPalette(c("yellow", "orange", "red", "darkred", "black"))(100), main="")
title(main = "Positive Control 01 - Heatmap")

#### Scaling ####
#With approximated data
pc01_unscaled <- scale(pc01.imp$ximp, center = TRUE, scale = FALSE)
pc01_scaled <- scale(pc01.imp$ximp, center = TRUE, scale = TRUE)
#With missing data
positive_cluster_01_unscaled <- scale(positive_cluster_01_numeric, center = TRUE, scale = FALSE)
positive_cluster_01_scaled <- scale(positive_cluster_01_numeric, center = TRUE, scale = TRUE)

#### Euclidean Distance computation ####
#With approximated data
pc01_d_unscaled = dist(pc01_unscaled)
pc01_d_scaled = dist(pc01_scaled)
#With missing data
positive_cluster_01_d_unscaled = dist(positive_cluster_01_unscaled)
positive_cluster_01_d_scaled = dist(positive_cluster_01_scaled)

#### Ward Clustering ####
#Scaled and unscaled approximated data clustered with ward.D2
pc01_dendro_unscaled <- hclust(pc01_d_unscaled, method = "ward.D2")
pc01_dendro_scaled <- hclust(pc01_d_scaled, method = "ward.D2")
#Scaled and unscaled missing data clustered with ward.D2
positive_cluster_01_dendro_unscaled <- hclust(positive_cluster_01_d_unscaled, method = "ward.D2")
positive_cluster_01_dendro_scaled <- hclust(positive_cluster_01_d_scaled, method = "ward.D2")

#Scaled and unscaled approximated data clustered with ward.D
pc01_dendro_unscaled_w <- hclust(pc01_d_unscaled, method = "ward.D")
pc01_dendro_scaled_w <- hclust(pc01_d_scaled, method = "ward.D")
#Scaled and unscaled missing data clustered with ward.D
positive_cluster_01_dendro_unscaled_w <- hclust(positive_cluster_01_d_unscaled, method = "ward.D")
positive_cluster_01_dendro_scaled_w <- hclust(positive_cluster_01_d_scaled, method = "ward.D")

#### Dendrogram computation ####
#Scaled and unscaled approximated data clustered with ward.D2
plot(pc01_dendro_unscaled, main = "")
title(main = "Unscaled Positive Cluster 01 - Dendrogram with Ward.D2")
plot(pc01_dendro_scaled, main = "")
title(main = "Scaled Positive Cluster 01 - Dendrogram with Ward.D2")

#Scaled and unscaled missing data clustered with ward.D2
plot(positive_cluster_01_dendro_unscaled, main = "")
```

```
title(main = "Unscaled Positive Cluster 01, Missing Data - Dendrogram with Ward.D2")
plot(positive_cluster_01_dendro_scaled, main = "")
title(main = "Scaled Positive Cluster 01, Missing Data - Dendrogram with Ward.D2")

#Scaled and unscaled approximated data clustered with ward.D
plot(pc01_dendro_unscaled_w, main = "")
title(main = "Unscaled Positive Cluster 01 - Dendrogram with Ward.D")
plot(pc01_dendro_scaled_w, main = "")
title(main = "Scaled Positive Cluster 01 - Dendrogram with Ward.D")

#Scaled and unscaled missing data clustered with ward.D
plot(positive_cluster_01_dendro_unscaled_w, main = "")
title(main = "Unscaled Positive Cluster 01, Missing Data - Dendrogram with Ward.D")
plot(positive_cluster_01_dendro_scaled_w, main = "")
title(main = "Scaled Positive Cluster 01, Missing Data - Dendrogram with Ward.D")

#### PCA ####
tree_PCA_pc01 <- pca(pc01.imp$ximp, scale = T, center = T)
plot(tree_PCA_pc01)
title(main = "Positive Control 01 - PCA")
plotScores(tree_PCA_pc01, comp = c(1, 2), show.labels = TRUE, main = "Positive Control 01 - PCA Scores" )
```

Supplement 9.2: R script used for the statistical analysis of the negative controls, yielding heatmaps, dendrograms, scatterplots and PCA results. Prior analysis, the data matrix was prepared to obtain numeric matrices.

```
##############################################
##                                          ##
##                BGC-Analysis              ##
##              Negative control            ##
##                                          ##
##############################################

#################################################################
#                written by Gabriel Alexander Vignolle          #
#                    modified by Denise Schaffer                #
#################################################################

library("readr"); packageVersion("readr")
library("stats"); packageVersion("stats")
library("outliers"); packageVersion("outliers")
library("dplyr"); packageVersion("dplyr")
library("ggplot2"); packageVersion("ggplot2")
library("gplots"); packageVersion("gplots")
library("car"); packageVersion("car")
library("moonBook"); packageVersion("moonBook")
library("e1071"); packageVersion("e1071")
library("lmtest"); packageVersion("lmtest")
library("metaheuristicOpt"); packageVersion("metaheuristicOpt")
library("factoextra"); packageVersion("factoextra")
library("pls"); packageVersion("pls")
library("idx2r"); packageVersion("idx2r")
library("mdatools"); packageVersion("mdatools")

##### Data preparation #####
random_cluster_01 <-read_delim("~/Neg_control/random_cluster_01.fasta.analysis/distance_matrix/random_cluster_01.csv",
                    ";", escape_double = FALSE, col_types = cols(X1 = col_skip()),
                    trim_ws = TRUE)
random_cluster_01_numeric <- as.matrix(sapply(random_cluster_01, as.numeric))
rownames(random_cluster_01_numeric) <- colnames(random_cluster_01)

#### Heatmap computation ####
heatmap.2(random_cluster_01_numeric, trace="none", margins = c(8, 8), srtRow=0, srtCol=45, keysize = 1.5, col=
colorRampPalette(c("yellow", "orange", "red", "darkred", "black"))(100), main = "")
title(main = "Negative Control 01 - Heatmap")

#### Scaling ####
random_cluster_01_unscaled <- scale(random_cluster_01_numeric, center = TRUE, scale = FALSE)
random_cluster_01_scaled <- scale(random_cluster_01_numeric, center = TRUE, scale = TRUE)

#### Euclidean Distance computation####

random_cluster_01_d_unscaled = dist(random_cluster_01_unscaled)
random_cluster_01_d_scaled = dist(random_cluster_01_scaled)

#### Ward Clustering ####
random_cluster_01_dendro_unscaled <- hclust(random_cluster_01_d_unscaled, method = "ward.D2")
random_cluster_01_dendro_scaled <- hclust(random_cluster_01_d_scaled, method = "ward.D2")

#### Dendrogram computation ####
plot(random_cluster_01_dendro_unscaled, main = "")
title(main = "Ward's minimum variance - random cluster 01 unscaled")
plot(random_cluster_01_dendro_scaled, main = "")
title(main = "Ward's minimum variance - random cluster 01 scaled")

#### Principal component analysis####
tree_PCA_rc01 <- pca(random_cluster_01_numeric, scale = T, center = T)
plot(tree_PCA_rc01, main = "")
title(main = "Negative Control 01 - PCA")
plotScores(tree_PCA_rc01, comp = c(1, 2), show.labels = TRUE, main="")
title(main = "Negative Control 01 - Scores")
```

Supplement 9.3: Manual evaluation measures (MEM) of the genes in the positive control 01, tetramic acid biosynthetic gene cluster (BGC) of Hapsidospora irregularis (GenBank® KP8352.02). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green.

Gene/ Gene	tasS	tasH	tasG	tas3	tasK	tasR	tasC	tasA
tasS	3.00	2.25	2.21	2.72	2.56	2.30	2.66	2.71
tasH	2.25	3.00						
tasG	2.21		3.00					2.20
tas3	2.72			3.00				
tasK	2.56				3.00			
tasR	2.30					3.00		
tasC	2.66						3.00	2.61
tasA	2.71		2.20				2.61	3.00

Supplement 9.4: Manual evaluation measures (MEM) of the genes in the positive control 02, mycophenolic acid biosynthetic gene cluster (BGC) of Penicillium brevicompactum (MIBiG BGC0000104). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	mpaF	mpaA	mpaB	mpaD	mpaE	mpaG	mpaH	mpaC
mpaF	3.00	1.88	2.21	1.90	1.85	2.35	2.23	2.40
mpaA	1.88	3.00	1.90		2.23	2.08		2.13
mpaB	2.21	1.90	3.00			2.06		2.52
mpaD	1.90			3.00	2.15			2.33
mpaE	1.85	2.23		2.15	3.00			2.11
mpaG	2.35	2.08	2.06			3.00	2.14	2.63
mpaH	2.23					2.14	3.00	1.94
mpaC	2.40	2.13	2.52	2.33	2.11	2.63	1.94	3.00

Supplement 9.5: Manual evaluation measures (MEM) of the genes in the positive control 03, mycophenolic acid biosynthetic gene cluster (BGC) of Penicillium roqueforti (MIBiG BGC0001360). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	mpaC	mpaA	mpaB	mpaDE	mpaF	mpaG	mpaH
mpaC	3.00	2.23	2.21	1.47	1.52	2.46	1.63
mpaA	2.23	3.00	2.23	2.04	1.63		
mpaB	2.21	2.23	3.00	1.88	2.47	2.30	
mpaDE	1.47	2.04	1.88	3.00	2.38		
mpaF	1.52	1.63	2.47	2.38	3.00	2.32	2.21
mpaG	2.46		2.30		2.32	3.00	1.44
mpaH	1.63				2.21	1.44	3.00

Supplement 9.6: Manual evaluation measures (MEM) of the genes in the positive control 04, mycophenolic acid biosynthetic gene cluster (BGC) of Penicillium roqueforti (MIBiG BGC0001677). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	mpaC	mpaA	mpaB	mpaDE	mpaF	mpaG	mpaH
mpaC	3.00	2.28	1.97	2.41	1.92	2.53	2.19
mpaA	2.28	3.00	1.93	2.31	1.66		
mpaB	1.97	1.93	3.00	1.91	2.54	2.54	
mpaDE	2.41	2.31	1.91	3.00	2.23		
mpaF	1.92	1.66	2.54	2.23	3.00	2.78	2.36
mpaG	2.53		2.54		2.78	3.00	2.48
mpaH	2.19				2.36	2.48	3.00

Supplement 9.7: Manual evaluation measures (MEM) of the genes in the positive control 05, botrydial biosynthetic gene cluster (BGC) of Botrytis cinera (MIBiG BGC0000631). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	bot1	bot2	bot3	bot4	bot5	bot6	bot7
bot1	3.00	2.50	2.73	2.55	2.44	2.16	2.50
bot2	2.50	3.00	2.38	1.83	2.13	2.19	2.54
bot3	2.73	2.38	3.00				
bot4	2.55	1.83		3.00			
bot5	2.44	2.13			3.00		
bot6	2.16	2.19				3.00	
bot7	2.50	2.54					3.00

Supplement 9.8: Manual evaluation measures (MEM) of the genes in the positive control 06, leporin B biosynthetic gene cluster (BGC) of Aspergillus flavus (MIBiG BGC0001445). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	lepA	lepB	Gen1	lepC	lepD	lepE	lepF	lepG	lepH	lepI
lepA	3.00	2.25	1.38	2.38	2.46	2.31	2.31	2.50	2.35	2.25
lepB	2.25	3.00								
Gen1	1.38		3.00							
lepC	2.38			3.00						
lepD	2.46				3.00	2.44			2.45	
lepE	2.31				2.44	3.00	2.02	2.63	1.79	
lepF	2.31					2.02	3.00	2.31		
lepG	2.50					2.63	2.31	3.00	2.73	
lepH	2.35				2.45	1.79		2.73	3.00	
lepI	2.25									3.00

Supplement 9.9: Manual evaluation measures (MEM) of the genes in the positive control 07, fumitremorgin biosynthetic gene cluster (BGC) of Aspergillus fumigatus (MIBiG BGC0000356). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	ftmA	ftmC	ftmD	ftmB	ftmE	ftmF	ftmG	ftmH	ftmI
ftmA	3.00	1.27	2.25	1.82	1.75	1.43	1.58	1.93	2.16
ftmC	1.27	3.00	0.00	1.91				1.69	
ftmD	2.25	0.00	3.00	2.09	1.88			1.88	
ftmB	1.82	1.91	2.09	3.00	1.94	2.11	1.70	2.11	1.81
ftmE	1.75		1.88	1.94	3.00		2.04	1.81	
ftmF	1.43			2.11		3.00		2.11	
ftmG	1.58			1.70	2.04		3.00	1.67	
ftmH	1.93	1.69	1.88	2.11	1.81	2.11	1.67	3.00	1.80
ftmI	2.16			1.81				1.80	3.00

Supplement 9.10: Manual evaluation measures (MEM) of the genes in the positive control 08, tenellin biosynthetic gene cluster (BGC) of Beauveria bassiana (MIBiG BGC0001049). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	tenS	ten1	tenA	tenB	tenC
tenS	3.00	1.20	2.03	2.04	2.33
ten1	1.20	3.00			
tenA	2.03		3.00	2.47	2.57
tenB	2.04		2.47	3.00	
tenC	2.33		2.57		3.00

Supplement 9.11: Manual evaluation measures (MEM) of the genes in the positive control 09, ilicicolin H biosynthetic gene cluster (BGC) of Neonectria sp. DH2 (MIBiG BGC0002035). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	iliA	gene1	iliE	iliB	iliD	iliC
iliA	3.00	1.79	2.33	2.27	2.04	2.41
gene1	1.79	3.00				
iliE	2.33		3.00			
iliB	2.27			3.00		2.65
iliD	2.04				3.00	2.22
iliC	2.41			2.65	2.22	3.00

Supplement 9.12: Manual evaluation measures (MEM) of the genes in the positive control 10, 2-pyridon-desmethylbassianin biosynthetic gene cluster (BGC) of Beaveria bassiana (MIBiG BGC0001136). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	dmbA	dmbB	dmbC	dmbS
dmbA	3.00	2.41	2.61	2.53
dmbB	2.41	3.00	2.11	2.06
dmbC	2.61	2.11	3.00	2.40
dmbS	2.53	2.06	2.40	3.00

Supplement 9.13: Manual evaluation measures (MEM) of the genes in the positive control 11, xanthocillin biosynthetic gene cluster (BGC) of Aspergillus fumigatus (MIBiG BGC0001990). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	xanB	xanG	xanF	xanE	xanD	xanC	xanA
xanB	3.00	2.59	1.75	1.63	2.53	1.83	2.31
xanG	2.59	3.00		1.69		2.16	1.94
xanF	1.75		3.00			1.63	
xanE	1.63	1.69		3.00		1.44	1.50
xanD	2.53				3.00	2.00	
xanC	1.83	2.16	1.63	1.44	2.00	3.00	1.61
xanA	2.31	1.94		1.50		1.61	3.00

Supplement 9.14: Manual evaluation measures (MEM) of the genes in the positive control 12, fumagillin biosynthetic gene cluster (BGC) of Aspergillus fumigatus (MIBiG BGC000107). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	Af370	Af380	Af390	Af400	Af410	Af420	Af430	Af440	Af460	Af470	Af480	Af490	Af500	Af510	Af520
Af370	3.00	2.28	1.63	1.31	1.71	1.48	1.75	2.06	1.66	0.00	2.08	1.88	1.50	2.13	1.38
Af380	2.28	3.00	1.69										1.63		1.13
Af390	1.63	1.69	3.00										2.68		1.86
Af400	1.31			3.00									2.68		
Af410	1.71				3.00								2.50		
Af420	1.48					3.00							1.53		
Af430	1.75						3.00						2.68		
Af440	2.06							3.00					2.60		
Af460	1.66								3.00				2.09		
Af470	0.00									3.00			1.95		
Af480	2.08										3.00		2.03		
Af490	1.88											3.00	2.48		
Af500	1.50	1.63	2.68	2.68	2.50	1.53	2.68	2.60	2.09	1.95	2.03	2.48	3.00	2.48	2.29
Af510	2.13												2.48	3.00	
Af520	1.38	1.13	1.86										2.29		3.00

Supplement 9.15: Manual evaluation measures (MEM) of the genes in the positive control 13, terrein biosynthetic gene cluster (BGC) of Aspergillus terreus (MIBiG BGC0000161). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	terA	terB	terC	terD	terE	terF	terR	terG	terH	terJ
terA	3.00	2.30	1.96	2.43	1.97	2.06	1.73	1.43	1.57	2.05
terB	2.30	3.00	1.75	1.83	1.63	1.50	2.53			
terC	1.96	1.75	3.00	2.24	1.56	1.79	2.00			
terD	2.43	1.83	2.24	3.00	0.00	1.81	1.72			
terE	1.97	1.63	1.56	0.00	3.00	0.00	1.75			
terF	2.06	1.50	1.79	1.81	0.00	3.00	1.44			
terR	1.73	2.53	2.00	1.72	1.75	1.44	3.00			
terG	1.43							3.00		
terH	1.57								3.00	
terJ	2.05									3.00

Supplement 9.16: Manual evaluation measures (MEM) of the genes in the positive control 15, fumonisin biosynthetic gene cluster (BGC) of Fusarium oxysporum (MIBiG BGC0000063). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene / Gene	Fum 1	Fum2 1	Fum 6	Fum 7	Fum 8	Fum 3	Fum1 0	Fum1 1	Fum 2	Fum1 3	Fum1 4	Fum1 5	Fum1 6	Fum1 7	Fum1 8	Fum1 9	Cpm 1
Fum 1	3.00	1.95	2.79	2.67	2.53	2.30	2.47	1.72	2.23	2.75	2.31	2.75	2.30	2.17	1.91	2.77	1.48
Fum 21	1.95	3.00			1.81		2.10				1.93		2.31				
Fum 6	2.79		3.00				2.94	1.94		2.92	2.46		2.56				
Fum 7	2.67			3.00			2.78	2.29	2.50		2.18		2.53				
Fum 8	2.53	1.81			3.00		2.47			2.50	2.45		2.44				
Fum 3	2.30					3.00	2.04		2.08		1.70		2.16				
Fum 10	2.47	2.10	2.94	2.78	2.47	2.04	3.00	2.19	2.08	2.98	2.63	2.96	2.72	2.28	2.52	2.98	1.50
Fum 11	1.72		1.94	2.29			2.19	3.00			1.65		2.23				
Fum 2	2.23			2.50		2.08	2.08		3.00		1.85		2.54				
Fum 13	2.75		2.92		2.50		2.98			3.00	2.50		2.69				
Fum 14	2.31	1.93	2.46	2.18	2.45	1.70	2.63	1.65	1.85	2.50	3.00	2.34	2.15	2.09	1.85	2.67	1.28
Fum 15	2.75						2.96				2.34	3.00	2.44				
Fum 16	2.30	2.31	2.56	2.53	2.44	2.16	2.72	2.23	2.54	2.69	2.15	2.44	3.00	2.18	2.21	2.63	1.25
Fum 17	2.17						2.28				2.09		2.18	3.00			
Fum 18	1.91						2.52				1.85		2.21		3.00		
Fum 19	2.77						2.98				2.67		2.63			3.00	
Cpm 1	1.48						1.50				1.28		1.25				3.00

Supplement 9. 17: Manual evaluation measures (MEM) of the genes in the positive control 14, pneumocandin biosynthetic gene cluster (BGC) of Glarea iozoyensis (MIBiG BGC0001035). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene	10034	10020	10021	10022	10023	10024	10026	10027	10028	10029	10030	10031	10032	10033	10035	10036	10037	10038	10039	10040	10041	10042	10043	10044	10045	10047	10049	10050
10034	3.00	0.00	0.00	1.50	1.75	1.54	1.88	1.88	1.63	0.00	1.50	1.29	1.25	1.59	1.25	2.13	1.63	1.88	0.00	1.75	0.00	1.63	2.05	2.09	1.00	1.38	1.25	1.96
10020	0.00	3.00													0.00				1.15				0.00					
10021	0.00		3.00												0.00				1.52				1.72					
10022	1.50			3.00											0.00				1.72				1.09					
10023	1.75				3.00										1.19				1.78				1.50					
10024	1.54					3.00									2.13				1.35				1.19					
10026	1.88						3.00								1.00				1.72				2.15					
10027	1.88							3.00							1.38				1.57				1.91					
10028	1.63								3.00						0.00				1.39				1.94					
10029	0.00									3.00					1.00				1.34				1.56					
10030	1.50										3.00				2.38				1.98				2.16					
10031	1.29											3.00			2.63				2.39				1.83					
10032	1.25												3.00		2.38				1.46				1.29					
10033	1.59													3.00	2.54				2.50				2.10					
10035	1.25	0.00	0.00	0.00	1.19	2.13	1.00	1.38	0.00	1.00	2.38	2.63	2.38	2.54	3.00	2.63	2.00	1.93	2.50	0.00	2.13	1.00	1.53	1.75	1.29	1.25	1.38	1.38
10036	2.13														2.63	3.00			2.50				2.09					
10037	1.63														2.00		3.00		2.23				1.88					
10038	1.88														1.93			3.00	2.63				1.83					
10039	0.00	1.15	1.52	1.72	1.78	1.35	1.72	1.57	1.39	1.34	1.98	2.39	1.46	2.50	2.50	2.50	2.23	2.63	3.00	1.55	2.38	2.10	1.98	2.04	1.54	1.10	0.00	1.08

100 40	1.75														0.00				1.55	3.00			0.00					
100 41	0.00														2.13				2.38		3.00		2.00					
100 42	1.63														1.00				2.10			3.00	1.97					
100 43	2.05	0.00	1.72	1.09	1.50	1.19	2.15	1.91	1.94	1.56	2.16	1.83	1.29	2.10	1.53	2.09	1.88	1.83	1.98	0.00	2.00	1.97	3.00	2.03	1.42	1.29	1.33	1.63
100 44	2.09														1.75				2.04				2.03	3.00				
100 45	1.00														1.29				1.54				1.42		3.00			
100 47	1.38														1.25				1.10				1.29			3.00		
100 49	1.25														1.38				0.00				1.33				3.00	
100 50	1.96														1.38				1.08				1.63					3.00

Supplement 9.18: Manual evaluation measures 34 of the genes in the positive control 16, destruxin biosynthetic gene cluster (BGC) of Metarhizium robertsii (MIBiG BGC0000337). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	DtxS1	MMA 10031	MMA 10032	MMA 10033	MMA 10034	MMA 10035	MMA 10036	MMA 10037	MMA 10038	MMA 10039	MMA 10040	MMA 10041	MMA 10042	DtxS2	DtxS3	DtxS4	MAA10 047	MAA10 048	MAA10 049	MAA10 050	MAA10 051
DtxS1	3.00	0.00	1.56	2.44	2.38	2.16	2.13	2.13	2.44	2.13	1.88	2.21	2.58	2.89	2.67	2.53	2.24	2.52	1.63	2.21	1.88
MMA 10031	0.00	3.00		1.13			0.00														
MMA 10032	1.56		3.00	1.80			2.08														
MMA 10033	2.44	1.13	1.80	3.00	2.76	2.08	2.32	2.23	2.73	2.31	1.29	2.16	2.25	2.38	1.94	2.68	1.92	2.34	2.28	2.28	1.83
MMA 10034	2.38			2.76	3.00		2.53														
MMA 10035	2.16			2.08		3.00	2.30														
MMA 10036	2.13	0.00	2.08	2.32	2.53	2.30	3.00	2.25	2.28	1.96	1.00	1.75	1.92	1.54	1.63	1.85	1.89	1.98	2.16	2.25	1.33
MMA 10037	2.13			2.23			2.25	3.00													
MMA 10038	2.44			2.73			2.28		3.00												
MMA 10039	2.13			2.31			1.96			3.00											
MMA 10040	1.88			1.29			1.00				3.00										
MMA 10041	2.21			2.16			1.75					3.00									
MMA 10042	2.58			2.25			1.92						3.00								
DtxS2	2.89			2.38			1.54							3.00							
DtxS3	2.67			1.94			1.63								3.00						
DtxS4	2.53			2.68			1.85									3.00					
MAA1 0047	2.24			1.92			1.89										3.00				
MAA1 0048	2.52			2.34			1.98											3.00			
MAA1 0049	1.63			2.28			2.16												3.00		
MAA1 0050	2.21			2.28			2.25													3.00	

MAA10051	1.88			1.83			1.33													3.00

Supplement 9.19: Manual evaluation measures (MEM) of the genes in the positive control 17, paxillin biosynthetic gene cluster (BGC) of Penicillium paxilli (MIBiG BGC0001082). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	PaxG	PaxA	PaxM	PaxB	PaxC	PaxP	PaxQ	PaxD
PaxG	3.00	1.17	2.31	1.78	2.05	1.64	1.72	1.55
PaxA	1.17	3.00			1.25			1.94
PaxM	2.31		3.00	2.48	2.73	2.16		1.69
PaxB	1.78		2.48	3.00	2.50	1.96		1.63
PaxC	2.05	1.25	2.73	2.50	3.00	2.04	2.00	1.52
PaxP	1.64		2.16	1.96	2.04	3.00	2.31	1.65
PaxQ	1.72				2.00	2.31	3.00	1.75
PaxD	1.55	1.94	1.69	1.63	1.52	1.65	1.75	3.00

Supplement 9.20: Manual evaluation measures (MEM) of the genes in the positive control 18, aflatoxin biosynthetic gene cluster (BGC) of Aspergillus flavus (MIBiG BGC0000008). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	PksA	AflT	Nor1	HexA	HexB	AflR	AflJ	AdhA	EstA	NorA	Ver1	VerA	AvnA	VerB	AvfA	OmtB	OmtA	OrdA	VBS	CypX	MoxY	OrdB	HypA
PksA	3.00	2.28	2.43	2.35	2.33	1.88	2.34	2.08	1.75	1.83	1.71	1.95	2.34	2.34	1.98	2.21	2.29	2.63	1.05	2.38	2.08	2.50	2.42
AflT	2.28	3.00																					
Nor1	2.43		3.00			1.79							2.38										
HexA	2.35			3.00		1.88																	
HexB	2.33				3.00	2.00																	
AflR	1.88		1.79	1.88	2.00	3.00		1.84	1.58	2.06	1.25		1.81	2.13	2.15	2.05	2.63	2.05	2.10	2.50	1.88	1.71	
AflJ	2.34						3.00																
AdhA	2.08					1.84		3.00		2.44			2.04		2.13					2.44			
EstA	1.75					1.58			3.00						2.00				2.08	2.06			
NorA	1.83					2.06		2.44		3.00			2.28		1.77					2.09			
Ver1	1.71					1.25					3.00			2.25		1.67					1.95		
VerA	1.95											3.00											
AvnA	2.34		2.38			1.81		2.04		2.28			3.00										
VerB	2.34					2.13					2.25			3.00					2.50		2.46		
AvfA	1.98					2.15		2.13	2.00	1.77					3.00					2.41			
OmtB	2.21					2.05					1.67					3.00	1.91				2.49		
OmtA	2.29					2.63										1.91	3.00	1.78				1.63	
OrdA	2.63					2.05											1.78	3.00					
VBS	1.05					2.10			2.08					2.50					3.00				
CypX	2.38					2.50		2.44	2.06	2.09					2.41					3.00			
MoxY	2.08					1.88					1.95			2.46		2.49					3.00		
OrdB	2.50					1.71											1.63					3.00	
HypA	2.42																						3.00

Supplement 9.21: Manual evaluation measures (MEM) of the genes in the positive control 19, pestheic acid biosynthetic gene cluster (BGC) of Pestalotiopsis fici (MIBiG BGC0000121). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	ptaA	ptaB	ptaC	ptaD	ptaE	ptaF	ptaG	ptaH	ptaR1	ptaR2	ptaI	ptaJ	ptaK	ptaL	ptaR3	ptaM	orf1	orf2
ptaA	3.00	2.23	2.00	2.11	2.38	2.05	2.46	2.08	1.75	1.28	1.52	2.15	1.75	2.06	1.63	1.50	1.33	1.44
ptaB	2.23	3.00	2.21															
ptaC	2.00	2.21	3.00					1.78			1.74							
ptaD	2.11			3.00														
ptaE	2.38				3.00			1.85			1.77					1.84		
ptaF	2.05					3.00		2.01			1.52	2.35						
ptaG	2.46						3.00											
ptaH	2.08		1.78		1.85	2.01		3.00			1.59	2.11				1.46		
ptaR1	1.75								3.00									
ptaR2	1.28									3.00								
ptaI	1.52		1.74		1.77	1.52		1.59			3.00	2.38				1.75		
ptaJ	2.15					2.35		2.11			2.38	3.00						
ptaK	1.75												3.00					
ptaL	2.06													3.00				
ptaR3	1.63														3.00			
ptaM	1.50				1.84			1.46			1.75					3.00		
orf1	1.33																3.00	
orf2	1.44																	3.00

Supplement 9.22: Manual evaluation measures [34] of the genes in the positive control 20, cephalosporine biosynthetic gene cluster (BGC) of Acremonium chrysogenum (MIBiG BGC0000317). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	pcbAB	cefG	cefD1	cefD2	pcbC	orf3	cefT
pcbAB	3.00	1.92	2.38	2.53	2.46	1.77	1.73
cefG	1.92	3.00	2.41				
cefD1	2.38	2.41	3.00	2.45	2.06	2.25	2.63
cefD2	2.53		2.45	3.00	2.28		
pcbC	2.46		2.06	2.28	3.00		
orf3	1.77		2.25			3.00	
cefT	1.73		2.63				3.00

Supplement 9.23: Manual evaluation measures [34] of the genes in the positive control 21, penicillin biosynthetic gene cluster (BGC) of Penicillium chrysogenum (MIBiG BGC0000404). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	pcbAB	ORF2	ORF3	ORF4	ORF5	ORF6	ORF7	pcbC	penDE	ORF11	ORF12	ORF13	ORF14	ORF15	ORF16
pcbAB	3.00	2.63	2.63	1.79	1.71	2.50	1.94	2.36	2.71	2.59	2.00	1.42	1.94	1.70	1.38
ORF2	2.63	3.00													
ORF3	2.63		3.00												
ORF4	1.79			3.00											
ORF5	1.71				3.00										
ORF6	2.50					3.00									
ORF7	1.94						3.00								
pcbC	2.36							3.00	2.68						
penDE	2.71							2.68	3.00						
ORF11	2.59									3.00					
ORF12	2.00										3.00				
ORF13	1.42											3.00			
ORF14	1.94												3.00		
ORF15	1.70													3.00	
ORF16	1.38														3.00

Supplement 9.24: Manual evaluation measures (MEM) of the genes in the positive control 22, penicillin biosynthetic gene cluster (BGC) of Penicillium chrysogenum (GenBank® EF601124.1). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	pcbAB	ORF10c	ORF30c	ORF40w	ORF50w	ORF60w	ORF70c	ORF80c	pcbC	penDE	ORF120c	ORF130w	ORF140w	ORF150w	ORF160
pcbAB	3.00	1.94	2.13	2.25	2.33	1.71	2.50	2.15	2.32	2.60	2.33	2.56	1.42	1.67	1.63
ORF10c	1.94	3.00													
ORF30c	2.13		3.00												
ORF40w	2.25			3.00											
ORF50w	2.33				3.00										
ORF60w	1.71					3.00									
ORF70c	2.50						3.00								
ORF80c	2.15							3.00							
pcbC	2.32								3.00	2.55					
penDE	2.60								2.55	3.00					
ORF120c	2.33										3.00				
ORF130w	2.56											3.00			
ORF140w	1.42												3.00		
ORF150w	1.67													3.00	
ORF160	1.63														3.00

Supplement 9.25: Manual evaluation measures (MEM) of the genes in the positive control 23, sorbicillin biosynthetic gene cluster (BGC) of Penicillium rubens (MIBiG BGC0001404). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	sorA	sorR1	sorC	sorB	sorR2	sorT	sorD
sorA	3.00	2.75	2.50	2.58	2.00	1.97	1.92
sorR1	2.75	3.00		2.31	1.83		
sorC	2.50		3.00	2.51			1.83
sorB	2.58	2.31	2.51	3.00	2.48	2.71	2.06
sorR2	2.00	1.83		2.48	3.00	2.47	
sorT	1.97			2.71	2.47	3.00	
sorD	1.92		1.83	2.06			3.00

Supplement 9.26: Manual evaluation measures (MEM) of the genes in the positive control 24, ustiloxin B biosynthetic gene cluster (BGC) of Aspergillus flavus (GenBank® NW_002477245). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	ustA	AFLA94900	AFLA94910	AFLA94920	ustO	ustF1	ustC	ustYa	ustP2	ustYb	ustH	ustD	ustF2	ustQ	ustT	ustR1	ustR2	ustM	ustS
ustA	3.00	1.25	1.75	0.00	1.88	1.13	2.00	1.52	2.28	1.38	1.44	1.48	2.00	2.38	2.38	1.83	2.10	2.04	1.31
AFLA94900	1.25	3.00					0.00	1.75	1.58		1.65								
AFLA94910	1.75		3.00				2.63	1.81	1.89		1.57								
AFLA94920	0.00			3.00			0.00	1.63	1.63		1.66								
ustO	1.88				3.00		1.78	1.31	1.69		2.38								
ustF1	1.13					3.00	0.00	0.00	2.20		2.46		1.99	1.75		1.44	1.31	1.88	2.46
ustC	2.00	0.00	2.63	0.00	1.78	0.00	3.00	2.38	2.63	2.00	2.63	2.00	1.88	2.13	2.13	2.63	2.63	2.19	0.00
ustYa	1.52	1.75	1.81	1.63	1.31	0.00	2.38	3.00	2.03	1.79	2.50	2.16	1.88	1.84	2.06	2.11	2.30	2.02	2.50
ustP2	2.28	1.58	1.89	1.63	1.69	2.20	2.63	2.03	3.00	1.83	2.43	2.38	2.39	2.69	2.28	2.03	2.25	2.45	2.29
ustYb	1.38						2.00	1.79	1.83	3.00	1.83								
ustH	1.44	1.65	1.57	1.66	2.38	2.46	2.63	2.50	2.43	1.83	3.00	2.25	2.48	1.94	1.73	2.63	2.63	1.91	2.57
ustD	1.48						2.00	2.16	2.38		2.25	3.00							
ustF2	2.00					1.99	1.88	1.88	2.39		2.48		3.00	1.50		1.75	1.63	1.48	2.39
ustQ	2.38					1.75	2.13	1.84	2.69		1.94		1.50	3.00		1.67	1.77		
ustT	2.38						2.13	2.06	2.28		1.73				3.00				
ustR1	1.83					1.44	2.63	2.11	2.03		2.63		1.75	1.67		3.00	1.87	2.22	2.63
ustR2	2.10					1.31	2.63	2.30	2.25		2.63		1.63	1.77		1.87	3.00	2.31	2.63
ustM	2.04					1.88	2.19	2.02	2.45		1.91		1.48			2.22	2.31	3.00	1.98
ustS	1.31					2.46	0.00	2.50	2.29		2.57		2.39			2.63	2.63	1.98	3.00

Supplement 9.27: Manual evaluation measures (MEM) of the genes in the positive control 25, lovastatin biosynthetic gene cluster (BGC) of Aspergillus terreus. The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	lovB	orf1	orf2	lovA	lovG	lovC	lovD	orf8	lovE	orf10	lovF	orf13	orf14	orf15	orf16	orf17	orf18
lovB	3.00	1.13	1.84	2.06	2.57	2.43	2.18	1.50	1.33	2.00	2.15	0.00	1.40	2.00	1.75	1.25	1.98
orf1	1.13	3.00									1.25						
orf2	1.84		3.00								1.56						
lovA	2.06			3.00	2.41						2.33						
lovG	2.57			2.41	3.00	2.41					2.20						
lovC	2.43				2.41	3.00					1.85						
lovD	2.18						3.00				2.13						
orf8	1.50							3.00			1.95						
lovE	1.33								3.00		1.50						
orf10	2.00									3.00	2.25						
lovF	2.15	1.25	1.56	2.33	2.20	1.85	2.13	1.95	1.50	2.25	3.00	1.56	1.25	2.03	1.75	1.63	1.96
orf13	0.00										1.56	3.00					
orf14	1.40										1.25		3.00				
orf15	2.00										2.03			3.00			
orf16	1.75										1.75				3.00		
orf17	1.25										1.63					3.00	
orf18	1.98										1.96						3.00

Supplement 9.28: Manual evaluation measures (MEM) of the genes in the positive control 26, compactin biosynthetic gene cluster (BGC) of Penicillium citrinum (MIBiG BGC0000039). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	mlcA	mlcB	mlcC	mlcD	mlcE	mlcF	mlcG	mlcH	mlcR
mlcA	3.00	2.80	2.61	1.69	2.00	2.73	2.68	2.50	1.21
mlcB	2.80	3.00	2.19	1.59	2.18	2.48	1.99	2.53	1.75
mlcC	2.61	2.19	3.00	1.52	2.23		2.35		
mlcD	1.69	1.59	1.52	3.00	1.81		1.54	1.63	1.00
mlcE	2.00	2.18	2.23	1.81	3.00		1.90	2.15	
mlcF	2.73	2.48				3.00			
mlcG	2.68	1.99	2.35	1.54	1.90		3.00		1.42
mlcH	2.50	2.53		1.63	2.15			3.00	1.56
mlcR	1.21	1.75		1.00			1.42	1.56	3.00

Supplement 9.29: Manual evaluation measures (MEM) of the genes in the positive control 27, sorbicillin biosynthetic gene cluster (BGC) of Trichoderma reesei (GenBank® GL985056). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	sor1	118	119	120	121	122	123	124	sor2	128	ypr2	sor4	ypr1
sor1	3.00	1.75	2.33	1.93	1.92	2.31	1.93	2.28	2.39	2.22	2.02	2.28	1.82
118	1.75	3.00						2.39	2.04				
119	2.33		3.00					2.64	2.63				
120	1.93			3.00				2.32	2.08				
121	1.92				3.00			2.41	2.63				
122	2.31					3.00		2.45	2.63				
123	1.93						3.00	2.18	2.63				
124	2.28	2.39	2.64	2.32	2.41	2.45	2.18	3.00	2.63	2.22	2.04	2.33	1.97
sor2	2.39	2.04	2.63	2.08	2.63	2.63	2.63	2.63	3.00	2.77	2.50	2.40	1.94
128	2.22							2.22	2.77	3.00			
ypr2	2.02							2.04	2.50		3.00		2.10
sor4	2.28							2.33	2.40			3.00	2.06
ypr1	1.82							1.97	1.94		2.10	2.06	3.00

Supplement 9.30: Manual evaluation measures (MEM) of the genes in the positive control 28, fumonisin biosynthetic gene cluster (BGC) of Fusarium verticillioides (MIBiG BGC0000062). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	fum1	npt1	wdr1	png1	znf1	zbd1	fum21	fum6	fum7	fum8	fum3	fum10	fum11	fum2	fum13	fum14	fum15	fum16	fum17	fum18	fum19	orf21	mpu1
fum1	3.00	2.00	2.38	2.00	2.43	2.16	2.15	2.53	2.30	2.30	2.16	2.50	1.91	2.34	2.30	2.27	2.39	2.30	1.63	2.11	2.43	2.38	2.00
npt1	2.00	3.00										1.78				1.31		2.09					
wdr1	2.38		3.00									2.30				1.74		2.77					
png1	2.00			3.00								2.16				1.69		2.12					
znf1	2.43				3.00							2.55				2.30		2.39					
zbd1	2.16					3.00						2.27				1.80		2.55					
fum21	2.15						3.00			1.83		2.10				1.98		2.28					
fum6	2.53							3.00				2.84	2.06		2.73	2.06		2.69					
fum7	2.30								3.00			2.47	2.16	2.17		2.25		2.33					
fum8	2.30						1.83			3.00		2.67			2.48	2.17		2.47					
fum3	2.16										3.00	2.08		1.75		1.88		2.23					
fum10	2.50	1.78	2.30	2.16	2.55	2.27	2.10	2.84	2.47	2.67	2.08	3.00	2.29	2.33	2.64	2.54	2.86	2.55	0.00	2.25	2.86	2.29	2.41
fum11	1.91							2.06	2.16			2.29	3.00			1.63		2.49					
fum2	2.34								2.17		1.75	2.33		3.00		2.28		2.46					
fum13	2.30							2.73		2.48		2.64			3.00	1.98		2.33					
fum14	2.27	1.31	1.74	1.69	2.30	1.80	1.98	2.06	2.25	2.17	1.88	2.54	1.63	2.28	1.98	3.00	2.38	1.89	1.00	1.86	2.45	2.17	1.88
fum15	2.39											2.86				2.38	3.00	2.69					
fum16	2.30	2.09	2.77	2.12	2.39	2.55	2.28	2.69	2.33	2.47	2.23	2.55	2.49	2.46	2.33	1.89	2.69	3.00	1.54	2.25	2.69	2.18	2.30
fum17	1.63											0.00				1.00		1.54	3.00				
fum18	2.11											2.25				1.86		2.25		3.00			
fum19	2.43											2.86				2.45		2.69			3.00		
orf21	2.38											2.29				2.17		2.18				3.00	
mpu1	2.00											2.41				1.88		2.30					3.00

Supplement 9.31: Manual evaluation measures (MEM) of the genes in the positive control 29, cyclosporine biosynthetic gene cluster (BGC) of Beauveria felina (MIBiG BGC0001565). The MEM values of the gene comparisons are represented in the matrix. Empirically verified gene comparisons necessary for the secondary metabolite production are highlighted green. False negatives according to the threshold of 2.0 and above are highlighted in red.

Gene/ Gene	simA	ATQ 39424	ATQ 39425	ATQ 39426	simB	simD	simE	simF	simG	simH	simI	simJ	simK
simA	3.00	1.77	1.13	1.27	1.56	1.81	1.96	1.48	1.28	1.65	1.34	1.13	1.29
ATQ39424	1.77	3.00	0.00						1.25				
ATQ39425	1.13	0.00	3.00	1.40	1.31	1.75	1.35	1.57	1.73	1.04	1.44	1.46	1.33
ATQ39426	1.27		1.40	3.00					1.00				
simB	1.56		1.31		3.00				2.19			1.63	
simD	1.81		1.75			3.00			2.23				
simE	1.96		1.35				3.00		1.56				
simF	1.48		1.57					3.00	1.75				
simG	1.28	1.25	1.73	1.00	2.19	2.23	1.56	1.75	3.00	1.59	2.67	1.75	1.48
simH	1.65		1.04						1.59	3.00			
simI	1.34		1.44						2.67		3.00	1.84	
simJ	1.13		1.46		1.63				1.75		1.84	3.00	
simK	1.29		1.33						1.48				3.00

Supplement 9.32: Manual evaluation measures (MEM) of each phylogenetic tree in the negative controls (number 1 – 60), represented as matrices. Their names of the trees were abbreviated and therefore some names occur more often. The matrices were used as the basis for the statistical evaluation. False Positives according to the threshold of 1.5 and below are highlighted in red.

	Tree/Tree	claim1	kil589	xp_7519	xp_0145	xp_0187
1	claim1	3.00	0.00	1.13	0.00	0.00
	kil589	0.00	3.00	1.13	0.00	0.00
	xp_7519	1.13	1.13	3.00	0.00	1.88
	xp_0145	0.00	0.00	0.00	3.00	0.00
	xp_0187	0.00	0.00	1.88	0.00	3.00

	Tree/Tree	exosp1	morco1	xp0023	xp0183
2	exosp1	3.00	0.00	1.64	1.31
	morco1	0.00	3.00	1.50	1.35
	xp0023	1.64	1.50	3.00	1.13
	xp0183	1.31	1.35	1.13	3.00

	Tree/Tree	acrchr1	exosp1	neute	paevar	rfu808	ryo654	xp_0160
3	acrchr1	3.00	2.25	2.39	0.00	1.88	0.00	0.00
	exosp1	2.25	3.00	2.19	1.63	1.63	1.71	1.13
	neute	2.39	2.19	3.00	0.00	1.84	1.00	1.69
	paevar	0.00	1.63	0.00	3.00	1.63	1.25	1.88
	rfu808	1.88	1.63	1.84	1.63	3.00	1.00	1.50
	ryo654	0.00	1.71	1.00	1.25	1.00	3.00	1.41
	xp_0160	0.00	1.13	1.69	1.88	1.50	1.41	3.00

	Tree/Tree	exosp1	necha2	xp0234	xp0245
4	exosp1	3.00	0.00	0.00	1.50
	necha2	0.00	3.00	1.77	0.00

	xp0234	0.00	1.77	3.00	0.00			
	xp0245	1.50	0.00	0.00	3.00			
5	Tree/Tree	cadsp1	phisc1	triru1	tubae1			
	cadsp1	3.00	1.35	1.69	0.00			
	phisc1	1.35	3.00	2.13	0.00			
	triru1	1.69	2.13	3.00	0.00			
	tubae1	0.00	0.00	0.00	3.00			
6	Tree/Tree	mat_a1	verga1	xylhe1				
	mat_a1	3.00	1.75	0.00				
	verga1	1.75	3.00	1.75				
	xylhe1	0.00	1.75	3.00				
7	Tree/Tree	magpo1	paevar	xp668	xp622			
	magpo1	3.00	0.00	1.28	1.31			
	paevar	0.00	3.00	0.00	0.00			
	xp668	1.28	0.00	3.00	2.50			
	xp622	1.31	0.00	2.50	3.00			
8	Tree/Tree	capse1	phisc1	xp_002				
	capse1	3.00	1.92	1.08				
	phisc1	1.92	3.00	1.13				
	xp_002	1.08	1.13	3.00				
9	Tree/Tree	conap1	dipse1	exool1	hypco275	tgo886	thv476	thv560
	conap1	3.0	2.3	1.7	0.0	1.6	1.4	1.0
	dipse1	2.3	3.0	0.0	2.3	1.3	1.9	0.0
	exool1	1.7	0.0	3.0	0.0	1.4	1.5	1.4
	hypco275	0.0	2.3	0.0	3.0	0.0	1.6	0.0
	tgo886	1.6	1.3	1.4	0.0	3.0	2.3	1.8
	thv476	1.4	1.9	1.5	1.6	2.3	3.0	2.0
	thv560	1.0	0.0	1.4	0.0	1.8	2.0	3.0
10	Tree/Tree	exoxe	ppr05	tgo22				
	exoxe	3.00	1.50	0.00				
	ppr05	1.50	3.00	1.00				
	tgo22	0.00	1.00	3.00				
11	Tree/Tree	cdm	colsi	dalec	glost	hypco	xp_013	xp_018
	cdm	3.00	1.00	0.00	1.88	0.00	0.00	0.00
	colsi	1.00	3.00	1.42	1.56	1.63	1.88	1.33
	dalec	0.00	1.42	3.00	1.44	1.38	1.25	1.44
	glost	1.88	1.56	1.44	3.00	2.16	0.00	0.00
	hypco	0.00	1.63	1.38	2.16	3.00	0.00	0.00
	xp_013	0.00	1.88	1.25	0.00	0.00	3.00	1.85
	xp_018	0.00	1.33	1.44	0.00	0.00	1.85	3.00
12	Tree/Tree	colny1_101	colny1_102	maggr1	neute	paevar		
	colny1_101	3.00	2.44	1.38	2.21	0.00		
	colny1_102	2.44	3.00	1.13	1.85	0.00		
	maggr1	1.38	1.13	3.00	0.00	1.00		
	neute	2.21	1.85	0.00	3.00	0.00		
	paevar	0.00	0.00	1.00	0.00	3.00		
13	Tree/Tree	color	melva	rdk	xp_002	xp_014		

	color	3.00	1.38	0.00	0.00	0.00	
	melva	1.38	3.00	0.00	0.00	0.00	
	rdk	0.00	0.00	3.00	2.44	0.00	
	xp_002	0.00	0.00	2.44	3.00	0.00	
	xp_014	0.00	0.00	0.00	0.00	3.00	
14	Tree/Tree	exool	gaegr	melbi2_5	melbi2_6	rii2	zymps
	exool	3.00	0.00	0.00	2.50	0.00	0.00
	gaegr	0.00	3.00	0.00	0.00	0.00	0.00
	melbi2_5	0.00	0.00	3.00	1.58	0.00	0.00
	melbi2_6	2.50	0.00	1.58	3.00	1.65	0.00
	rii2	0.00	0.00	0.00	1.65	3.00	0.00
	zymps	0.00	0.00	0.00	0.00	0.00	3.00
15	Tree/Tree	neucr2	oaa50	xp01			
	neucr2	3.00	1.53	0.00			
	oaa50	1.53	3.00	0.00			
	xp01	0.00	0.00	3.00			
16	Tree/Tree	colsa1	melva1	xp0113			
	colsa1	3.00	0.00	1.25			
	melva1	0.00	3.00	1.27			
	xp0113	1.25	1.27	3.00			
17	Tree/Tree	cadsp1	capse1	oidma1	tgo074		
	cadsp1	3.00	1.59	1.28	0.00		
	capse1	1.59	3.00	1.00	0.00		
	oidma1	1.28	1.00	3.00	1.50		
	tgo074	0.00	0.00	1.50	3.00		
18	Tree/Tree	conlig	eit7	eit8	melva	neucr	xp_001	xp_018
	conlig	3.00	0.00	0.00	1.34	0.00	0.00	2.13
	eit7	0.00	3.00	2.13	0.00	1.78	1.88	1.85
	eit8	0.00	2.13	3.00	1.88	0.00	2.15	0.00
	melva	1.34	0.00	1.88	3.00	1.33	1.25	0.00
	neucr	0.00	1.78	0.00	1.33	3.00	0.00	1.96
	xp_001	0.00	1.88	2.15	1.25	0.00	3.00	0.00
	xp_018	2.13	1.85	0.00	0.00	1.96	0.00	3.00
19	Tree/Tree	colny	kid	pseve	rii	xp_001	xp_011	
	colny	3.00	0.00	2.13	0.00	0.00	0.00	
	kid	0.00	3.00	1.13	1.19	1.21	1.43	
	pseve	2.13	1.13	3.00	0.00	0.00	1.43	
	rii	0.00	1.19	0.00	3.00	1.08	1.63	
	xp_001	0.00	1.21	0.00	1.08	3.00	1.82	
	xp_011	0.00	1.43	1.43	1.63	1.82	3.00	
20	Tree/Tree	cochec	cocvi	colsi	crl19	por37		
	cochec	3.00	2.03	0.00	1.04	0.00		
	cocvi	2.03	3.00	0.00	1.00	1.13		
	colsi	0.00	0.00	3.00	0.00	1.58		
	crl19	1.04	1.00	0.00	3.00	0.00		
	por37	0.00	1.13	1.58	0.00	3.00		
21	Tree/Tree	hgy_1_10	hgy_1_23	xp_0187				

	hgy_1_10	3.00	1.75	0.00					
	hgy_1_23	1.75	3.00	1.69					
	xp_0187	0.00	1.69	3.00					
22	Tree/Tree	claim	colsi	monha	xp_002	xp_003			
	claim	3.00	0.00	0.00	1.50	2.10			
	colsi	0.00	3.00	1.44	1.04	0.00			
	monha	0.00	1.44	3.00	1.15	1.88			
	xp_002	1.50	1.04	1.15	3.00	1.56			
	xp_003	2.10	0.00	1.88	1.56	3.00			
23	Tree/Tree	colny1	crl18	exoxe1	xp_0234				
	colny1	3.00	0.00	1.88	0.00				
	crl18	0.00	3.00	1.88	2.00				
	exoxe1	1.88	1.88	3.00	1.50				
	xp_0234	0.00	2.00	1.50	3.00				
24	Tree/Tree	gaegr	maggr	magpo	neucr	oqe	ptb	xp_007	
	gaegr	3.00	1.25	2.63	0.00	0.00	0.00	1.25	
	maggr	1.25	3.00	1.45	1.00	0.00	0.00	1.70	
	magpo	2.63	1.45	3.00	1.00	0.00	1.33	1.89	
	neucr	0.00	1.00	1.00	3.00	0.00	1.00	1.00	
	oqe	0.00	0.00	0.00	0.00	3.00	1.75	0.00	
	ptb	0.00	0.00	1.33	1.00	1.75	3.00	2.00	
	xp_007	1.25	1.70	1.89	1.00	0.00	2.00	3.00	
25	Tree/Tree	botdo	capse	chove	kho	perma	tgo	tubae	xp_028
	botdo	3.00	1.88	1.60	0.00	1.69	1.50	1.13	1.77
	capse	1.88	3.00	0.00	0.00	1.25	0.00	0.00	1.38
	chove	1.60	0.00	3.00	0.00	1.81	1.45	1.88	1.81
	kho	0.00	0.00	0.00	3.00	1.67	0.00	1.28	0.00
	perma	1.69	1.25	1.81	1.67	3.00	1.45	0.00	1.63
	tgo	1.50	0.00	1.45	0.00	1.45	3.00	1.23	1.40
	tubae	1.13	0.00	1.88	1.28	0.00	1.23	3.00	1.06
	xp_028	1.77	1.38	1.81	0.00	1.63	1.40	1.06	3.00
26	Tree/Tree	cochec	kil	paevar	phisc	tgo	xp_009	xp_025	
	cochec	3.00	0.00	0.00	1.13	0.00	1.38	0.00	
	kil	0.00	3.00	0.00	0.00	0.00	0.00	0.00	
	paevar	0.00	0.00	3.00	1.72	1.50	0.00	1.32	
	phisc	1.13	0.00	1.72	3.00	1.22	1.94	1.88	
	tgo	0.00	0.00	1.50	1.22	3.00	0.00	0.00	
	xp_009	1.38	0.00	0.00	1.94	0.00	3.00	0.00	
	xp_025	0.00	0.00	1.32	1.88	0.00	0.00	3.00	
27	Tree/Tree	color	corca	rfu74	venin	xp_007			
	color	3.00	1.31	1.71	2.08	1.28			
	corca	1.31	3.00	1.31	1.77	1.00			
	rfu74	1.71	1.31	3.00	0.00	1.46			
	venin	2.08	1.77	0.00	3.00	0.00			
	xp_007	1.28	1.00	1.46	0.00	3.00			
28	Tree/Tree	erynec	oidma	paevar	verda	verga	xp_01		
	erynec	3.00	1.06	0.00	1.31	1.75	1.09		

	oidma	1.06	3.00	0.00	1.88	1.38	1.47	
	paevar	0.00	0.00	3.00	0.00	0.00	1.38	
	verda	1.31	1.88	0.00	3.00	0.00	1.33	
	verga	1.75	1.38	0.00	0.00	3.00	0.00	
	xp_01	1.09	1.47	1.38	1.33	0.00	3.00	
29	Tree/Tree	cdm	exode	kzn	hgy_1_18	hgy_1_29	xp_002	xp_007
	cdm	3.00	1.33	2.08	1.60	1.19	1.85	0.00
	exode	1.33	3.00	1.75	1.25	0.00	0.00	0.00
	kzn	2.08	1.75	3.00	1.84	1.25	1.38	0.00
	hgy_1_18	1.60	1.25	1.84	3.00	1.75	1.30	0.00
	hgy_1_29	1.19	0.00	1.25	1.75	3.00	1.25	1.00
	xp_002	1.85	0.00	1.38	1.30	1.25	3.00	0.00
	xp_007	0.00	0.00	0.00	0.00	1.00	0.00	3.00
30	Tree/Tree	hgy_1_97	hgy_1_24	xp_003				
	hgy_1_97	3.00	2.09	2.20				
	hgy_1_24	2.09	3.00	1.96				
	xp_003	2.20	1.96	3.00				
31	Tree/Tree	cdm	knd	necha	paevar	rfu	xp_018	xp_025	xylhe
	cdm	3.00	0.00	0.00	1.88	0.00	0.00	1.63	0.00
	knd	0.00	3.00	1.96	0.00	2.13	1.80	0.00	1.97
	necha	0.00	1.96	3.00	0.00	1.67	2.25	0.00	1.50
	paevar	1.88	0.00	0.00	3.00	0.00	0.00	1.88	0.00
	rfu	0.00	2.13	1.67	0.00	3.00	1.69	0.00	1.67
	xp_018	0.00	1.80	2.25	0.00	1.69	3.00	0.00	1.50
	xp_025	1.63	0.00	0.00	1.88	0.00	0.00	3.00	0.00
	xylhe	0.00	1.97	1.50	0.00	1.67	1.50	0.00	3.00
32	Tree/Tree	penvul	xp_0183	xp_0245					
	penvul	3.00	0.00	1.38					
	xp_0183	0.00	3.00	0.00					
	xp_0245	1.38	0.00	3.00					
33	Tree/Tree	amore	colny	ryo	xp_007	xp_014			
	amore	3.00	1.83	0.00	2.00	1.75			
	colny	1.83	3.00	0.00	0.00	0.00			
	ryo	0.00	0.00	3.00	0.00	0.00			
	xp_007	2.00	0.00	0.00	3.00	1.21			
	xp_014	1.75	0.00	0.00	1.21	3.00			
34	Tree/Tree	capse	hypco	venin	xp_01	xp_02			
	capse	3.00	0.00	0.00	0.00	1.85			
	hypco	0.00	3.00	0.00	1.98	0.00			
	venin	0.00	0.00	3.00	0.00	1.19			
	xp_01	0.00	1.98	0.00	3.00	0.00			
	xp_02	1.85	0.00	1.19	0.00	3.00			
35	Tree/Tree	ascim	conlig	exode	phach	xp_02			
	ascim	3.00	0.00	0.00	0.00	0.00			
	conlig	0.00	3.00	1.65	2.04	1.75			
	exode	0.00	1.65	3.00	1.63	0.00			
	phach	0.00	2.04	1.63	3.00	0.00			

	xp_02	0.00	1.75	0.00	0.00	3.00		
36	Tree/Tree	colsi	knd	paevar	xp_01	xp_02		
	colsi	3.00	2.00	1.63	1.78	0.00		
	knd	2.00	3.00	0.00	1.94	1.70		
	paevar	1.63	0.00	3.00	0.00	0.00		
	xp_01	1.78	1.94	0.00	3.00	1.90		
	xp_02	0.00	1.70	0.00	1.90	3.00		
37	Tree/Tree	chagl	corca1	kid67	phach1			
	chagl	3.00	0.00	1.69	0.00			
	corca1	0.00	3.00	0.00	1.50			
	kid67	1.69	0.00	3.00	0.00			
	phach1	0.00	1.50	0.00	3.00			
38	Tree/Tree	ptb80	ryo62	xp_003	xp_013			
	ptb80	3.00	0.00	1.21	1.63			
	ryo62	0.00	3.00	0.00	1.28			
	xp_003	1.21	0.00	3.00	1.75			
	xp_013	1.63	1.28	1.75	3.00			
39	Tree/Tree	cadsp	dalec	kil	thv16	thv50		
	cadsp	3.00	1.75	1.75	1.75	1.88		
	dalec	1.75	3.00	0.00	0.00	0.00		
	kil	1.75	0.00	3.00	1.75	0.00		
	thv16	1.75	0.00	1.75	3.00	2.44		
	thv50	1.88	0.00	0.00	2.44	3.00		
40	Tree/Tree	acrchr	claba	melva	monha	oaa	paevar	
	acrchr	3.00	0.00	0.00	0.00	1.40	0.00	
	claba	0.00	3.00	1.13	0.00	0.00	1.48	
	melva	0.00	1.13	3.00	1.44	0.00	1.13	
	monha	0.00	0.00	1.44	3.00	0.00	0.00	
	oaa	1.40	0.00	0.00	0.00	3.00	0.00	
	paevar	0.00	1.48	1.13	0.00	0.00	3.00	
41	Tree/Tree	arb_038	claim1	neucr	rdk4	xp_007		
	arb_038	3.00	1.63	1.81	1.63	0.00		
	claim1	1.63	3.00	0.00	2.18	0.00		
	neucr	1.81	0.00	3.00	1.83	2.45		
	rdk4	1.63	2.18	1.83	3.00	0.00		
	xp_007	0.00	0.00	2.45	0.00	3.00		
42	Tree/Tree	hgy_1_262	xp_0187	xp_0247				
	hgy_1_262	3.00	0.00	0.00				
	xp_0187	0.00	3.00	1.48				
	xp_0247	0.00	1.48	3.00				
43	Tree/Tree	claim1	hgy_1	xp_001	xp_0255			
	claim1	3.00	1.13	0.00	0.00			
	hgy_1	1.13	3.00	1.28	1.68			
	xp_001	0.00	1.28	3.00	2.03			
	xp_0255	0.00	1.68	2.03	3.00			
44	Tree/Tree	magpo	paevar	verda	verga	xp_001	xp_014	xp_024
	magpo	3.00	0.00	2.13	1.88	0.00	0.00	0.00

	paevar	0.00	3.00	0.00	0.00	1.85	1.81	0.00
	verda	2.13	0.00	3.00	1.67	0.00	0.00	1.77
	verga	1.88	0.00	1.67	3.00	0.00	0.00	0.00
	xp_001	0.00	1.85	0.00	0.00	3.00	1.65	0.00
	xp_014	0.00	1.81	0.00	0.00	1.65	3.00	0.00
	xp_024	0.00	0.00	1.77	0.00	0.00	0.00	3.00

45	Tree/Tree	claim	oaa	paevar	xp_013	xp_014	xp_018	xp_025
	claim	3.00	0.00	1.19	0.00	0.00	0.00	1.04
	oaa	0.00	3.00	0.00	1.83	2.05	1.46	0.00
	paevar	1.19	0.00	3.00	0.00	0.00	0.00	1.80
	xp_013	0.00	1.83	0.00	3.00	1.75	2.00	0.00
	xp_014	0.00	2.05	0.00	1.75	3.00	1.50	0.00
	xp_018	0.00	1.46	0.00	2.00	1.50	3.00	0.00
	xp_025	1.04	0.00	1.80	0.00	0.00	0.00	3.00

46	Tree/Tree	ascim1	necha2	neute	verga1	xp_003	xp_018	xp_023	xp_024
	ascim1	3.00	0.00	0.00	1.25	0.00	0.00	0.00	0.00
	necha2	0.00	3.00	1.25	0.00	1.27	1.76	1.61	1.69
	neute	0.00	1.25	3.00	1.88	1.25	1.50	1.28	1.50
	verga1	1.25	0.00	1.88	3.00	0.00	1.94	1.75	1.50
	xp_003	0.00	1.27	1.25	0.00	3.00	1.13	1.10	1.15
	xp_018	0.00	1.76	1.50	1.94	1.13	3.00	2.45	1.99
	xp_023	0.00	1.61	1.28	1.75	1.10	2.45	3.00	1.75
	xp_024	0.00	1.69	1.50	1.50	1.15	1.99	1.75	3.00

47	Tree/Tree	crl27	hypec38	paevar	tgo822	xp_024
	crl27	3.00	0.00	1.81	0.00	0.00
	hypec38	0.00	3.00	0.00	0.00	1.00
	paevar	1.81	0.00	3.00	0.00	0.00
	tgo822	0.00	0.00	0.00	3.00	2.83
	xp_024	0.00	1.00	0.00	2.83	3.00

48	Tree/Tree	khn98	neute	hgy_1_2	hgy_1_3	xp_02
	khn98	3.00	0.00	0.00	0.00	1.23
	neute	0.00	3.00	0.00	0.00	1.78
	hgy_1_2	0.00	0.00	3.00	2.41	1.92
	hgy_1_3	0.00	0.00	2.41	3.00	1.63
	xp_02	1.23	1.78	1.92	1.63	3.00

49	Tree/Tree	acrchr1	corca1	paevar	xp_001	xp_024
	acrchr1	3.00	2.69	0.00	0.00	2.19
	corca1	2.69	3.00	1.88	1.75	3.00
	paevar	0.00	1.88	3.00	1.63	0.00
	xp_001	0.00	1.75	1.63	3.00	0.00
	xp_024	2.19	3.00	0.00	0.00	3.00

50	Tree/Tree	capse1	claba1	gaegr1	rii08
	capse1	3.00	1.99	1.77	1.00
	claba1	1.99	3.00	1.50	1.25
	gaegr1	1.77	1.50	3.00	1.41
	rii08	1.00	1.25	1.41	3.00

51	Tree/Tree	arb	ascim	capse	magpo	thv	xp_001

	arb	3.00	0.00	1.41	1.50	1.00	1.44		
	ascim	0.00	3.00	0.00	0.00	0.00	0.00		
	capse	1.41	0.00	3.00	1.00	1.00	1.88		
	magpo	1.50	0.00	1.00	3.00	0.00	0.00		
	thv	1.00	0.00	1.00	0.00	3.00	1.00		
	xp_001	1.44	0.00	1.88	0.00	1.00	3.00		
52	Tree/Tree	botdo	cadsp	chove	maggr	oaa	xp_013		
	botdo	3.00	1.50	0.00	1.40	1.19	1.31		
	cadsp	1.50	3.00	0.00	1.46	0.00	0.00		
	chove	0.00	0.00	3.00	0.00	0.00	0.00		
	maggr	1.40	1.46	0.00	3.00	1.77	1.13		
	oaa	1.19	0.00	0.00	1.77	3.00	1.24		
	xp_013	1.31	0.00	0.00	1.13	1.24	3.00		
53	Tree/Tree	conap1	fonmo1	hgy_1_1	hgy_1_2	xp_01			
	conap1	3.00	1.50	0.00	0.00	1.63			
	fonmo1	1.50	3.00	1.00	1.00	0.00			
	hgy_1_1	0.00	1.00	3.00	1.73	0.00			
	hgy_1_2	0.00	1.00	1.73	3.00	0.00			
	xp_01	1.63	0.00	0.00	0.00	3.00			
54	Tree/Tree	ryo557	xp_0013	xp_0077	xp_0078	xp_0187			
	ryo557	3.00	0.00	0.00	0.00	1.42			
	xp_0013	0.00	3.00	0.00	0.00	1.75			
	xp_0077	0.00	0.00	3.00	1.47	2.04			
	xp_0078	0.00	0.00	1.47	3.00	0.00			
	xp_0187	1.42	1.75	2.04	0.00	3.00			
55	Tree/Tree	exoxe1	Pseudest1	xp_016	xp_018	xp_023			
	exoxe1	3.00	2.00	0.00	1.88	0.00			
	Pseudest1	2.00	3.00	1.15	0.00	0.00			
	xp_016	0.00	1.15	3.00	0.00	0.00			
	xp_018	1.88	0.00	0.00	3.00	2.40			
	xp_023	0.00	0.00	0.00	2.40	3.00			
56	Tree/Tree	colny	conlig	exool	hgy_1_20	hgy_1_25	xp_009	xp_023	
	colny	3.00	1.73	1.21	0.00	0.00	1.28	1.27	
	conlig	1.73	3.00	1.71	0.00	0.00	1.25	1.38	
	exool	1.21	1.71	3.00	1.44	0.00	1.25	1.30	
	hgy_1_20	0.00	0.00	1.44	3.00	2.63	0.00	0.00	
	hgy_1_25	0.00	0.00	0.00	2.63	3.00	0.00	0.00	
	xp_009	1.28	1.25	1.25	0.00	0.00	3.00	1.63	
	xp_023	1.27	1.38	1.30	0.00	0.00	1.63	3.00	
57	Tree/Tree	cadsp	erynec	melva	hgy_1_19	hgy_1_21	xp_016	xp_025	zymps
	cadsp	3.00	0.00	1.19	1.17	0.00	0.00	0.00	1.25
	erynec	0.00	3.00	1.00	0.00	0.00	0.00	0.00	0.00
	melva	1.19	1.00	3.00	1.75	0.00	1.63	0.00	1.52
	hgy_1_19	1.17	0.00	1.75	3.00	1.92	1.00	1.19	1.48
	hgy_1_21	0.00	0.00	0.00	1.92	3.00	1.34	1.82	1.44
	xp_016	0.00	0.00	1.63	1.00	1.34	3.00	1.44	1.75
	xp_025	0.00	0.00	0.00	1.19	1.82	1.44	3.00	1.75

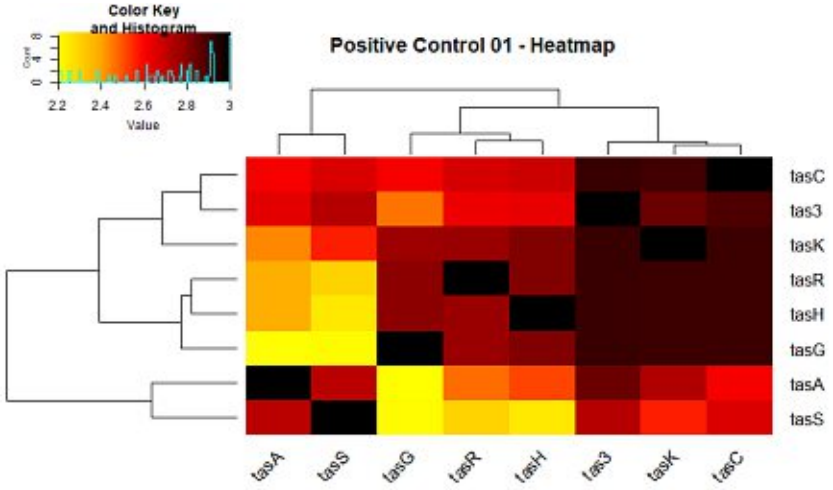
	zymps	1.25	0.00	1.52	1.48	1.44	1.75	1.75	3.00

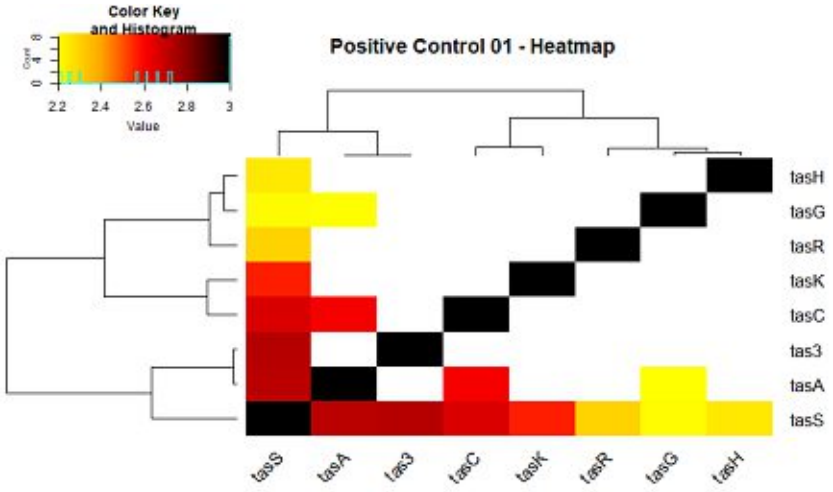
58	Tree/Tree	parbra1	phaal1	venin	xp0012	
	parbra1	3.00	1.63	1.75	1.66	
	phaal1	1.63	3.00	1.54	0.54	
	venin	1.75	1.54	3.00	0.91	
	xp0012	1.66	0.54	0.91	3.00	

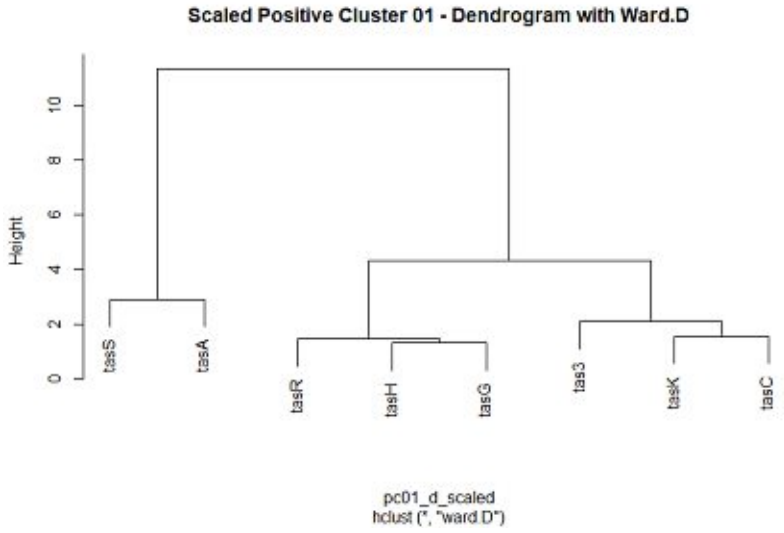
59	Tree/Tree	colsa	paevar	xp_001	xp_003	xp_007
	colsa	3.00	0.00	0.00	1.80	0.34
	paevar	0.00	3.00	1.71	0.00	2.13
	xp_001	0.00	1.71	3.00	0.00	0.00
	xp_003	1.80	0.00	0.00	3.00	1.00
	xp_007	0.34	2.13	0.00	1.00	3.00

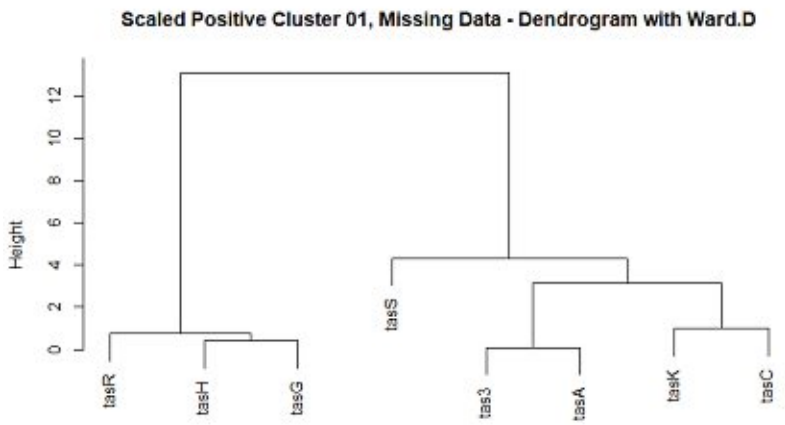
60	Tree/Tree	MeBi	MeVa
	MeBi	3.00	1.50
	MeVa	1.50	3.00

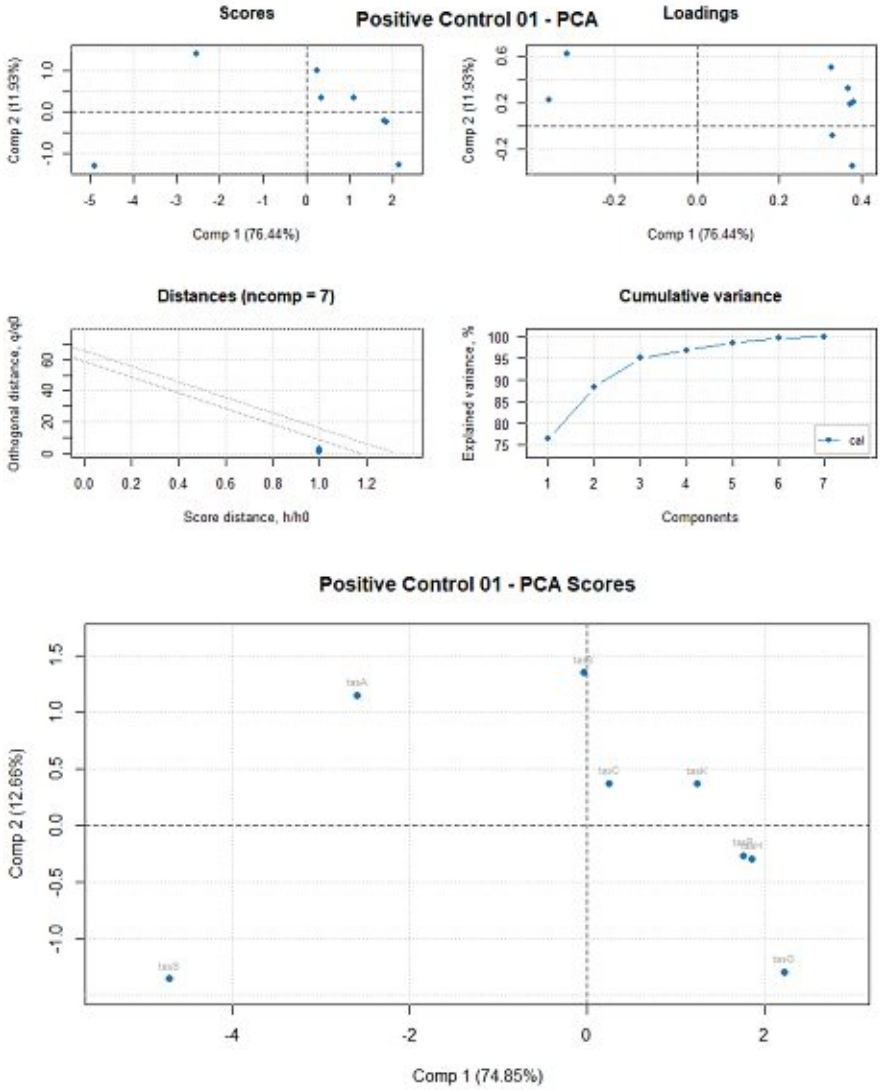
Supplement 9.33: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 01 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.



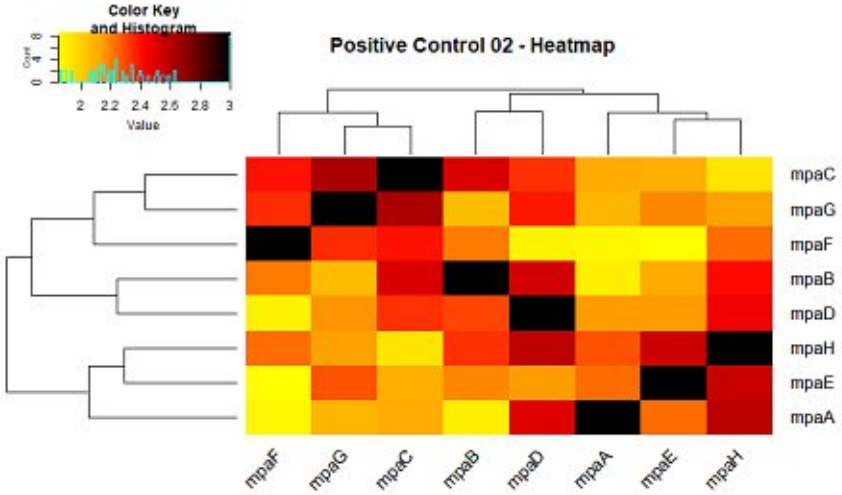


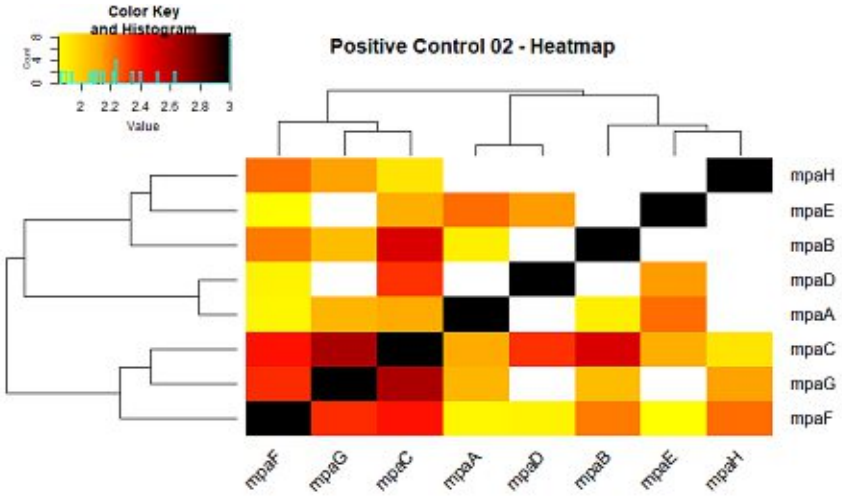


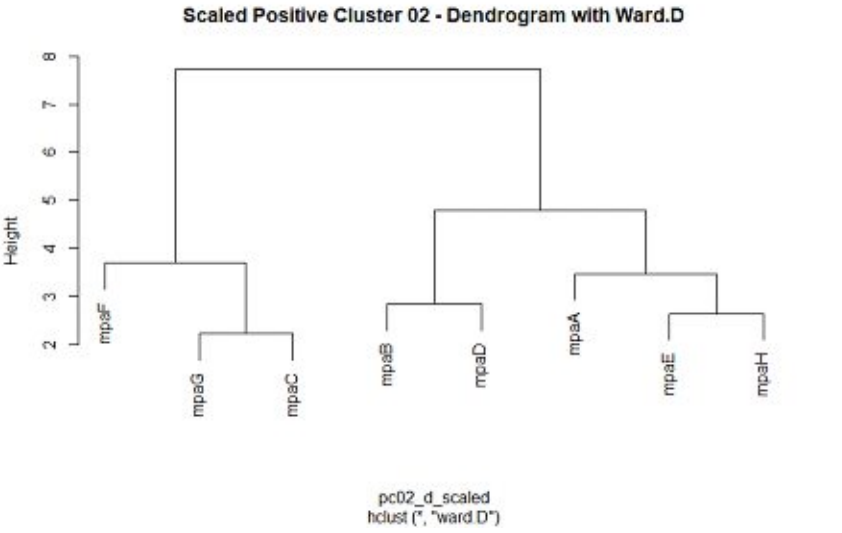


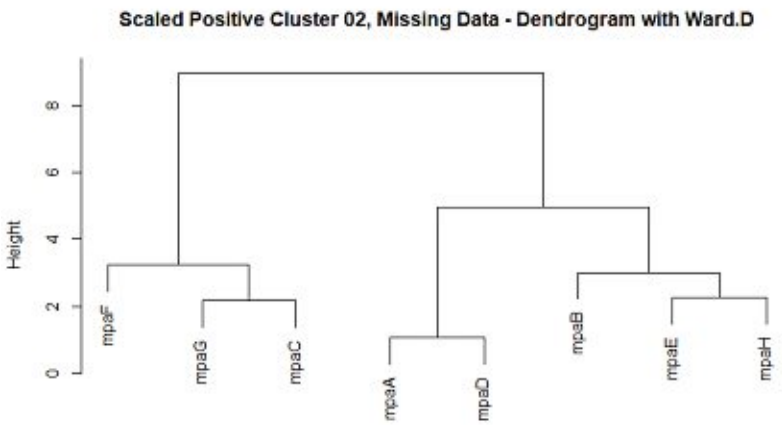


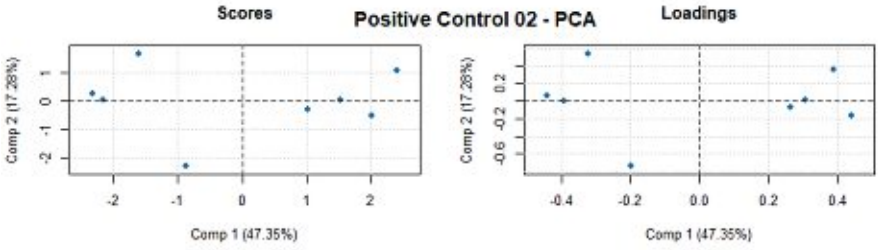
Supplement 9.34: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 02 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

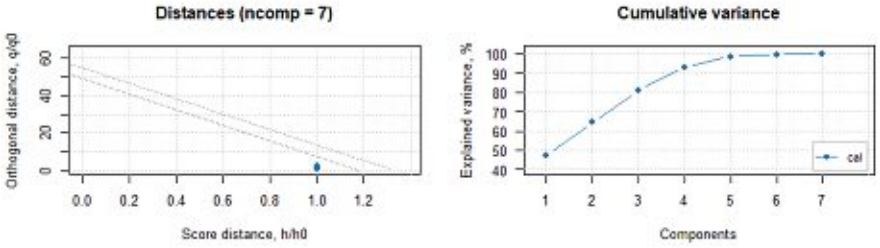


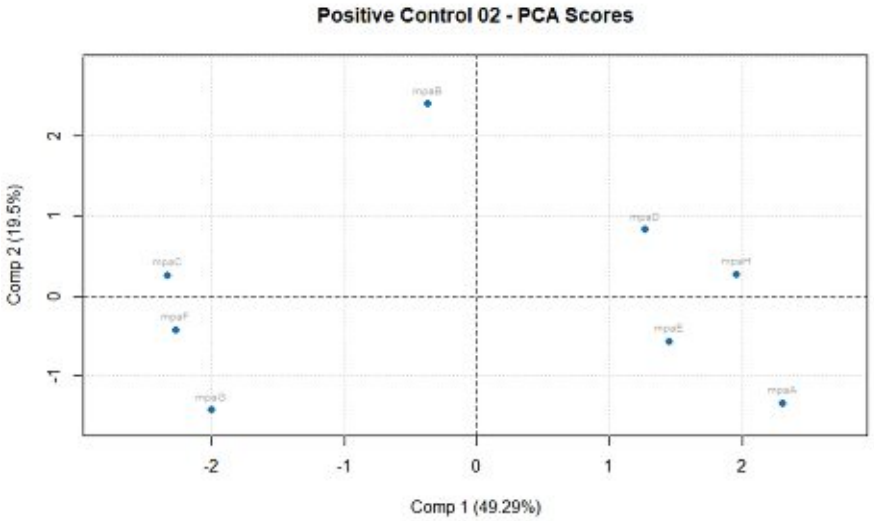




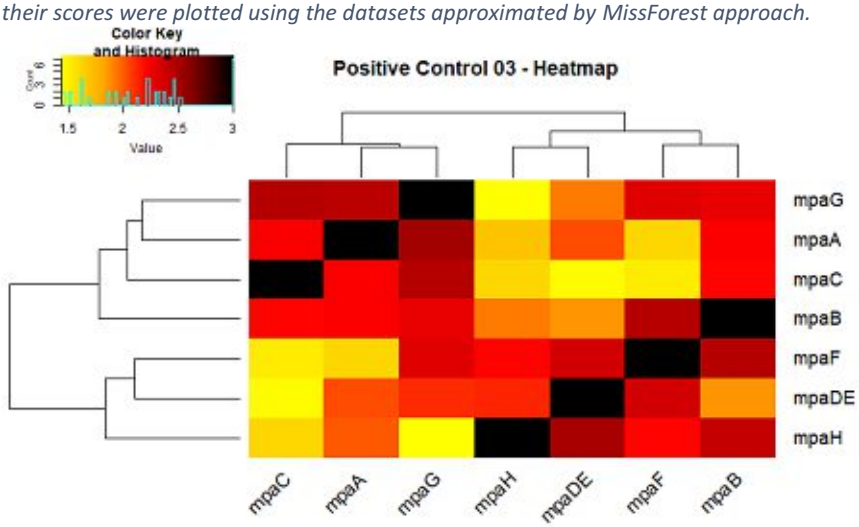


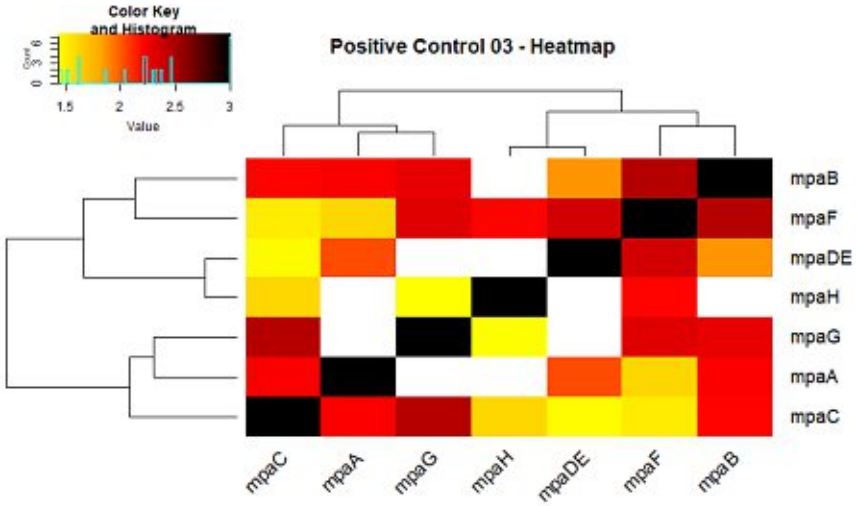


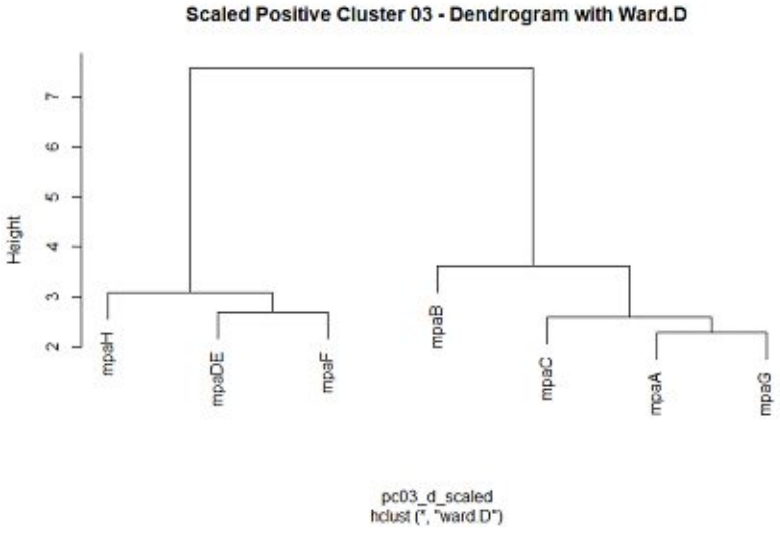


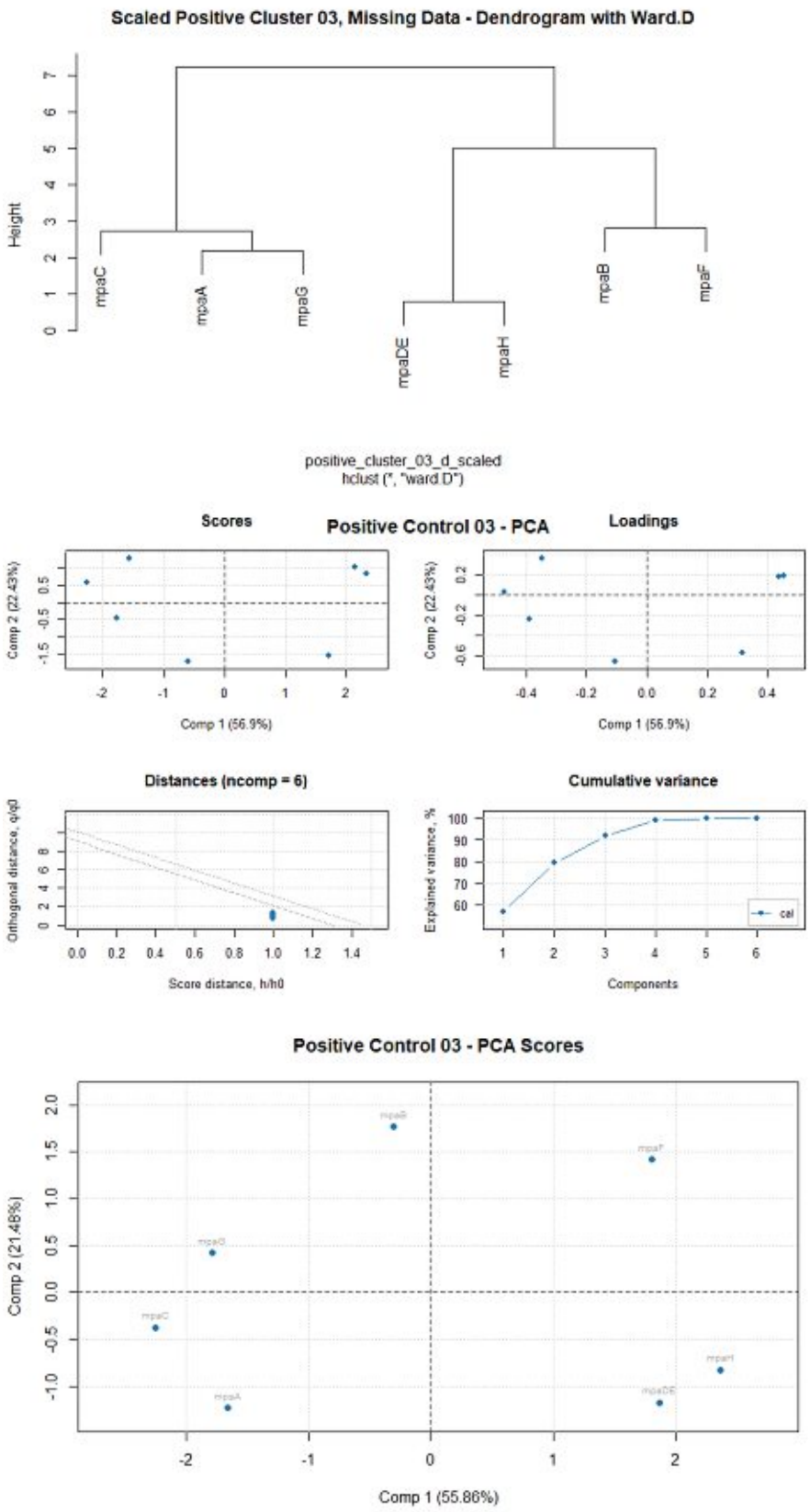


Supplement 9.35: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 03 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

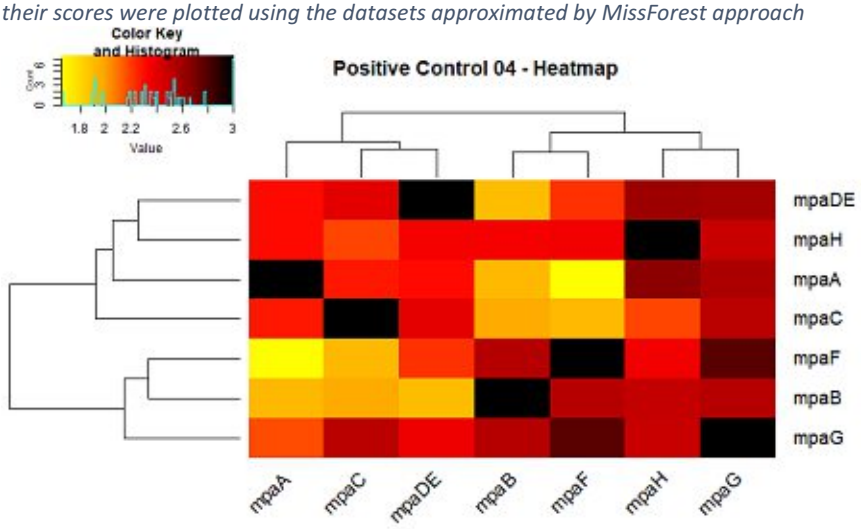


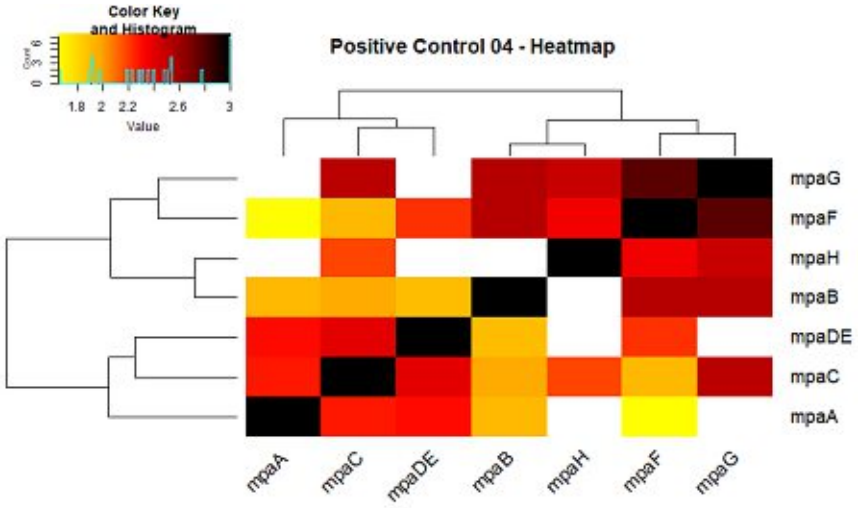


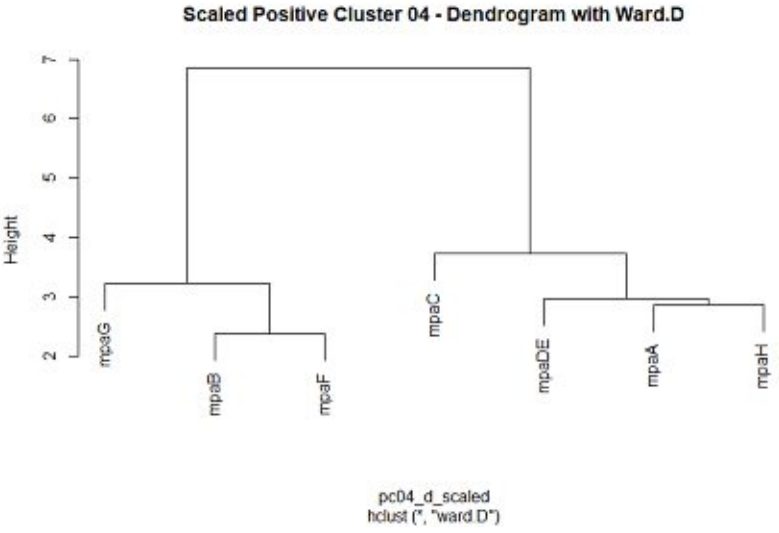


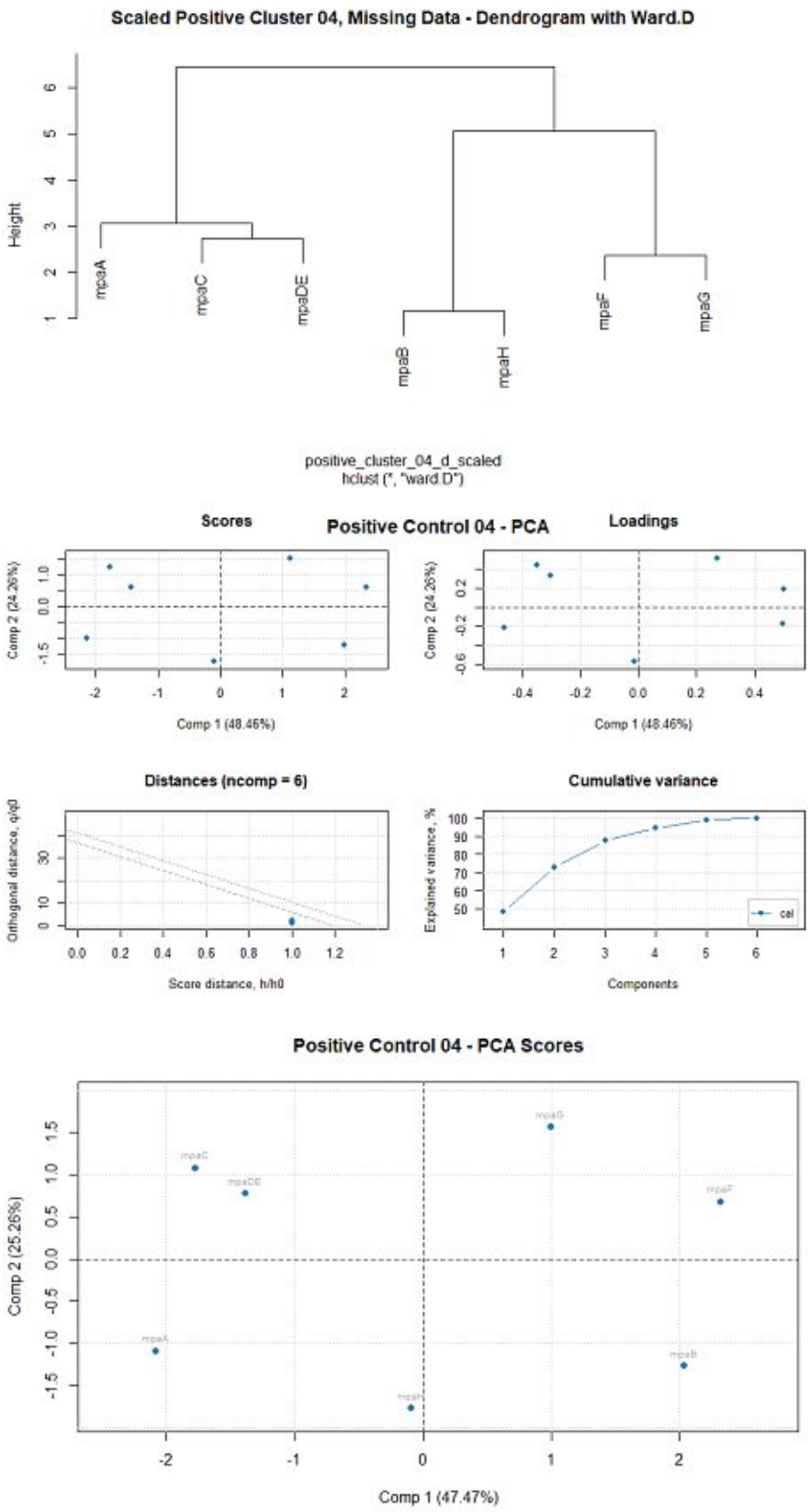


Supplement 9.36: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 04 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach

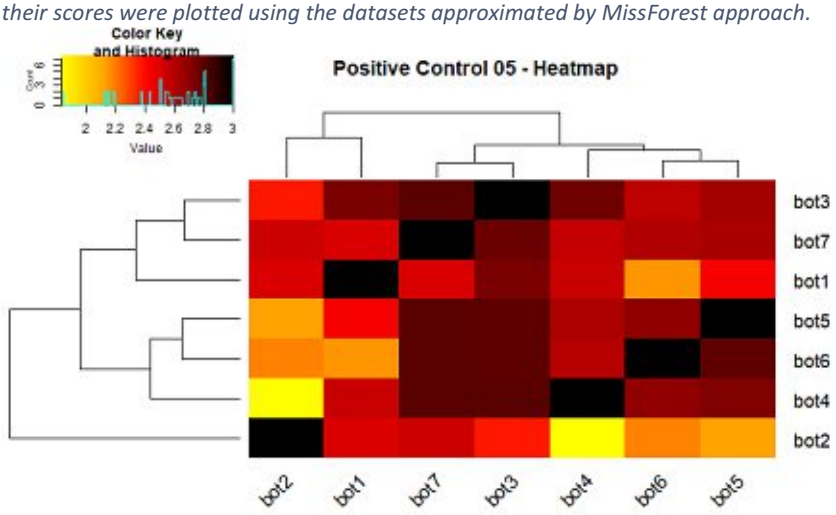


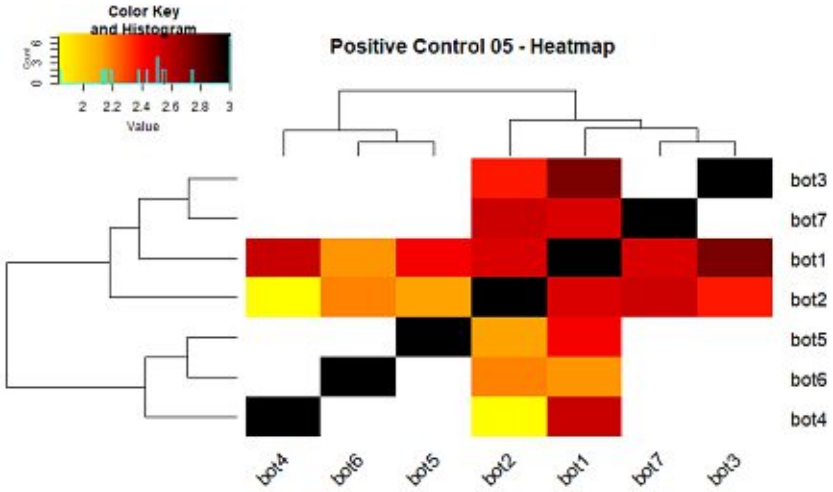


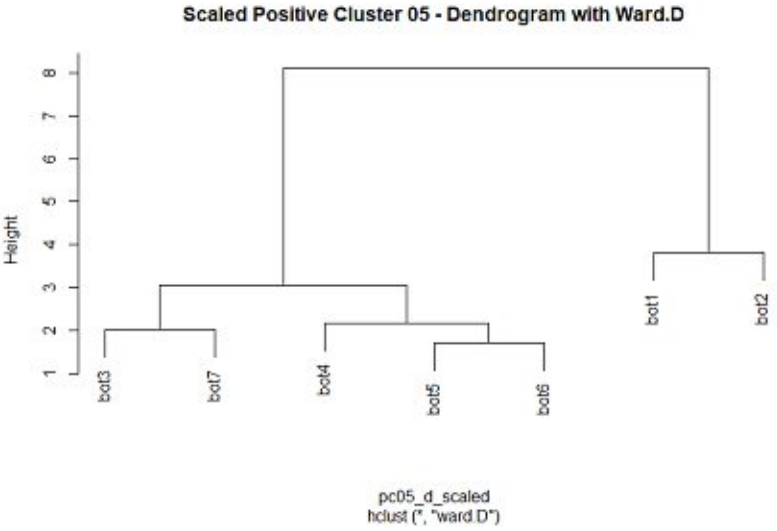


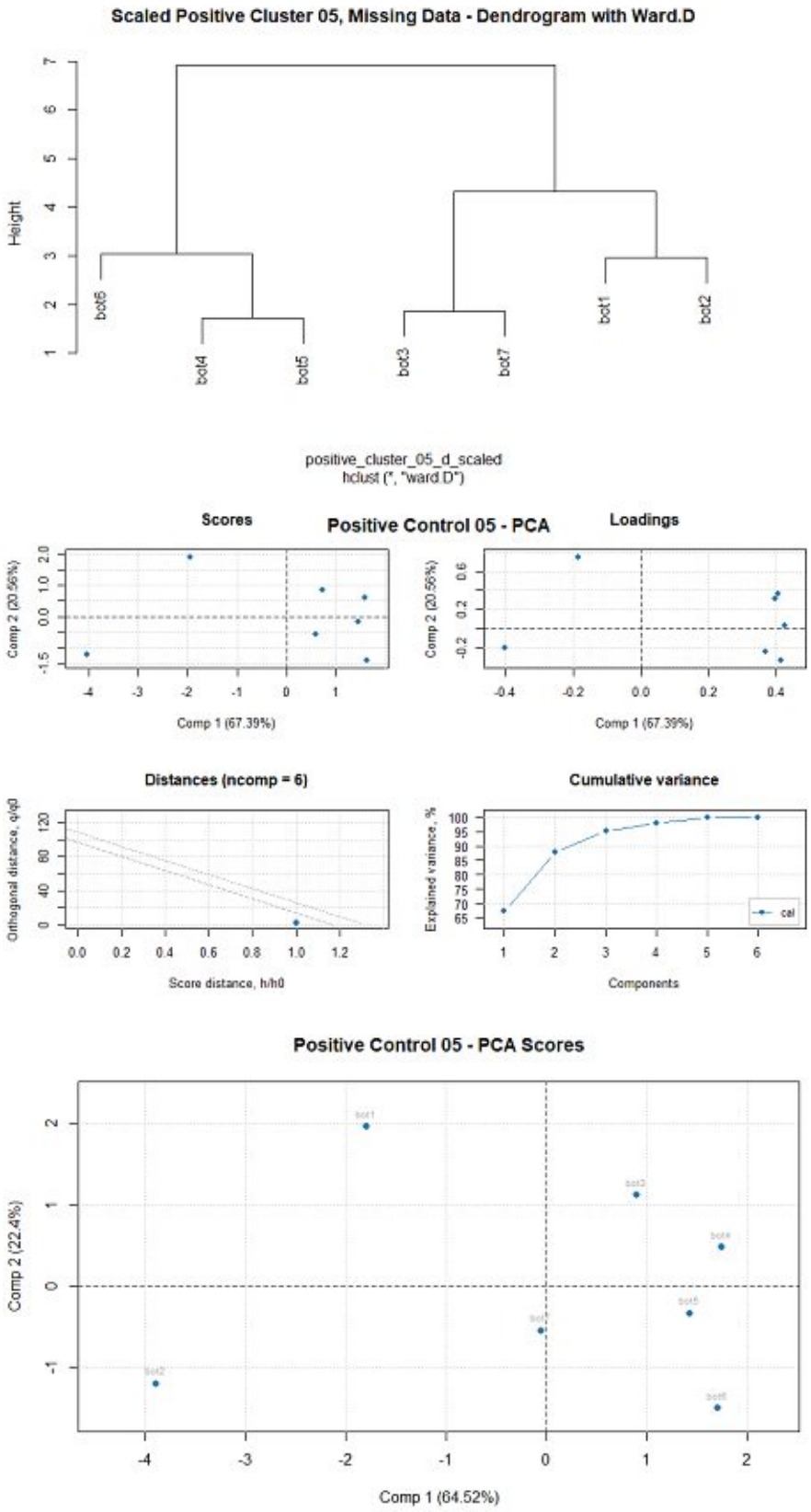


Supplement 9.37: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 05 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

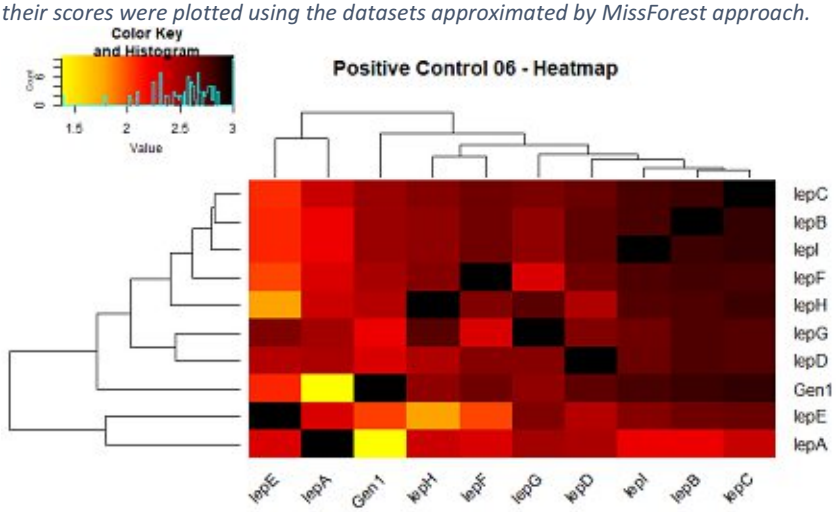


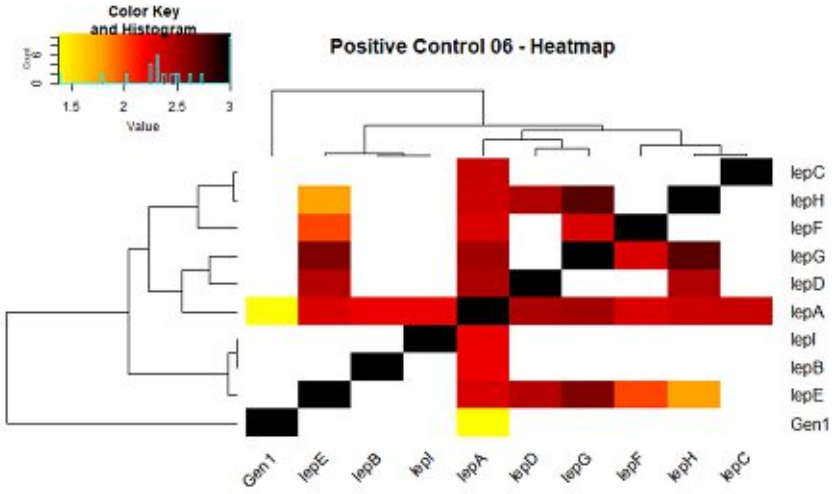


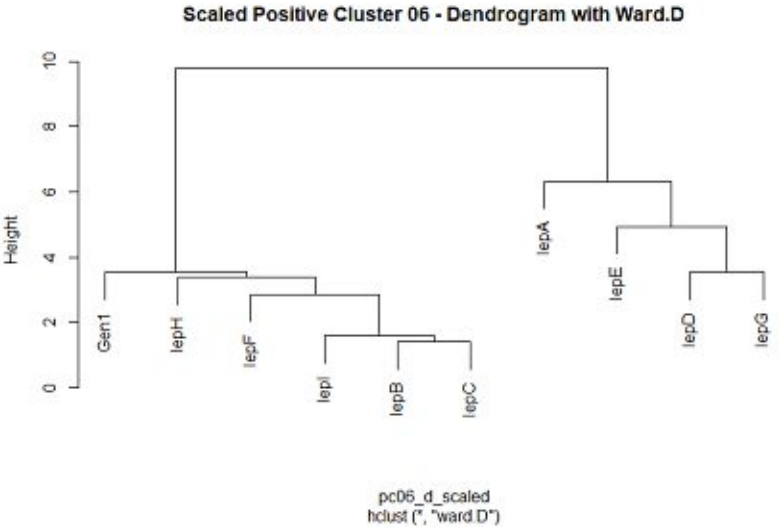


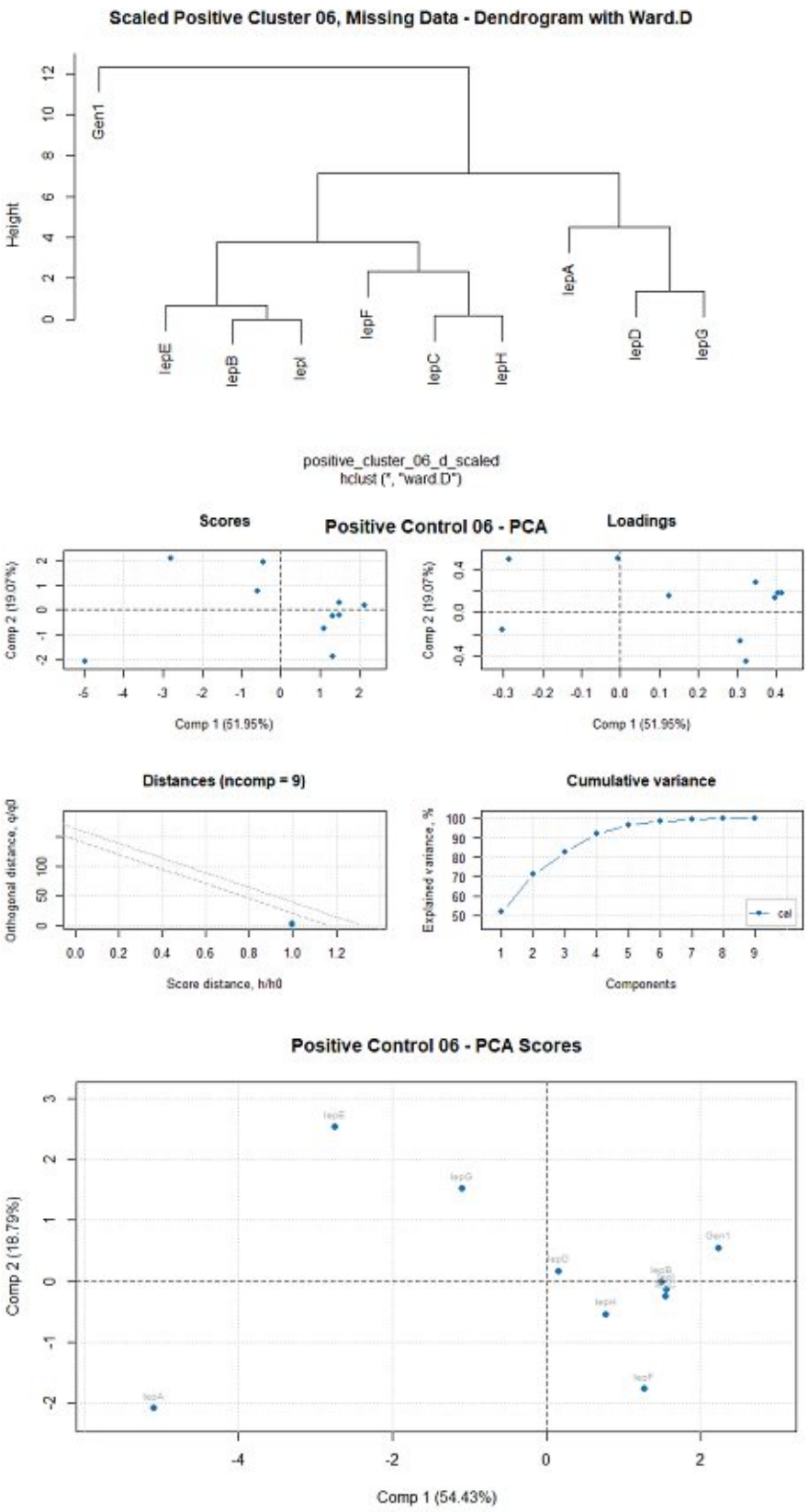


Supplement 9.38: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 06 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

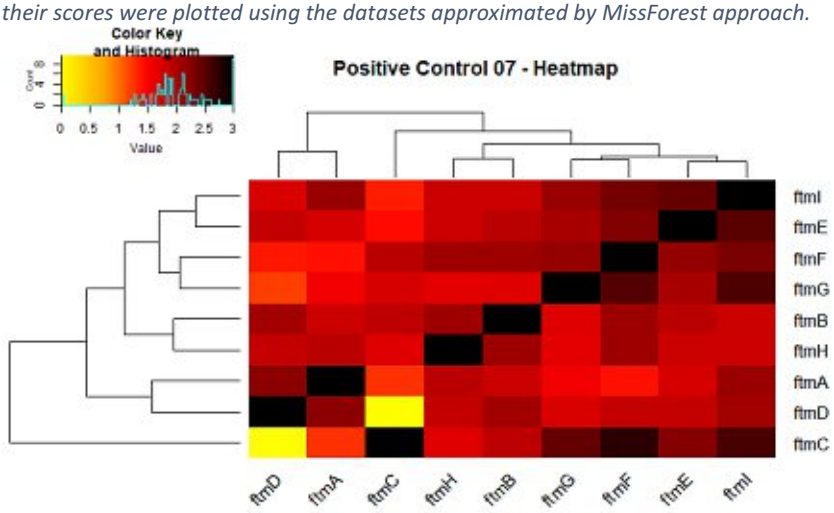


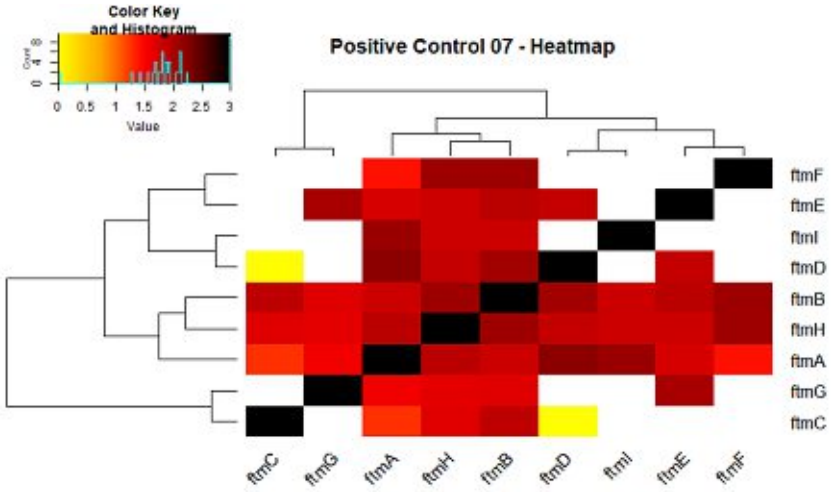


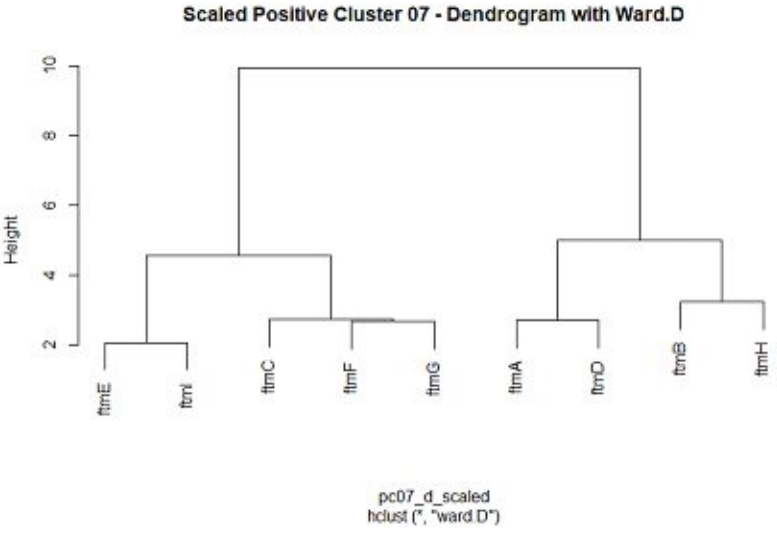


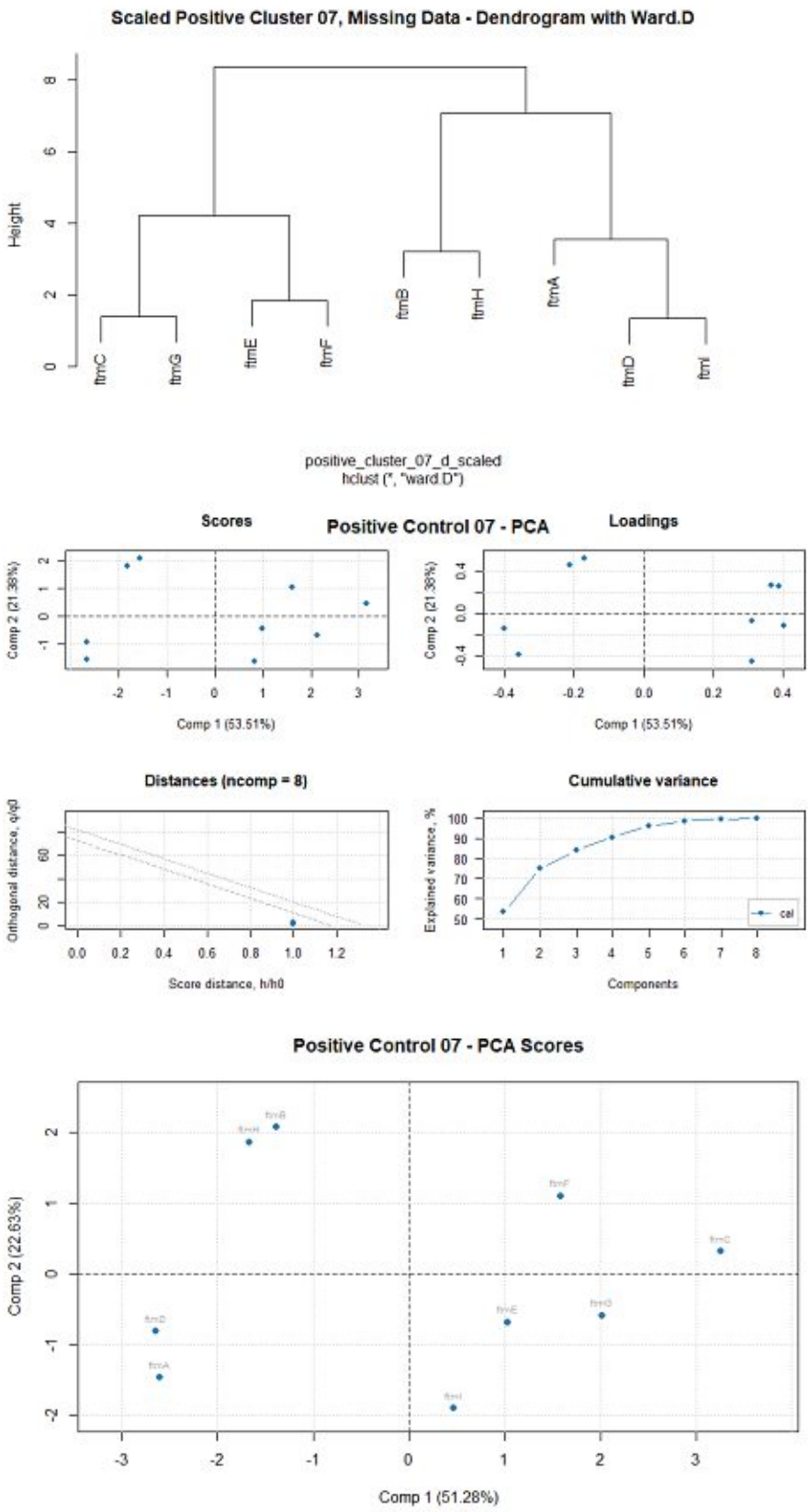


Supplement 9.39: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 07 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

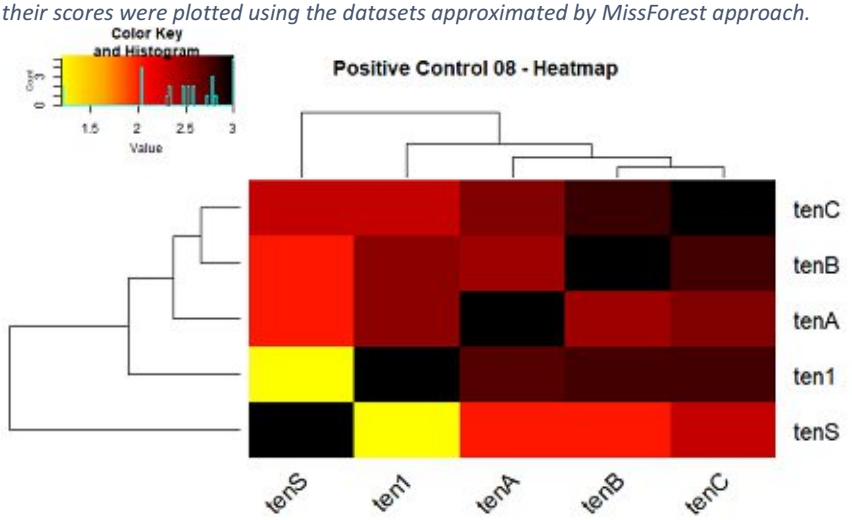


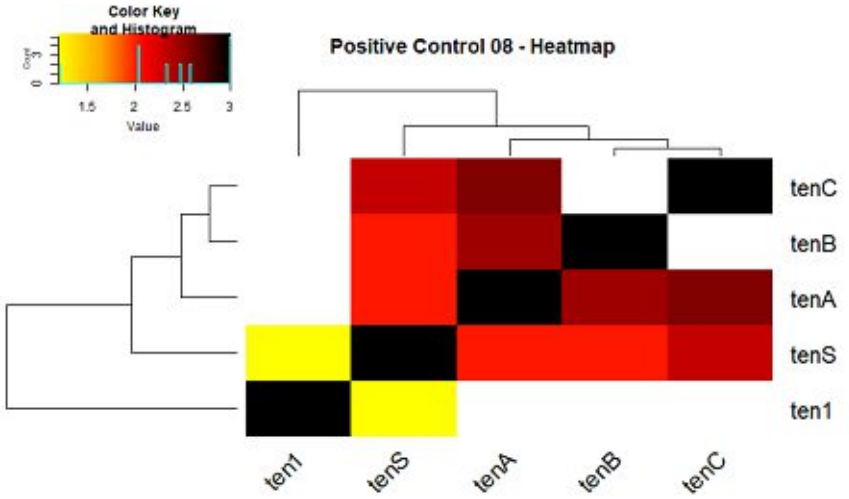


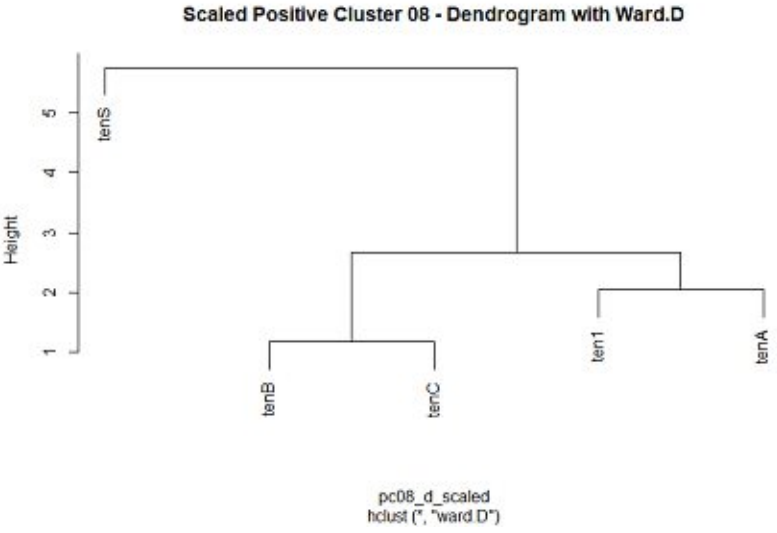


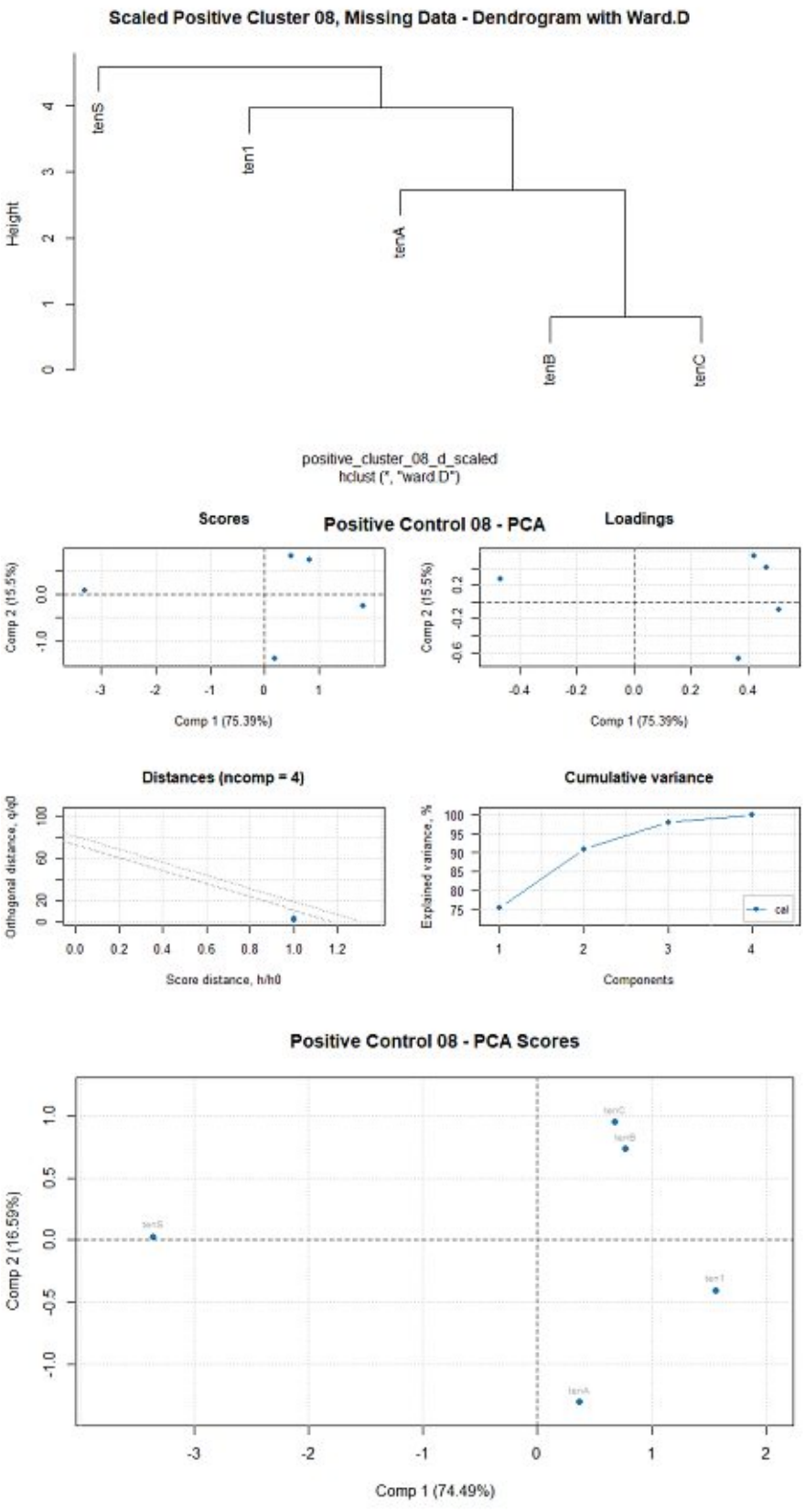


Supplement 9.40: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 08 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

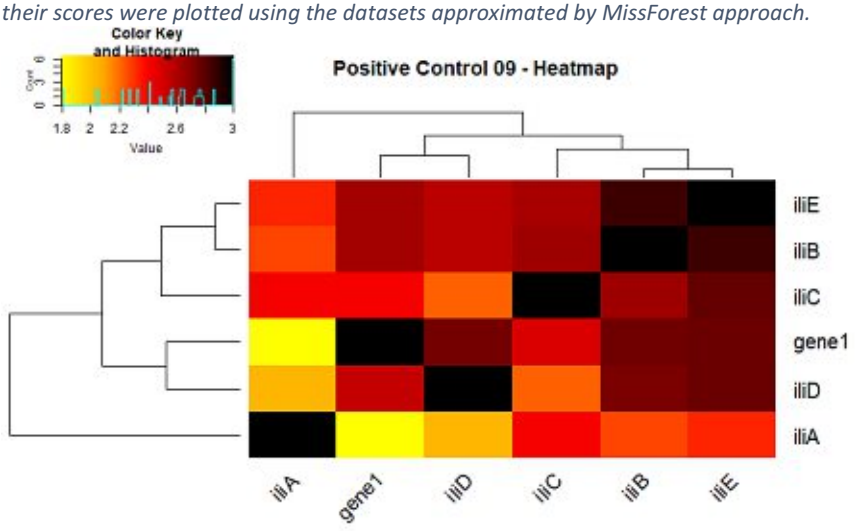


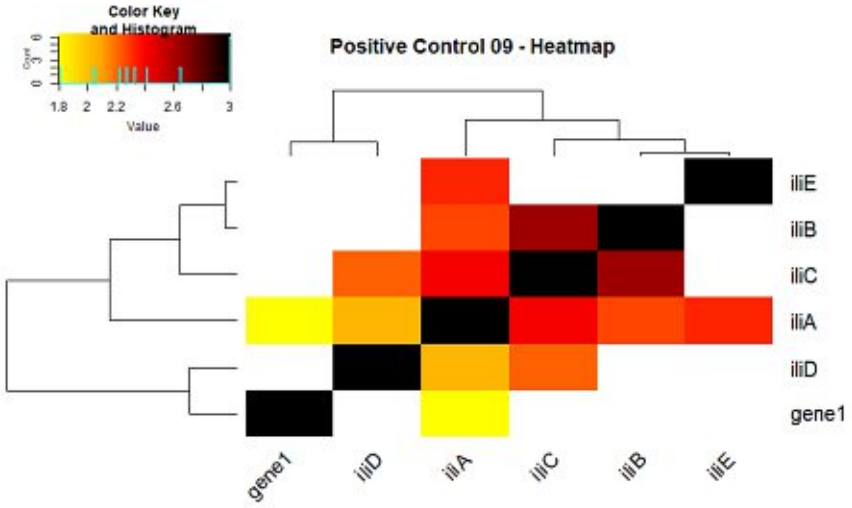


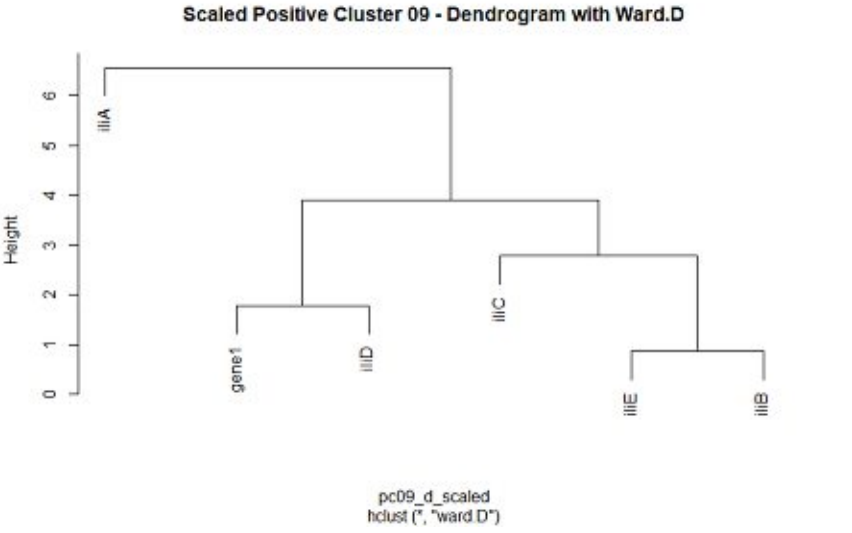


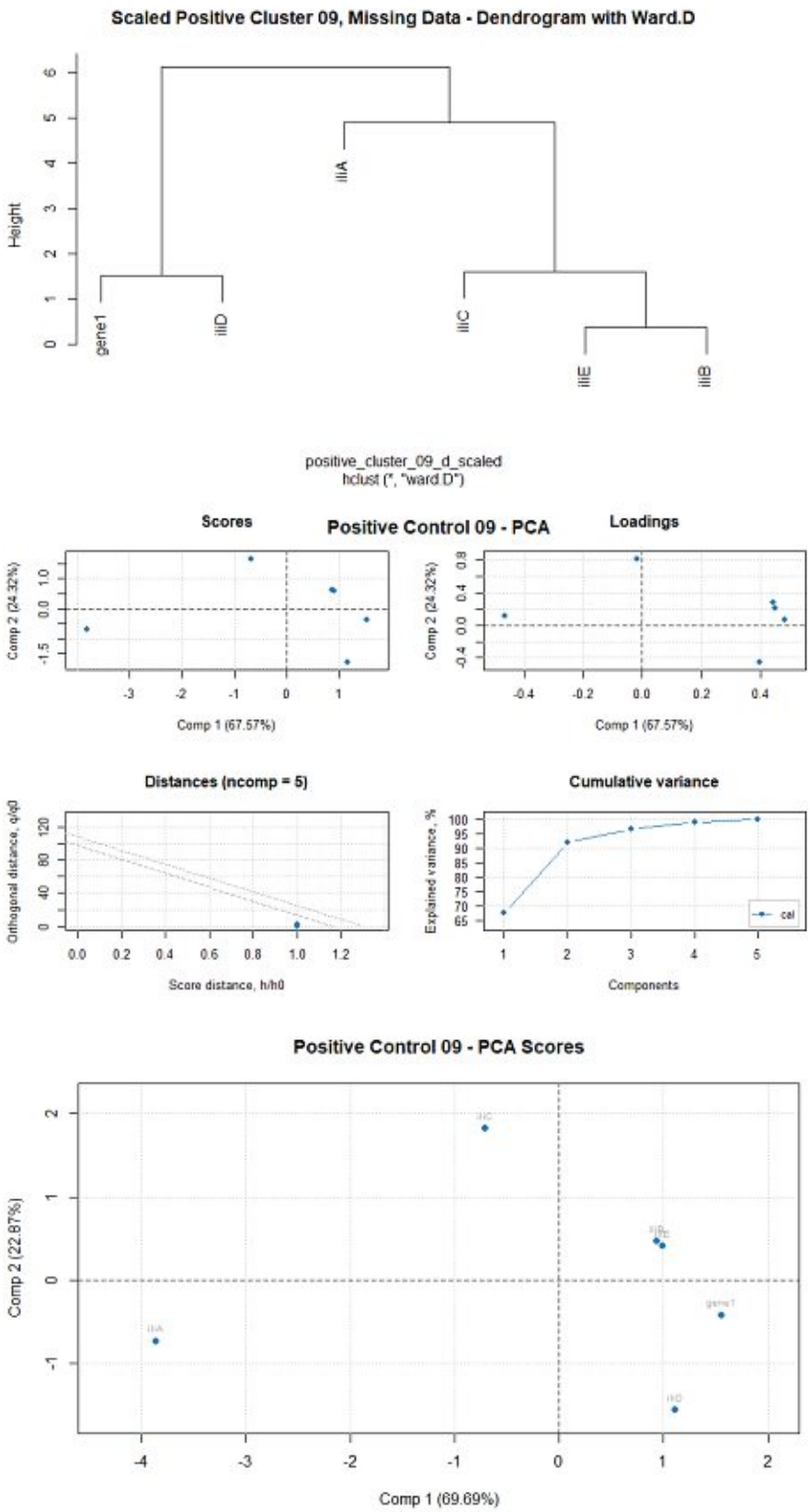


Supplement 9.41: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 09 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

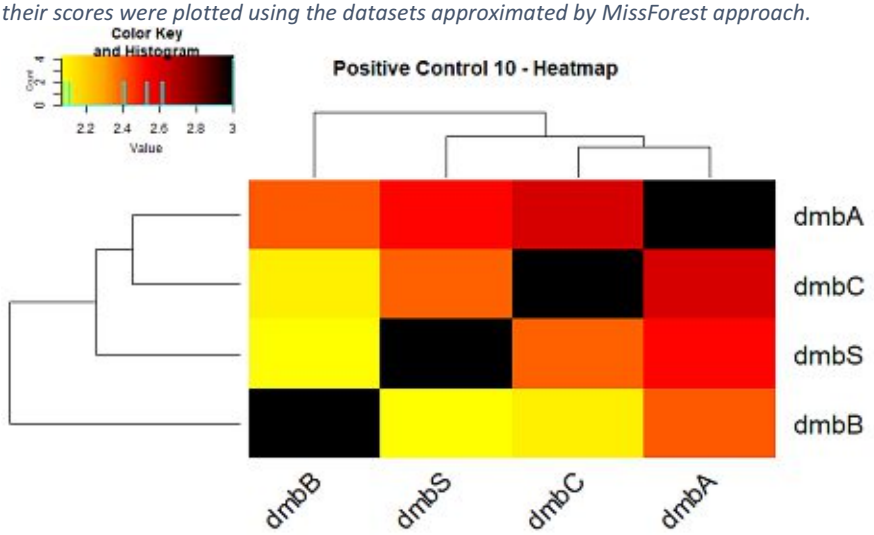


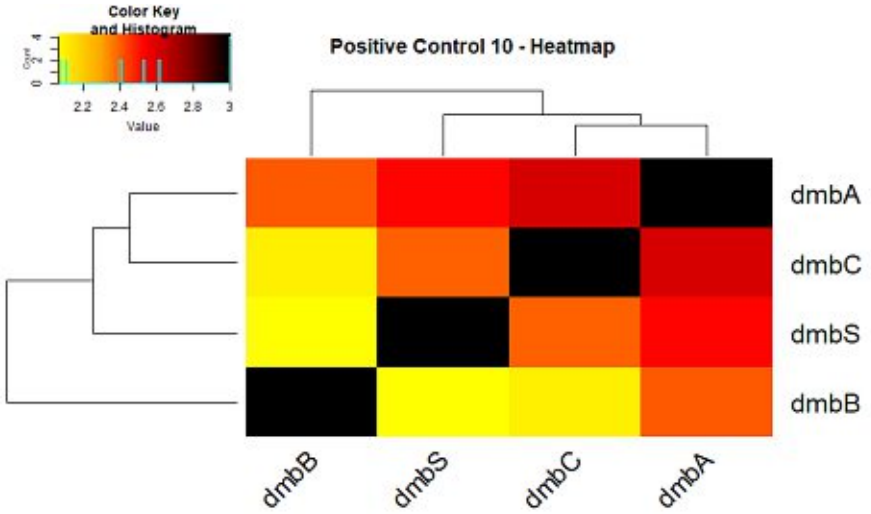


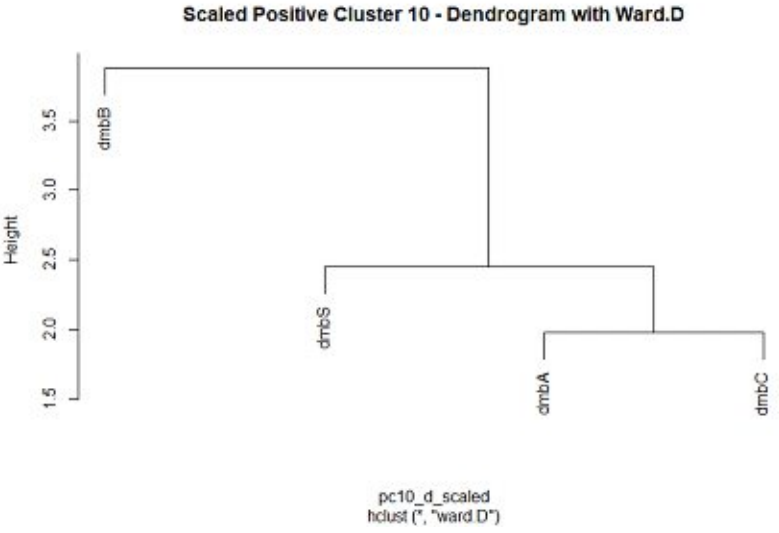


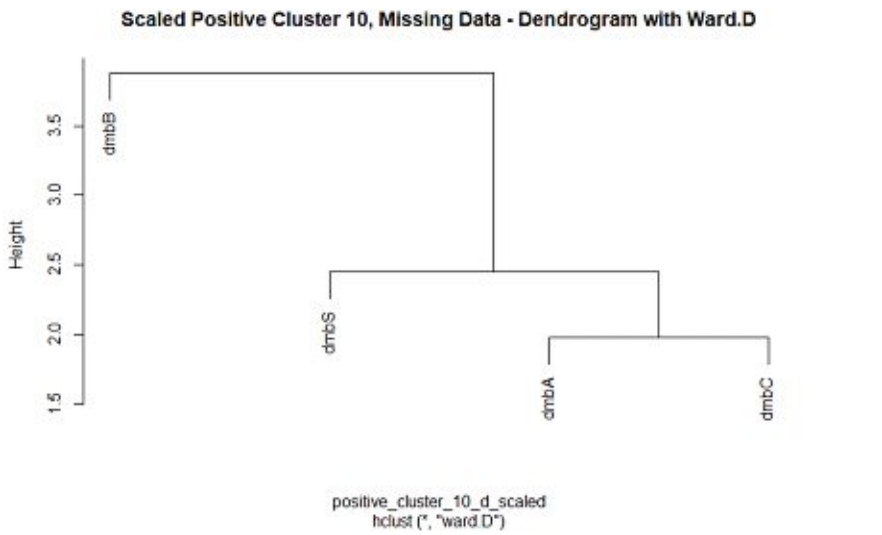


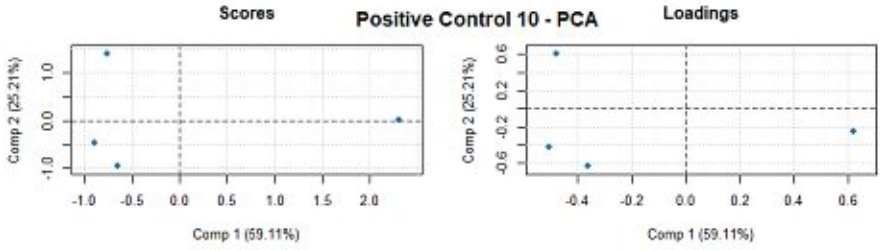
Supplement 9.42: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 10 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

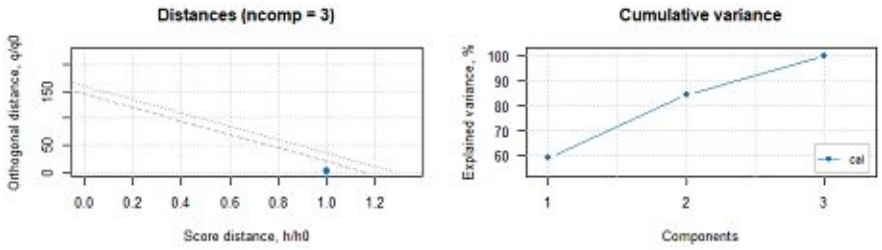


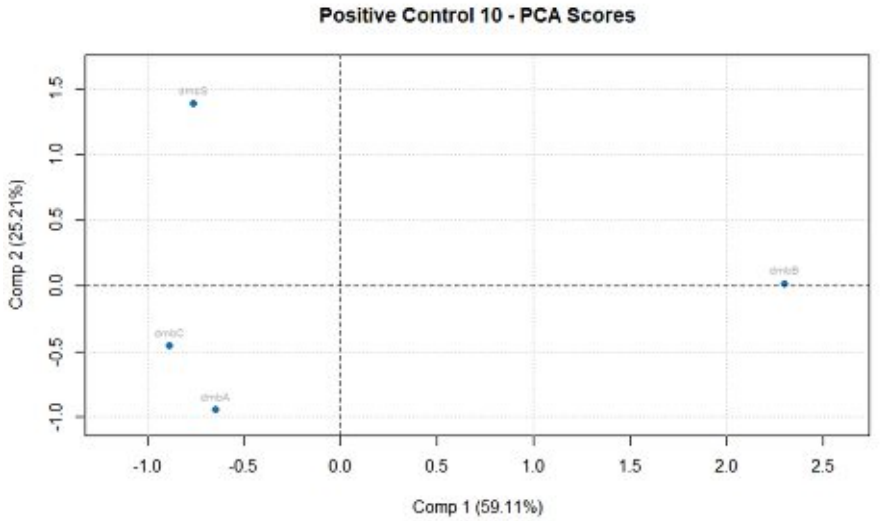




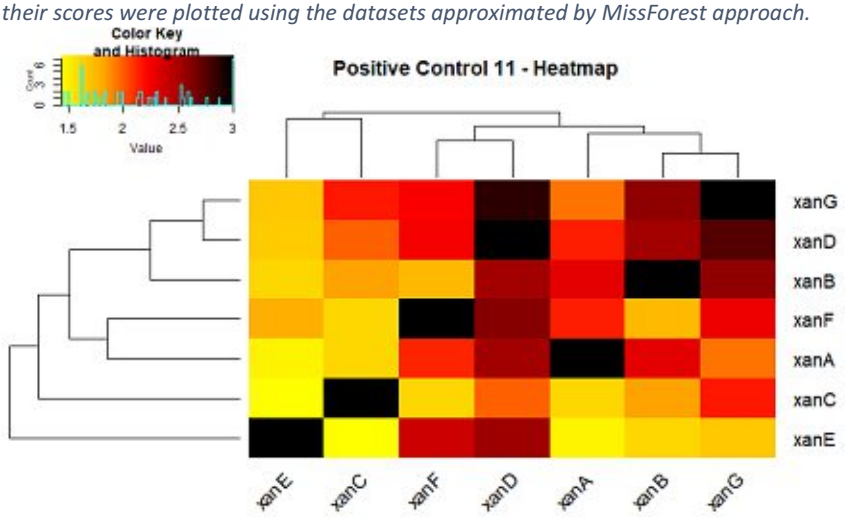


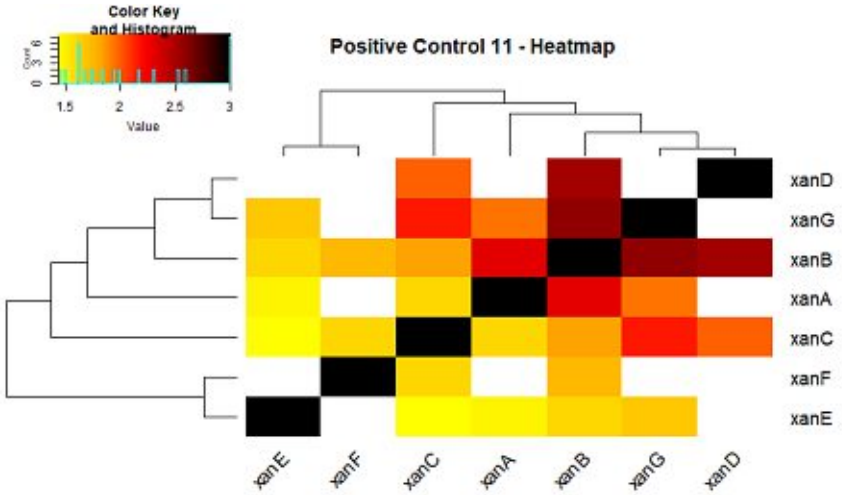


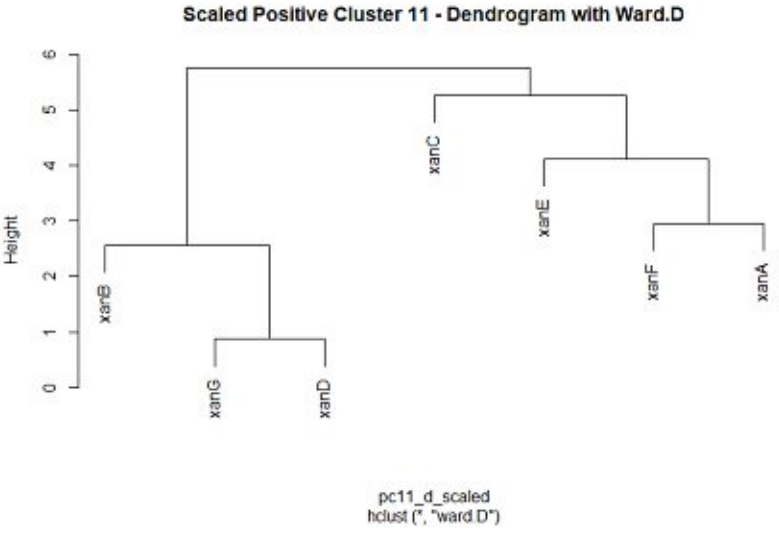


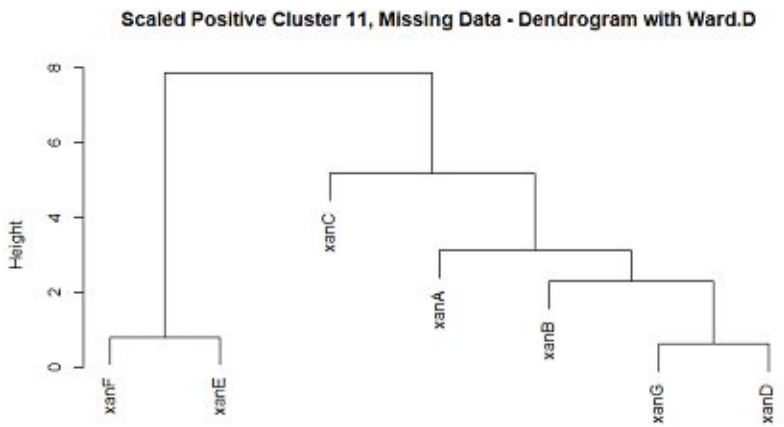


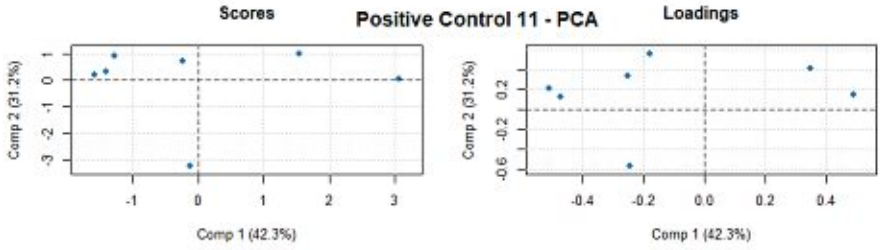
Supplement 9.43 Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 11 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

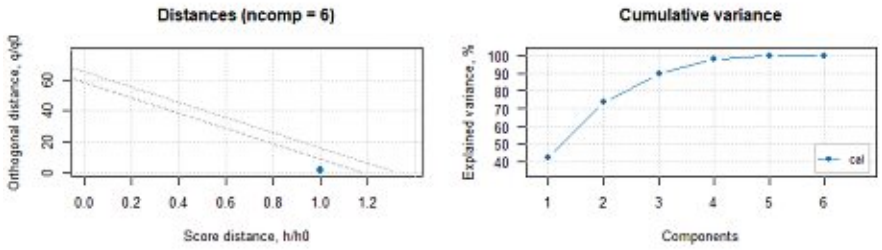


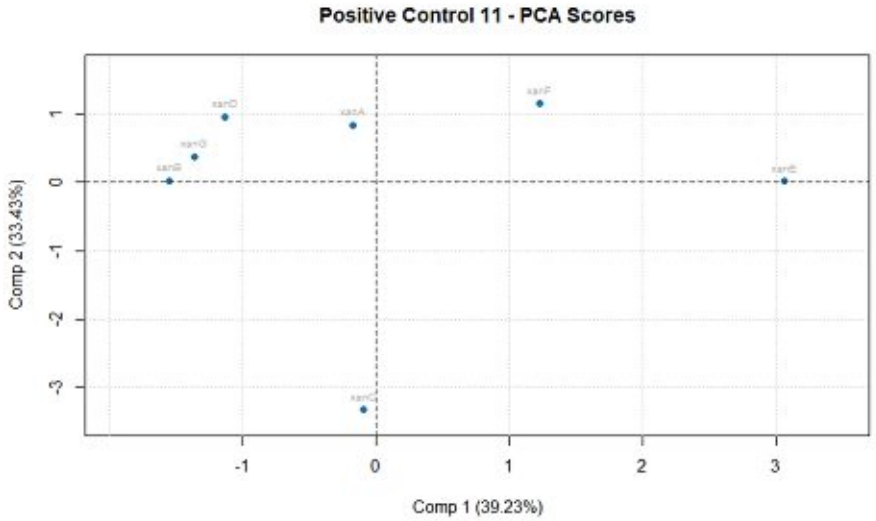




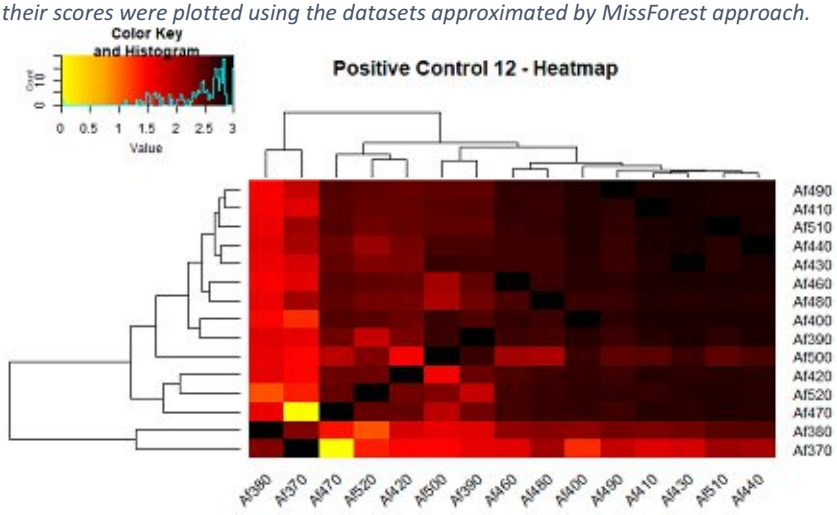


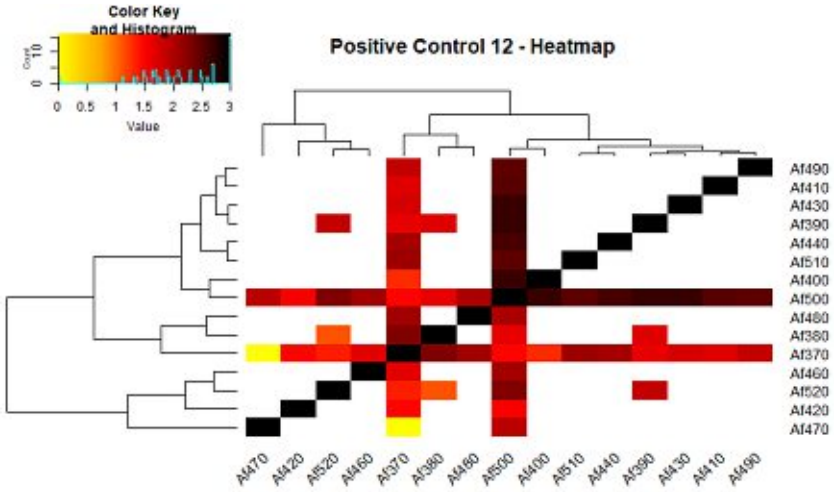


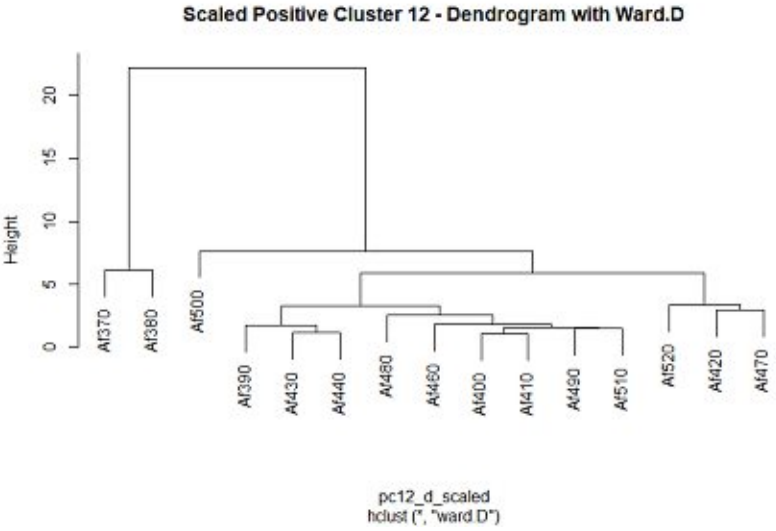


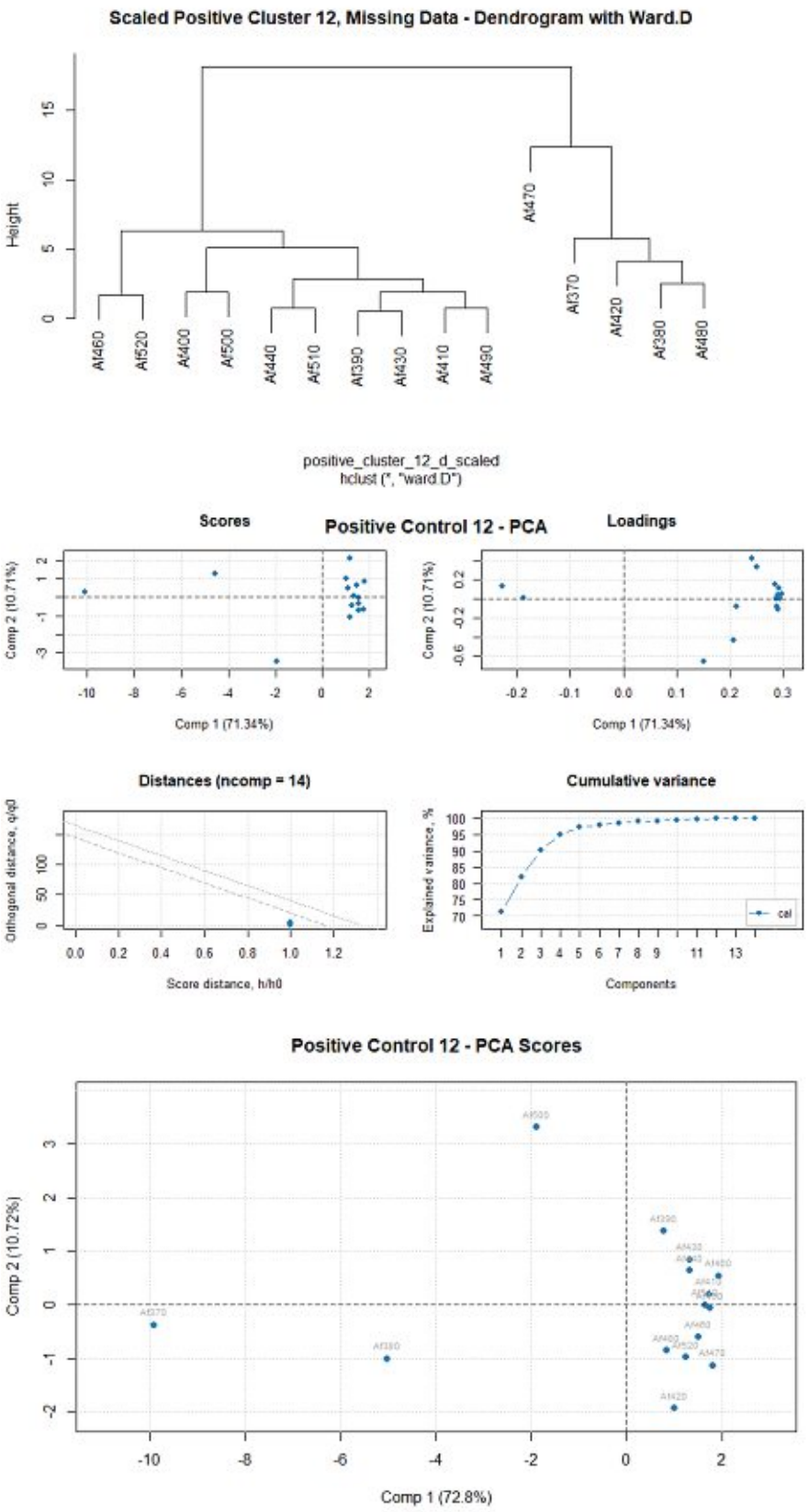


Supplement 9.44: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 12 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

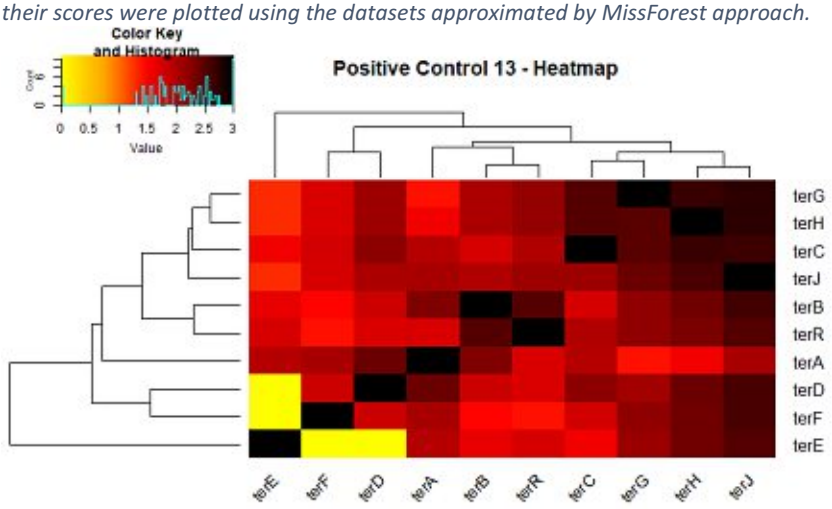


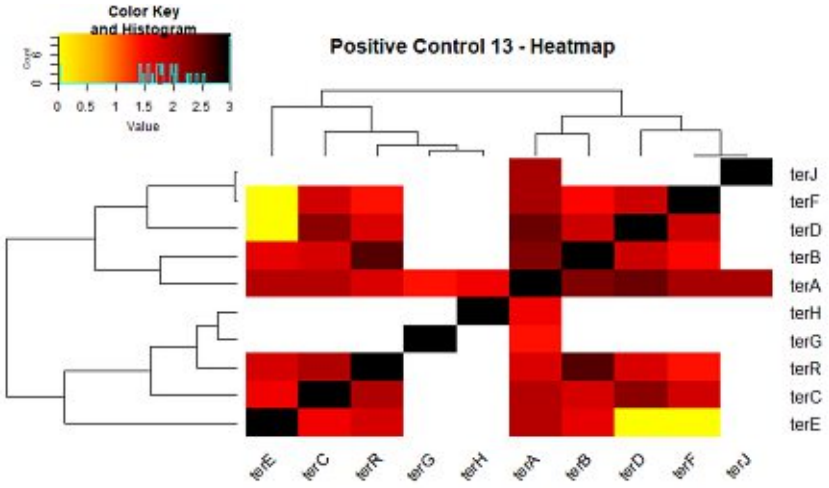


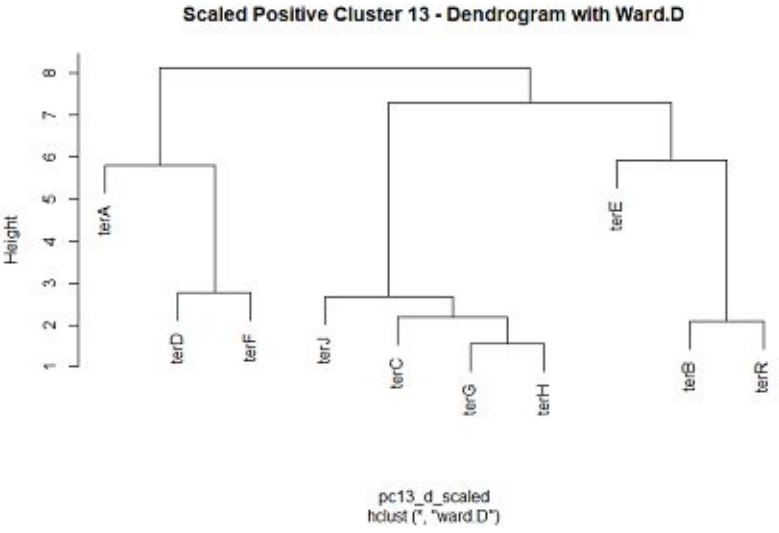


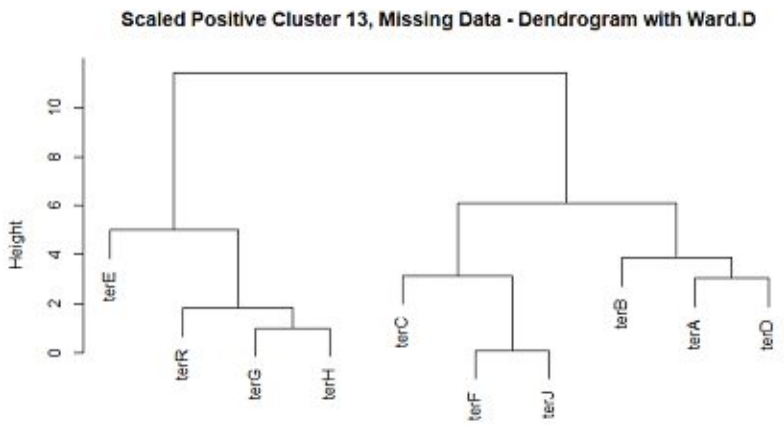


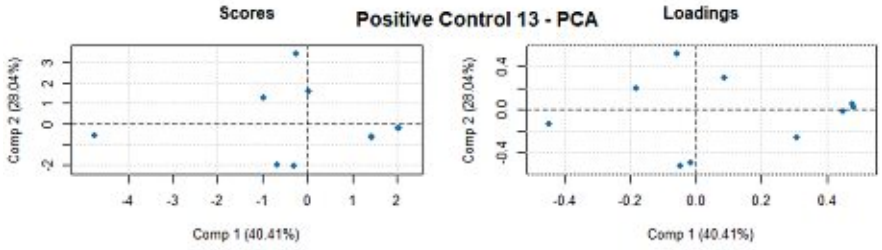
Supplement 9.45: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 13 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

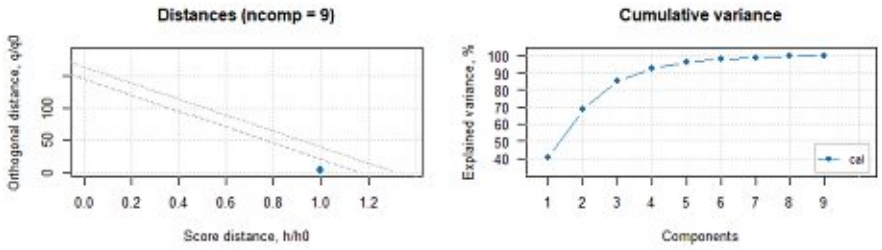


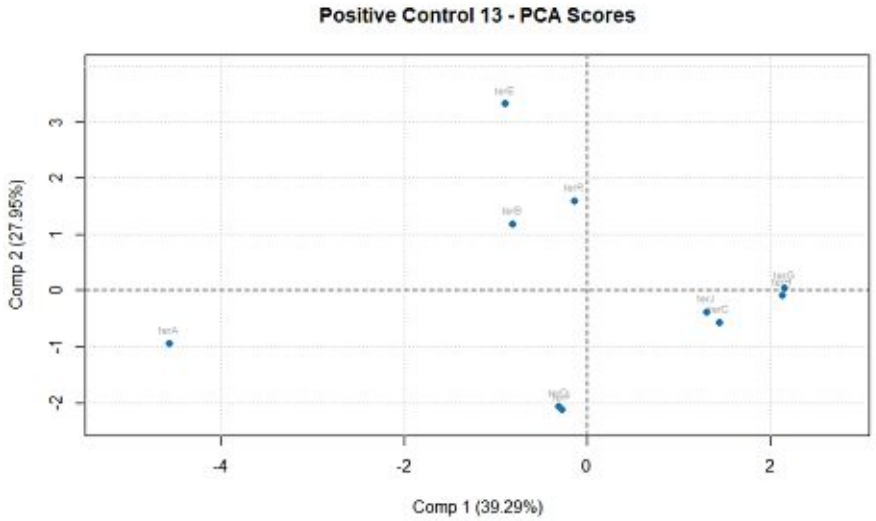




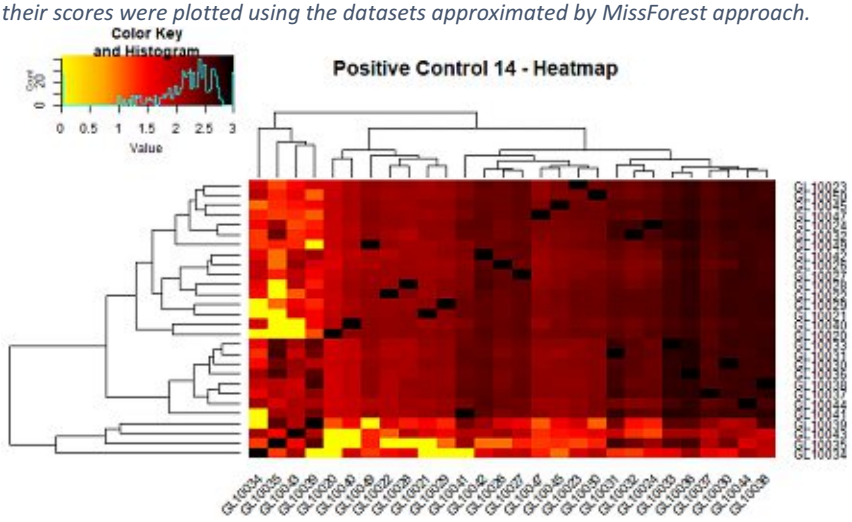


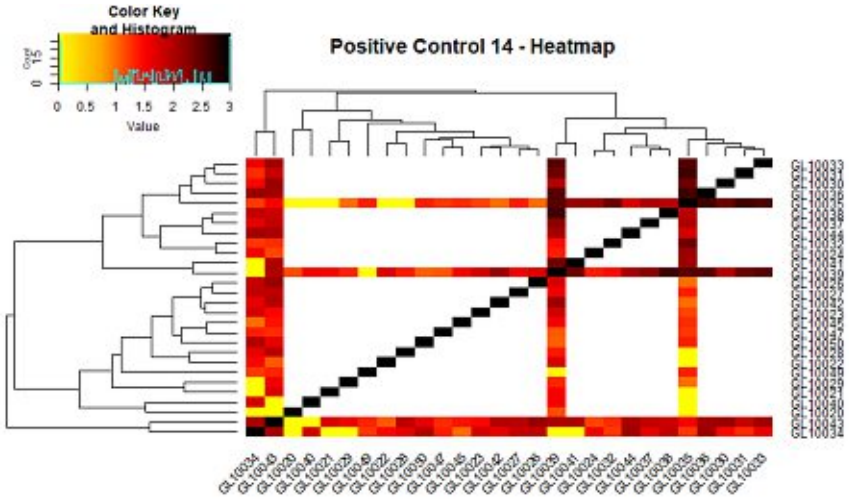


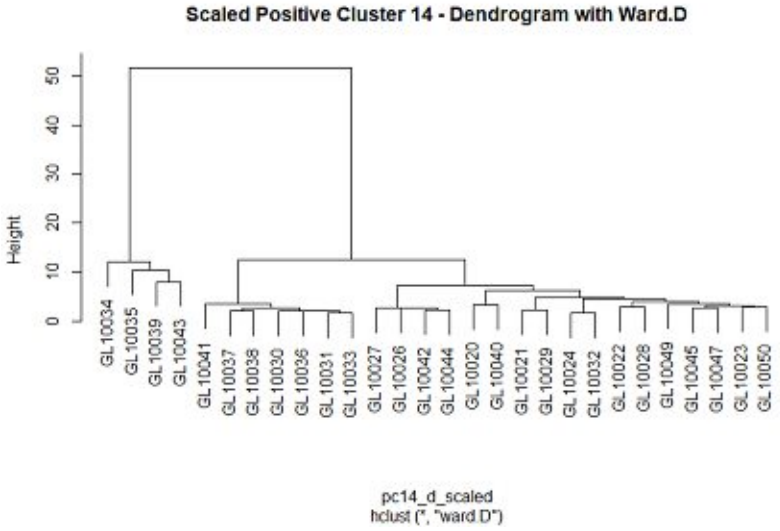


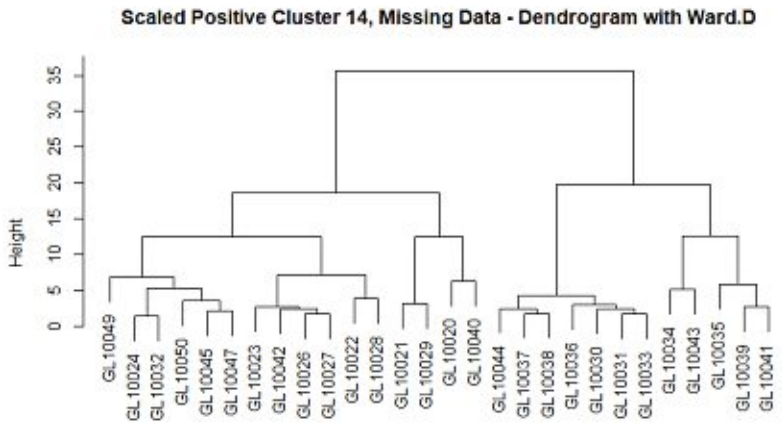


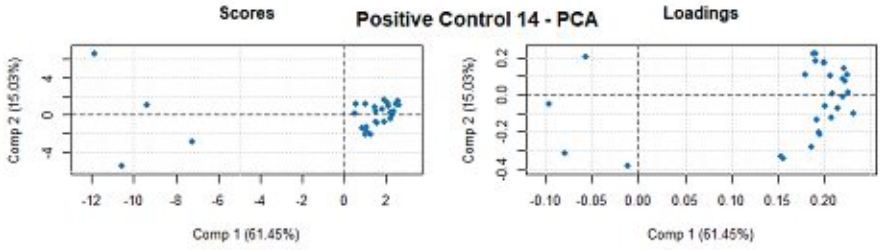
Supplement 9.46: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 14 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

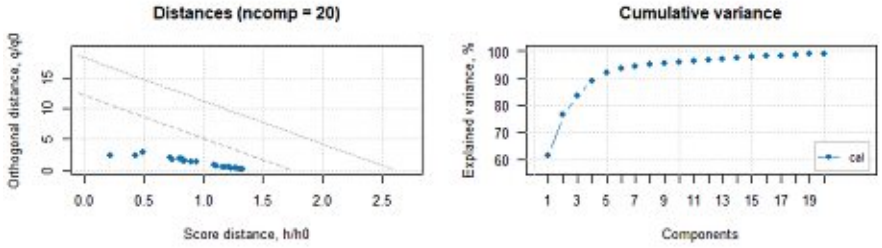


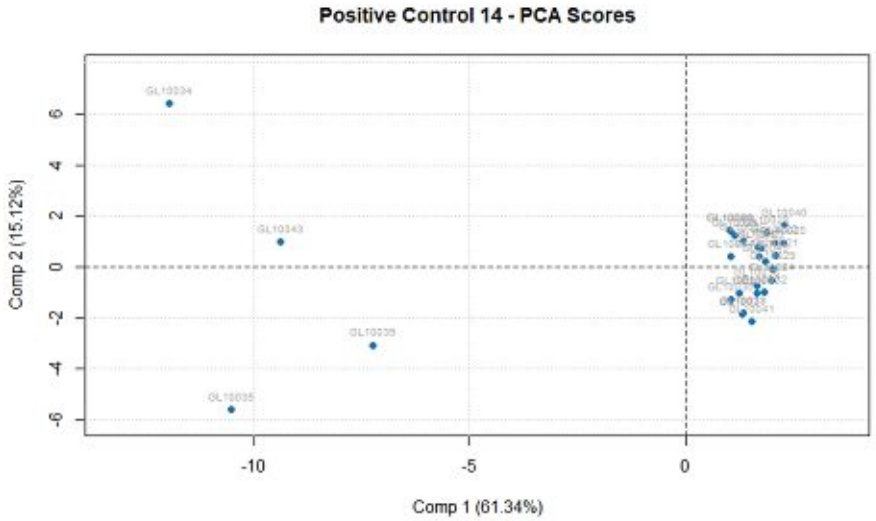




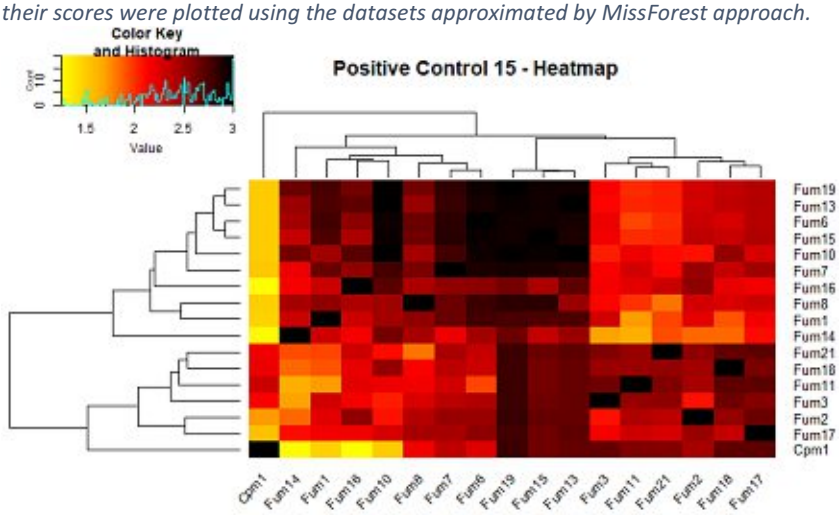


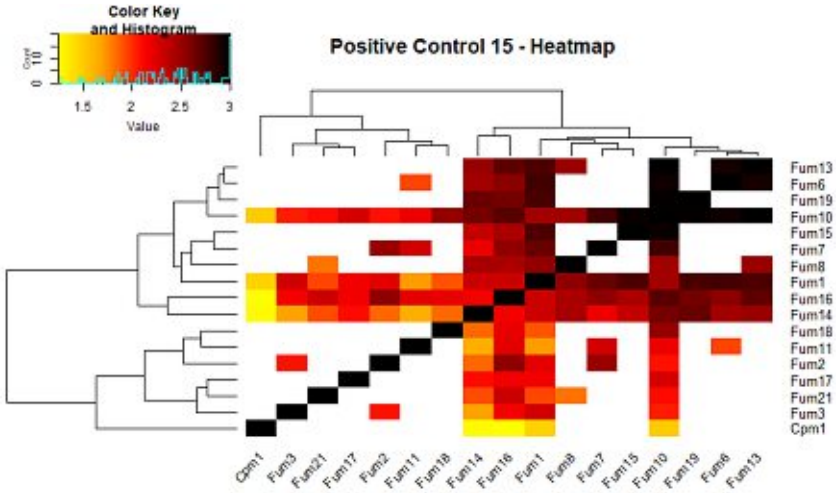


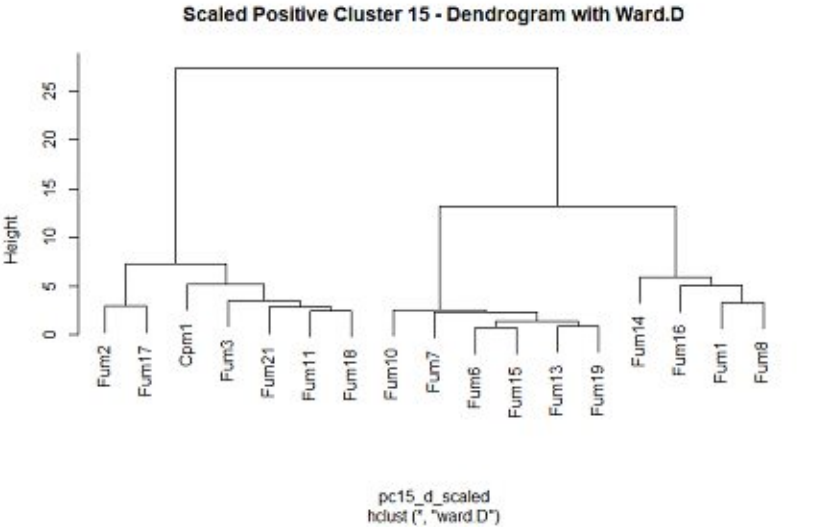


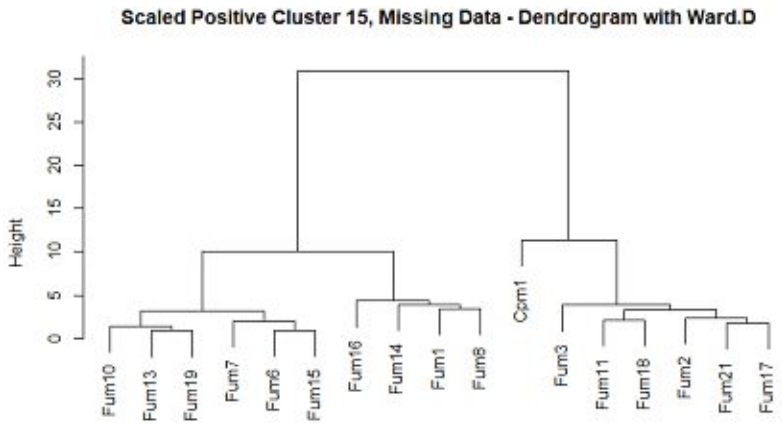


Supplement 9.47: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 15 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

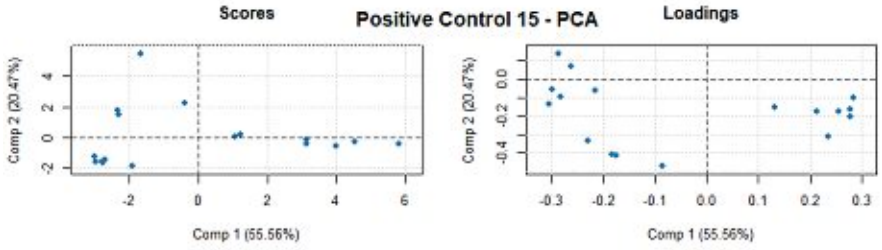


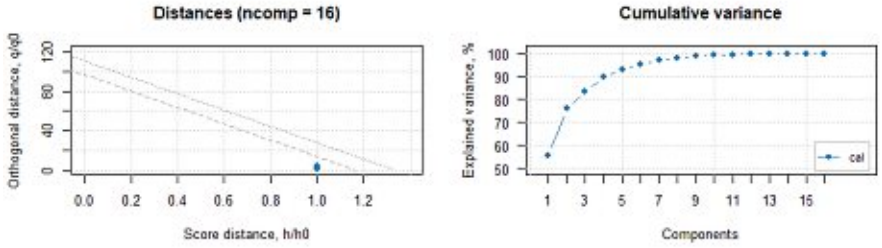


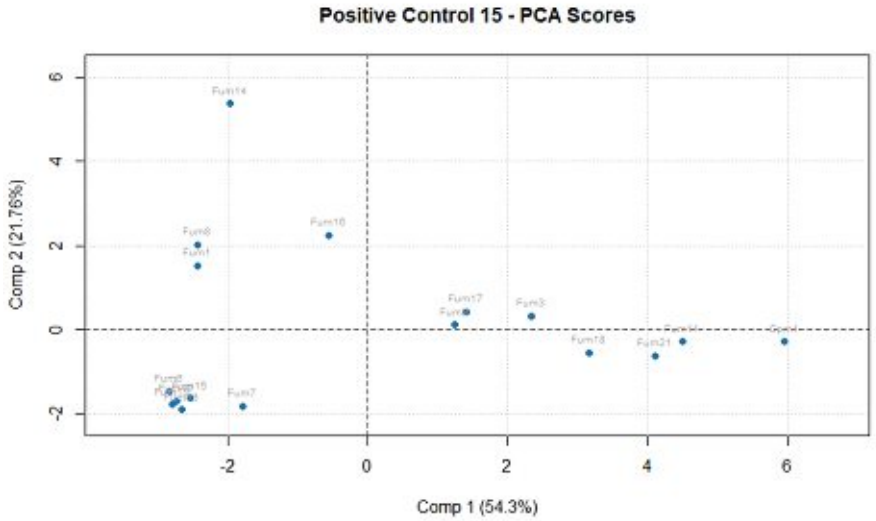




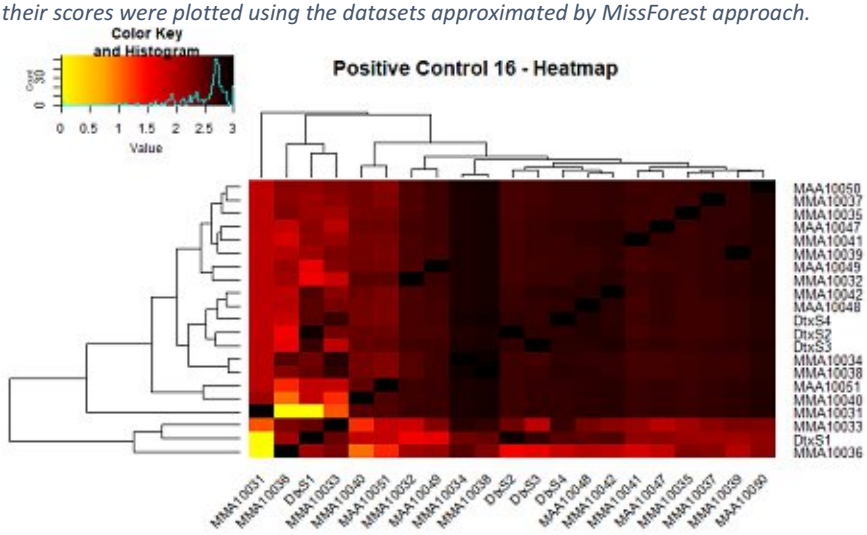
positive_cluster_15_d_scaled
hclust (*, "ward.D")

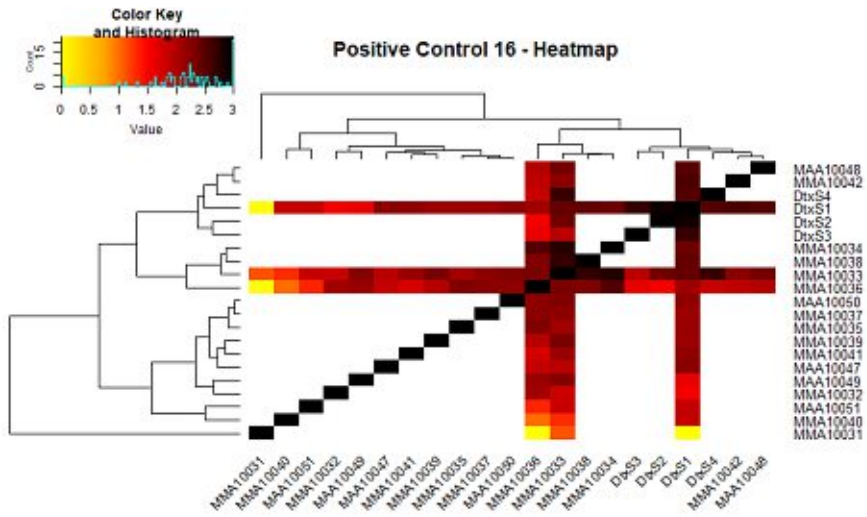


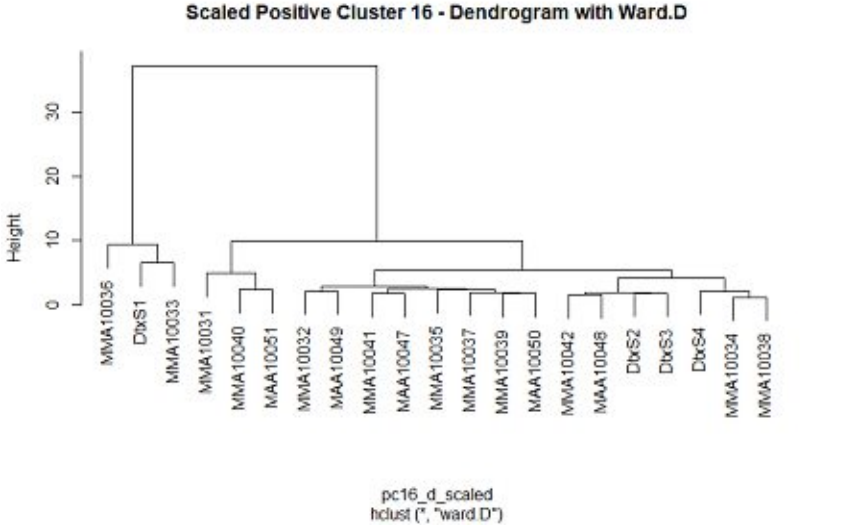


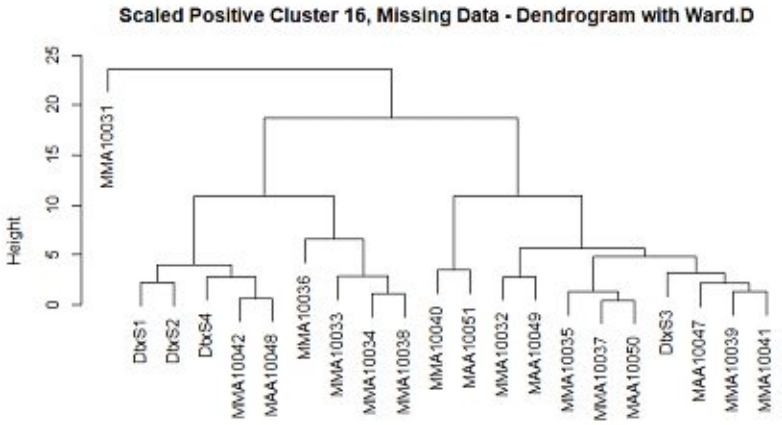


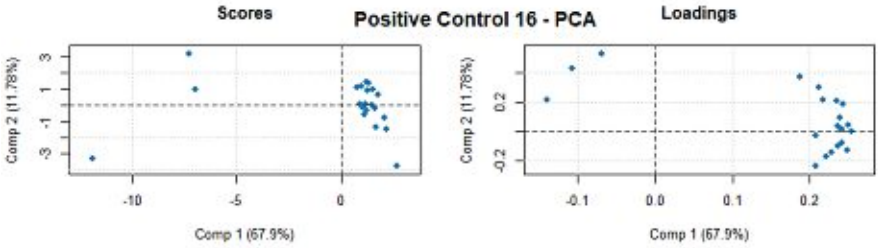
Supplement 9.48: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 16 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

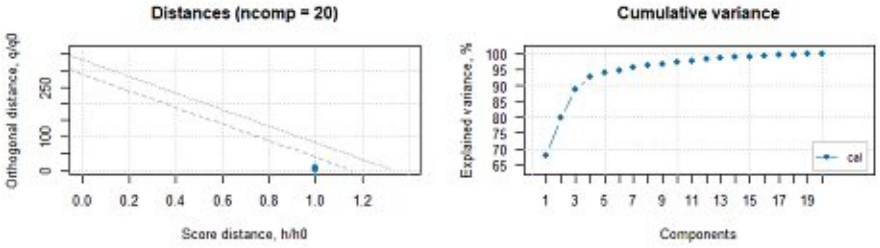


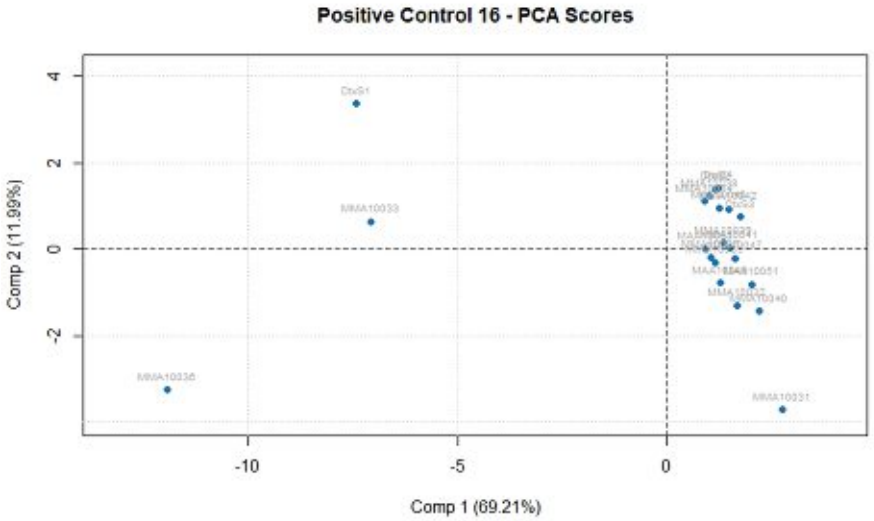




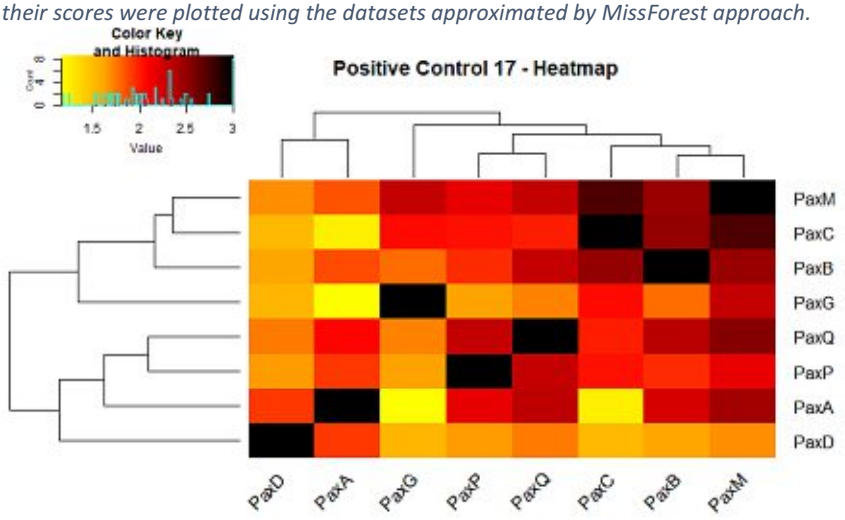


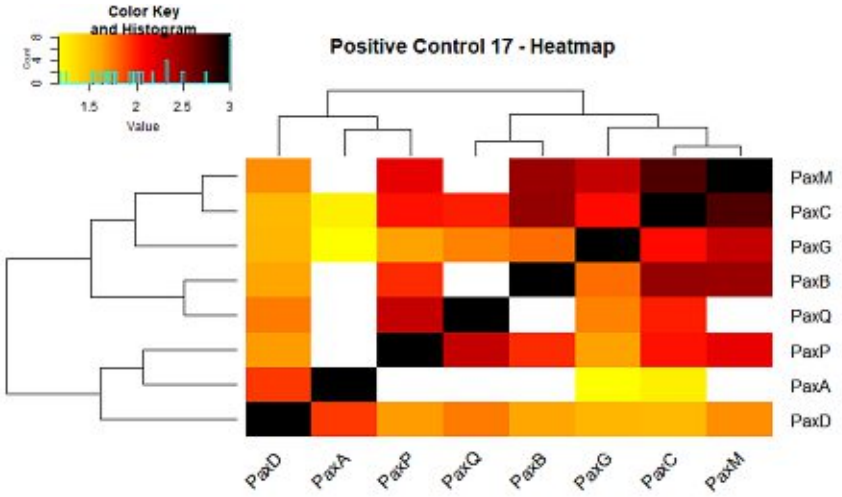


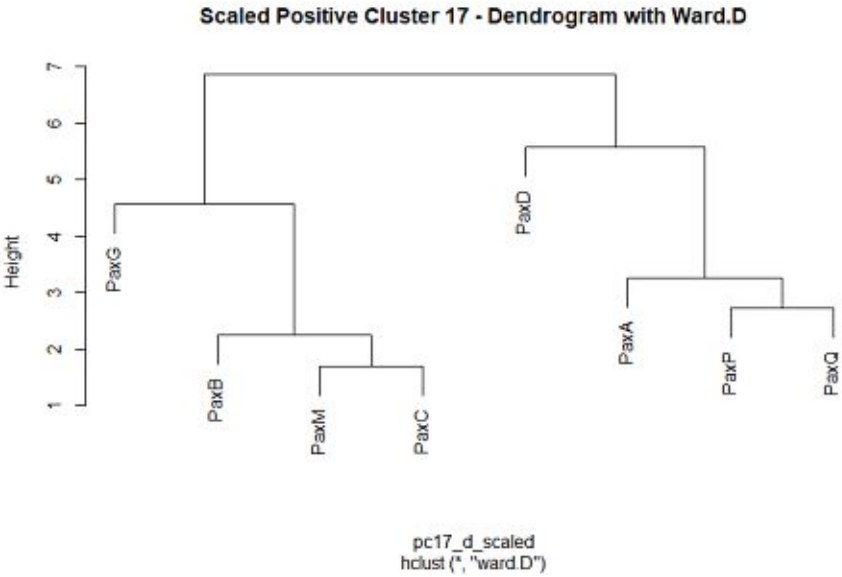


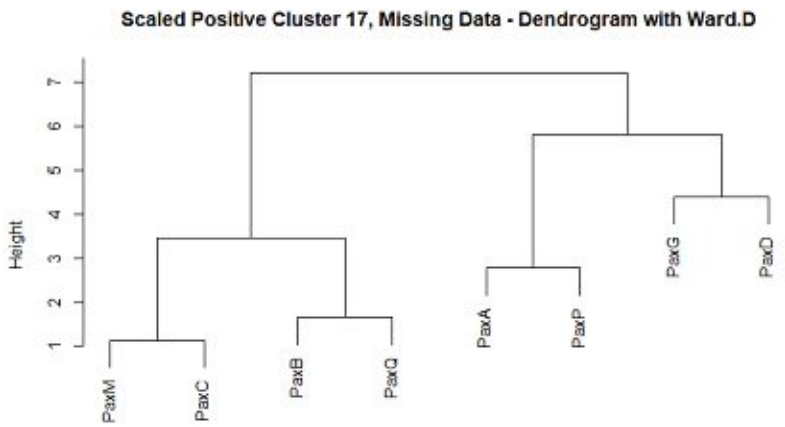


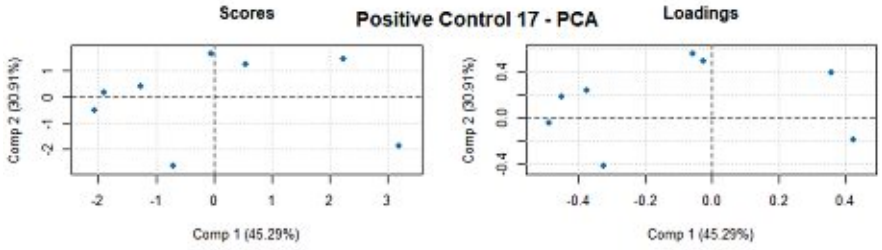
Supplement 9.49: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 17 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

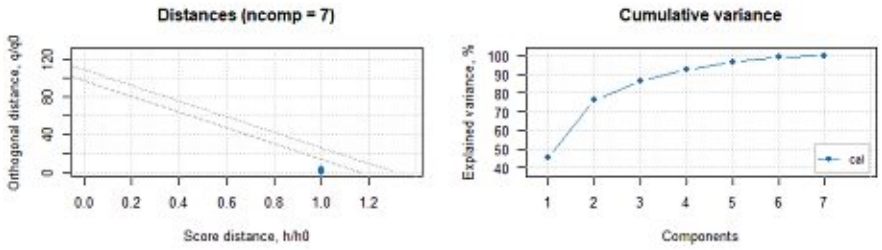


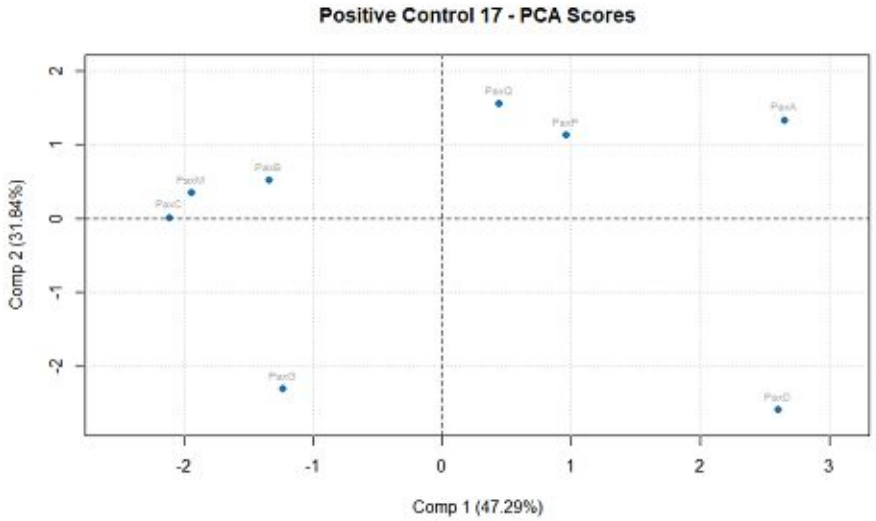




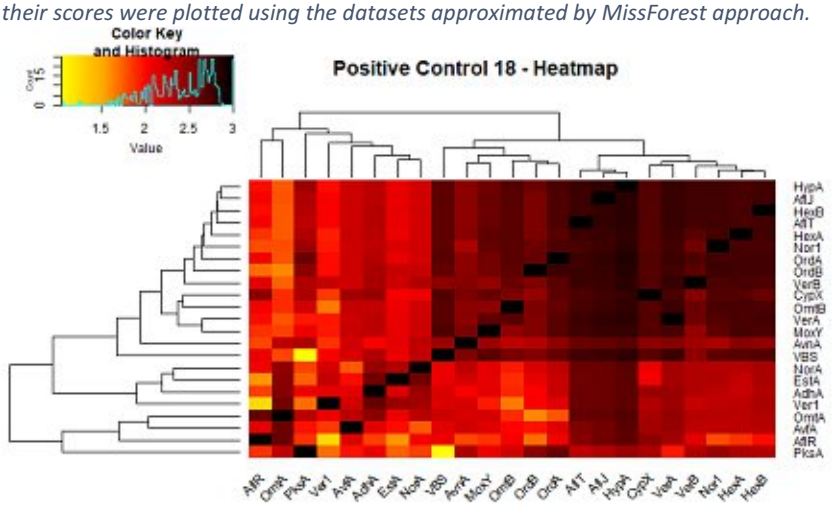


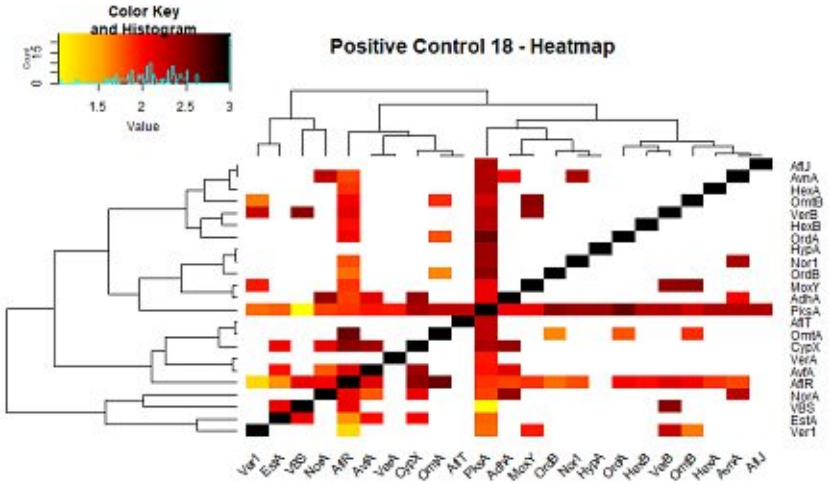


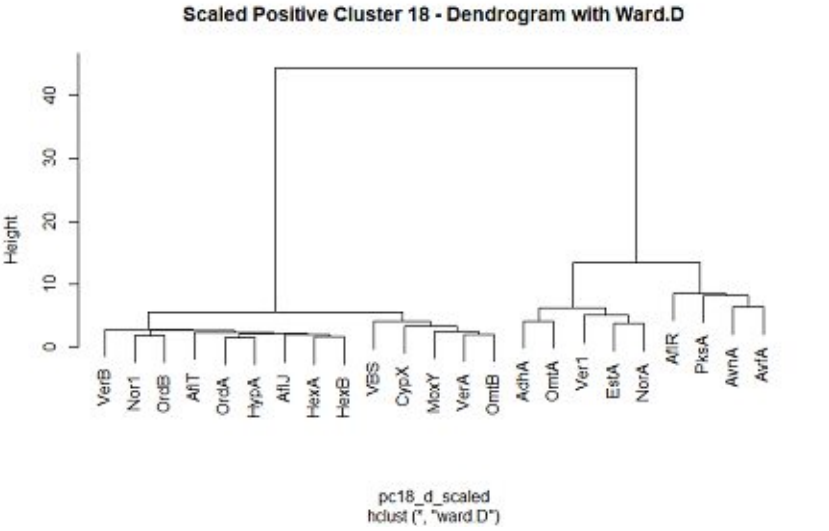


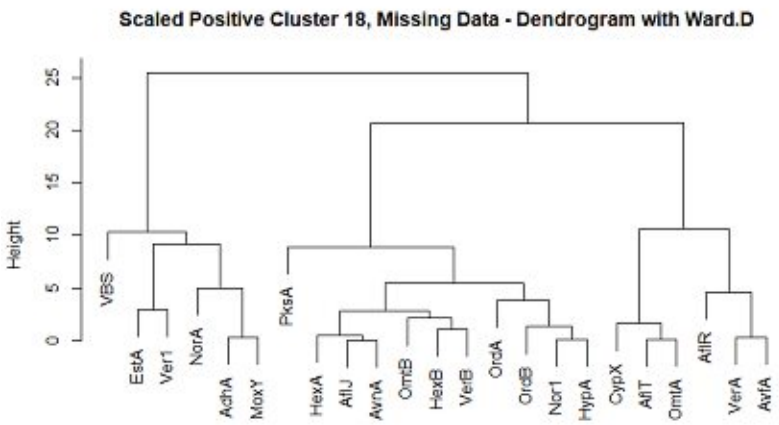


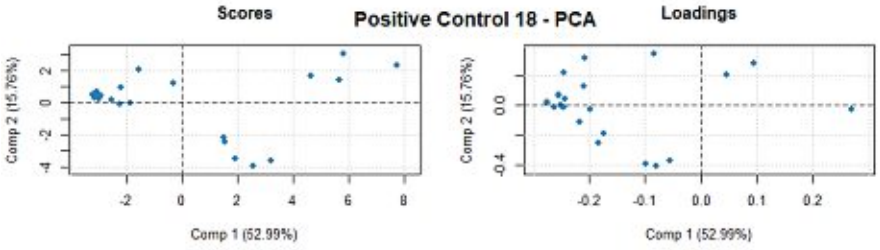
Supplement 9.50: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 18 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

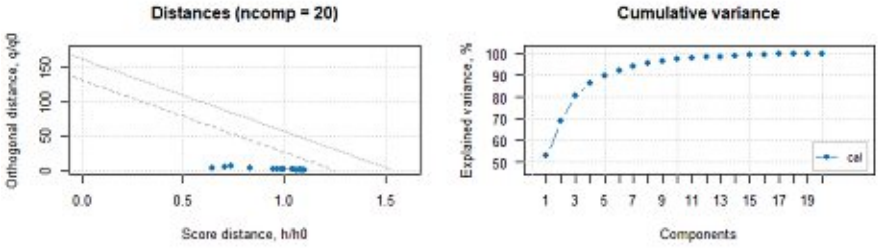


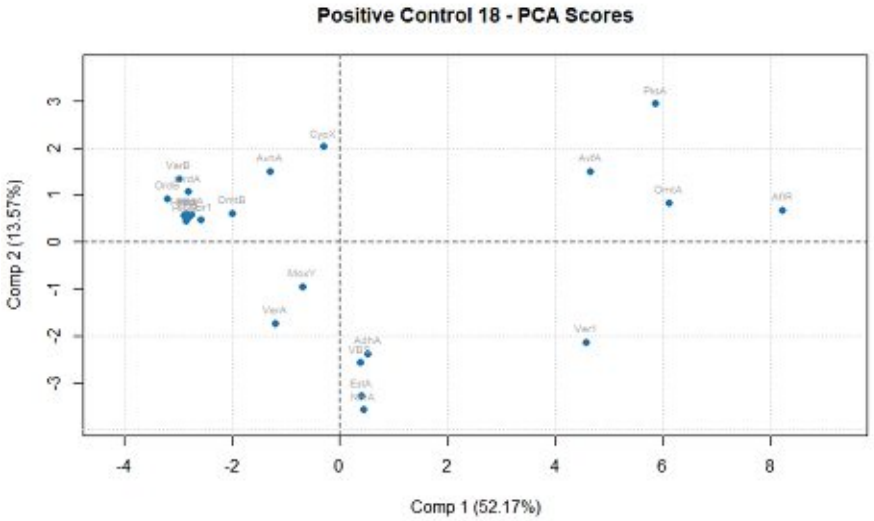




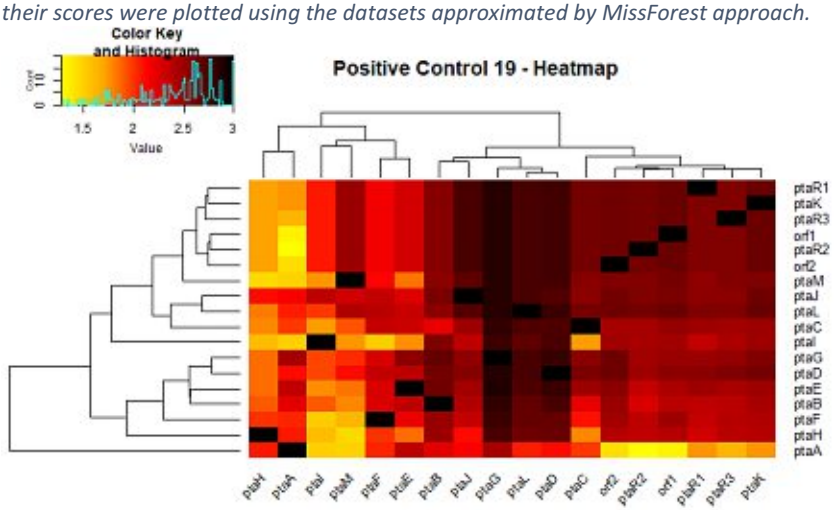


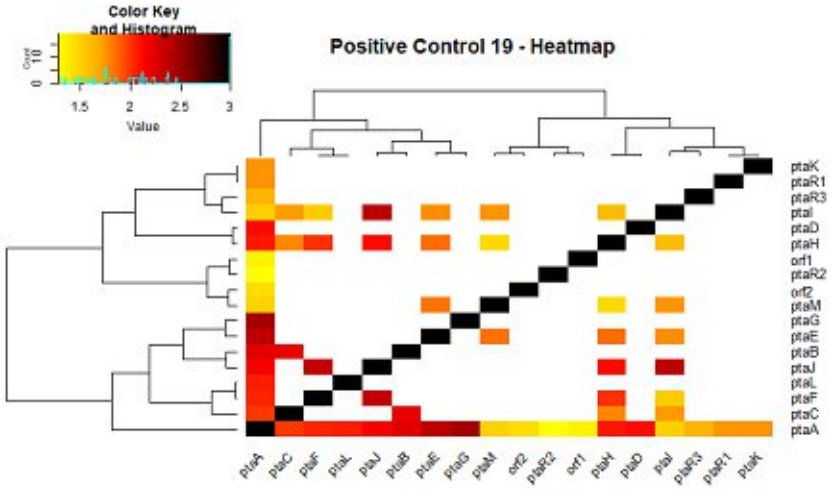


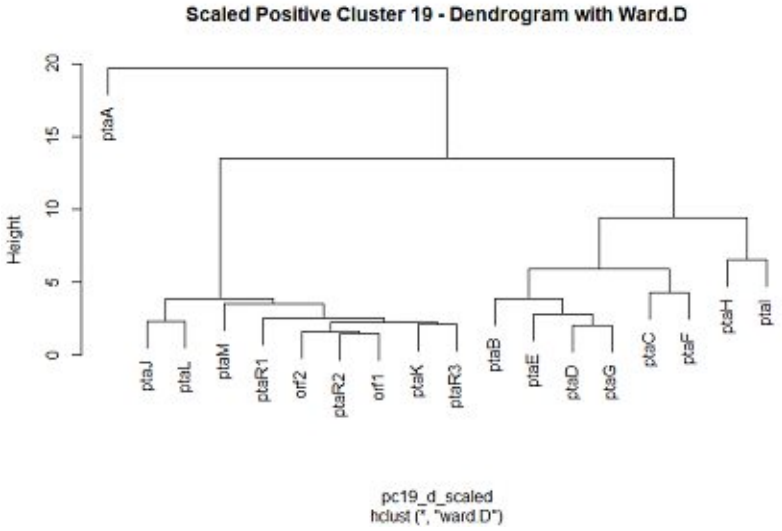


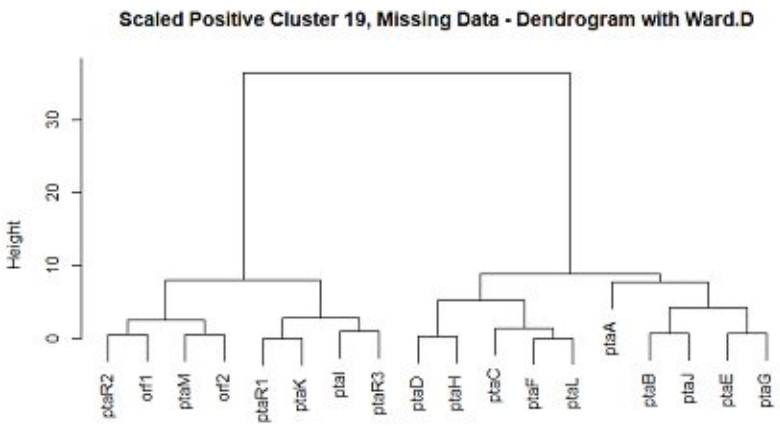


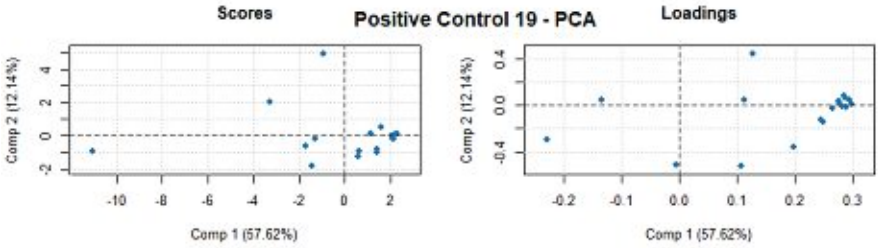
Supplement 9 51: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 19 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

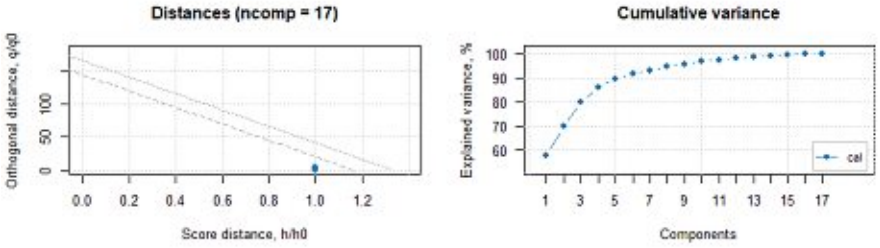


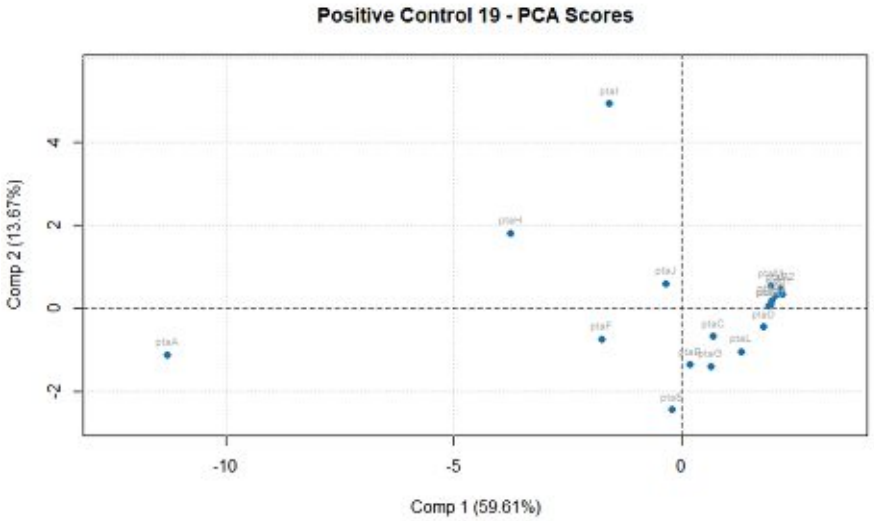




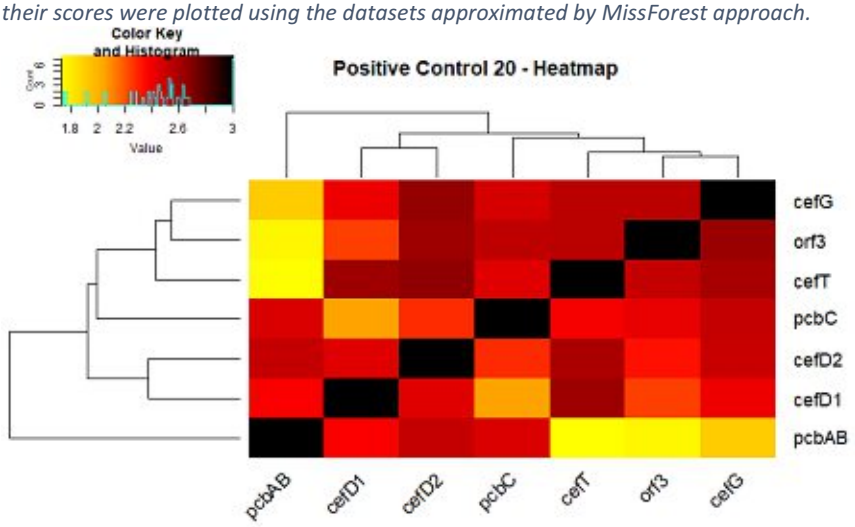


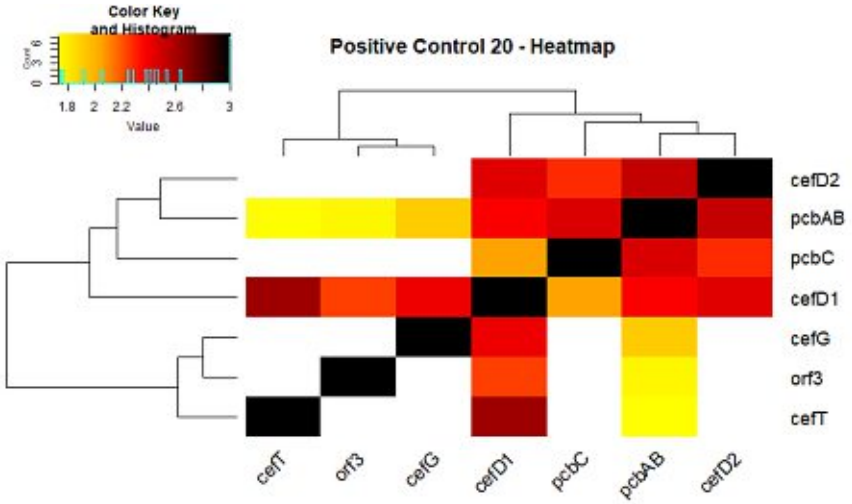


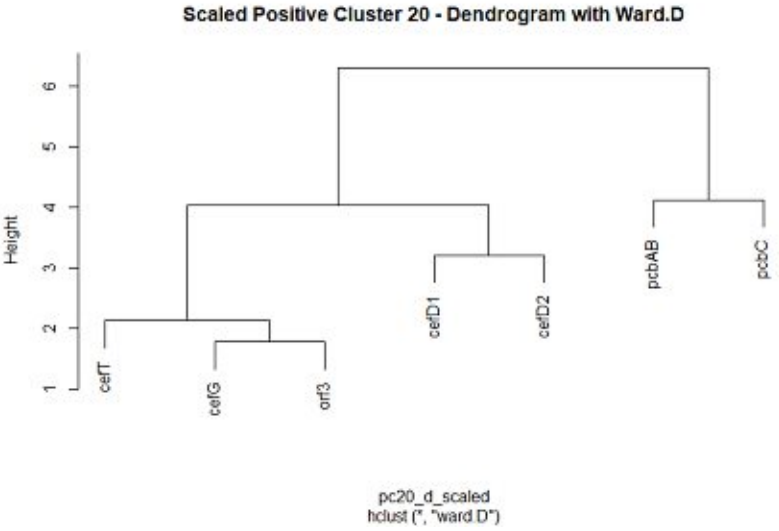


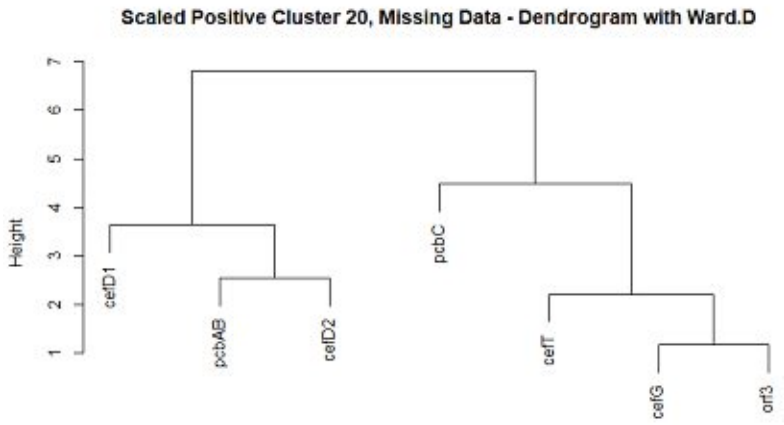


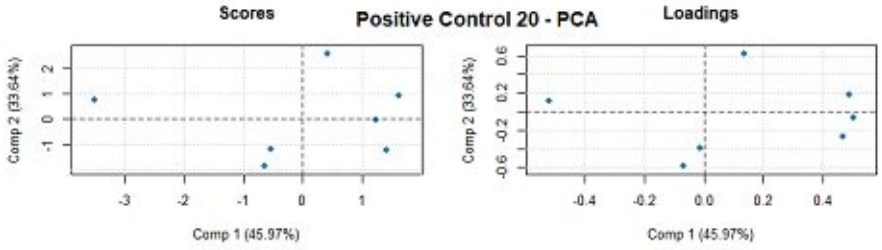
Supplement 9.52: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 20 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

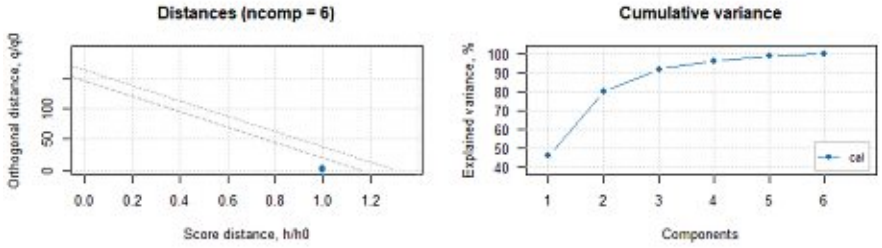


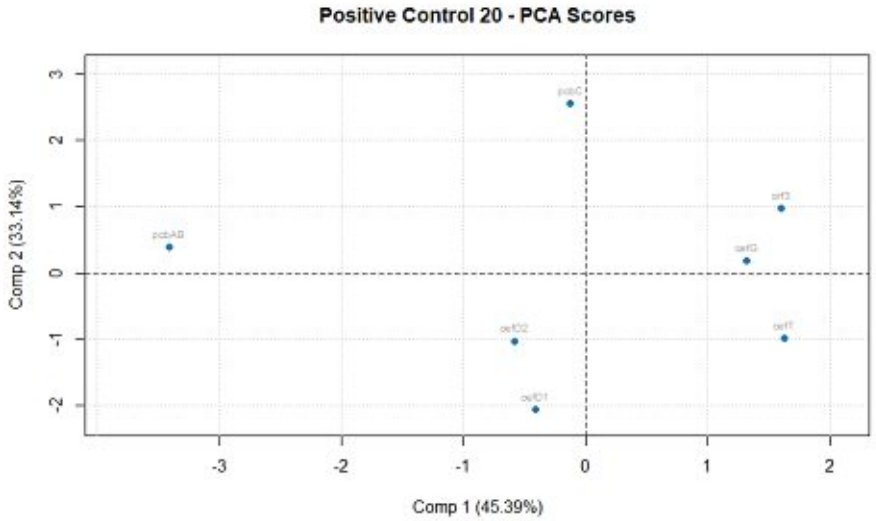




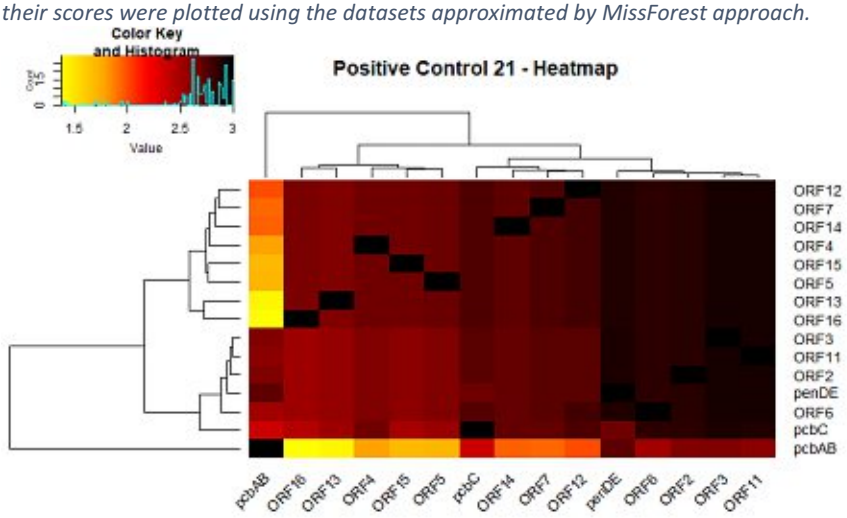


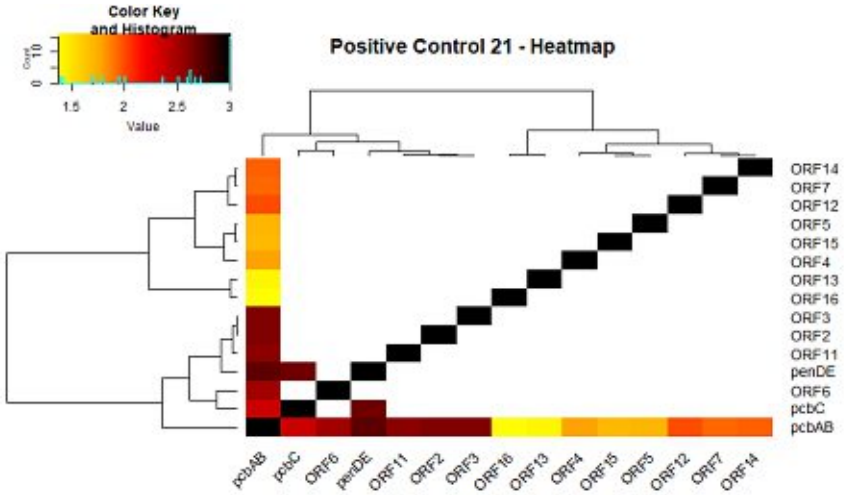


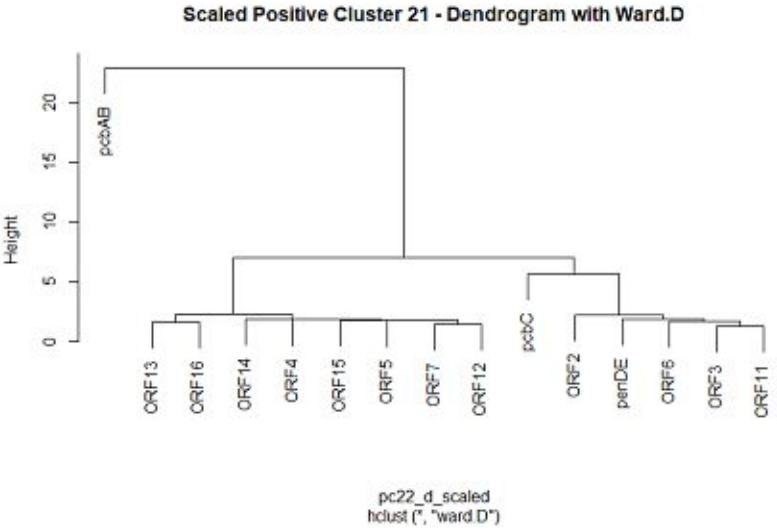


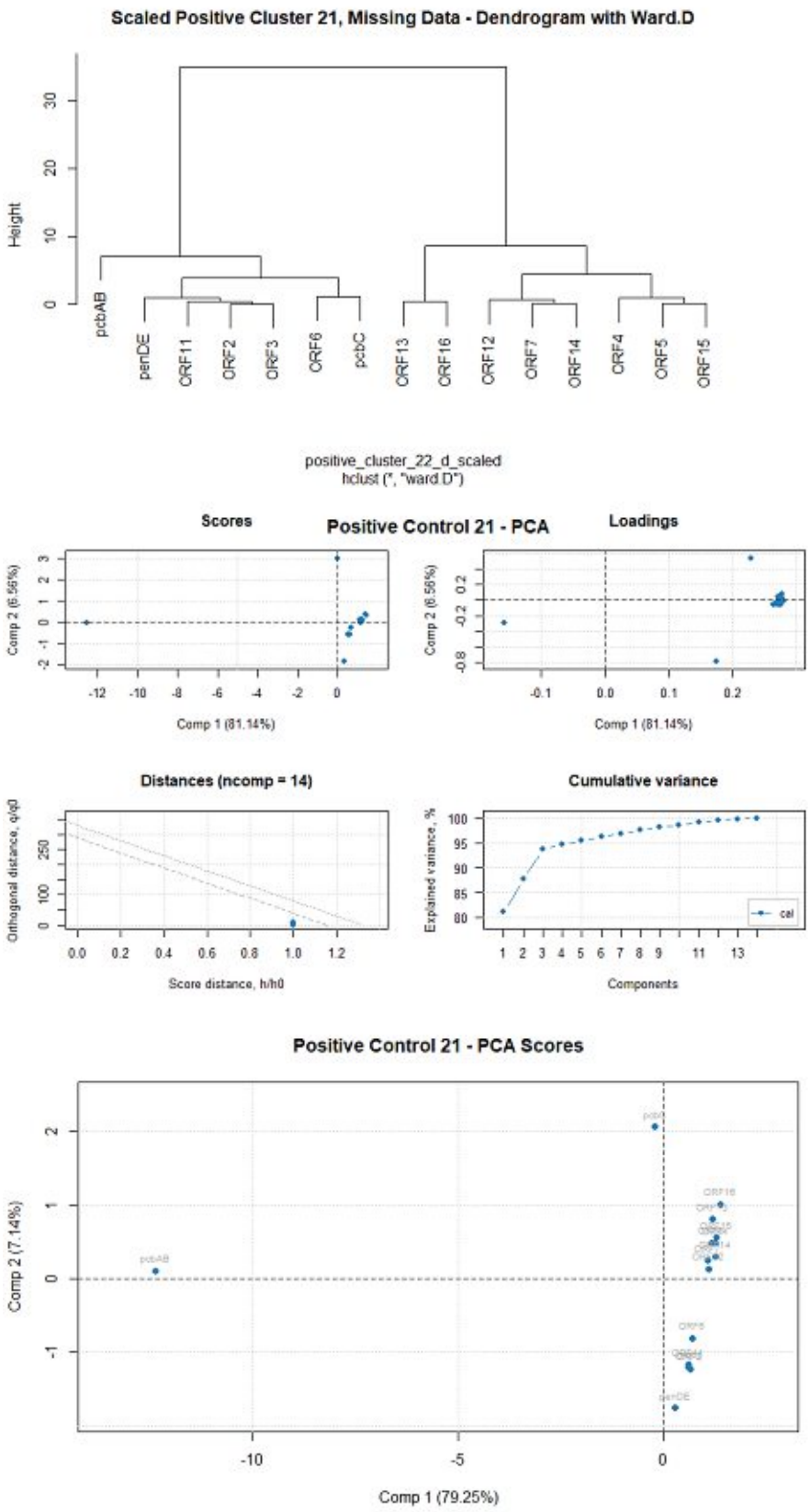


Supplement 9.53: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 21 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

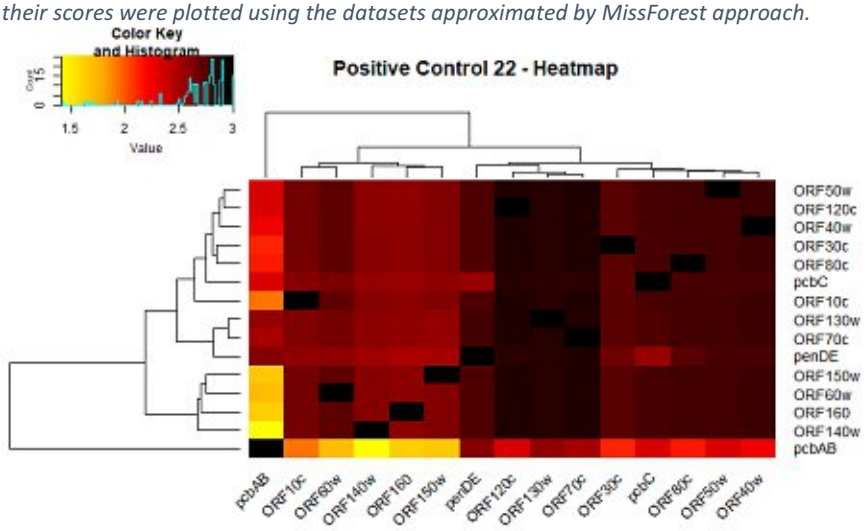


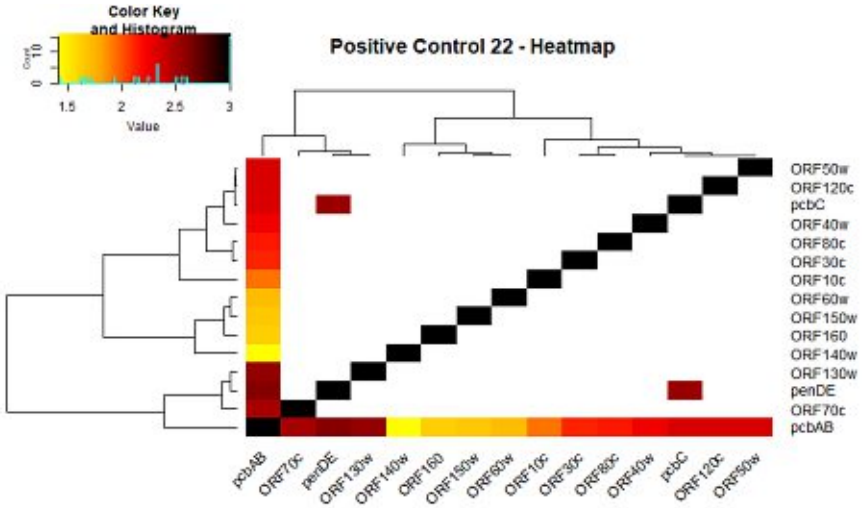


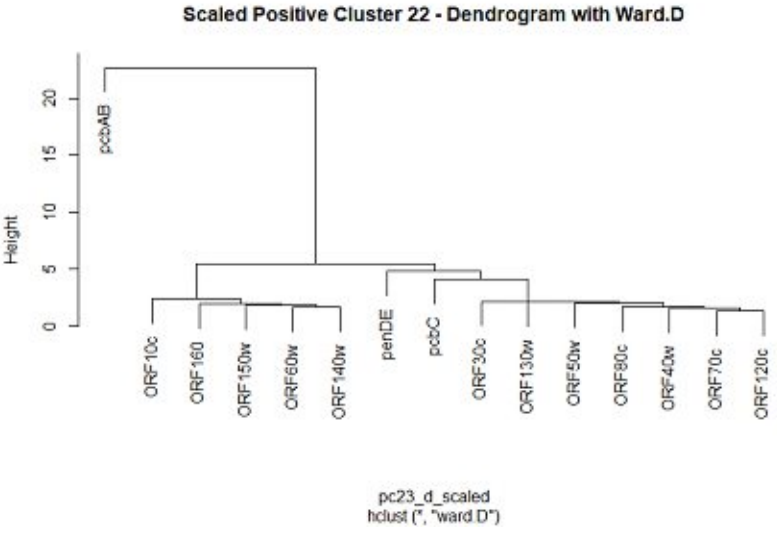


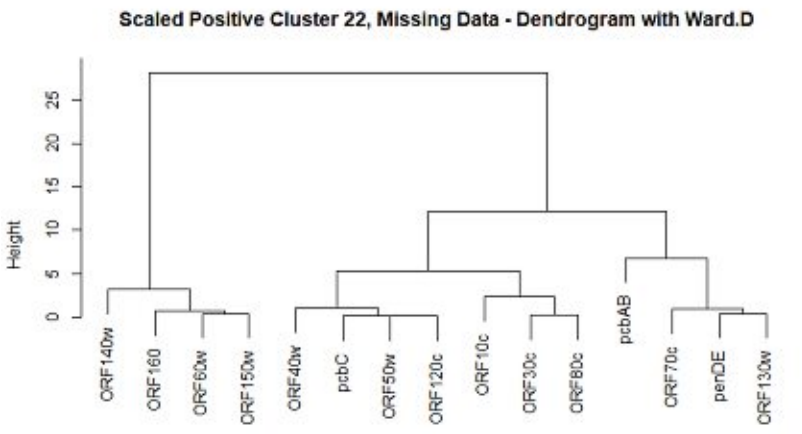


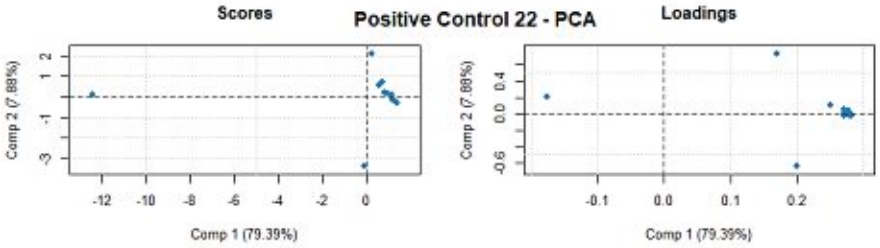
Supplement 9.54: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 22 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

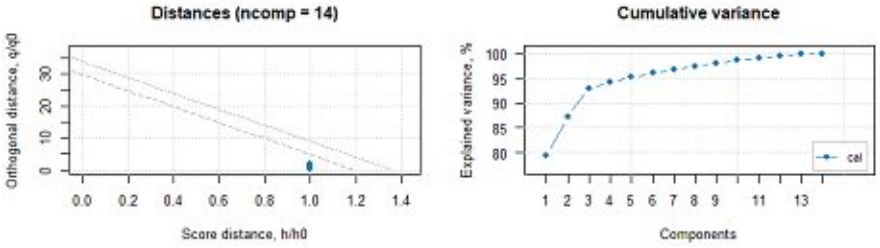


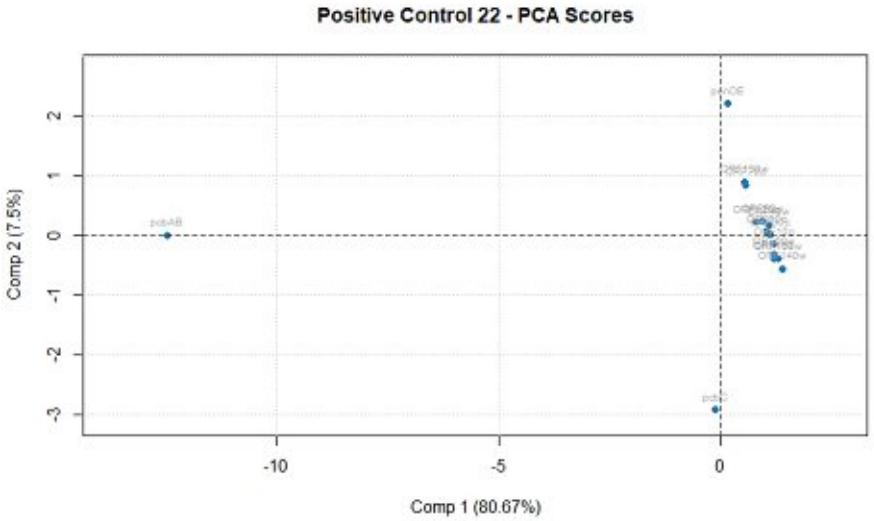




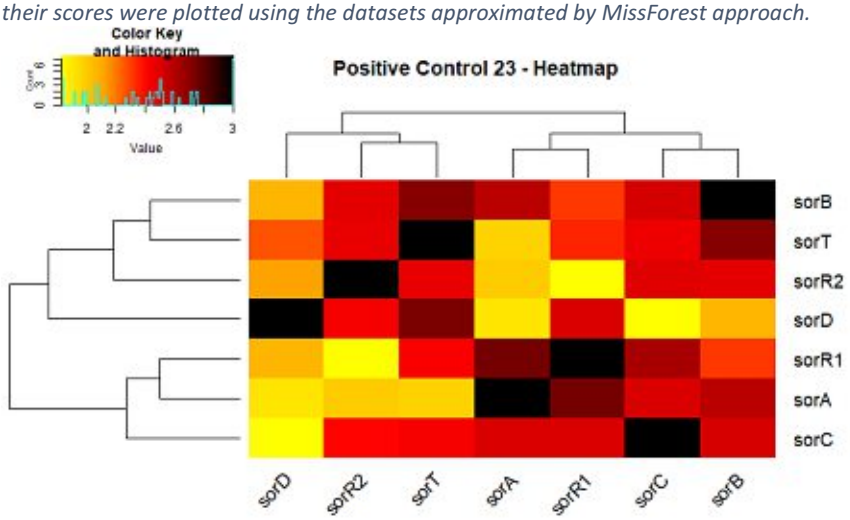


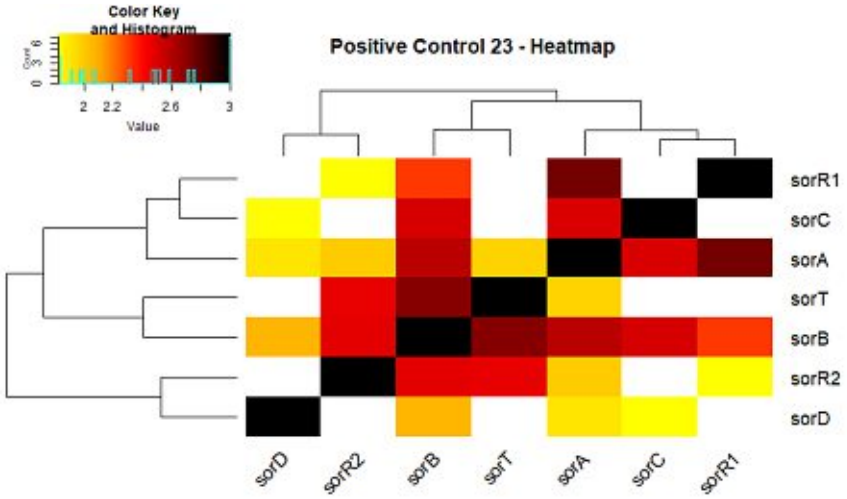


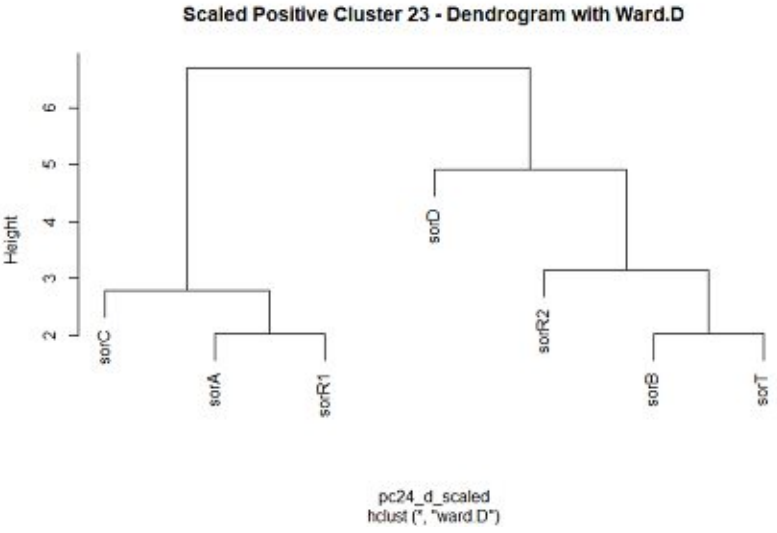


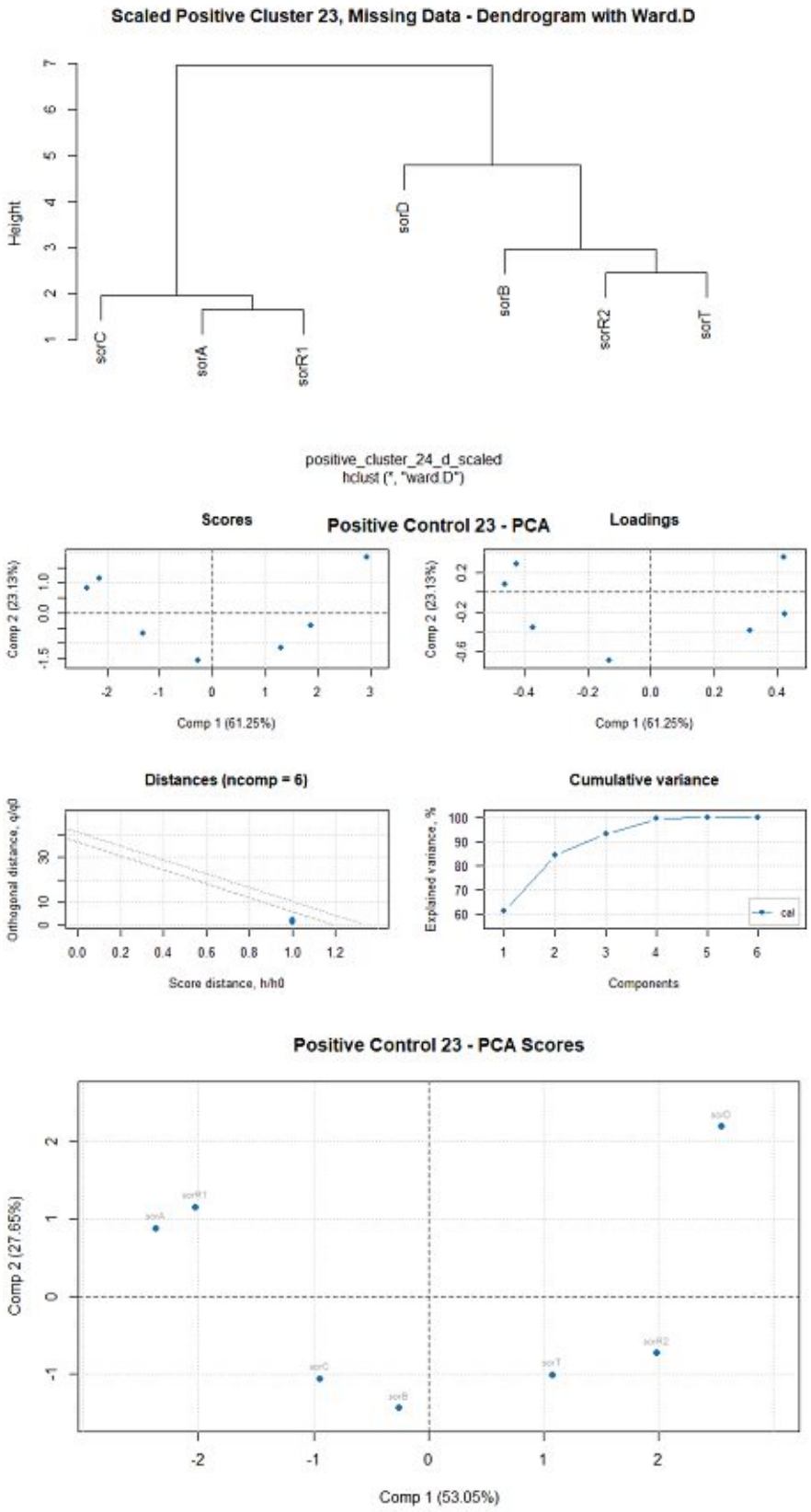


Supplement 9.55: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 23 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

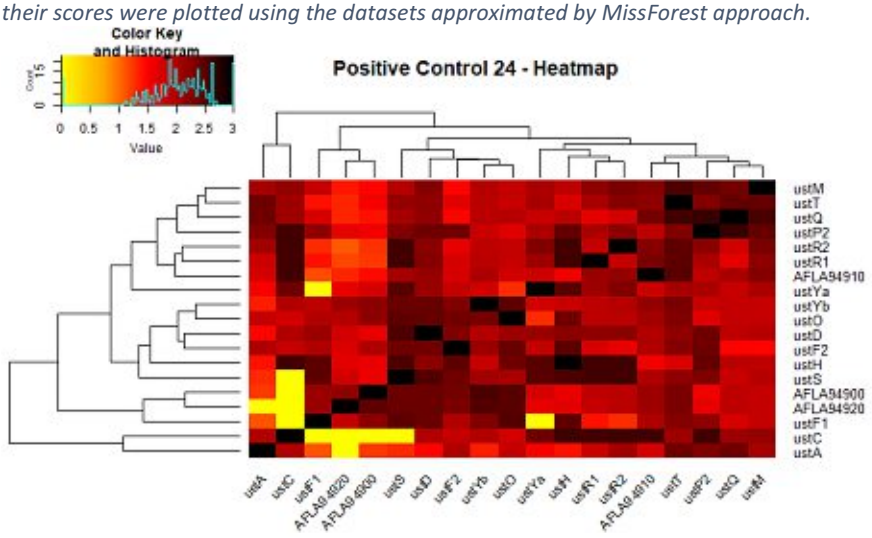


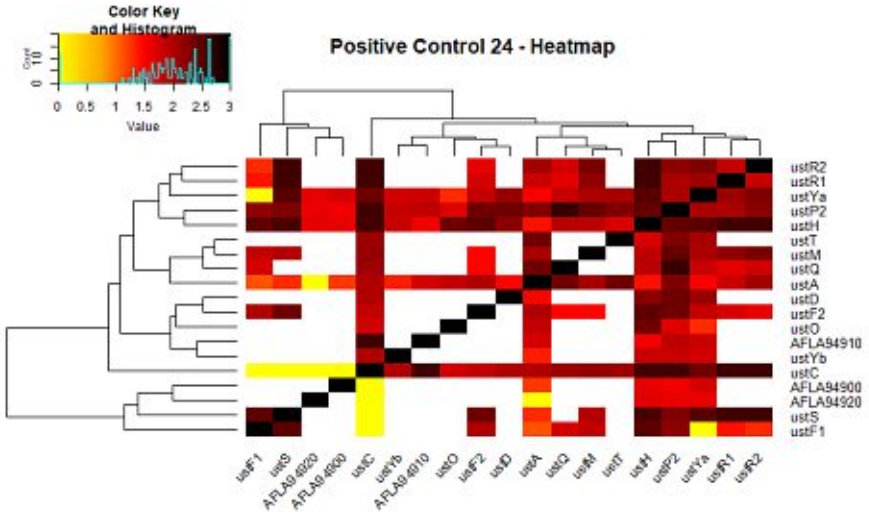


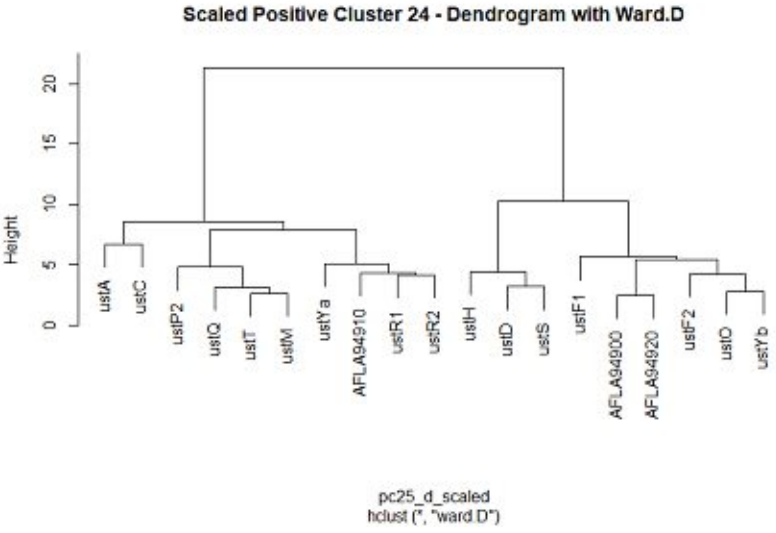


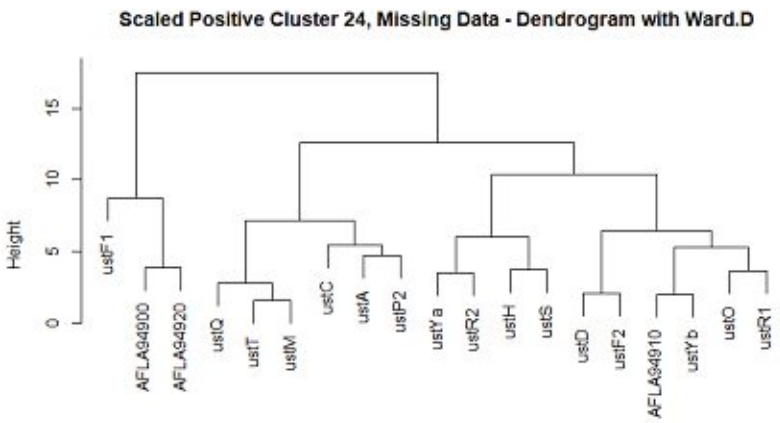


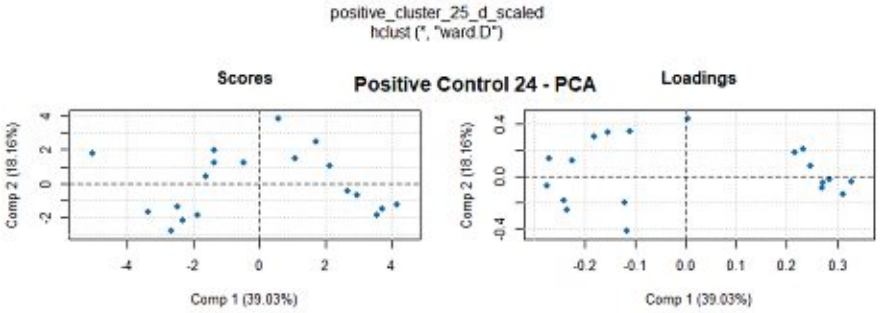
Supplement 9.56: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 24 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

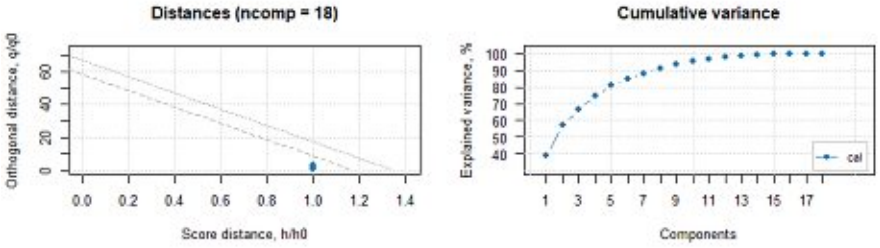


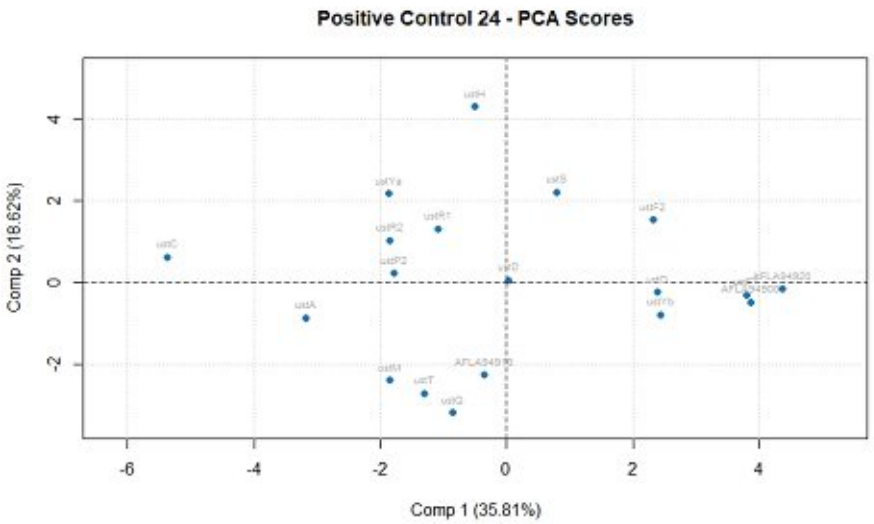




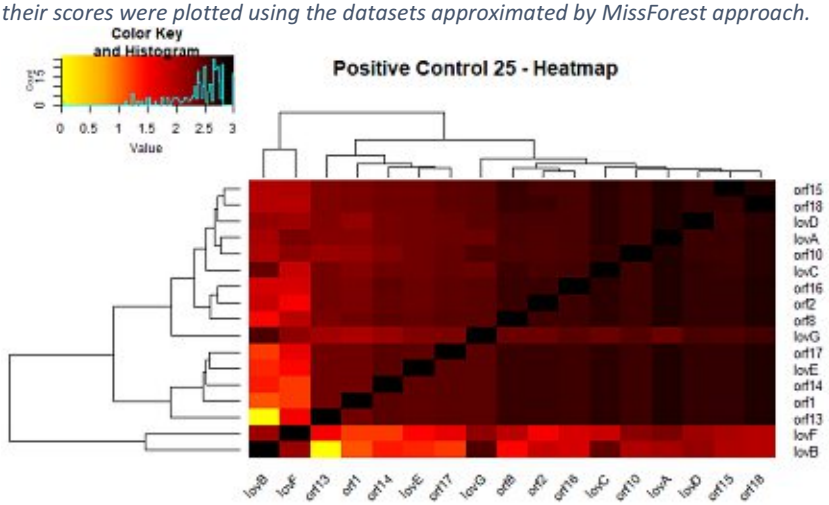


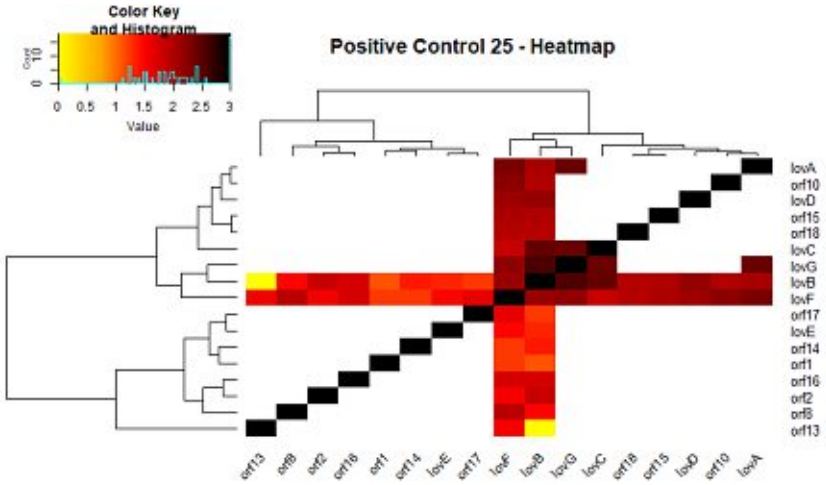


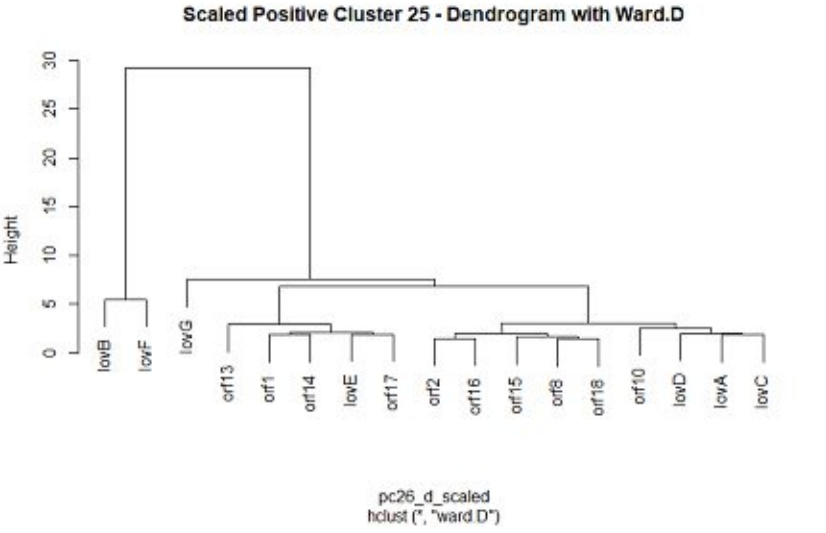


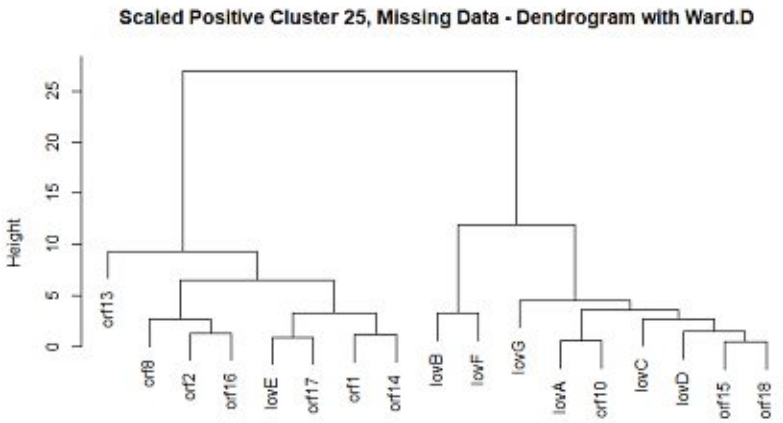


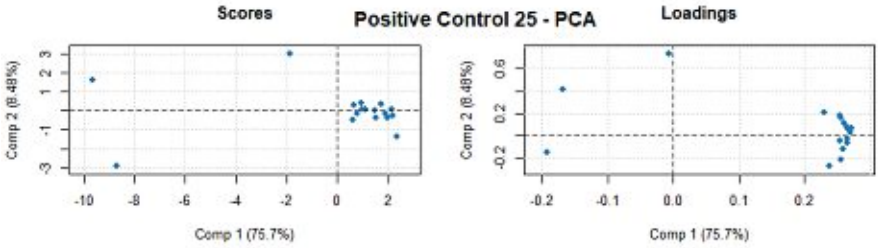
Supplement 9.57: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 25 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

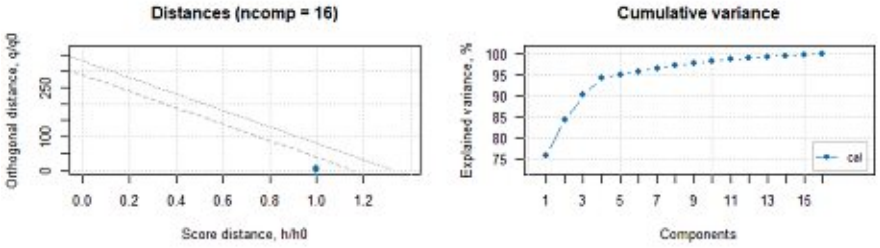


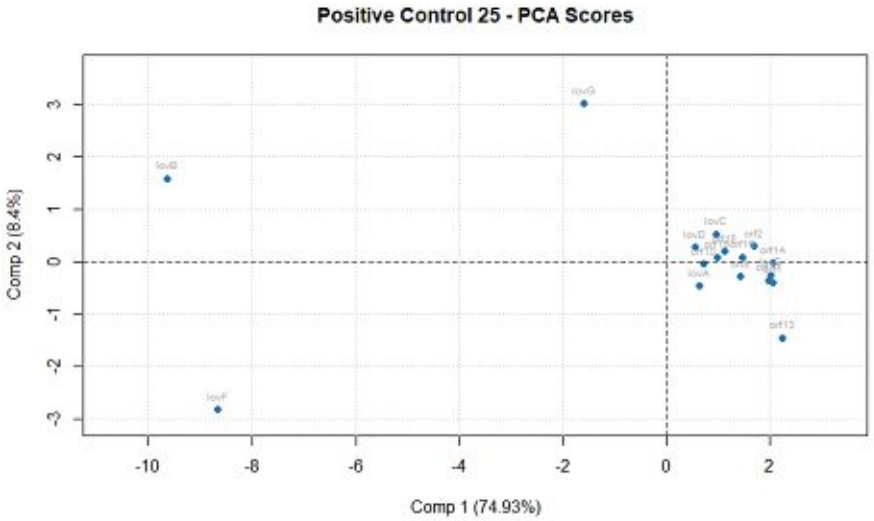




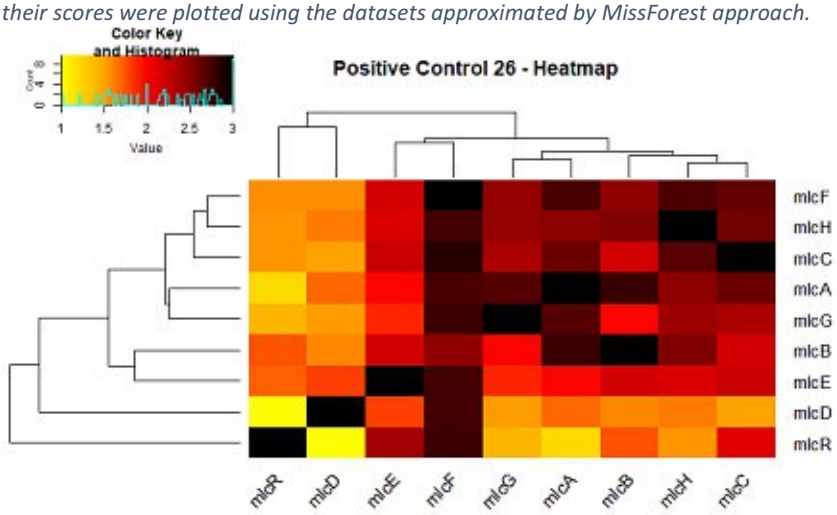


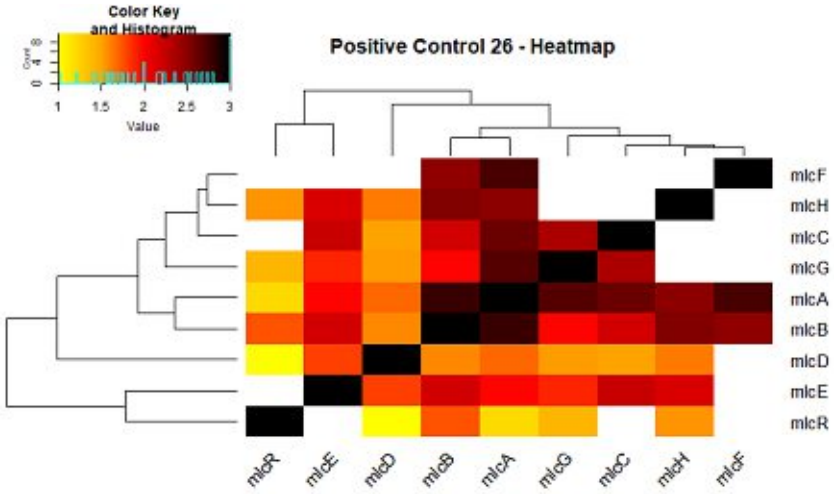


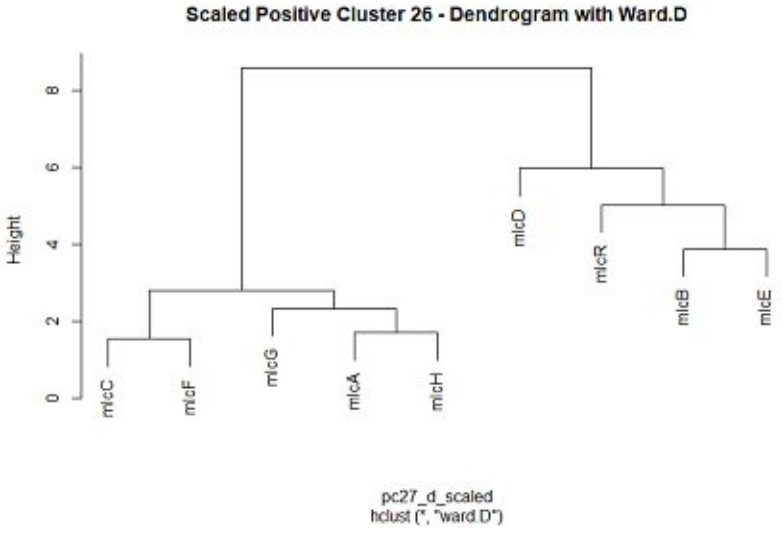


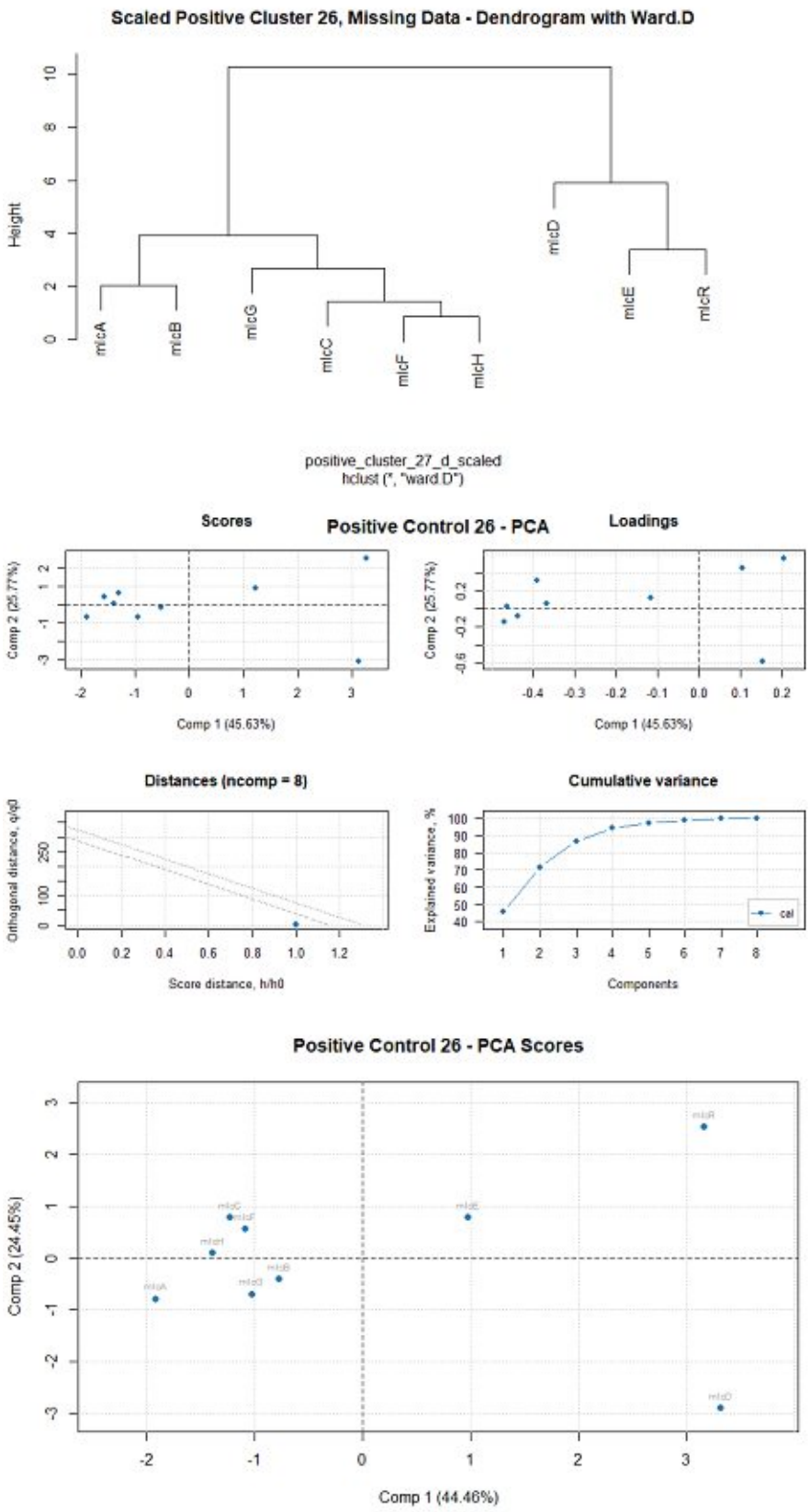


Supplement 9.58: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 26 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

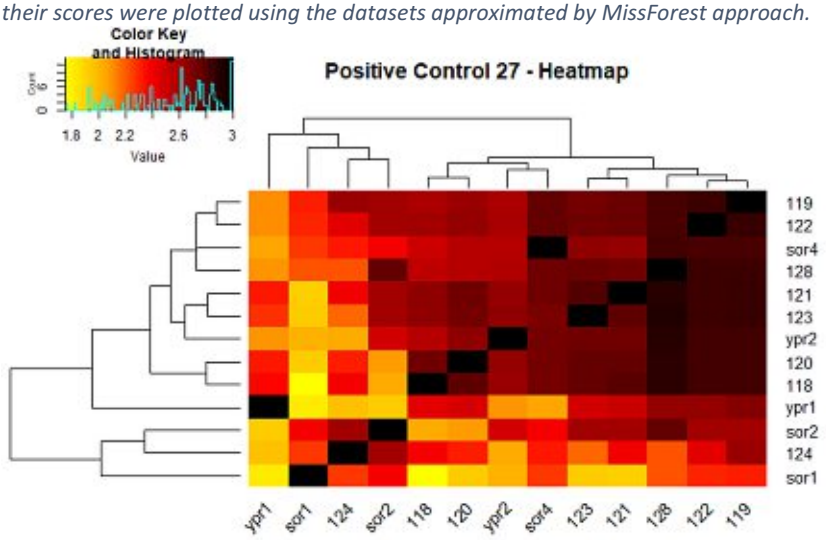


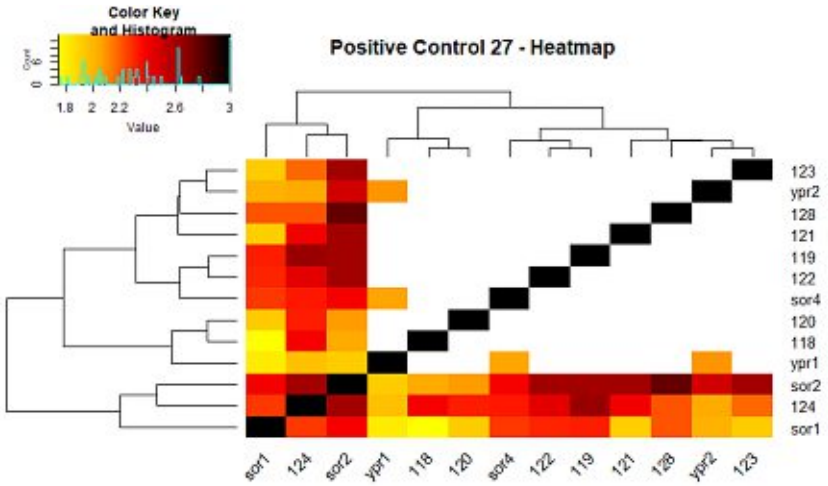


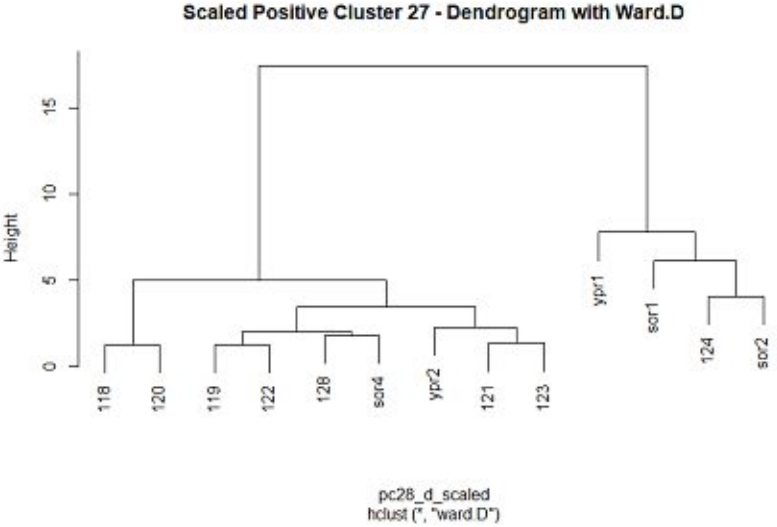


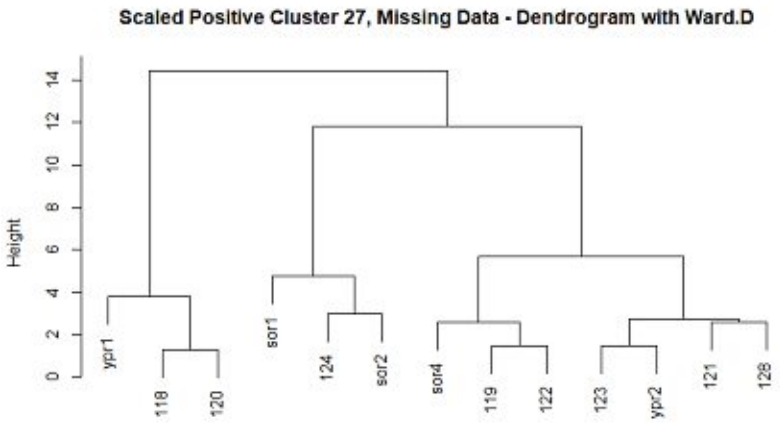


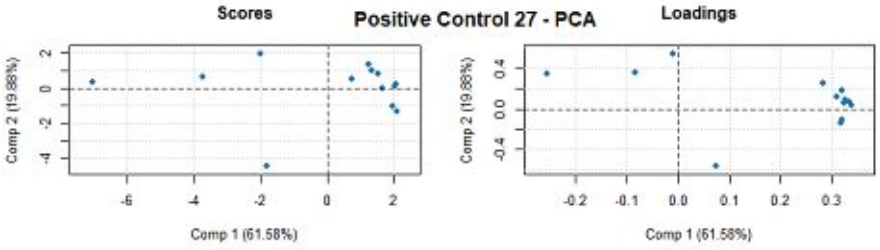
Supplement 9.59: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 27 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

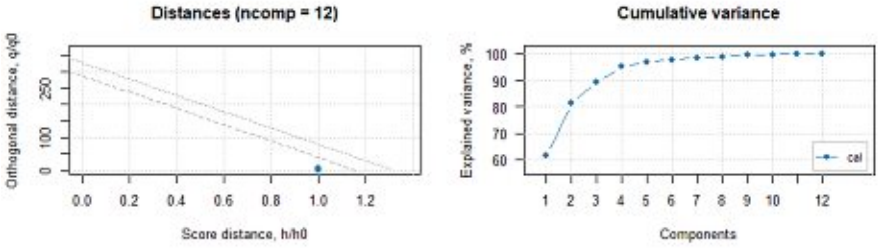


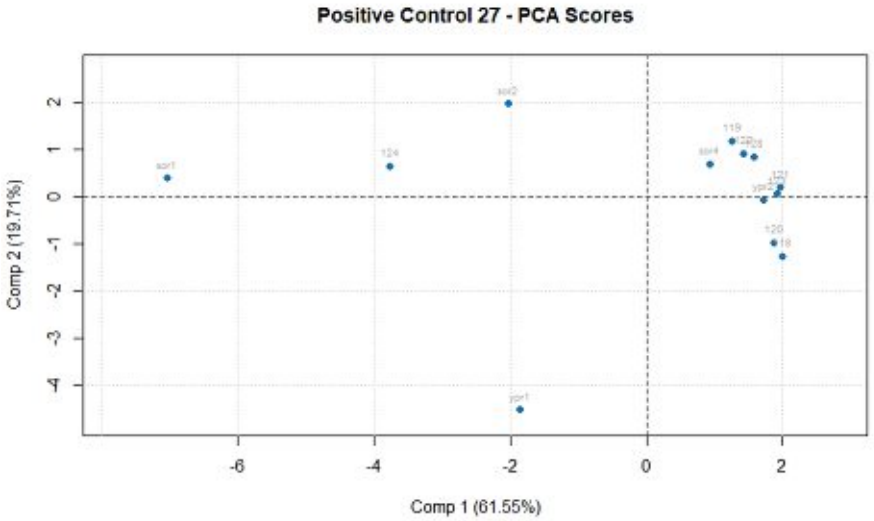




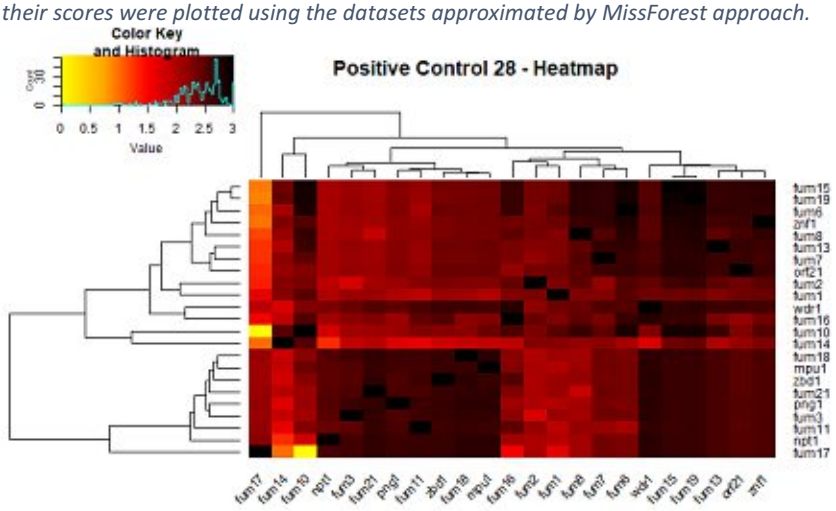


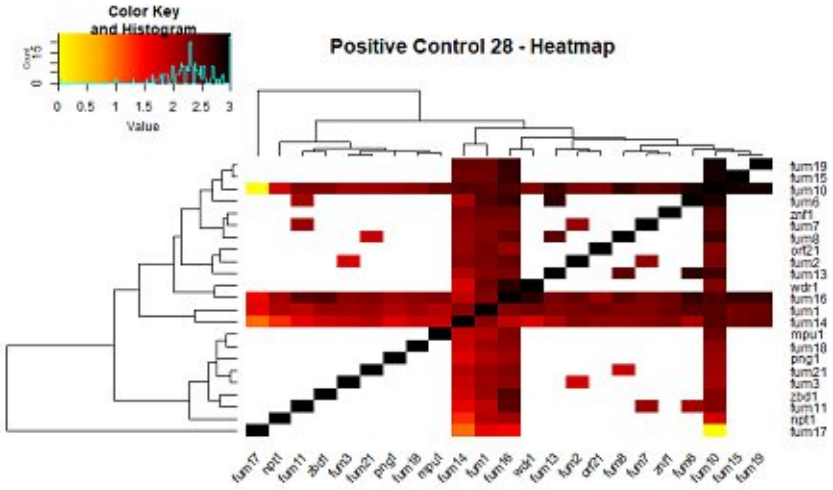


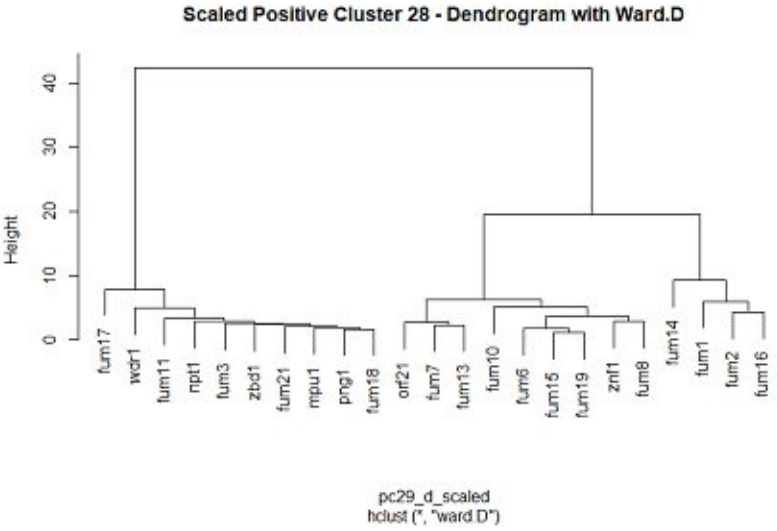


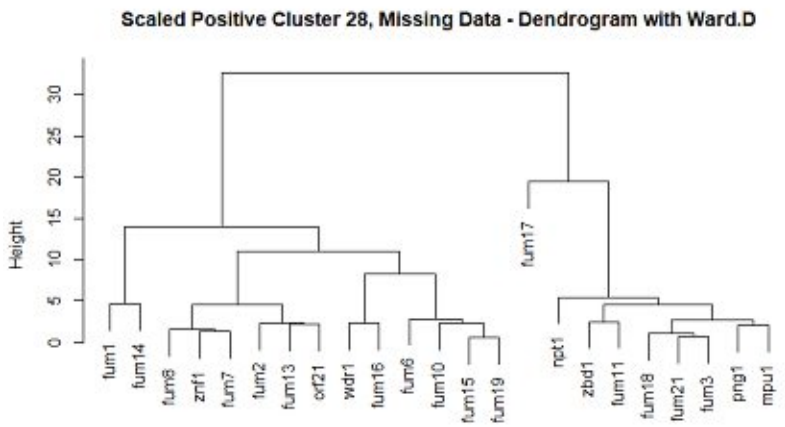


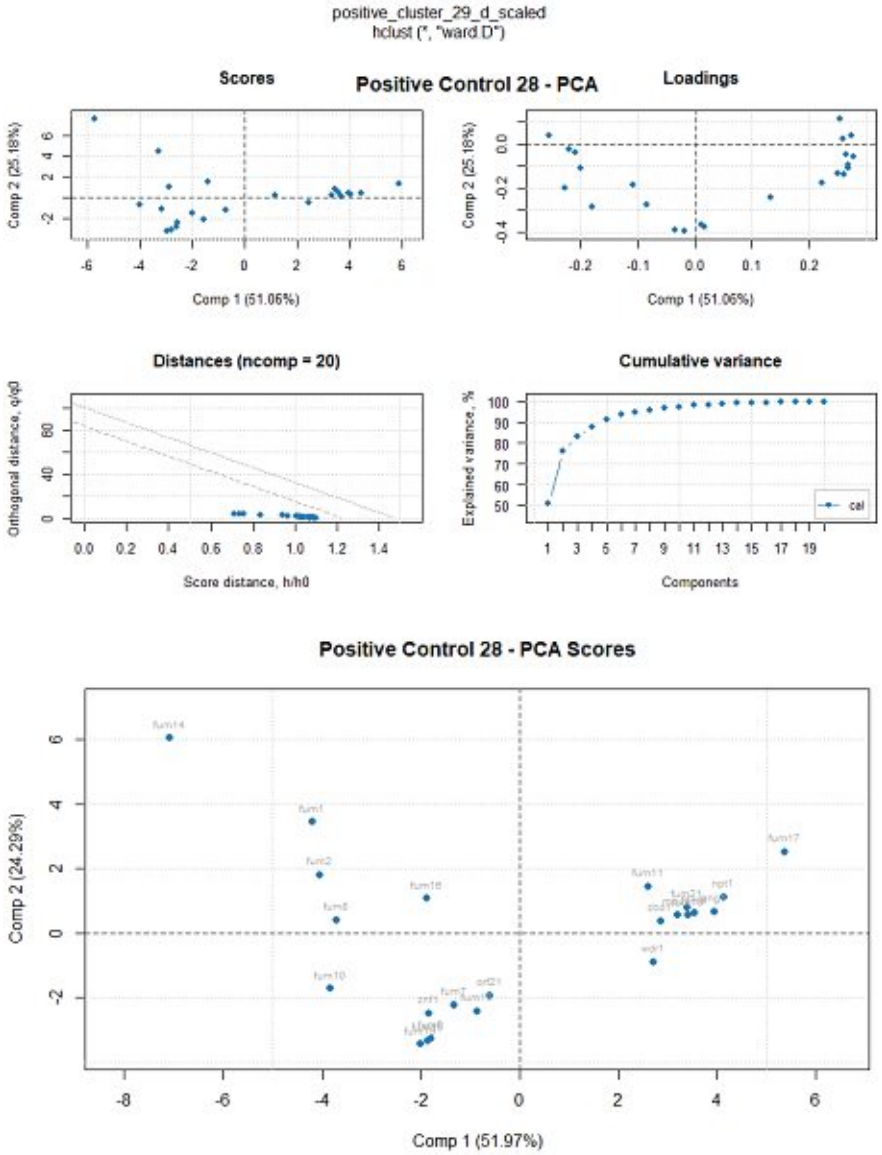
Supplement 9.60: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 28 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.



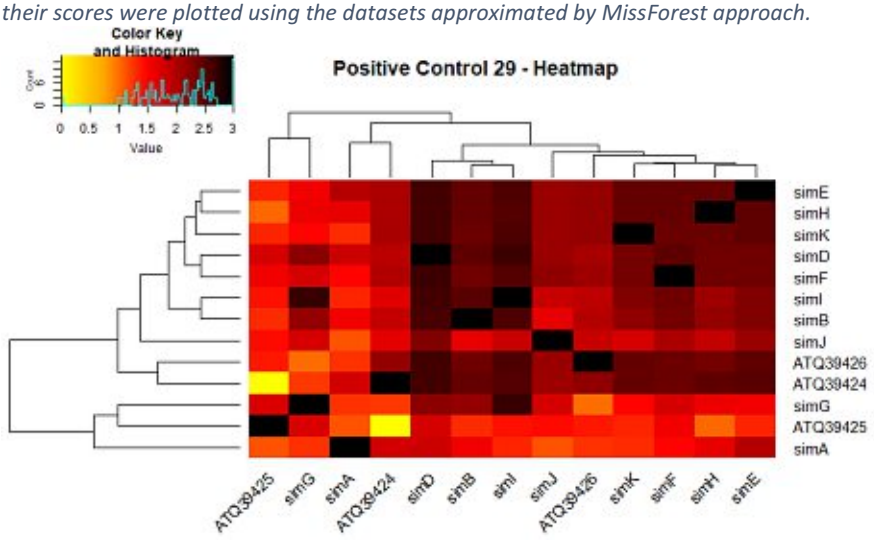


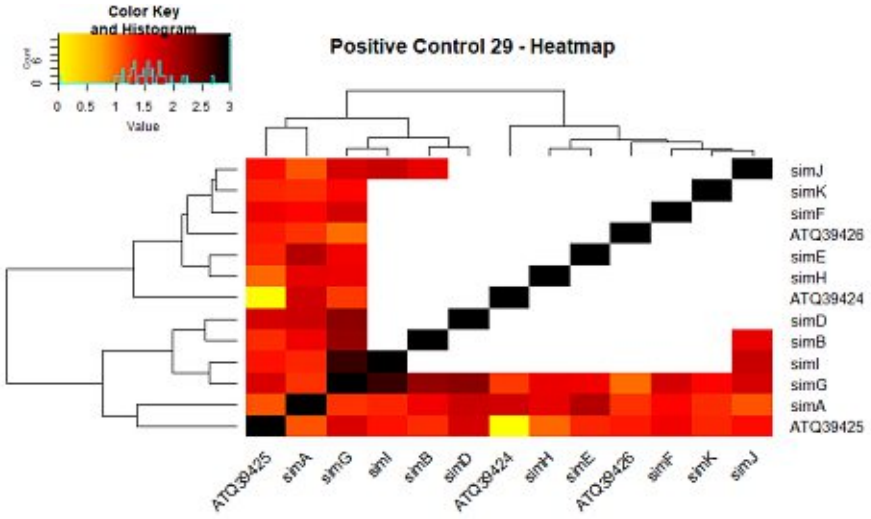


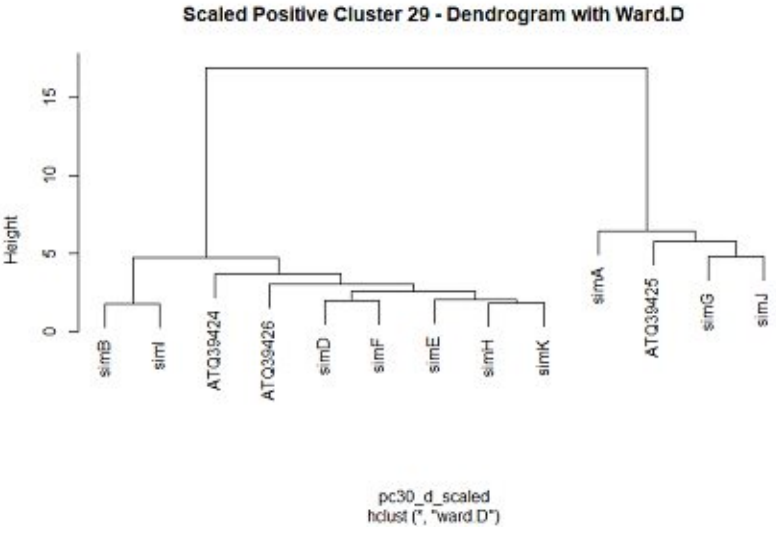


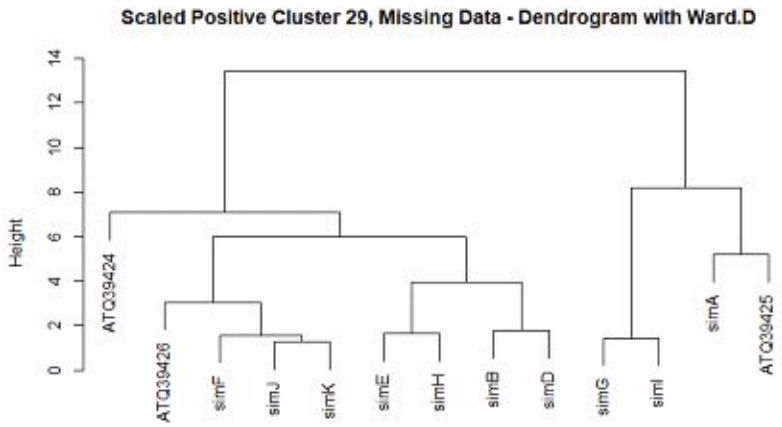


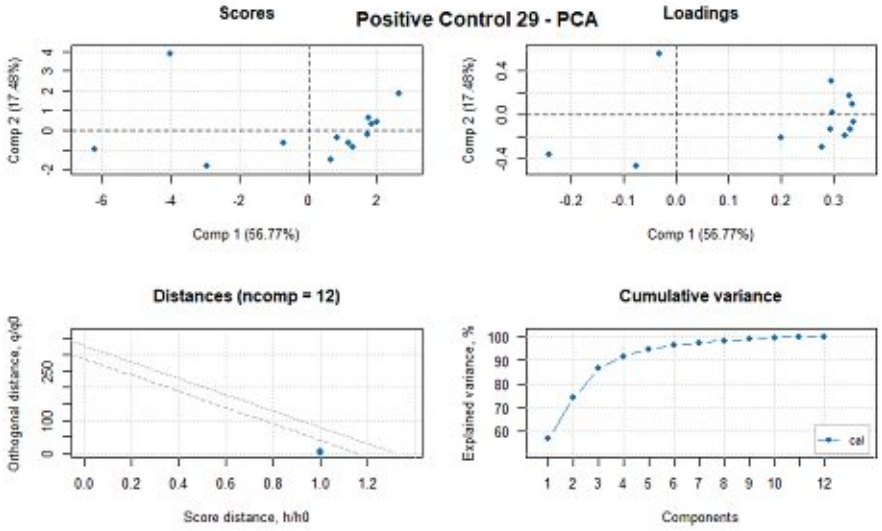
Supplement 9.61: Heatmaps, Dendrograms and principal component analysis (PCA) of Positive control 29 used for the statistical analysis. The heatmaps were calculated using approximated data (top) and the missing data (bottom). To obtain the dendrograms, the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA and their scores were plotted using the datasets approximated by MissForest approach.

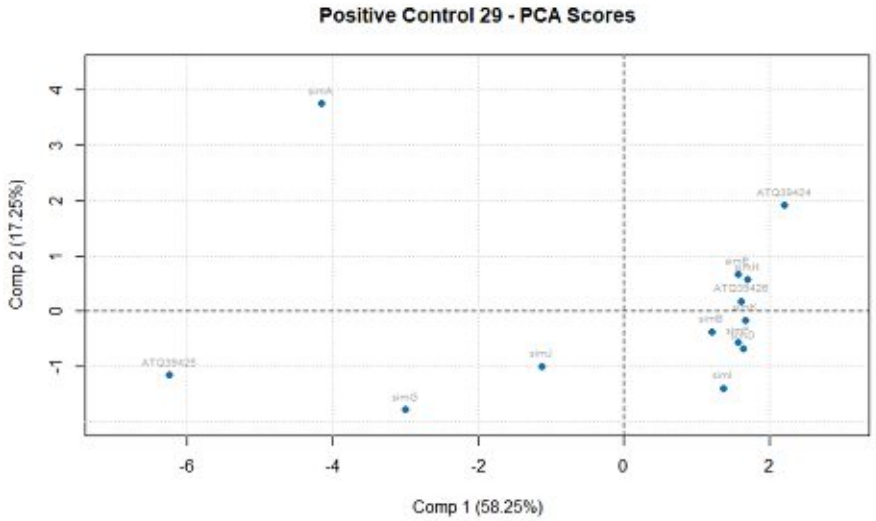




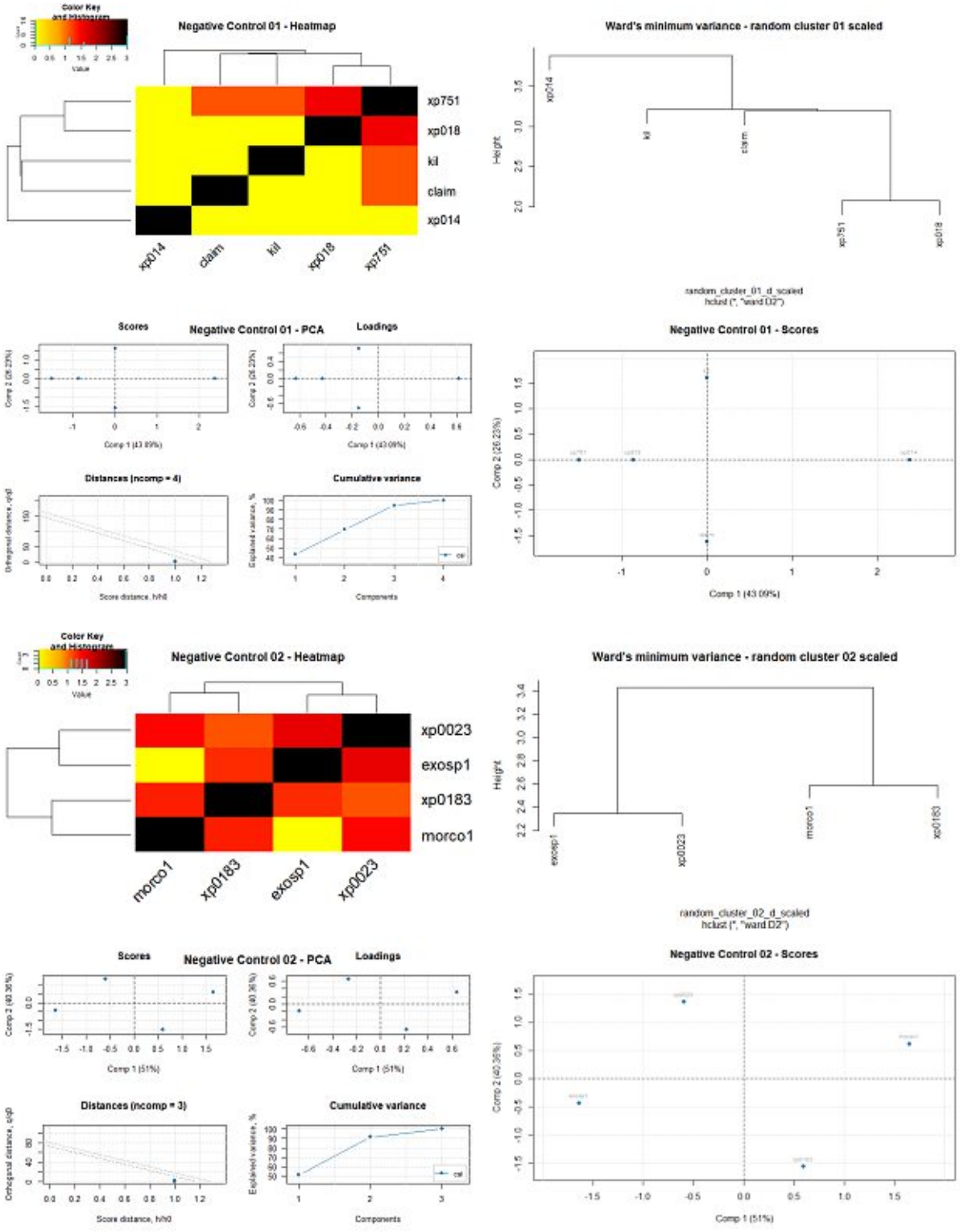


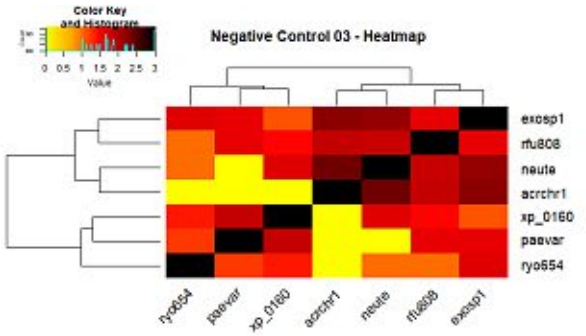


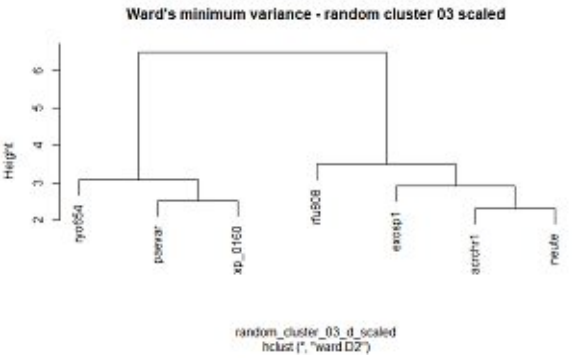


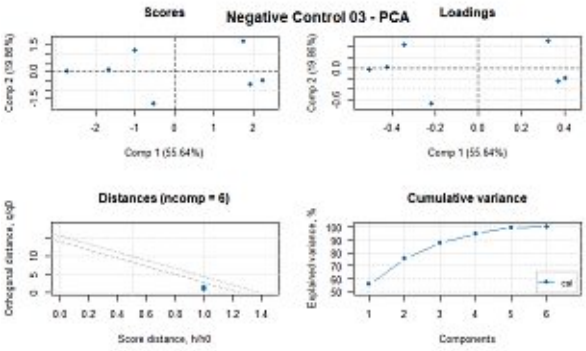


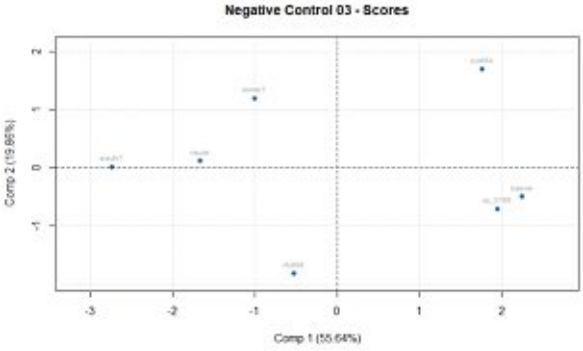
Supplement 9.62: Heatmaps, Dendrograms and principal component analysis (PCA) of all negative controls. As all of the phylogenetic trees obtained from FunOrder were evaluated, they include no missing data. Hence, only one heatmap plot per control were computed. To calculate the dendrograms (labelled as "Ward's minimum variance – random cluster"), the former conducted matrices were scaled, and a distance matrix were returned. The illustrated dendrograms were then computed using the hclust function and the ward clustering method ward.D in RStudio. PCA was calculated using the pca function of the package mdatools [99].

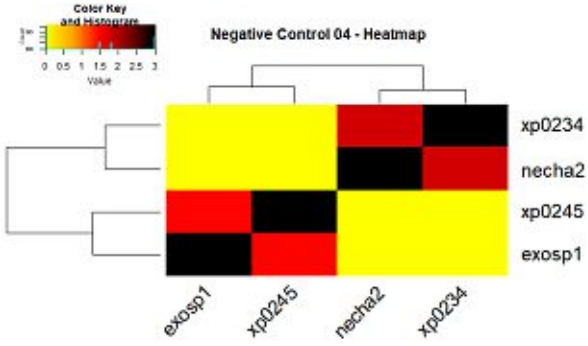


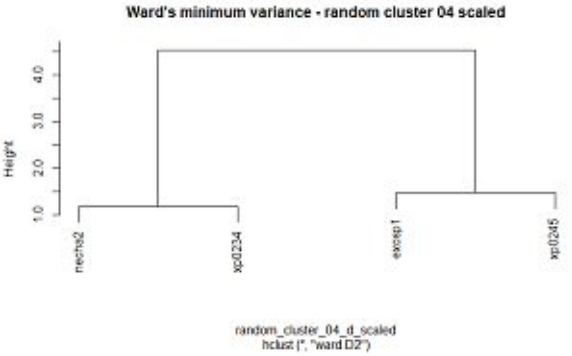


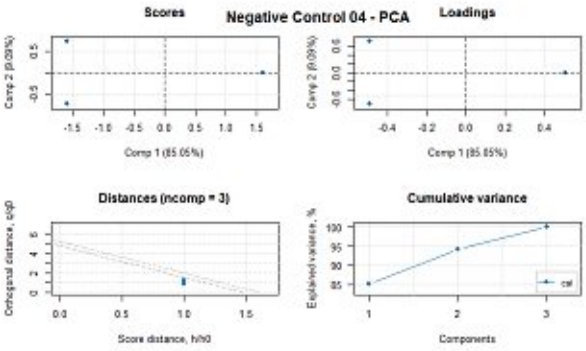


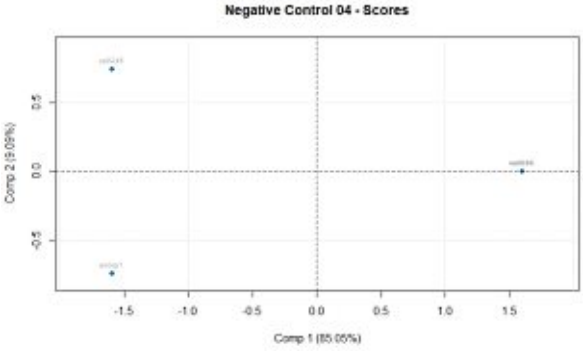


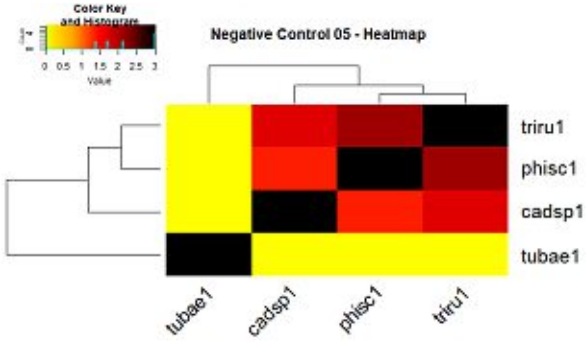


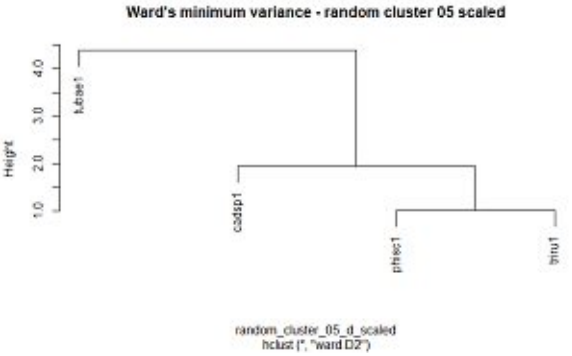


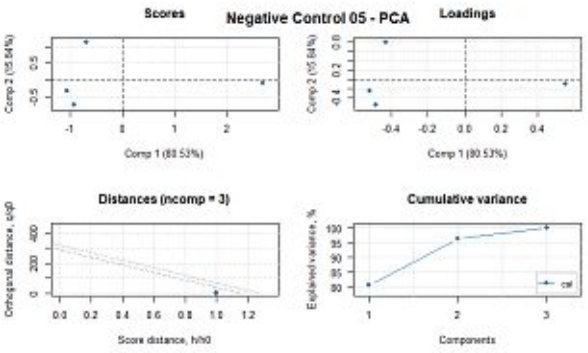


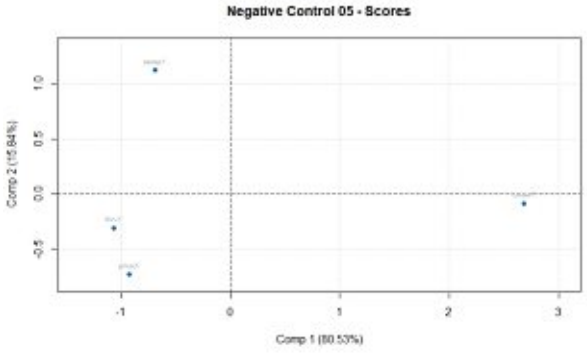


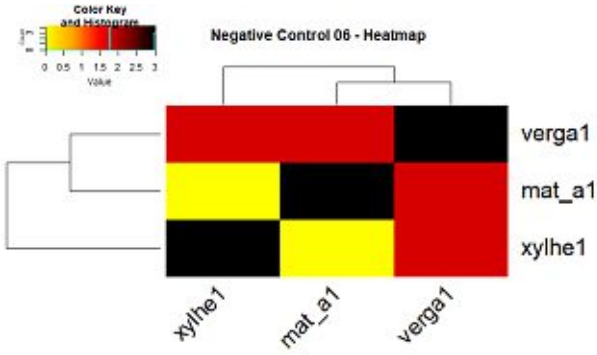


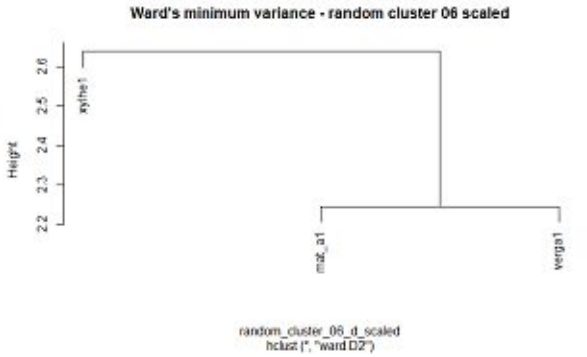


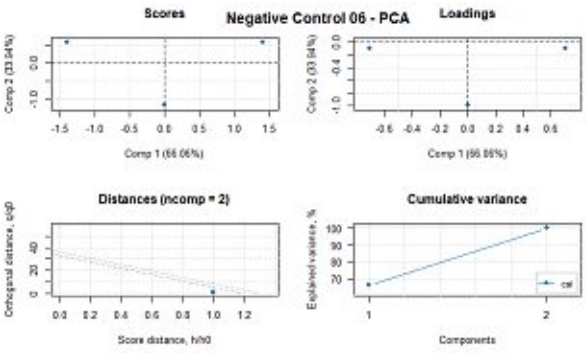


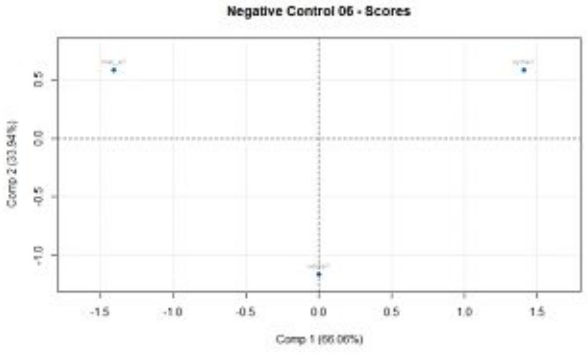


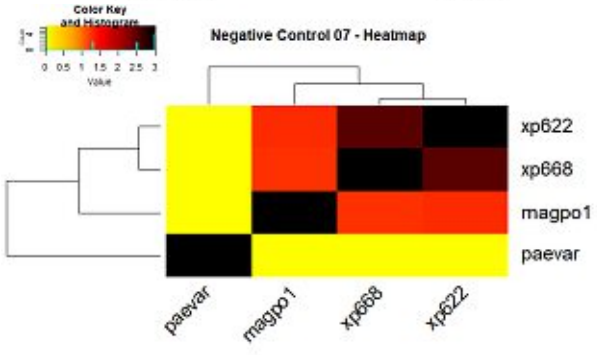


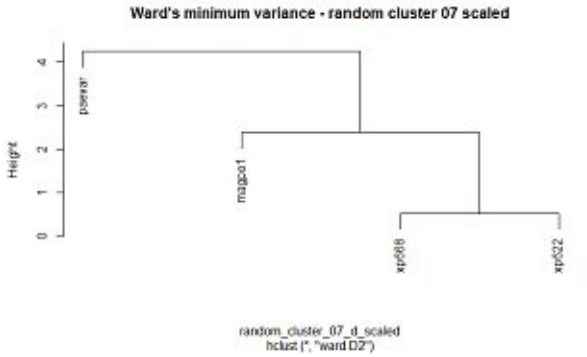


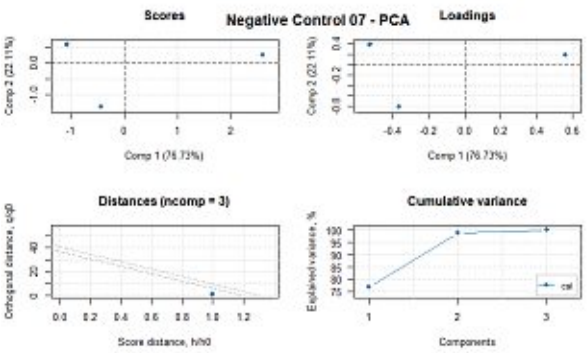


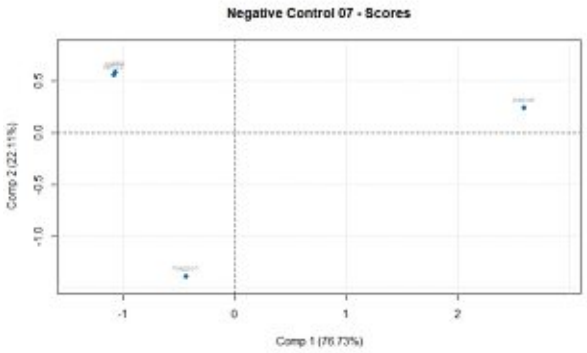


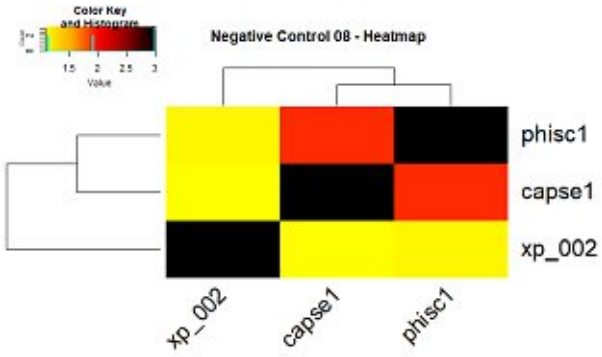


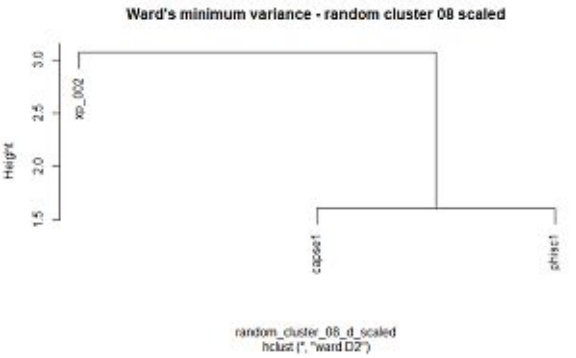


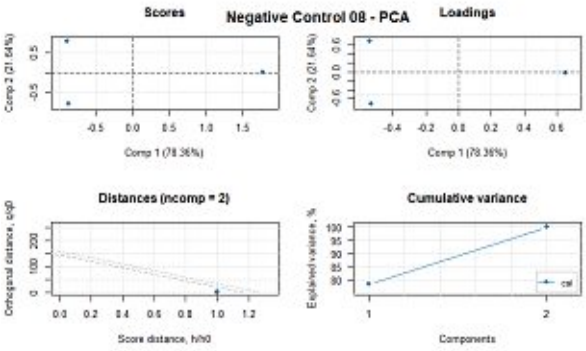


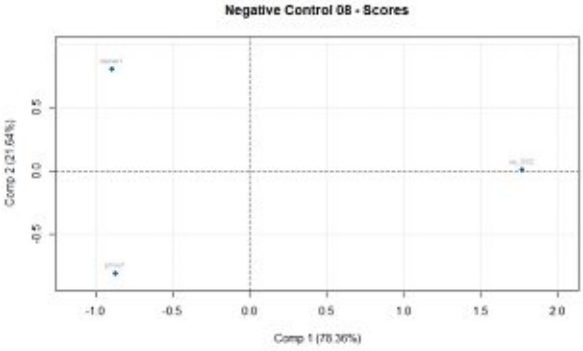


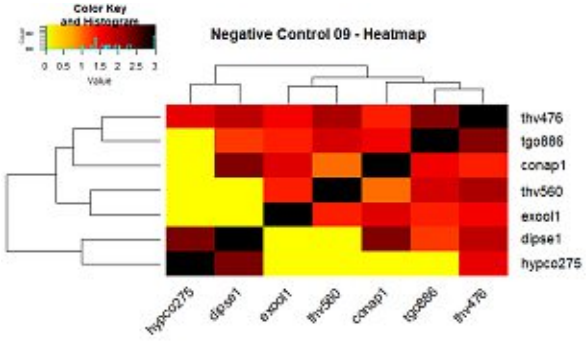


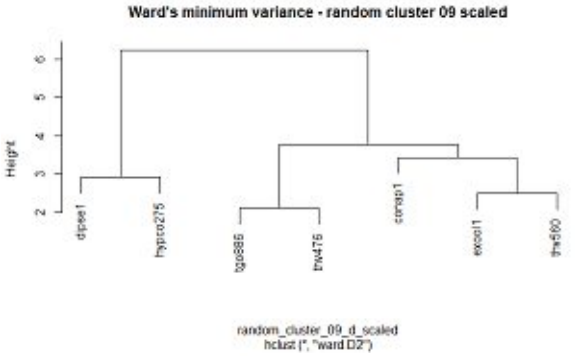


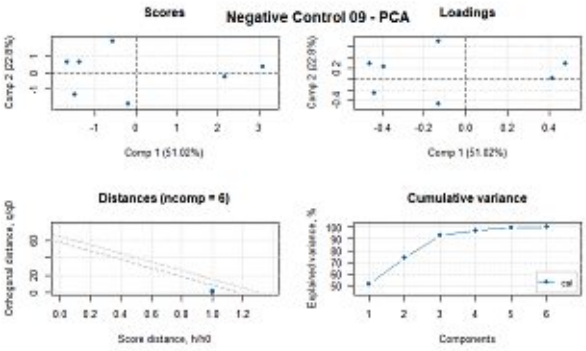


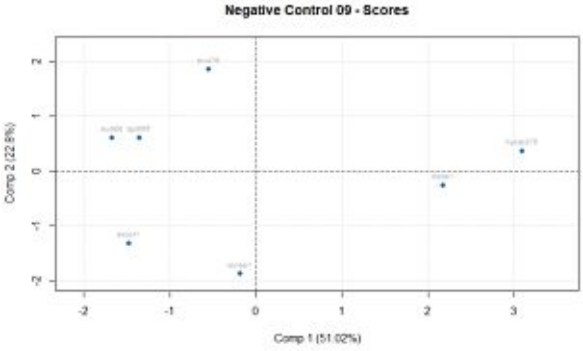


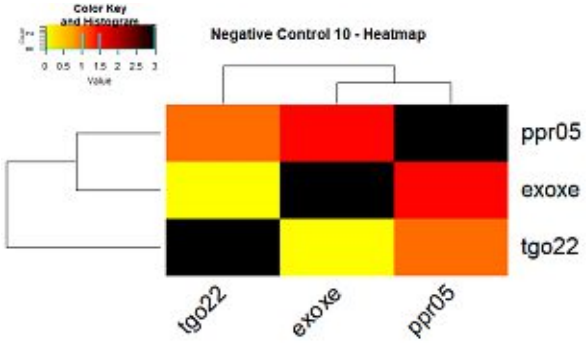


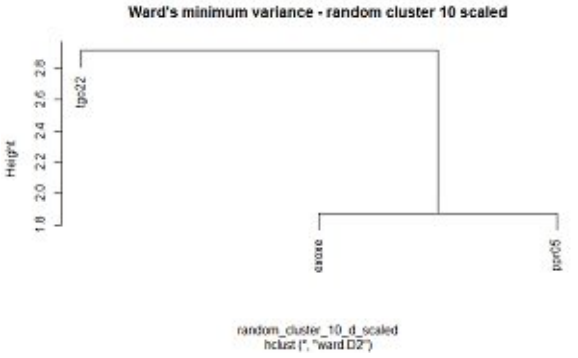


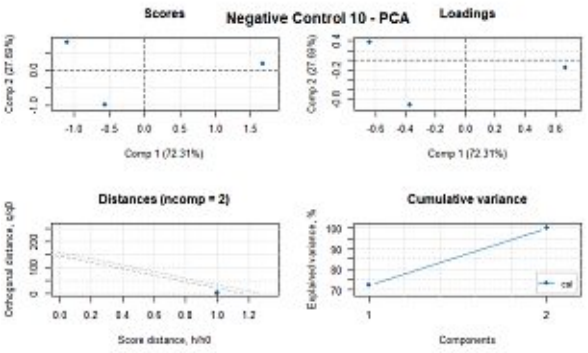


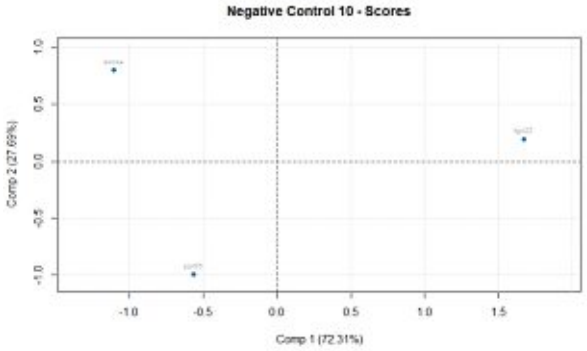


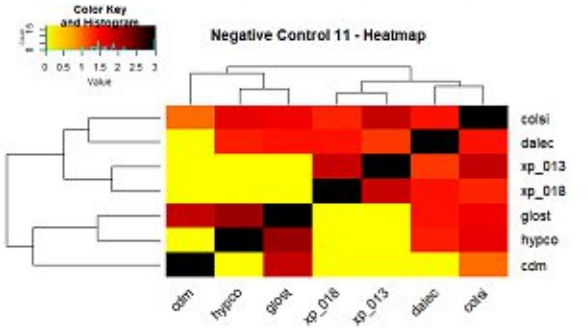


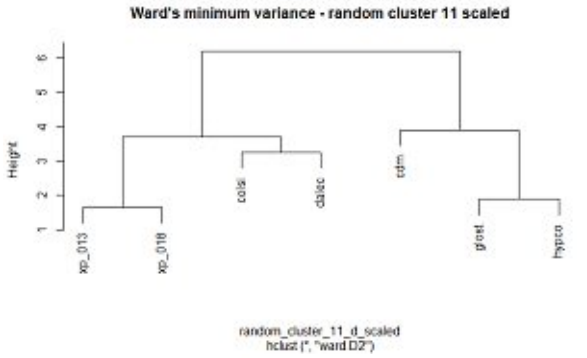


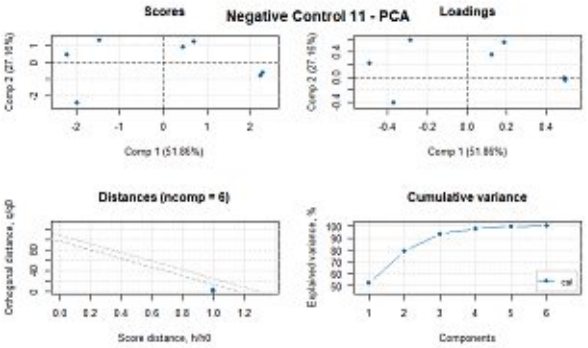


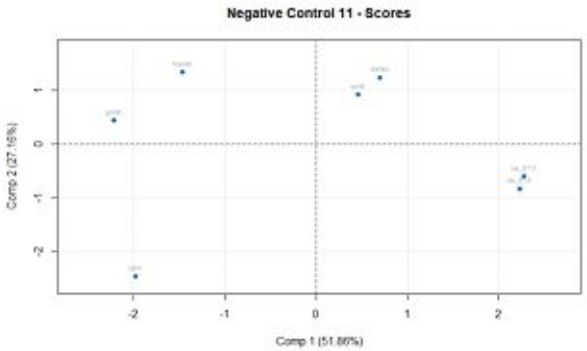


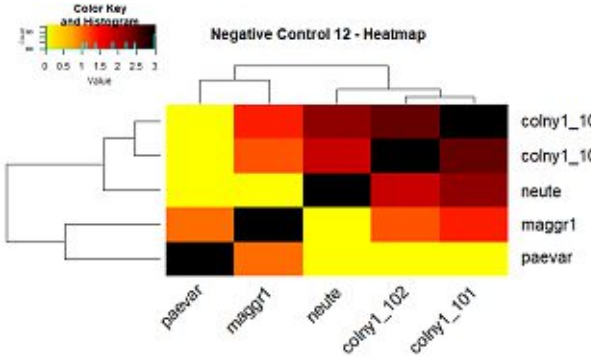


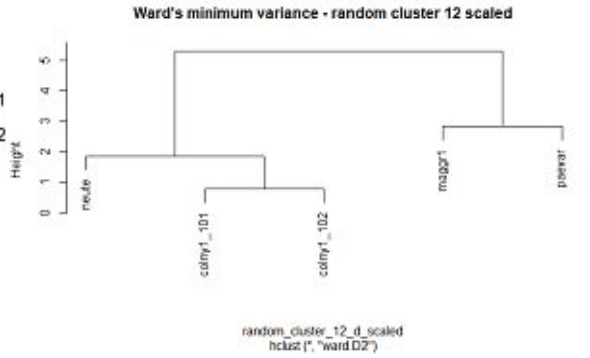


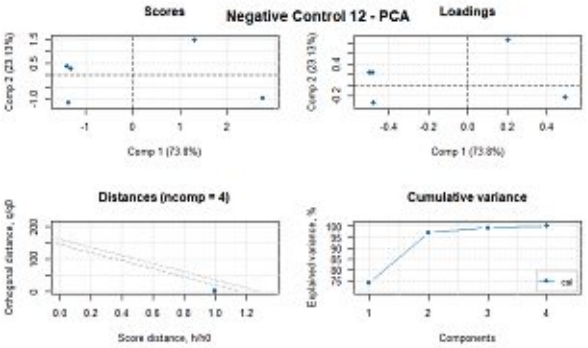


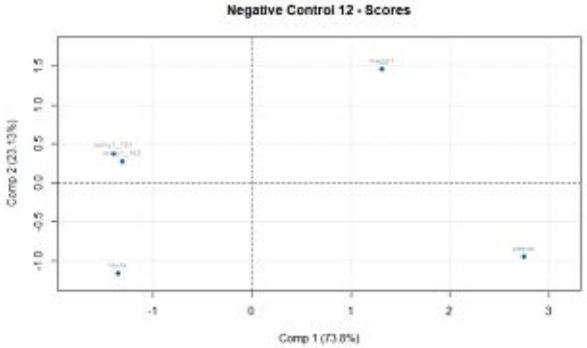


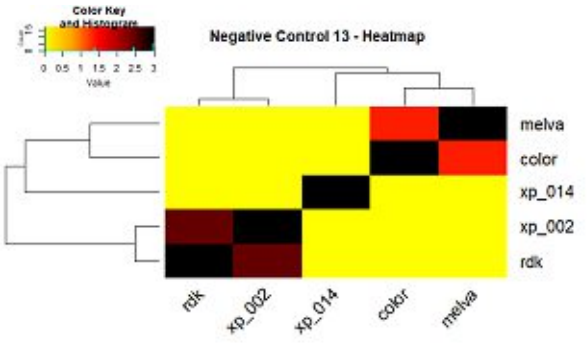


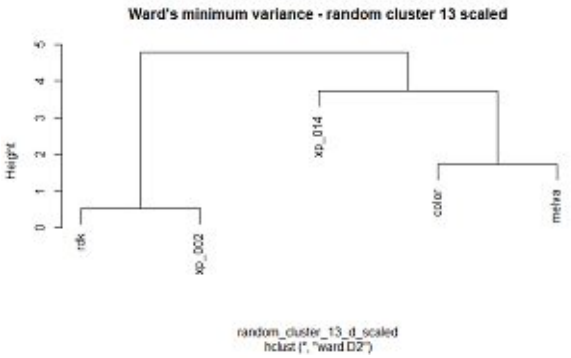


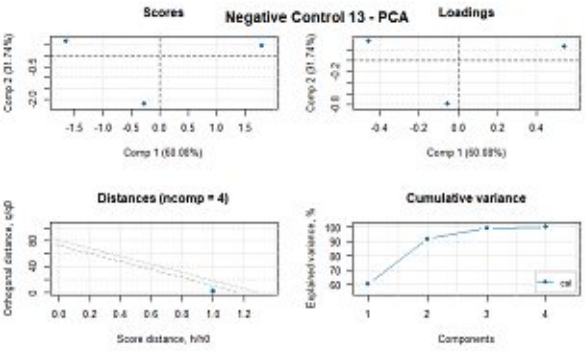


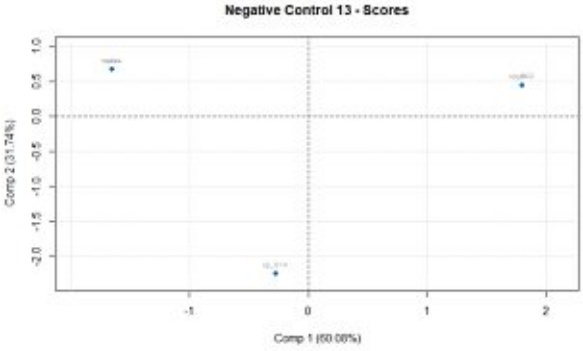


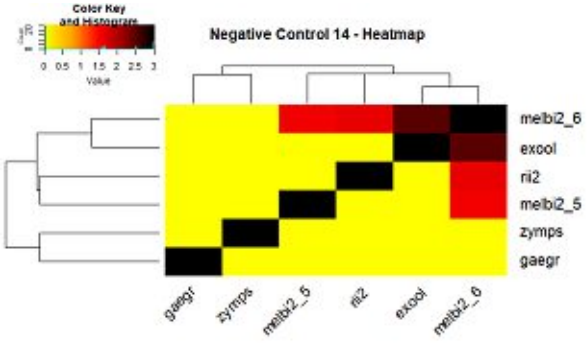


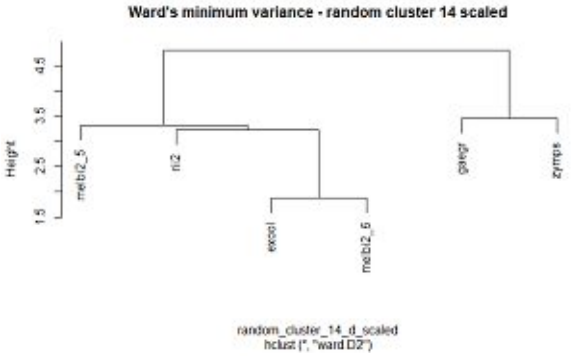


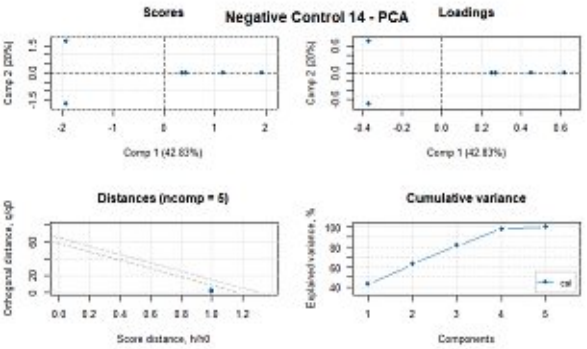


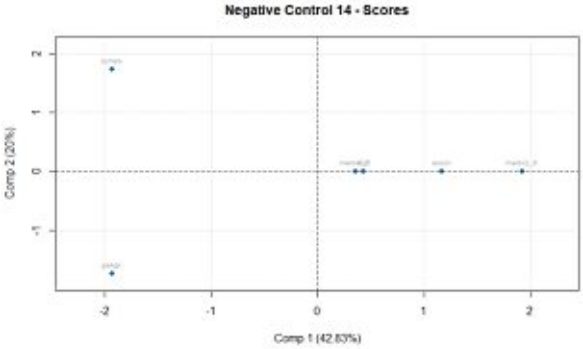


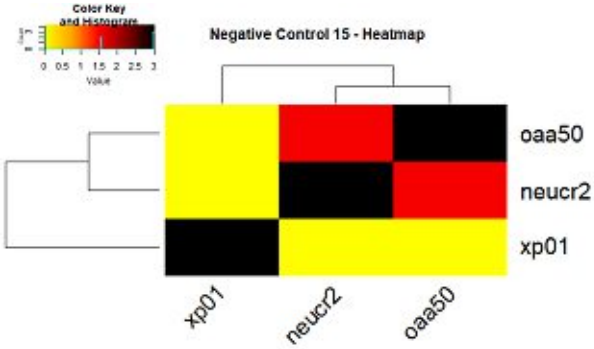


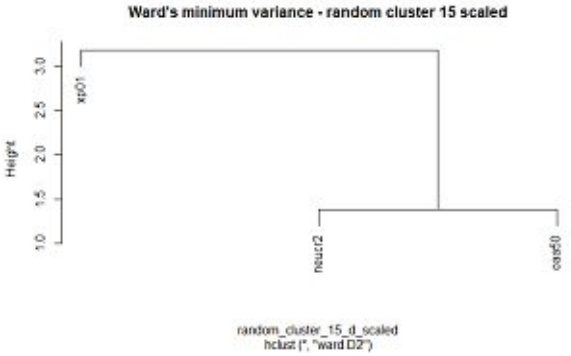


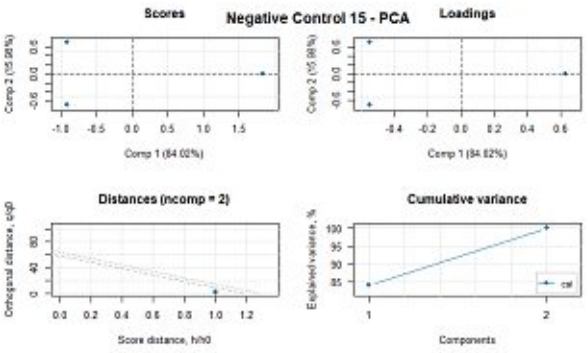


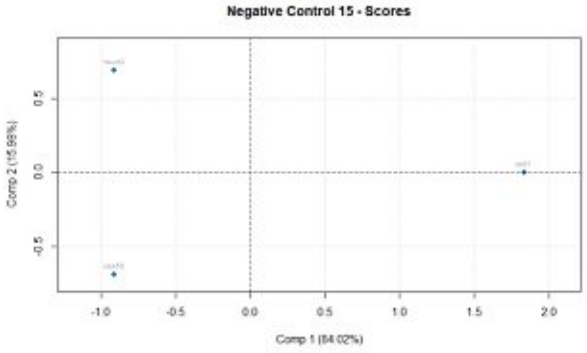


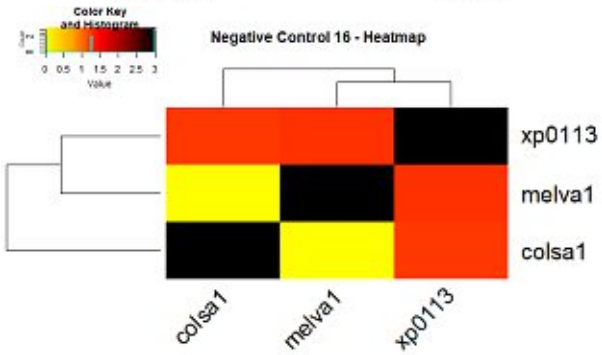


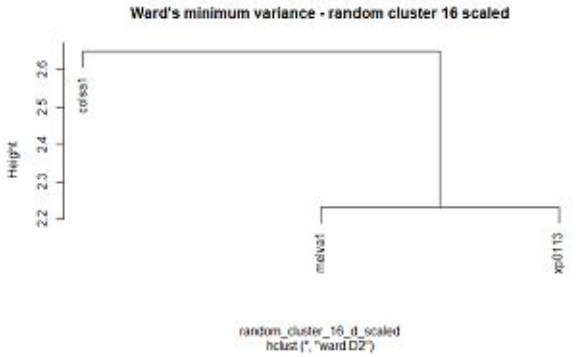


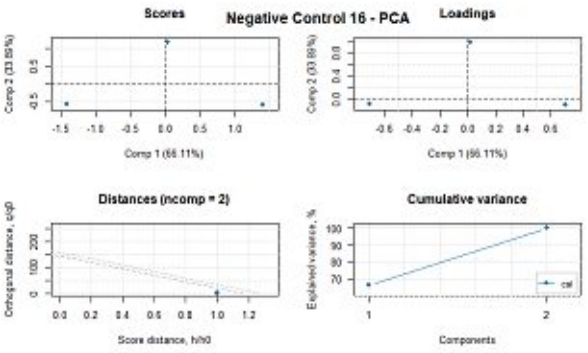


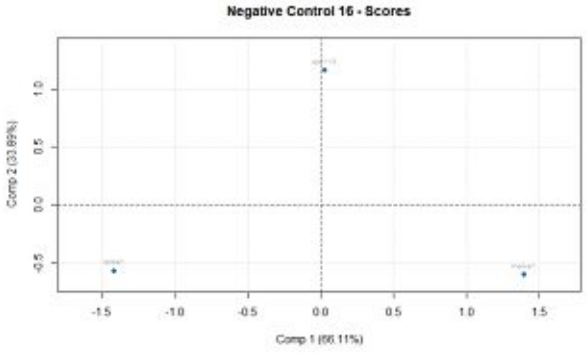


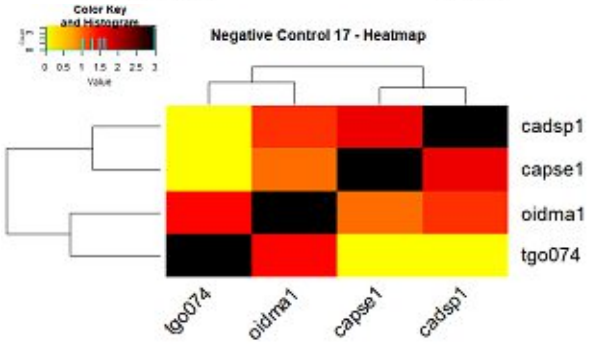


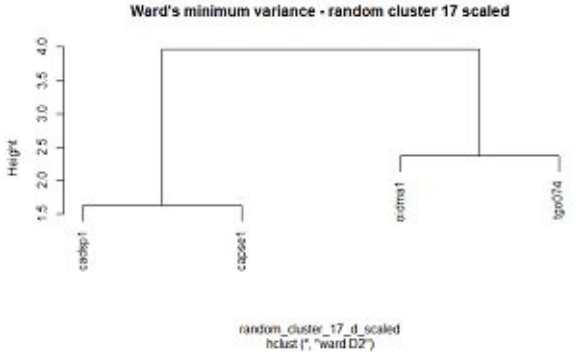


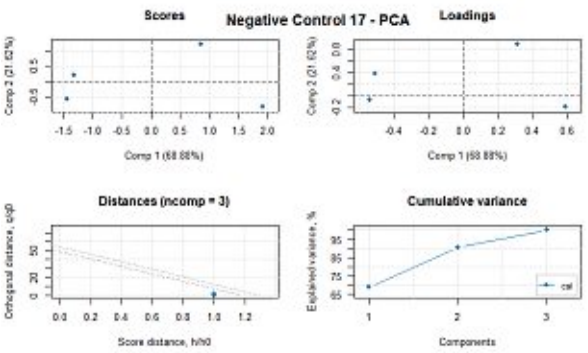


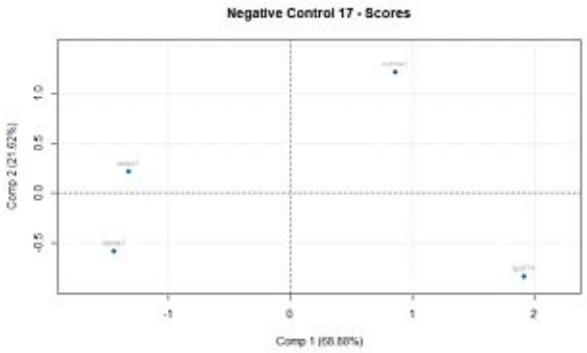


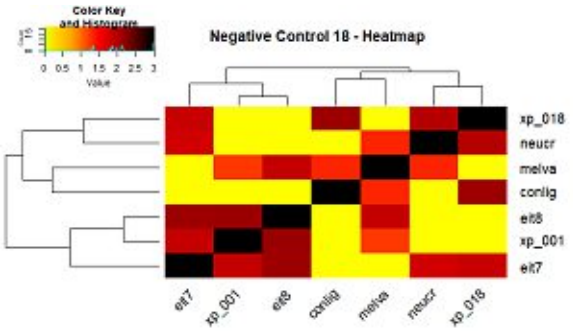


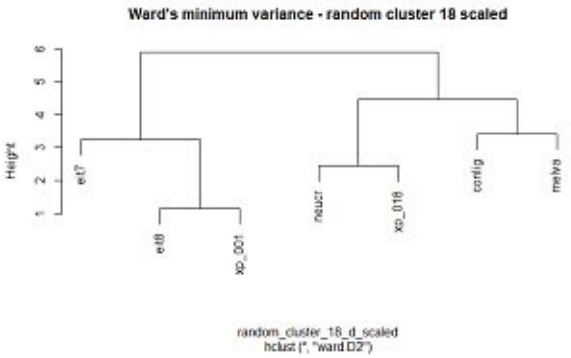


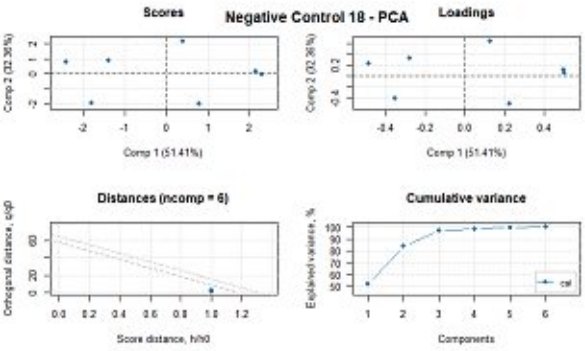


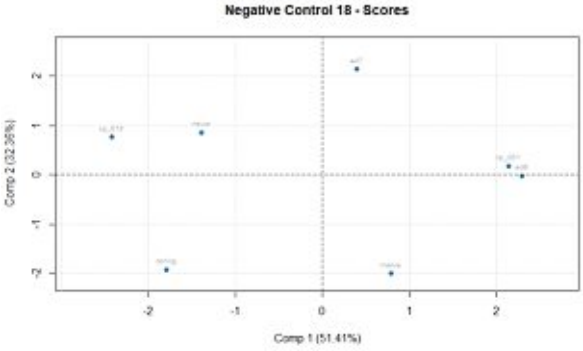


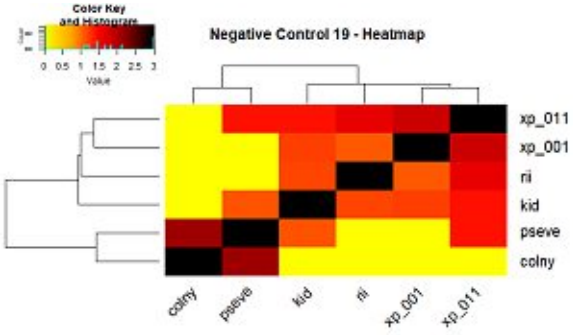


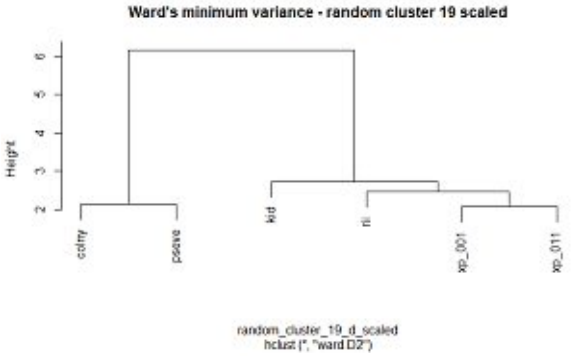


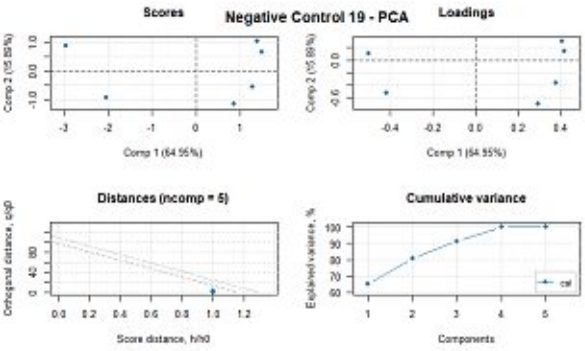


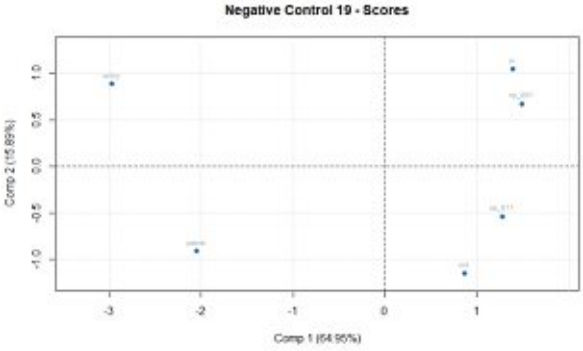


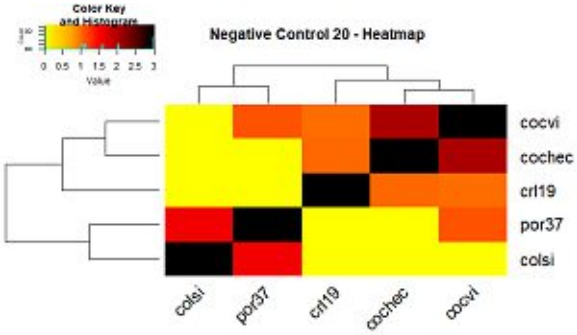


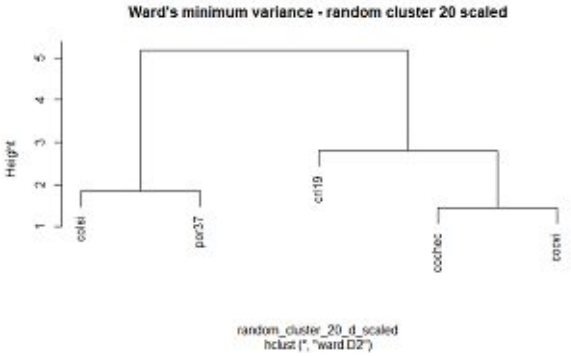


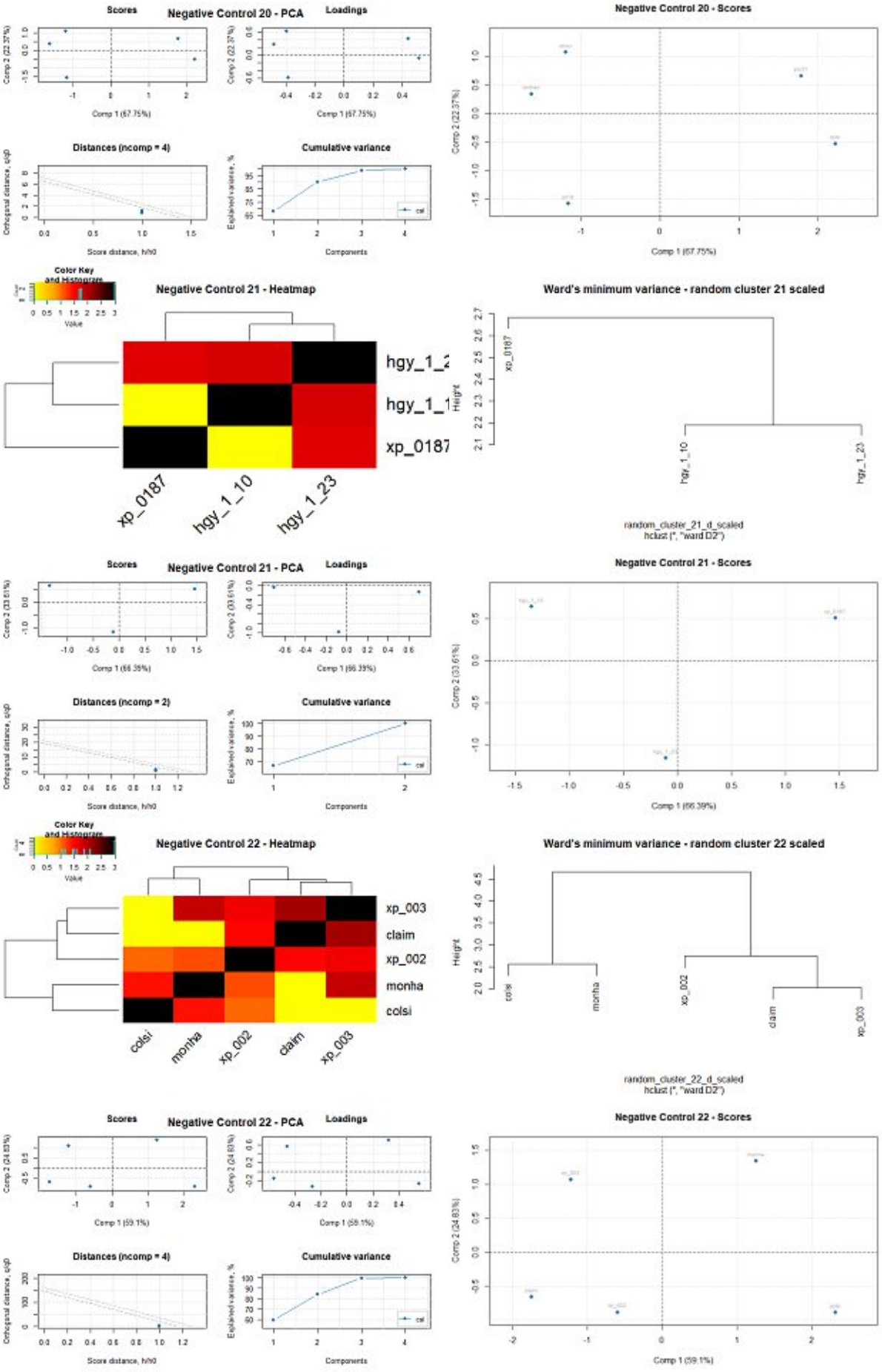


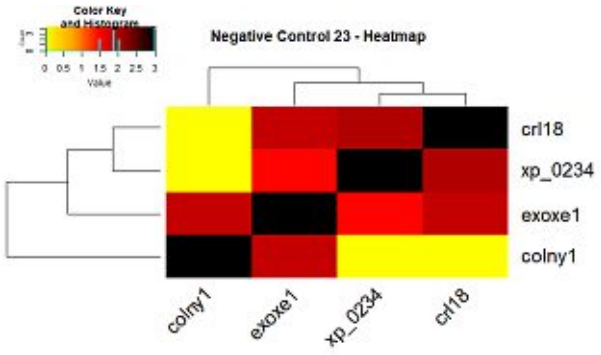


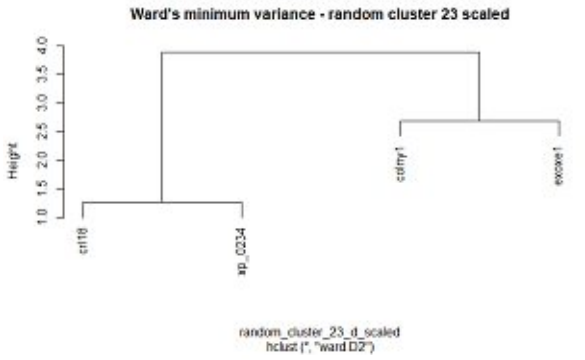


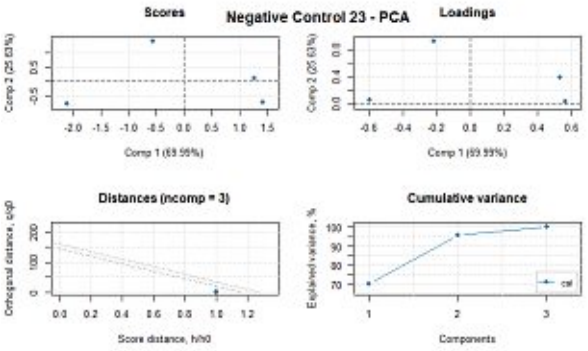


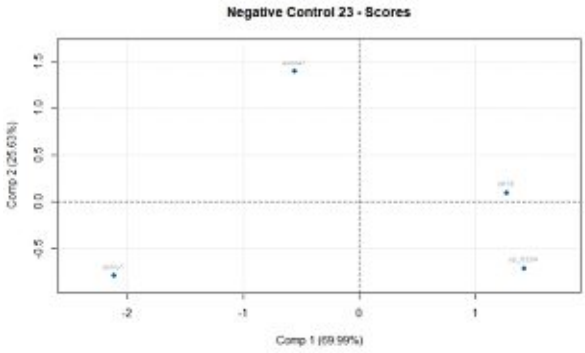


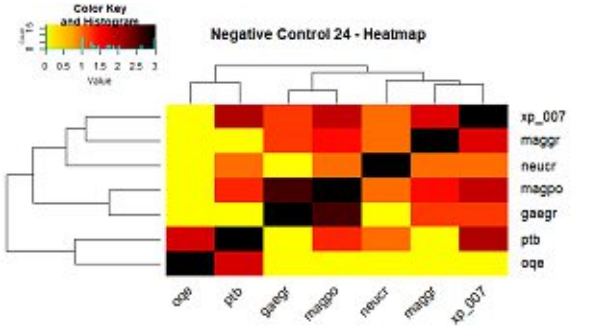


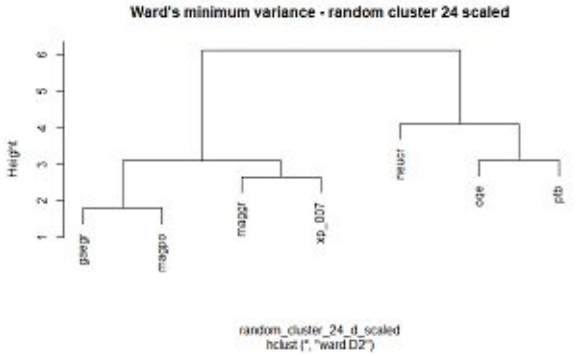


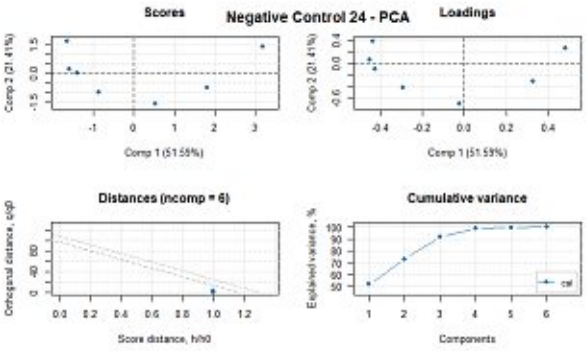


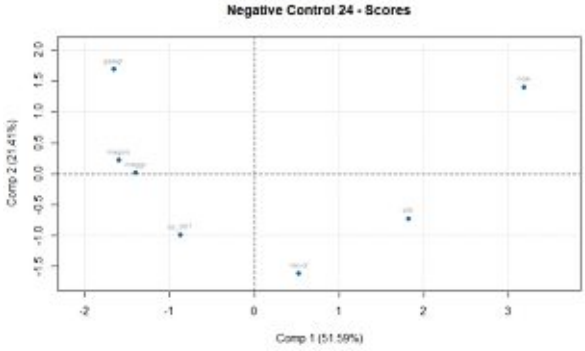


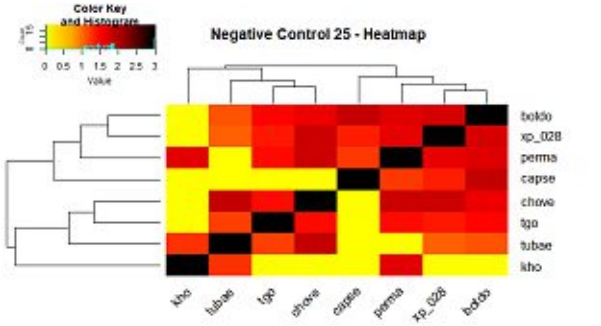


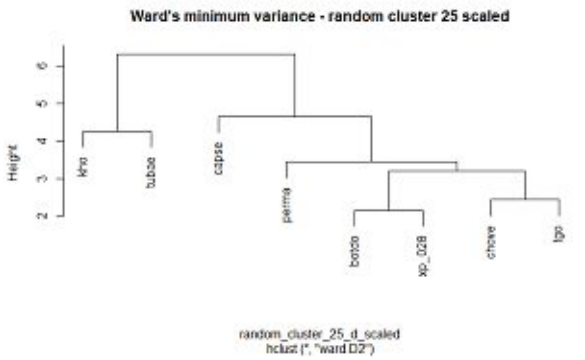


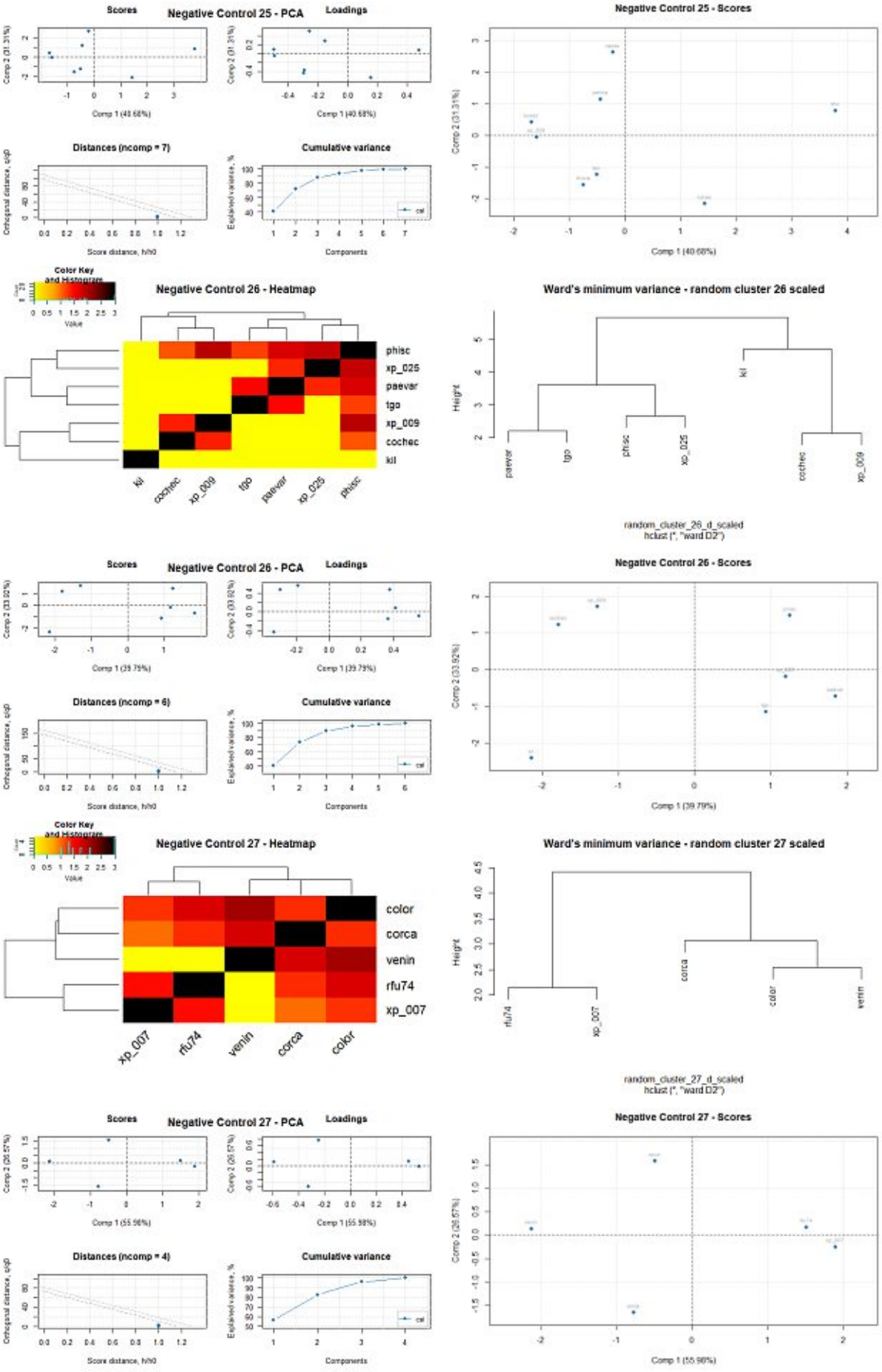


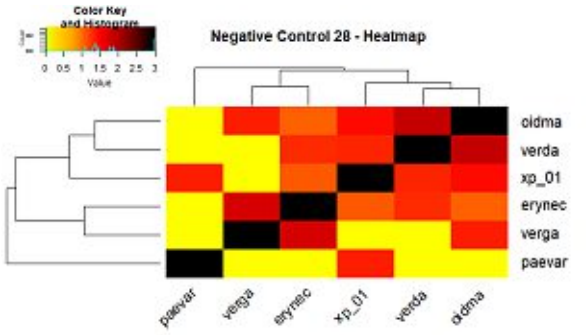


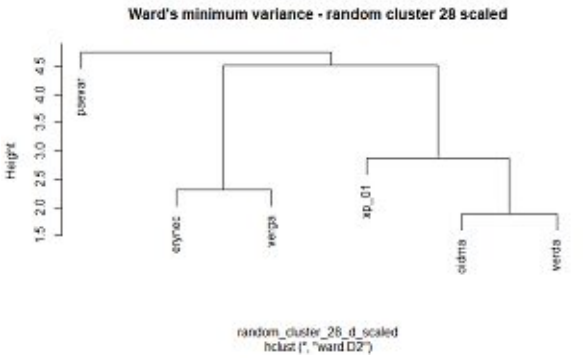


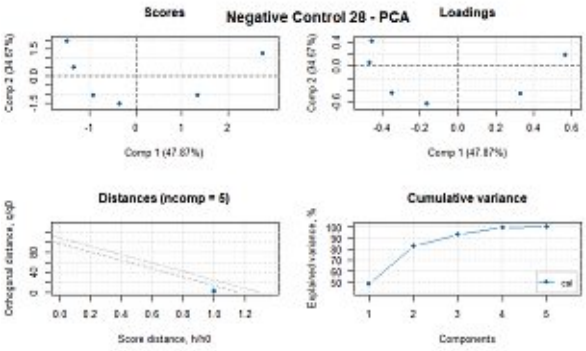


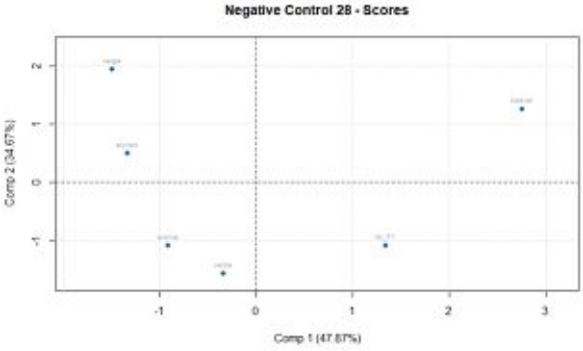


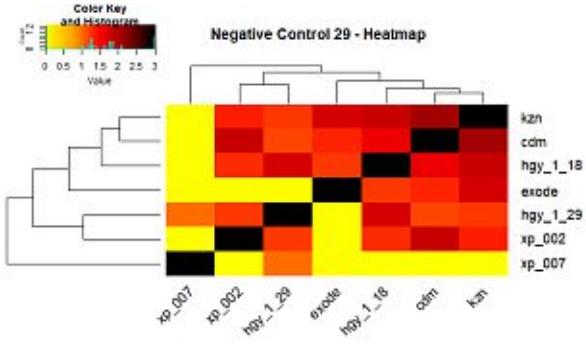


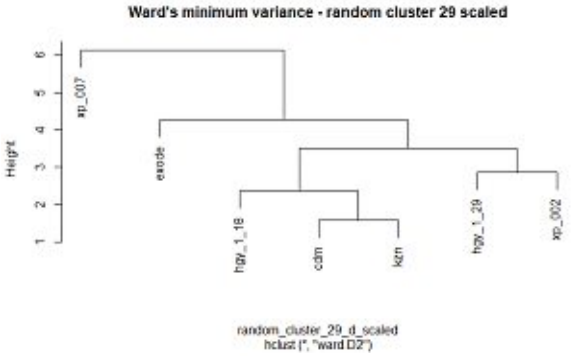


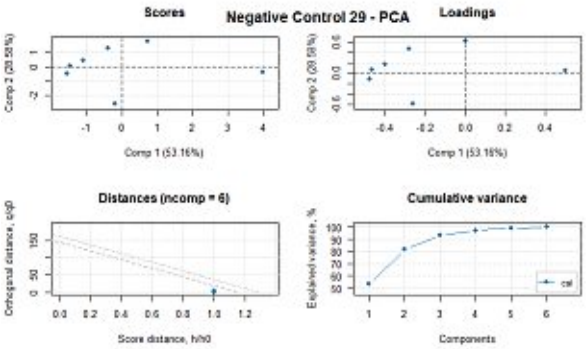


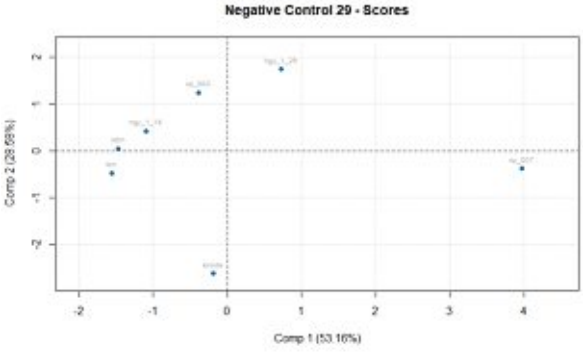


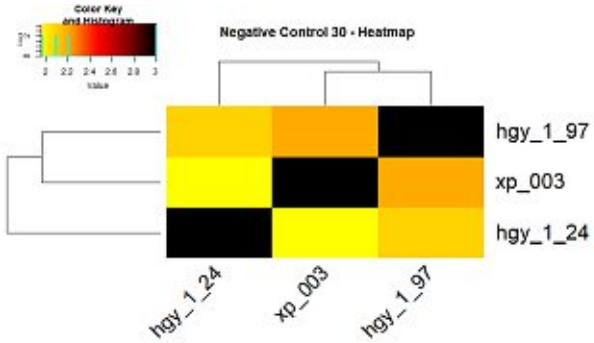


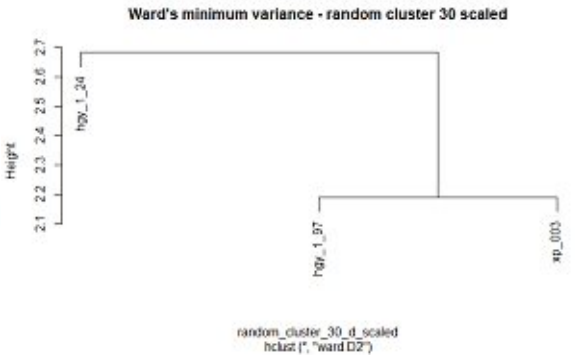


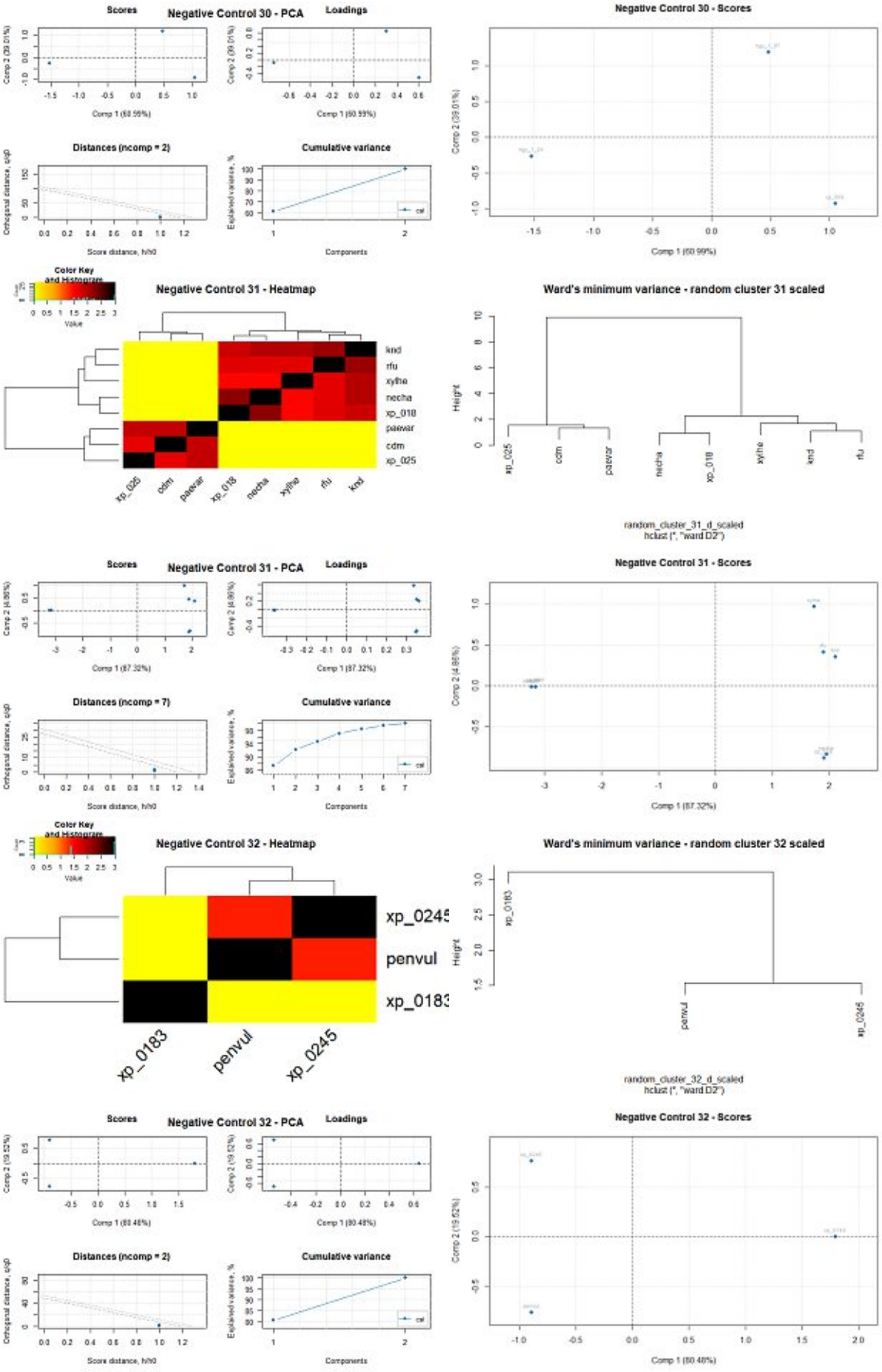


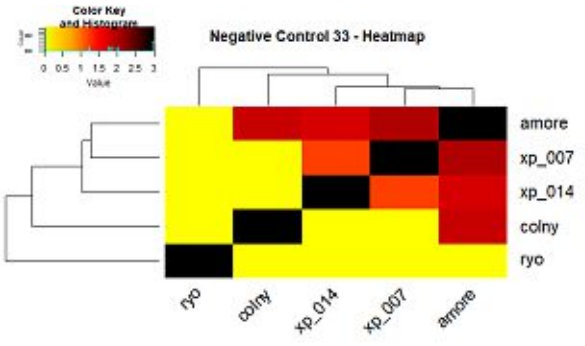


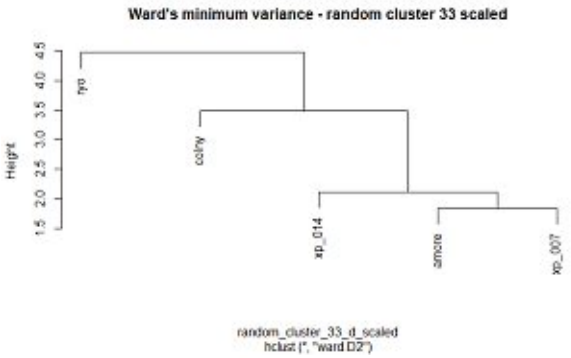


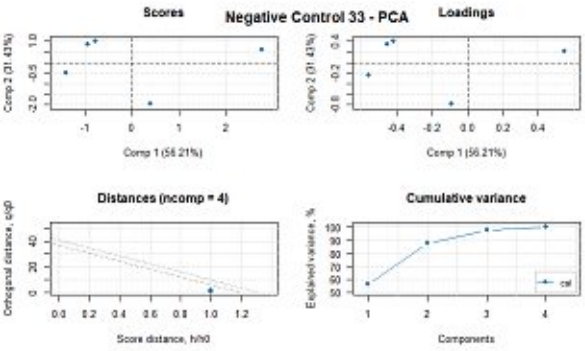


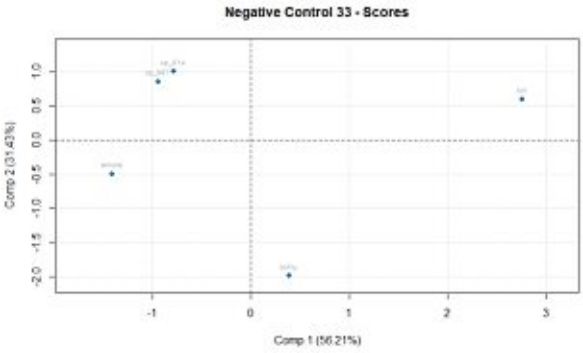


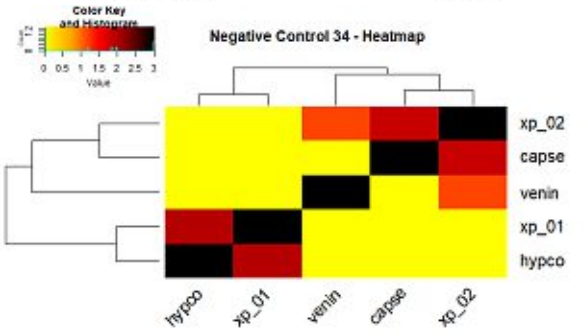


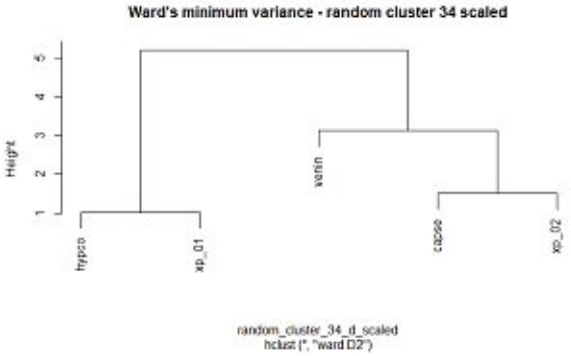


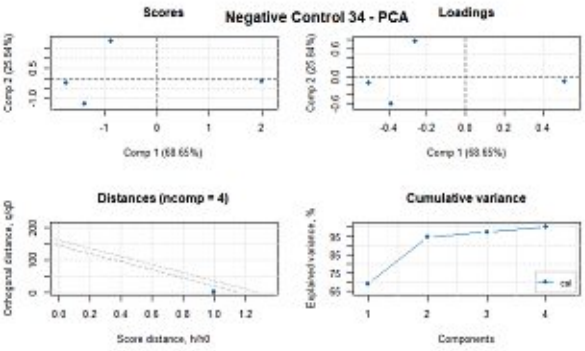


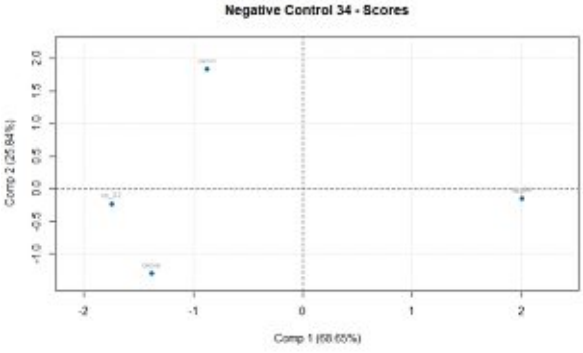


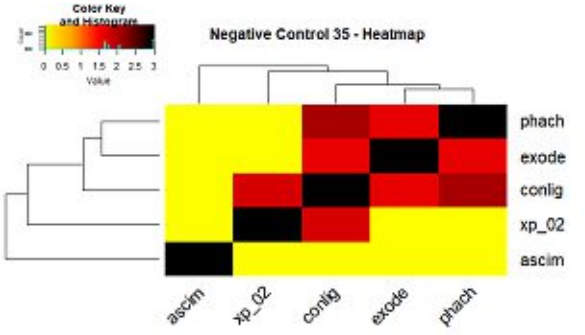


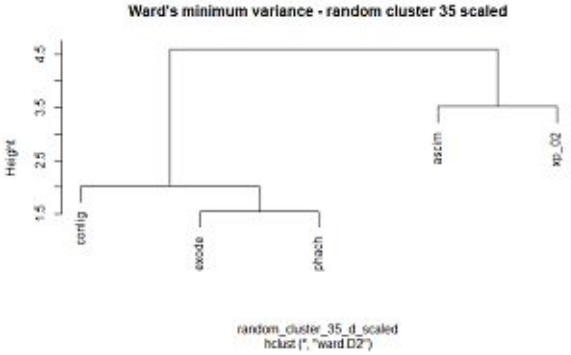


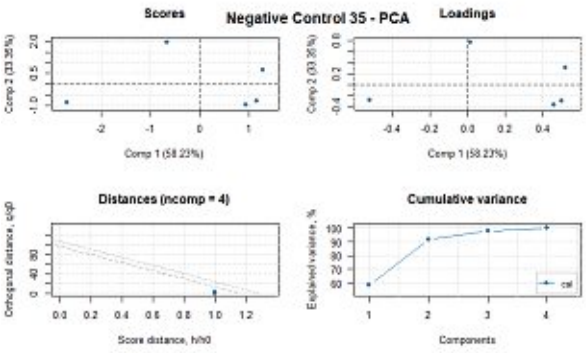


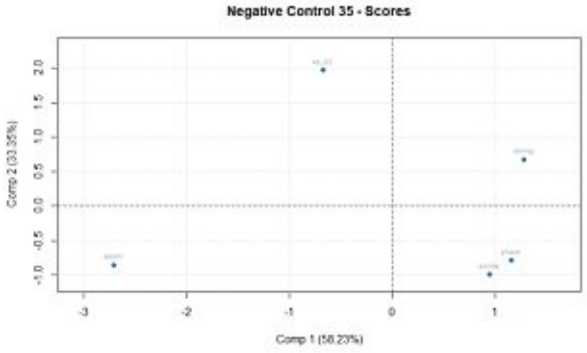


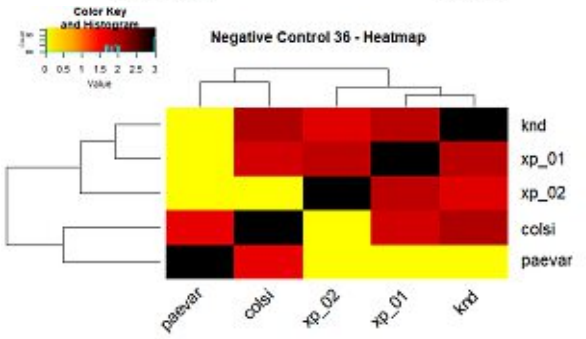


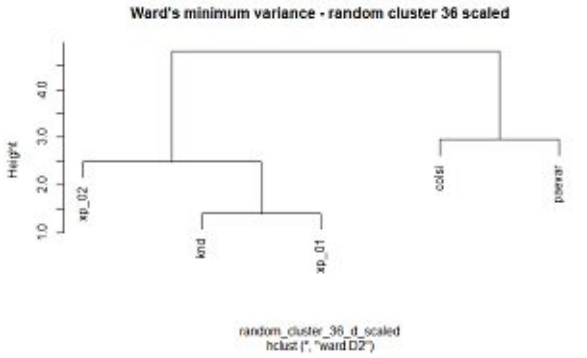


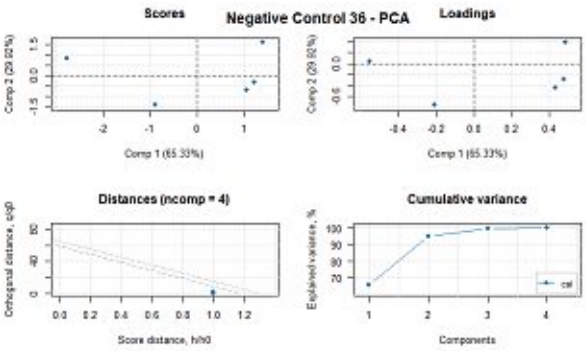


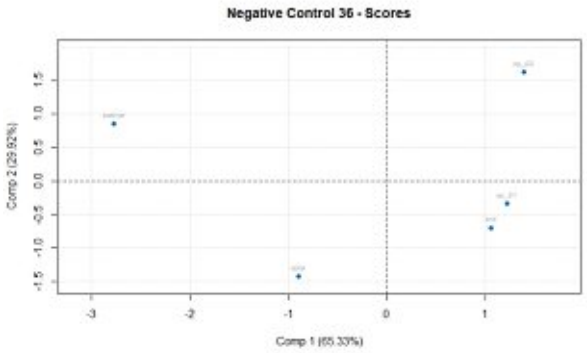


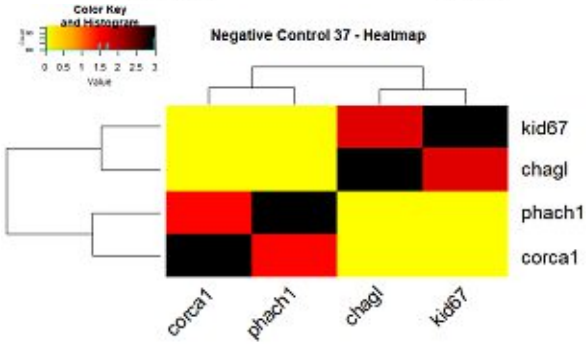


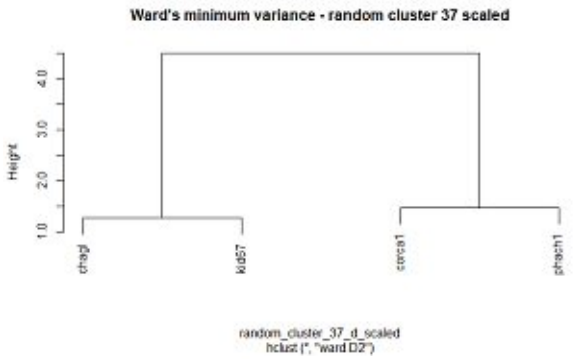


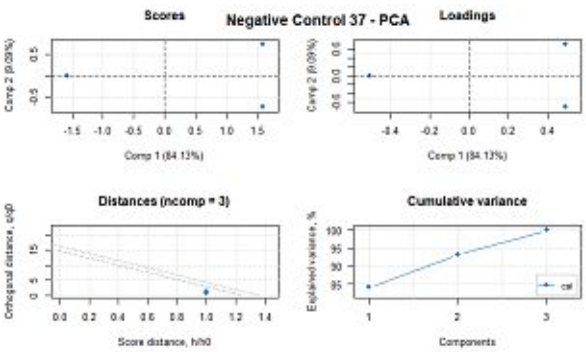


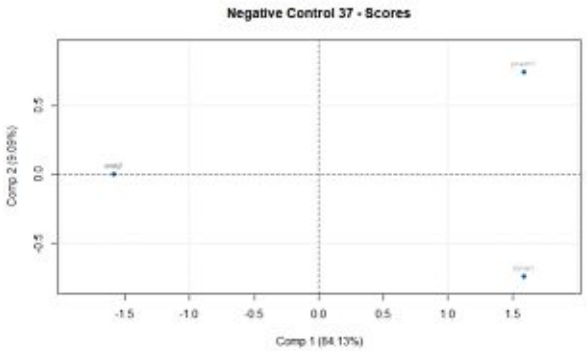


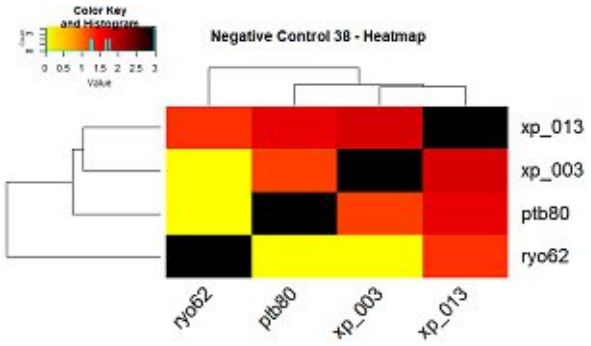


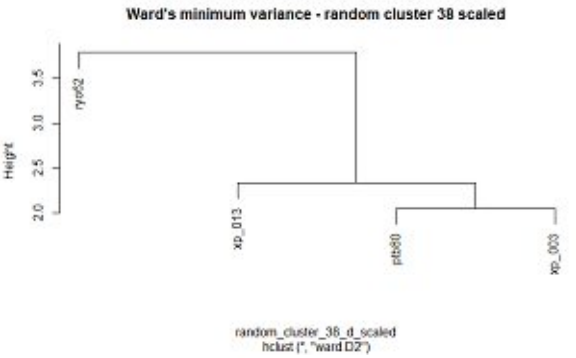


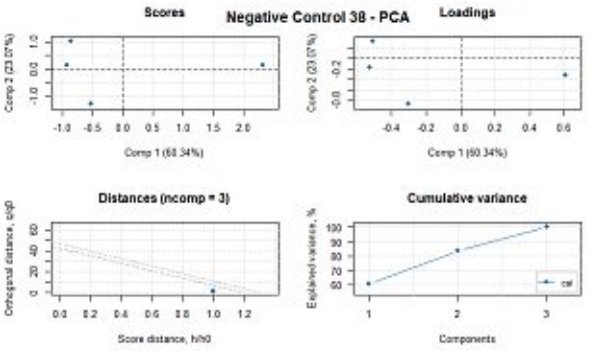


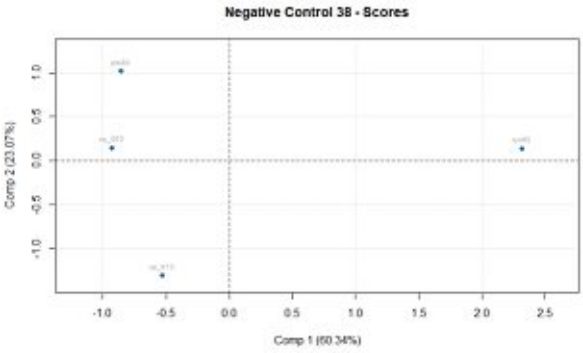


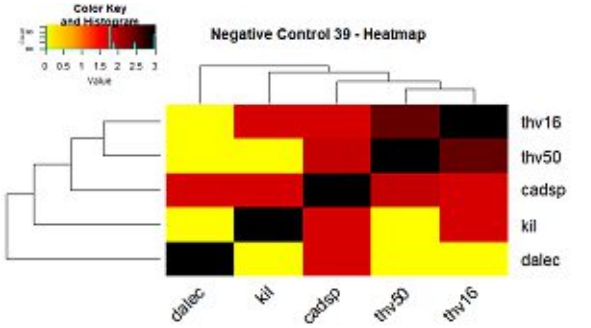


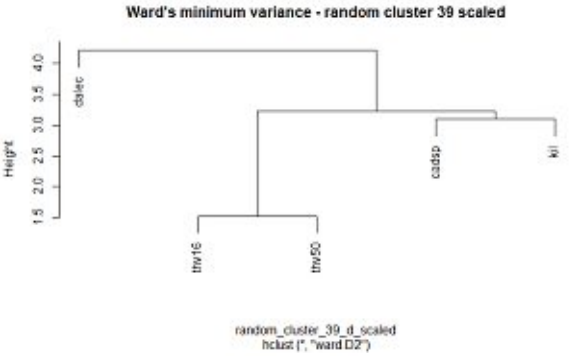


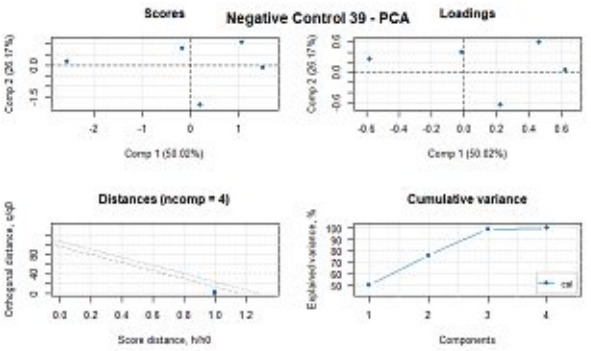


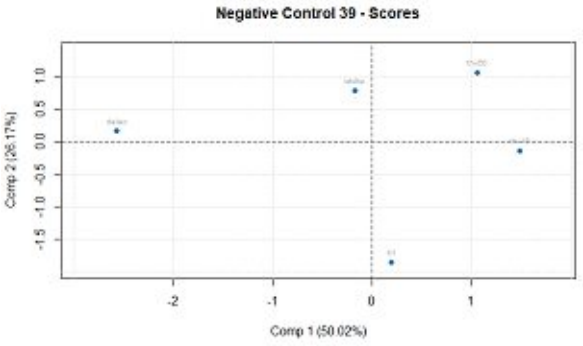


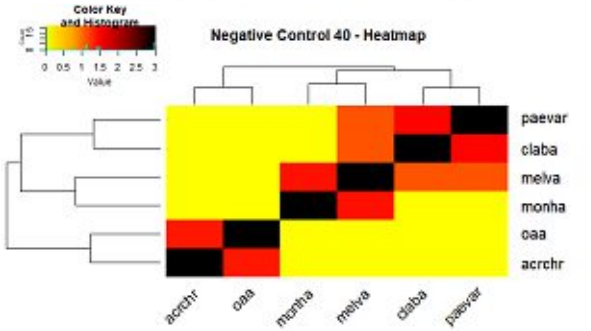


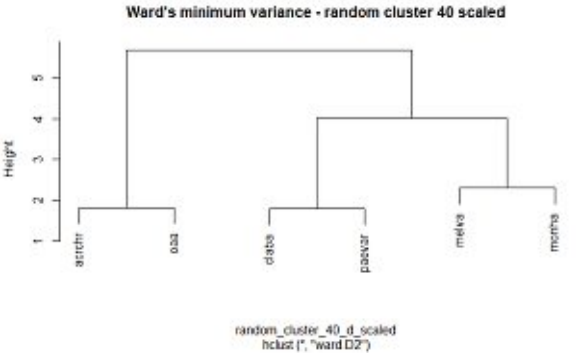


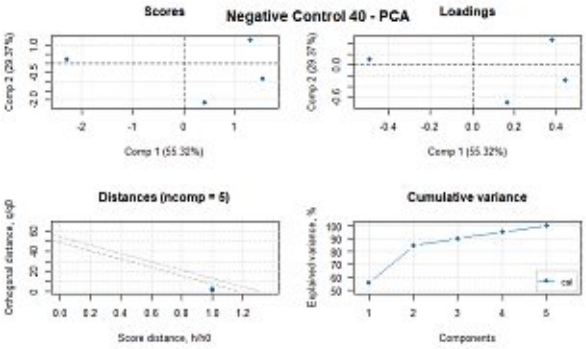


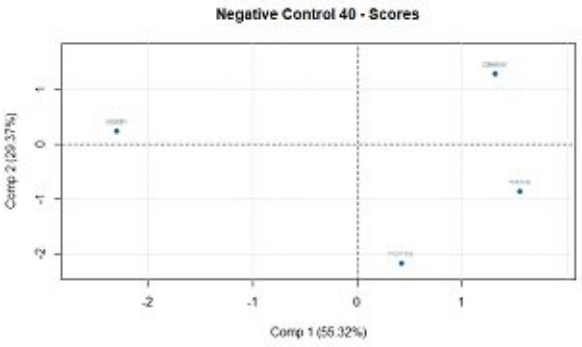


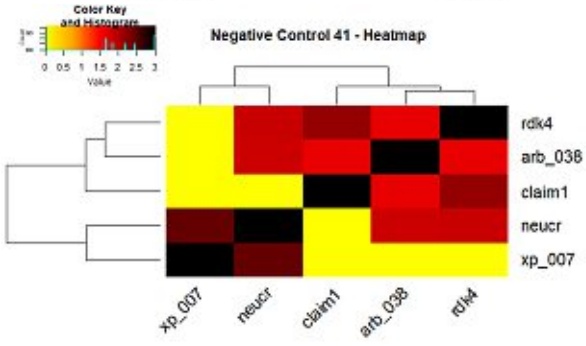


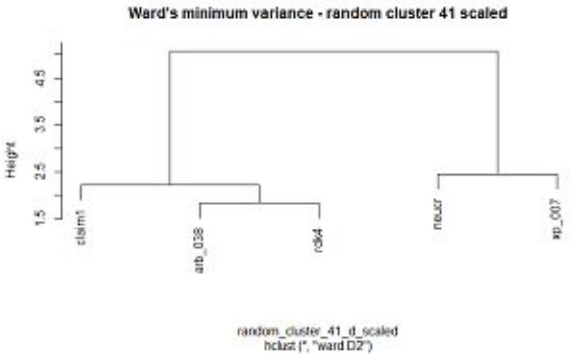


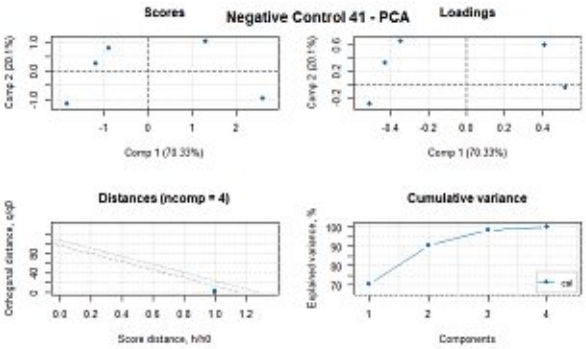


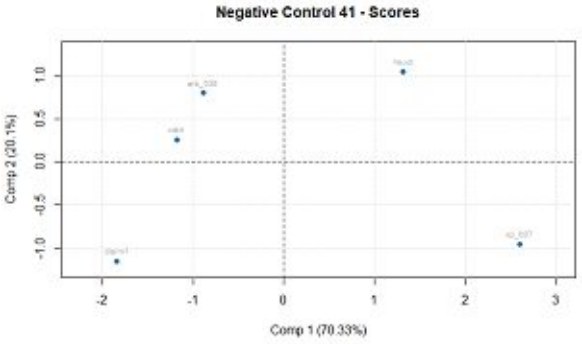


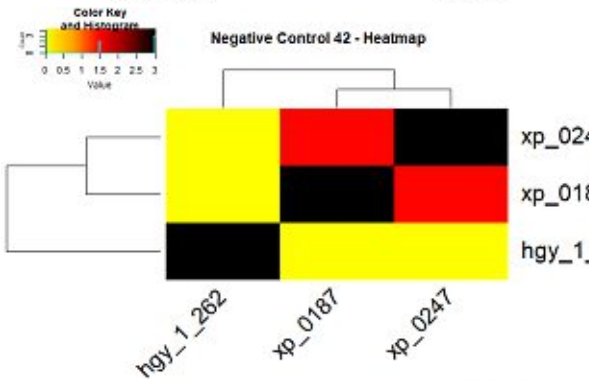


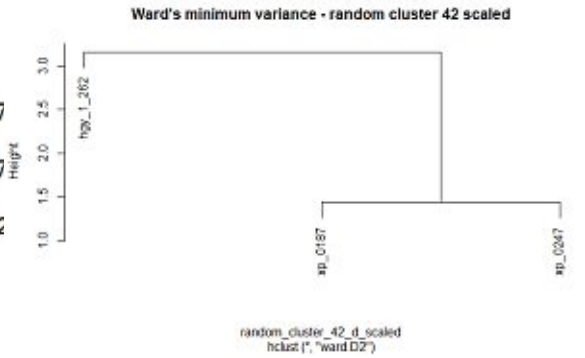


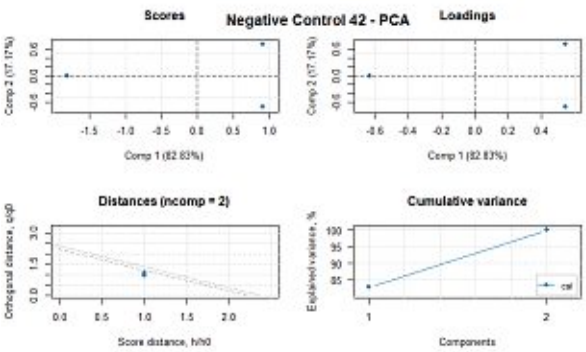


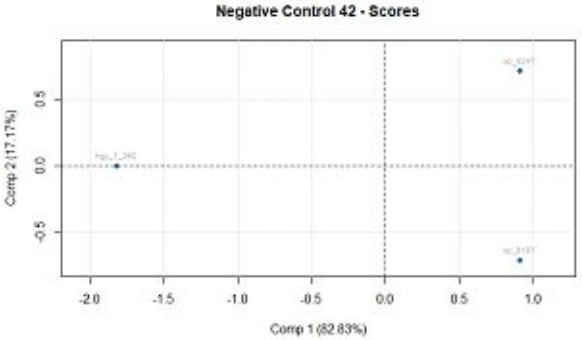


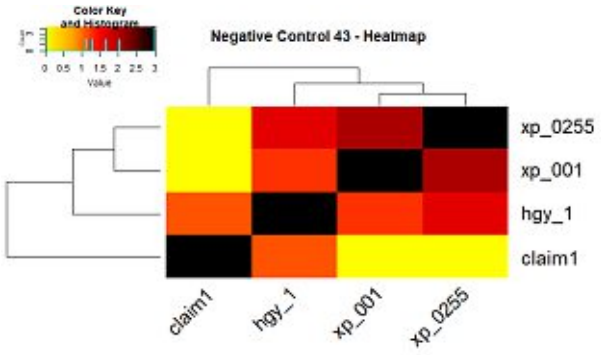


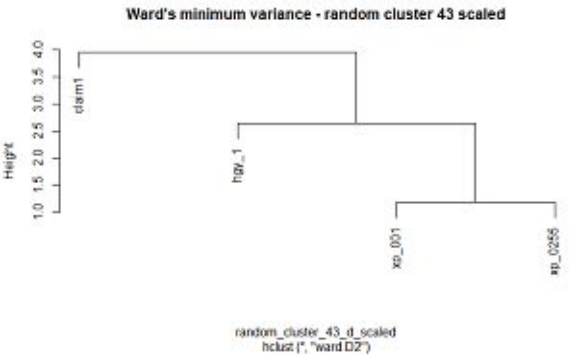


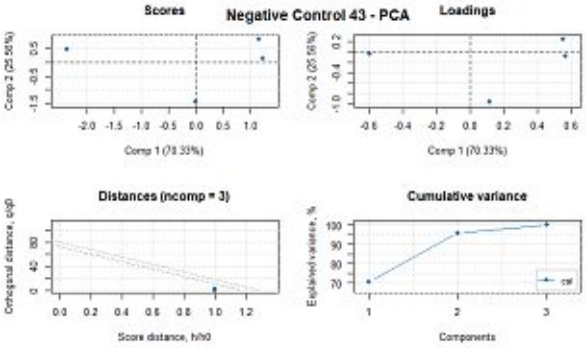


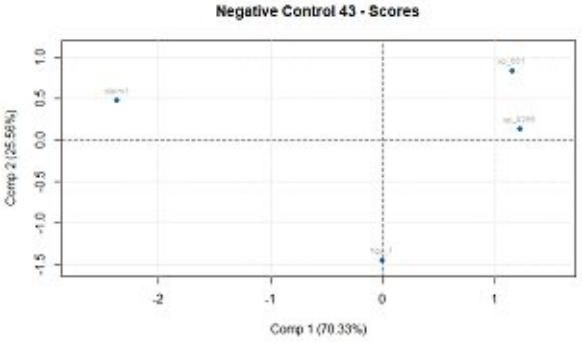


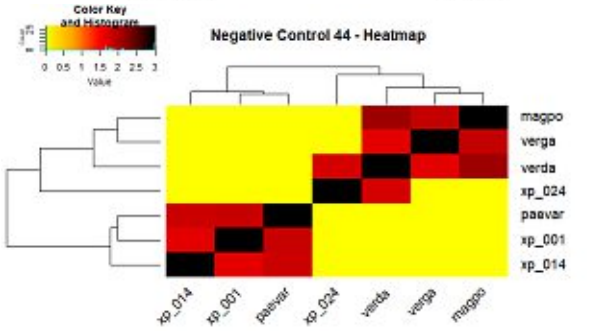


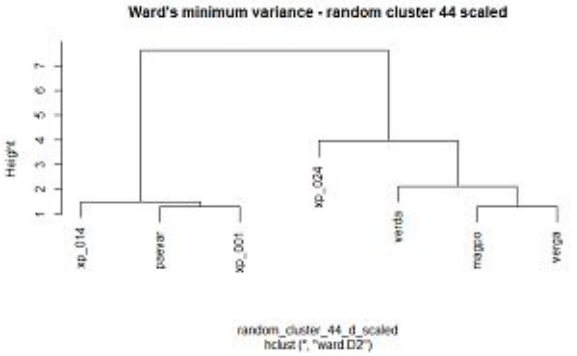


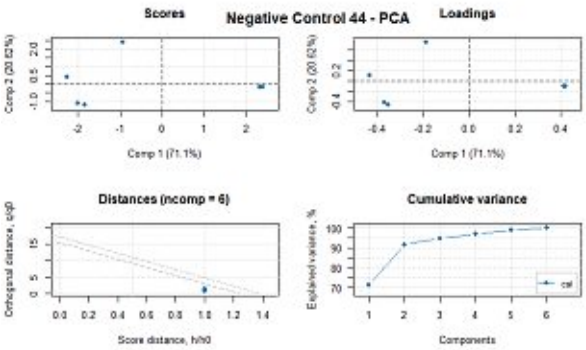


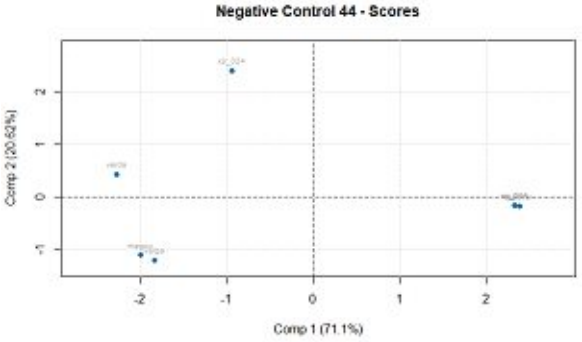


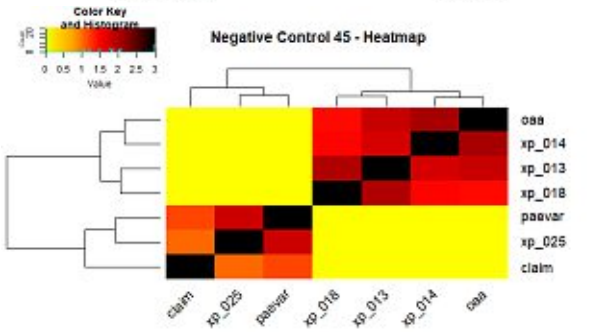


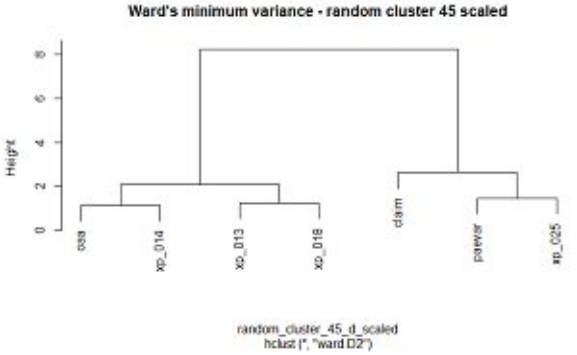


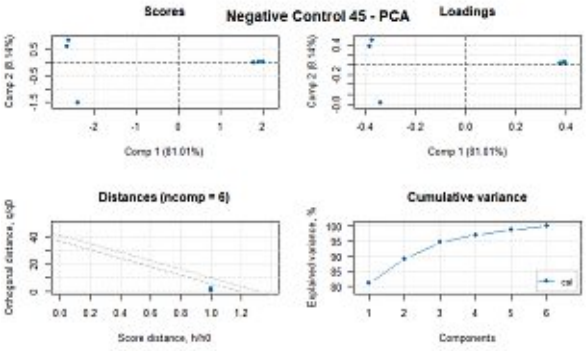


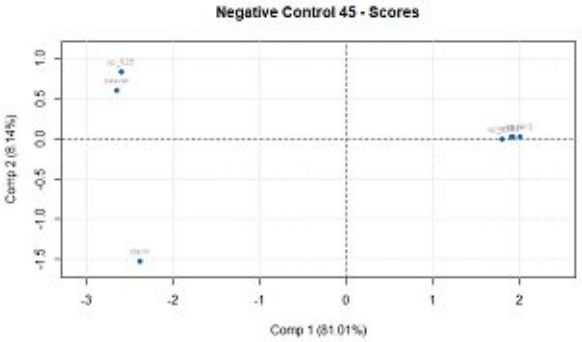


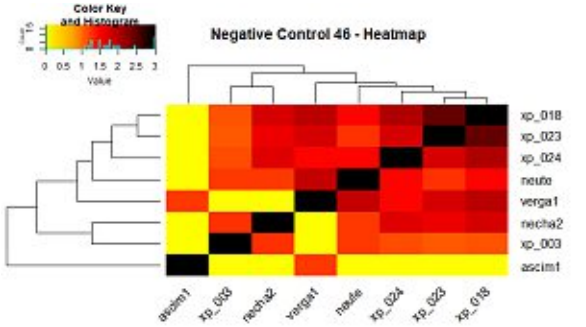


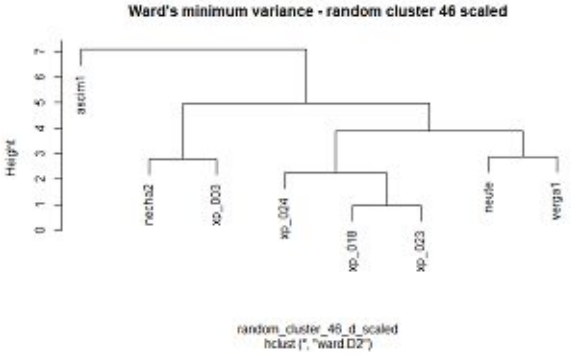


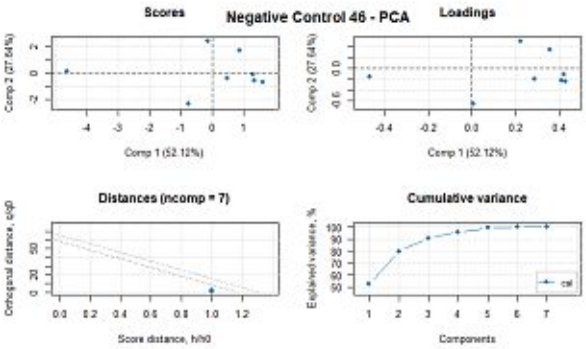


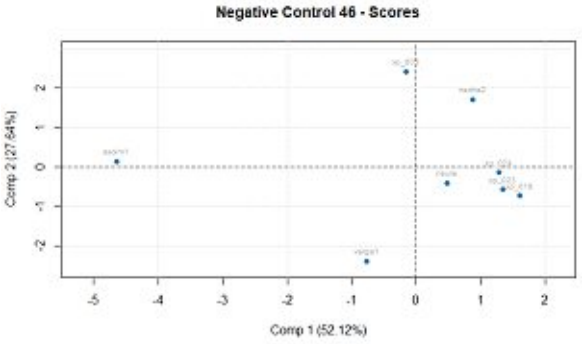


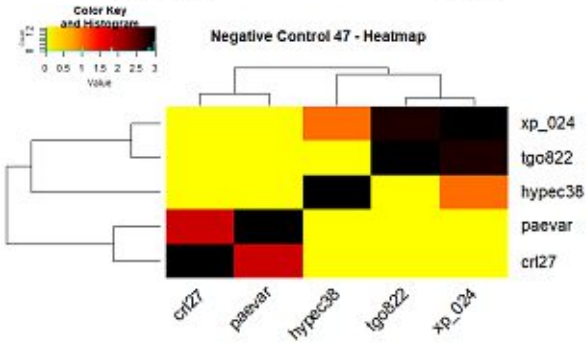


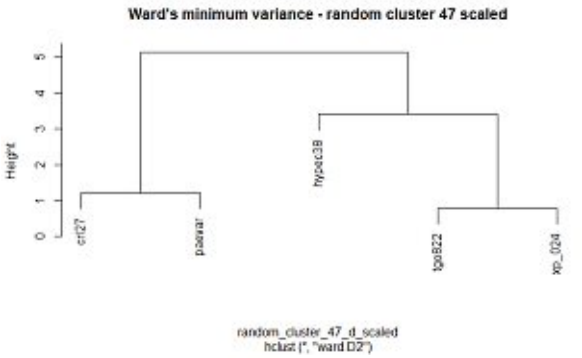


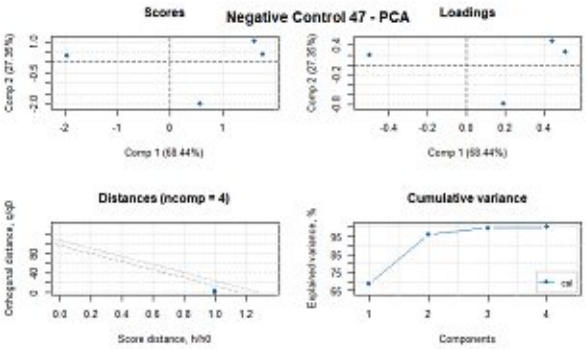


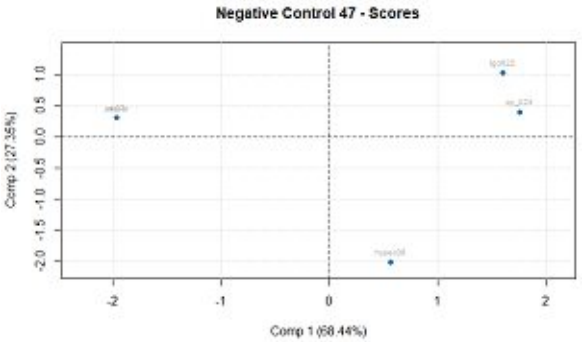


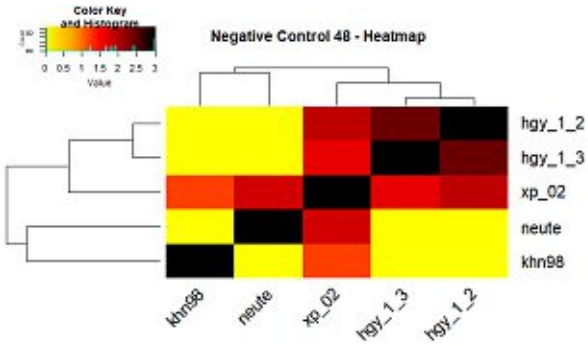


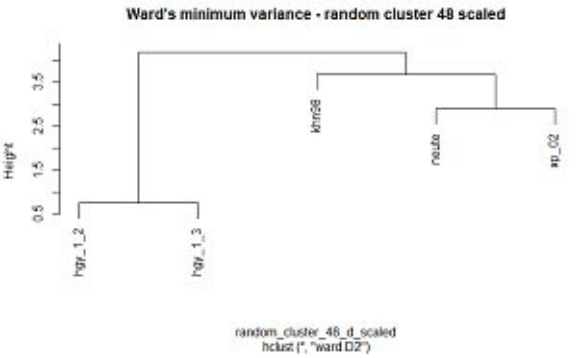


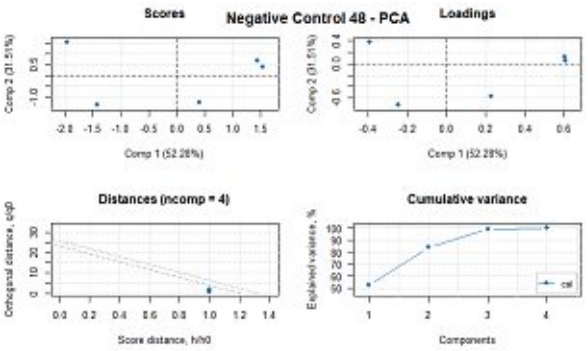


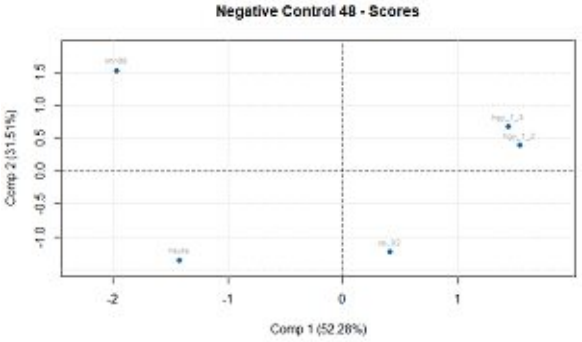


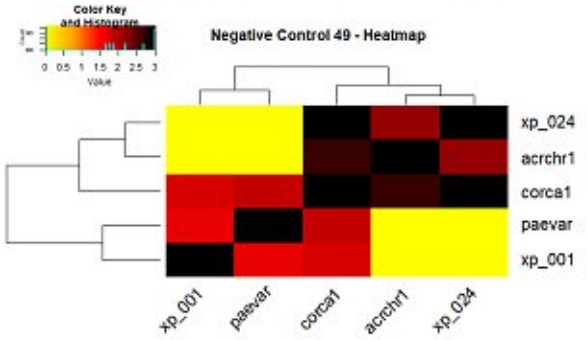


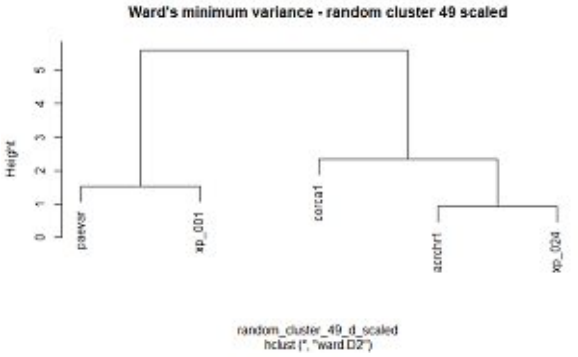


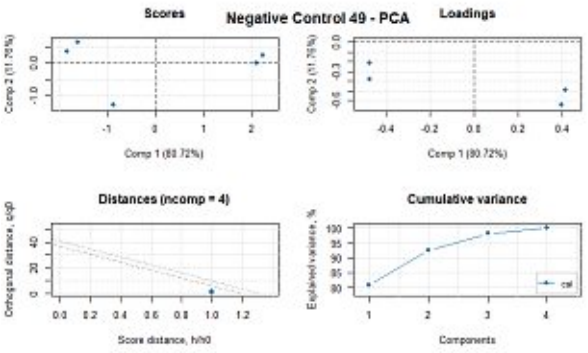


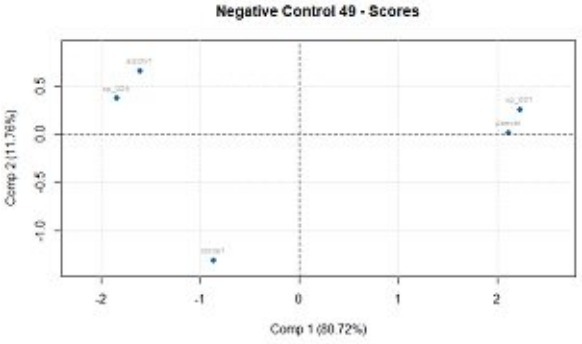


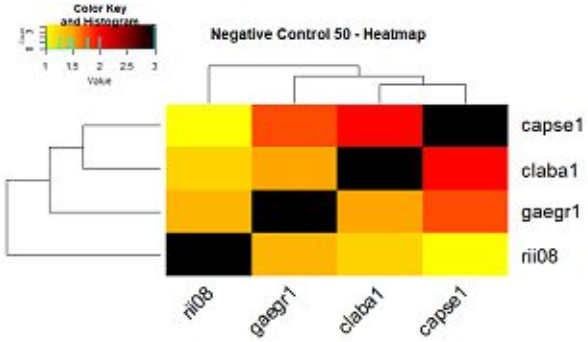


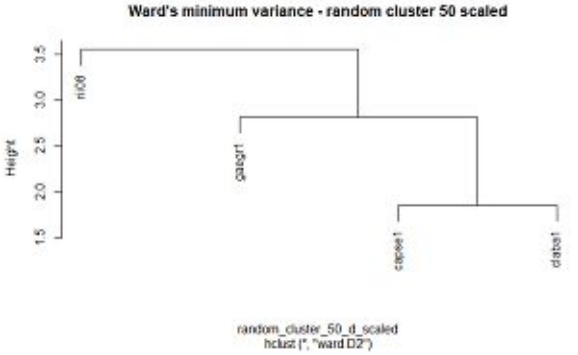


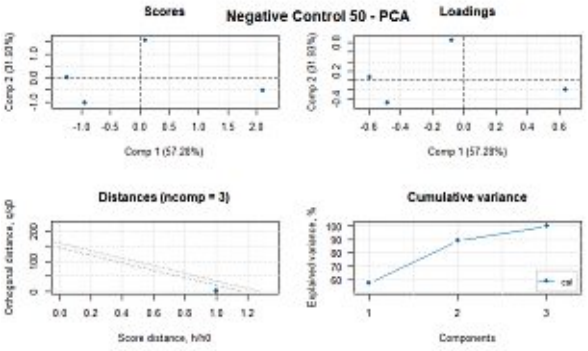


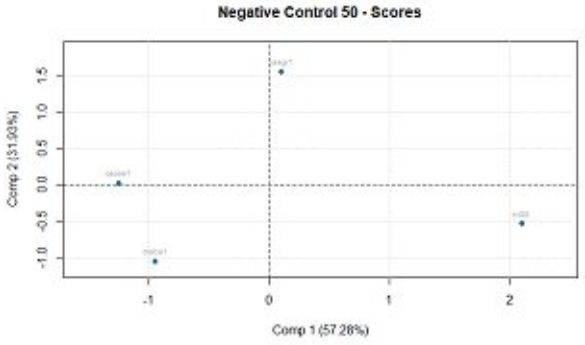


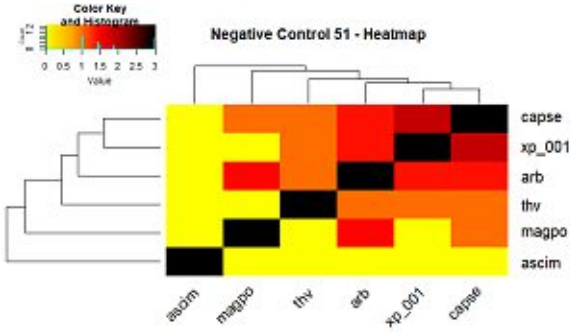


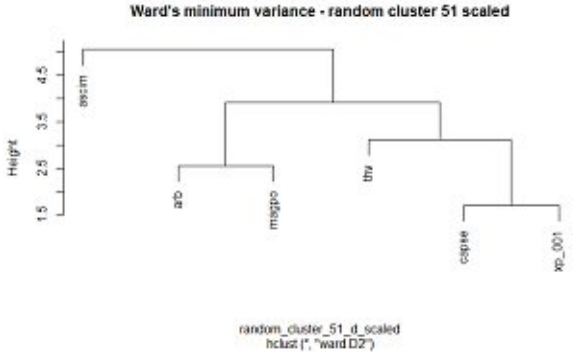


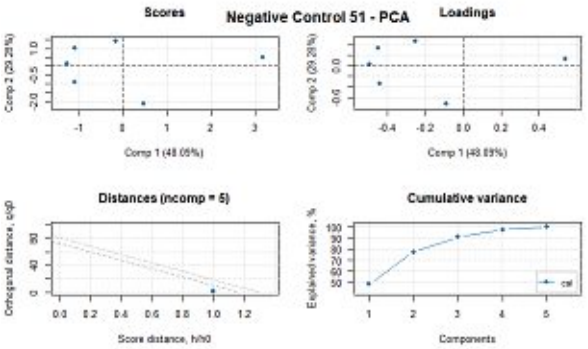


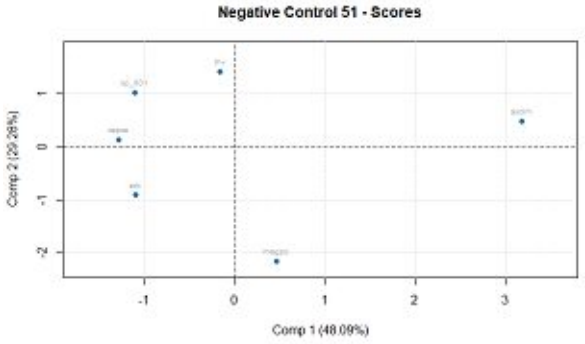


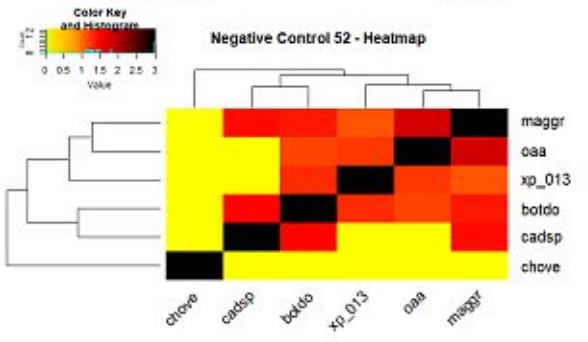


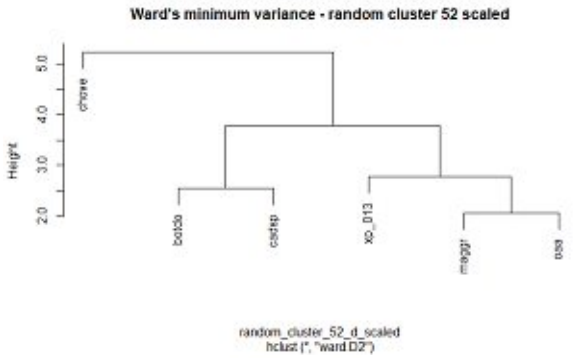


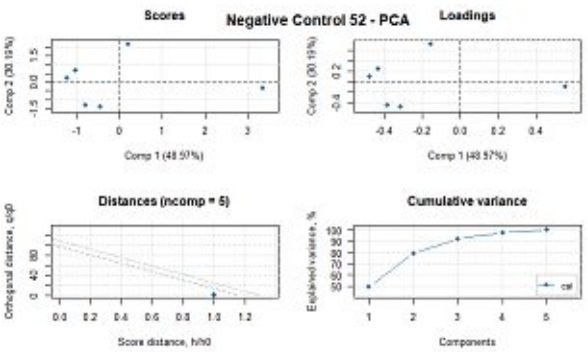


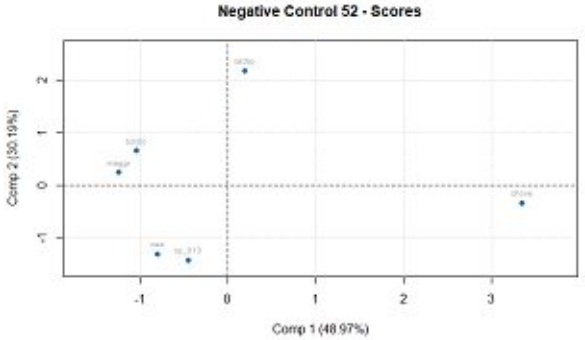


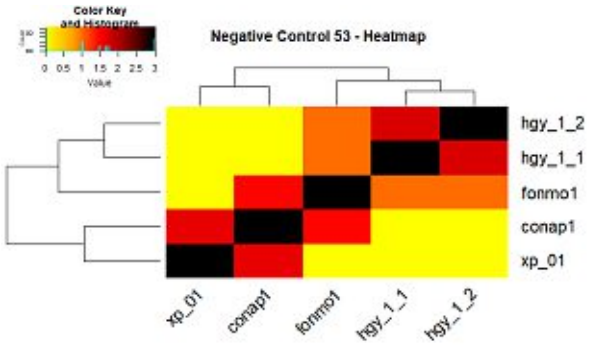


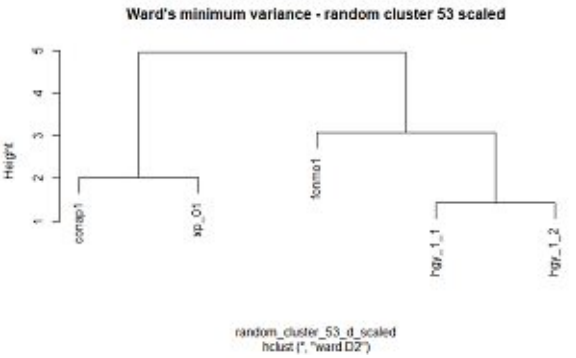


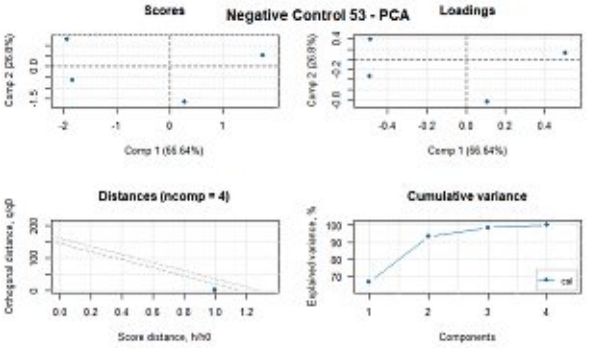


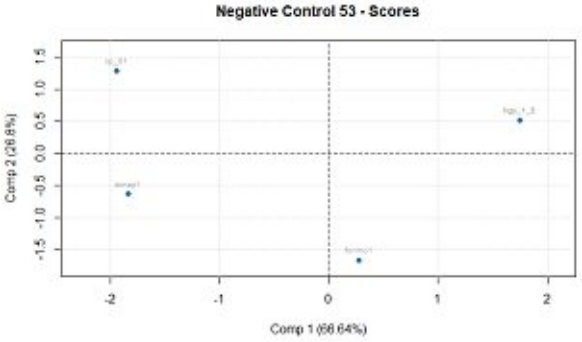


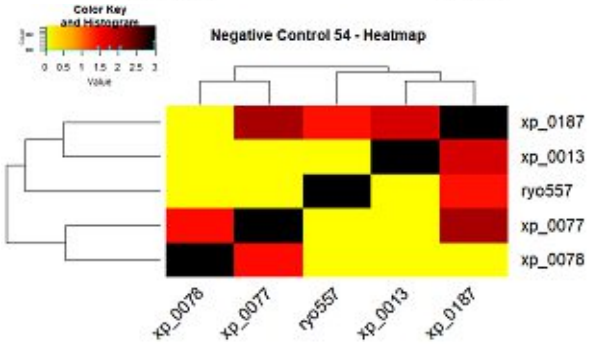


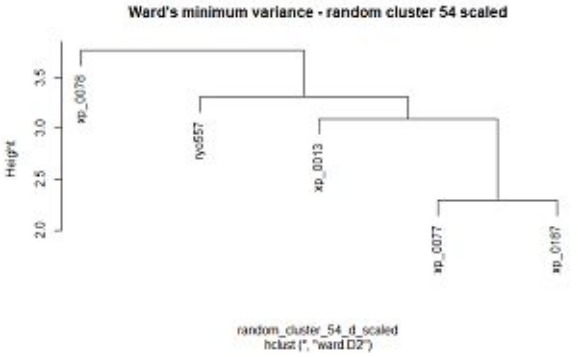


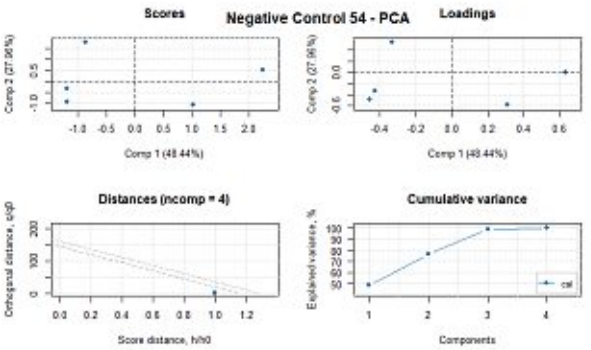


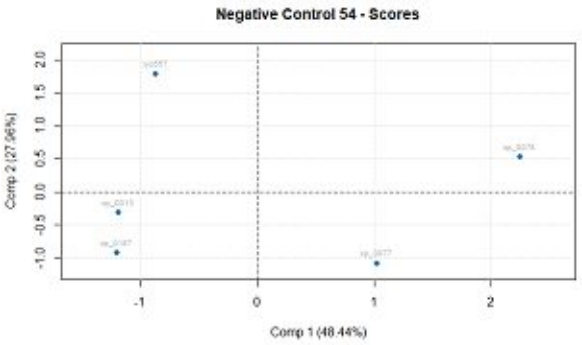


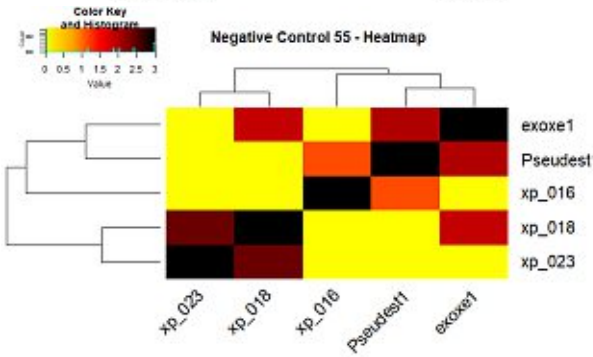


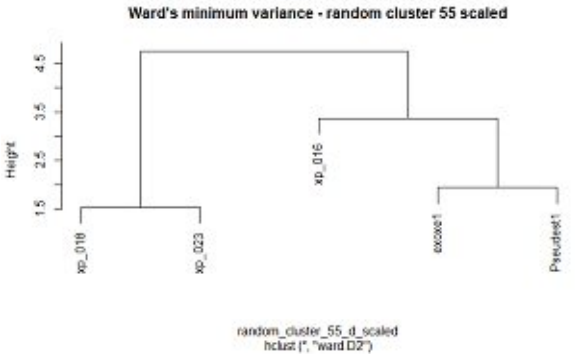


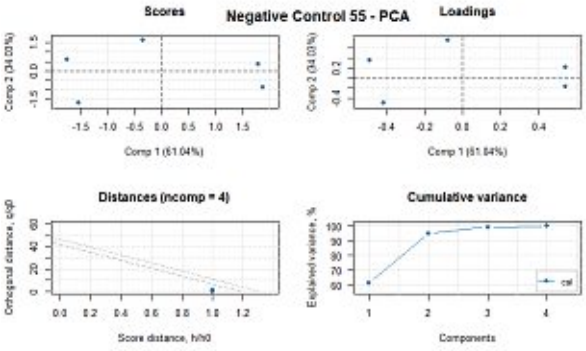


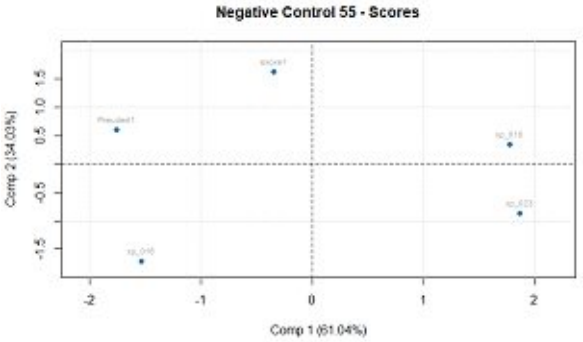


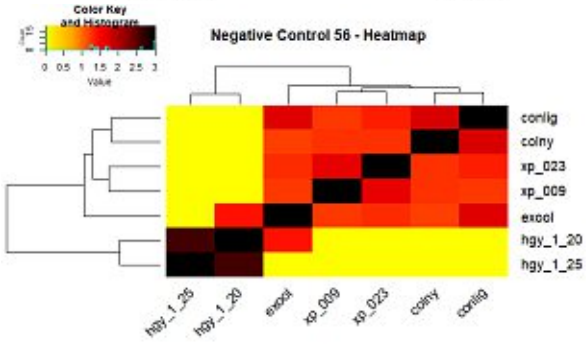


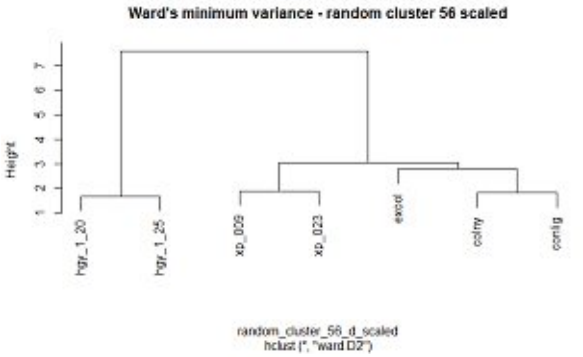


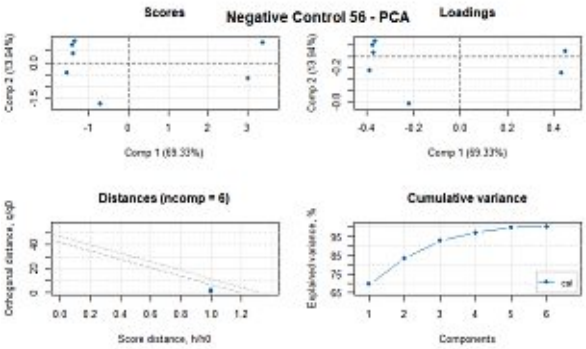


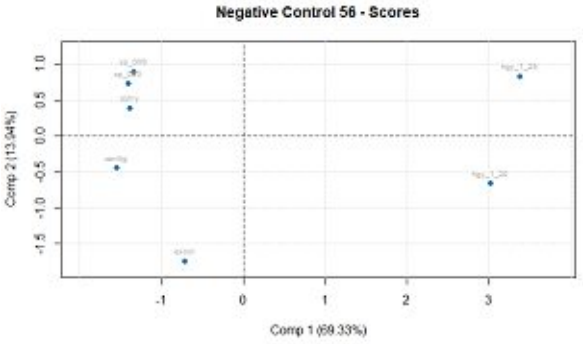


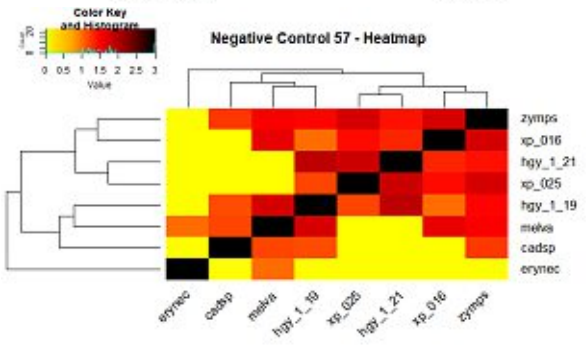


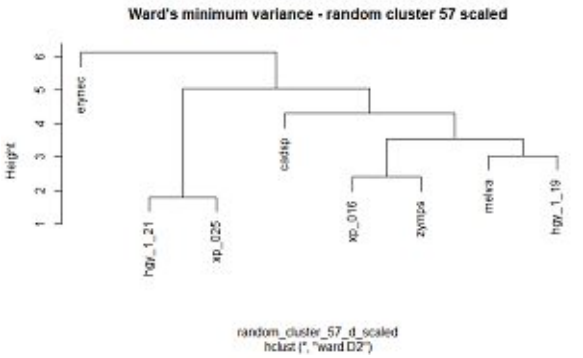


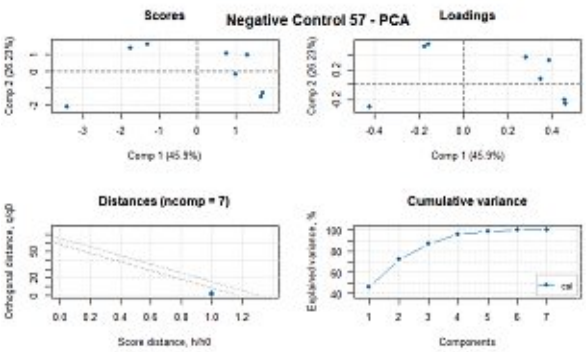


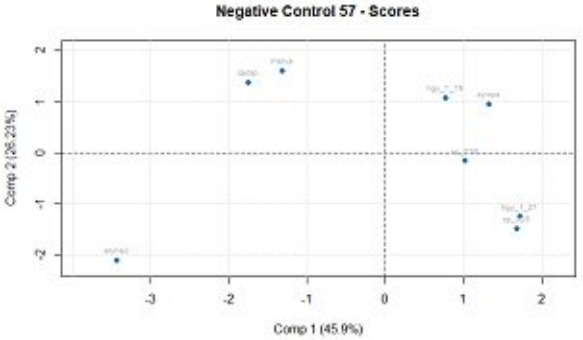


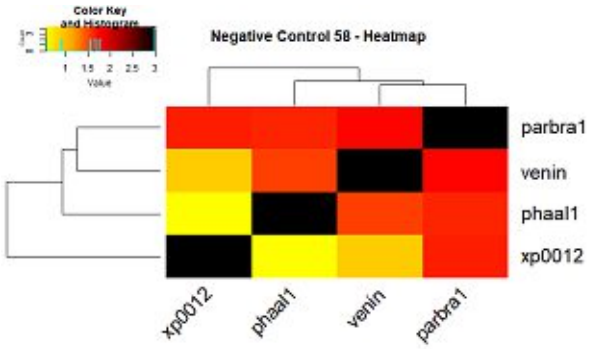


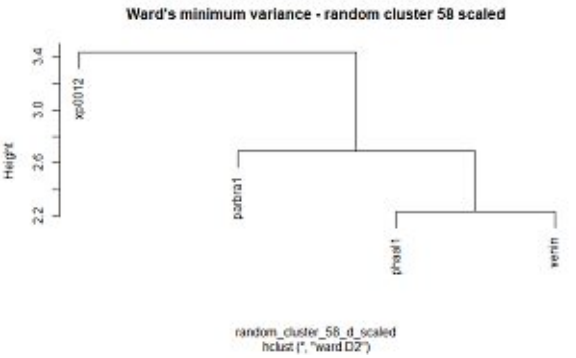


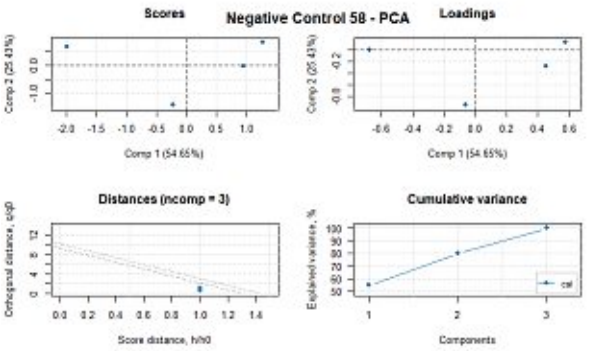


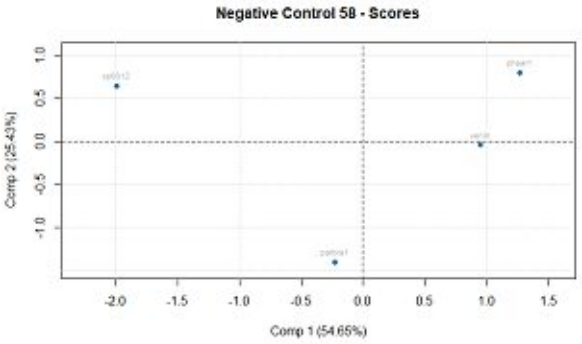


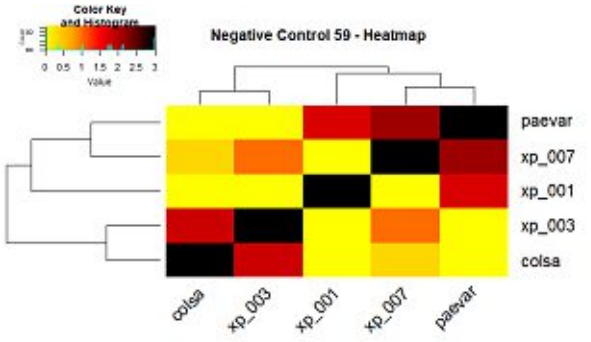


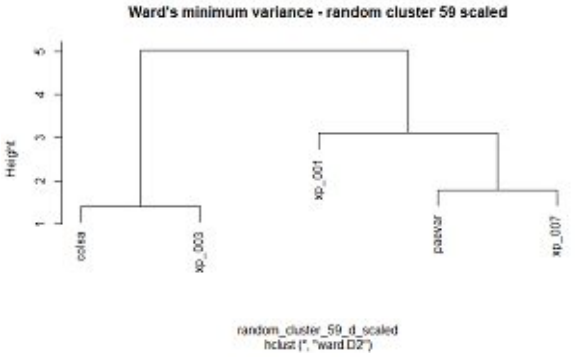


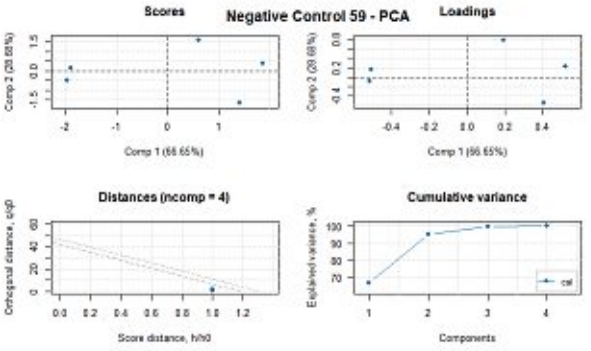


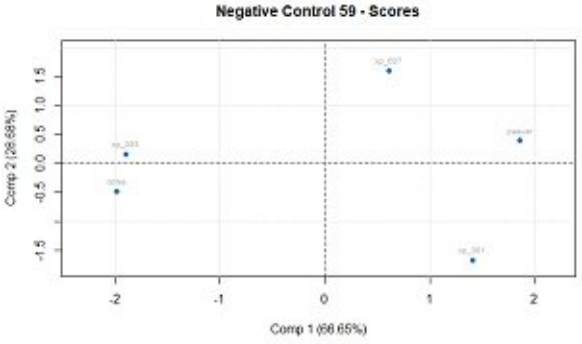












Supplement 9.63: Heatmaps, Dendrograms, and principal component analysis (PCA) of Positive control 30 which were automatically calculated by the TreeKO approach of FunOrder based on the manually evaluated data.