Die approbierte Originalversion dieser Dissertation ist in der Hauptbibliothek der Technischen Universität Wien aufgestellt und zugänglich.
http://www.ub.tuwien.ac.at

The approved original version of this thesis is available at the main library of the Vienna University of Technology.
http://www.ub.tuwien.ac.at/eng

Diese Dissertation haben begutachtet:

. . . .

DISSERTATION

Analysis of most complete biological datasets

Graph algorithms, combinatorics, GWAS, dimension reduction and classification in omics data

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

Privatdoz. Mag.rer.nat. Dr.rer.nat. Irina Druzhinina

E166 - Institut f. Verfahrenstechnik, Umwelttechnik und Techn. Biowissenschaften
eingereicht an der Technischen Universität Wien
Fakultät für Technische Chemie

von

Mag. Ing. Alexander Platzer

Matrikelnummer: 0101535
Kornhäuselgasse 3/2/15, 1200 Wien

Wien, am 30.09.2014 . . . .

Long title of the thesis

Graph algorithms, combinatorics, GWAS, dimension reduction and classification in populations of full genome sequenced data, full genome microarrays and quite comprehensive metabolic profiles.

Summary

Already a few years ago reports appeared in popular computer science magazines that molecular biology data is exponentially growing [1]. More recently, there have been concerns that NGS/sequencing data is growing faster than computer storage capacities, despite the exponential growth of this storage [2, 3]. This thesis deals with these large amounts of data, therefore this work is clearly located in the field of bioinformatics.

Additionally, the classic paradigm of 'one gene/protein – one function or phenotype' has shifted from being the main approach to just one of several options, with most of these combining large amounts of information to arrive at a conclusion [4-7]. Several terms exist for this: systems biology, the –omics field, integrative analysis, and a few more.

The present work makes a broad sweep of the field, from whole genome microarrays, through metabolomics, to sequencing data, with sidetracks into the complexities of a combinatorial problem, dimension reduction, and transposons. The steady goal is to gain general insights into the full data collection and/or to indicate other promising procedures.

The work on differentially expressed genes in tumors started with the diploma thesis of the applicant and was continued in a later article. The main result is that, although the overlaps of differential expressed genes in tumors from the same tumor type seem random, these gene lists share elements on a protein interaction network level.

The observation of metabolite levels from tissues is still an immature field; currently, several 100 different metabolites can be distinguished. At the time of the dataset for my analysis, about 100 metabolites had been safely identified. In comparison to the p>>n (i.e., many more variables than data records) problems in bioinformatics, this is rather a standard problem and a classification can be made with known machine learning methods. We were thus able to create classification models by which we could identify key metabolites in renal cell carcinoma.

NGS data is basically a paragon for p>>n data, and this situation will not change for a while since several million variations can be found in a population with feasible levels of effort but far fewer than a million individuals are usually sequenced. In some cases, more variations may be found than individuals that even exist for the sequenced species. These datasets present certain issues, which can be summarized by the curse of dimensionality and potential population structure. Since I have been working for the last few years on the 1001 Genomes Project [8], my main data source was the largest collection of sequenced *Arabidopsis thaliana*. As a model species, *A. thaliana* offers several advantages: it is fast growing; recombinant inbred lines are possible; the genome is quite small; and there are no ethical concerns. On the other hand, it is a 'mere weed'.

For such p>>n data, a subfield of machine learning, dimension reduction, is very helpful. We combined these fields for visualization and added a new measure of the 'quality' of the visualizations.

For the transposons hidden in the 1001 genomes data, we developed a new transposon caller tool, which leverages our data in a better way.

Additional challenges in a project of this scale are data collection, organization, development of other calling pipelines, a final consistency check, and of course selling it reasonable high as paper(s). Apart from the last point, where I was just one in a group of people involved, the remaining points were headed up by me for a longer phase in the project.

Another result that arose within the above mentioned data sets is the solution of the combinatorial problem of getting an exact p-value when putative regulations are inferred and the unbiased validation is a set of proven transcription factors (TRANSFAC database [9]). The outcome is that an exact solution is possible with a computational complexity of O(n^3).

This work resulted in some publications and several useful insights, which are unfortunately not enough for full papers. These latter are also described here.

Vollständige Überschrift dieser Arbeit

Graphenalgorithmen, Kombinatorik, GWAS, Dimensionsreduktion und Klassifikation in Populationen von komplett sequenzierten Genomdaten, vollständigen Microarrays und relativ umfassenden Metabolitdaten.

Kurzfassung

Bereits vor etlichen Jahren wurden in Computerzeitschriften molekularbiologische Daten als exponentiell wachsend aufgezählt [1]. In jüngerer Zeit wird dieses Wachstum bei Sequenzierungsdaten mit Sorge betrachtet, weil sie schneller wachsen als die Datenspeicher, obwohl deren Wachstum exponentiell ist [2, 3]. In dieser Arbeit werden diese riesigen Datenmengen behandelt und analysiert, damit fällt diese Arbeit eindeutig in das Forschungsgebiet Bioinformatik.

Zusätzlich hat sich der klassische Ansatz 'ein Gen/Protein – eine Funktion/Phänotyp' vom Hauptansatz zu einem Ansatz unter mehreren entwickelt, die meisten ([4-7]) davon kombinieren eine Menge Informationen für das Ergebnis. Für diese gibt es die Begriffe: Systembiologie, den Bereich der –omiks, integrative Analyse und einige mehr.

Diese Arbeit spannt einen großen Bogen von vollständigen Microarrays über Metabolitdaten zu Sequenzdaten, mit Seitensträngen in die Tiefe eines kombinatorischen Problems, Dimensionsreduktion und Transposons. Das Ziel ist dabei immer in der gesamten Datensammlung generelle Eigenschaften zu finden, bzw. aussichtsreiche weitere Verfahren.

Die Arbeit an differentiell exprimierten Genen von Tumoren fing mit meiner Diplomarbeit an und setzte sich zu einem Artikel nach Ende fort. Das Hauptergebnis darin: Obwohl die Schnittmengen von differentiell exprimierten Genen in Tumoren, von verschiedenen Artikeln zum gleichen Tumortyp, wie zufällig sind, haben diese Genlisten etwas im Proteininteraktionsnetzwerk gemeinsam.

Die Extraktion von Metabolitkonzentrationen von Geweben ist nach wie vor ein junges Feld, aktuell können einige 100 verschiedene Metaboliten unterschieden werden. Zur Zeit der Daten für meine Analyse waren es etwa 100. Im Vergleich zu den p>>n (das heißt einiges mehr an Variablen als Datensätze) Problemen in der Bioinformatik ist das eher ein Standardproblem und eine Klassifikation kann mit bekannten machine learning Methoden gemacht werden. Damit waren wir in der Lage Klassifikationsmodelle zu erzeugen mit denen wir Schlüsselmetabolite in Nierenkarzinome finden konnten.

NGS Daten sind mehr oder weniger ein Paradebeispiel für p>>n Daten und werden es noch einige Zeit bleiben, da in einer Population einige Millionen von Variationen mit vertretbaren Aufwand gefunden werden können, aber für gewöhnlich weit weniger als eine Million Individuen sequenziert werden. In einigen Fällen können mehr Variationen gefunden werden, als von der untersuchten Spezies überhaupt Individuen existieren. Mit diesen Daten gehen einige Probleme einher, welche folgendermaßen zusammengefasst werden können: Der Fluch der Dimensionalität und Populationsstruktur. Da ich in den letzten Jahren im 1001 Genomes Project gearbeitet habe ist meine Hauptdatenquelle die größte Sammlung von sequenzierten *Arabidopsis thaliana*. *A. thaliana* als Modellorganismus hat einige Vorteile: Wächst schnell, Inzuchtlinien sind einfach machbar, das Genom ist relativ klein und es gibt für diese Spezies keine ethischen Bedenken. Auf der anderen Seite könnte man zu *Arabidopsis thaliana* auch Unkraut sagen.

Für solche p>>n Daten ist eine Teildisziplin von machine learning, Dimensionsreduktion, sehr hilfreich. Wir kombinierten diese Disziplinen für Visualisierung und fanden eine neue Maßzahl für die Güte die Visualisierung.

Für die Transposons, die sich in den Sequenzdaten des 1001 Genomes Project verborgen hielten, entwickelten wir eine neue Methode die vorhandenen Daten besser nützt.

Auch eine Herausforderung in einem Projekt dieser Größe sind die Sammlung der Daten, die Organisation, die Entwicklung zusätzlicher Analysemodule, die Endprüfung der Konsistenz und das möglichst gute Verkaufen als Artikel(n). Abgesehen vom letzten Punkt an dem ich als einer unter einigen Leuten beteiligt war/bin, liefen/laufen die anderen Punkte über längere Phasen des Projekts hauptsächlich zu mir.

Eine weitere abgeschlossene Nebengeschichte ergab sich aus dem kombinatorischen Problem einen exakten p-Wert zu bekommen, wenn mögliche Regulationen generiert werden und die unverzerrte Validierung eine Liste von Transkriptionsfaktoren sind (TRANSFAC Datenbank [9]). Das Ergebnis ist dass eine Lösung in O(n^3) möglich ist.

Diese Arbeit führte zu einigen Artikeln und noch zu einigen mehr an Erkenntnissen die leider keine ganzen Artikel wert sind; hier wird auch letzteres präsentiert.

Table of Contents

SUMMARY **2**

KURZFASSUNG **4**

1. INTRODUCTION **8**

1.1 DATA SOURCES 8
- *1.1.1 Microarrays* *9*
- *1.1.2 Protein–protein interaction* *9*
- *1.1.3 Chromatin-Immunoprecipitation Chip* *9*
- *1.1.4 Metabolite profiling* *10*
- *1.1.5 NGS / sequencing data* *11*

1.2 GENERAL DEFINITIONS 15
1.3 FIELDS OF THE METHODS 17
- *1.3.1 Computational complexity* *17*
- *1.3.2 Graph theory* *18*
- *1.3.3 Combinatorics* *20*
- *1.3.4 Machine learning* *20*
- *1.3.5 String searching and sequence alignment* *20*
- *1.3.6 Efficient programming* *20*

1.4 ORGANIZATIONAL ASPECTS OF LARGE, DATA-INTENSIVE PROJECTS 21
1.5 ETHICS 23

2. AIMS OF THE THESIS **23**

3. GRAPH THEORY AND PROTEIN INTERACTION NETWORKS **24**

3.1 STARTING POINT 24
3.2 ALTERNATIVE AND NEWER DATA SOURCES 24
3.3 ARTICLE: **CHARACTERIZATION OF PROTEIN-INTERACTION NETWORKS IN TUMORS** 25

4. COMBINATORICS IN REGULATORY NETWORKS **37**

4.1 THE GENERAL PROBLEM 37
4.2 THE SPECIFIC SIDE PROBLEM 37
4.3 THE GENERAL SIDE PROBLEM DEFINITION 38
4.4 ARTICLE: **THE OCCURRENCE-IN-SUBTUPLE PROBLEM** 38

5. CLASSIFICATION METHODS AND METABOLOMICS **46**

5.1 CLASSICAL CLASSIFICATION 46
5.2 MACHINE LEARNING FOR CLASSIFICATION 46
5.3 ARTICLE: **METABOLIC PROFILING REVEALS KEY METABOLIC FEATURES OF RENAL CELL CARCINOMA** 47
5.4 RESULTS - ADDENDUM 58

6. CLASSICAL DATA PROCESSING AND MASSIVE DNA SEQUENCING **61**

6.1 THE MASSIVE DATA SOURCE FROM SEQUENCING 61
6.2 COMPUTER CLUSTER ARCHITECTURE 61
6.3 PROJECT MANAGEMENT 65
6.4 PIPELINES FOR CALLING EVENTS 65
6.5 THE EFFECT OF FILTERING AND OTHER BIASES 66
6.6 ARTICLE: **MASSIVE GENOMIC VARIATION AND STRONG SELECTION IN *ARABIDOPSIS THALIANA* LINES FROM SWEDEN** 67
6.7 RESULTS – ADDENDUM 109
- *6.7.1 LAE-finder* *109*

6.7.2 Imputation vs. masking 112
6.7.3 Uniqueness / mapping issues 113
6.7.4 SVs 114

7. DIMENSION REDUCTION AND SINGLE NUCLEOTIDE POLYMORPHISMS 117

7.1 THE CURSE OF DIMENSIONALITY 117
7.2 MEASURES FOR DIMENSION REDUCTION 117
7.3 ARTICLE: **VISUALIZATION OF SNPS WITH T-SNE** 118
7.4 RESULTS – ADDENDUM 125

8. GWAS AND CALLING TRANSPOSONS FROM PAIRED-END READS 126

8.1 THE CHALLENGE OF TRANSPOSONS IN CURRENT SEQUENCING TECHNOLOGIES 126
8.2 ROLES OF TRANSPOSONS 126
8.3 GWAS 126
8.4 ARTICLE: **TE-LOCATE** 127
8.5 RESULTS – DISCUSSION 144

9. CONCLUSIONS & FUTURE DIRECTIONS 145

9.1 …OF BIOLOGICAL BIG DATA IN GENERAL 145
9.2 …OF THE WORK PRESENTED IN THE THESIS 146

10. BIBLIOGRAPHY 149

11. APPENDIX 156

11.1 CURRICULUM VITAE WITH COMPLETE LIST OF PUBLICATIONS 156
11.2 LIST OF FIGURES 158
11.3 LIST OF TABLES 158

12. ACKNOWLEDGEMENTS 159

1. Introduction

Although this is an interdisciplinary thesis, I would locate it *precisely* in the field of bioinformatics. Of course, bioinformatics is seen as a broad area, consisting of anything that combines biology and computer science, and the field is always is in the area of tension between chemistry, biology, computer science, mathematics, and parts of other fields, besides these major ones. An older name for computer science is 'electronic data processing'[1], which was and is, literally, computer science. If we imagine some biological data in a small table where a simple function of standard statistics provides a yes or no answer to a question, this is also bioinformatics, only on a tiny scale. The relatively tiny scale is especially true when the generation of the biological data took months or years, while in this case the analysis takes only minutes.

However, as biological data grows due to standard high-throughput methods, when the underlying source is complex, as most biological systems are, and the questions aim at complex models because these can be inferred with the high resolution gained by a huge amount of data, then the electronic data processing aspect becomes the major effort.

A scientific field usually matures from the simple to the complex questions; for example, it seems quite difficult nowadays to find a new physical principle with just simple experiments and a sheet of paper. 'Complex' in 'complex questions' does not necessarily refer to a long question or one which is difficult to understand, but rather to the answer. A question like 'How do all genes regulate each other?' is short and simple to understand as a question, yet the answer is complex. A question like 'What is the result of this mutation in a gene for which we know already that it is for eye color?' is also not hard to understand and the answer will very likely be as simple. 'Complex' and 'simple' are not necessarily related to the effort of finding an answer. In biology, the simple open questions and the complex questions still exist alongside one another (for an arbitrary example of a solved simple question see [10], for a complex question see [11]). However, there has been a shift in the last years towards the complex questions.

The focus of this work is on the complex questions, that is, on data-intensive questions for which require large amounts of electronic data processing.

1.1 Data sources

Engineering, miniaturization and efficiency are also advancing in biology, resulting in a growing amount of data. While this at first only led to more interesting developments, combinations, and the field of bioinformatics, now a data flood threatens. The NGS/sequencing data amount is growing faster than the computer storage capacities [12], which results in the need of ever-larger computer clusters. If the trend continues, limitations will also arise from excess data, whereas in the past the main bottleneck was from insufficient sampling. Apart from that, noisy[2] data was, and remains, an issue. **Fehler! Hyperlink-Referenz ungültig.**

---

[1] 'Electronic data processing' also refers to analysis, algorithms, or everything done electronically with data.

[2] The term 'noise' comes here from signal processing, where it is referred as 'signal-to-noise ratio' [13]. The idea is that there is a signal in the data (=the information we want to know), and there is the (background) noise, which is overlapping the signal.

1.1.1 Microarrays

DNA Microarrays exist for different aims, but they all have in common the use of many specific DNA sequences which are used as probes and may hybridize to the complementary of the sample DNA [14]. The outcome of a microarray is the absolute amount of complementary sequences, the relative amount, or the simple existence as binary information. In our case, the microarrays were for differential gene expression. The main method used here is the two-color microarray, which follows these main steps:

- isolating the mRNA from two given samples, for example one cell pool from malign tissue, one cell pool from healthy tissue.
- translating the mRNA in cDNA and selective marking of the two samples; the fluorophores Cy3 (green) and Cy5 (red) are usually used.
- applying a mixture of both differently marked samples to the microarrays; this results in a sequence-dependent hybridization to the probes of the array (for cDNA arrays these probes are sequences of lengths ~200-300).
- reading the fluorescent signal and calculating the ratio red/green; these ratios give the relative concentration of the mRNA of sample A in comparison to sample B.

Several other protocols exist, each with slightly different strengths and weaknesses. For an overview see [15].

One of the main issues with microarrays is - and has always been - the normalization of the values. In the case of two sample comparisons, the standard normalization is simply global: the values are scaled so that the average ratio is 1.

A newer method, RNAseq [16], exists for the same purpose. Broadly speaking, the method is to sequence all mRNA of a sample and map it back to the genes. The coverage of the genes by sequences gives the concentration of mRNA. As the sources for differentially expressed genes in this work are exclusively from microarrays, RNAseq remains a side note here.

1.1.2 Protein–protein interaction

Protein-protein interactions, short PPI, are the core information for how the protein circuit in a cell functions. These interactions can be of relevance and are studied on several levels [17]: signal transduction, modifications, fold changes, formation of protein complexes, transport, and so on. Unfortunately, most levels of investigation cannot be done in high-throughput, which entails that the amount of knowledge is biased to the apparently most interesting PPIs [18]. The information, which is most likely unbiased for popular parts, is the binary information whether two proteins interact or not. This was done exhaustively with the yeast two-hybrid [19] and curated with certain other sources collected, for example, in the Interologous Interaction Database, formerly called OPHID [20]. All this binary PPI information forms the hopefully complete protein interaction network.

1.1.3 Chromatin-Immunoprecipitation Chip

The method, in brief, is to acquire the information on proteins and associated DNA/chromatin, which are temporarily bonded [21]. The starting point are living cells.

The particular methods of the chromatin immunoprecipitation (short: ChIP) differ in the amount and type of information gained from the fixed protein-DNA interaction (overview at [22]).

In this work, ChIP-based data is only used for a very specific problem, and only a specific kind of information is used: at the time of the analysis at least, the most likely unbiased complete information of this source was whether a protein interacts with DNA or not. The used source was the TRANSFAC [9] database's manually curated database of eukaryotic transcription factors.

Chromatin-Immunoprecipitation Chip is sometimes abbreviated to 'ChIP-chip' and is seen as belonging to epigenetics, because it can extract information which does not rely on the DNA sequence.

1.1.4 Metabolite profiling

For analyzing metabolites of cells, the full range of analytical chemistry can be used. For an overview of past and current large projects see [23-26]. For our samples, we used GC-TOF-MS (Gas Chromatography Time-Of-Flight Mass Spectrometry), specifically a Leco Pegasus 3 time-of-flight mass spectrometer (Leco, St. Joseph, MI, USA; see more in section 5.3).

Figures 1 and 2 show schematically how this method works in general.

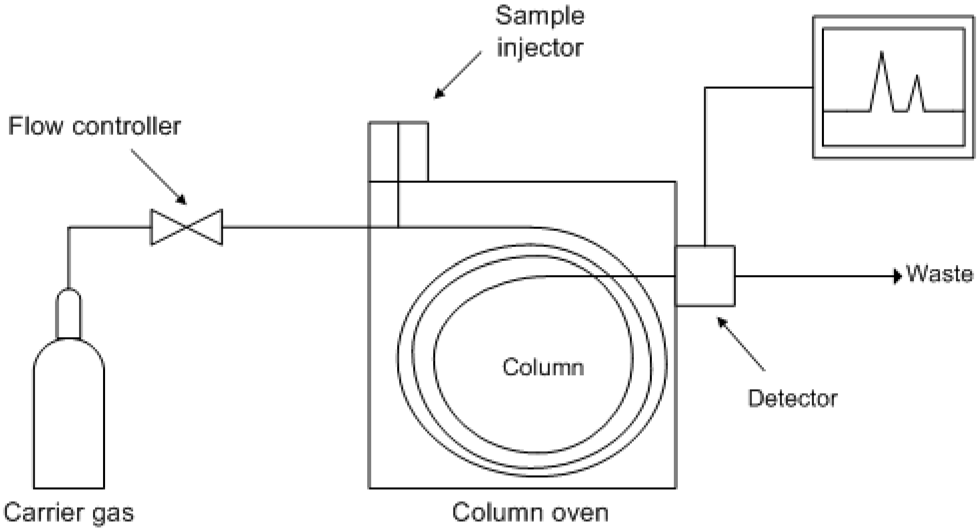


Figure 1. Diagram of a gas chromatograph. The column in the diagram can be anything suitable for a stationary phase. For a GC-TOF-MS the detector resembles that in Figure 2. Figure 1 is from [27]

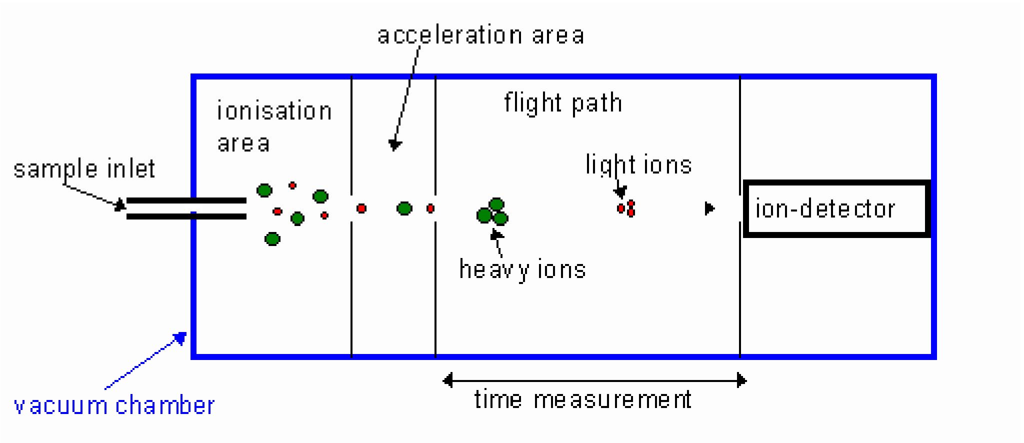


Figure 2. A TOF-MS detector. The minimum requirement for the sample inlet is to provide a continuous ~low flow of sample material; the sample can already be separated at the inlet. The sample molecules are ionized to enable the acceleration with a constant homogeneous electrostatic field. After acceleration, the time to the ion-detector and the amount of ions is measured. The time corresponds to the m/z ratio of the ion (m=mass, z=charge). Figure is from [28].

The resulting spectra is then compared with a database or classified as new. The main error source here is not the concentration of the metabolites, but the mixing-up of different metabolites, especially those with a very similar m/z ratio. At the time when our data was gathered, a few hundred metabolites could be safely distinguished.

1.1.5 NGS / sequencing data

Since 2000, when one human genome was sequenced [29], the amount of sequencing data has grossly increased. The vast majority of this data generation follows the shotgun sequencing [30] idea. An overview of current technologies for NGS/sequencing data generation is in Table 1.

In our case, the data of this type was generated with Illumina machines [31]. Beside the features in Table 1, several Illumina platforms, including that in our mostly contracted sequencing center, have the specific characteristic of producing read-pairs. The Illumina method is briefly shown in Figure 3.

Method	**Single-molecule real-time sequencing (Pacific Bio)**	**Ion semiconductor (Ion Torrent sequencing)**	**Pyro-sequencing (454)**	**Sequencing by synthesis (Illumina)**	**Sequencing by ligation (SOLiD sequencing)**	**Chain termination (Sanger sequencing)**
Read length (bp)	5,500 - 8,500	up to 400	700	50 - 300	50+35 or 50+50	400 - 900
Accuracy (%)	87, single-read accuracy	98	99.90	98	99.90	99.90
Reads per run	50,000	up to 80 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours	2 hours	24 hours	1 to 10 days	1 to 2 weeks	20 minutes to 3 hours
Cost per 1 million bp (US$)	0.33-1.00	1	10	0.05 to 0.15	0.13	2400
pros	Longest read length. Fast	Less expensive equipment. Fast	Long read size. Fast	Cheapest	Cheap	Long individual reads
cons	Moderate throughput. Expensive equipment	Homopolymer errors	Runs are expensive. Homopolymer errors	Expensive equipment. Requires high concentrations	Slower. Palindromic sequence errors	Expensive

Table 1. Overview of DNA sequencing technologies. The table is modified from Wikipedia [32].

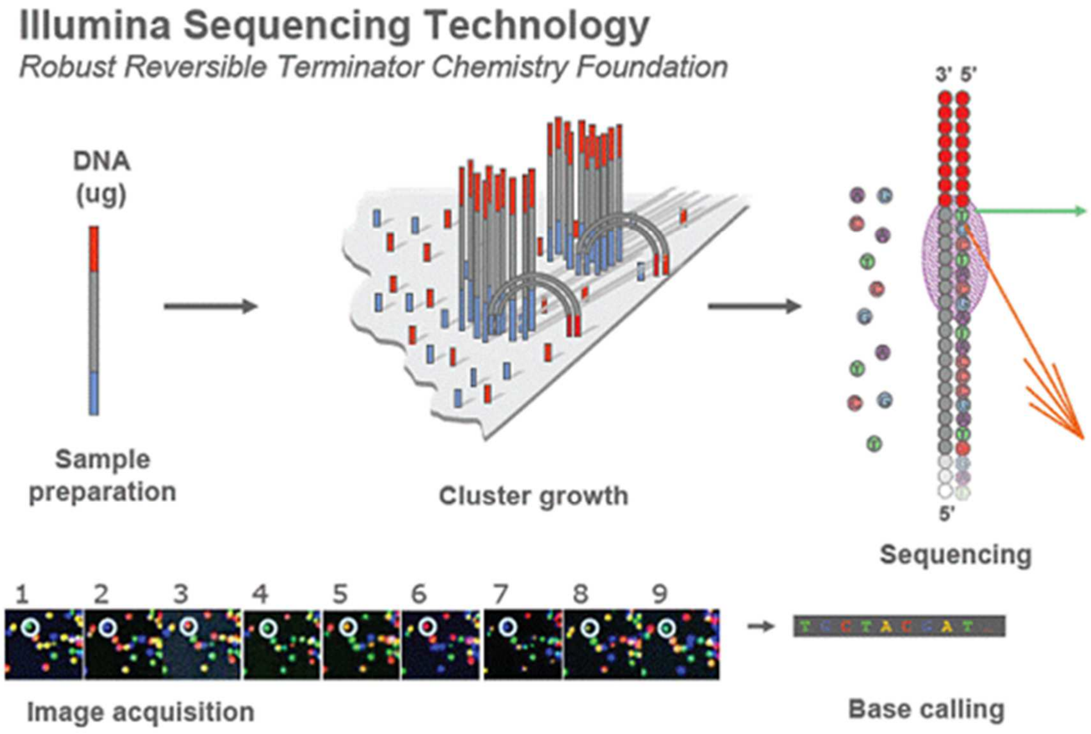


Figure 3. Illumina NGS method. The source for the pairs of reads are the two bound termini (here in blue and red). The given base of each round is determined from its fluorescence color. Figure from Illumina.

A fragment of DNA used for read-pair sequencing consists of several differently named segments. Their definitions are shown in Figure 4.

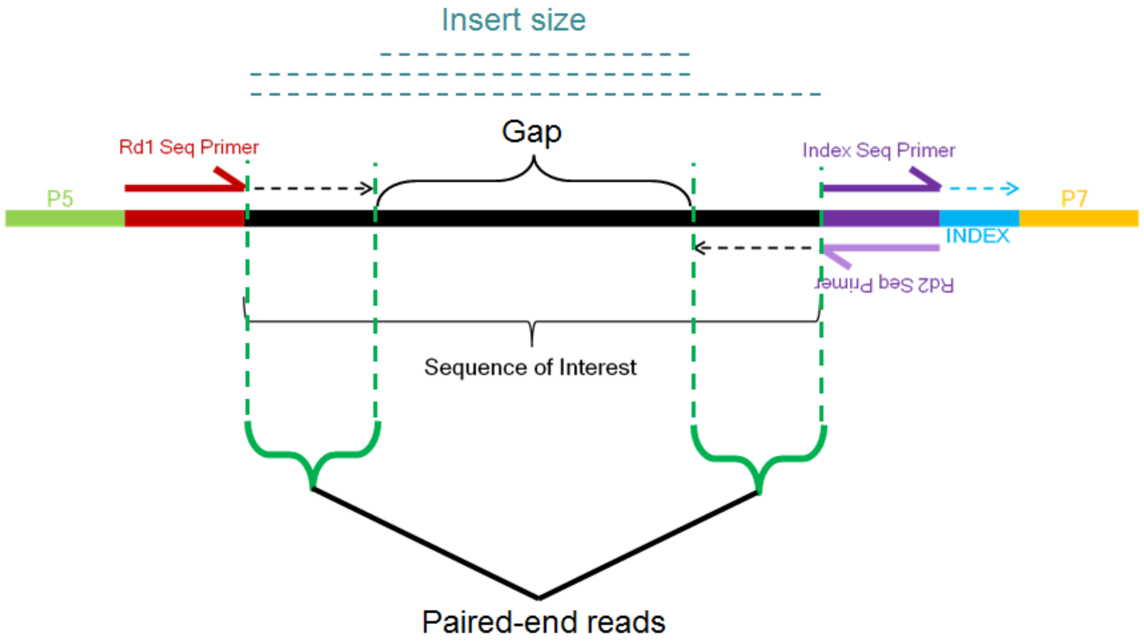


Figure 4. Segments of DNA, resp. their names in a fragment for read-pair sequencing. At the end, the 'raw data' of the sequencing are the two paired reads, here marked with the vertical curly brackets. Unfortunately, the term 'insert size' is inconsistently used in the literature.

Millions of sequence reads can be produced with reasonable effort today, and the effort per read is further decreasing [2]. Consequently, this leads to an enormous increase in the available data [3]. As shown in Figure 5, the increase is even more pronounced than the increase in data storage availability [12], which may lead to new challenges. Currently, the quantity of data is large enough to make full analysis impossible on a single computer; computer clusters are needed. Another implication is that efficient tools and implementations will grow in importance, which may cause a shift in the current ranking of programming languages used in bioinformatics (observed with colleagues as well as in a poll in [33]). A comparison of programming language in standard bioinformatic algorithms can be found at [34].

The optimum for sequencing a genome would likely be to sequence the full chromosomes just once without any noticeable error, but this is far from possible. The current sequencing situation (year 2014) is given by the fact that the generation of many short sequences is much cheaper than the generation of the same combined sequence length by longer sequences [32]. Moreover, there is always a noticeable error rate, which usually increases with the length of the sequence ([35] and is also visible in all of our data samples, see below in chapter 6). Therefore, it is always a trade-off between number of reads, lengths of reads, and error rates. A random error rate is, in some sense, treatable, though a fraction of the error is always systematic.

Special error sources from DNA sequencing reads:

- Duplicated sequences during library preparation: as there is often a PCR-step in the library preparation, it is possible that some fragments are amplified much more often than others (the goal is an even amplification), which can confuse the calling of events. For that reason, for a while it was recommended to remove reads that are aligned at exactly the same coordinates. On the other hand, the removal may mask the information of real sequence duplicates (as in [36], Supplementary Figure 23).
- The (error) noise is not evenly distributed (as an example see the GC bias in sequencing, as reported in [37]).
- The Q-values indicated from the sequencing are on average higher than the empirical Q-values (at least we saw this in our data): one reason is that several statistical models assume evenly distributed errors; another is that there is a kind of drift when longer sequences are produced, which explains why Q-value-recalibration has a positive effect [38]. One source of errors are alignment errors, which are difficult to tackle.
- Most sequencing is based on polymerases, which is also the case in 'independent' data sources for comparison, validations and replicas. That means that any bias is also present in different data sources (if a polymerase has a certain error pattern and the same polymerase is used), which lets the validation and comparison appear better than it actually is.

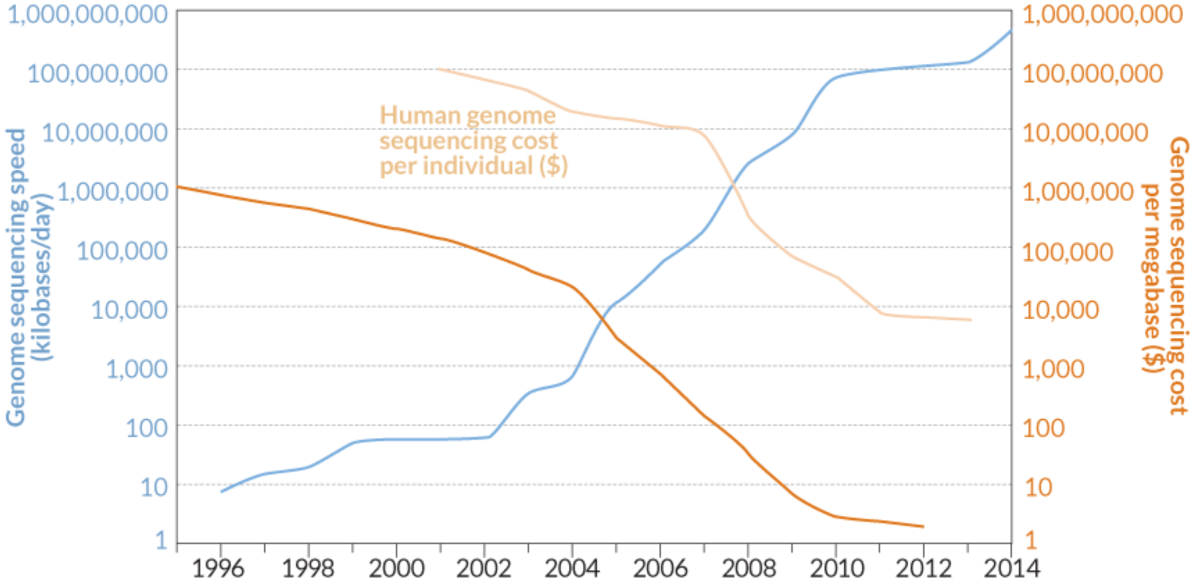


Figure 5. The amount of sequencing. Figure is from [39]

The current quantity of sequencing data with all its peculiarities makes the long-term discipline of alignments even more important; every little advantage in accuracy from noisy data and/or efficiency in processing has some impact. For a recent overview of alignment algorithms, see [40].

Given the trade-offs between number of reads, lengths of reads, error rates and algorithms there are some combinations of sources with certain gains:

- read-pairs with different insert sizes: as the two mates of a read-pair serve as an anchor, different but roughly known distances of these can tackle different sizes of structural variations between them [41].
- longer reads with a higher error rate on a sequence level are used as a skeleton, whereas shorter, more reliable reads are used to reduce the errors of the longer ones [42].
- regions of interest are improved with some single, quite costly additional efforts, for example, longer PCR-fragments.
- information of close samples and/or the population is used, like a consensus alignment within a set of close individuals or as a filter for very suspicious regions (which act very 'un-biologically', but always with the danger of filtering new patterns) [43, 44].

Besides the options on the data generating side, there is a lot of room left for algorithm improvement. More of this will be discussed further in section 1.3.

1.2 General definitions

In this section, the definitions of terms that are essential for the work below are presented; or, from another perspective, for definitions of terms that are precisely defined in their chapters, these are short versions. The relevant area is given in short brackets after the defined term, as some terms may be used differently depending on the area. IT denotes information technology.

Epigenetics: briefly, 'Epigenetics is genetics without genome(-sequence)', or, as in this lengthier quote, 'stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence' [45].

RIL: Recombinant Inbred Line or Recombinant Inbred Strain indicates a strain which, due to long-term inbreeding, has an essentially permanent set of recombination events. It often also refers to homozygous lines for organisms that have more than one set of chromosomes. Where the organism is capable of selfing, as in *Arabidopsis thaliana*, the generation of RILs requires less effort and is less complex. For theoretical background in the complex case, see [46].

NGS data: The abbreviation stands for Next Generation Sequencing. Also referred to as 'post-Sanger sequencing methods', it is an umbrella term for the newer sequencing methods that are producing ever-increasing quantities of data. The source of the data in Figure 5. In the text it is often written as 'NGS/sequencing', although the 'S' in 'NGS' already stands for sequencing; the '/sequencing' is because the general term, 'NGS', is just the current term for the methods which generate this large amount of data. A new term will likely arise for the next technological leap.

Library preparation (for NGS): the preparation of a biological sample for its sequencing. At the end of the preparation, the library consists of the reads in solution.

Coverage (NGS): when reads are aligned to a reference, coverage stands for the number of reads aligned to a locus. . It can also be intended as the average reads per bp of a region.

Genome coverage: The ratio of a reference genome covered with aligned reads or coverage with more reads than an arbitrary threshold.

Calling/calls/called events (NGS): the term 'call' is regularly used in various contexts; in the context of NGS, it signifies an event that is inferred from data.

phred scores: see [47]

Nucleotide ambiguity code: see [48]

SNP: Single-Nucleotide Polymorphism, pronounced ‘snip’, also SNV for single nucleotide variant, it signifies a single nucleotide present as more than one allele. It is somewhat confusing that, on the one hand, many tools have a 'variant-only' option to output only the non-reference calls and, on the other, that a SNP is not only the non-reference side, it is both the reference base and the alternative allele.

indel: a combination of the words INsertion and DELetion, indel is defined from the calling side, that is, relative to the reference; it is not meant in an evolutionary sense. In comparison to SVs, indels are shorter, usually just up to 50bps; the main distinction derives from their source: a linear alignment of a single read. Consequently, the coordinates of indels are accurate to single bps, but, due to gap costs, not larger than 50bps. Most tools follow this notation: the coordinate of an indel is the base before the event; if there are several equivalent possibilities (i.e.: an A inserted in AAAA) the leftmost is chosen.

SV: Structural Variation is a 'long indel' or any other variation of a chromosome, which is not a chromosome abnormality. The main distinction from indels is the manner of calling it: a SV is not called from a single read with a simple linear alignment. If the type of event is not SNP, insertion or deletion, such as inversions or translocations, it is denoted as SV.

Calling (NGS): the process of finding events (= SNPs, indels, SVs) in a genome. The word ‘calling’ also reflects the less than 100% certainty, as NGS methods always imply a certain noise level and also systematic errors [49-51].

Scale-free: a distribution is called scale-free when it follows a power law. The term derives from the unchanging shape when zoomed in (and assuming axes are rescaled). Mostly, it is used with the distribution of the degree of vertices in graphs. A more precise "scale-free metric" is described at [52]

Percent correctly classified (PCC): 100 * correctly classified instances / all instances, the simplest key figure for a classification model.

Lower border / zeroR: the ratio of the most frequent class. A classifier should always perform better than this.

Information entropy: the amount of information which is encoded in a certain sequence, see [53].

Information gain: the decrease of information entropy, which can be also negative. The term is used and better defined at 5.2 Machine learning for classification.

Masking (sequencing): a sequence is filtered by a binary array of the same length. This is the same as RepeatMasker [54] does, when replacing repetitive sequences with 'X'.

Masking (classification models): in trees, or other hierarchical models variables chosen first, make correlated variables much less likely to be chosen later. This term can be better understood in the context in 5.2 Machine learning for classification.

Cross-validation [55]: The given data is divided into n parts, n-1 parts are used for training a model, the remainder is used for testing. It is usually written as '10-fold cross-validation' when n = 10.

Transparent (IT): invisible from the user's perspective. For example, most users are entirely unaware of the existence and functioning of routing in the internet, a complicated topic that is not visible in the browser.

Job (IT): a limited process to be executed in the background. The background can be a single machine or a larger computer cluster.

Walltime (IT): a restriction for a job.

Ad hoc code: a ~neutral way of saying that code is in bad shape and likely not reusable, which could be either because it makes no difference at the level required for a journal, because it is being done for the first time, and/or it is simply not user-friendly.

Explorative data analysis: looking into data without much prior expectation or hypotheses. It can be positive in the sense of avoiding a bias in analysis, but it can be also translated to 'we have a lot of data, what are we going to do now?'. It should be followed by better defined analyses.

1.3 Fields of the methods

As data processing is a broad field whose analyses proceed in various directions, many scientific fields are touched on or even heavily used. This section gives short overviews of these fields.

1.3.1 Computational complexity

The computational complexity theory is a core field in computer science. 'Computational complexity' is the umbrella term covering many useful concepts for the amount of resources needed for a method given a certain amount of data. The short formula for this is

t = f(n), where t is time and n is the amount of data. Usually only the highest exponent to n is noted, the rest is ignored. The idea here is that only the highest exponent to n remains when n goes to infinity. This is quite often noted in the Big O notation [56]; see Figure 6 for common cases.

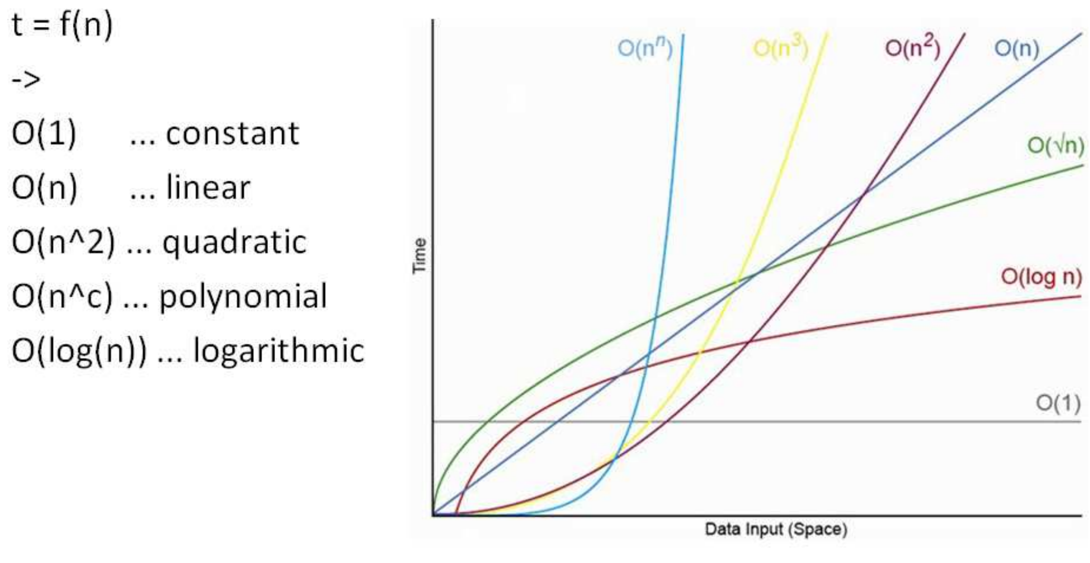
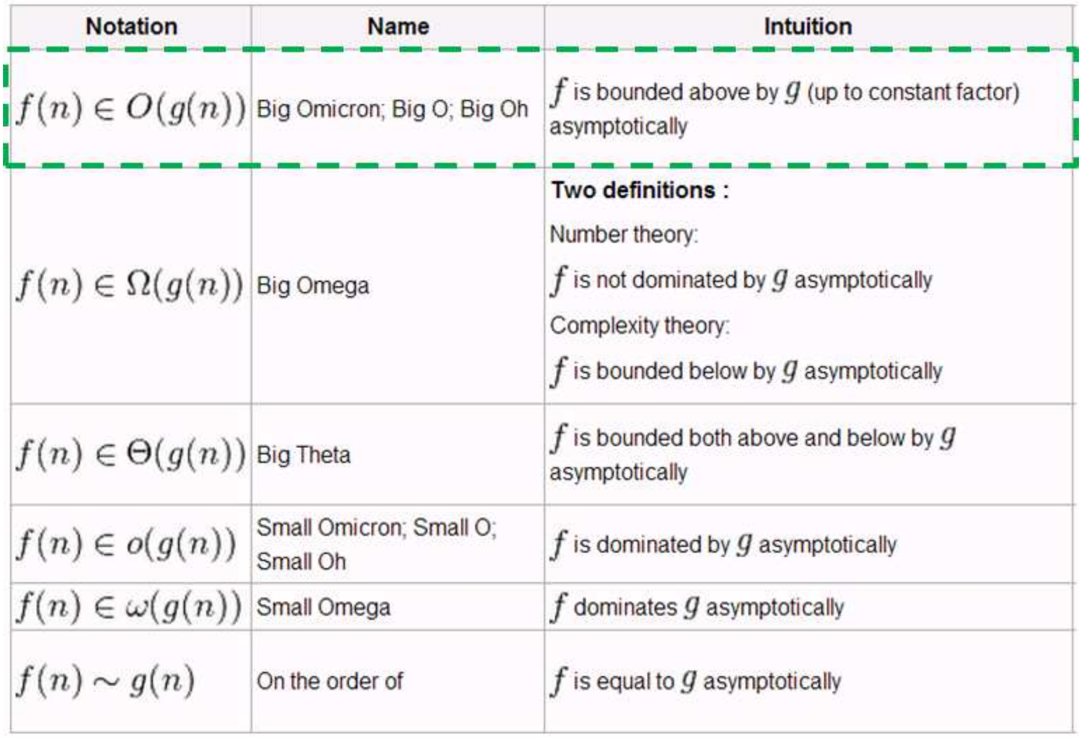


Figure 6. Computational complexity - Big O notation - Common cases

For the sake of completeness, it is important to mention that the O(n) as shown in Figure 6 is only one option to note computational complexity, where f(n) e O(g(n)) is defined as 'f is bounded above by g (up to a constant factor) asymptotically'. See Table 2 for differently defined notations. In methodological papers, as in this work also, the Big O is mainly used. For cases with more than one input dimension, more variables than n can be used, for instance if there are n reads (see section 1.1.5) of length z, then n and z are part of the function O.



Notation	Name	Intuition
$f(n) \in O(g(n))$	Big Omicron; Big O; Big Oh	f is bounded above by g (up to constant factor) asymptotically
$f(n) \in \Omega(g(n))$	Big Omega	**Two definitions :** Number theory: f is not dominated by g asymptotically Complexity theory: f is bounded below by g asymptotically
$f(n) \in \Theta(g(n))$	Big Theta	f is bounded both above and below by g asymptotically
$f(n) \in o(g(n))$	Small Omicron; Small O; Small Oh	f is dominated by g asymptotically
$f(n) \in \omega(g(n))$	Small Omega	f dominates g asymptotically
$f(n) \sim g(n)$	On the order of	f is equal to g asymptotically

Table 2. Family of Bachmann–Landau notations / asymptotic notations. Table is modified from [56]

1.3.2 Graph theory

Graph theory is the study of graphs, where graphs are defined as vertices and edges. This sounds simple, but some of the most complex problems occur in this area ([57] section '6 Problems in graph theory'). A graph can represent all types of networks, interactions, paths, flows, dependencies, etc. and is often the starting point for an optimization problem.

Definitions used in this work:

graph	consists of a vertex set and an edge set; the edges must connect two vertices in the vertex set; vertices on the other hand can be isolated.
simple graph	a graph which contains no multiple edges (=edges with the same start- and endpoint) and no loops.
vertex	is drawn as a node or a dot; is denoted as v for a single vertex and V for a vertex set.
edge	is an undirected connection between two vertices; when the two end-vertices/endpoints are the same, it is called a loop; is denoted as e for a single edge and E for an edge set.
arc	like an edge, but directed; is denoted as a for a single edge and A for an edge set.
subgraph	a subset of vertices of a graph and all edges between them within the graph
path	a *path*[1] between two vertices, where no vertices are repeated

---

[1] The term 'path' on the right side is used as it is commonly used outside graph theory. The path does not need to be the shortest possible path.

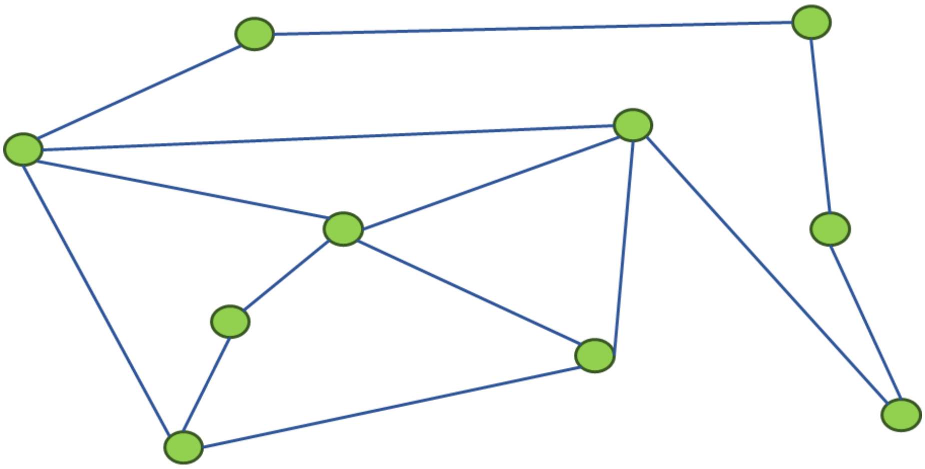


Figure 7. Example of a small graph

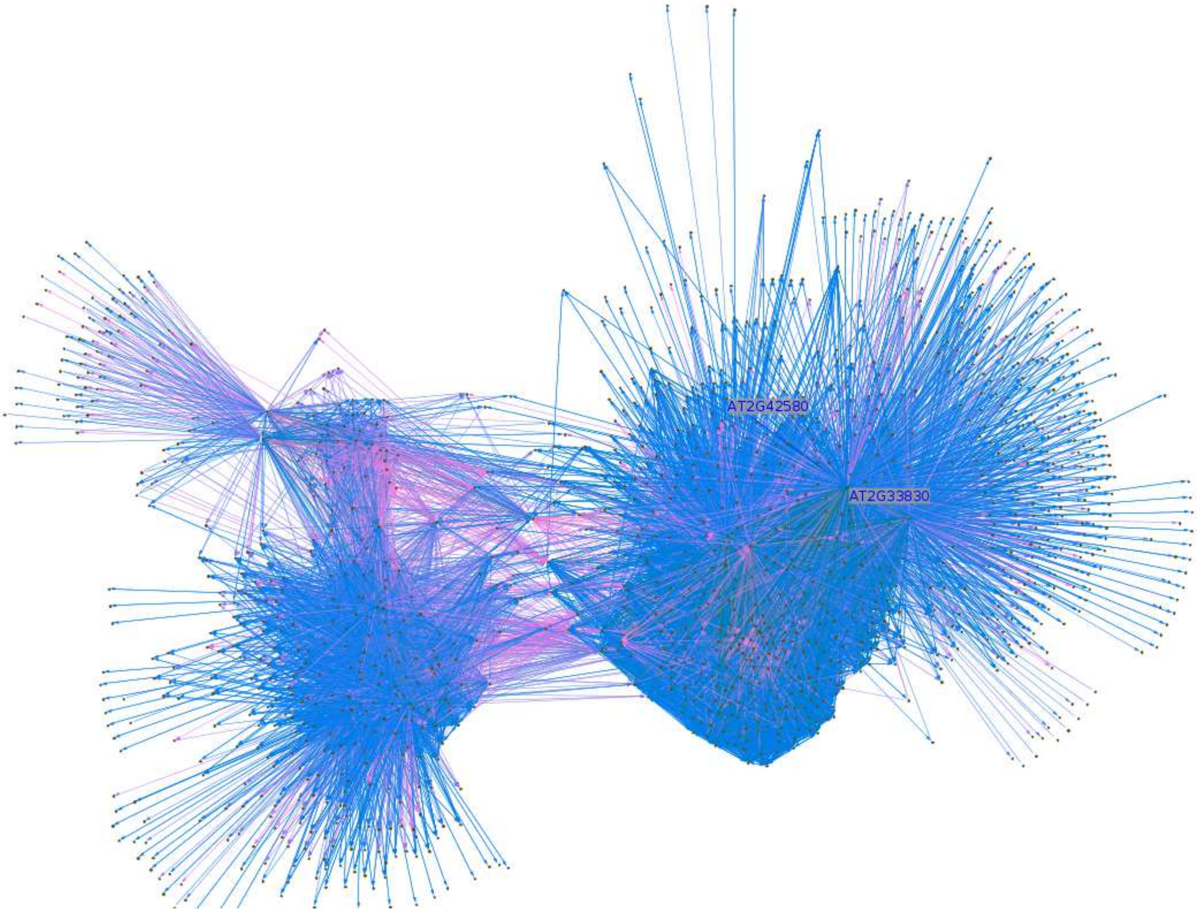


Figure 8. Example of a large graph. It represents a putative regulatory network.

Figures 7 and 8 show examples of a small graph and a large graph. Note that although both graphs are simple graphs (as per the definitions: there are no multiple edges and no loops), they seem different in multiple senses: whereas the small graph was manually drawn, the large one is directed, color-coded and generated with a layout algorithm. It was created with GUESS [58] with the layout algorithm spring embedder [59].

This will be further discussed in section 3.

1.3.3 Combinatorics

Combinatorics is the study of finite or countable discrete structures. For such a large field, Wikipedia is an appropriate starting point [60]. In this work, this field is explicitly named for calculating an exact p-value. However, graph theory and, in another sense, pattern search are subfields of combinatorics.

1.3.4 Machine learning

Machine learning is a large field which is in some areas difficult to strictly differentiate from neighboring areas. Its methods consist mainly of sophisticated heuristics to solve problems, where no analytical solution usually exists. Two other definitions are:

'A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.' [61]

'The core objective of a learner is to generalize from its experience. The training examples from its experience come from some generally unknown probability distribution and the learner has to extract from them something more general, something about that distribution, that allows it to produce useful answers in new cases.' [62]

The sources of the two quotes above are quite extensive bodies of knowledge for this field; a compact source is the material of a course I gave in 2012 [63].

1.3.5 String searching and sequence alignment

String searching is an older problem in computer science, older in the sense of having long been resolved for exact matching. For non-exact matching, the problem is now called (string) alignment and remains a major part in bioinformatics.

For string searching in the classical sense, see [64, 65]; for alignment algorithms a good overview is at [40].

1.3.6 Efficient programming

Efficient programming is a core area in computer science [66], which has been somewhat neglected in bioinformatics until now, perhaps because of a differentiation of the fields.

Efficiency can exist on various levels here: lower resources requirement including time, faster development, and simplicity of use and/or greater stability. These goals can be achieved by various means: algorithmic efficiency (see section 1.3.1); choosing a more efficient programming language; parallelization; and/or hardware. Unfortunately, at some point efficiency might result in a trade-off, which means that increasing efficiency in one sense entails decreasing efficiency in another.

As long as a problem is solved 'well enough', for example, when the data is not too much and it is not a combinatory problem, no more efficiency is needed. Nevertheless, since very large data volumes already exist in molecular biology and there is no end in sight for data generation, efficiency in this regard is growing in importance.

A traditional overview of programming efficiency in the sense of algorithmic efficiency can be found at [67], for parallelization see [68], for parallelization with graphic cards [69] and for comparisons of programming languages in usage and efficiency see [34, 70-73].

1.4 Organizational aspects of large, data-intensive projects

All data analyses need a certain level of organization, but when more data is involved than a single computer can handle and many people are involved in one way or another in trying to optimize the gain from a population of data records/close samples, then a higher level of organization and project management is needed. This section is not exceedingly complex, but its influence on the efficiency of science is often more significant than the science itself. Put another way, a lack of organization can waste more time than dead ends in science (assuming dead ends are published fast enough).

Computer science as 'electronic data processing' has a longer history of dealing with data-intensive issues than biology, and this led, besides the processing called 'software project management' (e.g. [74]), to the field of databases [75]. On the other hand Information Technology is also infamous for failed projects on small and large scales [76], from which may be concluded that enough was happening to learn from!

Several recent sources exist for bioinformatics, including a comprehensive recent review at [77].

For most scientists project management may sound uninteresting and of minor relevance for their actual work. But its major impact can be expressed as follows: in theory a paper with n authors should take 1/n time, which would result in more papers for each author, but in the real world this is not the case [78]. The cited source comes to a quite different conclusion from its extensive extracted data: if papers are counted 1/n for each author and given the clear trend over years to more authors per paper, 'the actual productivity per person has decreased significantly'. Seen from another perspective: with more authors per paper and given that scientists nowadays are not lazier on average than in the past, the necessary work for a paper is significantly higher than 1/n for each author. Alternatively: the work is not organized optimally.

A non-exhaustive list of issues follows which were, or could have been, harmful to projects. It is divided into broad areas:

People:

- no more than 10 people can efficiently interact on one topic in an ad-hoc manner (see [79, 80]).
- when a common resource is developed, such as a common data repository or a sample list, some form of hierarchical organization is required for at least this part, and a single person must be the 'owner' of the common resource; if this is not a single person or if it is not clear to everyone in the project who this ‘owner’ is, it will likely result in either the resource not being used much or the emergence of diverging versions.
- the 'owner' of something (from the previous point) indicates someone able to solely decide on and be responsible for it; more people who are now and then able to decide on the same thing all become 'owners', with confusion being the likely result.
- meetings should always have a previously decided agenda, and a single organizer holding a list of people who must/should/can attend and an action list should refer to it.
- sometimes a poorly organized meeting can be replaced by one or two well-formulated emails.

- in a larger project there should be a list of tasks and preferably also names written by the participants themselves; one person should keep track of it and this person is formally or informally the project leader.

Logical:

- the importance of versioning for software, and therefore also algorithm development, has already been known for a while (though not always done), but some adaptions must be made in case of expensive data analysis. Ideally, all data analysis is repeated whenever a change is made in any of the analyzing modules; practically this is only rarely done because of limitations **->** the version of code and of course the data used for an analysis must always be keep with it; in this manner the code can change, the data can change and the analysis can change, but for any single analysis all required elements must be present in one unit, in the simplest case in one folder, which is not touched again unless the analysis is redone.
- if directories or files are located in one place and have a naming scheme, they should be named so that the sorting makes sense: for example, if a folder name contains the date, it should start with the year and not with the day.

Technology:

- operating systems allow for permissions and permissions should be used.
- when more than one individual is working on a single file, collaborative tools like google docs have advantages over tools for single workstations.
- while data-intense projects might give rise to a storage explosion, it is usually indifferent if some megabytes of modules are present in 100s of slightly different copies; on the other hand the storage amount can be considerable in case of large data. For this, a concept called 'hard links' exists [81].
- backups are important and should be occasionally checked.
- there are several levels of good software development, the basic level is that developed modules run outside their initial setup. If this is not the case, these modules will vanish when development ends.
- files should always have a unique name, especially data files, to make it possible to find all their usages.
- IDs should be numbers and maybe additionally a name to know if it is the right list; in a list of worldwide items, the ID should not be a name since places and names with 'ö', 'ä', 'â', etc. exist.

Project planning:

- late changes of early decisions are very expensive [82].
- if the use case is not exactly defined and more than one person is expected to use it, it is too early to start with the implementation.
- 'milestones' are a useful concept in project management as certain things depend on others, and ignoring these dependencies or interweaving them, such as analyzing data before it is finished without enforcing a full repetition later, results in a mess and/or a waste of time.
- open issues usually remain open issues if no one solves them fully or at least until reliable conclusions can be made.

1.5 Ethics

Ethics is a general issue for personalized data, in genetics, and of course even more in a combination of these two. For this work, the main data sources are plants, which are treated harshly but are not genetically modified. For the protein-interaction networks in tumors human data is used, but only metabolite profiles of the samples are used, without sequencing or other information. The author of this work has not seen any names or any other labels than the tissue sample ID and tumor stage. The preprocessing, so to say, was done by physicians at the hospital.

Other data of ethical relevance were used, but only as far as they are published and freely available. One of the most apparently sensitive data was of openSNP [83], but this was not used as it turned out during analysis that this data is very noisy.

2. Aims of the thesis

As outlined at the beginning of the introduction and in section 1.1, the amount of data in biology is rapidly increasing. There are different levels with regard to the amount of data, that is, they are separated into: to keep and solve mentally; with a large table; with various tools on a computer or on a cluster with an abundance of storage and many cores.

This study aims to investigate the gains of these massive data piles, and also their challenges and issues.

To accomplish the goal of the thesis a self-contained and publishable analysis is done for each large data source. As the data sources are increasing in amount and number, it is not an exhaustive enumeration, but the majority is covered in terms of data volume.

The specific questions to the respective data:

We observe that the lists of genes expressed differently in tumors and healthy tissue from different publications for the same tumor type overlap like random gene lists. Can we find something significant in common in these published gene lists if we consider the complete protein interaction network? (Chapter 3)

If we have a putative regulatory network and a true transcription factor list as validation, what is the exact p-value for finding a certain amount of the transcription factors in the putative network? (Chapter 4 contains a more formal and clearer definition)

We have metabolite levels of cancer tissue and healthy tissue. How can we derive a simple and general model from that? (Chapter 5)

How can we obtain, deal with, and organize more than 1000 fully sequenced *Arabidopsis thaliana* samples? Moreover, what can we infer from them? (Chapter 6)

What can the field of dimension reduction contribute to SNP data? How can structuredness of transformed data be measured? (Chapter 7)

How can transposons be called from paired-end NGS data? What can be inferred from the calls? (Chapter 8)

3. Graph theory and protein interaction networks

3.1 Starting point

An initial observation was made by us (and also by [84, 85]) that lists exist of genes expressed differently in tumors and healthy tissue which are assumed causative or at least closely related, but that the overlap of lists in different publications for the same tumor types is not far from random. The idea was that even if one list of genes is not always responsible for this tumor, it might be that the genes in the lists have some common properties in the protein interaction network. At the time this analysis began, it was becoming fashionable to analyze networks according to their distribution class. Since then a common term has been 'scale-free' (such as in Figure 9), for example in the articles [86-90]. We followed another route: since the full interaction network is known, at least as binary information, we based everything on that, without assumptions of the distribution type. To have the properties of a random subgraph of a certain size, we simply generated many random subgraphs, instead of inferring the properties of an idealized distribution. The underlying rationale is that there is exactly one interaction network and we want to know how the genes there are – and not how they may be in a network of a certain ideal distribution.

For results not presented in the manuscript (section 3.3), see the master thesis of the applicant [91].

3.2 Alternative and newer data sources

In the following manuscript, the main source for differentially expressed genes are microarray data in www.oncomine.org [92]. Oncomine is the database of oncogenomic research that is still one of the largest collections for gene expression in tumors, including RNA-seq data.

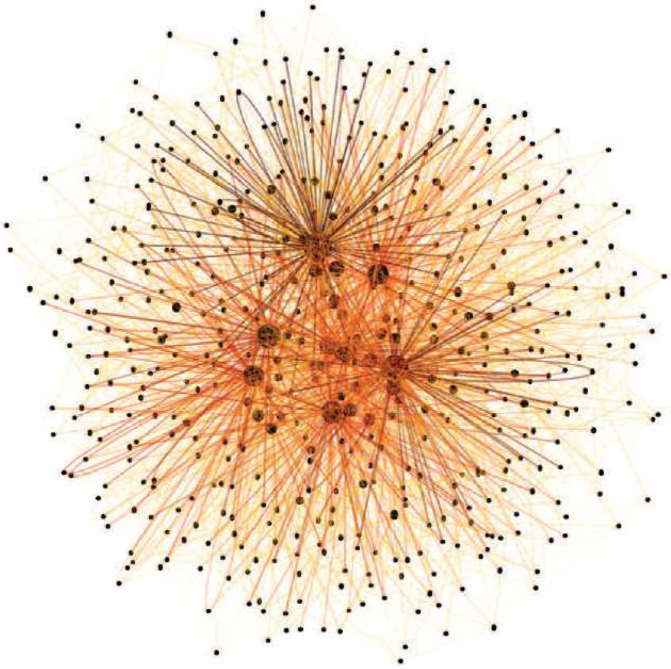


Figure 9. Example scale-free network. From [93].

3.3 Article: **Characterization of protein-interaction networks in tumors**

Alexander Platzer, Paul Perco, Arno Lukas and Bernd Mayer., Characterization of protein-interaction networks in tumors. **BMC Bioinformatics, 2007. 8: p. 224.**

This manuscript is, with some minor additions and the revision, the follow-up of the master thesis performed by the applicant. It offers a solid template for the analysis of similar data and modifications of graph measures, which is likely why it has often been cited.

OWN CONTRIBUTION IN [94]

(see page 35)

BMC Bioinformatics

BioMed Central

Research article

Open Access

Characterization of protein-interaction networks in tumors

Alexander Platzer[1], Paul Perco[1], Arno Lukas[2] and Bernd Mayer*[1,2]

Address: [1]Institute for Theoretical Chemistry, University of Vienna, Waehringer Strasse 17, A-1090 Vienna, Austria and [2]emergentec biodevelopment GmbH, Rathausstrasse 5/3, A-1010 Vienna, Austria

Email: Alexander Platzer - alexanderp@gmx.at; Paul Perco - paul.perco@univie.ac.at; Arno Lukas - arno.lukas@emergentec.com; Bernd Mayer* - bernd.mayer@emergentec.com

* Corresponding author

Published: 27 June 2007

BMC Bioinformatics 2007, **8**:224 doi:10.1186/1471-2105-8-224

Received: 22 December 2006
Accepted: 27 June 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/224

© 2007 Platzer et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Analyzing differential-gene-expression data in the context of protein-interaction networks (PINs) yields information on the functional cellular status. PINs can be formally represented as graphs, and approximating PINs as undirected graphs allows the network properties to be characterized using well-established graph measures.

This paper outlines features of PINs derived from 29 studies on differential gene expression in cancer. For each study the number of differentially regulated genes was determined and used as a basis for PIN construction utilizing the Online Predicted Human Interaction Database.

Results: Graph measures calculated for the largest subgraph of a PIN for a given differential-gene-expression data set comprised properties reflecting the size, distribution, biological relevance, density, modularity, and cycles. The values of a distinct set of graph measures, namely *Closeness Centrality*, *Graph Diameter*, *Index of Aggregation*, *Assortative Mixing Coefficient*, *Connectivity*, *Sum of the Wiener Number*, *modified Vertex Distance Number*, and *Eigenvalues* differed clearly between PINs derived on the basis of differential gene expression data sets characterizing malignant tissue and PINs derived on the basis of randomly selected protein lists.

Conclusion: Cancer PINs representing differentially regulated genes are larger than those of randomly selected protein lists, indicating functional dependencies among protein lists that can be identified on the basis of transcriptomics experiments. However, the prevalence of hub proteins was not increased in the presence of cancer. Interpretation of such graphs in the context of robustness may yield novel therapies based on synthetic lethality that are more effective than focusing on single-action drugs for cancer treatment.

Background

The "omics" revolution has dramatically increased the amount of data available for characterizing intracellular events at the cellular level. The main experimental methodologies responsible for this development have included differential gene expression analysis for recording mRNA concentration profiles, and proteomics for providing data on protein abundance [1,2]. Each technique generates data related to a defined intracellular aspect, such as differential-gene-expression profiles at the transcriptional level, and currently the main focus is on interlinking the various data sources generated by high-throughput screening and array technologies. The concept of systems biology is grounded on such heterogeneous data sources, and also includes the use of homolog information from other systems [3]. Methodologies following the framework of systems biology have increasingly been used to study complex diseases. For example, Hornberg and colleagues discussed the importance of the network topology of protein interactions to selecting drug targets for improving cancer therapy [4].

We have recently outlined a computational analysis workflow aimed at characterizing cellular events at a functional level, which includes the use of differential gene expression and proteomics data, analysis of transcriptional control, and coregulation via joint transcription factor modules, further complemented by protein interaction and functional pathway data [5]. A major goal of such analysis workflows is to decipher biological functioning at the level of protein interactions [6,7]; that is, to elucidate concerted processes by integrating diverse data sources that by themselves do not provide a functional context.

There are several experimental techniques for directly addressing protein-protein interactions, with the yeast two-hybrid system being the most commonly used [8]. The yeast two-hybrid approach can be used to identify protein interactions in vivo, with other techniques such as surface plasmon resonance being performed in a nonbiological environment, but still being useful for providing binding constants [9]. Other technologies involve protein arrays for parallel screening of protein interactions [10]. A recent review has discussed the different methodological approaches [11].

Public-domain databases have been established for making protein-protein-interaction data readily accessible. The Online Predicted Human Interaction Database (OPHID) is a collection of human protein-protein interactions assembled from other databases and complemented by homolog interactions identified in other organisms [12]. The OPHID database used in the present study (as at February 2006) included 41,785 interactions covering 8487 unique proteins of the human proteome. Unfortunately, the database contains only about 20% of the human proteome (presently representing about 39,000 sequences with a unique GI number). Generally, a literature bias is inherent in such interaction data due to disease associated genes and proteins being subject to more detailed analysis, also with respect to protein interactions.

Information on pairwise protein interactions as provided by the OPHID can be used to delineate protein interaction networks (PINs), which are usually represented as undirected graphs. Routines have been published for automatically generating and visualizing such interaction graphs [13,14], where the nearest-neighbor expansion as proposed by Chen and colleagues [15] is a useful approximation for extended graph construction when dealing with the sparse data sets typical of biological systems. Such routines can be used to directly extract PINs utilizing a list of proteins assembled on the basis of differentially expressed genes. If the functional context at the level of protein interactions is represented by the differential gene expression data, this should also be reflected by the characteristics of resulting PINs. Characteristics in this context include both quantitative measures (e.g., the number of nodes found for the largest subgraph) as well as qualitative measures in the biological context (e.g., the identification of hub proteins).

Like many real-world networks, biological networks are scale-free in nature, with the majority of nodes showing a low degree of connectivity, complemented by some highly connected nodes serving as hubs [16,17]. The connectivity, size, and topology of individual PINs are massively influenced by the number of hub proteins involved [18]. However, Lu and colleagues found in a murine asthma model that gene expression of the hub proteins tend to be less affected by disease [19]. The next-most-important factor to determining the overall PIN topology are the simple building blocks – such as a three-node "feedforward loop" motif or a four-node "bi-fan" motif – that have been detected more frequently in transcriptional gene regulatory networks than in networks generated from randomly selected genes [20]. PINs have been recently reviewed by Barabasi and Oltvai [21].

Various groups have applied network analysis to gene data sets associated with cancer. Jonsson and Bates reported very recently that proteins associated with cancer show an increased number of interacting partners in the interactome, reflecting their increased centrality in the PIN [22]. Wachi et al. specifically investigated the role of the interactome of genes differentially regulated in lung cancer [23]. That group found increased connectivity for these genes, in agreement with the findings of Jonsson and Bates. Tuck and colleagues analyzed transcriptional regulatory networks consisting of transcription factors and their target proteins [24]. Genes differentially regulated between acute myeloid leukemia and acute lymphoblastic leukemia were significantly closer in the network as compared to randomly generated gene lists. The analogous result was observed for genes differentially regulated in breast cancer patients. On a more general level, Xu and Li showed that disease-associated genes as listed in the OMIM database [25] tend to interact with other disease-associated genes [26].

The present paper provides a systematic analysis of properties computed for PINs represented as graphs, as exemplified by an extensive set of differential gene expression profiles covering various tumors. The primary hypothesis was that differential gene expression analysis provides systematic data on concerted events in malignant tissue [27], and these systematic data should also be present at the level of protein interactions, in contrast to network properties computed on the basis of randomly generated protein lists.

The formal representation of PINs as undirected graphs makes it possible to utilize a variety of well-established graph measures. Junker and colleagues recently presented a tool for exploring centralities in biological networks, named CentiBiN [28]. CentiBiN can calculate various graph measures, including closeness, betweenness, and eccentricity in protein networks. Jonsson and Bates demonstrated that proteins mutated in cancer showed an increased number of interactions [22]. Another study analyzed protein communities in PINs that were reported as being involved in metastatic processes [29]. Also, Jeong and colleagues were able to identify hub proteins in the PIN that are centrally linked to cell survival [30].

We have computed 22 individual graph measures for 29 tumor-associated differential gene expression data sets that reflect the following graph properties: size, distribution, relevance, density, modularity, and cycles. These graph measures provide a detailed characterization of the differential gene-expression data represented at the level of protein interactions.

Results

A mean of 90 genes (SD = 74 genes, range = 13–300 genes) were identified as significantly differentially regulated for each transcriptomics experiment, and these genes were selected for constructing the entire graph for each given data set. Table 1 lists the number of differentially regulated genes (*N*), the number of nodes in graph (*G*), as well as the number of nodes in the largest subgraph (*G*') for the 29 studies. Furthermore, the characteristics of the individual studies as included in the Oncomine database [31] are listed, including study author, tumor type, class comparison, and number of samples analyzed.

The mean number of nodes in *G* (after performing the nearest-neighbor expansion) was 140 (SD = 120 nodes, range = 14–469 nodes) for the 29 studies, with a mean of 109 nodes for the largest subgraph *G*' (SD = 110 nodes, range = 3–409 nodes). For seven of the studies there were less than 30 nodes in the largest subgraph. Measures related to size, distribution, biological relevance, density, modularity, and cycles were computed for each subgraph *G*'.

Size measures

We used three measures to characterize the graph size as reflected by the number of vertices, the graph expansion, and the length of the shortest path. All three measures – *Closeness Centrality*, *Graph Diameter*, and *Index of Aggregation* – were different for networks generated from gene lists derived from Oncomine than for randomly generated protein lists (Figure 1A,B and 1C), with networks derived on the basis of Oncomine data sets tending to be larger than networks derived on the basis of randomly generated protein sets.

Distribution measures

We used two distribution measures in our analysis: the *Assortative Mixing Coefficient* and the *entropy of the distribution of edges*. The *Assortative Mixing Coefficient* uses the edge-to-edge distribution, whereas the *entropy of the distribution of edges* uses an entropic term reflecting the distinct number of edges per node. We found that the *Assortative Mixing Coefficient* was significantly higher in Oncomine networks than in random networks (Figure 1D).

Biological-relevance measures

Three of the 22 computed measures focused on vertices in the network that were biologically relevant. All of the measures took the shortest path between two vertices in a given network into account. Highly connected proteins, frequently called hub proteins, usually show high *Betweenness*. Joy et al. demonstrated the importance of vertices with high *Betweenness* but low connectivity in the yeast PIN [32]. Interestingly, none of the three computed biological-relevance measures differed significantly between Oncomine networks and randomly generated networks.

Density measures

Eight of the 22 measures utilized in this study addressed aspects of graph density, including *Connectivity*, *Graph Centrality*, *Community*, and *Sum of the Wiener Number*. The numbers of edges and vertices, lengths of shortest paths, and walks on edges were key elements in calculating these measures. Two of the eight measures (*Connectivity* and the *Sum of the Wiener Number*) differed between Oncomine and random data sets (Figure 1E and 1F), and these are influenced by the size of the graph. Oncomine networks are generally larger but less dense than randomly generated networks.

Modularity measures

We calculated three measures reflecting modularity, mainly associated with the number of edges, dilation, and shortest path lengths. One of the computed measures, namely the *modified Vertex Distance Number*, differed between Oncomine networks and randomly generated networks (Figure 1G). This measure is highly correlated to

Table 1: Gene-expression studies and graph measures

Study no.	Study author	cancer type	class I	class II	No. of Samples	N	G	G'	Size (3)	distribution (2)	relevance (3)	density (8)	modularity (3)	circles (3)	total (22)
1	Rosenwald et al.	Leukemia	Blood B cell, Blood T cells, Cell Line, Cord Blood B cells, Cord Blood T cells, Diffuse Large Cell, Follicular Lymphoma, Nonblastic Cell Line, Thymic T cells, Tonsil GC B	Chronic Lymphocytic Leukemia	118	264	426	384	3	2	3	6	3	1	18
2	Segal et al.	Soft Tissue Cancer	Cell Line	Tumor	81	156	252	209	3	2	1	6	3	2	17
3	Rosenwald et al.	Diffuse Large B- Cell Lymphoma – Dlbcl Subgroup	Activated B-Cell-like DLBCL, Type III B-Cell-like DLBCL	Germinal-Center B- Cell-like	240	115	189	165	3	2	2	6	1	2	16
4	Rosenwald et al.	Diffuse Large B- Cell Lymphoma – Dlbcl Subgroup	Activated B-Cell-like DLBCL, Germinal-Center B-Cell-like	Type III B-Cell-like DLBCL	240	129	208	182	3	2	1	6	2	2	16
5	Welsh et al.	Ovary – Type	Normal Ovary	Ovarian Adenocarcinoma	32	96	153	128	3	2	1	6	1	1	14
6	Beer et al.	Lung – Type	Non-neoplastic Lung	Lung Adenocarcinoma	96	158	267	247	3	1	0	6	3	1	14
7	Notterman et al.	Colon – Type	Normal Colon	Ovarian Adenocarcinoma	36	41	62	44	3	1	1	5	1	2	13
8	Higgins et al.	Kidney – Type	Normal Kidney	Clear Renal Cell Carcinoma	29	62	96	76	3	1	2	5	1	1	13
9	Khan et al.	Small Round Blue Cell Tumor/Cell Line	Cell Line	Tumor Sample	86	126	196	155	3	0	1	5	2	1	12
10	Lancaster et al.	Ovary – Type	Ovary	Ovarian Adenocarcinoma	34	106	169	135	3	1	1	5	1	1	12
11	Welsh et al.	Prostate – Type	Normal Prostate	Prostate Cancer	34	50	77	58	3	1	0	4	1	2	11
12	Singh et al.	Prostate – Type	Prostate	Prostate Carcinoma	102	300	469	409	2	1	1	3	2	2	11
13	Liang et al.	Brain – Type	Normal Brain	Glioblastoma Multiforme	33	53	86	70	3	1	0	5	1	1	11
14	Higgins et al.	Kidney – Type	Angiomyolipoma, Chromophobe Renal Cell Carcinoma, Granular Renal Cell Carcinoma, Oncocytoma, Papillary Renal Cell Carcinoma	Normal Kidney	44	55	87	64	3	1	0	4	1	1	10
15	Sperger et al.	Germ Cell – Type	Normal Testis	Seminoma	37	219	342	279	3	1	0	4	1	1	10
16	Shai et al.	Brain – Type	Normal White Matter	Glioblastoma Multiforme	32	56	84	63	3	1	0	4	1	1	10
17	Rickman et al.	Brain – Type	Normal Neocortex of Temporal Lobe	Glioma	51	46	67	42	3	0	0	3	1	1	8
18	Rosenwald et al.	Lymphoid – Type	Normal Blood CD19+ B-Cells, Normal Germinal Center B-Cells	Diffuse Large B-Cell Lymphoma	284	37	60	32	2	0	0	4	1	0	7
19	Frierson et al.	Salivary Gland – Type	Normal Salivary Gland	Adenoid Cystic Carcinoma of Salivary Gland	22	70	104	72	1	1	0	2	1	1	6
20	Bhattacharjee et al.	Lung – Type	Normal Lung	Lung Adenocarcinoma	156	128	195	149	2	0	0	1	1	1	5
21	Bhattacharjee et al.	Lung – Type	Normal Lung	Squamous Cell Lung Carcinoma	38	111	167	123	0	1	0	0	1	1	3
22	Lenburg et al.	Kidney – Type	Normal Kidney	Renal Clear Cell Carcinoma	18	13	14	3	0	0	0	1	0	0	1
23	Garber et al.	Lung – Type	Normal Lung	Squamous Cell Carcinoma	19	26	34	5	0	0	0	0	1	0	1
24	Alon et al.	Colon – Type	Colon	Colon Adenocarcinoma	62	13	16	3	0	0	0	0	0	0	0
25	LaTulippe et al.	Prostate – Type	Non-neoplastic Prostate	Prostate Carcinoma	26	24	29	9	0	0	0	0	0	0	0
26	Iacobuzio-Donahue et al.	Pancreas – Type	Normal pancreas	Pancreatic Adenocarcinoma	17	80	106	35	0	0	0	0	0	0	0
27	Mutter et al.	Uterus – Type	Normal Endometrium	Endometrioid Adenocarcinoma	14	16	18	5	0	0	0	0	0	0	0
28	Bhattacharjee et al.	Lung – Type	Normal Lung	Small Cell Lung Cancer	23	17	20	7	0	0	0	0	0	0	0
29	Garber et al.	Lung – Type	Normal Lung	Lung Adenocarcinoma	46	45	58	9	0	0	0	0	0	0	0

Study number, study author, cancer type, class comparison, and number of samples for data from the Oncomine database. The number of differentially regulated genes (*N*), the number of nodes in graph *G*, the number of nodes in largest subgraph *G'*, and the number of measures per category outside the 2.5% lower and upper confidence limits as derived on the basis of randomly generated gene lists, and the total number of graph measures per study that fell outside the defined significance limits are also listed.

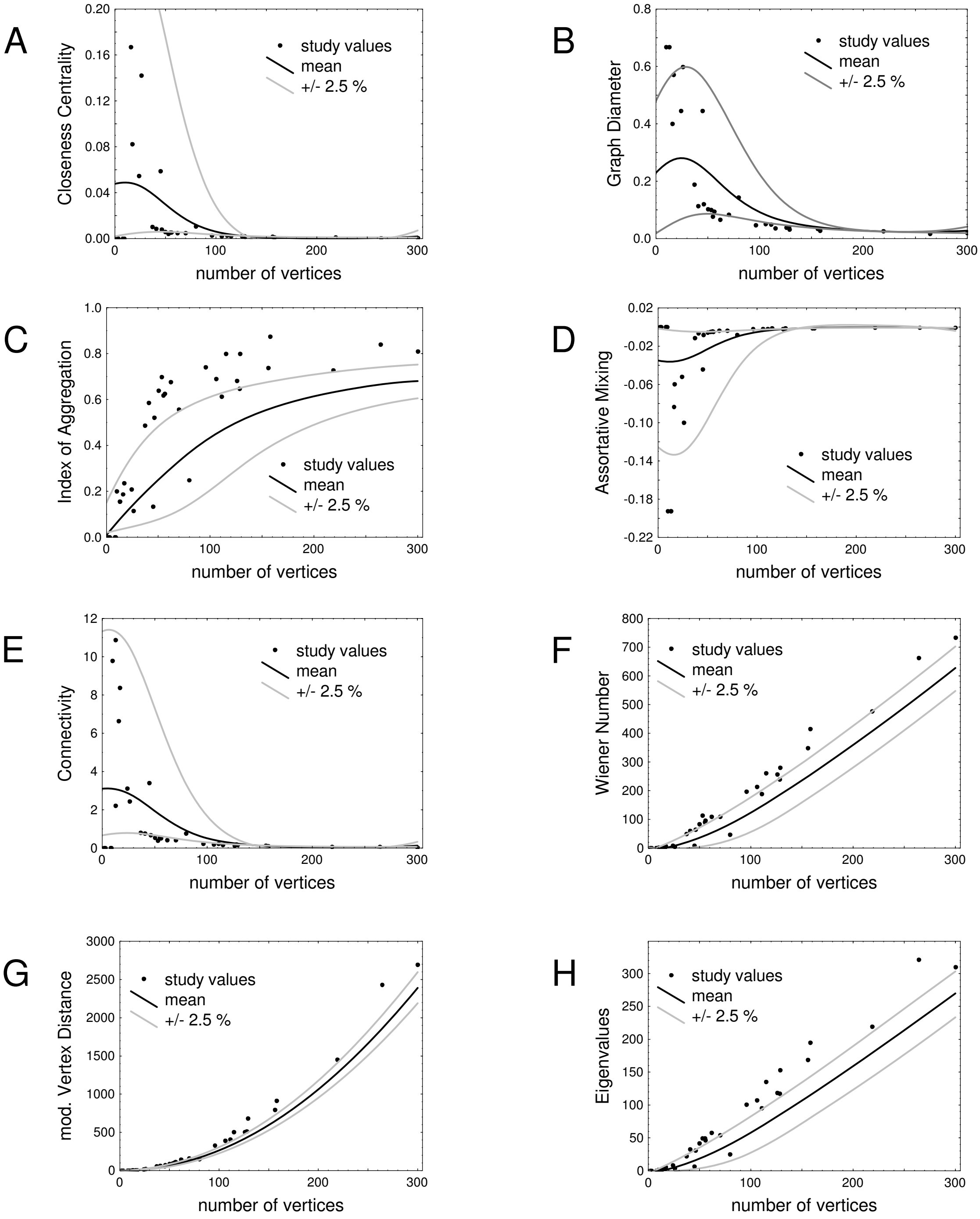


Figure 1

Graph measures. Graph measures (black dots) computed for the given differential gene expression data sets from 29 individual studies with between 10 and 300 genes. The following graph measures are presented: *Closeness Centrality* (**A**), *Graph Diameter* (**B**), *Index of Aggregation* (**C**), *Assortative Mixing Coefficient* (**D**), *Connectivity* (**E**), *Sum of the Wiener Number* (**F**), *modified Vertex Distance Number* (**G**) and *Eigenvalues* (**H**). The mean value (black curve) and the 2.5% lower and upper confidence limits (fitted graphs) based on randomly generated data sets are given for each graph measure.

Table 2: Formal representation of graph measures

Name	Class	Definition	Description	Ref.
Closeness Centrality	size	$CC_i = \frac{1}{\sum_j d(i,j)}$	*d*(*i,j*) is the length of the shortest path between vertices *i* and *j*. The sum of CC_i over all vertices gives the total *Closeness Centrality* of a given subgraph.	**[42]**
Graph Diameter	size	$GD = \frac{\max(d(i,j))}{N}$	*d*(*i,j*) is the length of the shortest path between vertices *i* and *j*. *GD* is computed for all pairs (*i,j*), and reflects the longest path identified.	**[43]**
Index of Aggregation	size	$IoA = \frac{A}{B}$	*A* is the total number of vertices in the subgraph, and *B* is the total number of all given vertices in the graph.	**[15]**
Assortative Mixing Coefficient	distribution	$r = \frac{4* < k_1 * k_2 > - < k_1 + k_2 >}{2* < {k_1}^2 + {k_2}^2 > - < k_1 + k_2 >^2}$	k_1 and k_2 are the counts of edges of two vertices connected by a given edge. This measure reflects the edge-to-edge distribution over all edges of a graph.	**[44]**
Entropy of the distribution of edges	distribution	$H = -\sum_k p(k) \ln p(k)$	*k* is the count of edges of one vertex, and *p*(*k*) is the ratio of vertices that have *k* edges.	**[45]**
Betweenness	biological relevance	$B = \frac{\sum_{i \in V} \sum_{j,k} \frac{\sigma(j,i,k)}{\sigma(j,k)}}{N}$	$\sigma(j,i,k)$ is the total number of shortest connections between vertices *j* and *k*, where each shortest connection has to pass vertex *i*, and $\sigma(j,k)$ is the total number of shortest connections between *j* and *k*. We computed $\sigma(j,i,k)$ and $\sigma(j,k)$ for the entire OPHID graph, but then only used vertices also present in the subgraph generated on the basis of a given gene-expression data set.	**[42]**
Betweenness of all selected Vertices	biological relevance		As for *Betweenness*, but considering all selected vertices.	**[42]**
Stress Centrality	biological Relevance	$StC = \sum_{i \in V} \sum_{j,k} \sigma(j,i,k)$	$\sigma(j,i,k)$ is the total number of shortest connections between vertices *j* and *k*, where each shortest connection has to pass vertex *i*.	**[42]**
Connectivity	density	$C = \frac{A}{B}$	*A* is the total number of edges realized in a given graph, and *B* is the maximum number of edges possible.	**[43]**
Clustering Coefficient	density	$CLUST_i = \frac{A}{B}$	*A* is the total number of edges between the nearest neighbors of vertex *i*, and *B* is the maximum number of possible edges between the nearest neighbors of vertex *i*. The sum of $CLUST_i$ over all vertices gives the total *Clustering Coefficient* of a given subgraph.	**[46]**
Number of edges divided by the number of vertices	density	$NeNv = \frac{A}{B}$	*A* is the total number of edges in a given graph, and *B* is the number of selected vertices in a given graph.	-
Community	density	$Comm = \frac{A}{B}$	*A* is the total number of edges, where both connected vertices are in the given subgraph, and *B* is the total number of edges, where one connected vertex is in the subgraph and the other vertex is outside it.	**[47]**
Entropy	density	$H(G) = \sum_{v \in V, i(v) >= 2} (i(v) - 1) * \log(\frac{\mid E \mid - \mid V \mid + 1}{i(v) - 1})$	where |*E*| is the total number of edges, |*V*| is the total number of vertices, and *i*(*v*) is the number of edges of vertex *v*.	**[48]**

Table 2: Formal representation of graph measures *(Continued)*

Graph Centrality	density	$GC_i = \frac{1}{\max(d(i,j))}$	max(*d*(*i*,*j*)) is the length of the shortest path between vertices *i* and *j* for a given vertex *i*.	**[42]**
Number of walks of length n	density	$NW = \sum NW_i$	NW_i is one walk with a length of *n* edges in the subgraph.	**[43]**
Sum of the Wiener Number	density	$W_i = \frac{1}{2} * \sum_{i,j} d(i,j)$	*d*(*i*,*j*) is the length of the shortest path between vertices *i* and *j*. We computed the *Sum of the Wiener Number* for each vertex.	**[43]**
Total number of triangles of a subgraph and its dilation	Modularity		Given a subgraph *g* of graph *G*, the complement of *g*, denoted as *g*, is the subgraph implied by the set of vertices *N*(*g*) = *N*(*G*)\N(*g*) The dilation of *g* is the subgraph *δ*(*g*) implied by the vertices in *g* plus the vertices directly connected to a vertex in *g*. The coat of nearest neighbors of the subgraph is defined as *DN*(*g*) = *δ*(*g*)\N(*g*) The set of all valid triangles for *g* is defined as *VT*(*g*) = {*x*,*y*,*z* | (*x*,*y*,*z* ∈ *N*(*δ*(*g*)) ^ (*x*,*y*),(*y*,*z*),(*z*,*x*) ∈ *E*(*δ*(*g*))) ∩ (*x* ∈ *N*(*g*) ^ *z* ∈ *DN*(*g*))} where *N* is the number of vertices and *E* is the number of edges in the graph. The result for a subgraph *g* is the total number of elements in *VT*(*g*).	**[42]**
Localized Modularity	modularity	$LM = \frac{\lvert \mathrm{E}_{\text{inside}} \rvert}{\lvert \mathrm{E}_{\text{within the (direct) neighbors}} \rvert} * \frac{\lvert \mathrm{E}_{\text{inside}} \rvert * \lvert \mathrm{E}_{\text{to the outside}} \rvert}{\lvert \mathrm{E}_{\text{within the (direct) neighbors}} \rvert^2}$	where |*E*| is the total number of edges.	**[49]**
modified Vertex Distance Number	modularity	$mVD = \sum_{i,j \in V, i \neq j}^{V} \frac{1}{d(i,j)^2}$	*d*(*i*,*j*) is the length of the shortest path between vertices *i* and *j*. For this measure, *i* and *j* are all selected from V.	-
Eigenvalues	cycles	$EV = \sum_j \lvert ER_j \rvert^2$	ER_j is the real part of the *j*-th *Eigenvalue* for the adjacency matrix of the given subgraph.	**[50]**
Subgraph Centrality	cycles	$SC = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{\infty} \frac{(A^k)ii}{k!}$	*A* is the adjacency matrix. We computed *SC* for *k* [1,99].	**[42]**
Cyclic Coefficient	cycles	$\theta(i) = \frac{2}{k_i * (k_i - 1)} * \sum_{j,k} \frac{1}{S_i(j,k)}$ $\theta = 1/N * \theta(i)$	S_i is the smallest possible cycle of vertex *i* and two of its neighboring vertices *k*. The total *Cyclic Coefficient* for all vertices *N* is then given as θ	**[42]**

Name, formal representation, and short description of graph measures computed for the categories of size, distribution, biological relevance, density, modularity, and cycles.

Closeness Centrality, which is also based on the sum of shortest paths between two vertices.

Cycles measures

The three measures implemented related to graph cycles were the *Cyclic Coefficient*, *Subgraph Centrality*, and *Eigenvalues*. The *Eigenvalues*, calculated from the adjacency matrix of the graph, differed between randomly generated data sets and Oncomine (Figure 1H). *Eigenvalues*, like *Subgraph Centrality*, mainly depend on all cycles of the graph, but the two methods differ in the scaling of cycle sizes. The *Cyclic Coefficient* mainly depends on local short cycles.

To study the data sets at the level of the graph-measure categories, the 22 graph properties of each data set were checked for measures that significantly deviated from those of random graphs. Results of this evaluation are listed in Table 1, where the individual studies are sorted by the total number of graph measures that deviated significantly from those derived from random gene selections. The study that deviated the most from random selections related to leukemia, in which 18 of the 22 graph measures were different. On the other hand, in six studies none of the graph measures differed significantly from random selections. Tests of the correlation between the number of graph measures deviating from their respective values for random selections and the total number of genes differentially regulated ($r^2 = 0.34$, $p < 0.05$), the total number of nodes in graph G ($r^2 = 0.38$, $p < 0.05$), and the total number of nodes in the largest subgraph G' ($r^2 = 0.43$, $p < 0.05$) revealed the dependence on number of nodes selected and the degree of deviation from random selections. This correlation was significantly affected by the small graphs analyzed, since studies resulting in subgraph sizes of less than 10 do not provide conclusive graph measures.

Interestingly, the number of samples analyzed for differential gene expression was not significantly correlated with the number of statistically significant differentially regulated genes found ($r^2 = 0.09$, $p = 0.12$), nor with the number of graph measures deviating from the randomly generated reference sets ($r^2 = 0.11$, $p > 0.05$).

Discussion

We characterized PINs derived from 29 gene-expression profiles of various tumors (as listed in Table 1) by computing 22 graph measures (as listed in Table 2). In general, the values of the graph measures did not depend on the type of microarray used in the analysis (cDNA arrays or Affymetrix Gene Chips). The small number of individual data sets per cancer type made it impossible to delineate a correlation between graph measures and tissue type. Interestingly, the number of samples used was not correlated with the number of statistically significant differentially expressed genes, and also not with the number of graph measures deviating from random selections. Under the assumption of comparable sample processing, expression results are strongly affected by the tissue and cancer type, and to a lesser extent on the number of samples per group.

We assigned the graph measures to the following categories: size, distribution, biological relevance, density, modularity, and cycles. The individual graph measures that showed significant differences (defined as identifying at least 50% of gene-expression experiments outside the 2.5% lower and upper confidence limits computed on the basis of randomly generated data sets) between cancer networks and networks based on randomly generated data sets were *Closeness Centrality*, *Graph Diameter*, *Index of Aggregation*, *Assortative Mixing Coefficient*, *Connectivity*, *Sum of the Wiener Number*, *modified Vertex Distance Number*, and *Eigenvalues*.

All three measures associated with the size of the graph differed significantly between tumor networks and randomly generated networks. The *Index of Aggregation* was on average higher in tumor networks, indicating dependencies between proteins involved in cancer, as also proposed by Chen et al. in the context of Alzheimer disease [15]. This increased connectivity is also consistent with data obtained by Jonsson et al. [22]. However, it is likely that the bias in OPHID interactions toward disease-associated genes contributes to these findings. The values of both *Graph Diameter* and *Closeness Centrality* were significantly lower in tumor networks. This finding was also reported by Yu and colleagues for networks solely including highly expressed genes in the yeast interactome [33]. Low *Closeness Centrality* values for tumor networks may initially appear surprising, but relative large size of the largest subgraphs in tumor networks (on average close to 80% of all nodes of G are also part of G') makes higher *Closeness Centrality* values harder to obtain. The largest subgraph of tumor networks also more elongated shortest paths between nodes.

One measure of the distribution category, the *Assortative Mixing Coefficient*, differed significantly in tumor networks. This coefficient is influenced by both the number of hub proteins and the number of edges, and a large number of hub proteins is correlated with an unequal distribution in the number of edges. The *Assortative Mixing Coefficient* is directly proportional to the number of edges and inversly proportional to the number of hub proteins. According to Jonsson and colleagues, tumor networks contain numerous hub proteins [22]. However, our data generally indicate the presence of a small number of edges per node, and no evidence for a large number of hub proteins.

The *Sum of the Wiener Number* characterizes the density of the graph. The significantly higher values of this measure in tumor networks indicate larger graphs, which is consistent with the observed *Index of Aggregation*. We found that the *Connectivity* was lower in the largest subgraphs of tumor networks. This may be also due to the largest subgraphs of tumor networks being on average larger than the subgraphs of randomly generated gene lists, corresponding to low values of *Closeness Centrality*.

The *modified Vertex Distance Number* is also influenced by the sum of shortest paths between two vertices, but in contrast to *Closeness Centrality*, all vertices in the OPHID network are considered. A higher *modified Vertex Distance Number* in tumor networks indicates higher connectivity and modularity in Oncomine networks. Finally, higher *Eigenvalues* values indicate the presence of fewer cycles in tumor networks.

Our analysis of 29 studies on differential gene expression in cancer has revealed a general tendency toward large subgraphs without the presence of explicit hubs. Comparing the graph measures between the individual gene expression studies and randomly selected genes provided a heterogeneous picture. Gene-expression studies resulting in a low number of statistically significant differentially regulated sequences (and consequently small subgraphs) do not support an interpretation at the level of PINs (see expression studies 22–29 in Table 1) as performed in this study: for small subgraphs the variance of graph measures determined for randomly selected gene lists is high, which prevents identification of significant differences of small subgraphs derived on the basis of differential gene-expression data.

Conclusion

The usefulness of analyzing topological characteristics of cancer networks for supporting drug targeting was recently highlighted by Hornberg and colleagues [4]. We based our study on a diverse set of cancer types, and have identified characteristics of cancer networks from differential-gene-expression data. In particular, measures of graph size deviated significantly from those for graphs constructed from random gene selections. Genes showing significant differential expressions in cancer appear to be interlinked also at the level of PINs. However, we were not able to identify hub proteins from the given data, or nodes exhibiting high *Betweenness*. Such nodes have been considered as primary targets for therapeutic interventions.

Extended graphs with a low density may indicate a network with high robustness – in contrast to networks containing hub proteins. This points to a different approach for identifying therapeutic intervention, namely synthetic lethality. This concept originates in classical genetics, where only the combination of two specific mutations leads to cell death. In metabolic networks a single node deletion can often be bypassed by different routes in the pathway. Combining this with a second deletion in that alternative pathway may only then result in lethality [34]. Analysis of the given PINs with respect to functional pathways and their potential bypass routes has the potential to identify synhetically lethal protein target combinations, as has been shown experimentally in yeast [35].

Methods

Databases

We used the OPHID [12] to derive information on human protein-protein interactions. This database contains information on protein-interaction pairs, where each protein is given by its Swiss-Prot identifier. We mapped the Swiss-Prot identifiers on the corresponding Gene Symbols so as to link gene-expression data sets, which mapped 8487 Swiss-Prot entries to 6033 different Gene Symbols. Among the protein-interaction sources used by the OPHID, we included HPRD (Human Protein Reference Database) [36], MINT (Molecular Interaction Database) [37], RikenBIND and RikenDIP [38], BIND (Biomolecular Interaction Network Database, [39], and MIPS (Munich Information Center for Protein Sequences) [40]. These data sets are mostly based on experimental evidence, which is further supported by expert reviews based on the scientific literature. We did not include interactions from other sources of low-to-medium quality that are also listed and indicated as such in the OPHID.

The OPHID provides interaction information in the form of object A interacting with object B. This information can be used to derive interaction graphs when providing an identifier list (A, B, ..., N), as resulting from the analysis of differential-gene-expression data.

We used Oncomine as a central repository for differential-gene-expression data [31]. This database provides an extensive collection of gene expression data on cancer, and compares various types and subgroups. A total of 962 raw data sets were identified in Oncomine (as at April 2006). We manually selected all gene expression studies where the malignant tissue was compared to a reference (either healthy tissue or a cell line). We initially selected 40 individual experiments covering tumors of 17 different tissues (4 B-cell, 1 bladder, 2 colon, 2 endometrium, 2 ovary, 5 brain, 1 liver, 1 leukemia, 9 lung, 1 multicancer, 3 kidney, 1 pancreas, 4 prostate, 1 salivary gland, 1 testis, 1 thyroid, and 1 soft-tissue tumor), of which 17 used cDNA arrays and 23 used Affymetrix Gene Chips. The mean number of available features per study was 11459 (range = 1988–44928 features).

We extracted each file and processed the raw data according to the following scheme: The two groups per study were analyzed at the level of individual genes by computing a probability value for the differential expression of a particular gene in that given experiment. Multiple testing was accounted for by using the Holm-Sidak step-down test and setting the significance level to 0.05 [41]. This procedure yield a mean of 278 genes from each study (range = 2–1838 genes). From the initial 40 gene expression data sets, 29 showed between 10 and 300 differentially expressed genes (mean = 90 genes), and these studies were included in subsequent analyses.

Each of the 29 selected differential gene expression studies was represented by a list of genes exhibiting significant differential regulation when comparing expression values for the group of tumor samples and the group of reference samples. Each gene on these lists was represented by its Gene Symbol, allowing a direct match with the protein interaction data as derived from the OPHID.

Graph construction

Protein interaction graphs (*G*) were constructed for each gene list of the 29 selected gene-expression studies based on OPHID interaction data utilizing the nearest-neighbor expansion. This procedure built edges between the nodes of entries A and B of a given gene list if the interaction between A and B was directly encoded in the OPHID, or if one element X was identified in the OPHID, allowing the construction of an interaction of the type A - X - B, where X was not listed in the gene expression data set [15].

For each gene list, entire graph *G* comprising *n* subgraphs *G'* was constructed on the basis of genes in the initial list and their nearest neighbors in the PIN. *G'* is defined as a graph whose vertices and edges form subsets of the vertices and edges of *G*.

Gene lists derived from analyzing differential gene expression might be linked on the level of coregulation and protein interactions. To quantitatively assess such dependencies, the graph properties of PINs derived on the basis of randomly selected gene lists were computed as follows: Proteins encoded by randomly selected gene lists exhibit a background level of protein interactions, and we analyzed graph measures characterizing gene expression data sets with respect to random data sets. One thousand random gene sets containing between 10 and 300 genes were picked in steps of 10. For each of these gene sets, the largest subgraph *G'* was generated again following the nearest-neighbor expansion as outlined above, and the graph measures were computed for each *G'*. This procedure yielded the mean value and 2.5% lower and upper confidence limits for each graph measure for each data set size represented by the 1000 individual data sets.

Graph measures and data evaluation

The graph measures for each largest subgraph *G'* were then determined for each Oncomine data set as well as for random data sets. Table 2 lists all of the applied graph measures. (Software for computing these properties on the basis of given Gene Symbol lists is available from the authors upon request.) The graph measures derived for Oncomine data sets were then interpreted in the context of the measure scales based on random data sets. A graph measure was considered as interesting in the context of cancer associated networks if at least 50% of the 29 Oncomine experiments showed this measure to be outside the 2.5% lower and upper confidence limits as computed on the basis of the randomly generated data sets.

Authors' contributions

BM and PP designed the study. AP extended the concept, developed the software, and performed all the calculations. AP, PP, AL, and BM contributed to data interpretation and writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This study was partly supported by the European Union (project number LSHC-CT-2005-018698).

References

1. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21:**33-37.
2. Tyers M, Mann M: **From genomics to proteomics.** *Nature* 2003, **422:**193-197.
3. Kitano H: **Systems biology: a brief overview.** *Science* 2002, **295:**1662-1664.
4. Hornberg JJ, Bruggeman FJ, Westerhoff HV, Lankelma J: **Cancer: a Systems Biology disease.** *Biosystems* 2006, **83:**81-90.
5. Perco P, Rapberger R, Siehs C, Lukas A, Oberbauer R, Mayer G, Mayer B: **Transforming omics data into context: bioinformatics on genomics and proteomics raw data.** *Electrophoresis* 2006, **27:**2659-2675.
6. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, de Atauri P, Aitchison JD, Hood L, Siegel AF, Bolouri H: **A data integration methodology for systems biology.** *Proc Natl Acad Sci USA* 2005, **102:**17296-17301.
7. Hwang D, Smith JJ, Leslie DM, Weston AD, Rust AG, Ramsey S, de Atauri P, Siegel AF, Bolouri H, Aitchison JD, Hood L: **A data integration methodology for systems biology: Experimental verification.** *Proc Natl Acad Sci USA* 2005, **102:**17302-17307.
8. Fields S, Song O: **A novel genetic system to detect protein-protein interactions.** *Nature* 1989, **340:**245-246.
9. Smith EA, Corn RM: **Surface plasmon resonance imaging as a tool to monitor biomolecular interactions in an array based format.** *Appl Spectrosc* 2003, **57:**320A-332A.
10. Kersten B, Wanker EE, Hoheisel JD, Angenendt P: **Multiplex approaches in protein microarray technology.** *Expert Rev Proteomics* 2005, **2:**499-510.
11. Stelzl U, Wanker EE: **The value of high quality protein-protein interaction networks for systems biology.** *Curr Opin Chem Biol* 2006, **10:**551-558.
12. Brown KR, Jurisica I: **Online predicted human interaction database.** *Bioinformatics* 2005, **21:**2076-2082.
13. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
14. Breitkreutz BJ, Stark C, Tyers M: **Osprey: a network visualization system.** *Genome Biol* 2004, **4:**R22.
15. Chen JY, Shen C, Sivachenko AY: **Mining alzheimer disease relevant proteins from integrated protein interactome data.** *Pac Symp Biocomput* 2006:367-378.

16. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407:**651-654.
17. Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proc Biol Sci* 2001, **268:**1803-1810.
18. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430:**88-93.
19. Lu X, Jain VV, Finn PW, Perkins DL: **Hubs in biological interaction networks exhibit low changes in expression in experimental asthma.** *Mol Syst Biol* 2007, **3:**98.
20. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks.** *Science* 2002, **298:**824-827.
21. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5:**101-113.
22. Jonsson PF, Bates PA: **Global topological features of cancer proteins in the human interactome.** *Bioinformatics* 2006, **22:**2291-2297.
23. Wachi S, Yoneda K, Wu R: **Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues.** *Bioinformatics* 2005, **21:**4205-4208.
24. Tuck DP, Kluger HM, Kluger Y: **Characterizing disease states from topological properties of transcriptional regulatory networks.** *BMC Bioinformatics* 2006, **7:**236.
25. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30:**52-55.
26. Xu J, Li Y: **Discovering disease-genes by topological features in human protein-protein interaction network.** *Bioinformatics* 2006, **22:**2800-2805.
27. Segal E, Friedman N, Kaminski N, Regev A, Koller D: **From signatures to models: understanding cancer using microarrays.** *Nat Genet* 2005, **37(Suppl):**S38-45.
28. Junker BH, Koschutzki D, Schreiber F: **Exploration of biological network centralities with CentiBiN.** *BMC Bioinformatics* 2006, **7:**219.
29. Jonsson PF, Cavanna T, Zicha D, Bates PA: **Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis.** *BMC Bioinformatics* 2006, **7:**2.
30. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42.
31. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **ONCOMINE: a cancer microarray database and integrated data-mining platform.** *Neoplasia* 2004, **6:**1-6.
32. Joy MP, Brock A, Ingber DE, Huang S: **High-betweenness proteins in the yeast protein interaction network.** *J Biomed Biotechnol* 2005, **2005:**96-103.
33. Yu H, Zhu X, Greenbaum D, Karro J, Gerstein M: **TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics.** *Nucleic Acids Res* 2004, **32:**328-337.
34. Ghim CM, Goh KI, Kahng B: **Lethality and synthetic lethality in the genome-wide metabolic network of Escherichia coli.** *J Theor Biol* 2005, **237:**401-411.
35. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS: **Gene function prediction from congruent synthetic lethal interactions in yeast.** *Mol Syst Biol* 2005, **1(2005):**0026-.
36. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TKB, Chandrika KN, Deshpande N, Suresh S, *et al.*: **Human protein reference database as a discovery resource for proteomics.** *Nucleic Acids Res* 2004, **32:**D497-501.
37. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2002, **513:**135-140.
38. Suzuki H, Fukunishi Y, Kagawa I, Saito R, Oda H, Endo T, Kondo S, Bono H, Okazaki Y, Hayashizaki Y: **Protein-protein interaction panel using mouse full-length cDNAs.** *Genome Res* 2001, **11:**1758-1765.
39. Bader GD, Betel D, Hogue CWV: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31:**248-250.
40. Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V: **MIPS: analysis and annotation of proteins from whole genomes in 2005.** *Nucleic Acids Res* 2006, **34:**D169-172.
41. Dupuy A, Simon RM: **Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting.** *J Natl Cancer Inst* 2007, **99:**147-157.
42. da Fontoura Costa L, Rodrigues FA, Travieso G, Boas PRV: **Characterization of complex networks: A survey of measurements.** 2005 [http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0505185].
43. Bonchev D: **Complexity Analysis of Yeast Proteome Network.** *Chem Biodivers* 2004, **1:**312-326.
44. Holme P: **Efficient local strategies for vaccination and network attack.** *Europhys Lett* 2004, **68:**908-914.
45. Claussen JC: **Offdiagonal Complexity: A computationally quick complexity measure for graphs and networks.** 2004 [http://www.citebase.org/abstract?id=oai:arXiv.org:q-bio/0410024].
46. Bader GD, Hogue CWV: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics* 2003, **4:**2.
47. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D: **Defining and identifying communities in networks.** *Proc Natl Acad Sci USA* 2004, **101:**2658-2663.
48. Kieffer J, Yang EH: **Ergodic behavior of graph entropy.** *ERA Amer Math Soc* 1997, **3:**11-16.
49. Muff S, Rao F, Caflisch A: **Local modularity measure for network clusterizations.** *Phys Rev E* 2005, **72(5 Pt 2):**056107-056111.
50. Chung F, Lu L, Vu V: **Spectra of random graphs with given expected degrees.** *Proc Natl Acad Sci USA* 2003, **100:**6313-6318.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

4. Combinatorics in regulatory networks

4.1 The general problem

This problem started with a few putative regulatory networks (like the network in Figure 8), which we wanted to compare and thus looked for validation. Unbiased experimental data is preferable here, but, unfortunately, the information on regulatory genes is likely rather biased: regulations of 'interesting' genes are known in detail, whereas other genes are rather neglected. 'Negative' information is less safe: a regulation observed in several experiments is quite reliable, but the opposite case, where the information that two genes definitely do not regulate each other, is less reliable. In fact, it is likely that two genes do not regulate each other, as it is a solid assumption that the regulatory network is quite sparse [95], that is, that most genes do not regulate each other, which also implies that any experiment cannot add much certainty here. The most reliable unbiased information is whether a gene is a transcription factor or not, because this is usually performed in a complete manner, for example with ChIP-on-Chip (see section1.1.3). A newer method for the same purpose is ChIP-sequencing [96]. Both methods have in common that they already result in gene-gene-regulation, and/or protein-DNA interaction, and that both have weaknesses which bias the results: common motifs, repetitive regions, non-specific nature of DNA binding proteins [97] and specific weaknesses of each respective method. Additionally, most compiled data sources of such experiments have additional prior knowledge of regulations, which is often called 'confirmation of functional relevancy' and the information about functional relevancy are likely biased in the same way as, for example, gene ontologies [98]. Nevertheless, whether a gene is a transcription factor or not is much less biased from these error sources. One good source here is TRANScription FACtor database (TRANSFAC [9]).

We also looked for validation key figures with the known gene-gene-regulation, even when they were biased, but the present chapter focus on the validation key figures with the transcription factor information of genes, because it turned out that this combinatorial problem was not so far resolved.

4.2 The specific side problem

From the description in the section before, we have on the one side putative regulations selected from n genes, that is, a subset of the full set of possible regulations of $n * n - 1$ elements (autoregulation of genes is excluded). From this set with elements in the form 'gene a regulates gene b', we have a set of regulating genes A. On the other hand, we have a set of validated transcription factors or regulating genes. These two sets should naturally overlap. Given an overlap, we want to know its p-value. One major use of this p-value is to compare it to p-values of other overlaps in case different putative regulations are inferred. Since all of the inferred sets of regulating genes should be highly significant, we would like to have an exact p-value for comparing them.

4.3 The general side problem definition

The specific side problem arising from the biological case can be defined generally:

SN ... a set of N different elements
SNN ... a tuple of N different elements occurring N-1 times
Sx ... a subtuple of SNN
Sy ... a subset of SN

We have N different elements occurring N-1 times. From this tuple we take x elements. We have a subset of set SN with the number of y elements. It is searched in Sx elements for the Sy elements and z elements well be found. How likely is it to find z or more elements of y in x?
(from [99], the paper of this chapter)

It was not difficult to get an approximate solution for this, at least sufficient to judge if the given set is significant, and it was not difficult to formulate an exact formula needing factorial time to compute[2], but it was quite a challenge to get a formula for an exact solution and of a polynomial complexity class. The computational complexity of the final formula in the following paper is O(x^3).

4.4 Article: **The Occurrence-in-subtuple problem**

Alexander Platzer. The Occurrence-in-subtuple problem. **Arxiv, 2008. arXiv:0811.4192 [math.CO].**

This manuscript was not put in an attractive location, inter alia because the solution/derivation is for a very specific problem, with limited use for other questions. On the other hand, the way to the problem and surrounding of this specific problem was never published, in part because this method prevented the completion of another paper showing that one putative regulatory network is not as solid as first thought. Nevertheless, this would be a poor reason for not publishing this method at all.

In retrospect, I realized that it is not only less common to cite Arxiv – although there are famous examples such as Grigori Perelman, a winner of the Fields Medal in 2006 – but also rather more inconvenient as there is no direct interface to EndNote^{TM}, for example.

OWN CONTRIBUTION IN [99]

Everything

---

[2] A computational complexity class higher than polynomial according to n, which means no computer can solve this in a lifetime for n > 1000. The naive formula already needed one day of a single computer for n = 5.

The Occurrence-in-subtuple problem

Alexander Platzer

November 25, 2008

Abstract

As we go along with a bioinformatics analysis, we stumbled over a new combinatorial question. Although the problem is a very special one, there are maybe more applications than only this one we have. This text is mainly about the general combinatorial problem, the exact solution and its derivation. An outline of a real problem of this type is in the discussion.

1. The problem

Definitions:

Given are:
SN ... a set of *N* different elements
SNN ... a tuple of *N* different elements occurring *N*-1 times
Sx ... a subtuple of *SNN*
Sy ... a subset of *SN*

Derived is:
Sz ... the subset of *Sy* of all elements of *Sy* occurring in *Sx*

With the variables:
N ... the number of different elements in *SN*
N-1 ... the count of occurrences of every different element in *SNN* $|SNN| = N*(N-1)$
$|Sx| = x$
$|Sy| = y$
$|Sz| = z$

The question:
We have *N* different elements occurring *N-1* times. From this tuple we take *x* elements. We have a subset of set *SN* with the number of *y* elements. It is searched in *Sx* elements for the *Sy* elements and there will be *z* elements found. How likely is it to find *z* or more elements of *y* in *x*?
(an illustration of this question is in figure 1)

Or with the general declarations:
What is the ratio of permutations of *SNN,* which results in an equal or higher *z* as given, to all possible permutations of *SNN*?

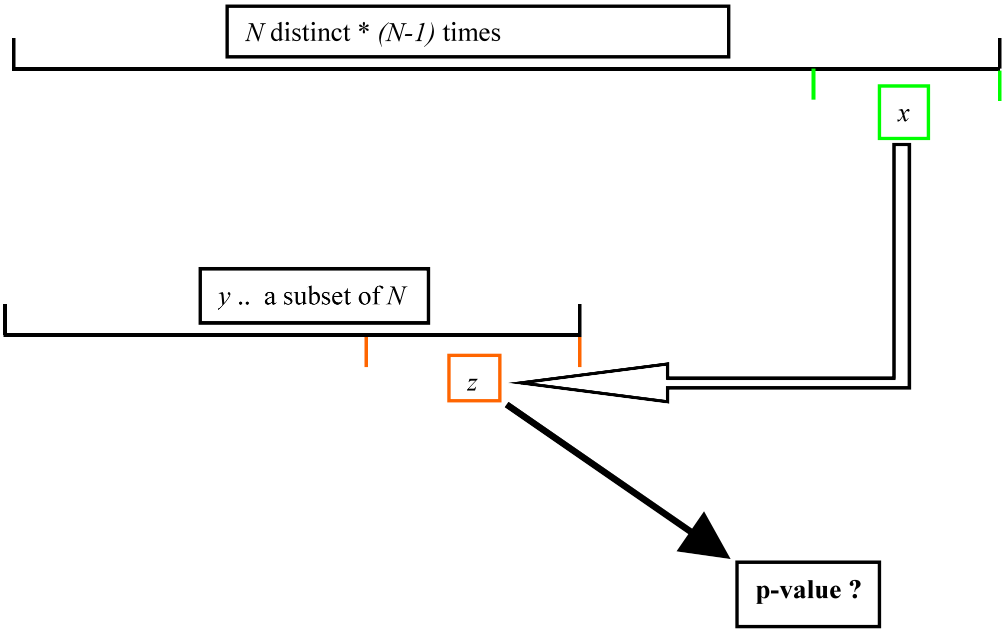


Figure 1. The combinatorial problem illustrated

2. Derivation

First of all it is important that the result should be a p-value, so we need the ratio of permutations fulfilling the condition (z or more elements of Sy are in Sx), not the combinations. To get permutations of combinations in Sx and the rest the factor $x!*(N*(N-1)-x)!$ is needed. And at the end we divide by all permutations possible $((N*(N-1))!$, the factor becomes

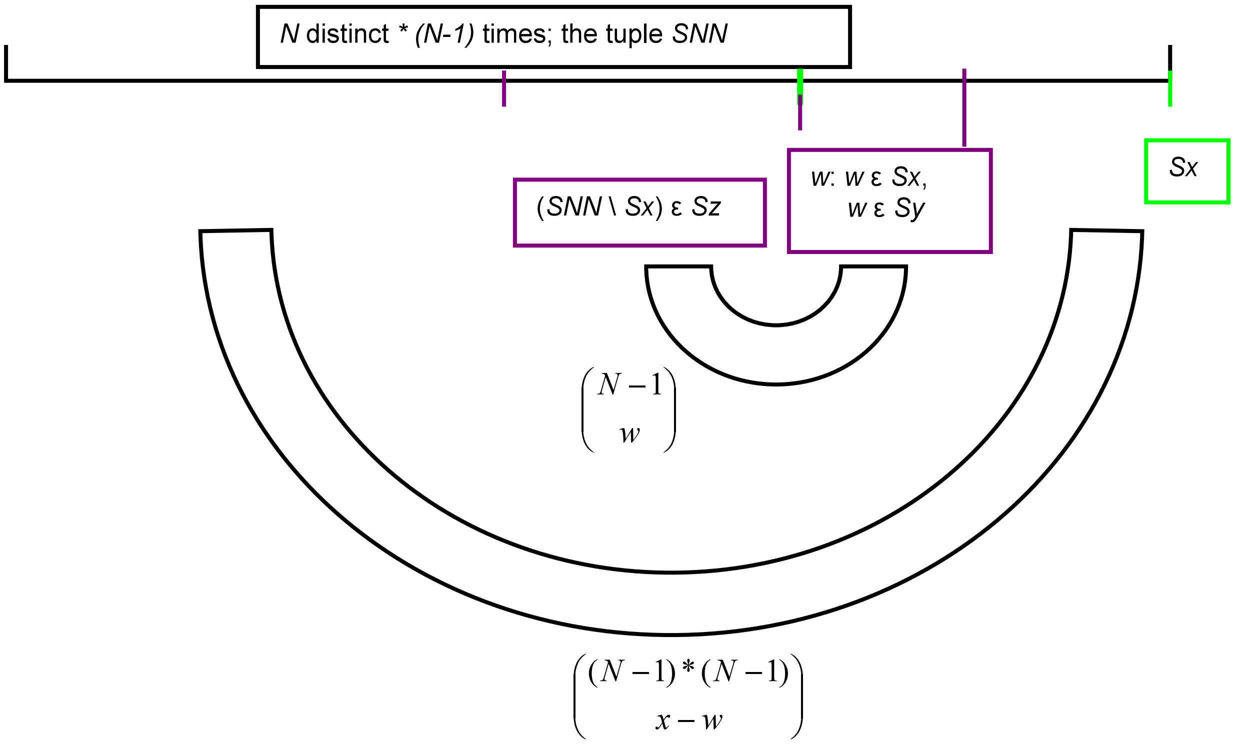
$$\frac{x!*(N*(N-1)-x)!}{(N*(N-1))!}$$ which is

$$\frac{1}{\binom{N*(N-1)}{x}} \tag{1}$$

if x=1, y=z=1 then this number of combinations is $N-1$ (because $N-1$ elements of one type exist) if x=2, y=z=1 for 2 times one element of Sz in Sx the number of combinations is $(N-1)*(N-2) / 2$; $(N-1)$ possibilities to choose the first and $(N-2)$ possibilities to choose the second, but the two chosen elements cannot be distinguished so divide by two) and so on ->

$$\binom{N-1}{how\,many\,times\;\;one\,element}$$ and for the question, how many combinations exist with one or more times one element of Sz in Sx (still x=2, y=z=1) it is the sum of $N-1$ and $(N-1)*(N-2)$, but notice that this are only the combinations of the elements of Sx which occur in Sz, each of this combination has another side (for the elements of Sx which are not occurring in Sz).

For a number of w elements equal to one element in Sz in Sx, it can be seen as 2 times k-combinations:


N distinct * (N-1) times; the tuple SNN
(SNN \ Sx) ε Sz
w: w ε Sx,
w ε Sy
Sx
$\binom{N-1}{w}$
$\binom{(N-1)*(N-1)}{x-w}$

so the count of combinations of w elements in SNN element in Sz in Sx, still y=z=1, is

$$\binom{N-1}{w}*\binom{(N-1)*(N-1)}{x-w}$$, for different w's just summing up.

If we change to more than one *z*, but still *y*=*z* then *w* is changed to a vector containing the count of every element of *Sz* in *Sx*. Moreover, there must be two constraints, first the sum of w must not be larger than *x* and no component of *w* may be larger than *n-1* (plus as before every component of *w* must be 1 or larger). If we have *k*=*z* and sum up all combinations (of possible values for each component of *w*):

$$\sum_{w1\ldots k=1}^{w1\ldots k\le n-1 \,\wedge\, \sum_{j=1}^{k} wj\le x} \left(\left(\prod_{j=1}^{k}\binom{n-1}{w_j}\right)*\binom{(n-\min(x;y))*(n-1)}{x-\sum_{j=1}^{k} w_j}\right) \tag{2}$$

for the case *y*>=*z* it is to include that a combination of *z* or more elements are chosen from *Sy*:
(vector w is renamed to vector i)

$$\sum_{k=z}^{\min(x;y)}\binom{y}{k}*\sum_{i1\ldots k=1}^{i1\ldots k\le n-1 \,\wedge\, \sum_{j=1}^{k} ij\le x} \left(\left(\prod_{j=1}^{k}\binom{n-1}{i_j}\right)*\binom{(n-\min(x;y))*(n-1)}{x-\sum_{j=1}^{k} ij}\right) \tag{3}$$

and with the factor for make permutations out of combinations (1) :

$$\frac{\sum_{k=z}^{\min(x;y)}\binom{y}{k}*\sum_{i1\ldots k=1}^{i1\ldots k\le n-1 \,\wedge\, \sum_{j=1}^{k} ij\le x} \left(\left(\prod_{j=1}^{k}\binom{n-1}{i_j}\right)*\binom{(n-\min(x;y))*(n-1)}{x-\sum_{j=1}^{k} ij}\right)}{\binom{n*(n-1)}{x}} \tag{4}$$

we have now a formula for the problem, but it can be seen easily that it will need large computing power to solve it with large parameter values. The order of this formula is O(*min(x;y)*(n-1)^ min(x;y)*), so we have only changed one combinatorial problem into another (faster, but still incalculable in terms of time).
Therefore, the next task is to transform this formula into one with a lower order.
The bad thing in it is the second sum, there the power to min(x;y) occurs.

First, we expand to:

$$\frac{\sum_{k=z}^{\min(x;y)}\binom{y}{k}*\sum_{s=k}^{x}\left(\sum_{i1\ldots k=1}^{i1\ldots k\le n-1 \,\wedge\, \sum_{j=1}^{k} ij=s} \left(\prod_{j=1}^{k}\binom{n-1}{i_j}\right)*\binom{(n-\min(x;y))*(n-1)}{x-s}\right)}{\binom{n*(n-1)}{x}} \tag{5}$$

One thing to remark is that the restriction $i_{1\ldots k}$ <= *n*-1 is only there for information, it is fulfilled of the formula alone because the binomial coefficient of $\binom{N-1}{a}$ with *a* > *N-1* is 0.

Now we look closer to

$$\sum_{i1\ldots k=1}^{i1\ldots k\le n-1 \,\wedge\, \sum_{j=1}^{k} ij=s} \left(\prod_{j=1}^{k}\binom{n-1}{i_j}\right) \tag{6}$$

If the vector components would not start with 1 but with 0 it would look a trifle better:

$$\sum_{i_{1\ldots k}=0}^{i_{1\ldots k}\le n-1 \,\wedge\, \sum_{j=1}^{k} i_j = s} (\prod_{j=1}^{k} \binom{n-1}{i_j}) \tag{7}$$

(Caution: In the next steps in between, the variable names are not always the same as before)

This we can transform with a generalization of the Vandermonde's identity.

Vandermonde's identity : $\sum_{j} \binom{m}{j} * \binom{n-m}{k-j} = \binom{n}{k}$ (8)

Generalization: $\sum_{k_{1\ldots y}=0}^{k_{1\ldots y}<=n} \binom{n}{k_1} * \binom{n}{k_2} * \binom{n}{k_3} *\ldots.* \binom{n}{x-\sum_{j=1}^{y} k_j} = \binom{(y+1)*n}{x}$ (9)

(generalization of Vandermonde's identity is made similar to its algebraic proof (2008), only with *k+1* polynomials instead of 2)

Here is to see that the conditions

- The sum of all lower parts of the binomial coefficients must be *x* (because of the last term with $x-\sum_{j=1}^{y} k_j$)
- if a lower part of the binomial coefficients is larger than *n* or lower than zero then the factor is zero and the corresponding addend is also zero (this holds also for $x-\sum_{j=1}^{y} k_j$)

With that we can transform (7) into $\binom{k*(n-1)}{s}$. This would be much faster for computation, but we have had (6), not (7).

$\sum_{i_{1\ldots k}=0}^{i_{1\ldots k}\le n-1 \,\wedge\, \sum_{j=1}^{k} i_j = s} ((\prod_{j=1}^{k} \binom{n-1}{i_j})$ *(7)* $>$ $\sum_{i_{1\ldots k}=1}^{i_{1\ldots k}\le n-1 \,\wedge\, \sum_{j=1}^{k} i_j = s} ((\prod_{j=1}^{k} \binom{n-1}{i_j})$ *(6)* so we can use the first formula and subtract the difference to the second.

The difference is somehow a similar problem because the lower part of the last binomial-coefficient i_k is zero if the sum of $i_{1\ldots k-1}$ is *s* (important in (9)). This is the same problem as before but this time with *k-1*.

However, the difference (7) - (6) is not: $k * \sum_{i_{1\ldots k-1}=0}^{i_{1\ldots k-1}\le n-1 \,\wedge\, \sum_{j=1}^{k-1} i_j = s} ((\prod_{j=1}^{k} \binom{n-1}{i_j})$ (10)

because some combinations will then be counted more than once.

To illustrate this:

We have the sum for every combination of *i* (the components of vector *i* are every combination of one value per column):

0	0	0	0	0
1	1	1	1	1
...	...	...	...	...
n-1	n-1	n-1	n-1	n-1

but we need only:

1	1	1	1	1
...	...	...	...	...
n-1	n-1	n-1	n-1	n-1

(Note that the constraint of the fixed sum of all components of *i* is fulfilled with a term $s - \sum_{j=1}^{k-1} k_j$, so the tables above have *k-1* colums)

Therefore, we have to substract every combination of *i* with at least one component is zero.

However, the result of $k * \sum_{i_{1\ldots k-1}=0}^{i_{1\ldots k-1}\le n-1 \wedge \sum_{j=1}^{k-1} i_j = s} ((\prod_{j=1}^{k} \binom{n-1}{i_j})$ *(10)* would be:

0	0	0	0	0
	1	1	1	1
	...	...	...	...
	n-1	n-1	n-1	n-1

0	0	0	0	0
1		1	1	1
...		...	...	...
n-1		n-1	n-1	n-1

0	0	0	0	0
1	1		1	1
...	...		...	...
n-1	n-1		n-1	n-1

.....

Then we count every combination of two times 0 more than once, so we have add again these combinations:

0	0	0	0	0
		1	1	1
		...	...	...
		n-1	n-1	n-1

0	0	0	0	0
	1		1	1
	...		...	...
	n-1		n-1	n-1

0	0	0	0	0
	1	1		1
	...	...		...
	n-1	n-1		n-1

.....

Here we have once again the similar problem with *k-2*, now $\binom{k}{2}$ times. If we go further and further with alternating add and subtract we end at $\binom{k}{k}$ where all combinations are counted properly ->

$$\sum_{i_{1\ldots k}=1}^{i_{1\ldots k}\le n-1 \wedge \sum_{j=1}^{k} i_j = s} ((\prod_{j=1}^{k} \binom{n-1}{i_j}) \qquad (6)$$

=

$$\binom{k*(n-1)}{s} + \sum_{j=1}^{k-1} (-1)^j * \binom{k}{j} * \binom{(k-j)*(n-1)}{s}) \qquad (11)$$

(11) combined with (5) results in:

$$\frac{\sum_{k=z}^{\min(x;y)} \binom{y}{k} * \sum_{s=k}^{x} \left(\binom{k*(n-1)}{s} + \sum_{j=1}^{k-1} (-1)^j * \binom{k}{j} * \binom{(k-j)*(n-1)}{s} \right) * \binom{(n-\min(x;y))*(n-1)}{x-s}}{\binom{n*(n-1)}{x}} \tag{12}$$

with the order O(*min(x,y)**x^2) and because the numbers are growing larger (but still needed to be computed as full integers) with a little term to the power of 4.

Because of the long drawn out derivation, the formula (12) was tested against (4) in a not too time-consuming parameter range (*N* < 20 and *y* < 20). The formula (4) again was tested against the full amount of permutations, but only *N* < 5 was feasible.

For the author of this derivation it seems remarkable that the constraints (which are indirect time-consuming) e.g. $i(1)...(k) \le n-1 \wedge \sum_{j=1}^{k} i(j) = s$ can be led back to the constraint of factorial(lower than 0) = 0.

3. Discussion

Even though the order of O(x^3) does not look so bad, there should be a faster approximate formula with a guaranteed bound of error. However, here the motivation was 'let us try to see if it is possible to solve it exactly'.
The first occurrence/use of this type of problem was in gene expression data (as far as we know). Assume we have n elements which are somehow (but unknown) regulating each other. So there are n*n possible regulations or n*(n-1) regulations without elements regulating itself. Of this n*(n-1) possible regulations we could take x regulations for some reason (e.g. it is the result of a method to filter all possible regulations). Additionally given is a list of y elements, which are known as regulators. (if we annotate the regulations as lines of 'A -> B' then the regulators are just the elements on the left side).
So we can search in (the first column of) the x chosen regulations for this y elements and we will find a count of z elements.
Now the question arises, how likely is this? Alternatively, the other way round, is it just by chance to get z of y elements in our filtered regulations?

Such a scenario is not quite frequent and it sounds strange to know that something is a regulator without knowing what is exactly regulated by it. Nevertheless, it is possible to know that a gene very likely is a transcription factor, even if it is not quite sure which other genes are exactly regulated of it.

References

(2008). "Vandermonde's identity." Retrieved 24.11., 2008, from http://en.wikipedia.org/wiki/Vandermonde's_identity#Algebraic_proof.

5. Classification methods and metabolomics

As mentioned in the introduction, one major challenge in biological data is their frequent p>>n shape, which means that more variables exist than samples. The classical data analysis is built on p < n problems and assumes that it should always be simpler to make the same measurement more times than to measure something different. While biology followed a different path with various multiple test chips like microarrays, with metabolites the variables are in most cases not more than the samples: only a few hundred can be distinguished and a few hundred samples are feasible. This makes this data source treatable by standard methods. Of course, standard methods were also further developed and today machine learning also falls into this category. The paper discussed in this chapter uses machine learning to infer models for kidney tumors.

5.1 Classical classification

The more classical approach for finding classes is to look at the distributions of the variables of the different classes, find and define the distribution type and select the threshold between the classes. If a linear combination is found to separate two classes, this leads to the Linear Discriminant Analysis [100], which is closely related to the simple machine learning method perceptron [101]. The two methods are not the identical due to their slightly different error function [102], but both result in a linear discriminant function. If the data records are sufficient for an exhaustive search of all variable combinations and the assumption of linearity is feasible, these classical approaches will find the optimum. The next section looks at a more elaborate search in the combinatorial space.

5.2 Machine learning for classification

Although machine learning is a diversified field, it contains the largest collection of classification methods, as machine learning is focused on known properties learned from training data. Neighboring areas are data mining, which focuses on discovery of (previously) unknown properties, analytics when an analytical solution is feasible, and a few others (for a general broad view see [103]).

The simplest case for classification is when numeric variables and two target classes are given; here most methods are available [63]. A few concepts for this type of problem are described in the following paragraphs.

Percent correctly classified (PCC): 100 * correctly classified instances / all instances. This is the simplest key figure for a classification model.

Lower border / zeroR: this is not a standard term, likely because it is too trivial. In the manuscript below it is termed 'lower border', in other literature and in one of the main machine learning tools WEKA [104] it is called 'zeroR' (for zero ratio). It is the ratio of the most frequent class, following the idea that a model can achieve this merely by always predicting this class. Therefore, the PCC rate of the model should exceed the lower border in order to be considered better than this overly simplistic prediction.

Information entropy: the amount of information which is encoded in a certain sequence, see [53].

Information gain: the change in information entropy, usually the difference of two models in this respect. When the model is build stepwise, the model before and after one step is compared.

Masking: whenever a classifier is built stepwise, there is the characteristic that variables chosen first make the later choice of correlated variables much less likely, because they are usually chosen for their information gain to the model to that point.

Cross-validation [55]: typically it is not only how well a model describes the data that is interesting but, because of possible overfitting, it is also interesting how well the model performs on independent test data. The simplest way to obtain this information is to divide the data into two parts: a training set and a test set. The training set is used to construct the model, the test set to validate its performance. With few samples, the size of the test set can be set smaller, but this would increase the noise level. In this case, it is done multiple times: the samples are divided into n parts, n-1 are used to construct a model, and the remainder is used for validation. This is done n times. In theory, the average PCC should increase with an increasing n, following the idea that more data should make the model more general and more likely to be correct on new data. In the usual case, and apart from the noise, the PCC approaches a certain value asymptotically, which is said to be the maximum possible achievable with the full data. It is a generally accepted carelessness in machine learning/classification that a model is constructed with all data, but the PCC is from 10-fold cross-validation. The paper below follows this standard procedure. The PCC of n-fold cross-validation, where n is variable, also allows us to see if the amount of data is enough. A steep curve until the maximal n (maximal = number of samples) indicates that the data is not enough for the best model. The curves need not be the same in absolute values, slope and bend for different methods.

5.3 Article: **Metabolic profiling reveals key metabolic features of renal cell carcinoma**

Gareth Catchpole*, **Alexander Platzer***, Cornelia Weikert, Carsten Kempkensteffen, Manfred Johannsen, Hans Krause, Klaus Jung, Kurt Miller, Lothar Willmitzer, Joachim Selbig, Steffen Weikert. *Metabolic profiling reveals key metabolic features of renal cell carcinoma*. **J. Cell. Mol. Med. Vol 15, No 1, 2011 pp. 109-118**

This manuscript is a relatively straightforward analysis, including standard statistics and machine learning. This was not common in medicine at time of publication and might not be standard today; nevertheless, this manuscript can be used as a template for similar analysis in the future.

OWN CONTRIBUTION IN [105]

AP created the decision tree models, contributed to the other data analysis and to writing the manuscript.

J. Cell. Mol. Med. Vol 15, No 1, 2011 pp. 109-118

Metabolic profiling reveals key metabolic features of renal cell carcinoma

Gareth Catchpole [a, #], Alexander Platzer [b, #], Cornelia Weikert [c], Carsten Kempkensteffen [d], Manfred Johannsen [d], Hans Krause [d], Klaus Jung [d], Kurt Miller [d], Lothar Willmitzer [a], Joachim Selbig [b], Steffen Weikert [d, *]

[a] *Department of Central Metabolism, Max-Planck-Institute of Molecular Plant Physiology, Golm, Germany*
[b] *Department of Bioinformatics, University of Potsdam, Potsdam, Germany*
[c] *Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany*
[d] *Department of Urology, Charité-University Medicine Berlin, Berlin, Germany*

Received: May 23, 2009; Accepted: October 5, 2009

Abstract

Recent evidence suggests that metabolic changes play a pivotal role in the biology of cancer and in particular renal cell carcinoma (RCC). Here, a global metabolite profiling approach was applied to characterize the metabolite pool of RCC and normal renal tissue. Advanced decision tree models were applied to characterize the metabolic signature of RCC and to explore features of metastasized tumours. The findings were validated in a second independent dataset. Vitamin E derivates and metabolites of glucose, fatty acid, and inositol phosphate metabolism determined the metabolic profile of RCC. α-tocopherol, hippuric acid, myoinositol, fructose-1-phosphate and glucose-1-phosphate contributed most to the tumour/normal discrimination and all showed pronounced concentration changes in RCC. The identified metabolic profile was characterized by a low recognition error of only 5% for tumour *versus* normal samples. Data on metastasized tumours suggested a key role for metabolic pathways involving arachidonic acid, free fatty acids, proline, uracil and the tricarboxylic acid cycle. These results illustrate the potential of mass spectroscopy based metabolomics in conjunction with sophisticated data analysis methods to uncover the metabolic phenotype of cancer. Differentially regulated metabolites, such as vitamin E compounds, hippuric acid and myoinositol, provide leads for the characterization of novel pathways in RCC.

Keywords: kidney cancer • metabolism • metabolomics • metastasis

Introduction

The metabolite pool of cells and tissues represents the end result of metabolism determined by genetic, environmental and nutritional factors. The metabolic profile of biological systems is closely related to the individual phenotype and reflects the biological endpoint of a multitude of pathways and their interaction with any confounding stimuli. Cancer cells exhibit activation of specific metabolic pathways to compensate for their extremely high energy demands. Indeed increased glucose uptake and lactate production and decreased respiration are key phenomena of tumour cell metabolism. In particular, the generation of an acidic microenvironment through increased lactate production, even under aerobic conditions, may confer extracellular matrix degeneration and exert toxic effects on surrounding cell populations, while being harmless for the cancer cell itself [1]. Thus, the metabolic adaptations may indeed be critical for the development of accelerated proliferation and the invasive growth of tumour cell populations [1, 2]. The molecular mechanisms underlying the metabolic hallmarks of cancer are still poorly understood, although genetic, epigenetic and environmental factors driving cancer development and progression will interact to determine the metabolic phenotype of tumour cells. Recent studies suggest that metabolic changes play a pivotal role in the biology of renal cell carcinoma (RCC) – a tumour entity that is largely resistant to conventional chemo- and radiotherapy. The metabolic profile of renal tumours may thus serve as a reliable biomarker of malignant transformation and biological behaviour.

[#] These authors contributed equally to this publication.

*Correspondence to: Steffen WEIKERT, M.D.,
Charité-Universitätsmedizin Berlin,
Hindenburgdamm 30, D-12200 Berlin, Germany.
Tel.: +49-30-8445-2577
Fax: +49-30-8445-4448
E-mail: steffen.weikert@charite.de

doi:10.1111/j.1582-4934.2009.00939.x

© 2011 The Author
Journal of Cellular and Molecular Medicine © 2011 Foundation for Cellular and Molecular Medicine/Blackwell Publishing Ltd

Recent advances in metabolic profiling technologies by providing quantitative measures of metabolite profiles from gas chromatography time-of-flight mass spectrometry (GC-TOF-MS) based technology present the opportunity to apply this technique in human specimens [3–5]. Global metabolic profiling has emerged as a promising approach to characterize the metabolite pool within a cell, tissue or bodily fluid under certain conditions, such as health or disease status [4, 6, 7]. Metabolic profiling is applied to monitor the health to disease continuum and has the potential of increasing our understanding of the mechanisms of disease [8]. Thus the characterization of the metabolic features in tumours is expected to provide a better understanding of the mechanisms of malignant transformation and progression and may lead to the identification of metabolic biomarkers for cancer detection and prognostication. However, comparative profiling of low molecular weight compounds, such as sugars, lipids and amino acids, in cancer as compared to the corresponding normal tissue is a rather unexplored area. The objective of this study was to characterize the key metabolic features of RCC using GC-TOF-MS and mutual information as well as decision tree-based data analysis.

Material and methods

Study population and sample collection

Tumour tissue and specimens of normal renal cortex tissue were collected from patients undergoing surgical treatment for primary RCC at the Department of Urology, Charité-University Medicine Berlin between November 1995 and November 2005. They included 29 female and 67 male patients with a mean age of 62 years (range 36–87). Their use was approved by the Ethics Committee of the Free University of Berlin, and all patients gave their informed consent prior to surgery. Tissue samples were obtained during radical nephrectomy following a standard operating procedure. All RCC specimens were derived from primary tumours. Tissue specimens were dissected in the operating room immediately after removal of the kidney, snap-frozen in liquid nitrogen and stored at −80°C until use. RCC samples were serially sectioned before further processing. Additional sections were stained with haematoxylin–eosin for histopathological evaluation. The histopathological classification and staging was based on the Q2 1997 World Health Organization and TNM classification guidelines (International Union Against Cancer, 1997): pT1 ($n = 53$), pT2 ($n = 13$), pT3 ($n = 30$); M0 ($n = 87$), M1 ($n = 9$). Primary tumour tissue samples and normal tissue samples from 57 patients (39 male; 18 female) were chosen for the first round of metabolic analyses. Tumour characteristics for these RCC patients were: pT1 ($n = 30$), pT2 ($n = 12$), pT3 ($n = 15$), G1 or 2 ($n = 36$) and G3 ($n = 21$). Of these, 36 patients had localized tumours and 21 had or developed metastasized RCC. Later, a second set of samples was put together from 39 patients (29 male; 10 female; RCC: $n = 39$; normal tissue: $n = 27$) for validation purposes. Tumour characteristics were: pT1 ($n = 16$), pT2 ($n = 8$), pT3 ($n = 15$), G1 or 2 ($n = 22$), G3 ($n = 17$), localized RCC ($n = 32$) and metastasized RCC ($n = 7$). Most of the tumour samples analysed belonged to the clear cell subtype of RCC ($n = 54$ in the first set; $n = 34$ in the second set).

Sample preparation and GC-TOF-MS analysis

Frozen biopsy tissue was processed under standard operating procedures. Samples were serially sectioned in a cryostat microtome to prevent thawing. A defined amount (30 mg) of sectioned tissue was then transferred to a 2 ml centrifuge tube and homogenized. Samples were centrifuged at 14,000 rpm for 2 min. and the supernatant taken and dried to complete dryness in a rotary evaporator in the glass vials used for GC-MS analysis.

GC-TOF-MS metabolite profiling was performed on a Leco Pegasus 3 time-of-flight mass spectrometer (Leco, St. Joseph, MI, USA) equipped with a Direct Thermal Desorption injector (ATAS GL International, The Netherlands) coupled to an HP 5890 gas chromatograph and a dual-arm autosampler with automatic derivatization and liner exchange. This eliminates both the impact of potential degradation/synthesis artefacts and sample carry-over and means that no phase separation of samples is necessary, thereby broadening the coverage of the profiling technique to non-polar compounds. The method allows relative quantification of metabolites which cover a large part of primary metabolism such as sugars, organic acids, amino acids and alcohols in addition to sterols and free lipids. Samples were derivatized in 10 μl methoxyamine hydrochloride in N, N-dimethylformamide diethyl acetal (40 mg/ml) at 42°C for 180 min. followed by 90 μl N-methyl-N-trimethylsilyltrifluoroacetamide at 37°C for 30 min.

A total of 1.5 μl samples were injected in splitless mode at 85°C, ramping to 290°C at 4°C/sec. The GC used a constant flow of 2 ml/min. helium as carrier gas and a 30 m 320 μm ID MDN35 column. The column temperature gradient was held at 85°C for 210 sec., followed by a linear gradient of 15°C/min. reaching a target temperature of 360°C. A 230-sec. acquisition delay was used and spectra subsequently acquired at the rate of 20/sec.

Chromatograms were processed using Leco ChromaTOF software (version 3.25) and peaks with a signal to noise ratio >10 were exported before using an algorithm developed in-house for dealing with the output.txt files [9]. Mass spectra were compared to an in-house mass spectral library for metabolite identification and peak heights expressed relative to an internal standard (^{13}C sorbitol-D).

Statistical analysis

In a univariate approach the non-parametric Mann-Whitney *U*-test was applied to search for significant differences in relative concentrations of metabolites between RCC and normal tissue samples, and between localized and metastasized primary tumours. For key metabolites associations of relative concentrations with tumour stage or grade were explored.

Metabolite profile data were normalized to an internal standard, log transformed and scaled according to [9]. Metabolites with more than 20% missing values were excluded and remaining missing values were estimated *via* BPCA using the R package pcaMethods [10]. Differences were expressed as median fold change and *P*-values Bonferroni corrected to address the problem of multiple testing. These pairwise comparisons were restricted to all metabolites that could be identified based on comparison to the mass spectral library.

For multivariate supervised classification, all metabolites, irrespective of their identified/non-identified status, were initially included. Data were normalized to the internal standard and any variables containing missing values were excluded. As an initial step the first dataset was used to determine metabolic signature differences in a two-group scenario between tumour present/absent groups. Subsequently this was expanded to a three-group scenario, in which tumour presence was further sub-divided into metastasized/non-metastasized.

Table 1 Performance of the decision tree models for the discrimination between normal and RCC tumour samples

	All metabolites				Identified metabolites only			
	First dataset		Validation dataset		First dataset		Validation dataset	
Method*	50%	crossfold10	50%	crossfold10	50%	crossfold10	50%	crossfold10
Random Forest	86%	94%	74%	71%	91%	91%	50%	78%
Random Tree	52%	71%	82%	65%	71%	70%	56%	66%
ADTree	95%	95%	68%	75%	89%	92%	47%	73%
SMO	92%	97%	85%	88%	95%	92%	59%	81%
Simple Logistic	95%	95%	91%	84%	94%	85%	56%	85%
lower border**								
most frequent class/total	66/132		40/68		65/130		39/67	
float of lower border	50%		59%		50%		58%	

*The correct classification returned by each of the five different classification methods used (random forest, random tree, ADTree, SMO and simple logistic) upon treating the two-class problem (tumour yes/no) is shown as percentage. Results are shown having divided the dataset into 50% training and 50% testing sub-groups and having used 10-fold cross-validation.

**The lower border is the percentage of the total sample number represented by the most numerous class. This percentage correct classification could therefore be achieved simply by always classifying unknown samples as belonging to this class. Therefore the success rate of the models should exceed the lower border in order to be considered better than this overly simplistic selection method.

A number of different mainly decision tree classification algorithms, available in the WEKA platform [11], was used (random forest, random tree, alternating decision tree (ADTree), sequential minimal optimization, simple logistic and C4.5).

Data were segregated into learning (50%) and testing (50%) subsets and models were validated using 10-fold cross-validation. As a second step, the most promising model was then further validated using the fully independent second dataset. In a further exploratory analysis, metabolites that contributed most to the classification of localized *versus* metastasized were tested as predictors of recurrence-free survival in Cox regression analyses. For further information the maximal information gains for a decision in the classification of tumour presence/absence and metastasized/non-metastasized were calculated for each single metabolite using mutual information. As decision tree methods generate minimal classification models metabolites with a high informational gain are not necessarily all contained in the decision tree models.

Results

Comparative metabolic profiling of RCC and normal renal tissue

In the first round of analyses, RCC tissue samples and control cortex specimens from 57 patients with RCC were investigated. The cohort consisted of 36 patients with localized disease and 21 patients with metastatic tumours either at time of diagnosis or who developed metastasis during follow-up. The mean follow-up was 41 months (range 2–113 months). Data matrices consisted of 188 metabolites, of which 74 could be identified, 25 putatively identified, 37 which could be assigned a possible metabolite class and 52 whose chemical structure remained unassigned.

All generated classification models describing tumour presence/absence performed satisfactorily on the first dataset. Alternating decision trees (ADTree) are preferred due to the fewer variables being necessary to yield the high prediction power of 95% correct assignment (Table 1, Fig. 1). Thus this model was selected for validation using the second dataset, where it resulted in 77% correct classification. This is less than the 95% observed for the first dataset, but is however, very similar to the classification reached within the second dataset itself (75% with the ADTree with 10-fold cross-validation).

Tree-based learning algorithms in particular, are designed to the selection of the single best classifier at each decision step and therefore occasionally prone to the exclusion of others which could themselves carry potentially meaningful information (Fig. 2). We therefore tested for differences between normal und RCC specimens by pairwise comparison of the relative concentration of all identifiable metabolites. These analyses were confined to all metabolites detected in >80% of samples. Metabolites with differences in relative concentrations are shown in Table 2. These include metabolites which may not have been detected in all samples and thus may not be contained in the decision tree models, which are intolerant of missing values. These compounds were subsequently assigned to common

© 2011 The Author

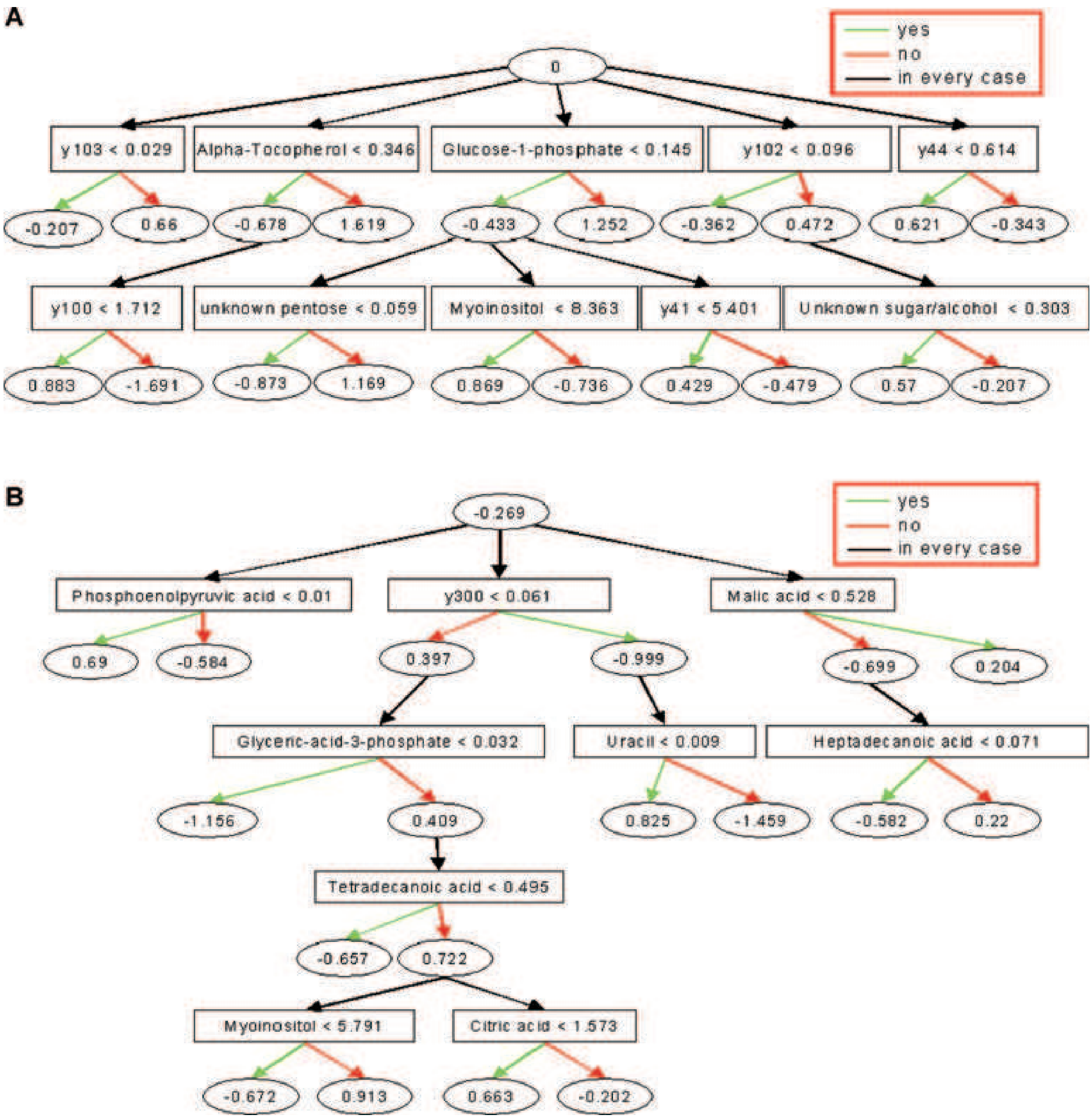


Fig. 1 Decision Tree Model (ADTree) generated for the two-class problem of discriminating RCC and normal renal tissue samples (**A**), and localized RCC and metastatic disease (**B**). Key metabolites are shown with the corresponding normalized relative peak intensity cut-offs. Each metabolite resembles a decision node that is linked to two prediction nodes with the corresponding prediction values. Classification of a hypothetical sample would be based on the sum of final attained prediction node values that are determined by applying the peak intensity cut-offs for all metabolites of the decision tree on the sample-specific data record. Any result < 0 means a class prediction of 0 (**A**: normal tissue; **B**: localized tumour), any result > 0 a class prediction of 1 (**A**: RCC, **B**: metastatic tumour). The model was trained with the first dataset and used all metabolites irrespective of identified status.

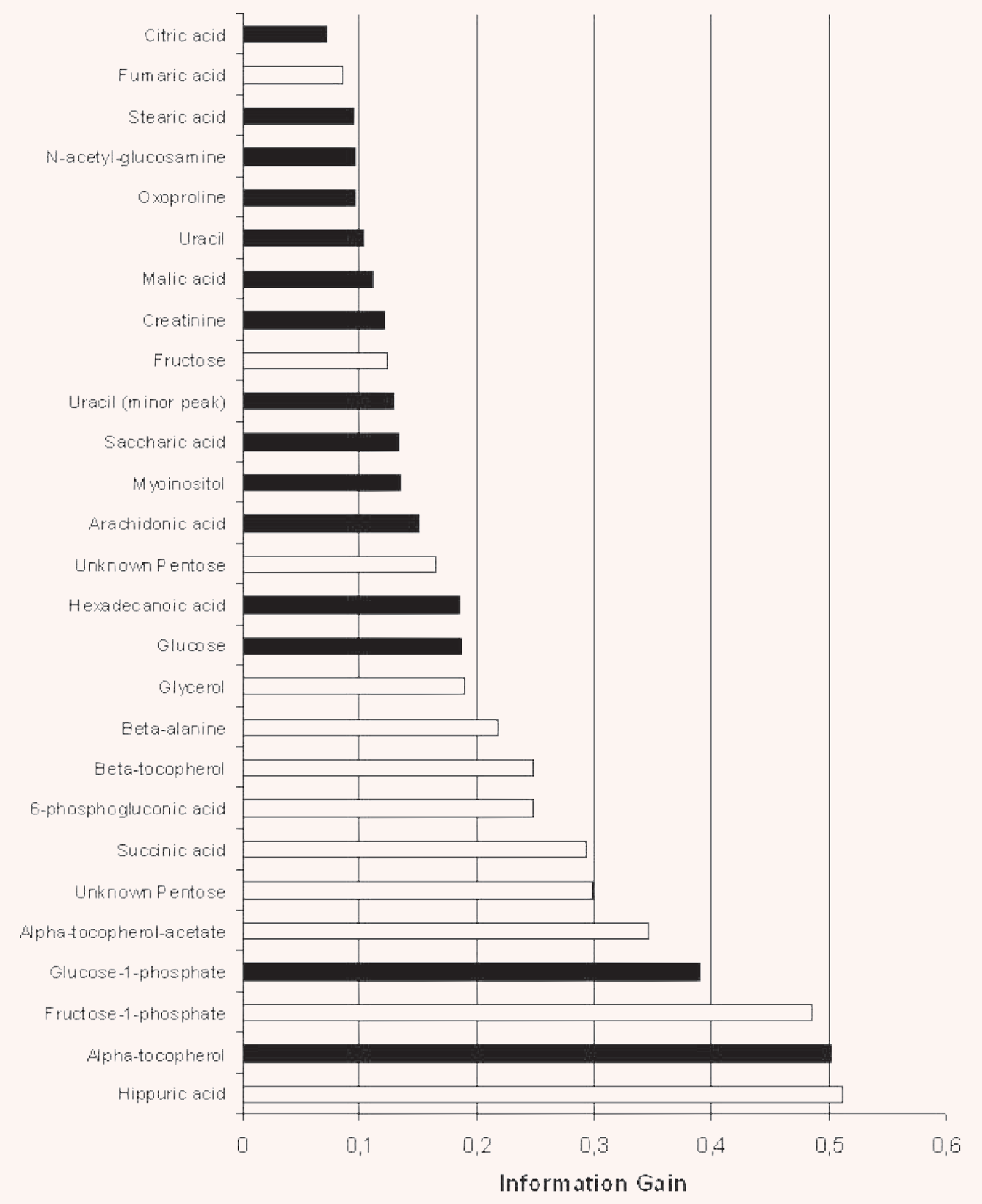


Fig. 2 The information gain for the two class discrimination between RCC and normal tissue by key metabolites. Metabolites with the highest gain contribute most to the correct discrimination. The theoretical maximum gain = 1. The black bars indicate metabolites that were not detectable in all samples and were therefore unable to be incorporated into the ADTree model, but all of these metabolites were detected in over 90% of samples, except for 6-phosphogluconic acid (88%).

pathways according to the Kyoto Encyclopedia of Genes and Genomes. The data indicate that the metabolic signature of RCC tends to be characterized by metabolites associated with glucose metabolism, such as glucose-1-phosphate, markers of fatty acid and phospholipids metabolism, such as palmitate, arachidonic acid and glycerol, and myoinositol belonging to the inositol polyphosphate family of cell signalling molecules. Interestingly, and consistent with the decision tree models, the metabolites revealing the largest relative RCC *versus* control differences were α-tocopherol and hippuric acid. Elevated levels of α-tocopherol were detected in RCC thereby pointing to a potential activation of vitamin E metabolism in tumour cells. When α-tocopherol was considered alone in a ROC analysis, correct classification of 84.8% of RCC samples and 92.4% of normal tissue samples was achieved (data not shown). Although similar accuracy could be achieved using hippuric acid as a marker, the relevance of the greatly decreased concentration of this metabolite in RCC is unknown. The descriptive statistics for selected metabolites are shown in Fig. 3. In further data exploration we tried to see whether or not metastasizing tumours could be

© 2011 The Author

Table 2 Metabolites displaying relative concentration differences in RCC and control renal tissue samples

Compound	**Median fold change***		***P*-value****		**Pathway**
		Training set	**Validation set**	**Combined set**	
α-tocopherol	5.2	<0.0007	<0.0007	<0.0007	Vitamin E metabolism
α-tocopherol acetate	4.0	<0.0007	n.s.	<0.0007	
β-tocopherol	3.1	<0.0007	0.004	<0.0007	
Arachidonic acid	–2.6	<0.0007	<0.0007	<0.0007	Arachidonic acid metabolism (involved in VEGF signalling pathway and angiogenesis)
Palmitate	–1.5	<0.0007	0.02	<0.0007	Fatty acid metabolism
Tridecanoic acid	–1.4	<0.0007	n.s.	0.0032	
Glycerol	–2.2	<0.0007	0.0008	<0.0007	Glycerolipid metabolism
Citric acid	1.6	0.001	n.s.	<0.0007	TCA cycle
Fumaric acid	–1.9	0.01	<0.0007	<0.0007	
Succinic acid	–3.4	<0.0007	0.003	<0.0007	
Malic acid	–1.7	<0.0007	0.01	<0.0007	
Glucose	5.0	<0.0007	n.s.	0.0008	Glycolysis, Pentose phosphate pathway
Glucose (minor peak)	4.8	<0.0007	n.s.	<0.0007	
Glucose-1-phosphate	3.0	<0.0007	n.s.	<0.0007	Glycolysis, Pentose phosphate pathway, Nucleotide sugars metabolism,
6-phosphogluconic acid	6.3	<0.0007	n.s.	<0.0007	Glycolysis, Pentose phosphate pathway, byproduct of tyrosine kinase acticity
Fructose	2.0	<0.0007	n.s.	<0.0007	Fructose and mannose metabolism
Fructose-1-phosphate	8.3	<0.0007	n.s.	<0.0007	
myo-Inositol	–1.5	<0.0007	<0.0007	<0.0007	Phosphatidylinositol signalling system, Inositol phosphate metabolism
Saccharic acid	–2.6	<0.0007	n.s.	<0.0007	Ascorbate and aldarate metabolism (linked to glycolysis)
N-Acetyl-D-glucosamine	–1.7	0.034	<0.0007	<0.0007	Glutamate metabolism, Aminosugars metabolism
β-alanine	2.5	<0.0007	n.s.	<0.0007	Pyrimidine metabolism
Uracil	–2.0	<0.0007	0.004	<0.0007	
Uracil (second peak)	–3.7	<0.0007	0.002	<0.0007	
Hippuric acid	–35.2	<0.0007	<0.0007	<0.0007	Phenylalanine metabolism
Oxoproline	1.4	<0.0007	n.s.	0.0037	Gluthathion metabolism (radical detoxification)

*Negative fold change indicates decreased relative concentration in RCC *versus* normal tissue.
**A *P*-value of <0.0007 indicates a significant difference upon Bonferroni correction for multiple testing.

differentiated from localized ones. Applying an ADTree model did not yield satisfactory results, largely due to the restrictive sample number, however a direct pairwise comparison of the relative metabolite concentrations in both tumour phenotypes suggested a number of differences (Table 3).

Independent validation of the RCC metabolic signature

Despite the reduced statistical power in this smaller second dataset, we repeated pairwise comparisons of relative concentrations of

© 2011 The Author
Journal of Cellular and Molecular Medicine © 2011 Foundation for Cellular and Molecular Medicine/Blackwell Publishing Ltd

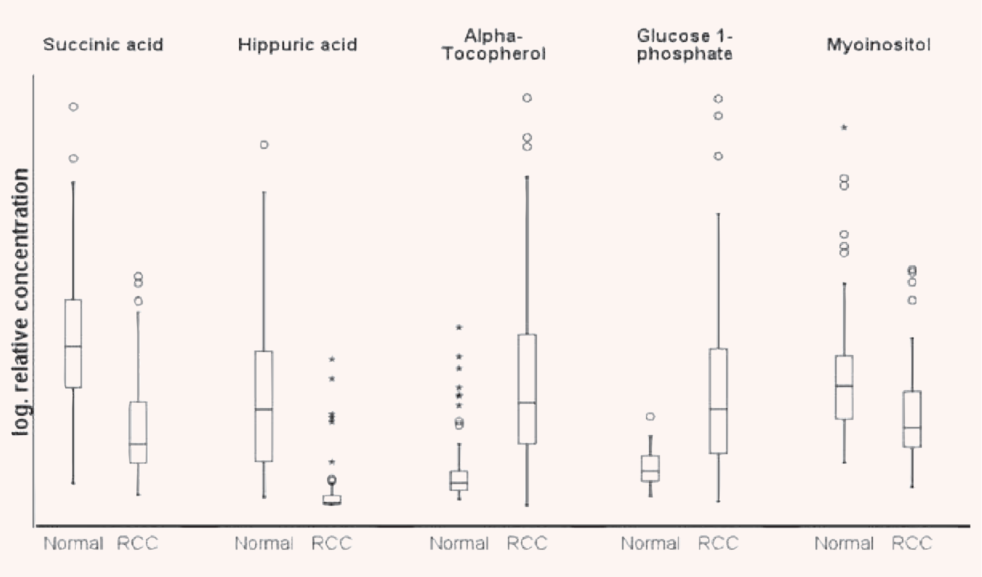


Fig. 3 Descriptive statistics of relative metabolite concentrations in tumour *versus* normal tissue. Select key metabolites are chosen based on their high informational gain for the tumour/normal discrimination and/or their identification in the decision tree analysis. Boxplots show median, 25th and 75th percentiles, range, and extreme values. For better illustration a logarithmic scale was chosen for the relative concentration; absolute concentrations cannot be calculated and therefore no precise scale is given.

Table 3 Metabolites with relative concentration differences in localized and metastasized RCC samples

Compound	Median fold change	*P*-value	Pathway
Uracil	1.9	<0.0007	Pyrimidine metabolism
Arachidonic acid	1.9	0.007	Arachidonic acid metabolism (involved in VEGF signalling pathway and angiogenesis)
Erythritol	1.7	0.002	Glycerolipid metabolism
3-Phospho-glycerate	1.9	0.005	
Heptadecanoic acid	1.5	0.001	Fatty acid metabolism
Hexadecanoic acid	1.3	0.008	
Tetradecanoic acid	1.4	0.01	
Isoleucine	2.9	0.008	Valine, leucine and isoleucine meatbolism
Phenylalanine	2.4	0.003	Phenylalanine metabolism
Proline	2.5	0.006	Arginine and Proline metabolism

*A *P*-value of <0.0007 indicates a significant difference upon Bonferroni correction for multiple testing.

known and interpretable metabolites to validate the findings of the first dataset (Table 2). As expected due to the smaller sample size, significant differences were seen in fewer metabolites in the second dataset. However, all differences observed in the first dataset were confirmed when the tests were repeated in the combined first and second dataset (Table 2). Key metabolites in both datasets were α-tocopherol, hippuric acid and myoinositol thus underlining the notion that these are of importance for the metabolic signature of RCC. The comparatively low number of metastasized samples ($n = 7$) in the second dataset hampered the validation of metabolic differences between these and non-metastasized tumours. Pairwise comparisons in the combined dataset however, confirmed uracil as a key metabolite in distinguishing metastasized and localized tumours. This metabolite is of relevance for the synthesis of nucleic acids and may indicate a metabolic adaptation to the increased transcriptional activity in aggressive, potentially lethal tumours. The increased fatty acid content adds weight to the theory that fatty acid degradation is reduced in tumour cells, but this may be particularly pronounced in aggressive metastasized tumours. The exploratory analysis revealed some other putative metabolites which characterize metastasized disease, such as myoinositol, arachidonic acid and several amino acids (isoleucine, phenylalanine and proline), but these findings require confirmation in independent datasets as false positive test results cannot be fully excluded. When differences in relative concentrations for key metabolites, *i.e.* α-tocopherol, hippuric acid, glucose-1-phosphate, myoinositol and succinic acid, by tumour stage (pT1–2 *versus* pT3) were explored α-tocopherol was increased in pT3 tumours ($P < 0.05$). No differences were observed by tumour grade (G1–2 *versus* G3). The ADTree models performed insufficiently when tumour stage and grade were studied as classifiers. The results were indicative of α-tocopherol, free fatty acids and uracil contributing to the metabolic signature of advanced (pT3) as compared to smaller tumours (pT1–2).

In addition we tested whether metabolites of the ADTree classifying metastasized tumours were associated with the outcome of RCC patients using univariate and multivariate Cox models. Only citric acid was independently related to recurrence-free survival (data not shown). Decreased citric acid concentrations could conceivably indicate a deteriorating prognosis and although this finding is in line with a switch towards increased glycolysis even under aerobic conditions and therefore seems plausible, the data are purely exploratory and, in view of the multiple testing problem, require confirmation.

Discussion

This study characterizes the metabolite pool of RCC as compared to control renal cortex tissue using non-targeted metabolic profiling and permitted the assignment of a specific metabolic signature to RCC. This signature was not only validated with common test procedures, but was also confirmed in an independent, subsequently compiled validation dataset. Thus a set of key metabolites representing relevant metabolic pathways of RCC was established. Our data together with a previous study [12] substantially extend the knowledge on the small molecule component of RCC tissue. These findings complement earlier studies on biomarker discovery in RCC using 'omics' platforms [13–15].

As the metabolomics methodology used in this study captures a large part of primary metabolism, our study for the first time gives a comprehensive overview of the metabolic phenotype of RCC tissue. This phenotype confirms presumed metabolic features of cancer cells in general and RCC in particular. The marked differential concentration of glucose 1-phosphate and metabolites of the tricarboxylic acid (TCA) cycle, such as succinate and malate, points to a pivotal role of altered glucose and energy metabolism in RCC. Remarkably, most substrates of the TCA cycle seemed to be notably down-regulated in RCC compared to control tissue. Since the TCA pathway is a catabolic pathway of aerobic respiration our findings may reflect the shift towards an anaerobic energy metabolism and reduced respiration even in the presence of oxygen, also referred to as aerobic glycolysis or as the Warburg effect [16]. Indeed, recent studies suggest that the up-regulation of hypoxia-inducible factors (HIF) mediates the reprogramming of glucose and energy metabolism including increased glycolysis and lactate production in renal cancer cells [17, 18]. Using a combination of transcriptomics and proteomics it has been recently confirmed that genes and proteins involved in cellular metabolism play a crucial part in the development and progression of RCC making them promising candidates for biomarker identification [15].

To compensate for their high energy demands, cancer cells are likely to exploit a multitude of energy sources including fatty acid oxidation and other non-glycolytic pathways [19, 20]. According to our findings, metabolites of fatty acid metabolism seem to play a key part in RCC metabolism. A number of fatty acids were found to be differentially concentrated, but uniformly down-regulated in RCC. This finding may be the consequence of increased fatty acid oxidation, which has also been described in other cancer types, in particular prostate cancer [21, 22]. Studies identifying fatty acid binding proteins (FABP) [23, 24] and fatty acid synthase [25] as tumour markers of RCC underline the importance of fatty acid metabolism in the biology of RCC. Interestingly, in our study up-regulation of fatty acid concentration seemed to be specifically associated with metastatic disease. This fact may indicate that an increase in de-novo fatty acid synthesis or increased fatty acid uptake and reduced mitochondrial β-oxidation of fatty acids may be rather late events in the progression of RCC to an invasive and metastasized phenotype. Indeed, the lipogenic phenotype has been linked to advanced and metastatic cancers [26], and the full pattern of metabolic reprogramming may be associated with advanced tumour progression. Our findings in metastasizing RCC, in particular the accumulation of fatty acids, glycerolipid compounds and TCA cycle intermediates such as succinate, are in line with the hypothesis that mitochondrial dysfunction has a role in tumour cell metastasis [27, 28].

Another remarkable finding was the profound up-regulation of α-tocopherol concentration in RCC and despite previous allusions to such an elevated vitamin E concentration [29, 30] this finding has as yet not received particular attention. Among all metabolites investigated in our study, α-tocopherol emerged as the most important classifier of normal *versus* tumorous tissue and therefore underlines the putative importance of vitamin E in RCC biology. The elevated concentration of vitamin E in RCC cells may just be an epiphenomenon und indicate an increased uptake of lipids and fatty acids through the up-regulation of rather unspecific transfer proteins, lipases or lipoprotein receptors [31]. The increased concentration of α-tocopherol has previously been observed in ovarian carcinomas by using similar metabolomics methodology and interpreted as an unspecific stress response [4]. As a potential alternative explanation, elevated vitamin E may indeed play a functional role and render the tumour cell resistant to increased oxidative stress toxic to surrounding normal cell populations. Vitamin E and α-tocopherol in particular, is a potent, lipid-soluble, chain-breaking antioxidant and additional vitamin E has been shown to prevent mitochondrial dysfunction in the presence of severe oxidative stress [32]. However, the specific role of vitamin E is likely not limited to its antioxidant function, but can rather be extended to α-tocopherol serving as a transcriptional regulator of gene expression [33]. Results which point to the importance of α-tocopherol would seem to indicate that further

© 2011 The Author
Journal of Cellular and Molecular Medicine © 2011 Foundation for Cellular and Molecular Medicine/Blackwell Publishing Ltd

studies are justified to clarify the phenomenon of tocopherol elevation in RCC, which may ultimately be exploited for establishing novel therapeutic targeting strategies [19]. Differentially regulated metabolites may also include intracellular signalling molecules, as indicated by the fact that myoinositol was one of the key metabolites identified in our study. This compound belongs to the inositol polyphosphate family of small cytosolic molecules involved in the control of a wide range of cellular processes [34]. Its down-regulation has also been described in prostate cancer [35].

In our exploratory analysis of the metabolic signature of metastatic tumours, intermediates of glucose metabolism, such as succinate and glucose, proved to be key classifiers. These findings are in line with the recent observation that the metastatic progression of RCC is associated with a shift toward non-oxidative glucose metabolism through the pentose phosphate pathway [36]. In our study, the metabolic profile of metastasized tumours could not be thoroughly validated as the number of metastasized tumours was restrictive in the test dataset. Nonetheless, it is worth mentioning that the concentration of arachidonic acid was elevated in metastasized tumours, whereas the concentration in RCC in general was lower than in normal renal tissue. The increase of arachidonic acid in aggressive metastasized tumours seems plausible, as this pro-inflammatory fatty acid has been linked to the VEGF-signalling pathway and tumour angiogenesis. Further, the activation of the inflammatory cascade may indeed increase the metastatic potential of RCC through dysregulation of the immune response in the tumour microenvironment. In this context, the observed elevation of proline levels in tumour tissue can be explained by the degradation of collagen in the microenvironmental extracellular matrix promoting invasive tumour growth [37]. The reduced proline oxidase expression, as described in RCC cell lines [38], would be an alternative explanation. Altogether, the findings in metastatic RCC merit further studies.

References

1. **Gillies RJ, Gatenby RA**. Hypoxia and adaptive landscapes in the evolution of carcinogenesis. *Cancer Metastasis Rev.* 2007; 26: 311–7.
2. **Pelicano H, Martin DS, Xu RH, *et al*.** Glycolysis inhibition for anticancer treatment. *Oncogene.* 2006; 25: 4633–46.
3. **Barba I, Fernandez-Montesinos R, Garcia-Dorado D, *et al*.** Alzheimer's disease beyond the genomic era: nuclear magnetic resonance (NMR) spectroscopy-based metabolomics. *J Cell Mol Med.* 2008; 12: 1477–85.
4. **Denkert C, Budczies J, Kind T, *et al.*** Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. *Cancer Res.* 2006; 66: 10795–804.
5. **Fiehn O, Kind T**. Metabolite profiling in blood plasma. *Methods Mol Biol.* 2007; 358: 3–17.
6. **Schlotterbeck G, Ross A, Dieterle F, *et al*.** Metabolic profiling technologies for biomarker discovery in biomedicine and drug development. *Pharmacogenomics.* 2006; 7: 1055–75.
7. **Wikoff WR, Pendyala G, Siuzdak G, *et al*.** Metabolomic analysis of the cerebrospinal fluid reveals changes in phospholipase expression in the CNS of SIV-infected macaques. *J Clin Invest.* 2008; 118: 2661–9.
8. **Schnackenberg LK, Beger RD**. Monitoring the health to disease continuum with global metabolic profiling and systems biology. *Pharmacogenomics.* 2006; 7: 1077–86.
9. **Lisec J, Schauer N, Kopka J, *et al.*** Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nat Protoc.* 2006; 1: 387–96.
10. **Stacklies W, Redestig H, Scholz M, *et al.*** pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics.* 2007; 23: 1164–7.
11. **Witten IH, Eibe F**. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
12. **Jung M, Mollenkopf HJ, Grimm C, *et al.*** MicroRNA profiling of clear cell renal cell cancer identifies a robust signature to define renal malignancy. *J Cell Mol Med.* 2009.
13. **Gao H, Dong B, Liu X, *et al.*** Metabonomic profiling of renal cell carcinoma: high-resolution proton nuclear magnetic resonance spectroscopy of human serum with multivariate data analysis. *Anal Chim Acta.* 2008; 624: 269–77.
14. **Kim K, Aronov P, Zakharkin SO, *et al.*** Urine metabolomics analysis for kidney cancer detection and biomarker discovery. *Mol Cell Proteomics.* 2009; 8: 558–70.
15. **Seliger B, Dressler SP, Wang E, *et al.*** Combined analysis of transcriptome land proteome data as a tool for the identification of candidate biomarkers in renal cell carcinoma. *Proteomics.* 2009; 9: 1567–81.
16. **Kim JW, Dang CV**. Cancer's molecular sweet tooth and the Warburg effect. *Cancer Res.* 2006; 66: 8927–30.
17. **Semenza GL**. HIF-1 mediates the Warburg effect in clear cell renal carcinoma. *J Bioenerg Biomembr.* 2007; 39: 231–4.
18. **Zhang H, Gao P, Fukuda R, *et al.*** HIF-1 inhibits mitochondrial biogenesis and cellular respiration in VHL-deficient renal cell carcinoma by repression of C-MYC activity. *Cancer Cell.* 2007; 11: 407–20.
19. **Pan JG, Mak TW**. Metabolic targeting as an anticancer strategy: dawn of a new era? *Sci STKE.* 2007; 2007: pe14.
20. **Buzzai M, Bauer DE, Jones RG, *et al.*** The glucose dependence of Akt-transformed cells can be reversed by pharmacologic activation of fatty acid beta-oxidation. *Oncogene.* 2005; 24: 4165–73.
21. **Liu Y**. Fatty acid oxidation is a dominant bioenergetic pathway in prostate cancer. *Prostate Cancer Prostatic Dis.* 2006; 9: 230–4.
22. **Zha S, Ferdinandusse S, Hicks JL, *et al.*** Peroxisomal branched chain fatty acid beta-oxidation pathway is upregulated in prostate cancer. *Prostate.* 2005; 63: 316–23.
23. **Seliger B, Lichtenfels R, Atkins D, *et al.*** Identification of fatty acid binding proteins as markers associated with the initiation and/or progression of renal cell carcinoma. *Proteomics.* 2005; 5: 2631–40.
24. **Teratani T, Domoto T, Kuriki K, *et al.*** Detection of transcript for brain-type fatty Acid-binding protein in tumor and urine of

© 2011 The Author
Journal of Cellular and Molecular Medicine © 2011 Foundation for Cellular and Molecular Medicine/Blackwell Publishing Ltd

patients with renal cell carcinoma. *Urology.* 2007; 69: 236–40.

25. **Horiguchi A, Asano T, Ito K, *et al.*** Fatty acid synthase over expression is an indicator of tumor aggressiveness and poor prognosis in renal cell carcinoma. *J Urol.* 2008; 180: 1137–40.
26. **Menendez JA, Lupu R**. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nat Rev Cancer.* 2007; 7: 763–77.
27. **Ishikawa K, Takenaga K, Akimoto M, *et al.*** ROS-generating mitochondrial DNA mutations can regulate tumor cell metastasis. *Science.* 2008; 320: 661–4.
28. **Lopez-Rios F, Sanchez-Arago M, Garcia-Garcia E, *et al.*** Loss of the mitochondrial bioenergetic capacity underlies the glucose avidity of carcinomas. *Cancer Res.* 2007; 67: 9013–7.
29. **Tosi MR, Rodriguez-Estrada MT, Lercker G, *et al.*** Magnetic resonance spectroscopy and chromatographic methods identify altered lipid composition in human renal neoplasms. *Int J Mol Med.* 2004; 14: 93–100.
30. **Nikiforova NV, Kirpatovsky VI, Darenkov AF, *et al.*** Liposoluble vitamins E and A in human renal cortex and renal cell carcinomas. *Nephron.* 1995; 69: 449–53.
31. **Mardones P, Rigotti A**. Cellular mechanisms of vitamin E uptake: relevance in alpha-tocopherol metabolism and potential implications for disease. *J Nutr Biochem.* 2004; 15: 252–60.
32. **Ham AJ, Liebler DC**. Antioxidant reactions of vitamin E in the perfused rat liver: product distribution and effect of dietary vitamin E supplementation. *Arch Biochem Biophys.* 1997; 339: 157–64.
33. **Azzi A, Gysin R, Kempna P, *et al.*** Regulation of gene expression by alpha-tocopherol. *Biol Chem.* 2004; 385: 585–91.
34. **Burton A, Hu X, Saiardi A**. Are inositol pyrophosphates signalling molecules? *J Cell Physiol.* 2009; 220: 8–15.
35. **Serkova NJ, Gamito EJ, Jones RH, *et al.*** The metabolites citrate, myo-inositol, and spermine are potential age-independent markers of prostate cancer in human expressed prostatic secretions. *Prostate.* 2008; 68: 620–8.
36. **Langbein S, Frederiks WM, zur Hausen A, *et al.*** Metastasis is promoted by a bioenergetic switch: new targets for progressive renal cell cancer. *Int J Cancer.* 2008; 122: 2422–8.
37. **Phang JM, Donald SP, Pandhare J, *et al*.** The metabolism of proline, a stress substrate, modulates carcinogenic pathways. *Amino Acids.* 2008; 35: 681–90.
38. **Maxwell SA, Rivera A**. Proline oxidase induces apoptosis in tumor cells, and its expression is frequently absent or reduced in renal carcinomas. *J Biol Chem.* 2003; 278: 9784–9.

© 2011 The Author
Journal of Cellular and Molecular Medicine © 2011 Foundation for Cellular and Molecular Medicine/Blackwell Publishing Ltd

5.4 Results - Addendum

The models of choice here are alternating decision trees. The choice has two reasons: firstly, the ADTrees [106] perform well on this problem; secondly, trees are simple to interpret. The interpretation is simpler because trees can be nicely visualized and provide a not too high number of predictive variables. For SVMs the result are weights for each variable, which is much more difficult to sum up.

From the n-fold cross-validation, in the paper just 2-fold (= split in half; see Table 1 therein) and 10-fold, it can be seen that the first dataset is abundant enough because the PCC does not increase much from 2-fold to 10-fold cross-validation, at least for ADTrees. Using the same argument, the second dataset has too little data due to the larger differences between 2-fold and 10-fold cross-validation and the varying algebraic sign of the differences.

One property of ADTrees is that they can summarize a set of trees in one tree. This means they can hold the same split as a forest in a single tree notation. The text representation of the tree in the paper Figure 1A is:

```
: 0
|  (1)vitamineE < 0.346: -0.678
|  |  (2)y_100 < 1.712: 0.883
|  |  (2)y_100 >= 1.712: -1.691
|  (1)vitamineE >= 0.346: 1.619
|  (3)glucose_1_phosphate < 0.145: -0.433
|  |  (4)ribonicacid2incorrectassignment < 0.059: -0.873
|  |  (4)ribonicacid2incorrectassignment >= 0.059: 1.169
|  |  (5)inositolput_49 < 8.363: 0.869
|  |  (5)inositolput_49 >= 8.363: -0.736
|  |  (9)y_41 < 5.401: 0.429
|  |  (9)y_41 >= 5.401: -0.479
|  (3)glucose_1_phosphate >= 0.145: 1.252
|  (6)y_103 < 0.029: -0.207
|  (6)y_103 >= 0.029: 0.66
|  (7)y_44 < 0.614: 0.621
|  (7)y_44 >= 0.614: -0.343
|  (8)y_1020 < 0.096: -0.362
|  (8)y_1020 >= 0.096: 0.472
|  |  (10)cho_alcohol_98 < 0.303: 0.57
|  |  (10)cho_alcohol_98 >= 0.303: -0.207
```

The direct representation of this in a diagram is shown in Figure 10. The corresponding binary tree can have up to the squared number of decision nodes.

All machine learning models are constructed with one of the standard tools in this area called WEKA [104].

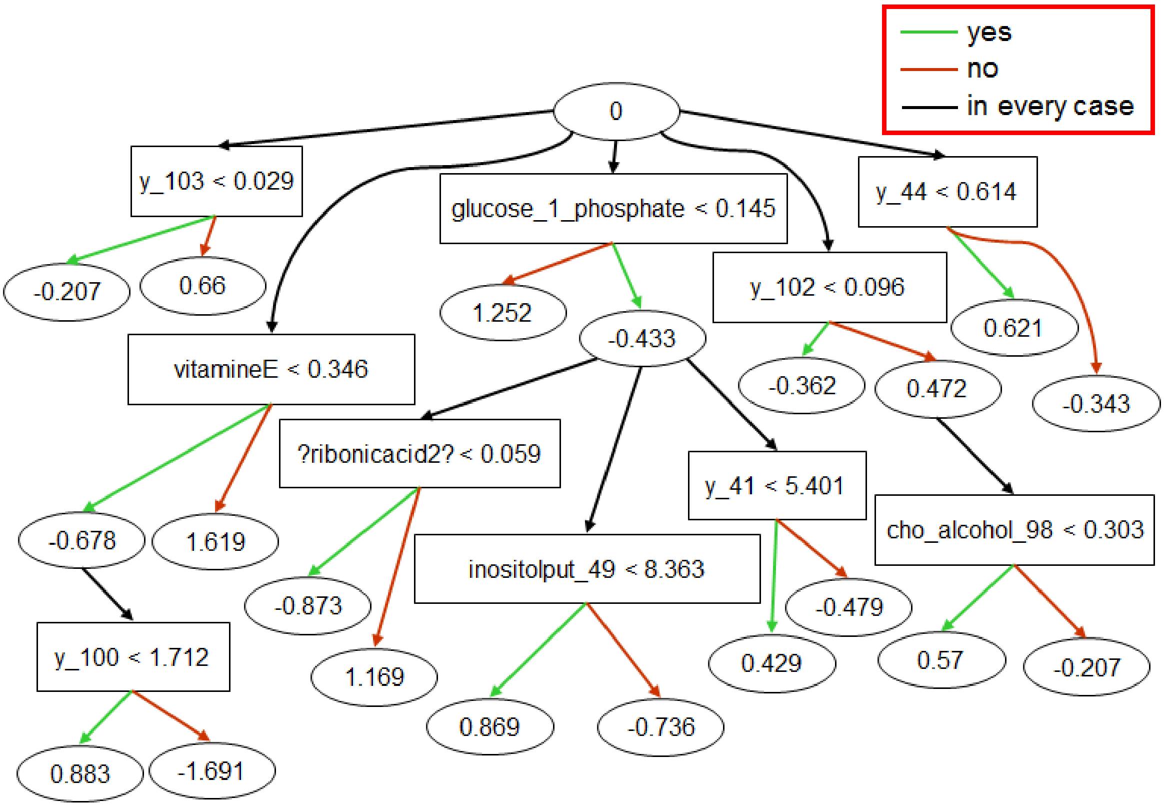


Figure 10. Decision Tree Model (ADTree) for RCC. It is generated for the two-class problem of discriminating RCC and normal renal tissue samples for the text representation in this section. Some labels are different from the paper's Figure 1A as alpha-tocopherol is vitamin E and we were less optimistic in safe metabolite identification in the paper (e.g. 'cho_alcohol_98' is just mentioned as 'Unknown sugar/alcohol' in the paper)

For Figure 1B it is only mentioned in the paper that its performance is not satisfactory. 10-fold cross-validation PCC is 73% for the model in Figure 1B, which is not remarkable with a lower limit of 64%. Since it is a 3-class problem (healthy tissue, non-metastasized and metastasized tumor), we also tried to build models for three classes at once. Again, several methods to build models were tried, but there the range of suitable methods was smaller, that is, ADTrees can only deal with 2-class problems. One similar method is the standard decision tree generator C4.5, which performs slightly worse than ADTrees on our data in the 2-class fashion. Its result for the 3-class problem is:

```
vitamineE <= 0.285613
|   glucose_1_phosphate <= 0.153917
|   |   ribonicacid2incorrectassignment <= 0.094326: 0
|   |   ribonicacid2incorrectassignment > 0.094326: 1
|   glucose_1_phosphate > 0.153917
|   |   124_trihydroxybutane <= 0.060255: 1
|   |   124_trihydroxybutane > 0.060255: 2
vitamineE > 0.285613
|   erythritolput <= 0.057747: 1
|   erythritolput > 0.057747
|   |   glycericacid_3_phosphate <= 0.03226
|   |   |   124_trihydroxybutane <= 0.067844: 1
|   |   |   124_trihydroxybutane > 0.067844: 0
|   |   glycericacid_3_phosphate > 0.03226
|   |   |   malicacid <= 0.870674: 2
|   |   |   malicacid > 0.870674: 1
```

With a 10-fold cross-validation PCC of 79% with a lower limit of 55%. This result lies between the two 2-class formulations and shows that a) non-metastasized vs metastasized tumor is a tougher problem with this data, where the small amount of data is also an issue, and b) that most methods are optimized for 2-class problems. What becomes clear here is that if you have a problem falling into a general problem class and there are methods designed for this class, these methods will likely perform well. A broader overview of problem classes in machine learning is published by the applicant in Platzer, A. Machine Learning - Overview. 2012; Available from: http://sourceforge.net/projects/machine-learning2012/files/ [63].

6. Classical data processing and massive DNA sequencing

6.1 The massive data source from sequencing

As mentioned in section 1.1.5, NGS/sequencing is now one of the largest data sources, and its rate is also increasing. It has changed or replaced several methods where the focus was on specific sequence features since the full sequence can now be captured with not much more effort and the costs are still decreasing (see Figure 5). As computer hard disks are not advancing as rapidly, the major efforts for gaining insights from this data are increasingly the analysis and the storage, resp. the IT-environment. Naturally, biology was not the first data-intensive research area that needed more effective and sophisticated data processing, but as it is a specific data source, it also has specific demands, in comparison with physics and astronomy for instance.

Computer clusters are needed to process and analyze these large amounts of data. This pressure for more hardware can in part also be addressed with more efficiently programmed modules (see section 1.3.6).

One quite expensive step which is almost always needed for NGS data is alignment. Most of the recent aligners are heuristics, which result in some inaccuracy but make the amount of NGS data treatable. For an overview of NGS aligners see [40]. There even exist recent approaches to avoid the alignment step in the analysis (see [107, 108]). Although it is of course always good to know which computationally expensive steps are really needed and/or what effect they have, skipping alignments in sequence analysis appears a bit like surrendering in the face of the required computational effort. On the other hand, there are also advantages in the standard full alignment, as in [109]. A rough estimate was that one sample could be processed in a few days with the library of this paper using our full current computer cluster. We have not tried this, but it is at least not out of scope.

6.2 Computer cluster architecture

Computer clusters are a long-term core topic in computer science. This topic is even older than personal computers, which appeared on the scene in larger quantities about 20 years after mainframe computers. Today’s personal computers have more computational power than the mainframes of several years ago, but as still more speed, processing power and storage are required, mainframes have not disappeared. There are several, not very clearly delineated terms for 'larger' computers, such as mainframe, computer cluster, server farm, and so on. For definitions see [110, 111].

Primarily, computer clusters are classified according to their main purpose. This can be:

- High-Availability: The services of the machine shall not fail at any time. To archive this, the components are redundant.
- Load balancing: The machine should handle processes in parallel, which should be served with equal performance. Usually this is done with several units, where a load balancer distributes the requests.
- High Performance Computing: This type should primarily have high computational power. The demands may be different: high-throughput for single jobs and/or many jobs in parallel.
- Storage: more storage is needed at once than a single hard disk has. The focus can be on throughput, size and/or data safety.

NGS data is moving towards more than one of these possible design goals of a computer cluster: a large storage to store all the data, with high performance computing for processing. For the latter both are useful: many jobs in parallel because the data is already divided in samples, which is perfect parallelization; and high-throughput for single jobs for methods like de novo assembly.

In the best case, the execution of programs on a cluster is done just as on the development machine, which is possible when the following conditions are met: the operating system is the same and the setup for the computational nodes, the distribution and job scheduling is done transparently. From the other perspective: a single job can be done on the development machine. This latter is a problem if a demonstration job already needs more resources than the development machine has; in this case, new modules must developed directly on the cluster.

Although the various clusters are quite different, for some design issues there is a kind of common agreement:

- The nodes are divided in computational nodes, management nodes and 'invisible' administration nodes; computational nodes are only accessed with the job scheduler
- The storage is an own entity and can be mounted from multiple computational entities.
- The storage is divided for projects and users. This can be done as quota or dedicated.
- One login node would be a single point of failure, so if a larger cluster is involved it is made redundant in a load-balancing way.
- The IO can be a bottleneck on several levels, so the cluster backbone, network, caching and general speed should be in tune with each other.
- Only the very basic software is available at login, software needed is loaded and unloaded with a module system. This is to make different versions of a software available.
- Jobs after submission are organized in queues, where it make sense that different queues exist, for example a debug queue for short jobs with high priority and queues for different types of machines.
- Generally, all computational nodes should be identical; if they are not, they are divided into classes with different job queues.
- Jobs get only a precisely defined amount of resources, if the job module requests or simply takes more, only the job will crash and not the computational node.
- On the one hand, the jobs should be spread across the computational nodes, because a node usually only gets a certain amount of IO; on the other hand, there should also be free nodes if a single job needs more resources at one node. This usually results in a bias to one side, depending on the other setup constraints.
- Moving a running job from one node to another is complicated for arbitrary types of programs, so this only exists as potential option.
- The job scheduler should distribute the computational resources fairly between the jobs and users and should use all computational resources.

It may not be immediately clear that this last point always involves a trade-off: that a perfect schedule, a fair share per user and no computational resources wasted by computational nodes being idle when jobs are waiting cannot all be optimally done at the same time for all combinations of jobs. Here is one simple example to visualize the competing demands: the memory is organized in slots of 4GB (with at least one core for it), user1 submits many jobs which need 4GB, user2 submits a single job for 8GB when the cluster is already filled with jobs of the user1; when one job of user1 is finished and 4GB are available, should the job scheduler wait for another job to finish for providing something for user2 (which would leave one slot idle), or should it use the free slot for a suitable job of user1 (so that user2 would have to wait until all single-slot jobs are complete)?

It should additionally be mentioned that finding the optimal job schedule is itself a computationally difficult problem (NP-hard, see [98]). This means that a cluster which is trying to find the optimal schedule will be only able to update the job queues once every few minutes.

The following computer clusters were available and used for the project of this section:

- The **GMI cluster** (Gregor Mendel Institute) until 2013: This cluster had no special name; identification was mainly by its login node with a generic name. It had a storage of 100TB and ~300 cores divided into 3 classes. Job scheduling was performed by the Sun Grid Engine. I had not seen this cluster from its beginning, so I likely missed its starting issues and saw it only in the mature state.
    - *advantages*:
        - a simple setup compared with other clusters
        - although we were warned that the main file system could crash and fail, there was no outage of the storage and almost none of the computational resources
        - job scheduling was a combination of not wasting computational resources and having a sophisticated priority value for sharing between users
    - *disadvantages*:
        - no direct connection to the other storages, which meant copying files from or to other storages flowed through the user machines
        - the storage/file system was not safer than expected; a few times a flipped bit was visible
        - the setup of a few nodes was sometimes changed for cluster experiments without taking this node out of the general computational nodes pool, resulting in jobs with strange behavior, for example jobs which were vastly slowed down
- The bios and the tarbell cluster of the Kenwood Data Center on the University of Chicago campus, short **UC cluster**: 100TB backed-up storage were dedicated for the institute where I work. The bios cluster was the first cluster there with ~1000 cores; it was transitioned into the tarbell cluster with ~2500 cores in 2014. It was a quite smooth transition as the computational cores only were introduced at one end and removed at the other; the storage remained identical. In sum, after several iterations, this cluster is quite mature now.
    - *advantages*:
        - from the storage aspect it is quite simple, and, since the full storage is backed up, it is the safest storage concept possible (likely with some overhead for hardware and administration which is not visible for users)

    - since the organization changed from a single part-time person managing large parts of it to a full team of people, the support and the possibilities have become very professional
    - there are almost no limits for jobs, neither in time or requested resources; queues for every demand are handled in a transparent way unless otherwise specified
  - *disadvantages*:
    - the cluster belongs to a larger university with all the usual bureaucracy; when there was just one part-time person some organizational tasks were rather Kafkaesque
    - the computational jobs were and still are somewhat too unrestricted: it periodically happens that indirect resource requests of jobs lead to crashing computational nodes
    - the job scheduling system attempts to find the optimal solution, which means that the job status is only updated once every few minutes: this also implies that it takes minutes to delete a job and up to half an hour to start or delete a group of jobs
    - as it is an external cluster at a distance, the network speed from or to it is less than would be the case in-house
- the **Mendel cluster** [112] of the GMI: This cluster is the successor of the unnamed GMI cluster which operated until 2013. The transition was a direct hand-over with the data copied from one cluster to the other. Initially it had ~340TB storage divided into two parts: a scratch storage for working (which is planned to be purged automatically; on the other hand it is the only storage visible to computational nodes); and a project storage for the results. There are ~2000 cores, 2 login nodes and 2 data mover nodes (and of course a certain amount of administration nodes). This cluster has a portal page, serving as an organizational overlay. It serves as a formalized system of resource management, makes it simple to look up which projects are running with which people; on the other hand it is an additional layer of bureaucracy beyond the cluster itself. As it is currently the newest cluster, its setup iterations are quite fresh in memory: half a year of iterations has led to a reasonably mature state now. This period is rather long for this, though not particularly long for a larger cluster.
  - *advantages*:
    - very solid, almost no jobs failed through the cluster
    - scalable, it should be able to grow ~8x, almost transparently
  - *disadvantages*:
    - it has a scheduled downtime of ~half a day each month
    - it is rather complex. The required cluster training of half a day before first use is a good idea and is also needed for this cluster
    - the need for staging (the term for the interplay of the project storage and the purged working storage for the jobs)
    - the behavior of the job scheduler, which intentionally prefers job arrays (i.e., does not lead to a fair share between users). This will likely be iterated further when the cluster is under high loads for a longer time

- several elements which could work transparently need special treatment/attention here: copying larger amount of data, queue selection, walltime limitations

One may ask why this issue is detailed to such extend in the thesis, as it is more recognized as a helping area and/or engineering and not as science. Firstly, some aspects of it are science, for example job shop scheduling for which about 60 papers can be found from IEEE (http://ieeexplore.ieee.org) since 2013; also computational clusters are the applied side of computability, a core field in computer science. Secondly, it creates the difference between smooth and fast analyses or long workarounds, where a workaround can range from small adaptions for a non-transparently working cluster to different architectures, modules and algorithms to make something computable. Occasionally, analyses are simply not done due to of the expected larger effort in a given setup.

6.3 Project management

The project of this chapter is the part of the 1001 Genomes Project - A Catalog of *Arabidopsis thaliana* Genetic Variation, http://1001genomes.org/ [8]. The manuscript of this chapter is only one paper of several in this project.

First, samples are collected and RILs generated. In our case, these are plants (mainly *Arabidopsis thaliana*) from all over the world. The samples received a name and a unique numerical ID. At least two IDs are intentionally created: one ID for the strain, which is called ecotypeid, and one for the ~individual. The strains are all RILs (= ~8 generations self-pollinating), nevertheless the second ID exists for the reason that a strain can change in generations, or to distinguish the samples if there were problems in self-pollination.

As this project is a collaboration between several groups, the plan was that all involved groups first collect samples, sequence and analyze a part of them, and then combine all into one resource. The GMI, where the applicant is currently employed, mainly collected samples from Sweden. These samples are the basis of the manuscript in section 6.6.

Later the sample data of the project were collected at the UC cluster (see page 63) and the additional data in one combined table. This will result subsequently in one manuscript combining all data.

6.4 Pipelines for calling events

See the course materials [113], where the applicant was one of 3 lecturers and had the part of calling events when a reference is available. For other overviews see [114-116]. [117] is better focused on de novo assembly.

As we worked with model organisms, a reference is always available; in these cases de novo assembly is merely an extension, when larger sequence changes are expected in certain strains.

The specific main pipeline used in section 6.6 is partly specified in the supplementary command listing 1 therein. The remainder is not fully specified therein, partly because of space limitations and partly because it honors the term 'ad hoc pipeline', which basically means it was not in good shape from a computer science angle. This other half is shown in Figure 10 from a preliminary version of the supplement. After a reviewer's comment, we decided to skip it.

The reason in this case for the non-optimal shape of the pipeline was not that the people making it were unable to do better (and I make this statement not merely because these people were two colleagues and myself); it was more a case of 'spending more time on it doesn't matter and isn't honored, especially if something is done the first time' (to cite a colleague, who will very likely forgive me if I do not mention his name here).

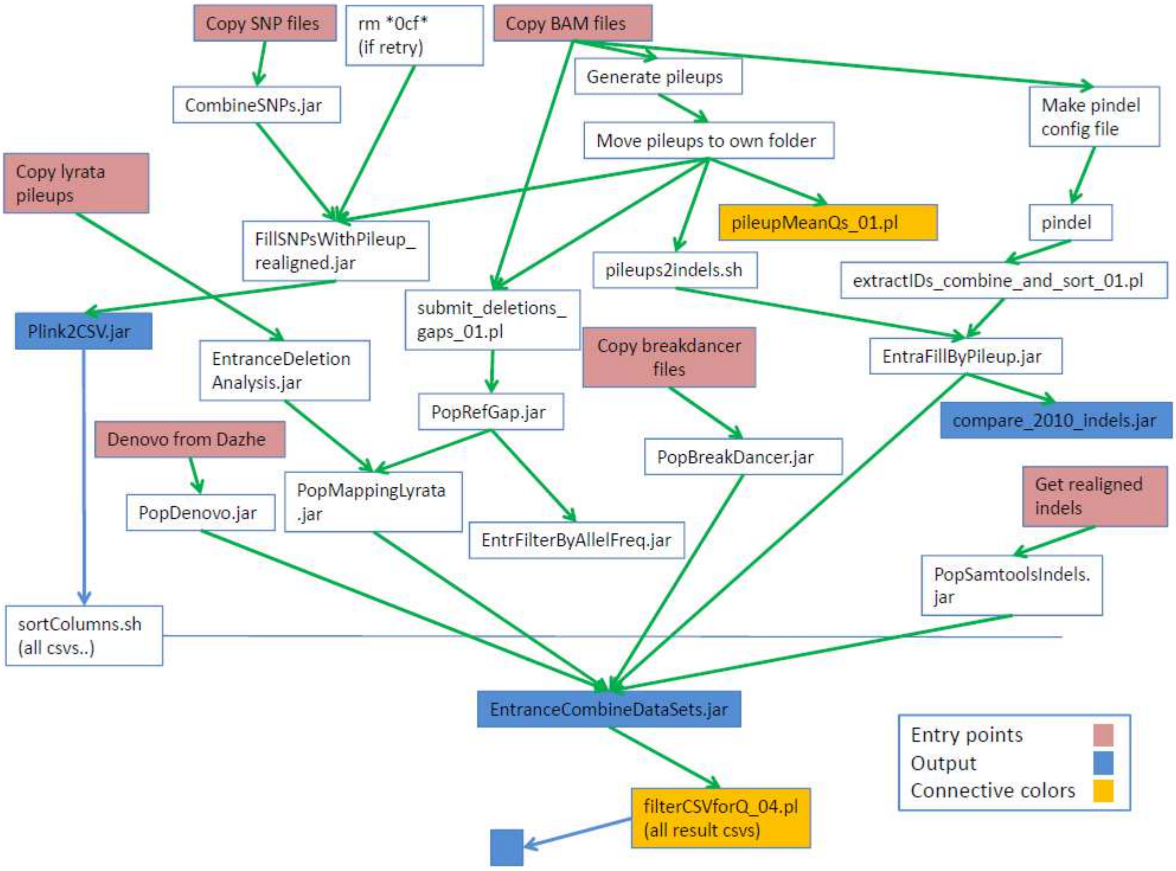


Figure 11. Pipeline for the full population of samples in section 6.6. The figure here is more a demonstration for the term 'ad hoc pipeline'.

6.5 The effect of filtering and other biases

It seems somehow ironic that the analyses of collections of samples, resp. populations, are affected by population structure [118-120]; however, on the whole population analyses are mainly affected by different filtering and other biases [121].

Generally, all measures depending on the full genome sequence are biased by filtering; for example, when calling with a reference sequence, the reference allele is always called more safely than a non-reference allele. It is also quite common that more events lie in filtered regions, but it is difficult to say where the noise starts. Examples for measures depending on the full genome sequence are polymorphisms in terms of pair-wise difference π, fixation index F_{st}, estimation of split times and dendrograms, and so on. It must be said that these measures within a set of samples, which are generated and filtered similarly, are quite correct and meaningful, but comparisons between sets of different quality, data amount and/or filtering often give arbitrary results. In this regard, samples sequenced with a large difference in coverage are also not similar as data source, since the sequence difference is biased if a reference genome is covered to 10% on one side and to 90% on the other: when taking the difference from a much smaller fraction it is either biased in the absolute number or it is biased in noise.

On the other hand, analyses like causal inference or associations are less dependent on having a full genome: either the source/causal/associated event is in the set or it is not. In the latter case, when the interesting information is filtered, there should be a negative result, not a biased one.

There are often clear improvements in processing and filtering NGS data, but usually a trade-off remains between calling safely and calling as much as possible. Non-population analyses require more the safe data and are biased by population structure, whereas population analyses would need both, safe *and* complete data.

An example of the effects of filtering is in supplementary Figure 6 in the paper of the next section. As mentioned there, 'Very conservative criteria were used to polarize the polymorphisms in order to avoid inflation at the right end of the plots.', which is indeed the case; yet it is hard to say where the noise, resp. the inflation starts. If the reference used for calling were the perfect ancestor sequence (assuming such a perfect ancestor exists for the samples), then the allele frequencies would monotonically drop to the right and inflation would more likely occur on the left side. If the reference is not close to the ancestor, the allele frequency might also rise to the right end and there is more likely to be inflation. In our case the species is *Arabidopsis thaliana*, the reference is Col-0 and the ancestor is *Arabidopsis lyrata*. As mentioned in the text, the divergence should also not be too large for determining the ancestral state. In any case, the ratio between low and high derived allele frequency should not be taken as strong argument.

One good hint on filtering from this paper is to remove duplicates: in the past, it was a standard step during library preparation to get rid of likely much amplified PCR products. Here it is shown that at least some of these duplications are real biology (for rDNA, see Supplementary Figure 23).

6.6 Article: **Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden**

Quan Long, Fernando A Rabanal, Dazhe Meng, Christian D Huber, Ashley Farlow, **Alexander Platzer**, Qingrun Zhang, Bjarni J Vilhjálmsson, Arthur Korte, Viktoria Nizhynska, Viktor Voronin, Pamela Korte, Laura Sedman, Terezie Mandáková, Martin A Lysak, Ümit Seren, Ines Hellmann & Magnus Nordborg. *Massive genomic variation and strong selection in* Arabidopsis thaliana *lines from Sweden*. **Nat Genet, 2013. 45(8): p. 884-90.**

Although it might frequently be cited because of its data, this paper does not only present data. It also contains many analyses, which might be somewhat squeezed together, at least near the end when we were compressing things, even the supplemental. The paper also provides several starting points for the next paper in the 1001 Genomes Project - besides the fact that the 1001 Genomes Project regards the complete project data set, while this paper regards only the Swedish subset.

OWN CONTRIBUTION IN [36]

Q.L., D.M. and A.P. performed primary analysis of the sequencing data, including all polymorphism detection and quality control.

Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden

Quan Long[1,5], Fernando A Rabanal[1,5], Dazhe Meng[2,5], Christian D Huber[3,5], Ashley Farlow[1,5], Alexander Platzer[1], Qingrun Zhang[1], Bjarni J Vilhjálmsson[2], Arthur Korte[1], Viktoria Nizhynska[1], Viktor Voronin[1], Pamela Korte[1], Laura Sedman[1], Terezie Mandáková[4], Martin A Lysak[4], Ümit Seren[1], Ines Hellmann[3] & Magnus Nordborg[1,2]

© 2013 Nature America, Inc. All rights reserved.

Despite advances in sequencing, the goal of obtaining a comprehensive view of genetic variation in populations is still far from reached. We sequenced 180 lines of *A. thaliana* from Sweden to obtain as complete a picture as possible of variation in a single region. Whereas simple polymorphisms in the unique portion of the genome are readily identified, other polymorphisms are not. The massive variation in genome size identified by flow cytometry seems largely to be due to 45S rDNA copy number variation, with lines from northern Sweden having particularly large numbers of copies. Strong selection is evident in the form of long-range linkage disequilibrium (LD), as well as in LD between nearby compensatory mutations. Many footprints of selective sweeps were found in lines from northern Sweden, and a massive global sweep was shown to have involved a 700-kb transposition.

The common weed *A. thaliana* is highly selfing and naturally exists as inbred lines that can be grown in replicate under controlled conditions. The species is widely distributed throughout the northern hemisphere and shows strong evidence of local adaptation[1,2]. The pattern of genetic polymorphism is compatible with isolation by distance on every scale[3]. Taken together, these features make *A. thaliana* an excellent model for studying the genetics of natural variation, and, indeed, shared inbred lines have been a resource for the *Arabidopsis* community since its inception[4]. More recently, over 1,300 lines have been genotyped for 250,000 SNPs using a custom Affymetrix SNP tiling array (AtSNPtile1) to facilitate genome-wide association studies (GWAS)[5,6], and efforts are underway to sequence over 1,000 lines[7–11].

Here we report the sequencing of 180 lines from Sweden. We contribute the largest sample by far from a single geographic region, which allows us to look for evidence of selection and to carry out GWAS in local populations for the first time. Our analysis emphasizes structural variation, which we show to be a major component of genetic variation.

RESULTS

Sequencing and polymorphism detection

The analyzed lines were selected on the basis of low-density SNP data[3] to obtain samples with distinct genotypes from both northern and southern Sweden (52 versus 128 lines, respectively; **Supplementary Fig. 1**). Using 76- or 100-bp Illumina paired-end reads and fragments of roughly 300 bp in size, we obtained an average of 39-fold coverage per line. We identified differences from the *A. thaliana* reference genome, including short insertion-deletion polymorphisms (indels) and other structural variants, using an *ad hoc* pipeline (Online Methods). This approach generated 4.5 million SNPs and almost 0.6 million structural variants, over 90% of which are indels shorter than 10 bp in length. The data had low error rates overall (**Supplementary Table 1**), but it is important to realize that the genome sequences are far from complete. Several important biases exist. First, we are only able to detect polymorphisms reliably in the roughly 85% of the genome that can be uniquely aligned to the reference genome (**Fig. 1a**). Second, some kinds of variants are easier to detect than others. For example, we estimated that false positive and false negative rates when detecting short indels (shorter than 15 bp) were roughly twice as high as when detecting SNPs (**Supplementary Table 1**), and these rates rose markedly with increasing length of the indel. Third, there are biases with respect to the reference genome. For example, we found more indels with the variant allele shorter than the reference genome than with the variant allele longer than the reference genome. These differences are unlikely to be real (as there is no evidence that the reference genome is unusually large) but can readily be explained by noting that alignment algorithms handle gaps better than inserted sequence. Consistent with this interpretation, such discrepancies were almost absent for polymorphisms of ≤4 bp but increased with the length of the indel (**Fig. 1b**).

The overlap between the SNPs identified here and by two previous resequencing efforts[9,10] is shown in **Figure 1c**. Whereas the majority of SNPs were present in all data sets, there was also a very large number

---

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria. [2]Molecular and Computational Biology, University of Southern California, Los Angeles, California, USA. [3]Max F. Perutz Laboratories, University of Vienna, Vienna, Austria. [4]Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [5]These authors contributed equally to this work. Correspondence should be addressed to M.N. (magnus.nordborg@gmi.oeaw.ac.at).

Received 19 November 2012; accepted 31 May 2013; published online 23 June 2013; doi:10.1038/ng.2678

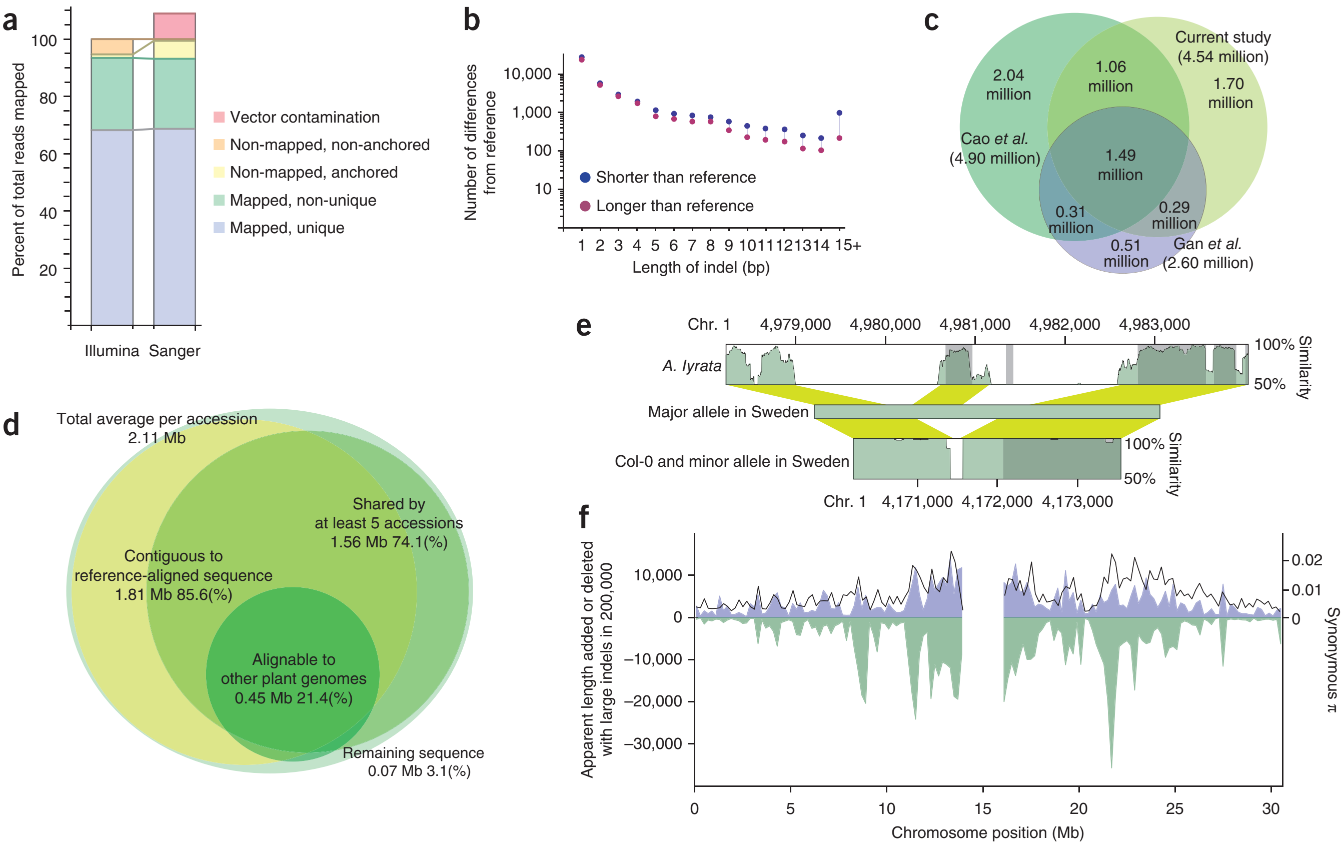


Figure 1 Polymorphism detection. (**a**) Comparison of Illumina reads and longer, dideoxy-sequenced, randomly cloned fragments (Sanger) with respect to how well they align to the reference genome. The distributions are very similar, except that longer reads that cannot be aligned are more likely to be anchored by a short stretch of presumably homologous sequence. (**b**) Average number of indels between the sequenced lines and the reference genome, divided into variants that are shorter and longer than the reference genome and shown as a function of the length of the variant. (**c**) Overlap between SNPs generated by this study and two previous resequencing studies[9,10]. (**d**) Characterization of new sequence identified by *de novo* assembly. (**e**) An example of a region containing new sequence. The graphs show sequence similarity (coding sequence in dark green, noncoding sequence in light green; yellow shows alignment) to the majority haplotype in Sweden, which contains a ~1-kb fragment of new sequence not found in the reference genome. The new fragment is also found in *A. lyrata*, indicating that it is ancestral; however, the region has been subject to several more rearrangements since the species diverged. The polymorphism may have functional consequences, as it affects putative coding sequence. (**f**) Distribution of large variants increasing length (blue; identified using *de novo* assembly), large variants decreasing length (green; inferred from sequencing coverage) and SNPs (synonymous nucleotide diversity, π; black line) along chromosome 1. Chromosomes 2–5 show an analogous pattern (**Supplementary Fig. 2**).

© 2013 Nature America, Inc. All rights reserved.

of new SNPs, as expected given that previous studies were smaller and did not include lines from Sweden. The total number identified was smaller than the number previously identified in 80 lines[10] (4.54 versus 4.90 million, respectively), reflecting a combination of differences in SNP calling and real differences between the samples (mostly in population structure, as the average number of pairwise differences per site between individuals did not differ greatly: 0.49% for the 180 lines sequenced here, 0.53% for the 80 previously sequenced lines; based on regions with high alignment scores in our data). A rigorous analysis of the nature of these differences will require reprocessing the raw sequence data using a common pipeline.

Detection and characterization of new sequence

The biases that arise from aligning to a reference genome apply to all resequencing studies, but there is reason to believe them to be more serious for *A. thaliana*, which has a genome half the size of its nearest relative, *Arabidopsis lyrata*, apparently owing to deletions in the *A. thaliana* lineage[12]. If this reduction in genome size is still ongoing, individual *A. thaliana* genomes will harbor many ancestral chromosomal segments not present in the reference genome (and will lack equally many that are). With this in mind, we assembled all our lines individually, *de novo*, identifying 1.3–3.3 Mb of new sequence per line (compared to 181 kb for Col-0, the line corresponding to the reference genome), largely in segments shorter than 10 kb. Most of this new sequence seemed to be genuine *A. thaliana* genomic sequence: 96.5% of the new sequence was either anchored by a sequence that aligned well with the reference genome or was shared by at least five of the Swedish lines (**Fig. 1d**). Furthermore, 21% of the sequence showed similarity to sequence from other plant genomes, usually *A. lyrata* (**Supplementary Note**), and thus likely represents retained ancestral fragments; however, closer examination often identified complex polymorphisms, making the precise mutational events difficult to infer (**Fig. 1e**). The genomic distribution of the new sequence is similar to that of regions of missing coverage, as would be expected if the latter reflect segregating longer deletions of ancestral sequence (that we largely did not detect). Both distributions resembled that of SNPs, suggesting that all three types of polymorphism are influenced by similar evolutionary forces (**Fig. 1f** and **Supplementary Fig. 2**). On the basis of available annotation and preliminary mRNA sequencing data, the identified new sequence seems to contain around 200–300 genes or gene fragments per line, in agreement with previous estimates[9]. One might expect rapidly

Table 1 Multiple regression of flow cytometry–based estimates

Feature	DF	SS	MS	F	*P* value	R^2
45S rDNA	1	739	739	94	7.7×10^{-17}	0.39
5S rDNA	1	42	42	5	0.023	0.022
Centromeres	1	114	114	14	2.3×10^{-4}	0.059
TEs	1	56	56	7	8.8×10^{-3}	0.029
Error	123	968	8			
Total	127	1,918				

DF, degrees of freedom; SS, sum of squares; MS, mean square; TEs, novel transposable element insertions. Total $R^2 = 0.50$; adjusted $R^2 = 0.48$.

evolving gene families, such as F-box and NB-LRR genes[12], to be overrepresented, but no evidence for this was found.

Massive variation in genome size

The above analyses suggest that, despite the recent marked decrease in the size of the *A. thaliana* genome, variation between lines is only on the order of 1%. Yet, flow cytometry analysis has suggested that there is up to 10% variation worldwide[13]. Using the same technique, we found that our lines varied by well over 10%, ranging from 161 Mb to 184 Mb in length. The estimate for the reference line, Col-0, was 166 Mb, making it one of the smallest, whereas the largest values were found exclusively in lines from northern Sweden. Extending the study by including 36 lines selected from the worldwide distribution of the species confirmed this impression: the variation in lines from southern Sweden was similar to that found worldwide, whereas the estimates in lines from northern Sweden were substantially greater (**Supplementary Fig. 3**).

Given the analyses above, it seemed unlikely that the cause of this variation would lie in the unique portion of the genome. To investigate the role of repetitive sequence, we used sequence coverage to estimate copy number variation for 45S rDNA, 5S rDNA and centromeric repeats, as well as for transposable elements, and used the results to predict the flow cytometry–based estimates of genome size using linear regression. In a multiple regression, all four classes of repeats were significantly positively correlated with the flow cytometry–based estimates; however, 45S rDNA made the largest contribution by far (**Table 1**). Notably, both the flow cytometry–based estimates and the 45S rDNA copy number estimates showed a strong geographic pattern, with larger estimates being more prevalent and the correlation between the estimates being much stronger ($R^2 = 0.73$) in lines from northern Sweden (**Fig. 2a**).

These results confirm that there is considerable natural variation in nuclear DNA content and demonstrate that this variation is mainly due to 45S rDNA, in agreement with findings from previous studies[14]. Because the flow cytometry and genome sequencing experiments used different plants as well as tissues (leaves and roots, respectively), it is clear that the variation is heritable. To investigate the genetics of this variation, we carried out a GWAS for the flow cytometry–based estimates of genome size. Unexpectedly, this analysis identified neither of the two known 45S rDNA clusters[15]. Instead, the scan identified a major locus in a euchromatic region of chromosome 1 that apparently explained 26% of the variation in genome size (**Fig. 2b**). Neither sequence analysis nor FISH found any evidence for new 45S rDNA clusters (**Supplementary Fig. 4**).

It would thus seem that the identified locus regulates DNA content in *trans* rather than in *cis*, and this, in turn, implies that the presumed 'genome size variation' should, at least partially, be regarded as a phenotype rather than a genotype. There is evidence of regulation of rDNA copy number in several organisms[16–18], including *A. thaliana*[19]. Notably, mapping of variation in cytosine methylation of 45S rDNA repeat arrays, which is strongly correlated with copy number[20], in a cross between two inbred lines has previously identified both *cis* and *trans* quantitative trait loci (QTLs)[21]. The two strongest QTLs corresponded to the 45S rDNA clusters, but the third strongest contained the GWAS peak reported here. These results are consistent with ours if the repeat number changes too rapidly to be mapped using GWAS but is inherited stably enough to be mapped in crosses. The *trans*-acting loci might modify the replication process, with different alleles effectively predisposing lines to large or small numbers of repeats. The peak of association contained at least three candidates that might affect replication (**Fig. 2c**)[22–25].

However, it must be emphasized that the association may simply be spurious. GWAS on subsets of the lines showed that the chromosome 1 association was due to a relatively small number of lines from northern Sweden with very large genome size estimates (**Supplementary Fig. 5**). Although our analysis takes confounding from genome-wide population structure into account, it does not necessarily handle confounding caused by a small number of genes of large effect[26,27]. A spurious correlation could arise due to LD with the true causal loci, for example, the 45S rDNA clusters themselves, which we think we are unable to map owing to allelic heterogeneity. In other words, the peak on chromosome 1 could be a so-called synthetic association[26,28]. To resolve this, multigeneration experiments will be required.

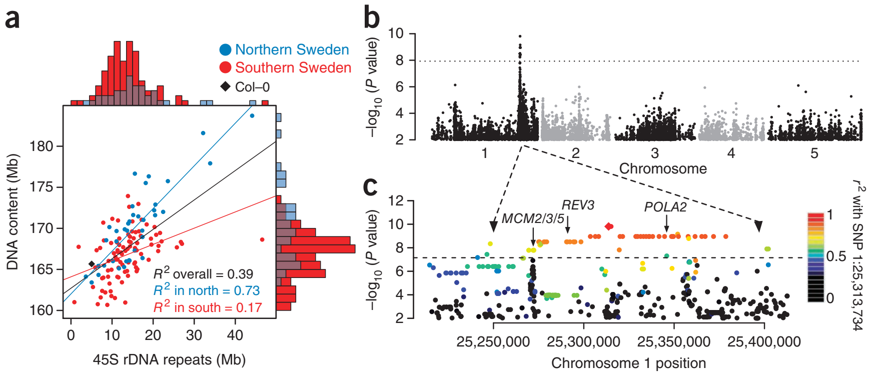


Figure 2 Genome size variation. (**a**) Joint distribution of nuclear DNA content (estimated using flow cytometry) and total amount of 45S rDNA (estimated using sequencing coverage). Marginal distributions are shown along the axes. (**b**) Manhattan plot of genome-wide association results for the flow cytometry–based estimates of genome size. The dotted horizontal line marks a significance level of 0.05 after Bonferroni correction for 4 million tests. The two known 45S rDNA clusters are close to the left ends of chromosomes 2 and 4 (ref. 15). (**c**) Magnified view of the chromosome 1 peak in **b** including a roughly 100-kb region of extensive LD. Colors indicate the extent of LD with the most significant SNP at position 25,313,734. The positions of three replication-related candidate genes are shown: *POLA2* (At1g67630), which encodes the B subunit of DNA polymerase α; *REV3* (At1g67500), which encodes recovery protein 3, the catalytic subunit of DNA polymerase ζ; and *MCM2/3/5* (At1g67460), which is related to the minichromosome maintenance family of proteins. Sequence analysis of these candidates identified no obvious candidate polymorphisms (multiple alignments are available on the project download site).

© 2013 Nature America, Inc. All rights reserved.

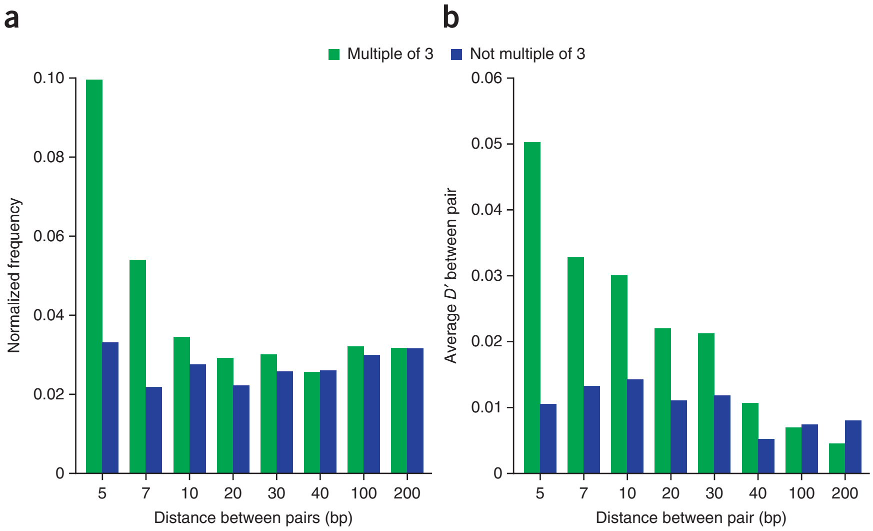


Figure 3 Compensatory indels. (**a**) Overrepresentation of compensatory pairs of indels compared to their genome-wide frequency, plotted as a function of the distance between the indels. Compensatory pairs of indels are those whose sum length is a multiple of 3, thus restoring the reading frame. (**b**) LD (*D*′) between compensatory pairs of indel alleles as a function of the distance between the indels. Positive LD indicates an excess of non-reference alleles.

Selection and LD

We searched for evidence that some of this genomic variation is adaptively important. With regard to the variation in nuclear DNA content, its marked geographic distribution (**Fig. 2a**) was suggestive of local adaptation, as the overall genetic divergence was much smaller. Less than 0.6% of SNPs showed a stronger correlation with location in northern versus southern Sweden than the flow cytometry–based estimates of genome size. However, if the variation is due to a very small number of genetic loci, then it might have been possible for genetic drift to cause the observed divergence in size. Resolving this will require further studies.

Given the apparent recent shrinkage of the *A. thaliana* genome, it is also natural to consider selection at indels. Previous work, using a small number of indels, has suggested that deletions are selectively favored relative to insertions, perhaps because of selection for a more compact genome[12]. Unfortunately, this kind of analysis is very sensitive to the kinds of biases we saw in our data and, even worse, depends on accurate inference of the ancestral state (that is, whether an indel is the result of a deletion or an insertion). Indels are often complex (see **Fig. 1e** for an example). For the 18% of indels we were able to classify unambiguously, there was no evidence of selection favoring deletions, in contradiction to previous results (**Supplementary Fig. 6**). However, it is dangerous to extrapolate from a biased minority of events, and our conclusion is that the divergence between *A. thaliana* and *A. lyrata* is probably too great for analyses that rely on the determination of the ancestral state of indels to be reliable.

However, we found several other clear signals of strong selection. Recent resequencing efforts have notably identified many new protein-coding alleles involving apparently disruptive frameshift mutations and closely linked compensatory changes[8–10]. With our larger sample size, we were able to show that selection has a role in creating this diversity. Closely linked alleles that restored the reading frame were greatly overrepresented compared to those that did not, and positive LD between such alleles ensures that aberrant proteins occur at a lower frequency than expected from the marginal allele frequencies (**Fig. 3**). How these kinds of variant haplotypes arise is far from clear, as the evolution of compensatory changes involves crossing an adaptive valley[29]. One possibility is that the population structure of *A. thaliana* leads to local fixation of weakly deleterious mutations during colonization of new patches, which is followed by compensatory evolution as the local population size increases.

Strong selection can also cause LD between unlinked loci, especially in conjunction with local adaptation (in which case, there is no requirement for epistatic interactions between the loci). In agreement with previous results[6,10,30–32], average LD in our sample decayed relatively quickly (on roughly the same scale as in humans) to high background levels that were largely determined by population structure (**Supplementary Fig. 7**). However, even after taking this structure into account, considerable long-range LD remained, including over 300,000 pairs of loci for which r^2 was >0.8, even though the loci were separated by more than 1 Mb (**Fig. 4a**). Especially notable was the prevalent LD between all centromeres. Because it is difficult to imagine selection maintaining LD between all centromeres, it seemed likely that most of these patterns must be artifactual, perhaps because the SNP loci, in fact, map to multiple regions. Indeed, strict filtering for uniqueness resulted in the elimination of all but around 70,000 pairs with long-range LD (corresponding to 7,973 loci). From these, we selected 4 centromeric and 2 non-centromeric sets of SNPs for genotyping in informative crosses (**Supplementary Table 2**). Of the centromeric pairs, one showed complete linkage, despite the SNPs supposedly being located on different chromosomes, and the other three failed PCR, perhaps because they are associated with repetitive regions. These results illustrate the danger inherent in assuming that SNPs are located where they are supposed to be located and show that population genetics analysis may assist in identifying unreliable ones. However, both non-centromeric pairs segregated independently in crosses, showing that at least some of the long-range LD we observed must be due to normal population genetics forces, whether chance or natural selection. In support of the latter explanation, there was a significant enrichment of the remaining loci among SNPs exhibiting signs of having been involved in local adaptation (**Fig. 4b** and **Supplementary Table 3**).

Global and local selective sweeps

Population structure in *A. thaliana* is generally characterized by varying degrees of isolation by distance[3]. In previous studies, samples from southern Sweden have seemed to be part of a European continuum, whereas those from northern Sweden were quite distinct[6,31]. Our data confirmed this distribution (**Supplementary Figs. 7** and **8**)[33] and further suggest that the divergence is due to changing allele frequencies rather than the accumulation of mutations (as would accompany ancient separation with little gene flow), as we found fewer private alleles in lines from northern Sweden than in lines from southern Sweden (18% versus 67%) and because pairwise sequence divergence was commensurate with the distance between the regions (**Supplementary Fig. 9**). However, within each region, the divergence increased more rapidly in lines from northern Sweden, consistent with previously reported greater population structure there[3,6,31], as well as with field observations: whereas *A. thaliana* is a common weed in southern Sweden, its distribution in northern Sweden

© 2013 Nature America, Inc. All rights reserved.
npg

© 2013 Nature America, Inc. All rights reserved.

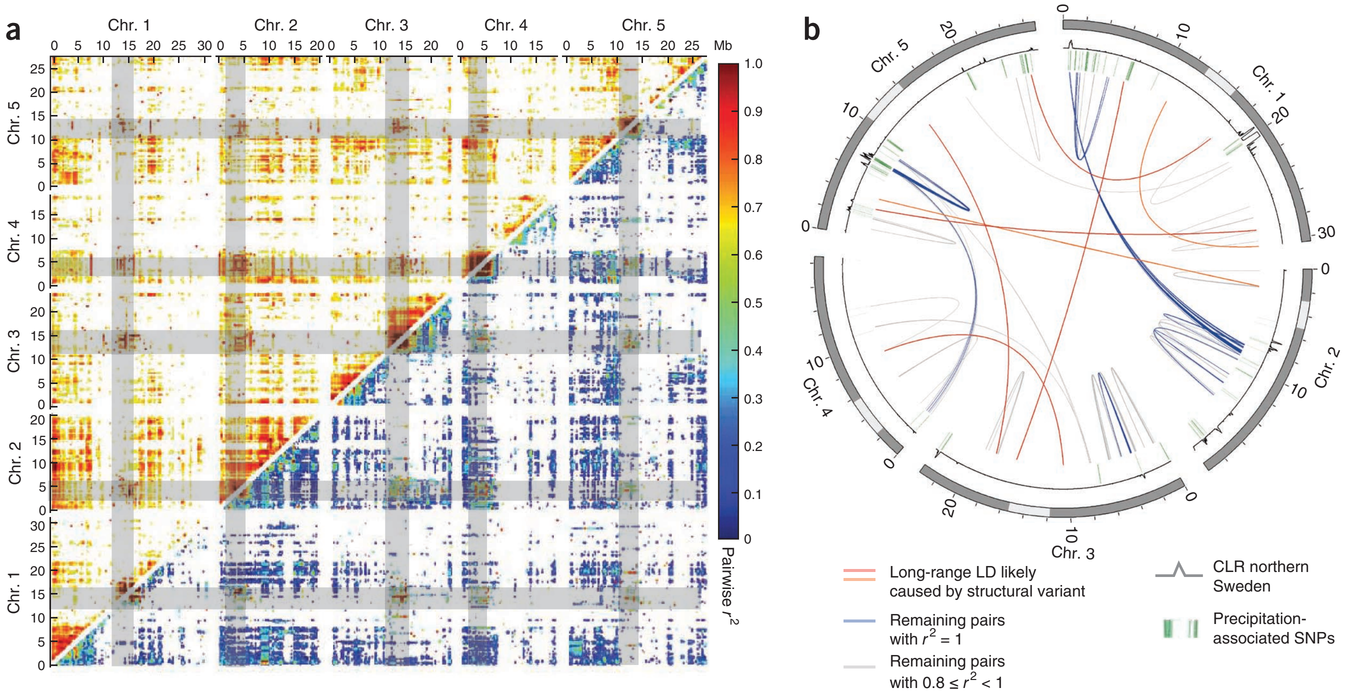


Figure 4 Long-range LD. (**a**) Genome-wide pairwise LD. Values before correcting for population structure are shown above the diagonal; for clarity, only values above 0.6 are shown. Values after applying a transformation to reduce the effects of population structure (related to the correction used in genome-wide association mapping; **Supplementary Note**) are shown below the diagonal. (**b**) Remaining long-range LD after extensive filtering, combined with positions of putatively selected loci. Green bars show the position of loci significantly associated with minimum precipitation and relative humidity in a global sample (**Supplementary Table 3**), and the gray curve indicates the signatures of local adaptation in the northern Swedish population (**Fig. 5**). Gray bars indicate centromeric regions.

is often restricted to eroded south-facing slopes and is much more patchy (M.N., unpublished observation). Whereas most of the divergence between lines from northern and southern Sweden is likely due to genetic drift, there are clear differences in many traits that are likely to be adaptive, such as seed dormancy and flowering time (M.N., unpublished observation), and we thus decided to search our data for evidence of selective sweeps.

The results for lines from northern and southern Sweden were markedly different. SweepFinder[34], an algorithm that uses the distribution of SNP allele frequencies to detect sweeps close to fixation, returned 22 strong signals in lines from northern Sweden and only a single signal in lines from southern Sweden (**Fig. 5a**, **Supplementary Figs. 10–14** and **Supplementary Note**). The signals were extremely strong: the SweepFinder composite likelihood ratio for the strongest selective sweep was 178 times that corresponding to background, and those for the other sweeps were 30 times stronger on average. Most selective sweeps exhibited strong population subdivision, quantifed using F_{ST} (**Fig. 5a**), and were found by the F_{ST}-like cross-population statistic XP-CLR[35], in agreement with the notion that they are due to local adaptation. The identified regions were also over-represented among SNPs that showed long-range LD as well as among SNPs that have previously been associated with environmental variables (**Fig. 4b** and **Supplementary Note**).

The reason for the much greater number of sweep signals in lines from northern Sweden is not clear. Distinguishing between real signals of selection and artifactual ones due to demography is, as always, very difficult. However, at least one of the identified selective sweeps is

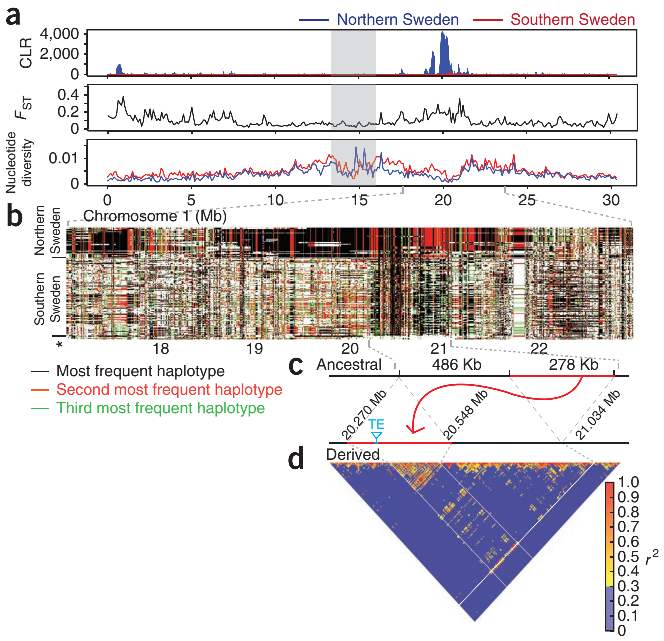


Figure 5 Characterization of selective sweeps on chromosome 1. (**a**) Values of three different statistics sensitive to selective sweeps plotted along the chromosome. Statistics were calculated separately for the lines from northern and southern Sweden. The CLR statistic clearly marks a strong sweep in the northern lines, and the same region also shows increased F_{ST} as well as decreased nucleotide diversity. The gray bar indicates the centromeric region. (**b**) Pattern of haplotype sharing underlying the major signal around 20 Mb. Shown are haplotypes derived from lines in northern and southern Sweden, as are the six presumed ancestral haplotypes (asterisk). Haplotype sharing is much more extensive in the lines from northern Sweden than in those from southern Sweden. (**c**) Schematic of the transposition event most likely responsible for the observed pattern. (**d**) Pattern of LD across the swept region (red bar in **c**).

almost certainly real. The single signal in lines from southern Sweden corresponded to the strongest signal in lines from northern Sweden (**Fig. 5a**), and the pattern of haplotype sharing showed that the selective sweep in lines from northern Sweden was simply more extensive (**Fig. 5b**). If the sweep signals in lines from northern Sweden were simply due to complicated demographics (for example, colonization bottlenecks and concomitant differences in local effective population size (**Supplementary Figs. 7** and **15**)), there would be no reason to expect them to overlap with sweep signals in lines from southern Sweden. Furthermore, the signal corresponded to a presumed global selective sweep, previously identified in a chip-based resequencing study of 20 lines, in which 18 of the 20 lines were found to share a haplotype of several hundred kilobases in length, with the remaining 2 lines hailing from Cape Verde and northern Sweden, at the southern and northern edges of the species range, respectively[36]. The simplest explanation for the observed pattern is thus that this is an ongoing global selective sweep and that the sweep is more recent in lines from northern Sweden than in those from southern Sweden. And, if this is true, then it seems likely that some of the other strong signals in lines from northern Sweden also represent genuine selective sweeps rather than artifacts due to demographic factors.

A curious feature of the previously reported selective sweep was that the shared haplotype appeared identical in all carriers[36], which is inconsistent with the random action of recombination. Furthermore, the extent of haplotype sharing seemed far too great given the average decay of LD in global samples of *A. thaliana*. An obvious explanation was that the selective sweep was associated with some kind of large-scale structural variant that suppressed recombination locally. With this in mind, we examined the region more closely and discovered that the swept haplotype was associated with an intrachromosomal conservative transposition of 278 kb containing 72 genes to a new position 486 kb away (**Fig. 5c** and **Supplementary Note**). The *A. thaliana* reference line Col-0 carried the swept haplotype, as did most members of the species: using genome-wide SNP data[6], we estimated that only 45 of 1,306 lines (3.4%) had escaped the selective sweep (**Supplementary Note**). Contrary to previous results, the ancestral haplotype was not just found at the extremes of the range but was also found at low frequency worldwide. Recombination in heterozygotes is likely to be effectively suppressed by selection against recombinants, given that crossing over within the region would lead to either duplication or deletion of the 72 transposed genes. The pattern of LD across the region was suggestive of the suppression of recombination (**Fig. 5d**). It should be noted that the strong signal of selection was not simply due to lack of recombination: it remained present even if we treated the entire transposed region as a single locus (SweepFinder scores based solely on SNPs outside the rearrangement decreases from 178 to 165 times the background).

The breakpoints of the identified transposition are consistent with the action of non-homologous end joining. Resealing at the donor site seems to have been facilitated by 5 bp of microhomology, leading to a 5-bp deletion, whereas a 9-bp target site deletion occurred at the receptor site (**Supplementary Fig. 16**). Although we do not know the selective agent, the transposition seems to contain a relatively small number of derived variants that tag the sweep globally, including roughly 30 SNPs and 2 helitron insertions. Attempts to date the selective sweep on the basis of polymorphism among the swept haplotypes yielded estimates of 43,000 years for lines from southern Sweden and 17,000 years for lines from northern Sweden (**Supplementary Note**), which predate the end of the last glaciation in Sweden and are consistent with the lack of geographic structuring of the sweep[3].

DISCUSSION

We have used next-generation sequencing to generate a high-quality polymorphism data set for a Swedish sample of *A. thaliana*. We provide a reasonable estimate of variation for SNPs and very short indels in the fraction of the genome that is accessible using these methods[10], and, although biases complicate many kinds of evolutionary analyses, the data comprise an important resource, in particular for GWAS. At the same time, our findings highlight how much we may be missing by simply employing standard pipelines for polymorphism detection. Perhaps most notably, we discovered massive variation in nuclear DNA content and showed that it may be possible to map genes regulating this variation, suggesting that what we had assumed to be part of the genotype should partly be viewed as a phenotype. It is also clear that we have very little idea of how many large structural variants (especially inversions and transpositions) exist. By combining population genetics analysis with manual searches for putative breakpoints in the sequencing data, we uncovered a very large structural variant that seems to have undergone extremely strong selection. Our attempt to search for such variants systematically, using a novel method based on *de novo* assembly, identified several other noteworthy examples, including the 1.17-Mb inversion that gave rise to a heterochromatic knob on chromosome 4 (**Supplementary Fig. 17**) (ref. 37). However, there is every reason to believe that there is more to be found. Of the roughly 13 million SNPs that distinguish the *A. thaliana* and *A. lyrata* reference genomes, roughly 4.4% are polymorphic in our sample of *A. thaliana* genomes. The corresponding percentage for short indels is of the same magnitude. If similar selection pressures affect large structural variants, a similar proportion of the very large number of structural rearrangements between these two genomes[11] should still be segregating. Many of these polymorphisms may be complex and very difficult to resolve using short-read sequencing data. Finally, our analyses found signs of selection at every level, from compensatory changes within single genes to local adaptation (giving rise to long-range LD and footprints of selective sweeps) and global selective sweeps. Even in an organism as well studied as *A. thaliana*, the genome is full of surprises.

URLs. 1001 Genomes Project, http://1001genomes.org/; interactive map of the lines used, http://goo.gl/2n6wp; download site for data from this paper, http://downloads.gmi.oeaw.ac.at; NCBI Sequence Read Archive (SRA), http://www.ncbi.nlm.nih.gov/Traces/sra/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Flat files of all polymorphism data as well as various lists and tables can be downloaded from the project website. Raw data have been deposited in the NCBI SRA under accession SRP012869. Seeds of all 180 lines have been submitted to the *Arabidopsis* Biological Resources Center stock center and will be available under accession CS78885.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank O. Mittelsten Scheid for comments on the manuscript, J. Dolezel for providing a size standard for flow cytometry, G. Schmauss for technical assistance with flow cytometry, N. Lettner for help with sample preparation, A. Sommer for help with sequencing and the Gregor Mendel Institute IT team (in particular, P. Forai) for excellent cluster support. This work was supported by European Research Council grant 268962 MAXMAP and European Community Framework

© 2013 Nature America, Inc. All rights reserved.

Programme 7 grant 283496 transPLANT to M.N., by the Austrian Science Fund (Vienna Graduate School of Population Genetics, FWF W1225) to I.H. and by Czech Science Foundation grants P501/12/G090 and P506/12/0668 to M.A.L.

AUTHOR CONTRIBUTIONS

M.N. supervised the project. V.N. generated the sequencing data. Q.L., D.M. and A.P. performed primary analysis of the sequencing data, including all polymorphism detection and quality control. D.M. carried out *de novo* assembly. F.A.R. and L.S. performed the genome size analyses. M.A.L. and T.M. carried out FISH analyses. Q.L., D.M., Q.Z. and B.J.V. analyzed the pattern of LD. C.D.H. and I.H. carried out population structure and selective sweep analyses. A.F., D.M., A.K., P.K. and V.V. analyzed the chromosome 1 transposition. Ü.S. contributed web tools and helped with data management. M.N. wrote the manuscript with major input from Q.L., F.A.R., D.M., C.D.H., A.F. and I.H.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Fournier-Level, A. *et al.* A map of local adaptation in *Arabidopsis thaliana*. *Science* **334**, 86–89 (2011).
2. Hancock, A.M. *et al.* Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* **334**, 83–86 (2011).
3. Platt, A. *et al.* The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000843 (2010).
4. Koornneef, M., Alonso-Blanco, C. & Vreugdenhil, D. Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu. Rev. Plant Biol.* **55**, 141–172 (2004).
5. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
6. Horton, M.W. *et al.* Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216 (2012).
7. Weigel, D. & Mott, R. The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biol.* **10**, 107 (2009).
8. Schneeberger, K. *et al.* Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc. Natl. Acad. Sci. USA* **108**, 10249–10254 (2011).
9. Gan, X. *et al.* Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011).
10. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011).
11. Schmitz, R.J. *et al.* Patterns of population epigenomic diversity. *Nature* **495**, 193–198 (2013).
12. Hu, T.T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–481 (2011).
13. Schmuths, H., Meister, A., Horres, R. & Bachmann, K. Genome size variation among accessions of *Arabidopsis thaliana*. *Ann. Bot.* **93**, 317–321 (2004).
14. Davison, J., Tyagi, A. & Comai, L. Large-scale polymorphism of heterochromatic repeats in the DNA of *Arabidopsis thaliana*. *BMC Plant Biol.* **7**, 44 (2007).
15. Copenhaver, G.P. & Pikaard, C.S. RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. *Plant J.* **9**, 259–272 (1996).
16. Brown, D.D. & Dawid, I.B. Specific gene amplification in oocytes. Oocyte nuclei contain extrachromosomal replicas of the genes for ribosomal RNA. *Science* **160**, 272–280 (1968).
17. Tartof, K.D. Increasing the multiplicity of ribosomal RNA genes in *Drosophila melanogaster*. *Science* **171**, 294–297 (1971).
18. Yao, M.C., Kimmel, A.R. & Gorovsky, M.A. A small number of cistrons for ribosomal RNA in the germinal nucleus of a eukaryote, *Tetrahymena pyriformis*. *Proc. Natl. Acad. Sci. USA* **71**, 3082–3086 (1974).
19. Pontvianne, F. *et al.* Histone methyltransferases regulating rRNA gene dose and dosage control in *Arabidopsis*. *Genes Dev.* **26**, 945–957 (2012).
20. Woo, H.R. & Richards, E.J. Natural variation in DNA methylation in ribosomal RNA genes of *Arabidopsis thaliana*. *BMC Plant Biol.* **8**, 92 (2008).
21. Riddle, N.C. & Richards, E.J. The control of natural variation in cytosine methylation in *Arabidopsis*. *Genetics* **162**, 355–363 (2002).
22. Casper, A.M., Mieczkowski, P.A., Gawel, M. & Petes, T.D. Low levels of DNA polymerase α induce mitotic and meiotic instability in the ribosomal DNA gene cluster of *Saccharomyces cerevisiae*. *PLoS Genet.* **4**, e1000105 (2008).
23. Sakamoto, A. *et al.* Disruption of the *AtREV3* gene causes hypersensitivity to ultraviolet B light and γ-rays in *Arabidopsis*: implication of the presence of a translesion synthesis mechanism in plants. *Plant Cell* **15**, 2042–2057 (2003).
24. Wittschieben, J.P., Reshmi, S.C., Gollin, S.M. & Wood, R.D. Loss of DNA polymerase ζ causes chromosomal instability in mammalian cells. *Cancer Res.* **66**, 134–142 (2006).
25. Forsburg, S.L. Eukaryotic MCM proteins: beyond replication initiation. *Microbiol. Mol. Biol. Rev.* **68**, 109–131 (2004).
26. Platt, A., Vilhjálmsson, B.J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045–1052 (2010).
27. Vilhjálmsson, B.J. & Nordborg, M. The nature of confounding in genome-wide association studies. *Nat. Rev. Genet.* **14**, 1–2 (2013).
28. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
29. Meer, M.V., Kondrashov, A.S., Artzy-Randrup, Y. & Kondrashov, F.A. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* **464**, 279–282 (2010).
30. Nordborg, M. *et al.* The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **30**, 190–193 (2002).
31. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3**, e196 (2005).
32. Kim, S. *et al.* Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **39**, 1151–1155 (2007).
33. Platzer, A. Visualization of SNPs with t-SNE. *PLoS ONE* **8**, e56883 (2013).
34. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575 (2005).
35. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
36. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
37. Fransz, P.F. *et al.* Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organization of heterochromatic knob and centromere region. *Cell* **100**, 367–376 (2000).

© 2013 Nature America, Inc. All rights reserved.

© 2013 Nature America, Inc. All rights reserved.

ONLINE METHODS

Sequencing and polymorphism detection. Genomic DNA was fragmented, size selected to between 450 and 800 bp and subjected to paired-end Illumina sequencing with read length of 76 or 100 bp. Reads were mapped with Burrows-Wheeler aligner (BWA)[38] to the TAIR 10 reference genome, allowing 4% mismatch and one indel. SNPs and short indels were called with SAMtools[39] and the Genome Analysis Toolkit (GATK)[40]. Larger structural variants were called using a variety of tools. For further details, see the **Supplementary Note**. Tables summarizing the results are available on the project download site.

Error estimates and quality control. Considerable effort was devoted to quality control. Notably, we were not simply trying to ensure that identified SNPs were called correctly but also tried to estimate the underlying sequence, paying as much attention to what was missed as to what was found. We compared our data to 4 different kinds of data to estimate error rates for SNPs and short indels: (i) the reference line was resequenced using our pipeline, and all variants called were assumed to be false positives; (ii) our results were compared with previously published SNP chip data[6] to provide estimates of the false negative rate (the rate at which we did not discover SNPs) and the genotyping error rate (the rate at which we made the wrong call for the ones we did detect); (iii) our results were directly compared with an old data set of close to 1,500 manually curated multiple alignments of PCR amplicons from Sanger sequencing of 95 lines[31]; and (iv) our results were directly compared with ~250 kb of sequence from a single accession that we generated by Sanger sequencing random shotgun clones. The results are summarized in **Supplementary Table 1** (for details, see the **Supplementary Note**). In general, error rates were higher close to centromeres and decreased markedly as the quality of the mapping (alignment *Q* value) increased (**Supplementary Figs. 18**–**20**).

Detection and characterization of new sequence. Each line was assembled *de novo* using SOAPdenovo to identify fragments longer than 100 bp that were absent from the reference genome (see the **Supplementary Note** for details). The majority of such fragments could either be anchored to the reference genome by flanking sequence or were shared by more than five lines (**Fig. 1d** and **Supplementary Fig. 21**). Summaries are available on the project download site.

Variation in genome size. Flow cytometry was carried out on 128 of the Swedish lines, the reference line (Col-0) and 36 randomly chosen worldwide lines using 2-week-old leaves. Copy number variation for 45S rDNA and 5S rDNA and centromeric repeat number were estimated via normalized read coverage across the appropriate region of the reference genome. In simple single-factor analysis, only 45S and 5S rDNA contributed significantly to the flow cytometry–based estimates (**Fig. 2a**, **Supplementary Fig. 22** and **Supplementary Table 4**). Estimates for 45S rDNA were validated via quantitative PCR (**Supplementary Fig. 23**). Further details are given in the **Supplementary Note**.

GWAS on genome size estimates were carried out with imputed SNP data from this study, accounting for population structure using a mixed model. The genome was scanned for new rDNA clusters bioinformatically (by using an algorithm that searches for read pairs with one read matching the relevant repeat and the other read anchored in unique sequence[41]), as well as by using FISH (**Supplementary Note**).

Selection on indels. Where alignment was possible, the ancestral state of each SNP and indel was defined using the *A. lyrata* genome as the outgroup. The criteria used are detailed in the **Supplementary Note**.

To test for selection on compensatory mutations (**Fig. 3**), the proportion of high-confidence indel pairs (no missing data, confirmed by GATK local realignment and less than three SNPs within 20 bp) within coding regions was normalized to the genome-wide count of indel pairs and binned for distance between the events. The count and LD for events that disrupted an ORF were compared with those for events that did not.

LD in structured populations. LD can be thought of as having three different sources: 'true' LD, population structure and chance or error[42]. The first source encompasses both short-range LD due to cosegregation of alleles (linkage) and LD (at any range) due to locus-specific deterministic forces (for example, selection). The other two sources act across the genome and are typically of no direct interest. However, if the sample is heavily structured, the second source will have a massive influence, making it difficult to draw conclusions about the first source. Two related methods for correcting LD estimates for population structure have been proposed[43,44]; however, the approach adopted in ref. 44 has the advantage that it results in symmetric r^2 values. Our approach (**Supplementary Note**) is similar to the one in ref. 44 (it is also superficially related to the mixed-model correction we used in GWAS), although the underlying assumptions necessary for the derivation are different. In particular, we make no assumptions about the existence of discrete subpopulations, something that would be inappropriate for an organism in which the pattern of variation is characterized by isolation by distance[3]. The transformed LD estimates were generally lower than the original ones, as the inflation caused by population structure had been removed (**Supplementary Fig. 24**); however, large numbers of pairs with long-range LD remained (**Fig. 4a**). Indeed, the presence of strong long-range LD, within as well as between chromosomes, between about 8,000 loci was robust to (i) correction for population structure; (ii) subdivision of our sample into northern and southern populations; (iii) SNP imputation and (iv) read mapping quality (**Supplementary Fig. 25**).

Global and local selective sweeps. Standard methods were used to describe the pattern of polymorphism within and between populations (**Supplementary Note**) and to confirm previously published results concerning population structure and the distinctiveness of the northern Swedish population (**Supplementary Figs. 7**–**9**, **15** and **Supplementary Table 5**). Two lines were identified as likely contaminants (**Supplementary Note**). We scanned the genome for signs of selective sweeps using five different statistics: (i) CLR, the composite likelihood ratio calculated by SweepFinder[34], which is sensitive to perturbations of the allele frequency distribution, was run separately for lines from northern and southern Sweden; (ii) F_{ST}, which simply measures divergence between populations (for example, due to fixation of locally adaptive alleles), was calculated in non-overlapping 100-kb windows; (iii) nucleotide diversity, which is expected to be reduced following a selective sweep, was similarly estimated in windows (but separately for lines from northern and southern Sweden); (iv) XP-CLR[35], which uses one population as a reference and searches for selective sweeps in the other, was used to look for sweeps in both lines from northern and southern Sweden; and (v) XP-EHH[45], which looks for extended haplotype sharing, was used search for evidence of selective sweeps in either population. Detailed results can be found on the project download site.

The chromosome 1 transposition was discovered via manual inspection of split reads in unswept lines. Distantly mapping read pairs were consistent with the transposition arrangement depicted in **Supplementary Figure 16**. PCR and Sanger sequencing of 5 unswept and 15 swept lines (including Col-0) confirmed the expected breakpoint arrangements. *De novo* assembly also correctly identified transposed breakpoint 4 (chromosome 1: 20,270,307 to 20,548,624) in five of six unswept accessions, breakpoint 5 (chromosome 1: 20,270,429 to 21,034,717) in five of six unswept accessions and breakpoint 6 (chromosome 1: 20,548,624 to 21,034,773) in two of six unswept accessions. The selective sweep was dated on the basis of divergence between swept haplotypes (**Supplementary Note**).

38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. Li, H. *et al.* The Sequence Alignment/Map format and SAM-tools. *Bioinformatics* **25**, 2078–2079 (2009).
40. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
41. Platzer, A., Nizhynska, V. & Long, Q. TE-Locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. *Biology* **1**, 395–410 (2012).
42. Ohta, T. Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc. Natl. Acad. Sci. USA* **79**, 1940–1944 (1982).
43. Cockram, J. *et al.* Genome-wide association mapping to candidate polymorphism resolution in the un-sequenced barley genome. *Proc. Natl. Acad. Sci. USA* **107**, 21611–21616 (2010).
44. Mangin, B. *et al.* Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity* **108**, 285–291 (2012).
45. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden

Q. Long, F. A. Rabanal, D. Meng, C. D. Huber, A. Farlow et al.

Supplementary figures

1 Map of sampling locations . . . 2
2 Chromosomal distribution of large indels . . . 3
3 Distribution of flow cytometry estimates . . . 4
4 FISH results for three lines with large estimated 45S rDNA copy number . . . 5
5 GWAS results for genome size using different samples and phenotypes . . . 6
6 Derived allele frequency distributions for indels . . . 7
7 Decay of LD for different sub-populations . . . 8
8 UPGMA clustering of the Swedish lines based on all SNPs . . . 9
9 Isolation by distance . . . 10
10 Sweep statistics for chromosome 1 . . . 11
11 Sweep statistics for chromosome 2 . . . 11
12 Sweep statistics for chromosome 3 . . . 12
13 Sweep statistics for chromosome 4 . . . 12
14 Sweep statistics for chromosome 5 . . . 13
15 Summaries of the pattern of polymorphism in north and south . . . 13
16 Validation of transposition . . . 14
17 LD surrounding putative major rearrangements . . . 15
18 False positive rates in reference re-sequencing as a function of Q-value cut-off . . . 16
19 Distribution of putative false positives in reference re-sequencing . . . 16
20 Error rates from comparison with SNP data as function of Q-value. . . . 16
21 Characterization of novel sequence . . . 17
22 Correlation between flow cytometry and repeat copy number estimates . . . 18
23 Correlation between flow cytometry and 45S rDNA . . . 18
24 Effect of LD transformation . . . 19
25 The robustness of long-range LD . . . 19

Supplementary tables

1 Error rates for SNPs and short indels . . . 20
2 Validation of long-range LD . . . 20
3 Enrichment of long-range LD SNPs among SNPs associated with climate . . . 20
4 Summary of simple regressions of flow cytometry estimates. . . . 21
5 Genetic and geographic distance between world-wide populations . . . 21

Supplementary note

1 Genome sequencing **22**

2 Polymorphism detection **22**

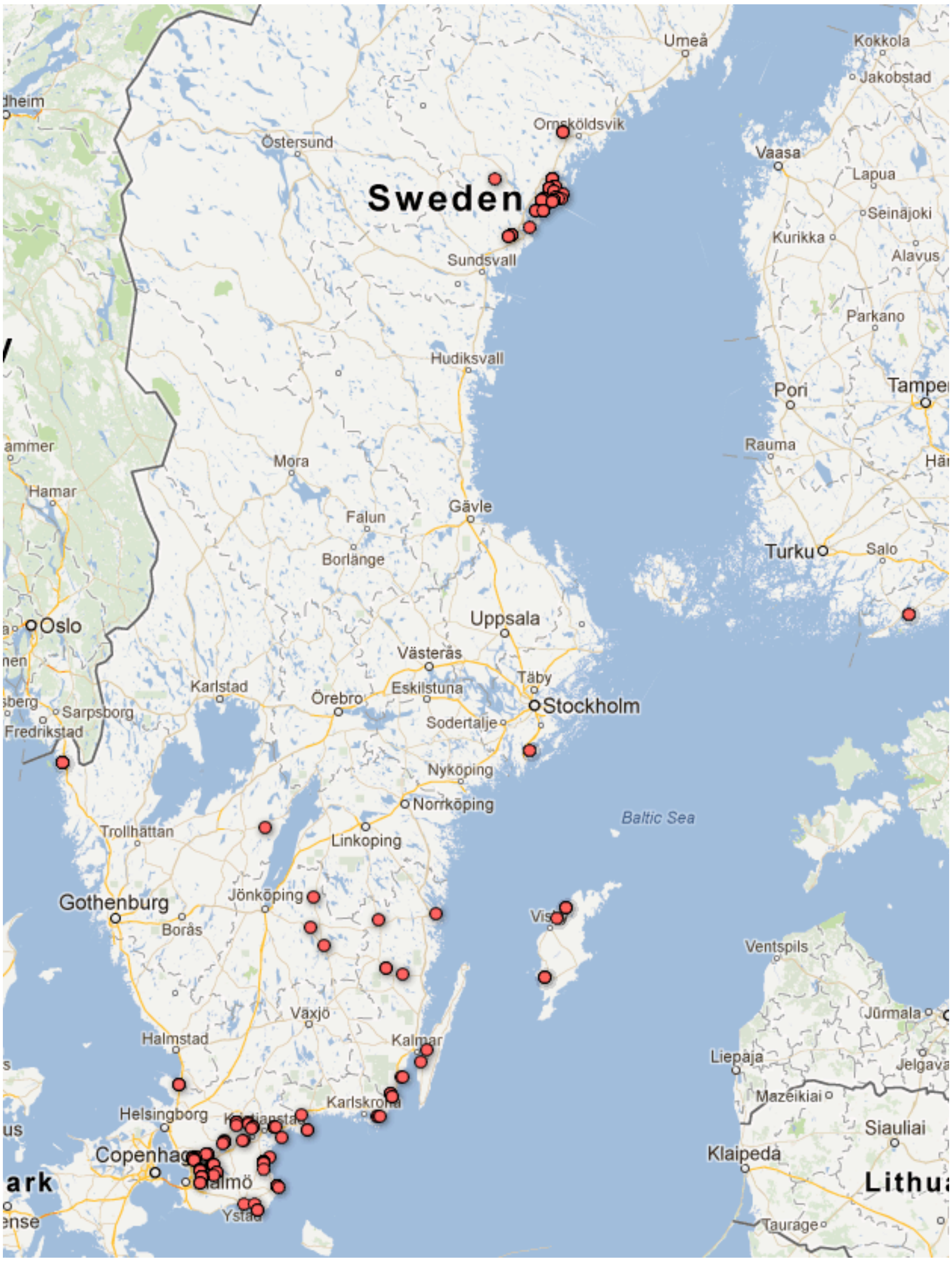
3 *De novo* assembly **25**

4 Variation in nuclear DNA content **27**

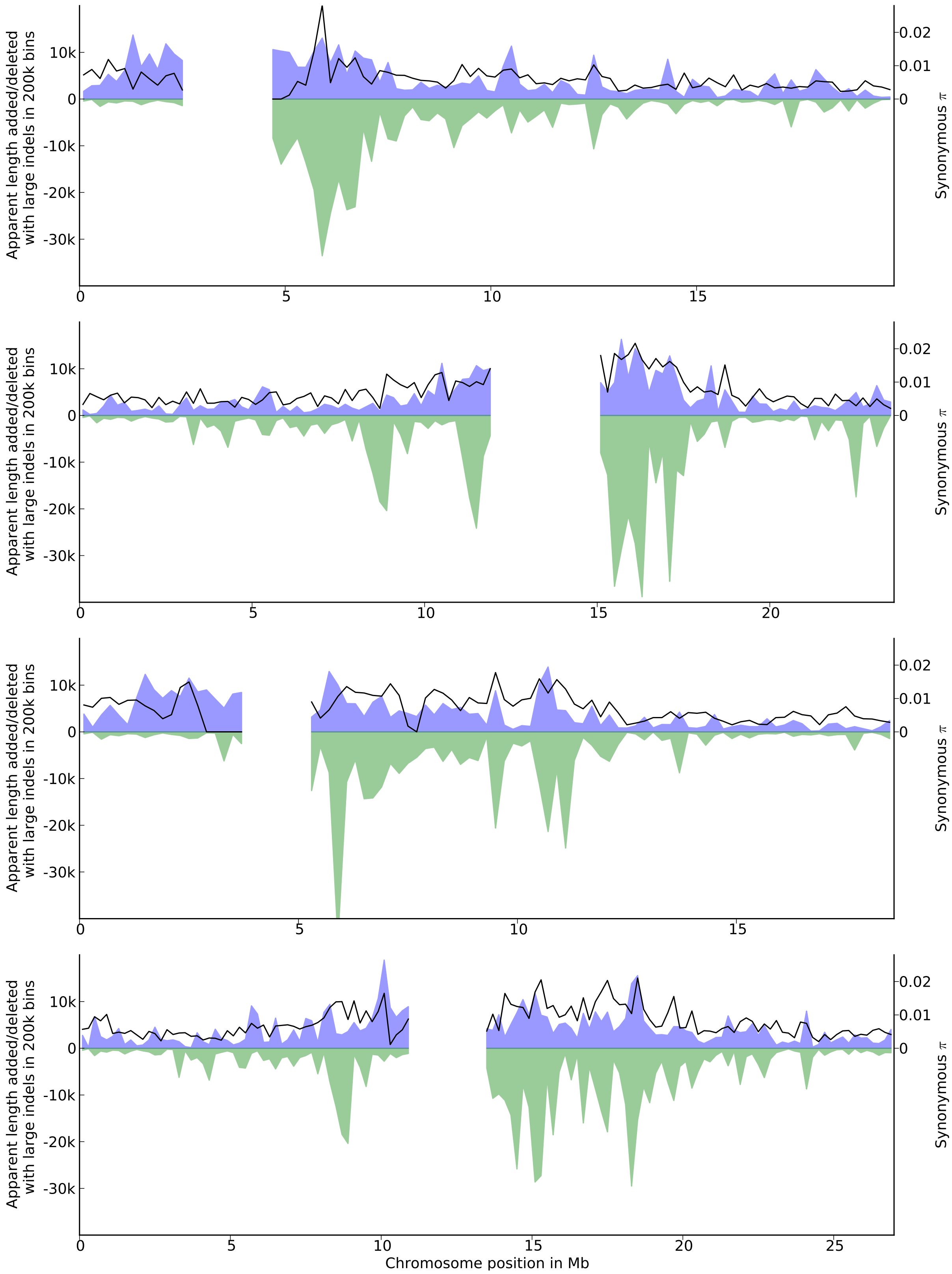
5 Population genetic analyses **28**

6 Data release **31**

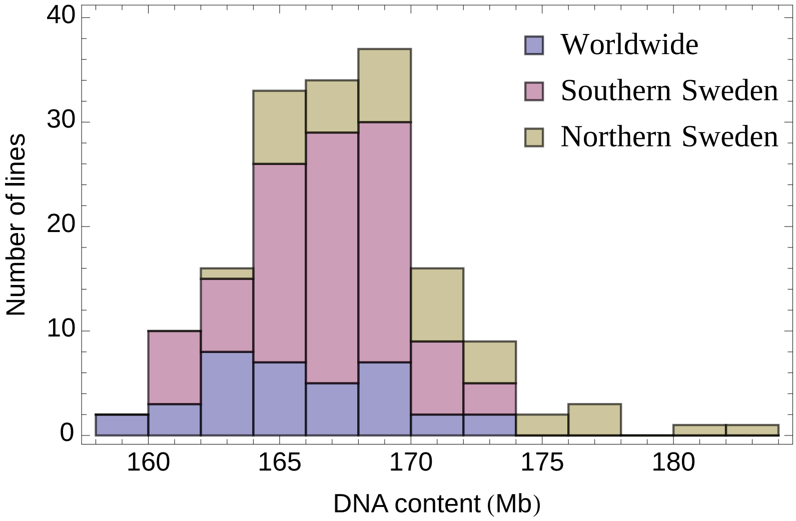
Nature Genetics: doi:10.1038/ng.2678



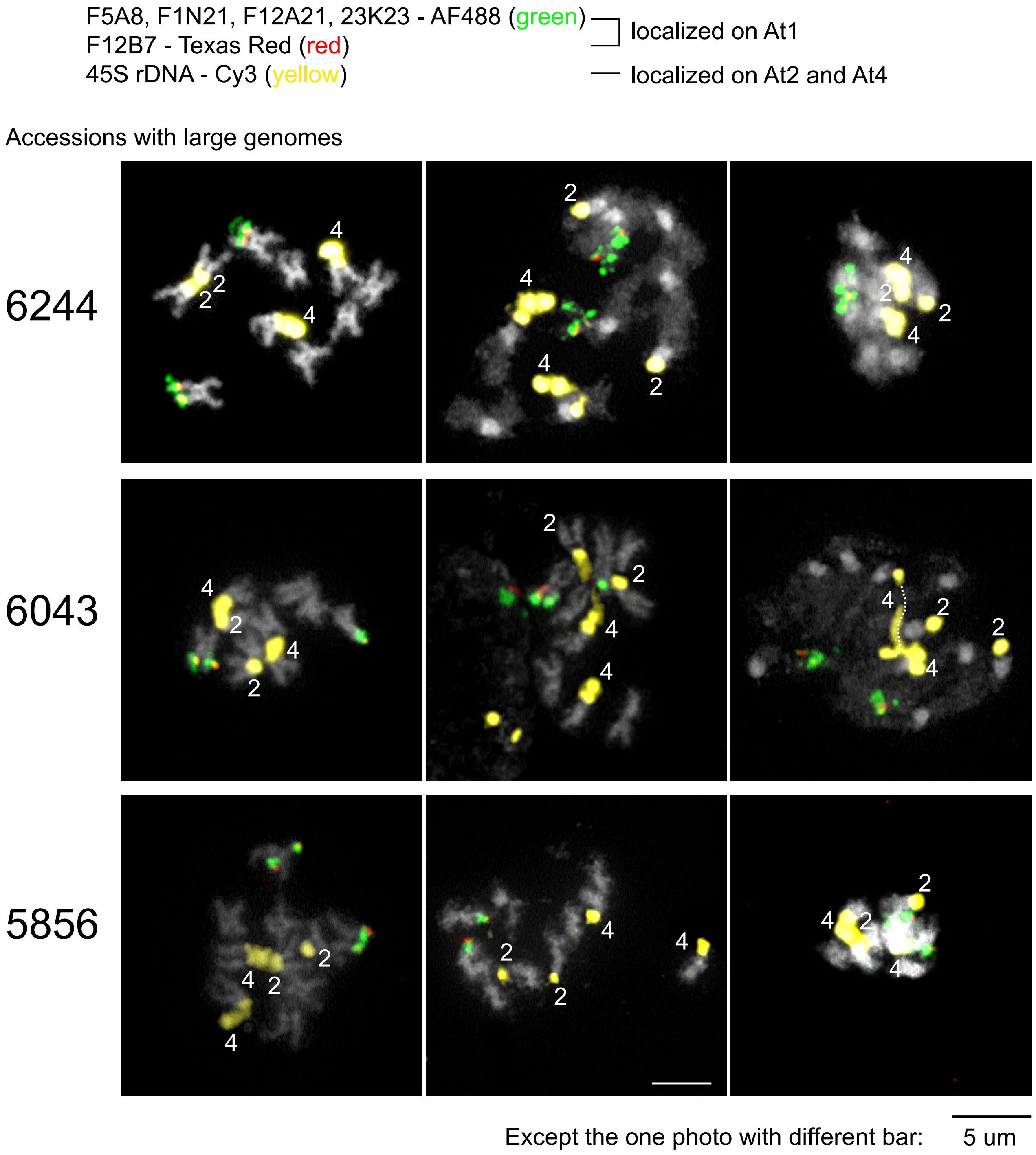
Supplementary Figure 1 Map of sampling locations. For a list of lines used, see project download site.



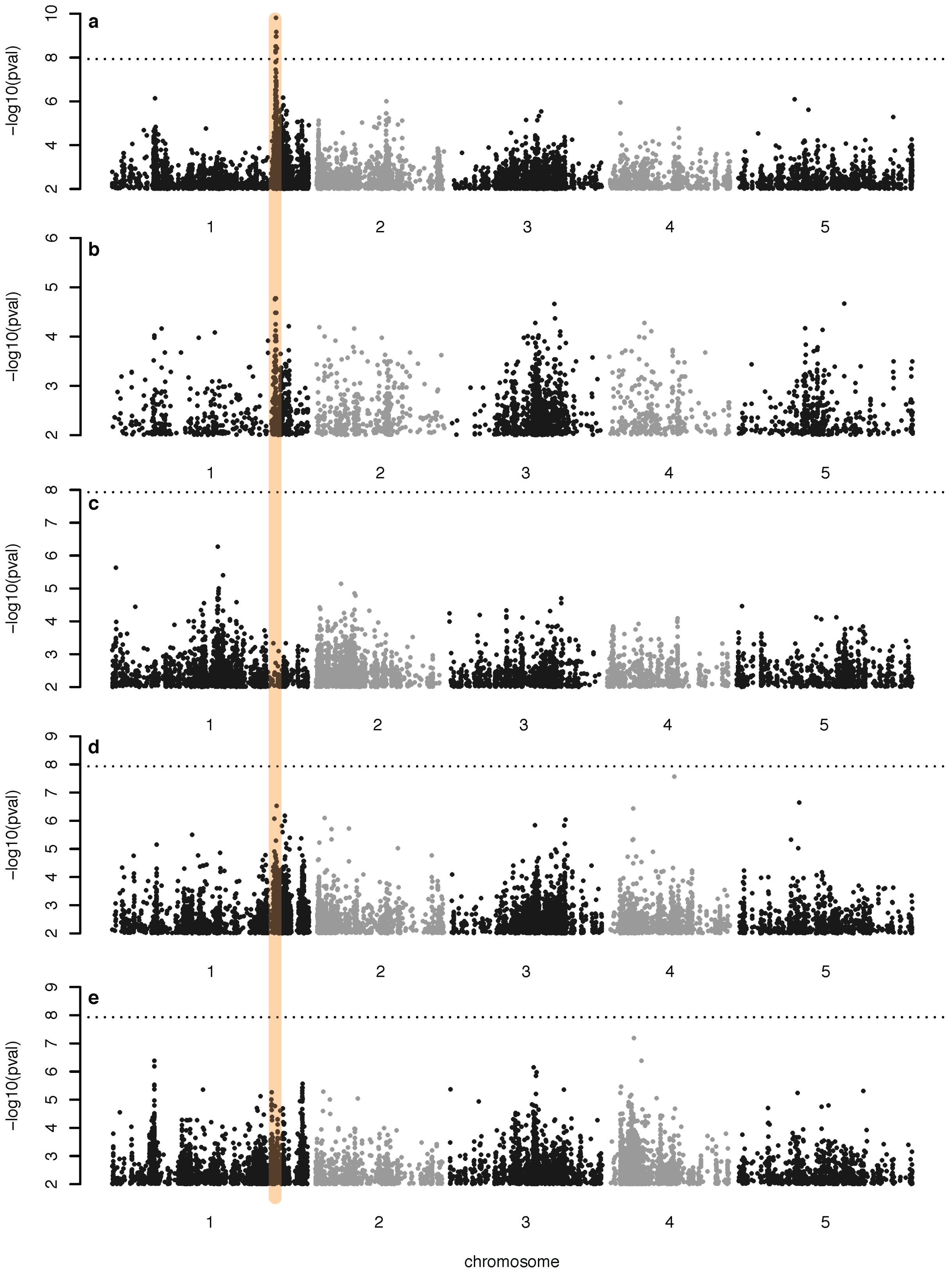
Supplementary Figure 2 Chromosomal distribution of positive (blue) and negative (green) indels, compared with SNP polymorphism levels (black). From top to bottom, chromosomes 2–5. Results for chromosome 1 are in the main text (**Fig. 1f**).



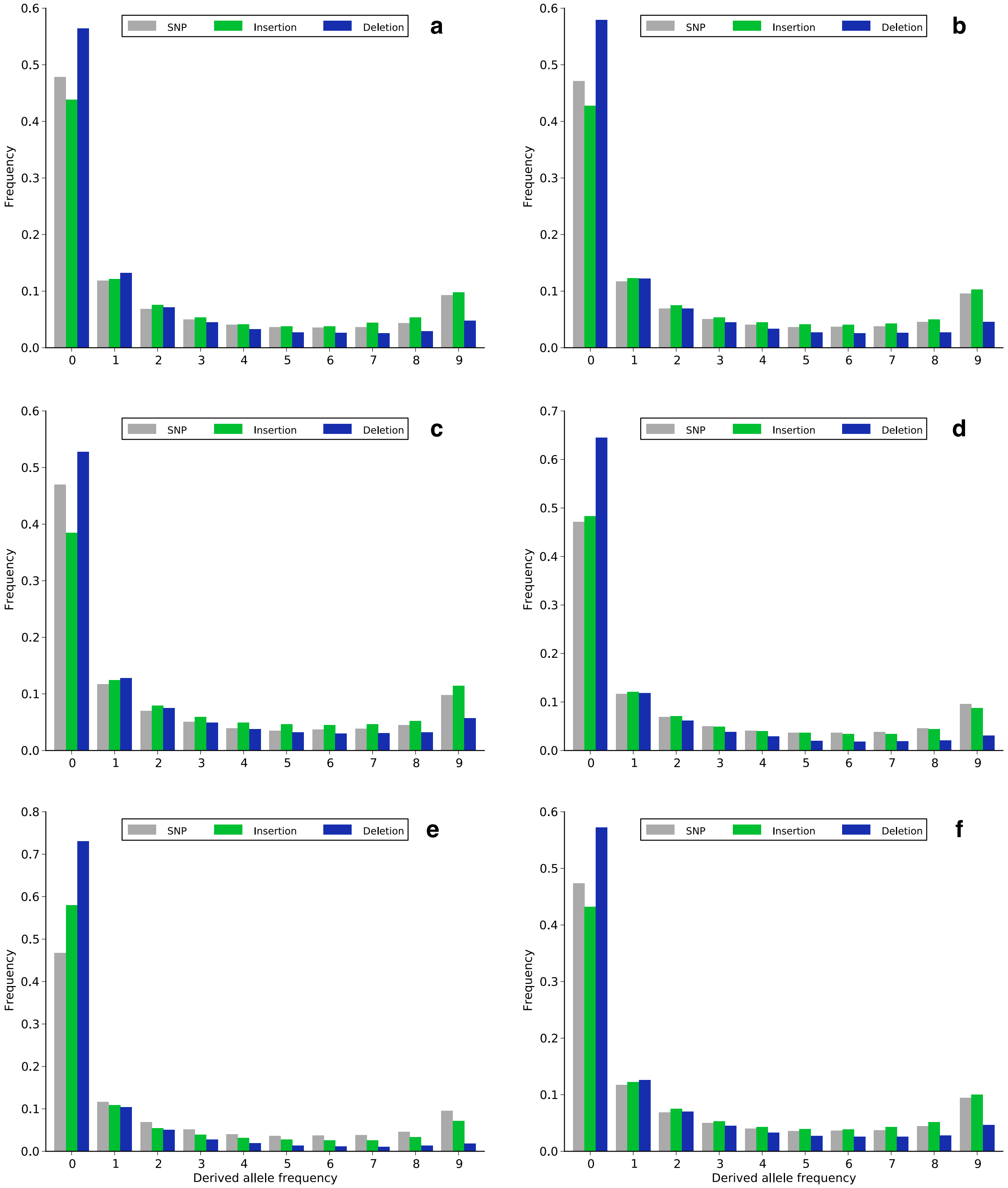
Supplementary Figure 3 The distribution of flow cytometry estimates for the global (36), southern (90), and northern (38) lines. The mean DNA content for the three samples are 166, 167, and 170, respectively. The northern measurements are significantly larger than those in the other two samples (Mann-Whitney p-value $< 10^{-4}$), whereas are southern sample is not significantly different from the world-wide sample.


F5A8, F1N21, F12A21, 23K23 - AF488 (green)
F12B7 - Texas Red (red)
45S rDNA - Cy3 (yellow)
localized on At1
localized on At2 and At4
Accessions with large genomes
6244
6043
5856
2
4
Except the one photo with different bar:
5 um

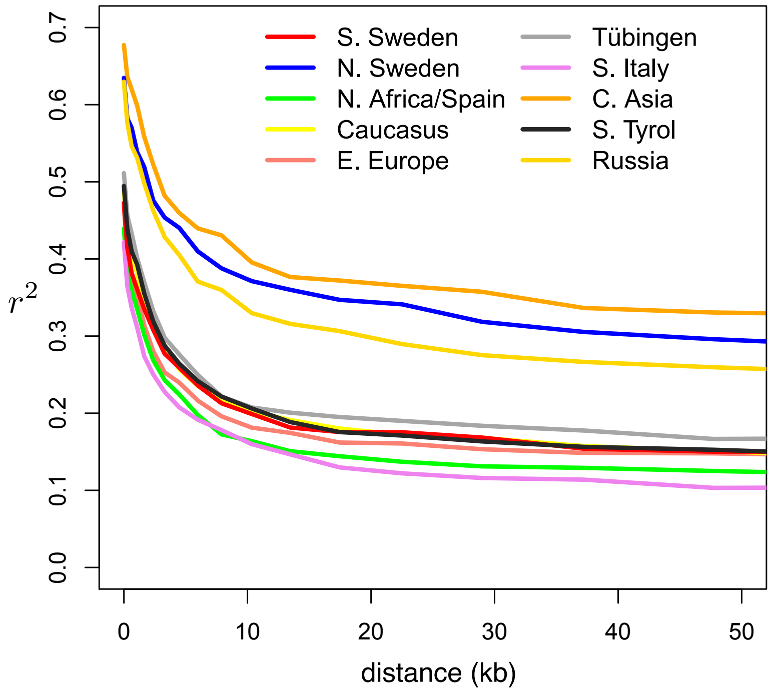
Supplementary Figure 4 FISH results for three lines with large estimated 45S rDNA copy number: 6244 (TRÄ 01); 6043 (Löv-1), and; 5856 (Dör-10). The known 45S rDNA clusters on chromosome 2 and 4 are highlighted (in yellow), as is chromosome 1 (green) and the region on chromosome 1 that contains the main association peak (red). There is no evidence for any novel 45S rDNA clusters.



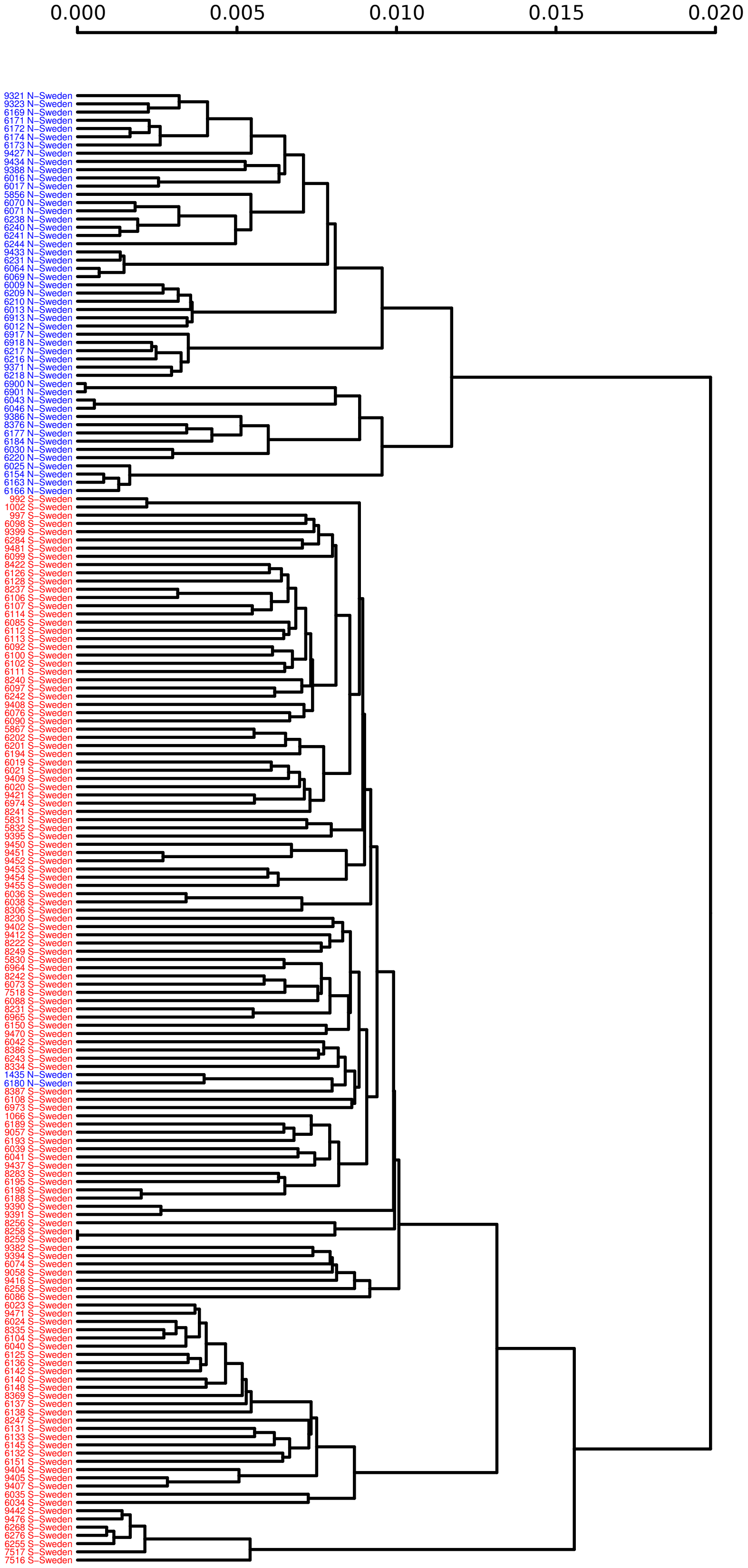
Supplementary Figure 5 Comparison of GWAS results for genome size using different samples and phenotypes. (**a**) GWAS using flow cytometry estimates for 128 lines (same as **Fig. 2b**). (**b**) GWAS using flow cytometry estimates for 38 northern lines. (**c**) GWAS using flow cytometry estimates for 90 southern lines. (**d**) GWAS using 45S coverage estimates for the 128 lines for which flow cytometry estimates are available. (**e**) GWAS using 45S coverage estimates for the full sample of 180 lines. The association peak discussed in the text is marked by the horizontal orange bar. It is clearly present in **a**, **b**, and **d**, but not in **c** and **e**.



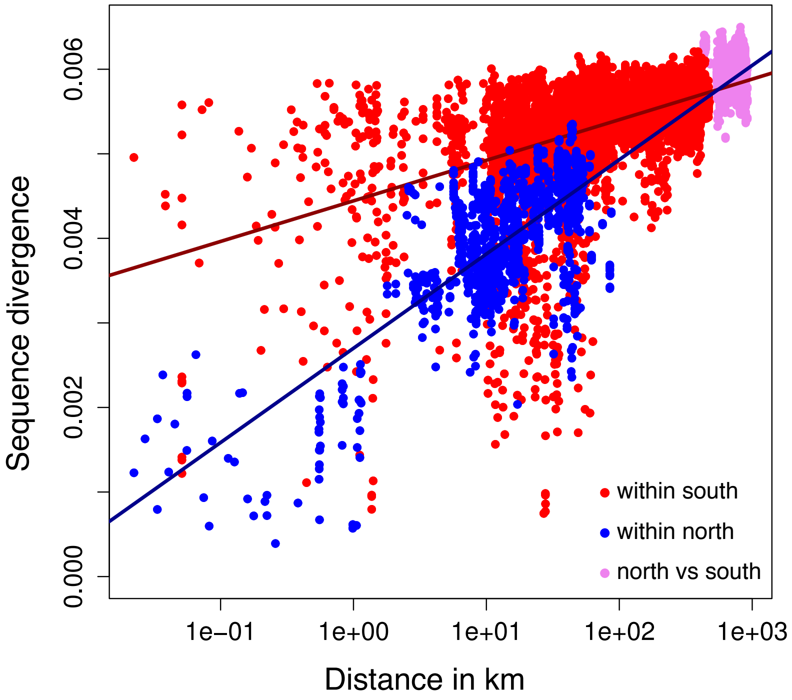
Supplementary Figure 6 Derived allele frequency distributions for different parts of the genome. The distributions were estimated by dividing the frequency into 100 bins, correcting the estimates using the hypergeometric distribution[1], and averaging across bins to create bins for the plot. Very conservative criteria were used to polarize the polymorphisms in order to avoid inflation at the right end of the plots. (**a**) Intergenic regions. (**b**) Genic regions. (**c**) Introns. (**d**) Exons. (**e**) CDS. (**f**) Whole genome. There is no evidence for high-frequency deletions in any category.



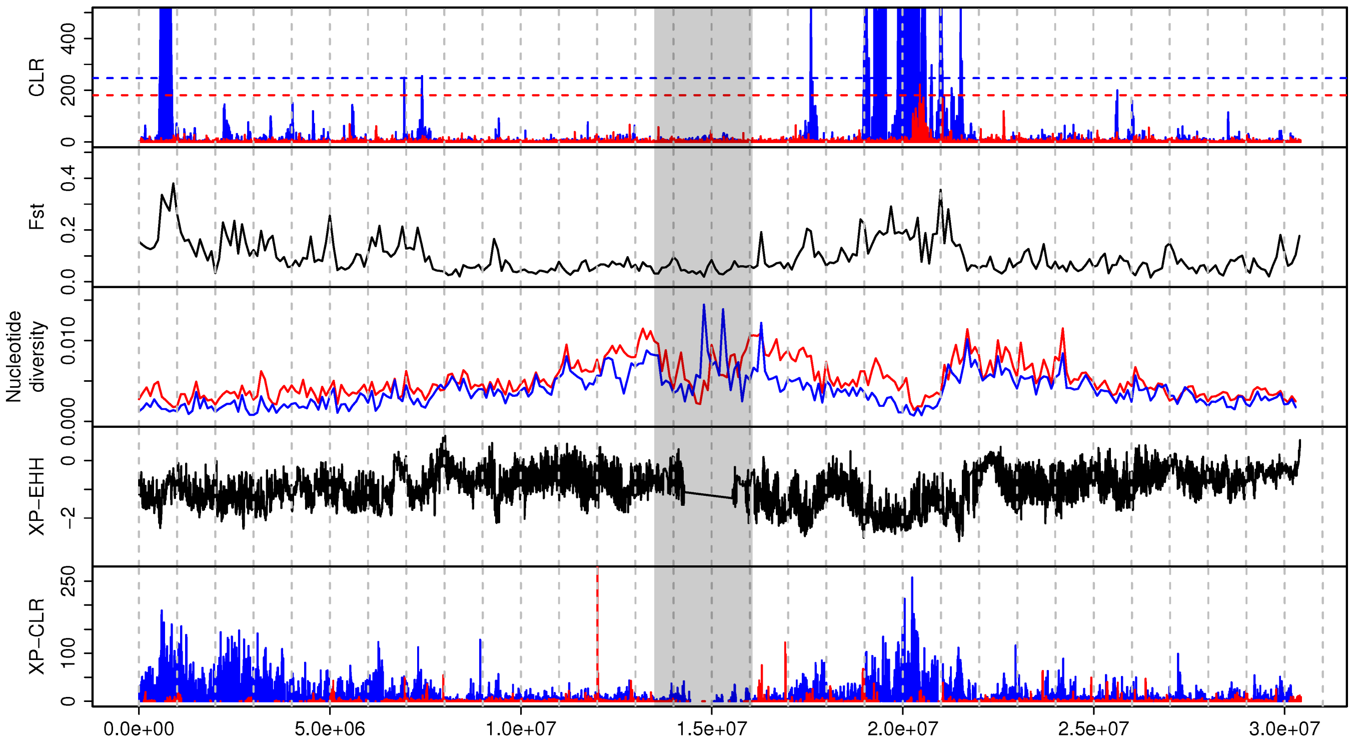
Supplementary Figure 7 Decay of LD for different sub-populations (*cf.* Fig. 1 in reference [2] and Fig. 3 in reference [3]). Sub-sampling was used to ensure equal sample sizes in all sub-populations.



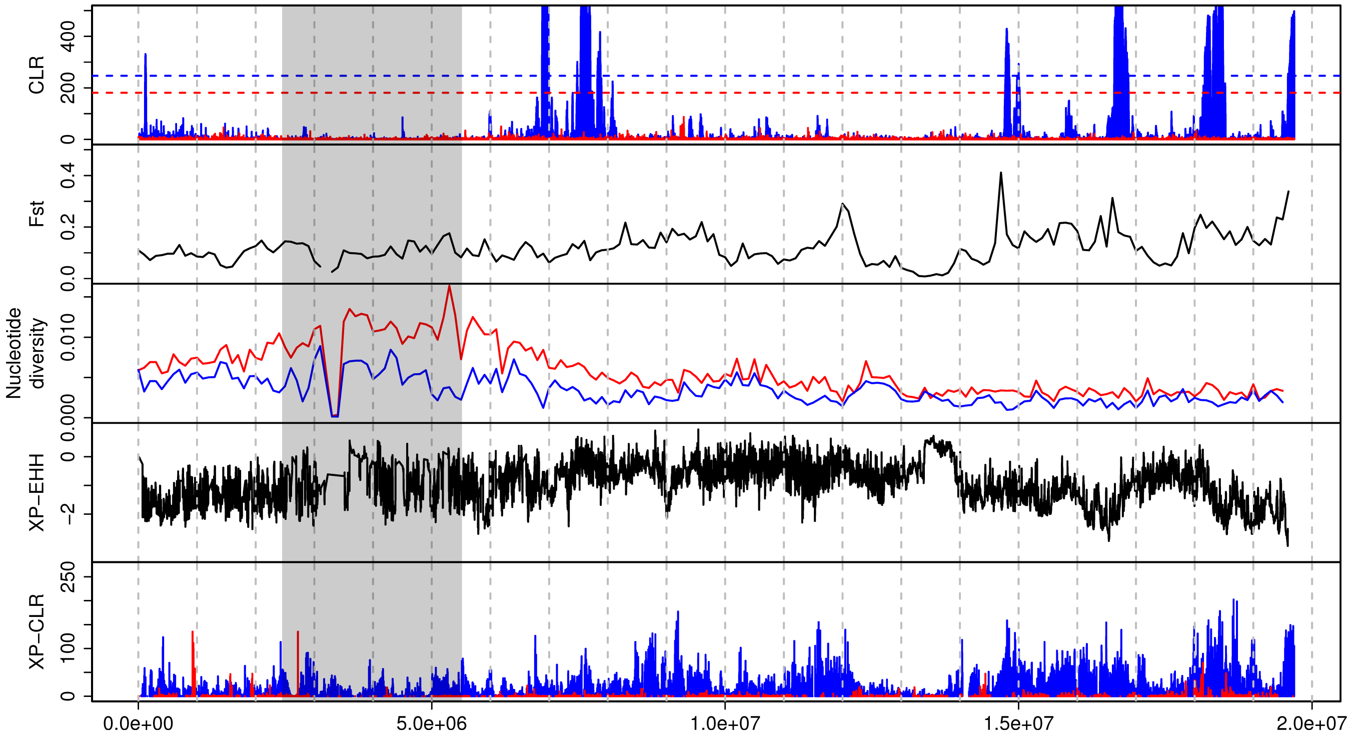
Supplementary Figure 8 UPGMA clustering of the Swedish lines based on all SNPs. Northern lines are labeled blue and southern red.



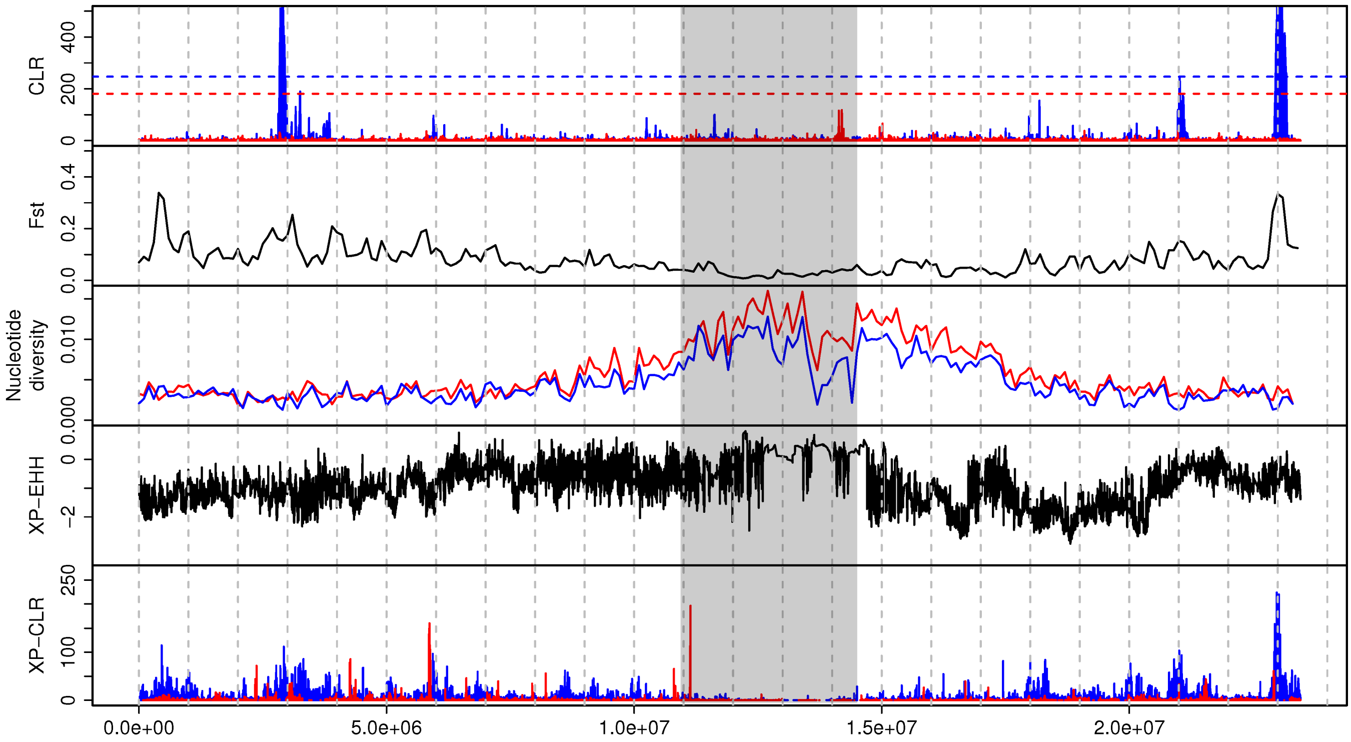
Supplementary Figure 9 Isolation by distance. Sequence divergence (proportion of sites that differ) between all pairs of Swedish lines as a function of the distance between the sample locations. Lines are fitted using least squares regression. The geographic distance between pairs was calculated from the geodesic distance using the Vincenty inverse formula for ellipsoids (function gdist of the R-package lmap).



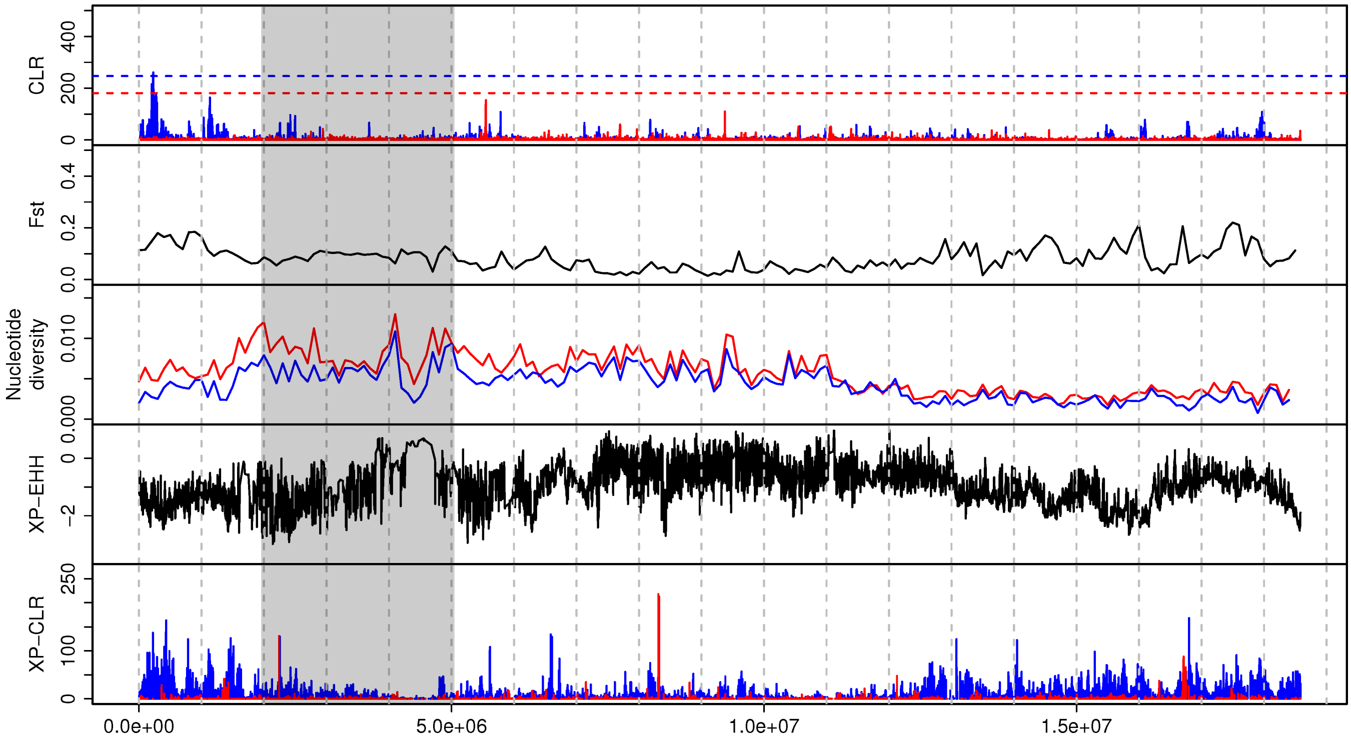
Supplementary Figure 10 Sweep statistics for chromosome 1. Blue curves are for northern Sweden, red for southern Sweden. For CLR, dashed lines indicate statistical cut-offs from simulations. XP-CLR and CLR were calculated for grid points with 1 kb spacing, XP-EHH was calculated for each SNP, and F_{ST} (between north and south) and nucleotide diversity were calculated in non-overlapping 100 kb windows.



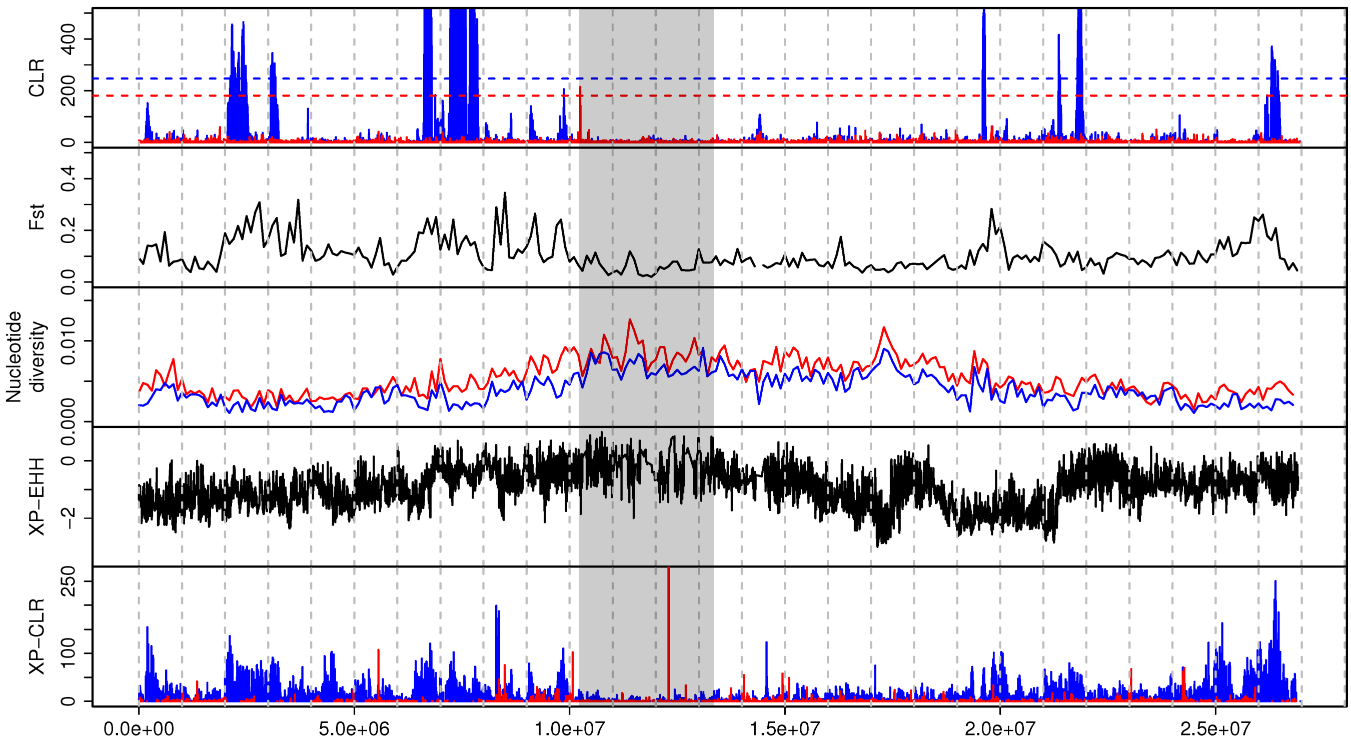
Supplementary Figure 11 Sweep statistics for chromosome 2. See **Supplementary Fig. 10** for details.



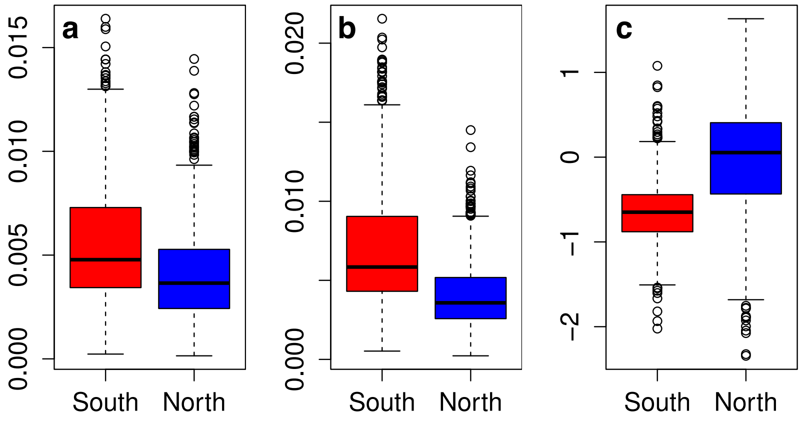
Supplementary Figure 12 Sweep statistics for chromosome 3. See **Supplementary Fig. 10** for details.



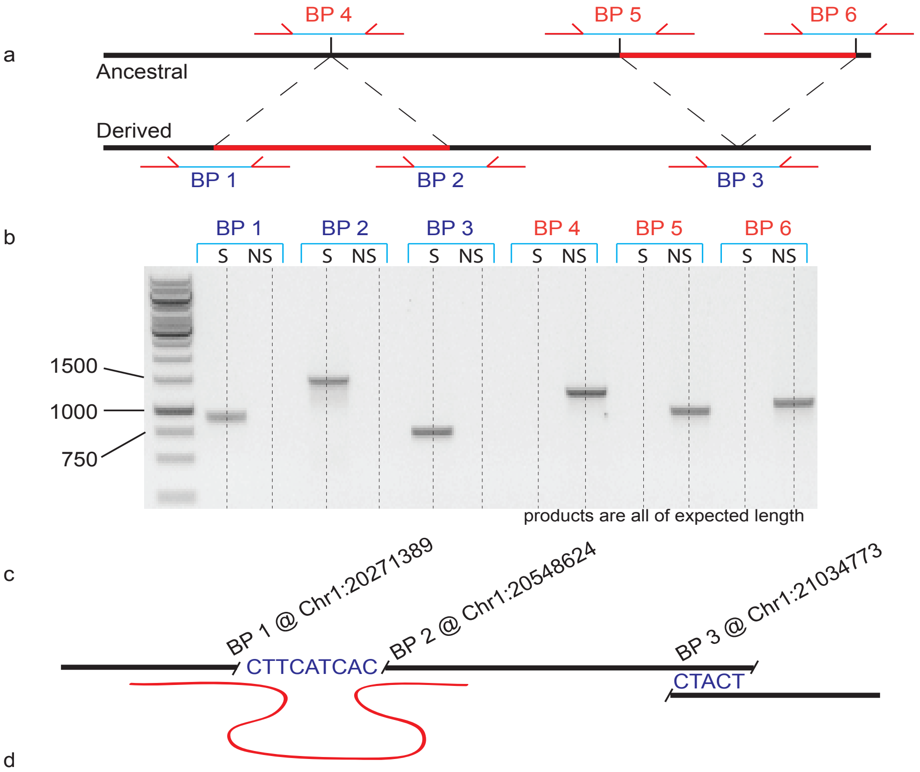
Supplementary Figure 13 Sweep statistics for chromosome 4. See **Supplementary Fig. 10** for details.



Supplementary Figure 14 Sweep statistics for chromosome 5. See **Supplementary Fig. 10** for details.



Supplementary Figure 15 Summaries of the pattern of polymorphism in north and south, in 100 kb windows across the genome. (**a**) Nucleotide diversity, π. (**b**) Watterson's θ. (**c**) Tajima's D. Only non-coding sites were used for these plots.



d

>Breakpoint 1, derived/ref. Chr1:20270340..20270440

```
gaattacttagcctcaaaatcctcacaacaactcttctcttgtagctacatctactttgctcttccctctctgttgtctcgcgtttcctctgctgtctcgt
```

>Breakpoint 2, derived/ref. Chr1:20548574..20548674

```
agttgggtatagtcttagagcttacaatgataaatttaaagtgaaatacaacattttcccgtagcacagataaagtctgtcgccctttacgtggactggtt
```

>Breakpoint 3, derived/ref. Chr1:21034721..21034821

```
tgttgatgaggtatatcgtctactttgctcaattctctctgtttctactccataatatgaagtattccgaaatttgtttttttcaaaatcccatcaattta
```

>Breakpoint 4 ancestral, unswept sequence from lov_5. ~9bp underlined was likely deleted during the translocation. Palindrome/repeat flanking BP(blue/green).

```
GAGGCTCTTTTTCACCTTGGAATTTCGAGAAAAGCTCTACCTAGCCTCAAAATCCTCACAACAACTCTTCTCTTGTAGCTACACTTCATCACACATTTTCC
CGTAGCACAGATAAAGTCTGTCGCCCTTTACGTGGACTGGTTTGTCATCCATTTTTAG
```

```
>Breakpoint 5 ancestral unswept sequence
AAGGAATCTCCAGGTGGCGGCGACCATAACAGATTTGTTGATGAGGGATATCGGCTACTTTGCTCAATCTCTCTGTTCTACTTTGCTCTTCCCTCTCTGTT
GTCTCGCGGTTCCTCTGGCTGTCTCGGGCGCTCTTTTCGGTCTGTTTTGG
```

>breakpoint 6 ancestral

```
AGACCTCGGATTTGTGAGTTGGGTATAGTCTTAGAGCTTACAATGATAAATTAAAAGTGAAATACAAATCTTTCCTGACTACTCCATAAATATGAAGTATT
CCGAAATTTGTTTTTTTCAAATTCCCATCAATTTACCTAAAATTGATAG
```

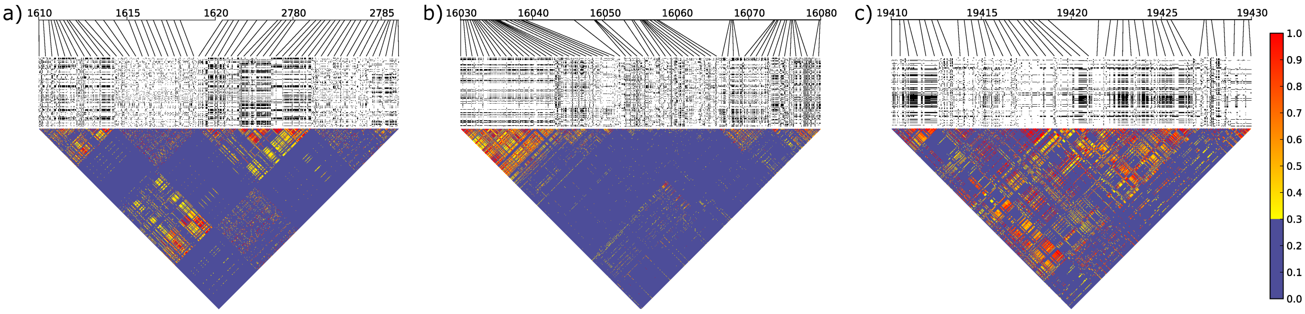
Alignment showing how BP5 and BP6 joined to generated BP3. 5bp of micro-homology (green) may have facilitated NHEJ, leading to a ~5bp deletion. Other repeat structures shown in red and yellow.

```
                                               /TTGCTCTTCCCTCTCTGTTGTCTCGCGGTTCCTCTG
BP5. TGAGGGATATCGGCTACTTTGCTCAA-TCTCTCTGTT-CTACT
BP3. TGAGGTATATCGTCTACTTTGCTCAATTCTCTCTGTTTCTACTCCATAA-TATGAAGTATTCCGAAATTTGTTTTTTTCAAAA
BP6.                                       CTACTCCATAAATATGAAGTATTCCGAAATTTGTTTTTTTCAAAT
        GATAAATTAAAAGTGAAATACAAATCTTTCCT-GA/
```

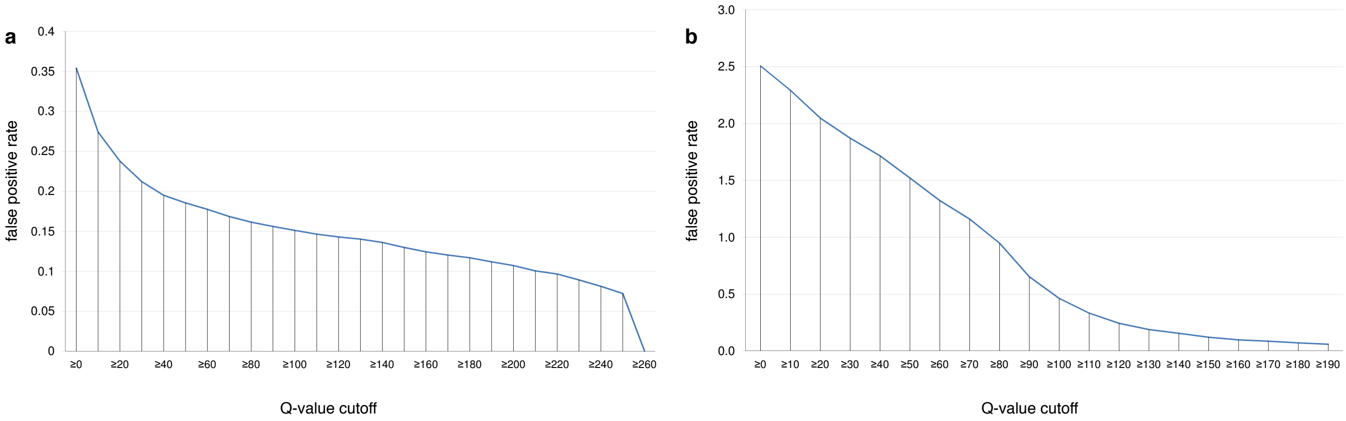
Alignment between the ancestral BP4 with the ends of the translocation 'insertion' (BP1 & BP2) showing the 9bp deletion (bold).

```
                                            /TCTACTTTGCTCTTCCCTCTCTGTTGTCTCGCGTTTCCTC
BP1. GCCTCAAAATCCTCACAACAACTCTTCTCTTGTAGCTACA
BP4. GCCTCAAAATCCTCACAACAACTCTTCTCTTGTAGCTACACTTCATCACACATTTTCCCGTAGCACAGATAAAGTCTGTCGC
BP2.                                                   ACATTTTCCCGTAGCACAGATAAAGTCTGTCGC
                      AGCTTACAATGATAAATTTAAAGTGAAATACA/
```

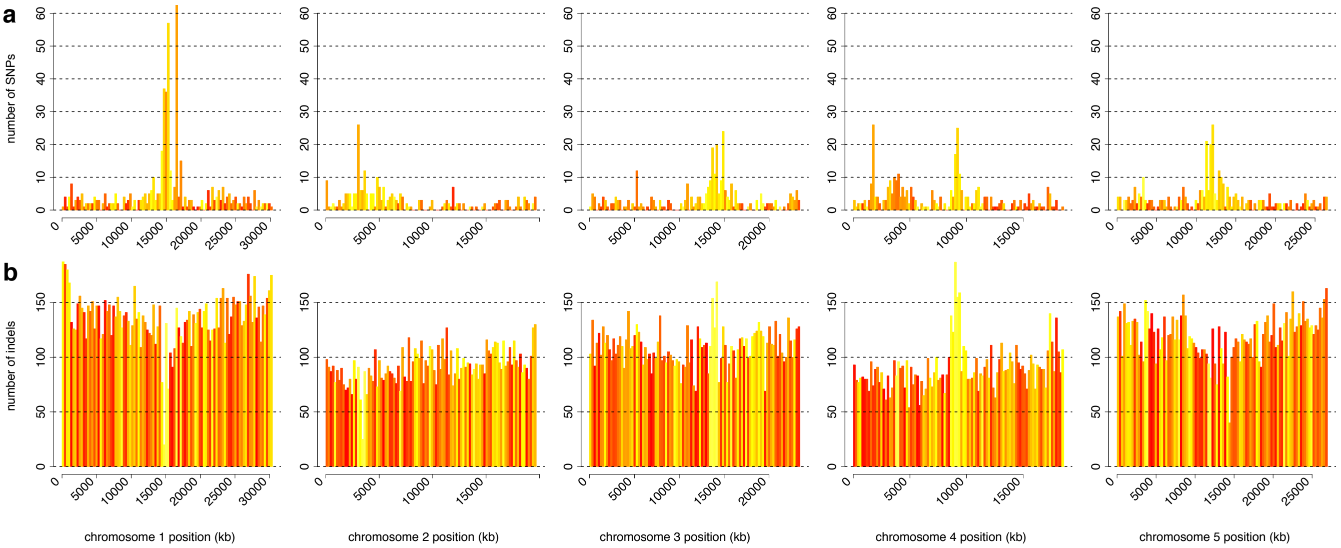
Supplementary Figure 16 Validation of transposition. (**a**) PCR primers used to confirm breakpoints. (**b**) PCR from lines 6124 (swept) and 6046 (unswept) showing products of expected length consistent with the predicted arrangement. Results were consistent in 15 swept and 5 unswept lines (not shown). (**c**) Likely mutational events that generated this transposition. Removal of the 278 kb and rejoining at breakpoint 3 was likely facilitated by 5 bp of microhomology, leading to the deletion of one copy. Insertion/capture of the 273 kb between breakpoint 1 and 2 lead to the deletion of 9 bp. (**d**) Sequence flanking all 6 breakpoints (1–3 from Col-0; 4–6 Sanger sequenced from 6046), and alignments in support of the predicted arrangement (with identity underlined and repeat structures in color).



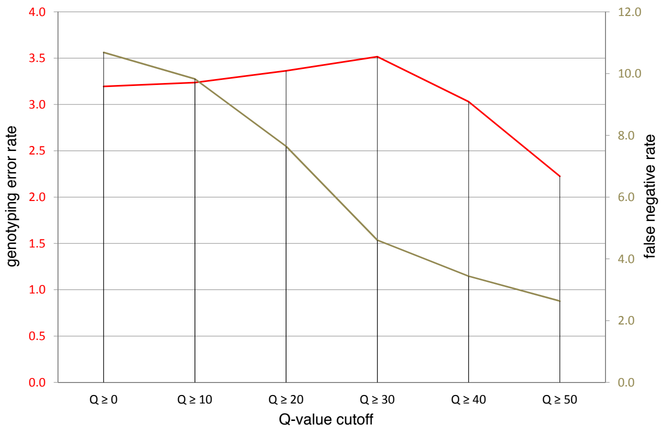
Supplementary Figure 17 LD surrounding putative major rearrangements identified by *de novo* assembly. (**a**) The 1170kb region between chr4:1612606 and chr4:2782618 appears to be inverted in 171 of our 180 Swedish lines, and comparison with *A. lyrata* shows that it is the *A. thaliana* reference genome that carries the inversion. (**b**) The 28 kb region between chr3:16039026 and chr3:16067070 appears to be missing with frequency of 106/180. The region contains several annotated transposable elements but also At3g44400, a possible resistance protein. (**c**) The 14–15 kb region between chr5:19412000 and chr5:19426000/19427000 appears to be inverted in 140/180 lines.



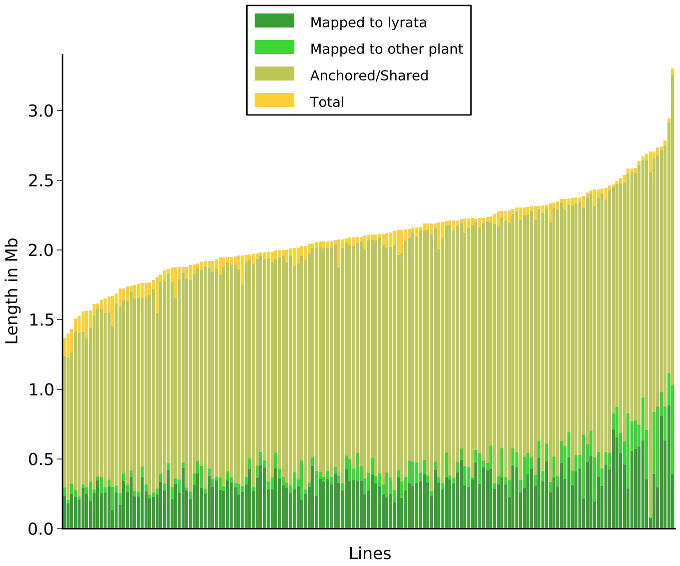
Supplementary Figure 18 False positive rates in reference re-sequencing as a function of Q-value cut-off. (**a**) SNPs. (**b**) Indels.



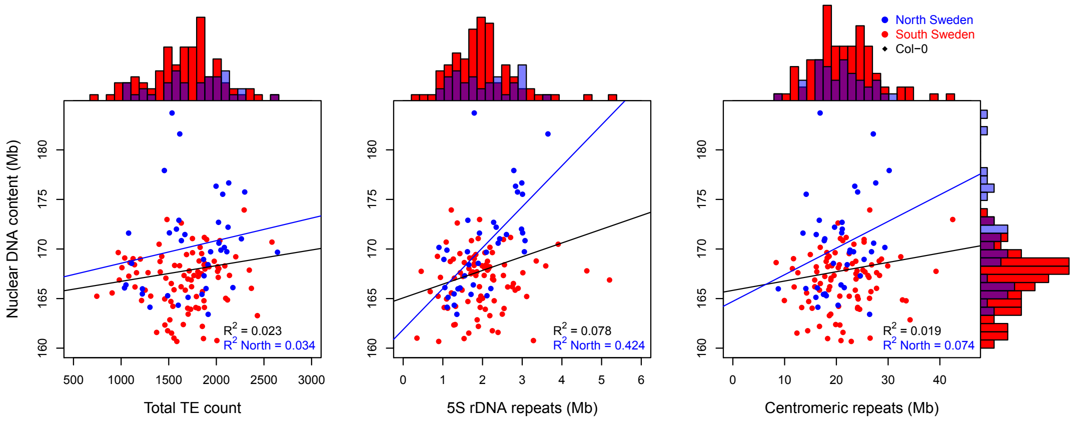
Supplementary Figure 19 The distribution of putative false positives from reference re-sequencing (for which all polymorphisms are assumed to be false) as a function of chromosomal position and alignment quality. (**a**) SNPs. (**b**) Indels. Colors indicate mapping quality (Q value: red is high, yellow low). For SNPs (but not indels), false positives are clustered near centromeres and tend to have low Q values.



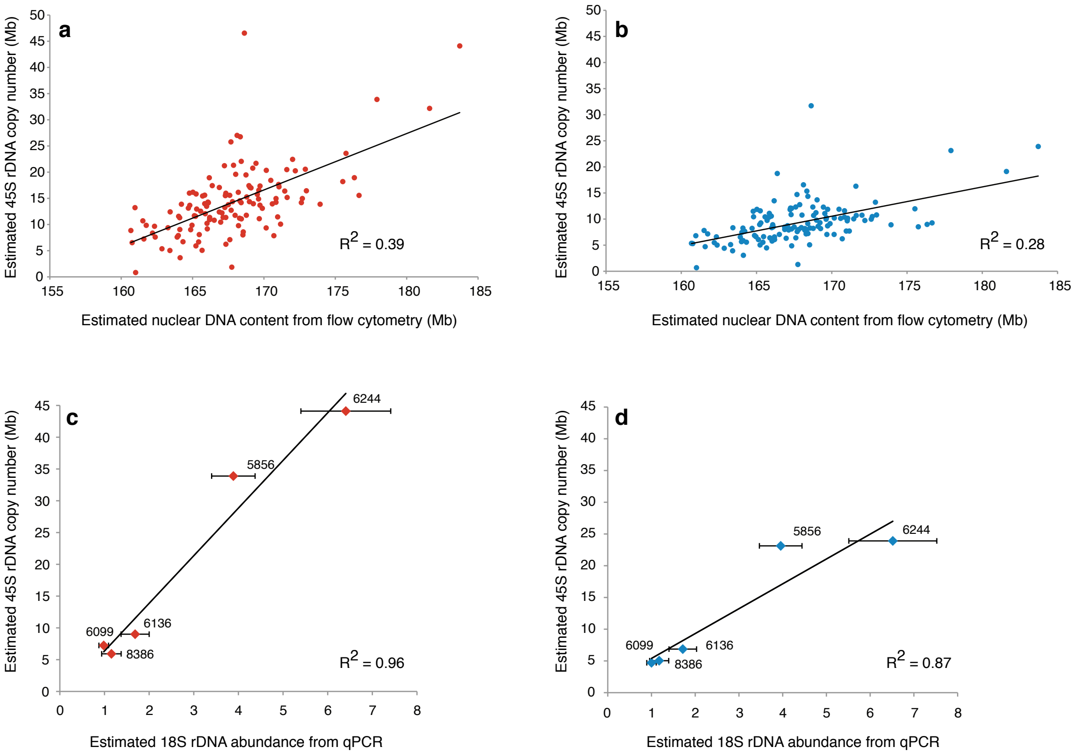
Supplementary Figure 20 Error rates from comparison with SNP data as function of Q-value.



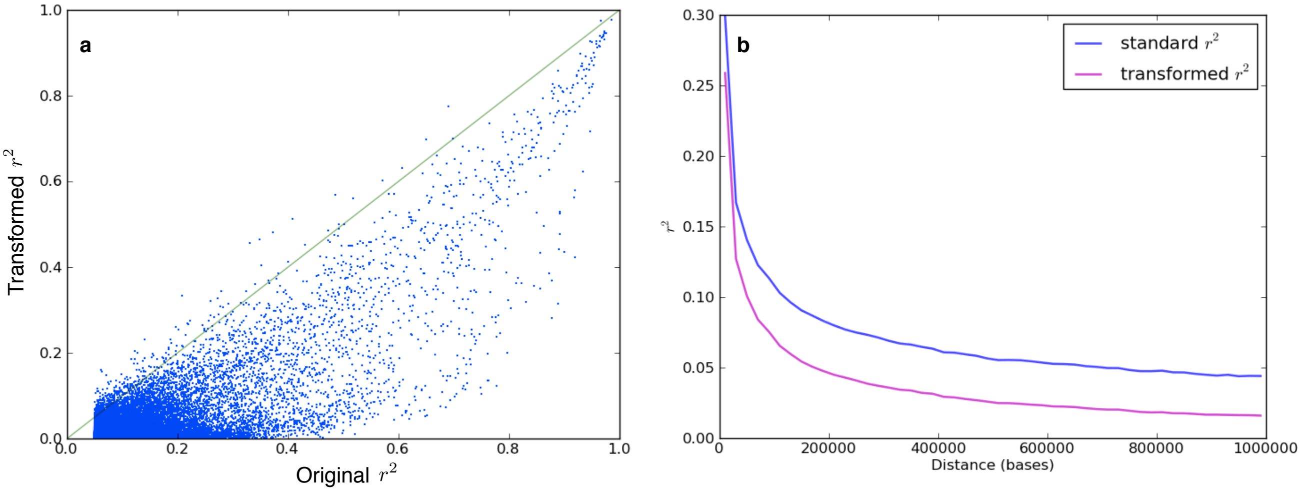
Supplementary Figure 21 Characterization of novel sequence. Note that the light/dark green bars are independent of the olive green bars. See **Supplementary Note** for details.



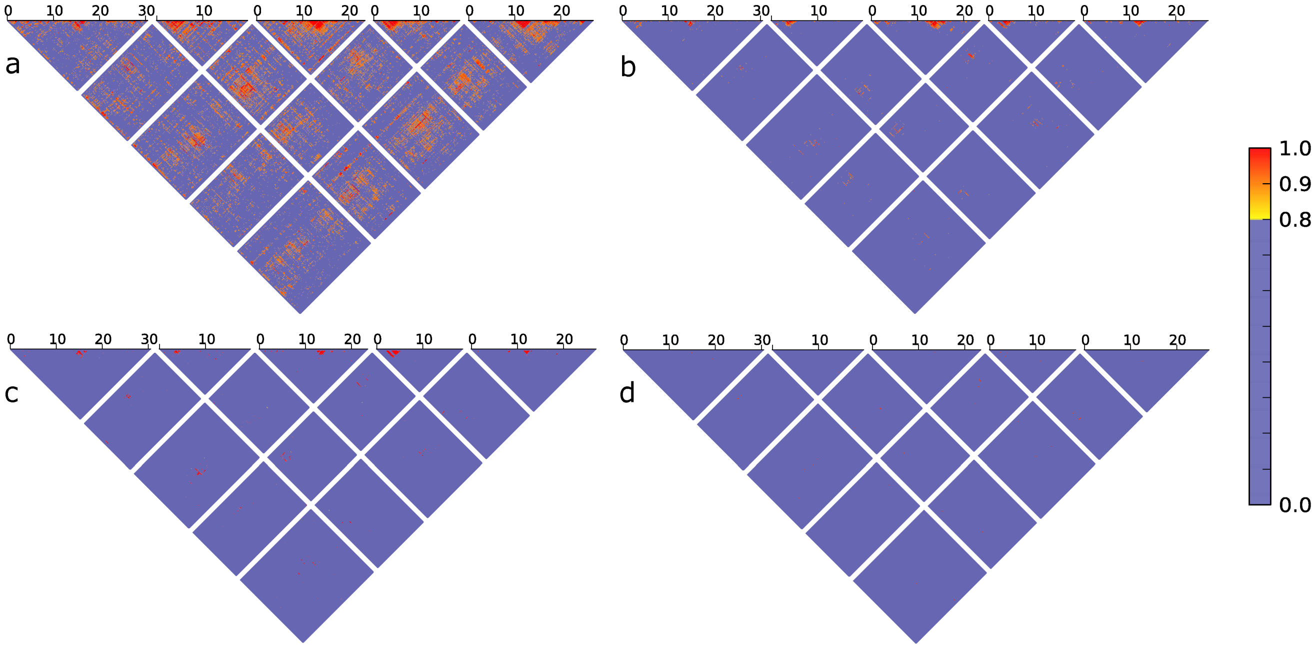
Supplementary Figure 22 Correlation between flow cytometry and repeat copy number estimates. For 45S rDNA, see **Fig. 2a**. Note that the TE count refers to the number of novel insertions detected.



Supplementary Figure 23 Correlation of flow cytometry and 45S rDNA. (**a**) Correlation between flow cytometry and 45S rDNA copy number estimates before removal (**a**) and after removal (**b**) of duplicated reads. Correlation between 45S rDNA copy number and 18S rDNA qPCR estimates (**c**) before and (**d**) after removal of duplicated reads.



Supplementary Figure 24 Effect of LD transformation. (**a**) The plot shows 16 million pairs of SNPs, selected at random from pairs with $r^2 > 0.05$ in the original data. (**b**) Average decay of LD.



Supplementary Figure 25 The robustness of long-range LD. The correction for population structure described in section 5.2.1 was used throughout. (**a**) LD in Northern Swedish population only. Due to smaller sample size, a higher minor allele frequency cutoff of 0.12 was used. (**b**) LD in Southern Swedish population only (minor allele frequency cutoff of 0.10). (**c**) LD in unimputed data, illustrating similarity to plot for imputed data (Fig. 4a in main text). (**d**) LD calculated from imputed data, but removing all high LD pairs from the previous plot. This “subtracted” LD plot, shows that imputation creates few additional high LD pairs.

Supplementary Table 1 Estimated error rates (%) for SNPs and short indels. In the case of indels, separate estimates are given for variants that are longer/shorter than the reference genome whenever enough polymorphisms were observed. For SNPs, all comparisons are with the data from the Q30 regions.

	False positives		False negatives		Genotyping	
Quality control data	SNPs	Indels +/-	SNPs	Indels +/-	SNPs	Indels +/-
Reference re-sequencing	0.21	2.5/1.4	NA	NA	NA	NA
SNP-chip	NA	NA	4.6	NA	3.5	NA
Sanger-sequenced PCR amplicons	1.8	0.8/1.6	3.7	8.3/5.1	0.64	3.9/1.8
Sanger-sequenced random clones	1.8	2.4	1.4	16.3	NA	NA

Supplementary Table 2 Validation of long-range LD. Six different sets of SNPs exhibiting long-range LD were tested in crosses: the number give r^2 in the natural as well as the F2 population (where $r^2 = 0$ is expected under independent segregation). "ND" means that PCR amplification failed so that genotyping was not possible.

							r^2					
	SNP 1		SNP 2		SNP 3		1 vs 2		1 vs 3		2 vs 3	
centromeric	chr	position	chr	position	chr	position	Sweden	F2	Sweden	F2	Sweden	F2
N	1	724,571	2	8,985,116	2	9,210,944	1	0.01	1	0.01	1	1
N	4	958,430	5	7,723,187	5	7,737,565	1	0.08	1	0.08	1	1
Y	3	13,649,143	1	13,341,722	1	13,359,206	0.94	1	0.94	1	1	1
Y	1	14,168,125	2	3,086,551	NA	NA	1	ND	NA	NA	NA	NA
Y	4	4,738,528	1	12,831,211	NA	NA	1	ND	NA	NA	NA	NA
Y	1	12,964,602	5	14,083,994	5	14,088,003	1	ND	1	ND	1	ND

Supplementary Table 3 p-Values for enrichment of long-range LD SNPs among SNPs associated with climate.

	Distance to peak	
Climate variable	5 kb	10 kb
consecutive frost free days	0.3476	0.2751
daylength	0.7466	0.6952
maximum temperature	0.0995	0.1151
minimum temperature	0.3534	0.4414
length of growing season	0.3224	0.3682
consecutive cold days	0.3467	0.4693
relative humidity	**0.0034**	**0.0033**
photosynthetically active radiation	0.4332	0.2992
temperature seasonality	0.5397	0.4812
maximum precipitation	0.4922	0.4503
precipitation seasonality	0.0924	0.0968
minimum precipitation	**0.0002**	**0.0011**
aridity	0.3373	0.3855

Supplementary Table 4 Summary of simple regressions of flow cytometry estimates.

	R	R^2	P-value
45S rDNA	0.62	0.39	5.6×10^{-15}
5S rDNA	0.28	0.078	1.4×10^{-3}
Centromeres	0.14	0.019	0.12
TEs	0.15	0.023	0.086

Supplementary Table 5 Genetic and geographic distance between world-wide populations. The distance between the population (in km) is above the diagonal, and F_{ST} is below.

	N. Sweden	S. Sweden	Spain	S. Italy	Tübingen	S. Tyrol	E. Europe	Caucasus	Russia	C. Asia
N. Sweden	—	796	3072	2589	1701	1884	2118	3129	2355	4046
S. Sweden	0.118	—	2361	1828	914	1092	1538	2907	2585	4352
Spain	0.110	0.038	—	1822	1546	1571	2582	4302	4705	6409
S. Italy	0.171	0.056	0.048	—	1113	836	985	2554	3291	4879
Tübingen	0.161	0.036	0.040	0.044	—	280	1337	3035	3187	4930
S. Tyrol	0.152	0.045	0.048	0.067	0.051	—	1139	2861	3152	4873
E. Europe	0.130	0.027	0.029	0.030	0.029	0.027	—	1801	2369	3979
Caucasus	0.158	0.057	0.039	0.043	0.055	0.064	0.029	—	1657	2601
Russia	0.185	0.104	0.079	0.116	0.120	0.087	0.049	0.101	—	1883
C. Asia	0.236	0.145	0.112	0.148	0.161	0.133	0.087	0.131	0.068	—

Supplementary note

Q. Long, F. A. Rabanal, D. Meng, C. D. Huber, A. Farlow *et al.*

May 31, 2013

1 Genome sequencing

Genomic DNA was extracted from the roots of young Arabidopsis seedlings (5-8 plants for each line were pooled) grown on sterile 1/2 MS plates with 1% sugar at room temperature. Roots were ground to a fine powder in liquid nitrogen and mixed with 2X CTAB DNA extraction buffer (100 mM Tris-HC1 pH 8, 1.4 M NaCl, 20 mM EDTA pH 8.0, 2% CTAB). After incubation at 65°C for 30 minutes, DNA was extracted with equal volume of chloroform and precipitated with 0.8 volume of isopropanol.

Libraries were prepared using slightly modified Illumina Genomic DNA Sample Prep protocol. Briefly, DNA was fragmented by sonication with Bioruptor (Diagenode); the peak of fragment sizes was about 400 bp. End-repair of sheared DNA fragments, A-tailing and adapter ligation were carried out with NEBNext DNA Sample Prep Reagent Set 1 (BioLabs). Adaptor-modified DNA was resolved on 2% low melt agarose (Peqlabs) gel (including SybrGold nucleic acid gel strain, Invitrogen), run for 90 minutes at 100 V. Library DNA was size-selected to 450–800 bp via gel extraction with a MinElute Gel Extraction Kit (Qiagen). The paired-end DNA libraries were amplified by PCR for 14–18 cycles with Illumina supplied PCR primers 1.1 and 1.2. Libraries were sequenced on Illumina GAII and HiSeq Analyzers using manufacturer's standard cluster generation and sequencing protocols, with either 76 or 100 bases read length.

2 Polymorphism detection

2.1 Initial read mapping

2.1.1 Read mapping and SNP discovery

We first mapped all reads to the TAIR10 reference genome using BWA (version 0.5.9)[4], allowing up to 4% mismatches and 1 gap. We tried trimming reads using different parameters before finally choosing the default parameters of BWA. After that, we used the rmdup function of Samtools (version 0.1.6)[5] to remove reads that are duplicated in library preparations or sequencing. Since it turned out that not removing duplicated reads in highly duplicated regions, i.e., ribosome repeats and centromere repeats, improved the quality of coverage estimates (see **Supplementary Fig. 23**), we retained the reads that are potentially library duplications in those regions.

SNPs and indels were initially called using GATK[6] with the default parameters of UnifiedGenotyper and IndelGenotyperV2, respectively. With these variants called, we run the GATK local realignment (function IndelRealigner) to refine the reads mapping in the presence of the variants. After that, we call SNPs using the pileup function of Samtools and run the varFilter function provided by samtools.pl in the Samtools package to filter low-quality calls. We found that the version of Samtools we were using had a small bug when setting heterozygosity to 0 (as is necessary for inbred lines). We worked around this by taking the SNP file provided by samtools (in pileup format) and counting the coverage of both alleles, then calling heterozygote when the minor allele count was at least 40% of the total. We do not make use of mpileup since we have our own population based pipeline (see below). Commands and detailed parameters used are listed in **Supplementary Command Listing 1**.

2.1.2 Quality calculations and filtering

The mapping quality (Q-value) calculated by BWA is intended to capture repetitiveness more than mismatches. For each base in the reference genome, we average the quality of all reads in all lines covering this position, resulting in a quality map. We then provide two versions of data: (i) the original version with Q-values attached; (ii) a filtered version where only $Q \geq 30$ is retained (see **Section 6** for details). The fraction of the genome retained in the filtered version is 87% and the fraction of SNPs and indels retained is 86% and 85%, respectively.

2.2 Reference-based structural variant discovery

We used several different reference-based methods to call indels and other structural variants (SVs), and we made use of population sharing to call variants in low-coverage regions. The different methods cover different sizes. We divide the indel/SV calling procedure into two phases: discovery and genotyping. The former is described in this section, and the latter in **Section 2.3**.

2.2.1 Short indels through local realignment

GATK calls short indels using local realignment. First-pass mapping will map different reads independently, which means that the same indel can be called in different ways for different reads within the same individual due to the mapper's local optimization. Re-alignment based on the first-pass information can alleviate this problem[6]. Thus, after running BWA, Samtools, and GATK, we use Samtools pileup and varFilter functions to call indels. Commands and detailed parameters are in **Supplementary Command Listing 1**.

It should be noted that we did not attempt to realign indels across lines. Instead, we simply filtered out with indels with more than two alleles.

2.2.2 Large SVs through paired-end reads

This method calls the big size structural variants based on abnormal read pairs. We ran BreakDancer[7] version 0.0.1r81 with default settings (**Supplementary Command Listing 1**).

2.2.3 All sizes of SVs through split reads

This method calls SVs of all lengths by re-mapping the reads that cross the breakpoints of the events. We applied Pindel[8]

BWA, Samtools and GATK (for each line)

```
% bwa aln -I -o 1 $ref $name.1.fastq > $name.1.sai
% bwa aln -I -o 1 $ref $name.2.fastq > $name.2.sai
% bwa sampe -a $insert -r $tag $ref $name.1.sai $name.2.sai $name.1.fastq $name.2.fastq
  -f $name.sam
% samtools view -bh -t $reflist -o $name.bam $name.sam
% samtools sort $name.bam $name.sort
% samtools index $name.sort.bam
% samtools rmdup $name.sort.bam $name.rmdup.bam
% samtools index $name.rmdup.bam
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T UnifiedGenotyper
  -I $name.rmdup.bam -o $name.gatk.snp.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T IndelGenotyperV2
  -I $name.rmdup.bam -o $name.gatk.indel.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -I $name.rmdup.bam
  -T RealignerTargetCreator -o $name.intervals -B:snps,VCF $name.gatk.snp.vcf
  -B:indels,VCF $name.gatk.indel.vcf
% java -Xmx2g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -I $name.rmdup.bam
  -T IndelRealigner -o $name.realigned.bam -targetIntervals $name.intervals
% samtools sort $name.realigned.bam $name.realigned.sort
% samtools index $name.realigned.sort.bam
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T UnifiedGenotyper
  -I $name.realigned.sort.bam -o $name.gatk.realigned.snp.vcf
% java -Xmx4g -jar /path/to/GATK/GenomeAnalysisTK.jar -R $ref -T IndelGenotyperV2
  -I $name.realigned.sort.bam -o $name.gatk.realigned.indel.vcf
% samtools pileup -c -s -r $r -f $ref $name.realigned.sort.bam > $name.realigned.pileup
% samtools.pl varFilter $name.realigned.pileup > $name.$r.realigned.var
% java -Xmx4g -jar correct_single_samtools_snp.jar $name.0.realigned.snp $name.0c.realigned.snp
```

Pindel (run for each chromosome separately)

```
% pindel -f $ref -i realigned_bam_config_files -c Chr$i -w 0.5 -u 0.1 -e 0.1 -b breakDancer_output
  -o output_Chr$i -Q confirm_Chr$i.txt
```

BreakDancer (for each individual)

```
% perl /path/to/breakdancer/bam2cfg.pl $name.sort.bam > $out_folder/bkdancer/$name.cfg
% perl /path/to/breakdancer/BreakDancerMax.pl $out_folder/bkdancer/$name.cfg
  > $out_folder/bkdancer/$name.bkd
```

TE-Locate

```
% perl TE_hierarchy.pl TAIR/TAIR10_GFF3_transposable_element.gff TAIR/family2superfamily.dat Alias
% perl TE_locate.pl 9 SAM/ TAIR/TAIR10_GFF3_transposable_element_HL.gff ref/at.fa TE 1000 5 1
  > temp.out 2 >&1
```

LAE-finder

for all BAM files:

```
% LAE-finder nothing SVs_02.dat data-folder/$name.realigned.sort.bam
% LAE-finder filterAndSVcall SVs_02.dat data-folder/$name.realigned.sort.bam
```

followed by:

```
% perl groupINVs.pl
% perl groupTLs.pl
```

Mach

Run through a custom file reformatting and submission script that is available online.

Supplementary Command Listing 1 Command lines used in sequencing pipeline.

to the BAM files that resulted from GATK local realignment. Given the nature of our data, i.e., high coverage and inbred lines, we can call the events rather aggressively, and we adjusted the default parameters accordingly (**Supplementary Command Listing 1**). In addition, we used the output from BreakDancer as prior knowledge to help Pindel to find more breakpoints.

2.2.4 Copy number variants through coverage

We called CNVs by calculating the coverage directly after the BWA reads mapping and Samtools pileup. To avoid bias by factors like GC content[9], the coverage was normalized. Coverage was estimated by summing the coverage in 1 kb windows and then normalizing them using the total coverage in the surrounding 3 Mb window. Regions with extremely high coverage, e.g., centromeric or rDNA repeats were not used in the normalization.

Windows were classified as have more copies than the reference genome if the estimated coverage was at least two-fold higher than background in at least 5 individuals.

Regions with zero coverage (i.e., no copies, where the reference genome has one) were also called from coverage, but we also required paired-end reads spanning the putative breakpoint, or sharing in at least 5 individuals. Software is available upon request.

2.2.5 Segregating deletions

Given the apparent shrinkage of the *A. thaliana* genome[10], it made sense to focus on finding segregating deletions. If the deleted sequence is present in the reference genome, then the polymorphism should be found using the method in the previous section, but if the deleted sequence is not in the reference genome, then we are looking for additional sequence present in some of the other *A. thaliana* lines, and perhaps also in the outgroup species *A. lyrata*.

With this in mind, we took reads that could not be mapped to the *A. thaliana* reference genome, and mapped them onto the *A. lyrata* genome. Once we found a contiguous region this way, we mapped the start and end coordinates back to the *A. thaliana* genome. If they mapped to a homologous region, we concluded that this is a genuine segregating deletion. Software is available on request.

2.2.6 TE polymorphism

Additional TEs copies not present in the reference genome were identified by looking for paired-end reads with one end in an existing TE, the other in a unique sequence. The software used is described elsewhere[11]. Copy number variation was analyzed at the level of superfamily (using the nomenclature in The Gypsy Database (GyDB) of Mobile Genetic Elements).

2.2.7 Large ancestral events

Very large structural rearrangements can be quite difficult to discover. However, we took advantage of the fact that many structural rearrangements separating the *A. thaliana* and *A. lyrata* reference genomes are segregating in the former species, and focused on finding these. An tool, LAE-finder, was developed for this purpose, in particular for finding inversions and translocations. The program, which is available for download, collects breakpoint information from the divergence data, and checks for presence-absence of these using paired-end reads.

2.3 Population-based SNP/indel genotyping

There is a general trade-off when calling SNPs and indels using short reads. If we call the variants aggressively, there will be many false positives; on the other hand, conservative parameters lead to unnecessary false negatives. The default setting of Samtools is quite conservative, at least for our data where the density of variants is high. In this project, we try to minimize the trade-off by separating the pipeline into two phases: discovery and genotyping. During the discovery phase, we discover events based on the individuals that have good coverage and read mapping quality, and during genotyping, we call the discovered variants with less conservative thresholds.

2.3.1 Recovering low coverage regions

We composed a list of relatively high-quality SNPs discovered by GATK and Samtools as described above, and then revisited the pileup files and genotype SNPs that were filtered out or not called due to local low coverage. Our criterion for calling a SNP was that it was supported by one read, and that the allele was present in at least 5 other individuals. For indels, we also genotyped alternative (i.e., non-reference) alleles from the consensus sequences that have at least only one read, but added the restriction that the number of reads supporting this event should be greater than the number of reads that did not. The results after this step comprise the original of SNPs/indels used in imputation.

2.3.2 Imputation of SNPs and short indels

In addition to allele sharing, we also leveraged LD across markers. We used MaCH[12] version 1.0 to impute SNP and short indel sites without sequence coverage. The result is the imputed version of released data (see **Section 6**, which is primarily intended for GWAS. Before imputation, all heterozygous calls were converted into missing calls, and only SNPs with exactly 2 alleles were kept. The latter step resulted in the removal of 281,942 SNPs.

As input to MaCH, we encoded each line as a homozygous diploid individual in the Merlin format. We then carried out imputation in windows of 20,000 markers, with an overlap of 2,000 markers between consecutive windows. 30 iterations of MCMC were used for each window, and the resulting probabilistic genotypes were converted into homozygous calls at each position.

2.3.3 Large SVs

Large ($>$ 200 bp) SVs were called using several different methods (section 2.2). Since methods have different resolution, we tolerated a 10% shift of breakpoints when combining events called from different individuals or methods. In the final dataset, an event was accepted if it is: (i) called by multiple pipelines, or; (ii) supported by at least 5% of the individuals (using a single method).

Given the inaccuracies inherent in calling large SVs, we did not try to impute them.

2.4 Error estimates and quality control

2.4.1 Re-sequencing the reference genome

The reference line, Col-0, was re-sequenced using the same methods as for our Swedish sample. Under the assumption that all variants called are errors, and restricting ourselves to regions with Q-value ≥ 30, we estimate false-positive rates of 0.21% and 1.9% for SNPs and indels, respectively. The estimated rates decrease dramatically as the quality of the mapping increases (**Supplementary Fig. 18**). Putatively false positive SNPs are aggregated near the centromeres (**Supplementary Fig. 19**), and tend to have low Q-values.

2.4.2 Array-based SNP genotyping

In contrast to reference re-sequencing, SNP-chip data provides estimates of the false negative rate (i.e., the rate at which we fail to discover SNPs), and the genotyping error rate (i.e., the rate at which me make the wrong call for the ones we did detect). The overlap between the previously published SNP data[13] and our data was 173 lines. As expected, both rates decreased with the alignment Q-value (**Supplementary Fig. 20**). Restricting ourselves to regions with Q-value ≥ 30, we found that we failed to discover 4.6% of SNPs, and made the wrong call for 3.5% of discovered SNPs. These estimates are conservative in that we ignore errors in the SNP data[14].

2.4.3 Sanger sequencing of PCR products

A subset of 45 of our lines overlapped an old data set of close to 1,500 manually curated multiple-alignments of Sanger-sequenced PCR-amplicons from 95 lines[15]. Since these regions were PCR-amplified, they do not represent a random sample for the genome, but they are useful nonetheless in that the quality of the data is extremely high. After eliminating complex regions with overlapping SNPs and indels, we estimated false positive and false negative rates for SNPs to be 1.8% and 3.7%, respectively, and the overall genotyping error rate (conditional on discovery) to be 0.64%. For indels, corresponding rates were 1.3%, 6.9%, and 2.7%.

2.4.4 Sanger sequencing of random clones

To avoid the biases mentioned on the previous section, we also sequenced randomly generated shotgun clones from a randomly chosen accession. Genomic DNA from the roots of Arabidopsis seedlings (DNeasy Plant Mini Kit, Qiagen) was sonicated (size range 300–1200 bp), gel-extracted (size range 700–800 bp), randomly cloned into the pJET1.2/blunt cloning vector (CloneJet PCR Cloning Kit, Fermentas) and transformed into competent *E. coli* cells. Plasmid DNA was isolated from overnight cultures. Inserts were amplified with T7 Promoter Sequencing primer and pJET1.2 Reverse Sequencing, and sequenced in both directions (Applied Biosystems 3130xL Genetic). The resulting chromatogram files were pre-processed as follows:

1. Remove vector sequence.
2. Apply sequence quality filter with a threshold of 0.0001 using Richard Mott's trimming algorithm (as implemented in CLC).
3. Eliminate reads shorter than 150 bp.
4. Align reads from complementary strands using SMALT (http://www.sanger.ac.uk/resources/software/smalt/).

All sequences were then aligned to the reference with BWA[4]. The mapping properties of the data types are consistent, the main difference being that Sanger reads that could not be aligned uniquely, could more often be anchored, due to their greater length (cf. **Fig. 1a**). Only sequences with a unique hit were used to calculate error rates. SNPs and indels were called using Samtools[5] in the same way as for the main data (except that there is no threshold for minimal coverage, i.e., a single Sanger sequence is always sufficient).

The above procedure generated ~250 kb of overlapping Sanger and Illumina data. After trimming a further 5 bp from the ends of each alignment to avoid artificial mismatches due to alignment problems, error rates were calculated for the Q30 data, assuming that the Sanger result were perfect. This analysis quickly revealed that two further steps of data filtering were required to obtain reasonable error rates. First, we eliminated putatively heterozygous polymorphisms. Although some tracts of genuine heterozygosity were observed in the data, the vast majority of heterozygous calls were shared between lines, which is extremely unlikely in a highly selfing species. Thus, most of these calls are dubious. Second, for indels, we eliminated mono- or di-nucleotide repeats, as these are very difficult to sequence (and are at least as likely to be called incorrectly in the Sanger data). After these filtering steps (released with the data, see **Section 6**), the error rates were comparable to the ones described above, except for the indel false negative rate (**Supplementary Table 1**). Note that we defined this rate as fraction of indels observed using the Sanger data that had not been observed using the Illumina data in *any* other line (if it had been observed in lines other than the right one, it would be a genotyping error). The indels we fail to call are thus most likely singletons.

3 *De novo* assembly

3.1 Assembly pipeline

We performed *de novo* assembly of the lines using two sets of tools: SOAPDenovo[16] v1.05 and clc_novo_assemble in the 4.0.1beta version of CLC Assembly Cell, the command-line backend to the CLCGenomics Workbench. For both tools, we used the raw fastq files from each line as input. For SOAPDenovo, we first tested a variety of different parameter options on 3 lines. Based on these results, we decided to carry out assembly of all lines at three different k-mer size setting: 27, 33 and 41. For other options, we used map_len=32, pair_num_cutoff=2, and avg_ins (average insert size for paired end reads) estimated from BreakDancer[7] in the configuration file. We enabled all optional procedures, including using reads to solve small repeats, remove low-frequency K-mers, remove low frequency edges and intra-scaffold gap closure. We used the defaults for all other options and precompiled parameters. See also **Supplementary Command Listing 2**.

In order to choose the best k-mer setting for each line, the resulting scaffolds were mapped back to the reference genome using BLAST, and the alignment results parsed to eliminate multiple hits. We used several criteria for this step: first, we only considered blast hits at above 85% similarity; second, if the same part of the contig aligned to multiple locations, we picked the highest scoring one (which in BLAST usually corresponds to the longest alignment); third, if two contigs mapped to the same location, we picked the longer alignment. We then evaluated the assemblies based on several criteria, the most important ones being: the proportion of the reference genome covered; the minimum length of scaffold to cover 50% of the genome (N50); and whether (based on manual inspection) the total length of the scaffolds differed too much from the population average. Finally, we chose the best among the three sets of scaffolds created using different k-mer settings as the assembly for that line.

For CLC assembly cell, we used the command line shown in **Supplementary Command Listing 2**, with insert sizes again estimated from BreakDancer. The results were comparable to those from SOAPDenovo, and we will not discuss them further.

SOAPdenovo

```
% SOAPdenovo63mer all -p 8 -K $kmersize -a 30 -R -d -D -F -s $configfile -o $outputfile
```

CLC Assembler

```
% clc_novo_assemble all -q -p fb ee $min_insert_size $max_insert_size -i $fastqfiles
  -o $outputfile --cpus 24
```

mafft

```
% mafft --thread 8 --localpair --maxiterate 20 --inputorder INPUTFILE > OUTPUTFILE
```

Supplementary Command Listing 2 Command lines used in *de novo* assembly.

3.2 Detection of structural variants

To complement the read-pair and split-read approaches to structural variant detection, we used a novel method based on our *de novo* assembly. Our method is similar in spirit to soapsv[17], based on alignment between scaffold and the reference genome, however, we focus on large events (>200 bp), which are poorly captured by standard read alignment algorithms.

3.2.1 Brief outline of algorithm

Our method detects structural variant breakpoints from irregularities in the alignment of scaffolds to the reference genome. For a region harboring a larger structural variant, we would expect to see the scaffolds covering the breakpoints to contain fragments from different regions on the reference genome, or different strands in the case of an inversion. Identifying such patterns, however, is complicated by the fact that, even in the absence of structural variation, a scaffold will sometimes map to multiple locations due to repetitiveness and small polymorphisms.

Our algorithm utilizes dynamic programming to search all possible alignments to identify scaffolds that contains genuine breakpoints, and further tries to discern the nature of each breakpoint, including where the flanking regions came from. The algorithm is still under development, and will be described in detail in a separate publication, however, the software implementation used here is available on request.

3.2.2 Quality of calls

When applied naively, our algorithm identified over 200,000 putative distinct breakpoints. However, when applied to our reference re-sequencing data, we obtained around 1/4 the average number of events for a line, indicating a false positive rate of at least 25%, and a rough analysis of the number of "missing" breakpoints in pair suggests a false negative rate of at least 20%. Thus, the results from this algorithm should be interpreted with care, and we include only small subset of the ones judged most reliable (most notably positive length variants) in the data release. All other types of events are released separately in the form of putative breakpoints (see **Section 6**). Although error rates are high, we note that our algorithm identifies several interesting examples of large structural variation that are readily validated by local patterns of LD, as illustrated in **Supplementary Fig. 17**.

3.3 Detection of novel sequence

Scaffolds and singleton contigs were first filtered by read coverage (those showing coverage less than 20% of average were eliminated) and then aligned to the reference using BLAST with the set of parameters described in previous section. All scaffolds, contigs, and sufficiently long (≥100bp) parts thereof that aligned poorly (or not at all: the criterion was 80% similarity and a minimum BLAST score of 200) to the reference genome were extracted.

The resulting sequences were aligned to genomic sequences from the Refseq-genomic database in order to identify their origin. Any sequence that was found to map to non-plant genomes was considered to be the result of contamination.

15 lines exhibited an aberrantly high amount of putative novel sequences. In some cases, this appeared to be due to contamination; in others it was probably due to a lower sequencing quality. After these lines were removed, the rest showed little sign of contamination. We assessed whether the remaining novel sequence was likely to be real *A. thaliana* sequence in two ways. First, we asked whether a given segment was part of a scaffold or contig that also contained sequence that clearly did match the reference genome (using the BLAST criteria given above). Second, we aligned novel sequences from different lines against each other, and looked for sharing between lines (as would be expected for a segregating indel polymorphism). **Supplementary Fig. 21** summarizes the results of these analyses across all lines. We found 1.5–2.5 Mb of novel sequence per line, almost of which was either anchored in the reference genome or shared among at least 5 lines. As for the origin of the novel sequence, about 250 kb per line was clearly of plant origin (total combined BLAST score for the best linear alignment >400). In general, the greatest similarity was to *A. lyrata*.

Overall, our attempts to identify large indels, identified many more positive (w.r.t. the reference genome) than negative length variants. If this difference were real, then it would imply that the reference genome comes from a line with unusually small genome. The alternative explanation is that it is simply due to bias: from a statistical point of view, it is easier to detect presence (i.e., novel sequence) than absence (missing reads, which could be due to chance). This explanation is supported by the observation that the spatial distribution of novel sequence along the chromosomes closely mirrors that of missing coverage (**Fig. 1f** and **Supplementary Fig. 2**).

We examined some polymorphisms in detail by aligning the scaffolds spanning the region with the homologous regions from the *A. thaliana* and *A. lyrata* reference genomes using mafft[18] (**Supplementary Command Listing 2**). An example can be seen in **Fig. 1e** in the main text.

We tested the hypothesis that NB-LRR and F-box proteins, two gene families with high birth and death rates, contribute disproportionally to the novel sequences. First, we used Hmmer (v3.0)[19] to create hmm profiles using multiple alignments of the NBS and F-box domains in *Arabidopsis thaliana* with assistance of previously published sequence of the domains[10]. Then, we searched in the novel sequences for the motifs (hmmsearch). We detected similar motif/length of sequence ratio using the reference genome as using either gene family (no significant enrichment). One caveat is that sequences containing the motifs could potentially have been removed due to similarity to the reference genome in the filtering steps for identifying novel sequence.

4 Variation in nuclear DNA content

4.1 Flow cytometry

Flow cytometry was carried out on 129 lines (128 Swedish plus Col-0) split into two sets with 11 overlapping lines. Each set was further divided into three blocks of replicates that were measured on a different days with 1–2 biological replicates per line within each block. In addition, a set of 36 world-wide lines (plus a single line overlapping the set of 129) divided into three blocks of replicates (no replication within blocks) were measured on different days.

Seeds were stratified directly on soil for 5 days. Plants were grown under long day conditions (16 h light at 21°C, and 8 h dark at 16°C), watered twice a week, and rotated within trays daily. At 2 weeks the first two true leaves of each plant were finely chopped with a razor blade together with an approximately 0.125cm^2 piece of leaf of the internal standard, *Solanum lycopersicum* cv. Stupicke (2C = 1.96 pg DNA[20]), in 250 μl of extraction buffer (kit PARTEC CyStain PI Absolute P no. 05-5022). 1 ml of staining solution (with 6 μl of propidium iodine (PI) and 3 μl of RNase [3.33 mg/ml]) was added, and the resulting suspension was passed through a 30-micron filter (Partec CellTrics no. 04-0042-2316). Samples were stored for 2–4 h at 4°C in the dark prior to DNA content evaluation.

Genome size was measured with a LSRFortessa special order research product equipped with a 561 nm yellow-green laser (110 mW) and a 488 nm blue laser (100 mW), for PI (610/20 nm) and side scattering (SSC; 488/10 nm) detection, respectively. Events representing debris were excluded by selecting only the major cluster when plotting the PI-area versus SSC-area for 10,000 events. Data was analyzed with the flowClust R package[21]. The mean position of the 2C peak for each sample was normalized to the 2C peak of the internal standard and converted into base pairs[22].

Simple linear regression models were fitted for each set in order to account for the block effect and obtain a single flow cytometry estimate for each line. The mean and standard deviation of these estimates are available online (see **Section 6**). The results were generally highly reproducible, with the standard deviation being on the order of a percent of the mean. The distribution of estimates is shown in **Supplementary Fig. 3**.

4.2 Repeat-number estimation through coverage

Sequence coverage for each individual was calculated by summing normalized read coverage in 1 kb windows (as described in **Section 2.2.4**) across the entire genome. Note that, for this analysis, we did not remove reads that were supposedly due to library duplication, as this seemed to removed actual repeats (leading to poorer agreement with flow cytometry estimates, see **Supplementary Fig. 23**). The contribution to genome size by 45S rDNA, 5S rDNA and centromeric repeat elements was estimated by summing read coverage across the appropriate regions of the reference sequence. For 45S rDNA, the two ∼10 Kb 45S rDNA loci, in the beginning of chromosome 2 and at 14.2 Mb of chromosome 3, respectively, were considered. For 5S rDNA, the locations were determined by BLASTing the transcribed region consensus sequences identified for the major and minor loci on chromosomes 4 and 5, as well as loci 1, 2 and 3 on chromosome 3[23]. Similarly, centromeric regions were located via BLAST with two centromeric variants, clones 22_At178 (GenBank: EU359499.1[24]) and AS1 (GenBank: X04320.1[25]). For TEs, the total count of *novel* TEs was used (see **Section 2.2.6**). All estimates are available online (see **Section 6**). The correlation between the flow cytometry and repeat copy number estimates was analyzed using standard regression methods as described in the text. The results are presented in **Table 1**, **Supplementary Table 4**, **Fig. 2a**, and **Supplementary Fig. 22**.

4.3 Estimating rDNA copy number through qPCR

45S rDNA copy number can also be estimated through quantitative PCR (qPCR) of either the 18S or 25S subunit[26]. We carried out qPCR in technical triplicate and biological replicate for each of five *A. thaliana* lines (ids: 5856, 6099, 6136, 6244, 8386) in a 25 μl total reaction volume using 2X SensiMix SYBR & Fluorescein Kit (Bioline No. QT615-05).

An iQ5 light cycler (Bio-Rad) was employed with the following thermal profile: 95°C for 600 seconds; 40 cycles at 95°C for 10 seconds, 60°C for 30 seconds and 72°C for 30 seconds; and a final cycle at 72°C for 60 seconds. One standard curve based on serial 10-fold dilutions was made for each sample. No primer dimmers were detected in the melting curve.

45S ribosomal DNA (rDNA) abundance was estimated by comparing 18S rDNA to two single copy genes (At3g18780, At4g38740; see relevant file in **Section 6** for a list of primers) according to:

$$\text{rDNA abundance} = 2^{\text{Ct(single copy gene)}-\text{Ct(18S rDNA)}}$$

where $\text{Ct}(x)$ stands for the threshold cycle for x. Estimates for all lines were then normalized to the line with the lowest estimated copy number of 18S rDNA. The qPCR results showed excellent agreement with the coverage-based estimates (**Supplementary Fig. 23c–d**).

4.4 Genome-wide association mapping

Genome-wide association mapping was carried out using the imputed SNP data (**Section 6**) using an approximation of the kinship model[27]. Minor alleles below 5% MAF were filtered out prior to the analysis, leaving around 1.8M SNPs. Both flow cytometry and 45S copy numbers estimates were used as phenotype, and we tried different subsamples of the lines in order to evaluate the robustness of the association (**Supplementary Fig. 5**). Analysis of the northern and southern lines separately demonstrated that the association is due to variation among the northern lines: the association is not present in the southern sample (**Supplementary Fig. 5a–c**). Association mapping using 45S copy number directly as a phenotype revealed that, while the peak is still present, it is much less distinct (**Supplementary Fig. 5d**). Furthermore, when we increase the sample size to the full sample for which we have sequence data, the peak vanishes (**Supplementary Fig. 5e**), which is troubling. However, the 45S-rDNA coverage data vary greatly between lines and replicates, and may be strongly affected by both alignment and sequencing artifacts. Thus, counter-intuitively, the flow cytometry data may actually be a better estimate of the true number of 45S-rDNA repeat copies.

4.5 FISH

An obvious explanation for putatively *trans* GWAS peaks related to 45S copy number is that they are linked to novel 45S rDNA clusters (i.e., they are, in fact, *cis*). Thus, although the 100 kb region on chromosome 1 that contains the most significant associations does not show any evidence for large structural variants, we decided to look genome wide. First, TE-Locate was used with the standard settings to map novel 45S rDNA repeat insertions, and none were found. Second, we used Fluorescence In Situ Hybridization (FISH) to look for clusters directly.

Actively growing, young roots were pretreated with ice-cold water for 12 hrs, fixed in ethanol:acetic acid (3:1) at 4°C for 24 hrs. Selected root tips were rinsed in distilled water and citrate buffer (10 mM sodium citrate, pH 4.8), and digested by 0.3% cellulase, cytohelicase and pectolyase (all Sigma-Aldrich) in citrate buffer at 37°C for 90 min. Individual root tips were dissected in ca. 10 μl of acetic acid on a microscopic slide. The cell material was covered with a cover slip, evenly spread by tapping, and the slide gently heated over a flame. Then the slide was frozen in liquid nitrogen, cover slip flicked off, fixed in ethanol:acetic acid (3:1) and air-dried. *A. thaliana* BAC clone T15P10 (AF167571) containing 45S rRNA genes was used to identify 45S rDNA loci. BAC clones F5A8, F1N21, F12A21, T23K23, and F12B7 were used to paint chromosome 1, and localize the candidate region. All DNA probes were labeled either with biotin-dUTP, digoxigenin-dUTP, or Cy3-dUTP by nick translation, pooled, ethanol precipitated and pipeted on ready-to-use slides. The slides were heated to 80°C for 2 min and incubated at 37°C overnight. Hybridized DNA probes were visualised either as the direct fluorescence of Cy3-dUTP (yellow) or through fluorescently labeled antibodies against biotin-dUTP (red) and digoxigenin-dUTP (green). DNA labeling and fluorescence signal detection was carried out using a previously published step-by-step protocol[28]. Chromosomes were counterstained with 4,6-diamidino-2-phenylindole (2 μg/ml) in Vectashield (Vector Laboratories). Fluorescence signals were analyzed and photographed using a Zeiss Axioimager epifluorescence microscope and a CoolCube camera (MetaSystems), and pseudocolored/inverted using Adobe Photoshop CS2 software (Adobe Systems).

Results for three accessions with large estimated 45S rDNA copy number are shown in **Supplementary Fig. 4**. The known clusters on chromosomes 2 and 4 are clearly visible, and there is no evidence for any other clusters.

5 Population genetic analyses

5.1 Selection on indels

All polymorphisms were annotated with respect to function, using the reference genome. As expected under the assumption that most mutations disrupt function, structural variants are relatively more common outside genes, and relatively rare in exons.

In order to test whether selection is driving deletions to fixation, as has been suggested[10], we used the global alignment between *A. thaliana* and *A. lyrata*[10] to determine the ancestral state of SNPs and indels. The derived allele frequency distribution can be a powerful tool when looking for signal of selection, essentially because it makes it possible to identify the high-frequency derived allele that are expected to be rare unless selectively favored. Very conservative criteria were used in the polarization, the reason being that the conclusions may be severely biased if alleles are misclassified with respect to ancestral status. Because derived alleles are almost always rare, misclassification of them as ancestral will cause a large inflation of supposedly high-frequency derived alleles[29]. With this in mind, a polarization was accepted only if the identify of surrounding region passed stringent criteria. For SNPs, we required that the identity of the surrounding 30 bp window should be at least 90%. Short indels ($\leq$ 200 bp) were classified as deletions if *A. lyrata* had what appeared to be the longer allele, and the additional sequence was identical in all carriers (including *A. lyrata*). Indels were classified as insertions if *A. lyrata* had the shorter allele and the additional, putatively inserted sequence was identical in all carriers. We did not try to polarize other structural variants.

Given that the overall divergence between the two reference genomes is greater than 10%, the above criteria are very stringent. The estimated allele frequency distribution for SNPs, deletions, and insertions is shown in **Supplementary Figure 6**. Contrary to previously observations[10], there is no evidence for an excess of high-frequency deletions. It is not clear what to conclude from this. While our analysis is based on many more events, it is worrisome that we polarize only 18% of all (observed) events. Experimenting with less conservative criteria demonstrated that the allele frequency distribution is extremely sensitive to the procedure used, however, in no cases did we get results consistent with strong selection favoring deletions (not shown).

For the SNP-based selective sweep analysis in **Section 5.4**, we employed no filtering when estimating ancestral state. Both polarizations, as well as the function annotation, are part of the released data (**Section 6**).

5.2 Long-range LD

5.2.1 LD in structured populations

Let n denote the number of individuals and m the number of SNPs. Given a genotype matrix in which each of the SNP vectors, S_i, $i = 1, \ldots, m$, is coded numerically, i.e., 0, 1, or 2 for each of the n diploid individuals, we can obtain normalized SNPs $X_i = (S_i - \overline{S_i})/\sqrt{\mathrm{Var}}(S_i)$ with mean 0 and variance 1. Under the assumption that these SNPs are independently sampled from a distribution with covariance matrix $\mathbf{V}$, we can obtain an unbiased estimate of this covariance matrix as

$$\mathrm{Cov}_{\mathrm{est}}(X_i) = \hat{\mathbf{V}} = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})(X_i - \overline{X})^\top \ . \quad (1)$$

If we assume that the average allele value for each accession is equal (i.e. alleles are labeled so that all accessions have roughly equal numbers of 0 and 1 alleles), then this covariance matrix estimate is exactly Fisher's correlation kinship matrix[30]. Like Mangin et al., we then obtain pseudo-SNPs, $T_i = \hat{\mathbf{V}}^{-\frac{1}{2}}X_i$, with values that are expected to be independent across individuals. Now we can proceed to obtain r^2 values that have been corrected for genetic correlations between individuals, or, in other words, population structure. The correlation becomes strikingly simple when written out as

$$\mathrm{Cor}_{\mathrm{est}}(T_i, T_j) = \frac{\mathrm{Cov}_{\mathrm{est}}(T_i, T_j)}{\sqrt{\mathrm{Var}_{\mathrm{est}}(T_i)\mathrm{Var}_{\mathrm{est}}(T_j)}} \ , \quad (2)$$

where

$$\mathrm{Var}_{\mathrm{est}}(T_i) = \frac{1}{m}(T_i - \frac{1}{m}\sum_k T_{ik})^\top (T_i - \frac{1}{m}\sum_k T_{ik}) \ , \quad (3)$$

and

$$\mathrm{Cov}_{\mathrm{est}}(T_i) = \frac{1}{m}(T_i - \frac{1}{m}\sum_k T_{ik})^\top (T_j - \frac{1}{m}\sum_k T_{jk}) \ . \quad (4)$$

If the genotype data does not contain any missing values then we can speed up the calculation for the adjusted correlation by calculating the pseudo-SNPs and normalizing them beforehand. Consider the normalized pseudo-SNP

$$W_i = \frac{T_i - \frac{1}{m}\sum_k T_{ik}}{\mathrm{Var}_{\mathrm{est}}(T_i)} \ . \quad (5)$$

We can now obtain the correlation estimate by simple vector multiplication

$$\mathrm{Cor}_{\mathrm{est}}(T_i, T_j) = W_i^\top W_j \ , \quad (6)$$

and hence $r^2(T_i, T_j) = \mathrm{Cor}_{\mathrm{est}}(T_i, T_j)^2$. For data with no missing values (or with missing data imputed), the time complexity for estimating r^2 for all SNP pairs is $\mathcal{O}(mn^3 + m^2n)$. An obvious extension to this would be to use some of the approximations proposed by Lippert et al.[31], to obtain an approximate r^2 in sub-cubic n time.

The transformed LD estimates are generally lower than the original ones, since the inflation caused by population structure is removed (**Supplementary Fig. 24**), however, large numbers of long-range LD pairs remain (**Fig. 4a**).

5.2.2 Potential causes of long-range LD

Major population subdivision Since the divergence between north and south in our sample is substantial (see **Section 5.3**), we evaluated its effect on LD separately, by applying the relatedness correction separately to the northern and southern subsamples. The LD pattern in the north contain many more high r^2 pairs (**Supplementary Fig. 25a**), a result of smaller sample size as well as higher relatedness, whereas the LD pattern in the South looks very similar to that of the full sample (**Supplementary Fig. 25b**). Thus we conclude that the vast excess of long-range LD is not simply due to the north-south division.

Imputation We used imputed data for the LD calculation, primarily to speed up computation, but also to make sample size even across all pairs. To ensure that this did not cause the pattern observed, we re-calculated r^2 for all high-LD pairs in the unimputed data. The results show that imputation is clearly not a source of long-range LD (**Supplementary Fig. 25c–d**).

Other artifacts Long-range LD might also various kinds of mapping and genotyping artifacts, i.e., the SNP loci do not segregate normally. To try to eliminate these problems, we applied several stringent filtering steps. To begin with, all analysis was based on the high-quality Q30 SNPs. To be even more stringent, we aligned 75 bp surrounding each SNP to the reference genome using BLAST, using a less stringent criterion (90% similarity) than was used in the original read mapping. Any SNP that could be aligned to more than one region on the reference genome was filtered out. This should remove simple mapping artifacts.

After transforming and filtering, we were still left with over 70,000 SNP pairs exhibiting strong long-range LD, especially between centromeric regions (**Fig. 4a**). To test whether they segregated normally, 4 centromeric and 2 non-centromeric set of SNPs were genotyped in informative F2 crosses (**Supplementary Table 2**). Two informative crosses (6035 × 9433 and 6136 × 6064) were carried out, with 10 F2 seedlings genotyped using PCR and dideoxy-sequencing. Of the four centromeric pairs, only one yielded reliable PCR fragments for both SNPs (**Supplementary Table 2**). In summary, 2 out of 2 between-chromosome, non-centromeric comparisons segregated normally, and 1 out of 1 between-chromosome, centromeric comparison did not.

Based on these results, we conservatively decided to remove all centromeric SNPs, leaving only 2509 pairs. To further test whether some of these might also be closely linked, we aligned short sequences (both 75 and 150 bp was tried) surrounding each SNP to the scaffolds generated by *de novo* assembly. If sequences flanking both SNP in a pair mapped near each other one the same scaffold, we considered them linked in that line. Out of the 2509 pairs tested, we found 17 that could be due to this kind of structural variation (**Fig. 4b**).

Selection For the remaining long-range LD pairs, we first tested whether the corresponding pairs of loci were overrepresented in published protein interaction data[32]. No significant overrepresentation was found. Next, we looked for overrepresentation of individual loci among those identified as having signals of local adaptation[33]. We asked whether SNPs involved in long-range LD were close to the peak SNPs (those with p < 0.001) for 13 climatic traits, and calculated the p-value using a permutation scheme that maintain the LD structure in the data[15]. We detected significant (Bonferroni-corrected p <0.05) enrichment in two of the traits, relative humidity and minimum precipitation (**Supplementary Table 3**). The same method was used to test for overlap with SNPs implicated in selective sweeps (**Section 5.4**). In northern Sweden, there is a significant overrepresentation of long-range LD SNPs and SNPs within 1 kb of a SNP with a Sweepfinder CLR above 50 (p = 0.0336).

5.3 Population structure

5.3.1 Analysis

Global population structure in *A. thaliana* has been described several times[13,15,34]. We compared our Swedish sample (which we divided into a northern and a southern population based on latitude 60° N, see **Supplementary Fig. 1**) with other samples for which whole-genome sequencing data are available, namely the 8 smaller populations (10 individuals in each) sequenced by Cao et al.[3]. Three different statistics were calculated: PCA, t-SNE, and F_{ST}. The results confirmed the distinctiveness of the northern Swedish sample (**Supplementary Table 5**)[35]. Clustering based on the 250k SNP data[13] identified two likely contaminants among the northern lines: 6180 and 1435. These two lines clearly cluster with southern (or even other European) lines, and were excluded from further analysis. These conclusions are supported by standard hierarchical clustering analysis as well (**Supplementary Fig. 8**).

To further characterize the pattern of variation in Sweden, we plotted the sequence divergence between all pairs of lines in our samples as a function of the distance separating the original sampling locations (**Supplementary Fig. 9**), and tested for isolation by distance using the Mantel test (function mantel of the R-package vegan). We found a significant correlation within both the northern and the southern sample, although the correlation within northern Sweden was much stronger (Spearman's r=0.6109, p<0.001 and r=0.4525, p<0.001, respectively). The 95% bootstrap confidence interval for the difference in Spearman's r between north and south was (0.115, 0.173). The north also has lower levels of polymorphism and higher Tajima's D[36] (**Supplementary Fig. 15**), as well as more extensive LD (**Supplementary Fig. 7**). Taken together, these results are consistent with the observation that the north seem to have a much more patchy population structure, with quite small local populations.

5.3.2 Methods

F_{ST} We used only bi-allelic SNPs for which we had complete information for all 260 (180+80) lines. Each SNP was coded as 0 for the major allele, and 1 for the minor allele. The standard method was used[37]. In addition to genome-wide averages, we also calculated F_{ST} between northern and southern Sweden in non-overlapping windows of 100 kb.

θ, π, and D We calculated three standard statistics describing aspects of nucleotide diversity, separately for north and south, and in 100 kb windows across the genome: Watterson's θ; nucleotide diversity, π; and Tajima's D. All three statistics are intended for complete sequence data rather than SNPs, and we therefore used BamTable[38], which essentially tries to integrate over the uncertainty in polymorphism detection, generating probabilistic calls. We ran BamTable on sorted BAM files to call SNPs, using the standard options. The most common base was called at each site and each line. Bases that were supported by less than 5 reads were excluded. If two different bases were supported with more than 4 reads for a certain site, then this site was excluded as potentially due to incorrect alignment or genuine heterozygosity site.

In order to accommodate missing data, we calculated the summary statistics separately for all sample sizes, then carried out weighted averaging similar to what has previously been suggested for Watterson's θ[39]. For the other two statistics, we use

$$\pi = \sum_{i=2}^{n} \frac{L_i}{L_T} \pi_i \tag{7}$$

and

$$D = \frac{\sum_{i=2}^{n} L_i a_i D_i}{\sum_{i=2}^{n} L_i a_i} , \tag{8}$$

where π_i and D_i, $i = \{2, ..., 180\}$, are the values of the sample-size specific statistics, L_i is the sequence length of sites with sample size i, and a_i is the ith harmonic number. These estimates turned out to be superior in terms of bias and root mean squared error when compared to alternative formulas in neutral simulations with random missing data added in a way that reflects the observed data (results not shown).

5.4 Selective sweeps and local adaptation

5.4.1 Genome scans

CLR Sweepfinder[40] is intended for polarized SNPs, but can handle missing data. SNPs were polarized as described in **Section 5.1**, but without the conservative filtering. The 46% of SNPs that could not be polarized were nonetheless used in the analysis. The output of Sweepfinder is a CLR (composite likelihood ratio) statistic for a grid of positions with distance of 1,000 bp between successive positions. To arrive at a significance threshold, we use the coalescent simulator msms[41], which also allows selection.

We used standard neutral simulations to determine the critical CLR values (**Supplementary Command Listing 3**). The population mutation rate was set to 0.005/bp, which corresponds to the average observed diversity. An average recombination rate of 4.6 cM/Mb[2] was assumed, and a selfing rate of 97%[34]. There is no evidence that recombination rates in the sweep regions deviate strongly from this value, and Sweepfinder has been shown to be relatively robust to deviations from the true recombination rate[40]. Recombination and diversity were scaled to correspond to a sequence length of 1Mb. The sample size was set to the mean northern and southern sample size (averaged over all SNPs): 43 and 111, respectively. The total number of simulations was 12,000, corresponding to the simulation of about 100 whole genomes. The cut-off was then calculated for a family-wise error rate of 5% per genome, so that a false positive signal only occurs in one out of 20 whole genome analyses on average.

Simulations with selection at a single locus showed that a single selective sweep leads to multiple significant peaks in the CLR in about 80% of cases. In those cases, significant regions span an average of 170 kb, with a 99% quantile of 430 kb. The 95% confidence interval of the true position of the selected site is 160 kb centered around the largest peak. Thus, it is not unlikely that multiple significant peaks created by a single sweep are relatively far apart from each other, although the largest peak is almost always nearer to the true selected locus than smaller peaks. Therefore, we treat multiple peaks within a region of 430 kb as single events in our analyses, and designate the peak with the largest CLR as the center of the sweep.

msms (neutral model)

```
% msms -ms 111 12000 -t 5000 -r 994
```

msms (with selection)

```
% msms -ms 50 1000 -t 5000 -r 800 -SAA 1000 -SaA 500 -N 100000 -SF 0 -Sp .5
```

XP-CLR (repeat for chromosomes 2–5)

```
% XPCLR -xpclr SouthChr1.geno NorthChr1.geno Chr1.map Chr1.xpclr -w1 0.005 200 1000 1 -p1 0.95
% XPCLR -xpclr NorthChr1.geno SouthChr1.geno Chr1.map Chr1.xpclr -w1 0.005 200 1000 1 -p1 0.95
```

XP-EHH (repeat for chromosomes 2–5)

```
% xpehh -m Chr1.map -h SouthChr1.hap NorthChr1.hap > Chr1.xpehh
```

Supplementary Command Listing 3 Command lines used in sweep finding.

In total there were 22 significant sweep locations in the northern and 1 in the southern Swedish sample. The single significant southern Swedish signal overlaps with the strongest signal in the north. For each location, 160 kb large regions centered on the largest CLR peak were selected, and subjected to further analysis (see below).

The sweep haplotype corresponding to the shared sweep signal on chromosome 1 turned out to be associated with an intrachromosomal transposition of 278 kb to a new position 486 kb away. During the selective sweep, this configuration likely prevented any recombination event between the ancestral and rearranged haplotypes in 764 kb region. To ensure that the sweep signal was not simply due to repression of recombination, we reran Sweepfinder on a data set where the entire 764 kb region was replaced by a single base pair. The sweep signal from this analysis was still the strongest in both northern Sweden and southern Sweden, and Sweepfinders CLR value for northern Sweden was still 165 fold larger than the average CLR value in the rest of the genome.

XP-CLR and XP-EHH XP-CLR[42] was calculated between the northern and the southern populations, looking for sweeps in the north with the south as reference, as well as vice versa. XP-EHH[43] (downloaded from http://hgdp.uchicago.edu/Software/) just returns a single value for the comparison of the two populations. Since it cannot handle missing data, all SNPs with missing individuals were removed.

5.4.2 Environmental correlations and GO terms

A file available online (see **Section 6**) summarizes the 22 significant sweep regions. For each region, GO terms were collected, and a test for GO category enrichment was carried out using func[44]. Terms that reached reached a significance threshold of $p < 0.01$ and had at least three genes in the category are reported. Only biological process (GO:0008150) sub-categories were used.

The data from Hancock et al.[33] were used to look for enrichment of significant SNP-environment correlations in the swept regions. For each environmental variable (aridity, consecutive cold days, consecutive frost-free days, day-length, length of growing season, maximum precipitation, maximum temperature, minimum precipitation, minimum temperature, PAR, precipitation seasonality, relative humidity and temperature seasonality), the 1% tail of largest correlation coefficients was selected and tested for enrichment in each sweep region. The p-value was calculated by deriving an empirical null distribution of the proportion of tail signals (number of SNPs that are in the tail of the correlation coefficient distribution, divided by all the SNPs in that interval) for a randomly placed interval. The p-value is the probability of having an as high or higher proportion of tail signals in a randomly placed interval compared to the actual sweep region.

5.4.3 Dating the sweep

The sweep was dated utilizing the average amount of polymorphism separating two swept haplotypes. To do this we utilized sequence data from a segment within the transposition (20.35-20.45Mb). To account for the fact that we only use SNPs without missing data for the age estimation, we reduce the sequence length by the ratio of the number of SNPs without missing data (1158) and the total number (2221), leading to an effective sequence length of 100000 × 1158/2221 = 52139 bp. The SNPs for which there were no missing individuals, and which were are also monomorphic in the six lines with the ancestral (unswept) configuration were then used to calculate the average number of differences for a randomly selected pair of northern and southern lines, respectively. These numbers were 12.7 and 31.6.

Assuming an approximately star-like tree, we estimate the age of the sweep by calculating the average divergence time of two sweep haplotypes. We do this by dividing the average number of differences with a factor of $2 \times 52139 \times 7 \times 10^{-9}$, where 7×10^{-9} is the estimated mutation rate per bp and generation[45]. This resulted in an average divergence time of 17,390 years for northern Sweden, and 43,282 years for southern Sweden.

6 Data release

The raw data has been deposit to NCBI trace SRA with accession number SRP012869. Processed data are available through the project download site.

6.1 Genotype files

In total we identified around 4.5M SNPs, 576k short indels, 23k transposable elements, 7.7k CNVs, as well as 3.8k other structural variants (larger than 200 bp). The following files are available:

- SNPs (Original and imputed)
- SNP mapping information
- SNP annotations
- Small indels (Original and imputed)

- Small indel annotations
- Large structural variants (several files from multiple pipelines)

6.2 Other files

The following files are also available from the website:

- list of lines used in this study (an interactive version is available).
- alignment and assembly statistics for all sequenced lines.
- flow cytometry and repeat copy number estimates for the 180 lines.
- flow cytometry estimates for 36 worldwide lines.
- multiple alignments for the candidate genes from **Fig. 2**.
- summaries of putative sweep regions.
- all PCR primers used in this study.
- predicted genotype for 1306 lines with respect to the transposition on chromosome 1, and the large inversion on chromosome 4 (**Supplementary Fig. 17**).
- all genes in the swept transposition on chromosome 1.

References

1. Marth, G. T., Czabarka, E., Murvai, J. & Sherry, S. T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166,** 351–372 (2004).
2. Kim, S. *et al.* Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genet.* **39,** 1151–1155 (2007).
3. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genet.* **43,** 956–963 (2011).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
5. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
6. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
7. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6,** 677–681 (2009).
8. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).
9. Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330,** 641–646 (2010).
10. Hu, T. T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genet.* **43,** 476–481 (2011).
11. Platzer, A., Nizhynska, V. & Long, Q. TE-Locate: A tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. *Biology* **1,** 395–410 (2012).
12. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epi.* **34,** 816–834 (2010).
13. Horton, M. W. *et al.* Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genet.* **44,** 212–216 (2012).
14. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465,** 627–631 (2010).
15. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3,** e196 (2005).
16. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–72 (2010).
17. Li, Y. *et al.* Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nature Biotech.* **29,** 723–730 (2011).
18. Katoh, K., Kuma, K.-i., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33,** 511–518 (2005).
19. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–37 (2011).
20. Dolezel, J., Sgorbati, S. & Lucretti, S. Comparison of 3 DNA fluorochromes for flow cytometric estimation of Nuclear-DNA Content in Plants. *Physiol. Plant.* **85,** 625–631 (1992).
21. Lo, K., Hahne, F., Brinkman, R. R. & Gottardo, R. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics* **10,** 145 (2009).
22. Dolezel, J., Bartos, J., Voglmayr, H. & Greilhuber, J. Nuclear DNA content and genome size of trout and human. *Cytometry A* **51,** 127–128 (2003).
23. Cloix, C. *et al.* Analysis of the 5S RNA pool in *Arabidopsis thaliana*: RNAs are heterogeneous and only two of the genomic 5S loci produce mature 5S RNA. *Genome Res.* **12,** 132–144 (2002).
24. Martinez-Zapater, J. M., Estelle, M. A. & Somerville, C. R. A highly repeated DNA sequence in *Arabidopsis thaliana*. *Mol. Gen. Genet.* **204,** 417–423 (1986).
25. Zhang, W., Lee, H. R., Koo, D. H. & Jiang, J. Epigenetic modification of centromeric chromatin: hypomethylation of DNA sequences in the *CENH3*-associated chromatin in *Arabidopsis thaliana* and maize. *Plant Cell* **20,** 25–34 (2008).
26. Davison, J., Tyagi, A. & Comai, L. Large-scale polymorphism of heterochromatic repeats in the DNA of *Arabidopsis thaliana*. *BMC Plant Biol.* **7,** 44 (2007).
27. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genet.* **42,** 348–354 (2010).
28. Lysak, M. A. & Mandáková, T. in *Plant Meiosis: Methods and Protocols* (eds Pawlowski, W. P., Grelon, M. & Armstrong, S.) (Humana Press, 2013).
29. Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Biol. Evol.* **24,** 1792–1800 (2007).

30. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinburg* **52,** 399–433 (1918).

31. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8,** 833–835 (2011).

32. Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an *Arabidopsis* interactome map. *Science* **333,** 601–607 (2011).

33. Hancock, A. M. *et al.* Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* **334,** 83–86 (2011).

34. Platt, A. *et al.* The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.* **6,** e1000843 (2010).

35. Platzer, A. Visualization of SNPs with t-SNE. *PLoS ONE* **8,** e56883 (2013).

36. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).

37. Hudson, R. R., Boos, D. D. & Kaplan, N. L. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* **9,** 138–151 (1992).

38. Siebauer, M., Fischer, A., Kelso, J. & Prufer, K. *BAMTable: A rapid filter for BAM files.* In preparation. 2012.

39. Hellmann, I. *et al.* Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res.* **18,** 1020–1029 (2008).

40. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15,** 1566–1575 (2005).

41. Ewing, G. & Hermisson, J. MSMS: A coalescent simulation program including recombination, demographic structure, and selection at a Single Locus. *Bioinformatics* **26,** 2064–2065 (2010).

42. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20,** 393–402 (2010).

43. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449,** 913–918 (2007).

44. Prufer, K. *et al.* FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8,** 41 (2007).

45. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327,** 92–94 (2010).

6.7 Results – addendum

6.7.1 LAE-finder

The Large-Ancestral-Events-finder is a tool for confirming known events. On its own, 'confirming known events' sounds a little meaningless; intended is gaining certainty about events in an individual which are either called in a rather unsafe way or which are known to occur in a population. The specific goals of this tool are to get all possible information out of read-pairs for and against specific events. The latter comprises its difference from most other tools, where only presence is called. The tool name contains 'Ancestral' because the comparisons between the reference sequence and ancestral references are the likeliest sources for safe large variations (see also section 6.7.4 SVs). Initially the main motivation was the existence of a lot of tools and methods to call SVs but all with incredibly high error rate.

In the paper, the tool is mentioned in the supplement at '2.2.7 Large ancestral events'. Only in one sentence is it mentioned in a misleading, where it is not clear that the events must come from elsewhere, but need not to be safe in any manner. The function of the LAE-finder is to make things safer. The event finding was done here by a colleague in a partly manual way, which was never fully automatized. The lack of automatization and some hints of biases in the event finding explain why this did not go into an own methods paper.

Figures 12 to 16 describe what the tool is designed for.

The read-pair information, resp. the coordinates, are used in three different ways. Firstly, to call the presence of the event (see Figure 12 and 14), which is what most other tools also do. Secondly, to call the absence of the event with a concept called 'continuous coverage': if trimmed reads or read-pairs span breakpoint coordinates it is a sign of absence (see Figures 13 and 16). The read-pairs must not have a breakpoint between them to count as continuous coverage. Thirdly, one mate of the read-pairs can be used at the first breakpoint in translocations ('BP' in Figure 15) to get a precise coordinate and/or to know if different possibilities exist for it.

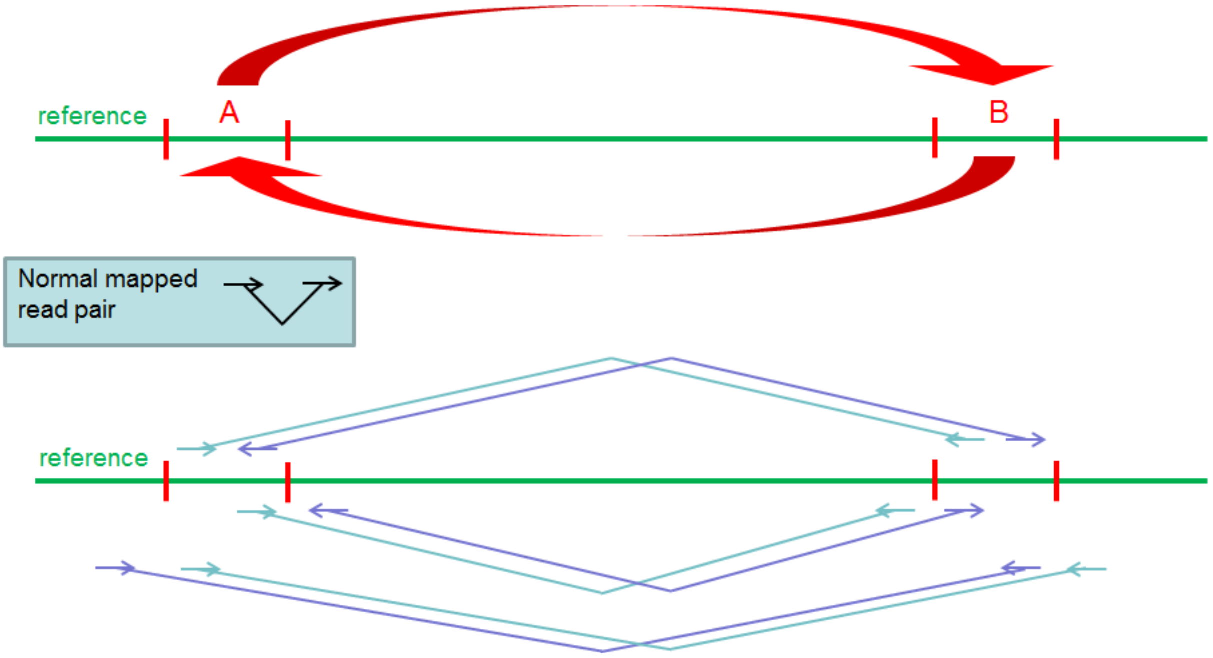


Figure 12. LAE-finder – Inversion – Event support.

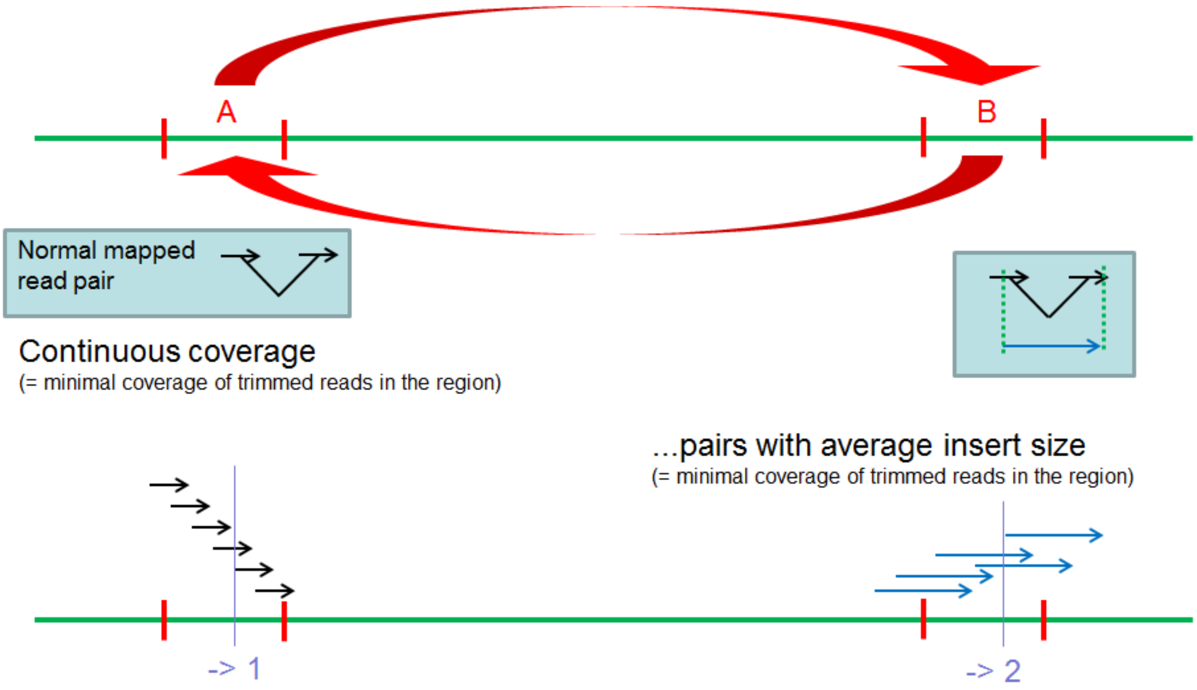

A
B
Normal mapped read pair
Continuous coverage
(= minimal coverage of trimmed reads in the region)
...pairs with average insert size
(= minimal coverage of trimmed reads in the region)
-> 1
-> 2

Figure 13. LAE-finder - Inversion – Negative support.

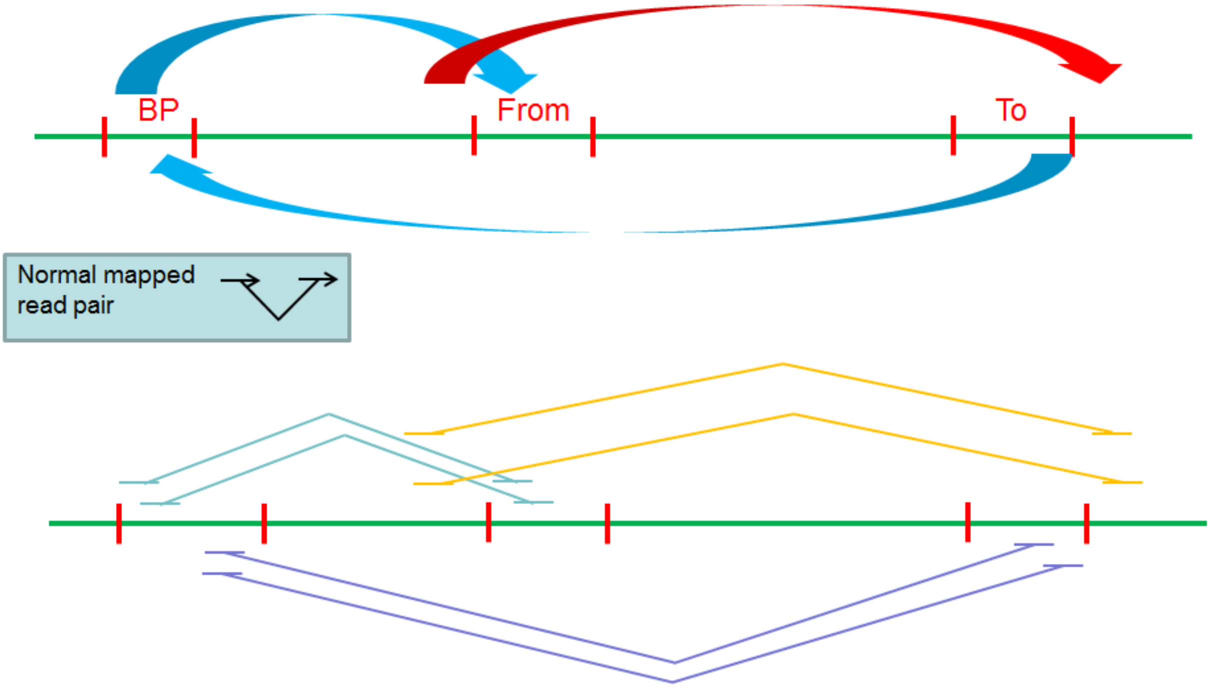


Figure 14. LAE-finder - Translocation – Event support. The arrows in the upper part are for the linear sequence: the sequence starts on the left and then jumps from BP to From, continues to the right and jumps back from To to BP, and so on. The lower part shows two example read-pairs for each jump.

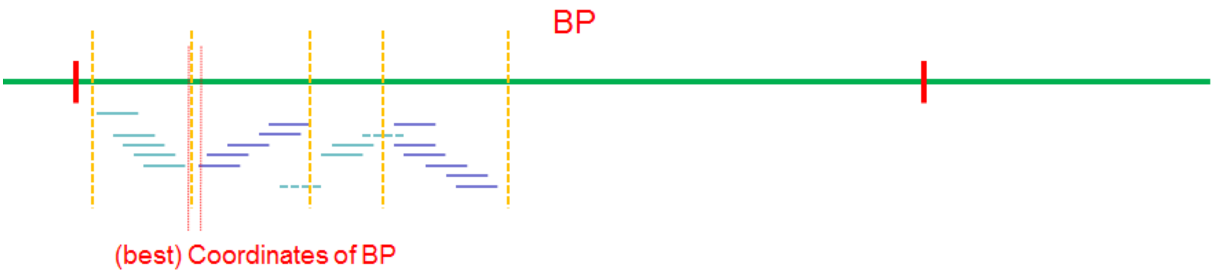


Figure 15. LAE-finder – Translocation – Finding the best breakpoint. The colors correspond to Figure 14. There are four stacks of endpoints from the read-pairs. The best breakpoint for the event is chosen from the count, being non-overlapping and the ratio of orientations. With perfect alignment, there can be only two stacks.

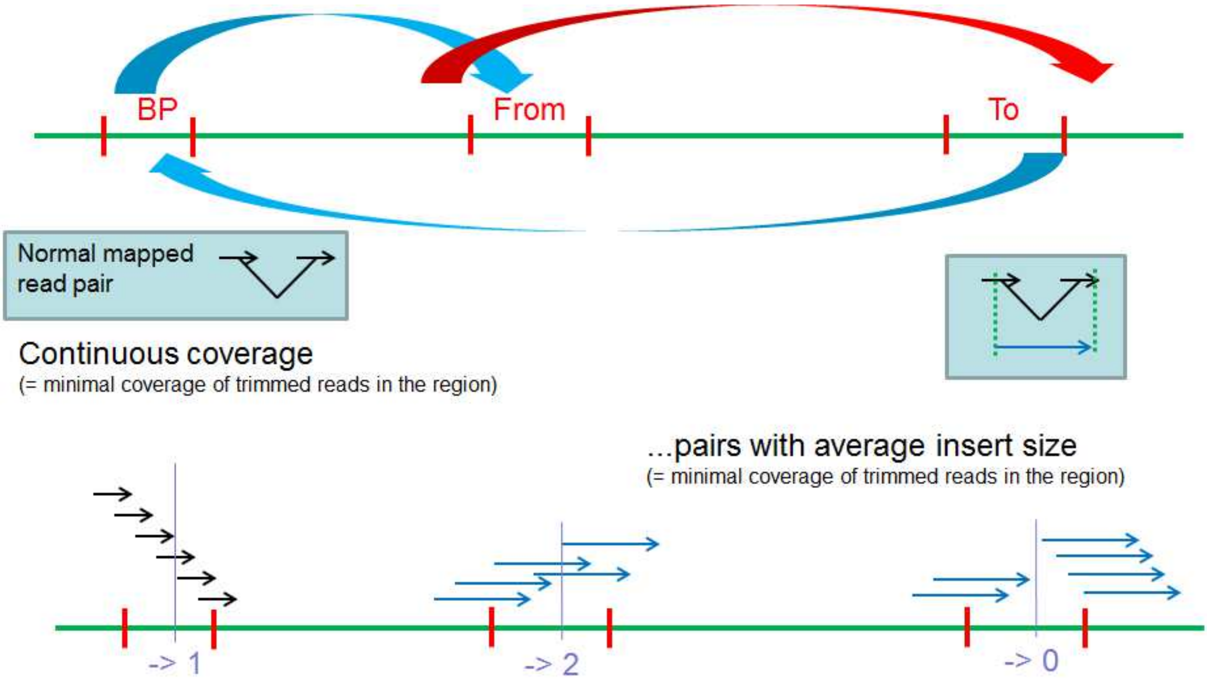

BP
From
To
Normal mapped read pair
Continuous coverage
(= minimal coverage of trimmed reads in the region)
...pairs with average insert size
(= minimal coverage of trimmed reads in the region)
-> 1
-> 2
-> 0

Figure 16. LAE-finder – Translocation – Negative support. The numbers at the bottom are the negative support for the corresponding breakpoints.

The output format of the tool is documented in the readme file at its location [122]. Primarily it counts all signals for presence and absence. This means that there could also be signals for presence and absence at the same time, which would be a strong hint for a suspicious region and/or alignment problems.

For this paper, the major effect was the limitation of the over 200000 putative breakpoints (mentioned in '3.2.2 Quality of calls' in the supplement) to a small fraction of safe calls, reporting 2/3 as wrong and the hint that the linkage disequilibrium is a better guide to large events.

The tool is available at (http://downloads.gmi.oeaw.ac.at/downloads/nordborg/data-release-swedish-lines/programs [122]).

6.7.2 Imputation vs. masking

This topic is mainly applicable when the samples share regions; in other words, it is not applicable for regions which are simply not present in a sample.

Missing data is quite frequent due to the properties of the genome together with the manner of sequencing. Missing data in this respect means unknown; in the case of reference calling it corresponds to 'N' (= it is known that a base is there, but it is not known which one). The sources for it are alignment problems, low coverage, uneven distribution of reads, and so on. Random missing data throughout all loci and all samples would perhaps have minor effects on analyses, but some sources for missing data are systematic, as for example coverage. Several methods would be biased by systematically missing data and some are simply not designed for any missing data, so a general way of dealing with that is needed. The two main directions taken here are: filtering for the parts of the matrix where no data is missing, or imputing the missing data.

The first option, filtering, is also called masking, because parts of genomes are masked in a binary way. Usually, thresholds are defined for where it was possible to call something. It should be noted that many tools only call presence, resp. non-reference, but do not distinguish between reference and unknown. A workaround is then to define thresholds for coverage and quality where it was possible to call something. To use a method which cannot deal with missing data, the resulting masks can be intersected and only the resulting part of the genome used (Figure 17). The advantage of this is that only real data is used, at the cost of using only a fraction of the genome and, of course, removing much data, where the amount of removal is mainly given by the 'weakest' sample.

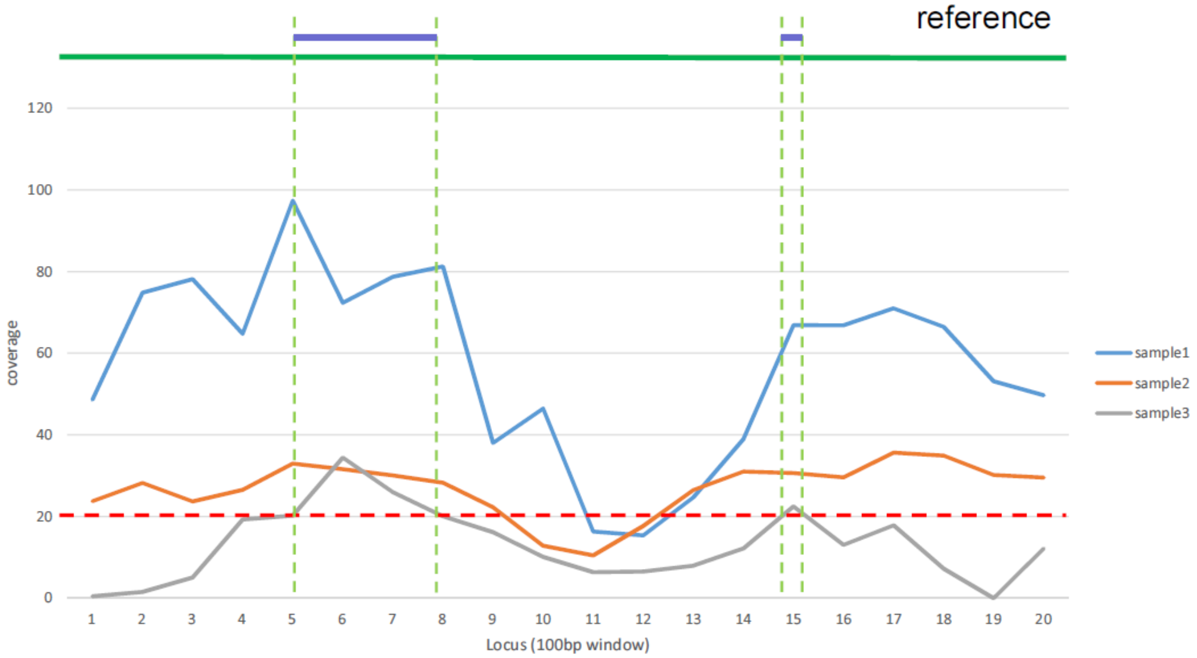


Figure 17. Combining masks of samples. A threshold of coverage >= 20 is defined and only regions of the genome are taken where this is the case for all three samples. The purple bars above the reference in green are the remaining regions. It can be seen that it is mainly the 'weakest' sample (=sample 3) which reduces the remaining genome.

The second option, imputation, approaches the problem from the other side: the missing data is somehow generated. Methods working on sequences build on the fact that events nearby are not independent but in strong linkage disequilibrium. For every missing locus, the methods look for the next present data on both sides and look for the most similar other sample with present data there. The data is then taken for the sample with missing data. The effort and the differences of the imputation methods lies in the fine-tuning, for instance the number of events on both sides, how many close samples, estimate and take error rates and recombination rates into account, and so on. In the paper we used MaCH [123] for this.

6.7.3 Uniqueness / mapping issues

All calling and later analyses are dependent on the alignments. This means that all problems of alignment can affect later analyses. A major property of a small read aligned to a reference is if it is aligned unambiguously or not. If it is unambiguously aligned then it is called unique mapping. There is usually more than one number/flag for this from the aligner, e.g. bwa [124] has a quality score and a flag. The quality score should be zero for ambiguous alignments and larger for unique ones. The flag is a binary information based on the alignment heuristic. If events should be called in a safer manner, it is usually filtered for uniqueness.

Here we present an extension of this concept: the full uniqueness property. It follows the simple idea that one bp can be hit by a certain amount of different reads without sequence errors by exactly n different reads. Each of the reads can be unique for the reference or not, from which a factor of uniqueness can be calculated. The concept is shown in Figure 18.

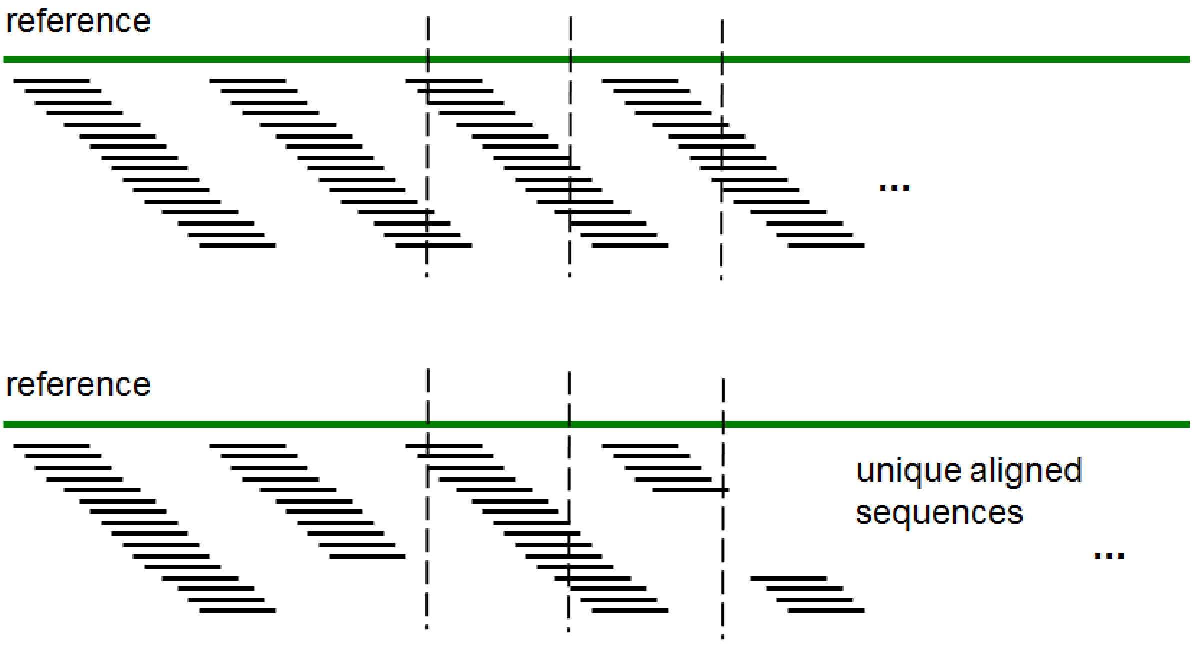
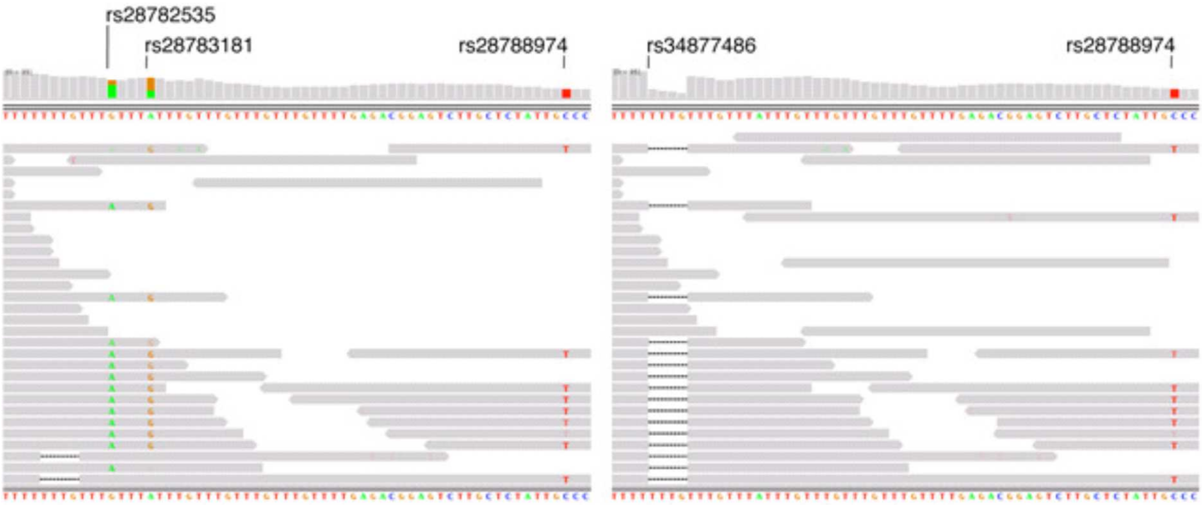


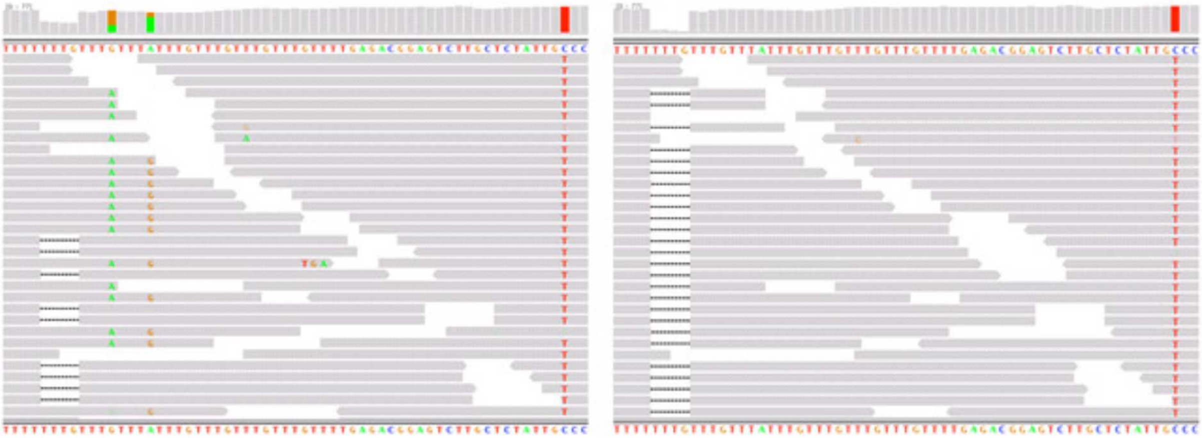
Figure 18. Full-uniqueness property. All possible reads are extracted from the reference. In the upper part, they are just shown as extracted, in the lower part they are filtered for being mapped uniquely. It can clearly be seen that the bp of the dashed line in the middle is hit by more uniquely mapped reads than the bp of the left and the right dashed line. The factor of uniqueness is the number of (perfect) unique reads / all reads.

Another issue is different but equivalent combinations of indels and SNPs. Some of these combinations can be resolved by looking at several reads at once and changing their alignment slightly, a process called realignment. An example of it from the tool GATK [115] is shown in Figure 19. We did this with all of our samples.



1,000 Genomes Pilot 2 data, raw MAQ alignments

1,000 Genomes Pilot 2 data, after MSA



HiSeq data, raw BWA alignments

HiSeq data, after MSA

Figure 19. Sequence realignment. On the left, the mapped reads are shown before realignment, on the right after it. The diagram is from [115].

6.7.4 SVs

As defined in 1.2 General definitions, a SV is an event involving usually a longer sequence, which is not called with a linear alignment from a single read. The following types of SV events exist:

- (long) Insertion
- (long) Deletion
- Duplication
- Copy-number variant
- Inversion
- Translocation

The simplest way of calling SVs with read-pairs is to derive breakpoints from the mates mapping distantly. In case of indels, this can appear as in Figure 20, for inversions see Figure 12.

In case of duplications, the coverage can be used as shown in Figure 21.

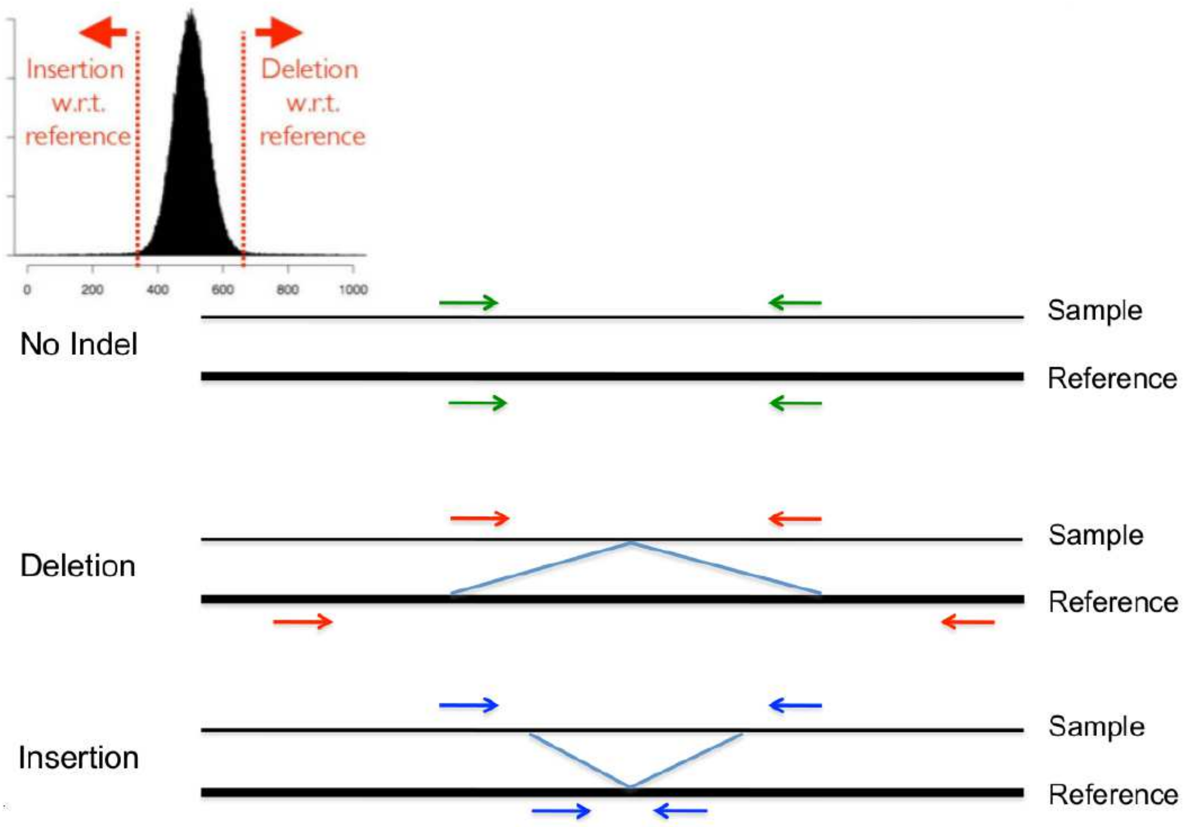


Figure 20. Using read-pairs for inferring insertions or deletions.

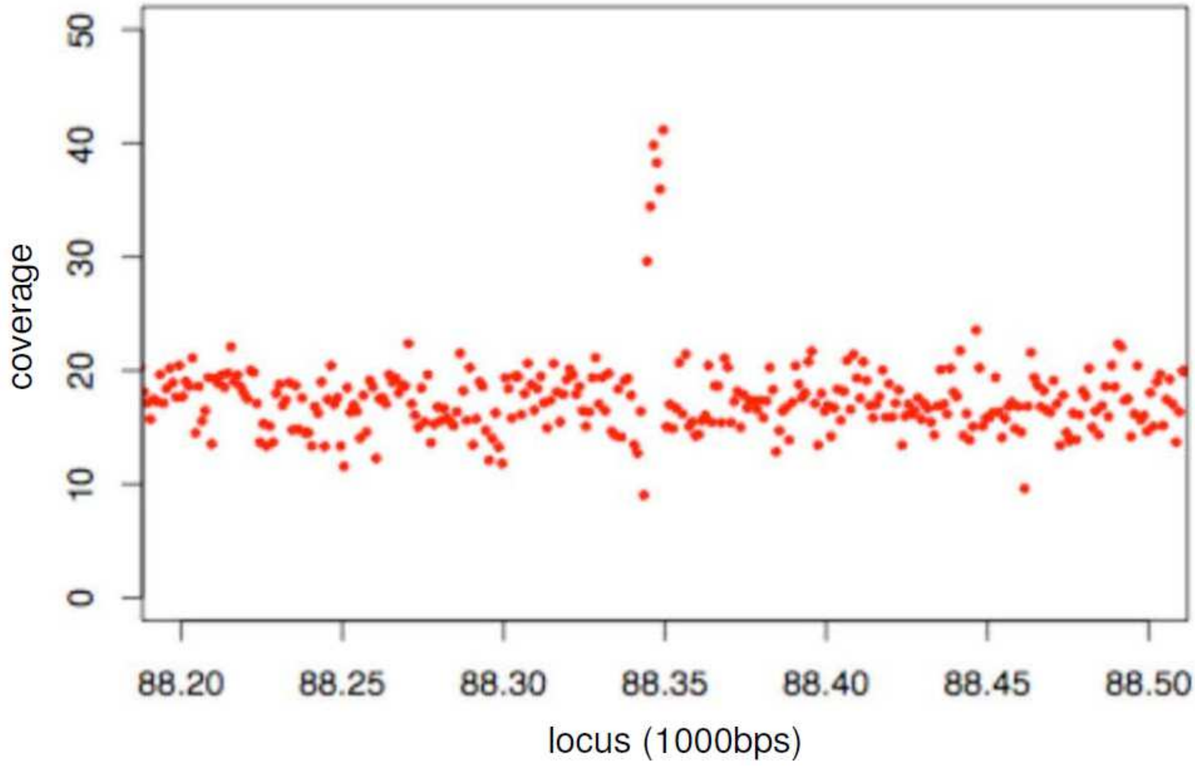


Figure 21. Example coverage for a likely duplication in the middle.

A fuzzy issue, in terms of definitions, are split reads. An example is shown in Figure 22. The difference to ‘normal’ alignments with gap costs is that the gap between the first and the second part can be very large and need not even be linear. Non-linear in this context means one part of the read can be inverted, in the ‘wrong’ order, or on another chromosome. A complete alignment with gap extension cost zero would overlap in the result with the heuristic calling with exact matching of short fragments. This would also mean that the differentiation between indels and SVs is not very precise in general.

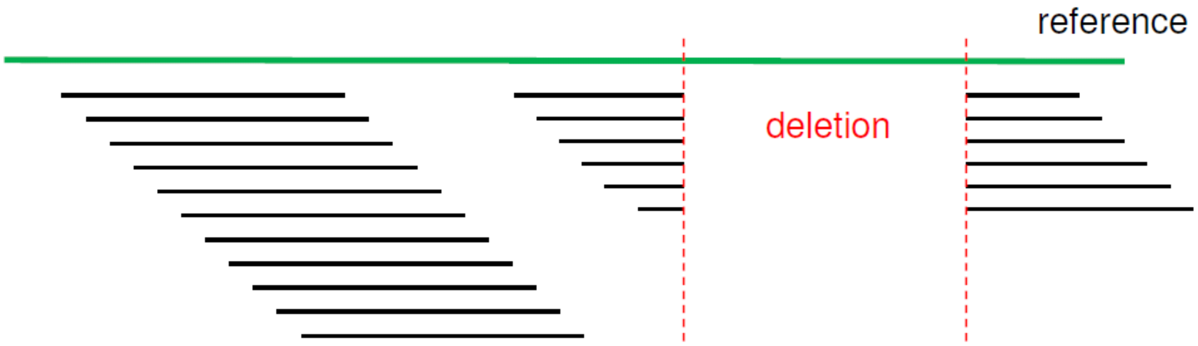


Figure 22. Split reads. The distance between the first and the second part of a read can be large or on different chromosomes.

In the manuscript in section 6.6 [36] all of the mentioned approaches are used.

7. Dimension reduction and Single Nucleotide Polymorphisms

7.1 The curse of dimensionality

The more dimensions the more complex the problem usually appears. This starts with the lack of visualization: more than 3 dimensions are difficult to show in a diagram. An example in this work is in the manuscript of section 8.4 [125], at Figure 3, where 5 dimensions are visible in one diagram; however, this is rather an exception and only possible because 3 dimensions have only discrete values.

Another issue is the observation that the difference between largest and smallest distance in comparison to the smallest distance gets smaller with more dimensions [126], formalized as

$$\lim_{d \to \infty} \frac{dist_{max} - dist_{min}}{dist_{min}} \to 0$$

This is more of an empiric statement than a law. More recently, this argument was partly limited for additional dimensions which do not add information; for dimensions with new information the case is otherwise [127].

The issues for analyzing in more dimensions can be various: loss of generalizability, non-converging results, much larger search space for nominal variables, exploding number of hypercubes to divide the space, more complex manifolds, and so on. One weakness is, of course, also the visualization, the content of the paper of this chapter.

The scientific area dealing with this problem is called ‘dimension reduction’ [128].

Quite in line with this, NGS is one of the largest sources of such data. In this regard, NGS data is not only large with respect to the amount of data, but also in the degree of p larger n: there are easily several millions of variables (e.g. SNPs) and only some 100s of samples.

7.2 Measures for dimension reduction

The task of dimension reduction is, as the term says, the transformation into a space of fewer dimensions. This is also called feature extraction.

Clearly, this cannot be done perfectly in every case as some information must be lost with a reduction in the number of dimensions and not all topological features may be possible in fewer dimensions. Several methods exist for this problem [129]. The surrounding problem for these methods is to judge different transformations. This can be done internally or with external information. Without external information, it is mainly about preserving measuring distance. As the distances cannot be the same in low-dimensional space (unless the data used only a hyperplane), the similarity in distance ranks or the distribution is measured. Examples for this are the Kullback–Leibler divergence [130] and other cases of the f-divergence [131], rank preservation [132], and differential entropy [133].

When external data is available more can be done on judging the dimension reduction. This external information must be one of the largest effects, because it is usually demanded that the reduction in dimensions should preserve the most important effects/information in the data. This can be seen as a similar problem to cluster validation. The difference is in the source of the data, that is, the labels for the data records are not given by a clustering method but are given by external data. Examples for cluster validation methods are Dunn's Validity Index [134] and Silhouette Validation Method [135], as used in the paper in the next section.

More generally, the reduction should preserve the data structure according to the ability to build classification models on that, which is the topic of the next section.

7.3 Article: **Visualization of SNPs with t-SNE**

Platzer, A., *Visualization of SNPs with t-SNE*. **PLoS One, 2013. 8(2): p. e56883.**

This paper regards the general but small task of visualizing high dimensional data. t-SNE [136] was chosen as one new method, where the main contribution of this paper is the combination of data, a new but known method, and new validation measure. It will likely not convince most biologists to replace their favorite method PCA with anything else, yet the topic is conceptually rich in terms of presentation esthetics and simple application to different data.

OWN CONTRIBUTION IN [137]

Everything.

Visualization of SNPs with t-SNE

Alexander Platzer*

Gregor Mendel Institute, Vienna, Austria

Abstract

Background: Single Nucleotide Polymorphisms (SNPs) are one of the largest sources of new data in biology. In most papers, SNPs between individuals are visualized with Principal Component Analysis (PCA), an older method for this purpose.

Principal Findings: We compare PCA, an aging method for this purpose, with a newer method, t-Distributed Stochastic Neighbor Embedding (t-SNE) for the visualization of large SNP datasets. We also propose a set of key figures for evaluating these visualizations; in all of these t-SNE performs better.

Significance: To transform data PCA remains a reasonably good method, but for visualization it should be replaced by a method from the subfield of dimension reduction. To evaluate the performance of visualization, we propose key figures of cross-validation with machine learning methods, as well as indices of cluster validity.

Citation: Platzer A (2013) Visualization of SNPs with t-SNE. PLoS ONE 8(2): e56883. doi:10.1371/journal.pone.0056883

Editor: Vladimir Brusic, Dana-Farber Cancer Institute, United States of America

Received November 9, 2012; **Accepted** January 18, 2013; **Published** February 15, 2013

Copyright: © 2013 Alexander Platzer. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Funding was provided by the Austrian Academy of Science (http://www.oeaw.ac.at/english/home.html). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: alexander.platzer@gmi.oeaw.ac.at

Introduction

SNPs are a major part of the extracted information from individual genomes. With the vast amount of NGS data, 100.000 s of SNPs can be found in a population today; naturally these are somewhat difficult to visualize. The most traditional method is PCA [1], which is still used in the majority of biology articles (e.g. [2–4]). PCA is designed for an orthogonal transformation, resulting in a number of components equal than or less to the number of original variables. These components are usually sorted for their explained variance.

At that point the assignment of a causing effect to the first components is attempted (e.g. [5–7]). For a correct assignment several constraints should be fulfilled [8].

Another usage is to plot the data with 2–3 higher components with primarily the first two or three principal components being plotted [2–4]. Due to the occasionally somewhat unsightly diagrams, several approaches to improve visualization with PCA have been developed (e.g. [9]).

In another field, that of machine learning, this problem of data reduction, often especially for visualization, has developed into its own subfield, 'dimension reduction', which was first outlined with the introduction of the term 'the curse of dimensionality' [10]. In this field several other methods have been developed since PCA, such as Sammon mapping [11], Isomap [12], Locally Linear Embedding [13], Classical multidimensional scaling [14], Laplacian Eigenmap [15], m-SNE [16], t-SNE [17], and others.

In this article we will focus on t-SNE as one of these newer methods and compare it with PCA in several ways.

The first step in comparing visualizing methods is of course to take several complex data sets, make diagrams, and discuss them. Decisions on aesthetic or artistic value may be made, but naturally more or less solid key figures for contrasting would be desirable.

The question of the quality of a visualization can be split in two parts: how well is the data structured; and how much (correct) insight can be obtained from it?

For biological data, the second question can often get out of hand; we will rather focus here on the first question.

Regarding the question of the structuredness of data, there exist long-known indices of cluster validity, such as Dunn's Validity Index [18], Silhouette Validation Method [19], and others (for an overview see [20]). But the property of structure can also be approached from another perspective: How easily may a model be built for the transformed data?

This question can be answered with splitting the data in two parts, use one part for constructing a model and the other to test it. The easier the structure of the data, the higher should be the validation key figure, assuming the model learning method makes an equal effort. We choose several machine learning methods for this purpose and compare their results for the different transformed data.

Here we show a comparison between the common PCA and the newer t-SNE on several large SNP datasets with a number of evaluating key figures.

Results

In Figure 1 and Figure 2 we show the visual results of our chosen large SNP data sources transformed with PCA and with t-SNE. In light of the good separation, we should repeat here that both methods are unsupervised, that is, neither methods received labels and the colors were added after transformation. Visually, the t-SNE transformed data looks 'nicer' (our opinion and that of nearly all colleagues). The only mentioned drawback is that no extra biological information can be seen from the diagrams on the right. Here we will leave for discussion a final conclusion on the amount of biological impact and focus on the structuredness of the data. For this purpose, Table 1 contains the respective cluster validity indices for all diagrams, the values of Dunn's Validity Index, and the Silhouette Validation Method. The higher these values, the better the cluster separation, which corresponds to the structuredness. Both methods rely on the pairwise distances of the data points. Dunn's index can be used in different ways: We took the average function (the diameter of a cluster and the distance of clusters are defined generically at this method; for both we choose the average of pairwise distances). It is to mention that a Silhouette value lower than zero makes not much sense in terms of validation, as it means a random label assignment would be 'better'. This occurs for the rice data (Figure 2b) because several clusters appear to consist of more than one real cluster, which are surrounding other clusters. Either the label (= country) is a too low resolution or the location is not one of the largest effects in this data.

As a more general measure of the structuredness of the transformed data, we formulate it as a supervised classification problem. The underlying rationale is that, if the data is well-structured, it should be easier for any method to construct a good model for it. We choose here C4.5, PART, Neural Networks, and naïve Bayes [21–24]. To judge a method's performance on a dataset we use the percent correctly classified of the 10-fold cross-validation. These results are presented in Table 2. Here, the difference in this key figure between PCA and t-SNE transformed data of the same source using the same learner should express the difference in structuredness of the two transformations. Mean values and standard deviations are only there per population as these populations are not fully comparable.

For our large SNP data sources we selected the 1001 genomes project [25], the RegMap panel [26], hapmap3 release 2/3 [27] and the Rice Haplotype Map Project [28]. We picked a subset from the 1001 genomes project, firstly because it seemed at the time of analyzing that the data would not be fully complete in the near future, and secondly, for the equal class sizes. With very unequal class sizes, both PCA and t-SNE suffer, as we could see with different unequal subsampling (not shown), and as also stated by [29,30]. *Class* is here and later referring to the labels of the data records, which are geographic locations in this paper. The next three datasets are taken as they were released, whereas the last dataset (Rice) was filtered for wild rice and for available labels (= country).

The species of the first two datasets is *Arabidopsis thaliana*, the species of the next two is human (they are just different releases of the same effort). The last dataset is from a collection of rice. For all species a solid assumption seems to be that a large effect in the genomes is linked with their geographic location [31,32].

As can be seen, all key figures to measure the structuredness of the transformed data point in the same direction (except for the RegMap data, where the Dunn Index is the same). A clear answer to the question of which transformation leads to better structured data thus materializes: there needs to be a movement away from PCA.

Discussion

The main purpose of this paper is to show an approach for testing possible transformations of SNP/biological data to 2 dimensions for visualization. Many more methods exist than t-SNE and PCA [33], though some do seem theoretically and practically outperformed by others.

SNPs are one of the largest sources of new data in biology, but until now none of this data has been in main machine learning repositories (e.g. [34]). This will change in the future as certain SNP data generating projects finalize.

We made two attempts to measure structuredness, which strongly correlates to what most consider the better scatter plot. The sources here are, on the one hand, merely much discussion with no exhaustive survey. But our intention, on the other hand, was to express this in numbers from the start. Our first approach uses cluster validity key figures, despite their known weaknesses [35]. Our second approach uses machine learning methods, following the rationale: if a moderately complex algorithm can more easily gain some 'understanding', and/or build a relatively better internally validated model, possible human insight should correlate to that. As measure for the machine learning methods we use the percent correctly classified.

For machine learning methods themselves, there is of course only little gain, since other approaches [33] exist to deal with (too) many dimensions than to transform them to exactly two. Newer methods of this type are usually able to perform better with the full data, or with data not more than *sufficiently* reduced [36].

By performance we mean the result, not the computational effort, which can sometimes overload the frame. In the context of machine learning method performance, transformation of the data to two dimensions can be seen as a loss of information, which could be described by how much these methods lose in constructing models. The measure here could again be the percent correctly classified of the 10-fold cross-validation. The transformation that lets to the smallest decrease for all methods eligible for this classification problem should be judged better.

The four machine learning methods were not the only tested methods; we choose these four because of their high performance.

As mentioned above the structuredness of the transformed data is merely the first part of the various biological questions, the second always regarding the biological impact. There are several systematic attempts to directly translate it into biological information [2,5–7]. Some appear quite convincing, while others seem more tweaks of the transformation. Nonetheless, with these attempts some insights have been obtained in this manner (e.g. [5–7]).

There may be other constraints that a dimension reduction method should fulfill to gain biological insight in other than standard classification problems. Like clustering in general this may simply remain ill-defined [37,38].

Materials and Methods

PCA

For PCA, the build-in R function *prcomp()* is used.

t-SNE

This method is presented in [17].

Beside the pseudocode of the simple version (Figure algorithm 1 in [17]), several 'tricks' and heuristics are used to make the results more attractive and/or the computation faster. All parameters for these 'tricks' are set within the method.

In the article ([17]), several other methods are compared and likely reasons for their worse performance are discussed. Some weaknesses also remain for t-SNE:

Dimension reduction for more than 3 dimensions: This was not a topic in designing t-SNE or in first testing, as it is irrelevant for visualization.

Curse of intrinsic dimensionality: Besides the general issue that dimension reduction always means that some information is lost, this targets the local linearity assumption of the method.

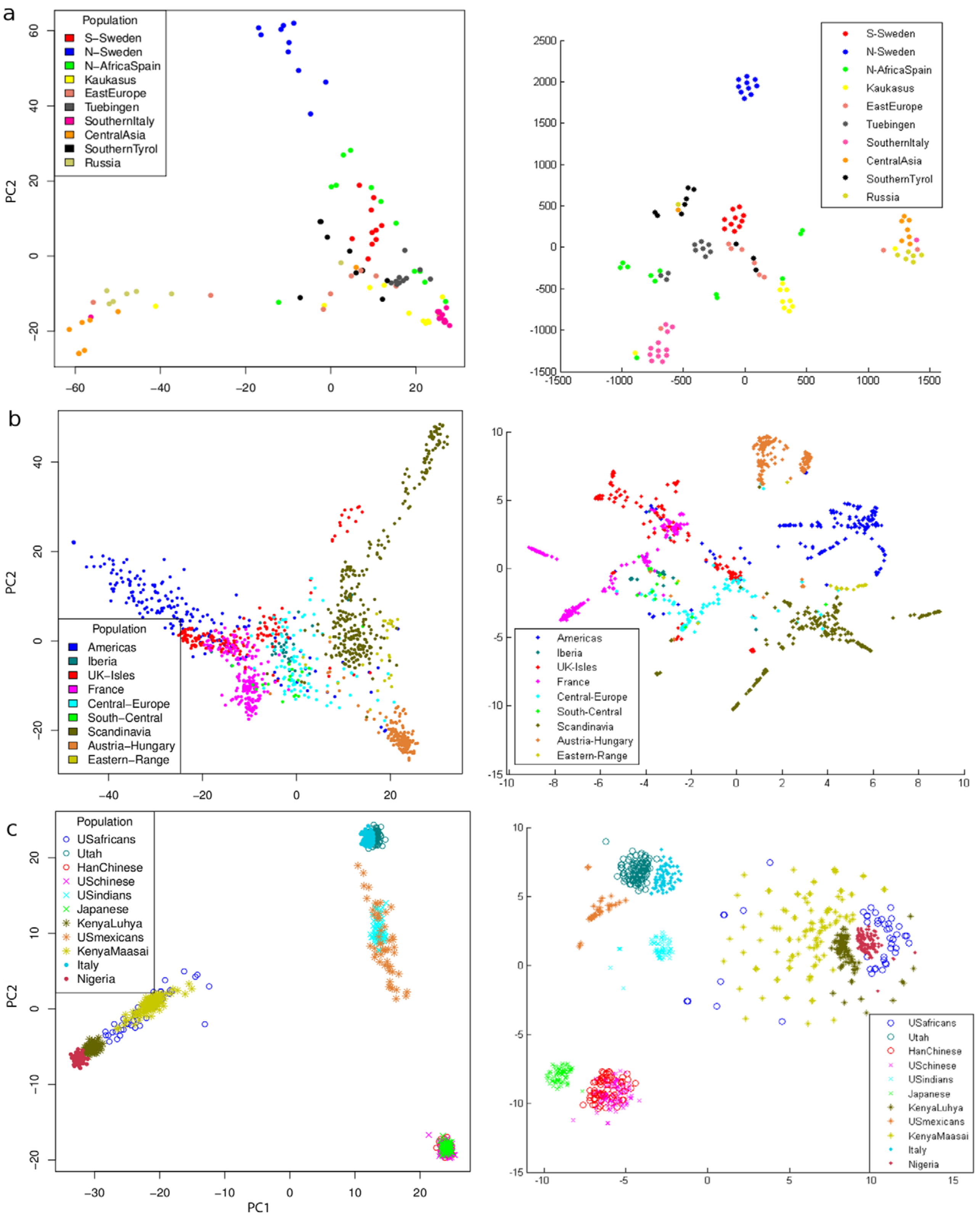


Figure 1. SNP data transformed with PCA and t-SNE 1/2. On the left is a PCA-plot with the first two components, on the right a t-SNE-plot of the very same data from each data source. Data sources: Panel (a) is from the 1001 genomes project, (b) from the RegMap panel and (c) from hapmap3 r2.
doi:10.1371/journal.pone.0056883.g001

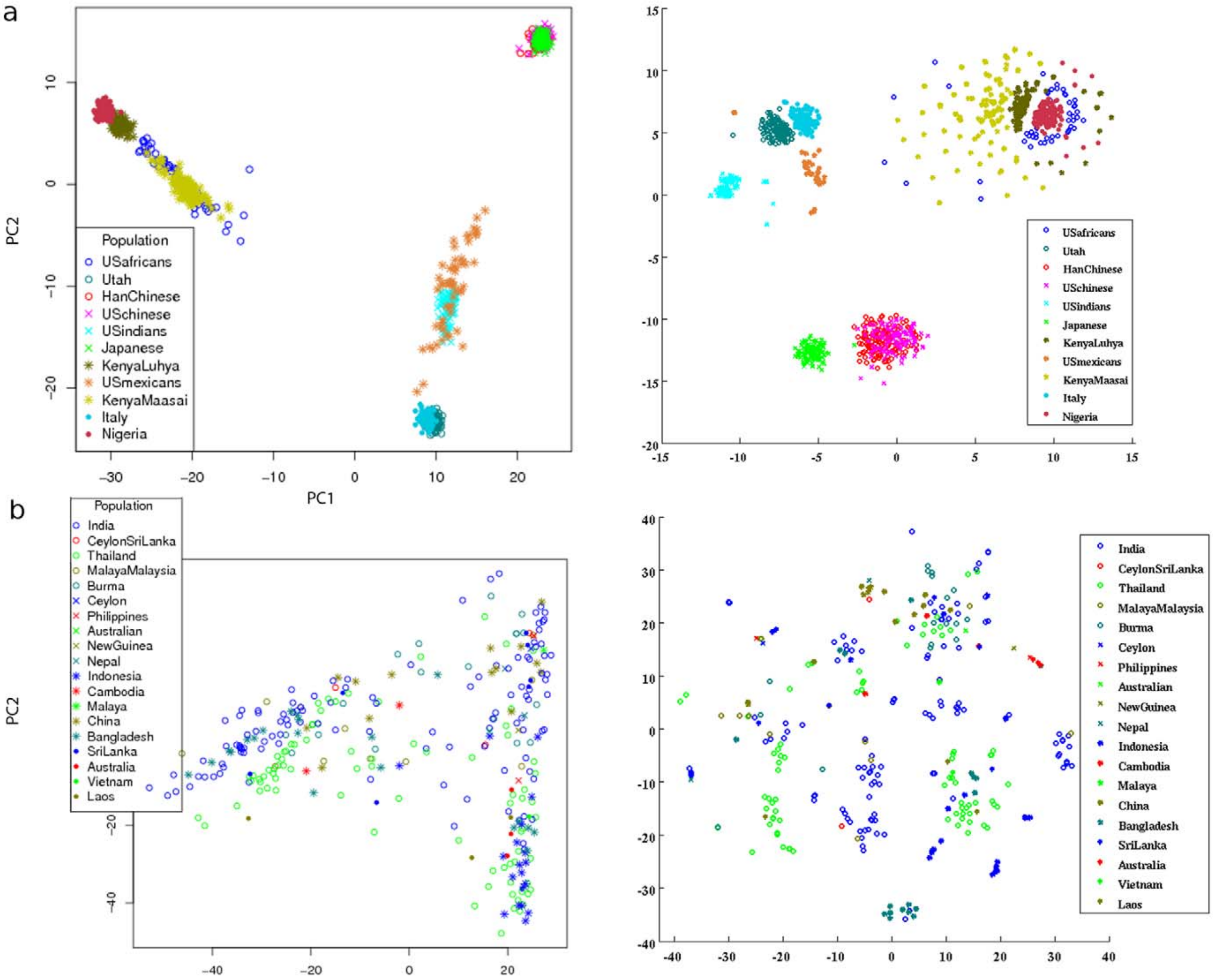


Figure 2. SNP data transformed with PCA and t-SNE 2/2. On the left is a PCA-plot with the first two components, on the right a t-SNE-plot of the very same data from each data source. Data sources: Panel (a) from hapmap3 r3 (compare with Fig. 1c) and (b) from the Rice Haplotype Map Project (only wild type where the label information was available).
doi:10.1371/journal.pone.0056883.g002

Table 1. Dunn's Validity Index and Silhouette Validation Method of the transformed SNP data.

Data	Dunn's Validity Index			Silhouette Validation Method		
	PCA	*t-SNE*	*Diff*	*PCA*	*t-SNE*	*Diff*
1001 genomes	0.52 (0.09)	0.61 (0.07)	0.09	0.07 (0.04)	0.22 (0.04)	0.15
RegMap	0.50 (0.06)	0.50 (0.04)	0.00	0.08 (0.02)	0.15 (0.02)	0.07
Hapmap3R2	0.16 (0.01)	0.25 (0.02)	0.09	0.27 (0.02)	0.31 (0.02)	0.04
Hapmap3R3	0.16 (0.01)	0.35 (0.01)	0.19	0.26 (0.02)	0.32 (0.02)	0.06
Rice	0.06 (0.07)	0.10 (0.10)	0.04	−0.54 (0.04)	−0.46 (0.04)	0.08

The values of two indices of cluster validity as a measure for structuredness of the different transformed data. As a comparison between PCA and t-SNE the diff(erence) column is expressive. The number in brackets is the standard deviation of the index with 1000 permutations of the labels.
doi:10.1371/journal.pone.0056883.t001

Non-convexity of the t-SNE cost function: This is one reason for the need for heuristics and tricks in the computation and the risk of not ending in the global optimum.

For t-SNE the matlab reference implementation is used [39].

There are two parameters for this implementation: *init_dims* and *perplexity*. *init_dims* is a preprocessing reduction with PCA to eliminate the most likely noise with skipping components with virtually no variance; it makes the computation faster. *perplexity* is used as defined in information theory, for example in [40]. Perplexity can be interpreted in this method as a smooth measure of the effective number of neighbors.

Unfortunately this version is restricted to 32bit, which entails a 2GB memory limit. There are other reference implementations, but all are restricted in memory usage at the moment.

Our chosen data sources would have required more total memory; to still allow the analysis, the data was downsampled to fit in 2GB memory. The same downsampled data was used also for the PCA.

Table 2. Percent correctly classified with various machine learning methods acting on transformed SNP data.

1001 genomes project			**RegMap**		**hapmap3 r2**		**hapmap3 r3**		**Rice**	
%	*PCA*	*t-SNE*	*PCA*	*t-SNE*	*PCA*	*t-SNE*	*PCA*	*t-SNE*	*PCA*	*t-SNE*
C4.5	55.6	72.7	79.2	89.7	72.9	90.5	72.9	87.5	41.3	66.6
PART	60.6	76.8	77.6	89.1	72.7	90.9	73.3	87.6	39.7	64.9
Perceptron	67.7	76.8	80.7	85.8	70.3	85.1	72.2	84.8	50.5	56.4
Naive Bayes	62.6	75.8	75.2	80.3	74.6	87.2	71.8	84.1	40.7	42.3
Mean diff.	13.9		8.1		15.8		13.5		14.5	
St.dev.	3.6		3.4		2.6		1.2		12.5	

The percent correctly classified as a measure how easy a model can be learned. As comparison between PCA and t-SNE, the respectively difference between these two columns is expressive. All models are better than random.

doi:10.1371/journal.pone.0056883.t002

PCA vs. t-SNE

The most notable differences between the methods PCA [1] and t-SNE [17]:

- PCA splits the data into n components, sorted for variance (where n is the number of variables), whereas t-SNE squeezes all information in m components (where m is freely to choose, in case of plots m = 2)
- PCA is a static transformation: with one input there is always exactly one output (conditions for ambiguous cases are also precisely defined) t-SNE is a non-static transformation: with the same input there are different outputs possible, especially as the method is till now only feasible as more or less stepwise optimization; but also if the best value in terms of the cost function is always found, there will be several results because the method/optimization-criteria is rotation and scale-invariant
- PCA has several constraints [8], which are tackled in t-SNE
- PCA is an orthogonal linear transformation, whereas t-SNE is a nonlinear reduction, which 'components' are not constrained to be orthogonal.

The first of these points is the main convincing reason why PCA should not be the only plot in case of high dimensional data. As long as the number of dimensions is not too high, it is more likely that the first few (for a plot = 2 or 3) PCA components explain a lot of the data variance. If the first few components explain only little variance, then there is a big gain if a method integrates the rest of the data well, or put it in a different way: in a PCA plot there is always the information of n-2 components left out, where in t-SNE all information is tried to be combined.

Of course, if one of the largest effects in the data is perfectly correlating with the first two PCA components, then this transformation would be 'better' in terms of this effect. In SNP data this is usually not the case, otherwise a lot of published plots would look different and also the conclusion of this paper would be the opposite.

Data sources

The 1001 genomes project [25] is one of the largest sources of SNP/genomic data for *Arabidopsis thaliana*, even though the data generation of this project is not finished. We used a subset of 99 individuals, selected for equal class sizes.

The Regional Mapping Project is another source for *Arabidopsis thaliana*. Though it has a lower resolution of SNPs, it is already finished. We have taken the same 1090 individuals as in the article's [26] PCA-plot.

The HapMap Project [27] is a large source of human genetic variation. We used the data from the second release of phase III, 988 individuals' sets of SNPs. We used also the third release of phase III as own dataset, because it was not sure at last if all issues were already resolved within (1198 individuals, should be a superset of the second release).

The Rice Haplotype Map Project [28] is the largest source of SNP/genomic data for rice. We have filtered here for the wild rice (species Oryza rufipogon) where the country of origin was available in the database (305 individuals).

Indices of cluster validity

The transformed and labeled data can be seen as a result of a clustering method, although it is not gained in that manner: As result from clustering the labels would be assigned through the clustering method, whereas in our case the labels are the true (external) classes and the values of the variables are 'generated' (= transformed original values). That means that the problem of judging the structuredness of our transformed data with the true classes is similar to judging the result of a clustering. For this reason we are able to use internal evaluation methods, although it is an external validation.

We choose Dunn's Validity Index [18] and the Silhouette Validation Method [19] for this purpose.

Dunn's Validity Index and Silhouette Validation Method

A good short description of both methods can be found in Wikipedia ([41,42]). Both methods rely on the pairwise distances of the data points within a cluster in comparison with distances within different clusters. Beside the chosen distance/dissimilarity, the main difference is that the Dunn Index looks for the worst combination (the maximal intra-cluster distance to the minimal inter-cluster distance), whereas the Silhouette Validation Method is taking the average of all cluster combinations (more precisely, the Silhouette Validation Method is originally defined for two clusters and we (/our chosen implementation) took the arithmetic mean of all combinations).

For Dunn's Validity Index we used the R package 'clv' [43] and for the Silhouette Validation Method we used the R package 'cluster' [44].

Classification methods

For constructing models for classification we use four standard machine learning methods:

- The well-known tree learner C4.5 [21] and the not very widely used method PART [23] relying on C4.5.
- A Neural Network [22] with one hidden layer (5–7 hidden nodes).
- Naïve Bayes [24]

The analysis with these classification methods was performed with WEKA [45].

Acknowledgments

We are grateful to Quan Long and Matthew Horton for discussions and to Thomas Friese for proof-reading.

Author Contributions

Analyzed the data: AP. Contributed reagents/materials/analysis tools: AP. Wrote the paper: AP.

References

1. Pearson K (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine 2: 559–572.
2. Sun Z, Chai HS, Wu Y, White WM, Donkena KV, et al. (2011) Batch effect correction for genome-wide methylation data with Illumina Infinium platform. BMC Med Genomics 4: 84.
3. Swingley WD, Meyer-Dombard DR, Shock EL, Alsop EB, Falenski HD, et al. (2012) Coordinating environmental genomics and geochemistry reveals metabolic transitions in a hot spring ecosystem. PLoS One 7: e38108.
4. Zhou H, Muehlbauer G, Steffenson B (2012) Population structure and linkage disequilibrium in elite barley breeding germplasm from the United States. J Zhejiang Univ Sci B 13: 438–451.
5. Hurtado MA, Racotta IS, Arcos F, Morales-Bojorquez E, Moal J, et al. (2012) Seasonal variations of biochemical, pigment, fatty acid, and sterol compositions in female Crassostrea corteziensis oysters in relation to the reproductive cycle. Comp Biochem Physiol B Biochem Mol Biol.
6. Jarzynska G, Falandysz J (2011) Selenium and 17 other largely essential and toxic metals in muscle and organ meats of Red Deer (Cervus elaphus)–consequences to human health. Environ Int 37: 882–888.
7. Yu Z, Tan BK, Dainty S, Mattey DL, Davies SJ (2012) Hypoalbuminaemia, systemic albumin leak and endothelial dysfunction in peritoneal dialysis patients. Nephrol Dial Transplant.
8. A Tutorial on Principal Component Analysis. Available: http://www.snl.salk.edu/~shlens/pca.pdf. Accessed: 2013 Jan 21.
9. Lu H, Plataniotis KN, Venetsanopoulos AN (2008) MPCA: Multilinear Principal Component Analysis of Tensor Objects. IEEE Trans Neural Netw 19: 18–39.
10. Bellman RE (1961) Adaptive Control Processes. Princeton, NJ: Princeton University Press.
11. Sammon JW (1969) A Nonlinear Mapping for Data Structure Analysis. Ieee Transactions on Computers C 18: 401-&.
12. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290: 2319-+.
13. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290: 2323–2326.
14. Torgerson WS (1952) Multidimensional Scaling: I. Theory and Method. Psychometrika 17: 401–419.
15. Belkin M, Niyogi P (2002) Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing Systems 14, Vols 1 and 2 14: 585–591.
16. Xie B, Mu Y, Tao DC (2010) m-SNE: Multiview Stochastic Neighbor Embedding. Neural Information Processing: Theory and Algorithms, Pt I 6443: 338–346.
17. van der Maaten L, Hinton G (2008) Visualizing Data using t-SNE. Journal of Machine Learning Research 9: 2579–2605.
18. Dunn JC (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics 3: 32–57.
19. Rousseeuw PJ (1987) Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. Journal of Computational and Applied Mathematics 20: 53–65.
20. Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. Journal of Intelligent Information Systems 17: 107–145.
21. Quinlan JR (1993) Programs for Machine Learning: Morgan Kaufmann Publishers.
22. Rumelhart DE, Geoffrey E . Hinton, and R. J . Williams (1986) Learning Internal Representations by Error Propagation. In: Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge: MIT Press. pp. 318–362.
23. Frank E, Witten I. Generating Accurate Rule Sets Without Global Optimization; 1998; Shavlik. Morgan Kaufmann Publishers, San Francisco, CA.
24. Zhang H (2005) Exploring conditions for the optimality of Naive bayes. International Journal of Pattern Recognition and Artificial Intelligence 19: 183–198.
25. Weigel D, Mott R (2009) The 1001 Genomes Project for Arabidopsis thaliana. Genome Biology 10.
26. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, et al. (2012) Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nature Genetics 44: 212–216.
27. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.
28. Huang X, Kurata N, Wei X, Wang ZX, Wang A, et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature 490: 497–501.
29. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. Bmc Medical Informatics and Decision Making 11.
30. Lin WJ, Chen JJ (2012) Class-imbalanced classifiers for high-dimensional data. Brief Bioinform.
31. Nothnagel M, Lu TT, Kayser M, Krawczak M (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet 19: 2927–2935.
32. Sharbel TF, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. Mol Ecol 9: 2109–2118.
33. Laurens van der Maaten EP, Jaap van den Herik (2009) Dimensionality Reduction: A Comparative Review. Tilburg: Tilburg University.
34. Asuncion AFaA (2010) UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.
35. Christopher D. Manning PR, Hinrich Schütze (2009) An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press.
36. Amir Globerson NT (2003) Sufficient Dimensionality Reduction. Machine Learning Research 3: 1307–1331.
37. Ackerman M, Ben-David S. Clusterability: A Theoretical Study; 2009; Clearwater Beach, Florida, USA.
38. Pau G (2010) Clustering and classification with applications to microarrays and cellular phenotypes. Bressanone-Brixen, Italy: Computational Statistics for Genome Biology 2010.
39. t-Distributed Stochastic Neighbor Embedding - Implementations. Available: http://homepage.tudelft.nl/19j49/t-SNE.html. Accessed: 21 June 2012.
40. Brown PL, Della Pietra V, Lai JC, Mercer RL (1992) An estimate of an upper bound for the entropy of English. Computational Linguistics 18: 31–40.
41. Dunn index – Wikipedia, The Free Encyclopedia. Available: http://en.wikipedia.org/w/index.php?title = Dunn_index&oldid = 511861769. Accessed: 7 Jan 2013.
42. Silhouette (clustering) – Wikipedia, The Free Encyclopedia. Available: http://en.wikipedia.org/w/index.php?title = Silhouette_(clustering)&oldid = 528712368. Accessed: 2013 Jan 7.
43. Tibshirani R, Walther G, Hastie T (2000) Estimating the number of clusters in a dataset via the Gap statistic. 63: 411–423.
44. Struyf A, Hubert M, Rousseeuw P (1997) Clustering in an Object-Oriented Environment. Journal of Statistical Software 1: 1–30.
45. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20: 2479–2481.

7.4 Results – addendum

The contribution of this chapter's manuscript was the combination of biological data with a newer dimension reduction method (t-SNE) and the introduction of a new measurement for structuredness. As the chosen method is designed explicitly for a reduction to not more than 3 dimensions, it is entirely concerned with visualization. Although the amount of datasets appears small, it was all the reliable data of this kind we were able to find within the time. One reviewer, clearly competent in analysis but not familiar with this kind of data, requested more datasets. Only the datasets of Figure 1 were in the paper at the first submissions; after a great effort looking for more data, Figure 2 was the result. This was not strikingly more as hapmap r3 and r2 are closely related (some colleagues told me that r3 is later, but slightly less reliable), and the rice data itself appeared not very solid as there is much less structure in the data and more then 1/3 was removed because the label was not available. The reasons for the few datasets to date are the demands of the analysis: the raw calls of a larger population sequenced in the same way and a label of one of the largest effects available for them. A large proportion of sequenced data is of humans, but often there are no raw calls freely available because of privacy concerns. One exception besides the hapmap project is openSNP [83], which is extensive and free but unfortunately unreliable. We observed the latter weakness when looking for the labels of the data records: there are data records with extensive information, some with almost no additional information, and some seem to be almost comical entries. We looked also into data collections from corn (*Zea mays* subsp. *mays*; data unpublished) and fruit fly - *Drosophila simulans* [138], but both sources were unsuitable for our purpose. Nevertheless, if sequencing continues as now, there will soon be much more and also more reliable data sets.

The measurements for structuredness in the paper are rather sufficient as there are not many. On the other hand, there are many more dimension reduction methods than just pca and t-SNE. In only one more extensive (Matlab) Toolbox for Dimensionality Reduction [129] 34 methods are implemented. For a poster for an event during a course at the TU Vienna we extended the paper's main findings for 2 more methods: Isomap [139] and LLE [140]. As expected from their design, the two methods performed worse than t-SNE, but surprisingly PCA performed similar well as Isomap and LLE. The poster is available at [141].

8. GWAS and calling transposons from paired-end reads

8.1 The challenge of transposons in current sequencing technologies

As described in section 1.1.5, the huge current amount of sequencing data is in the form of plenty of short reads. All later analyses depend either on the correct alignment of these reads to a reference sequence or to the correct de novo assembly. Both approaches suffer from duplicated and highly similar regions. Some TEs are present in more than one copy and several TEs are quite similar in sequence. We focus here on the reference-based-assembly, because the de novo assembly with highly similar and duplicated regions is even more difficult.

To get a translocation (a jumping TE is either a translocation or a duplication), a linearly sequenced piece of DNA must be aligned to more than one location on the reference. In the case of a single read, this is called 'split read alignment' (see Figure 22). In the case of a read-pair, the sequence is already split into two mates, which are usually aligned individually. This can be used to call events as for example shown for an inversion in Figure 12. During our search for tools for our quantity of read-pair data, we were surprised that no tool existed that use the read-pair information to call TEs. For this reason and because it is/was not much algorithmic effort, we created such a tool (-> TE-locate, section 8.4).

8.2 Roles of transposons

From the manuscript of this chapter:

'Transposable elements (TEs) have made themselves a great career, from being junk DNA [142] when first discovered [143], to having important roles in development [144], evolution [145, 146], and disease [147] through direct genome rejoining [148], epigenetic control [149, 150], or other known [151] or to-be-tested mechanisms [152].'

8.3 GWAS

Nowadays, genome-wide association studies are standard, although the term is a generic one for doing just what the name says: finding associations in all events called in a set of genomes. Usually the attempt is made to associate these events with a phenotype, or the opposite, the label or value to be associated with is called phenotype. This may include, for instance, expression, metabolite levels, and environmental states; the major overall constraint is that it is not the sequence itself or at least does not overlap with the information for calling events. Grey areas are for example copy numbers. The phenotype, resp. the variable to associate with can only be one value per individual; in the case of more variables of interest, as for example in gene expression, there is a GWAS for each variable.

In the simplest case, the phenotype is divided into 2 classes, the events are given binary (= not more than two alleles), and the GWAS is the result of one statistical test per event to phenotype combination. As the numbers of the combinations are contingency tables, the test can be Fisher's exact test or chi-square test, as examples. The main assumption, which is violated in this simple approach, is that the variables (= the events) are independent. This was already mentioned in one of the first GWAS papers [153], with 'One criticism of case-control association studies such as ours is that population stratification can result in false-positive results.' Today the most common way to deal with population structure are mixed models [154, 155]. For overviews of GWAS, reviews and recommendations see [156-159].

Extending beyond the simple case, mixed models can deal with numerical phenotypes but most implementations cannot deal with more than two alleles per event.

In general, GWAS remains an open field for at least two reasons. Firstly, the problem is a prime example of p>>n, i.e. the number of variables is much larger than the number of samples, which makes this problem ill-posed and frequently unstable. Secondly, GWAS is often used by biologists in this manner: the GWAS is made, the results are plotted and the peaks or the significant p-values are taken. Genes which seem interesting are looked for in the extended region of the peaks. These genes are then reported for the phenotype. Since these peaks are often quite broad, the genes considered interesting are decided by the GO-terms and these genes are also accepted if they are somehow nearby; it is thus difficult to estimate the significance of the finding as there are many interesting genes around. It is rather the exception that a GWAS finds precisely the causal variants and not much more [156].

8.4 Article: **TE-Locate**

Platzer, A., V. Nizhynska, and Q. Long, *TE-Locate: A Tool to Locate and Group Transposable Element Occurrences Using Paired-End Next-Generation Sequencing Data*. **Biology (Basel), 2012. 1(2): p. 395-410.**

The tool for this manuscript was made in the course of the research presented in section 6.6. As no other tool was available for read-pairs and because we were invited to submit to a special Issue "Next Generation Sequencing Approaches in Biology", we extended this method into an own paper. The paper can be categorized as a methods paper, according to [160].

OWN CONTRIBUTION IN [125]

AP and QL designed and implemented the tool, the analyses and wrote the paper. VN generated the sequencing data.

Biology **2012**, *1*, 395-410; doi:10.3390/biology1020395

OPEN ACCESS

biology

ISSN 2079-7737

www.mdpi.com/journal/biology

Article

TE-Locate: A Tool to Locate and Group Transposable Element Occurrences Using Paired-End Next-Generation Sequencing Data

Alexander Platzer, Viktoria Nizhynska and Quan Long *

Gregor Mendel Institute (GMI), Dr. Bohr-Gasse 3, 1030 Vienna, Austria;
E-Mails: alexander.platzer@gmi.oeaw.ac.at (A.P.); viktoria.nizhynska@gmi.oeaw.ac.at (V.N.)

* Author to whom correspondence should be addressed; E-Mail: quan.long@gmi.oeaw.ac.at;
Tel.: +43-1-79044-9904; Fax: +43-1-79044-9001.

Received: 27 July 2012; in revised form: 22 August 2012 / Accepted: 31 August 2012 /
Published: 12 September 2012

Abstract: Transposable elements (TEs) are common mobile DNA elements present in nearly all genomes. Since the movement of TEs within a genome can sometimes have phenotypic consequences, an accurate report of TE actions is desirable. To this end, we developed TE-Locate, a computational tool that uses paired-end reads to identify the novel locations of known TEs. TE-Locate can utilize either a database of TE sequences, or annotated TEs within the reference sequence of interest. This makes TE-Locate useful in the search for any mobile sequence, including retrotransposed gene copies. One major concern is to act on the correct hierarchy level, thereby avoiding an incorrect calling of a single insertion as multiple events of TEs with high sequence similarity. We used the (super)family level, but TE-Locate can also use any other level, right down to the individual transposable element. As an example of analysis with TE-Locate, we used the Swedish population in the 1,001 Arabidopsis genomes project, and presented the biological insights gained from the novel TEs, inducing the association between different TE superfamilies. The program is freely available, and the URL is provided in the end of the paper.

Keywords: transposable element; NGS data; calling TEs; paired-end reads; structural variation discovery; GWAS

1. Introduction

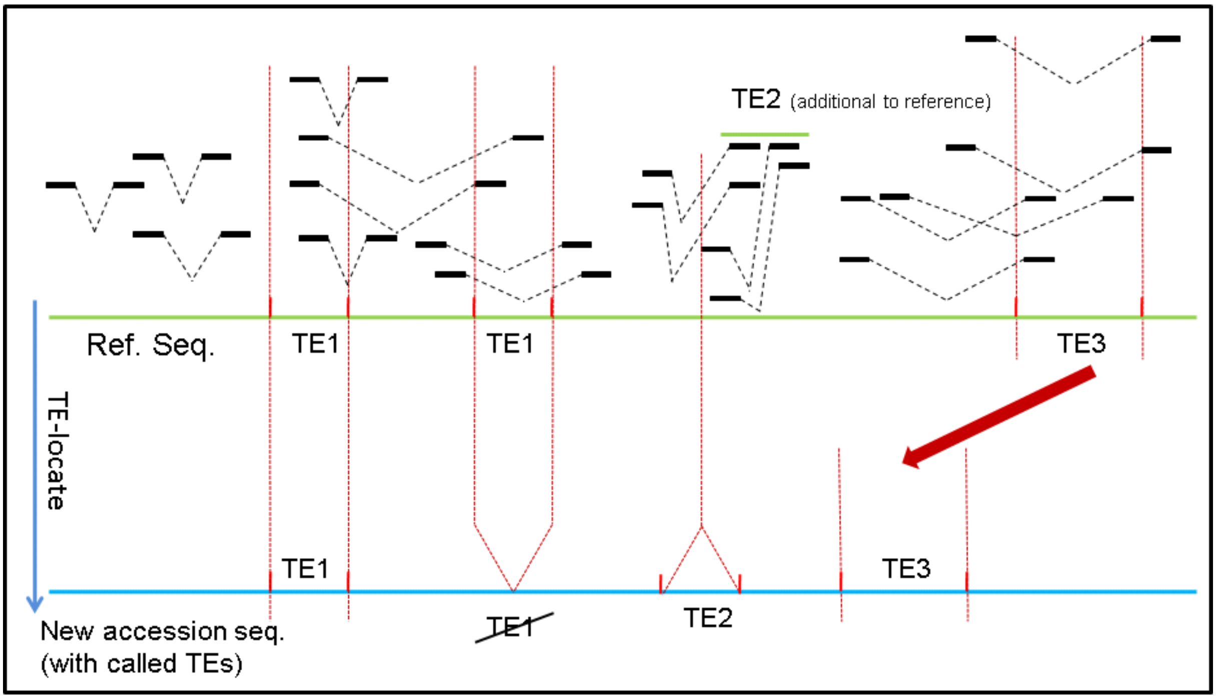
Transposable elements (TEs) have made themselves a great career, from being junk DNA [1] when first discovered [2], to having important roles in development [3], evolution [4,5], and disease [6] through direct genome rejoining [7], epigenetic control [8,9], or other known [10] or to-be-tested mechanisms [11].

The new quantity of next generation sequencing (NGS) data allows the discovery of structural variations (SVs) per individual and even intra-individual [12]. As TEs are an important source of SVs, their exact movements and copy number are of interest (e.g., studies [13–16]). One pitfall of TEs is their high sequence similarity, which causes alignment difficulties, especially for the short reads of most NGS platforms. This issue runs like a common thread beside the main method and analysis in this paper.

Given the difficulties of discovering TEs in general, we restricted ourselves to TEs with given sequences. Assuming the availability of a reference genome and the annotation of existing TEs in this reference genome, we developed TE-Locate, a computational tool that can call the newly-inserted copy of known TEs in sequenced individuals.

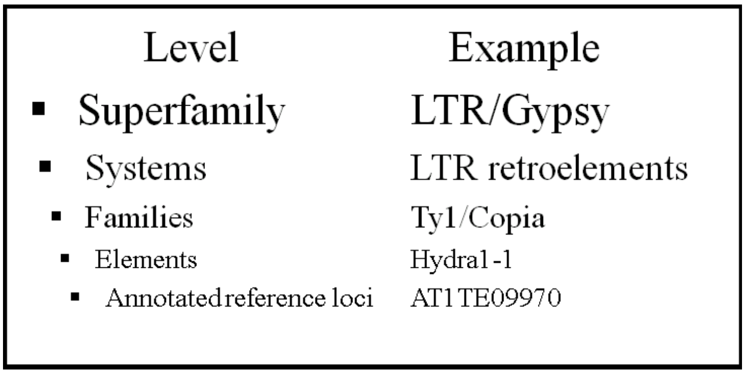
Two important insights into how TE-Locate functions should be noted. The first rationale underlying TE-Locate is the use of paired-end information. Although sequences of different TEs may be quite similar, the newly inserted regions should still somehow be divergent. Therefore, if a pair of reads is mapped across the breakpoint, we could observe one end of the mate-pair mapped onto the flanking sequences of the newly-inserted region with reasonably good quality, with the other end on the jumping TE (Figure 1).

Figure 1. How TE-Locate makes the callings with read pairs. In this scenario one element of TE1 has vanished from one locus (while the other is retained), one TE2 was inserted, and TE3 has moved to another nearby locus (*i.e.*, cut and paste).



However, although we can assume the read mapped to the flanking sequence of the new regions is uniquely mapped, we may ask if the read mapped to TE itself still suffers from repetitiveness. This would result in many different mistaken TE callings in the same spot due to their similarity in sequence content. In fact, this is true, and leads to the second insight underlying TE-Locate: although different TEs from a similar template may not be easily distinguishable, one can look at the level of difference within TE families or even superfamilies (Figure 2). For example, we may be able to conclude a new TE from a particular TE family that is inserted into a certain region, without specifying what exactly the TE gene is. The level of detailed information is thereby somewhat reduced, but a more reliable result is produced. In TE-Locate, we provide different levels of abstraction so that users can balance the trade-off between specificity and reliability.

Figure 2. TE hierarchies in The Gypsy Database (GyDB) of Mobile Genetic Elements.



Level	Example
▪ Superfamily	LTR/Gypsy
▪ Systems	LTR retroelements
▪ Families	Ty1/Copia
▪ Elements	Hydra1-1
▪ Annotated reference loci	AT1TE09970

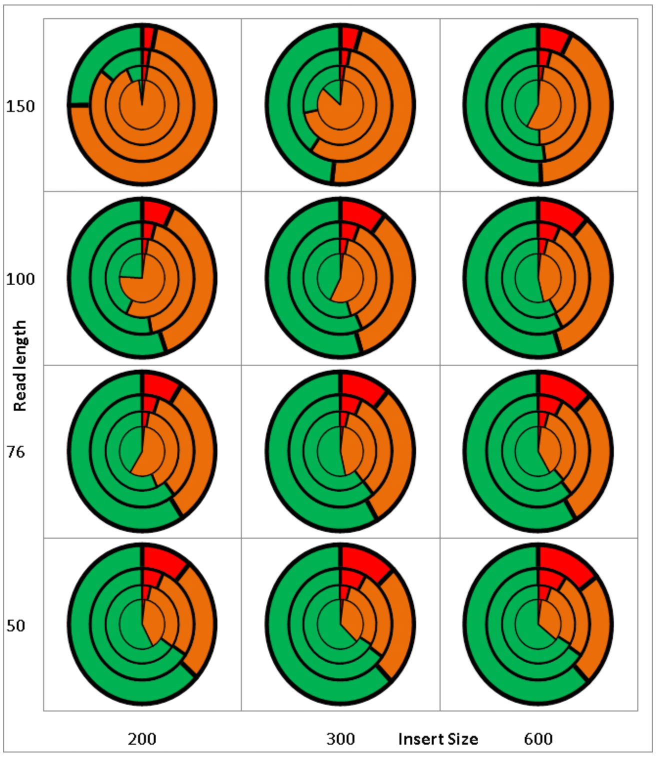
In addition to locating new copies of TEs, TE-Locate can also be used for calling insertions of other known sequences that are not TEs. In the general case, as long as a list of known to-be-likely inserted sequences is present as a template, TE-Locate can locate their new copies in the genome of the focal individual(s). A straightforward example is positioning the insertions of a virus to the host genome [17]; a less obvious application could be to chase the known ribosomal cluster sequences in the genome [18], which is what we are attempting using *Arabidopsis* data.

2. Results

2.1. Validation/Simulation

The outcome of TE-Locate is highly dependent on the aligner and the chosen hierarchy level (Figure 2). Nevertheless, we make an attempt at validation with simulated data. Firstly, a virtual reference genome is constructed starting from the *Arabidopsis thaliana* reference and its TE annotation [19]: the annotated TE regions are extracted and taken as additional sequences beside the (TE-free) chromosomes. This new reference is used later for analysis. For generation of the samples, the TE sequences are inserted back into the (TE-free) reference chromosomes, but at random locations. 500,000 SNPs (Single Nucleotide Polymorphism) (=0.4% of the whole genome) are mutated in this virtual individual genome. Based on that artificial sample, read pairs are generated with wgsim (part of Samtools [20]) for all combinations of coverages of 2×, 5×, 10× and 20×, insert sizes of 200, 300 and 600 bp (±100 bp standard deviation), and read lengths of 50, 76, 100 and 150 bp. The parameters for the real population data [21,22] which we later used for demonstrating analyses (insert size = 300 bp, read length = 76/100 bp, #SNP = 494,000, coverage = 20×) fit well to the simulations. The generated read pairs of the virtual individual genome are then aligned with BWA [23] to the virtual reference genome. The results with respect to error rates of TE-Locate with this data are shown in Figure 3. We choose superfamily as the hierarchic level. The calls are counted as correct if the right superfamily is called within 3-fold of the standard deviation of the read pair's insert size. The results are divided into chromosomal arms and pericentromeric regions (there are nearly no calls in the centromeres). Only the arms regions are depicted in Figure 3; the other diagram for pericentromeric regions, which shows slightly higher error rates, is the Supplementary Figure S1. One can see several trends in Figure 3: the False Positives (FP) decrease and the False Negatives (FN) increase with higher read lengths. This is expected, since very small TEs are missing when the read length decreases, at least with our chosen aligner. An efficient aligner that is able to deal with split reads would be helpful. There is an opposite effect with larger insert sizes and higher coverage (if the thresholds of calling the variants are fixed for any coverage). We also tried the same simulated data with BreakDancer [24], and depicted results in the Supplementary Figures S2 and S3. TE-Locate clearly outperforms BreakDancer at calling TEs. However, we do acknowledge that TE-Locate leverages TE annotations and uses hierarchy levels that general SV tools such as BreakDancer do not.

Figure 3. Results of TE-Locate with a virtual genome with known TEs. The X-axis denotes different insert sizes; the Y-axis denotes different read length; the concentric circles denote different coverage: from inner to outer circles, the coverages are 2×, 5×, 10× and 20× respectively. The red, orange, and green colors denote the proportion of false positives, false negatives and the rest. Here the false positive is defined as the ratio between false calls and all calls, the false negative is defined as the ratio between missing calls and all TEs inserted.

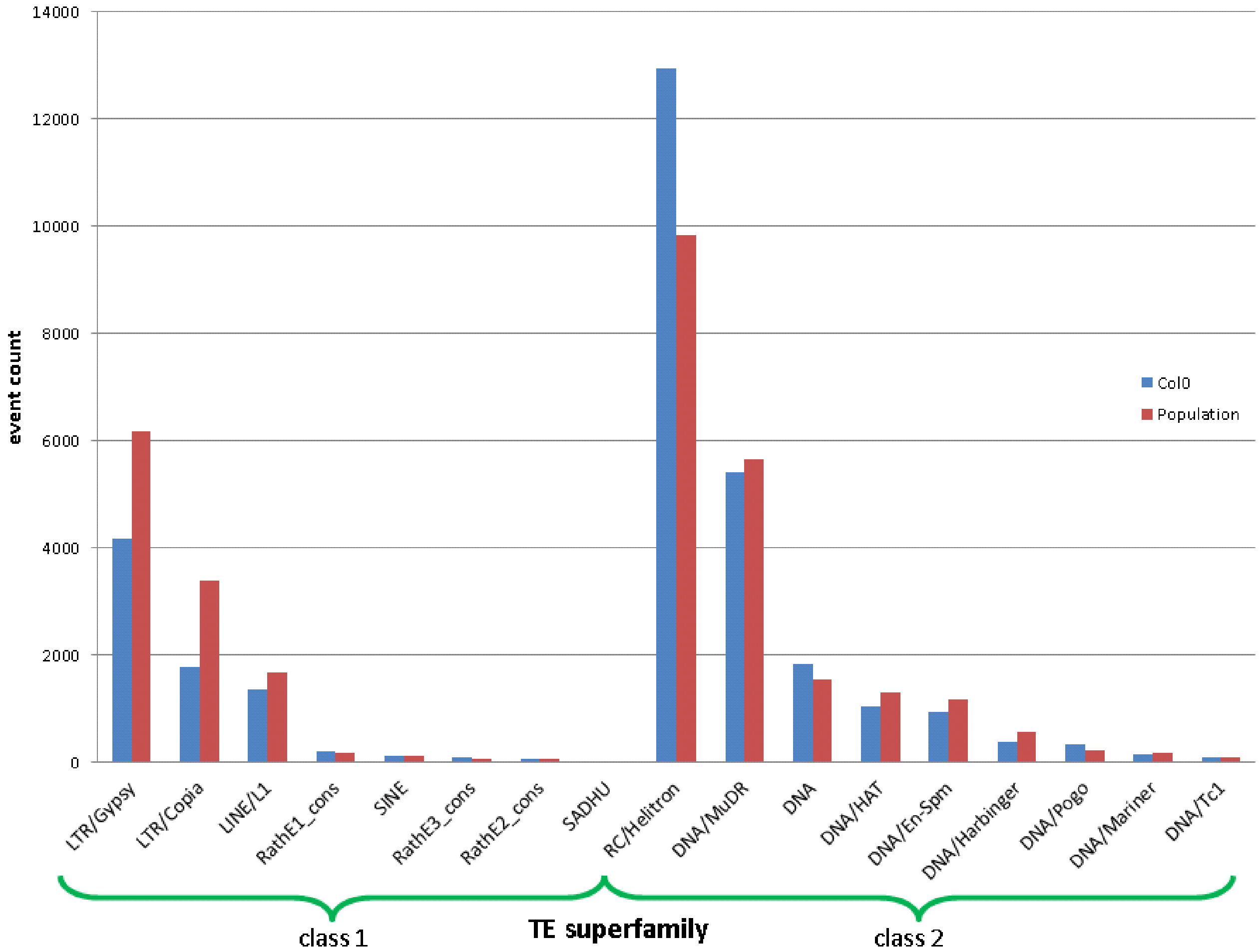


2.2. Real Data

To demonstrate the tool and some subsequent analysis, we applied it to NGS data of ~200 Swedish *Arabidopsis thaliana* lines sequenced in our group [25], which is part of the 1,001 genomes project [21,22]. The terms 'population', 'individuals', and 'real data' later in the text refer to this source.

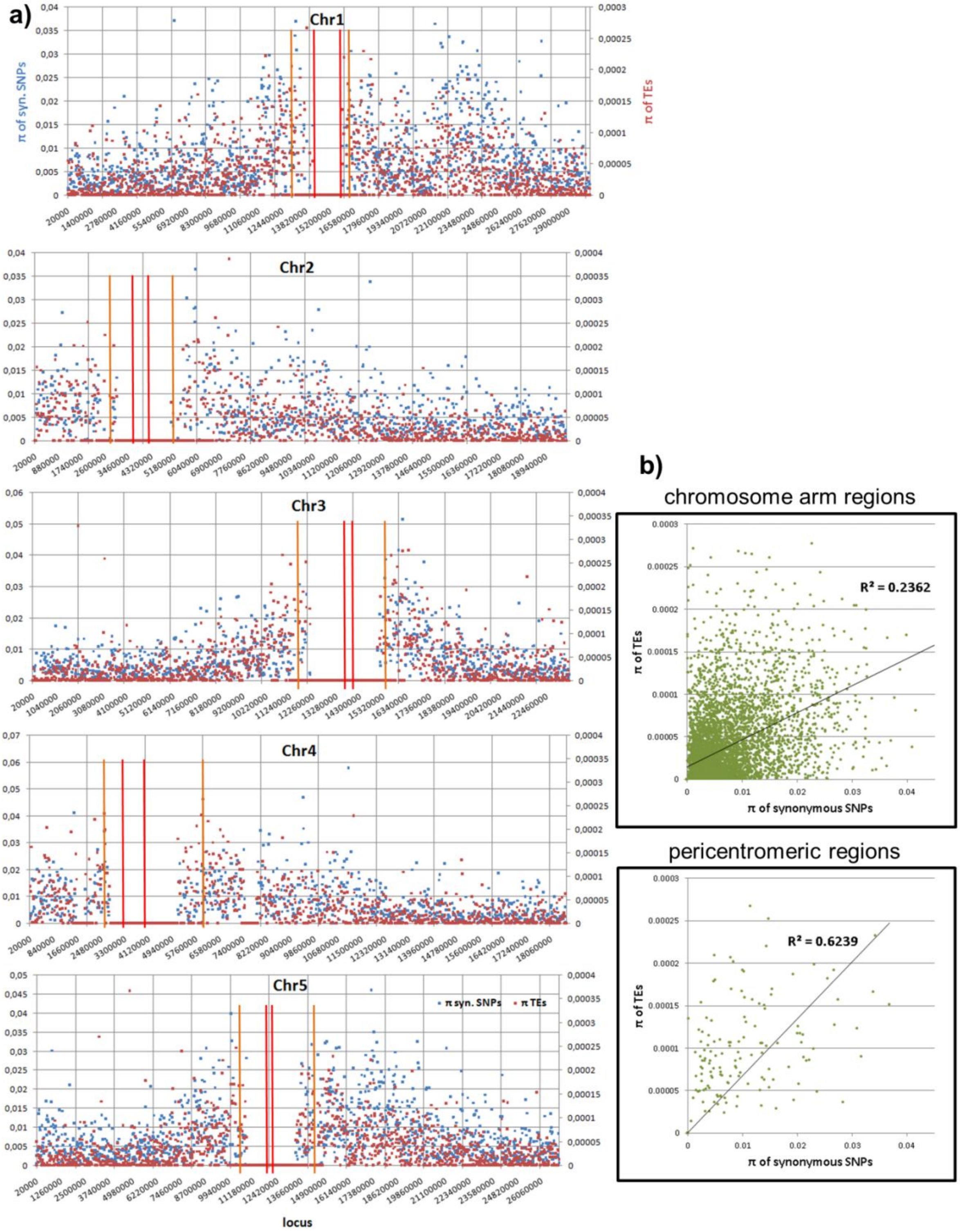
In total, we called about 40,000 TEs in the population on the superfamily level (on other hierarchical levels, it called other quantities of events). By contrasting the number of TE events called and that are annotated in the reference, we see a clear difference between Class I and II ("copy-paste" and "cut-paste") TEs (see Figure 4).

Figure 4. The event counts per TE superfamily annotated in the reference (blue) and newly discovered from the population. An event for the population is counted if it occurs in any individual. Class I and II TEs ("copy-paste" and "cut-paste") are depicted separately.



For comparative purposes, the distribution of polymorphism in terms of pair-wise difference, π, is shown in Figure 5 for TEs and for SNPs. We found that the polymorphism of SNPs is correlated to the density of new TEs (Figure 5b) in both chromosomal arms and pericentromeric regions, which might indicate an interesting mutation or selection mechanism, if not simply an effect of a deeper coalescence time.

Figure 5. Distribution of polymorphism in terms of pair-wise difference π (in terms of the number of events without being weighted by the lengths) of the TE calls in the population against π of SNPs. Both π are computed with a window size of 20 Kb and normalized to 1 bp. (**a**) The π distribution in the chromosomes. We use red and orange bars to indicate the centeromeric and pericentromeric regions. (**b**) The correlation between TE and SNP π's in both chromosomal arms and pericentromeric regions. If there is not even a single event in one of both windows (TE or SNP), this locus is skipped. Both correlations are highly significant (*p*-value = 0 due to machine precision).



We also looked for the distribution of the copy numbers to the geographic location. The sequenced samples were divided up between the north and south of Sweden (Figure 6). The question here is whether this classification could be replicated by observing the TE variations. Based on TE-Locate results, we tried several machine learning techniques (with Weka [26]). On the superfamily level there was no result better than chance at 10× cross-fold validation. On the TE-family level, there are good classifications with a true prediction rate of 92%–98% and a lower limit *i.e.*, zero ratio of 71% (zero ratio = the ratio of the more frequent class). The result of the C4.5 algorithm [27] is shown in Figure 7. With respect to the true prediction rate, this is not the best model, but trees are easier to interpret than, for example, the weights of SVMs (Support Vector Machine) [28]. As one can see in this tree, although all TE-families were used as variables, only *Copia* families are enough to sufficiently split the classes. We did not go into detail on why the copy numbers of *Copia* families are clearly different between north and south; the simplest explanation could be merely a temperature dependency in them (see the related, but not so recent [29,30]).

Figure 6. The geographic distribution of the *Arabidopsis thaliana* lines used for our analysis. The red line indicates the border between the later-used north and the south class.

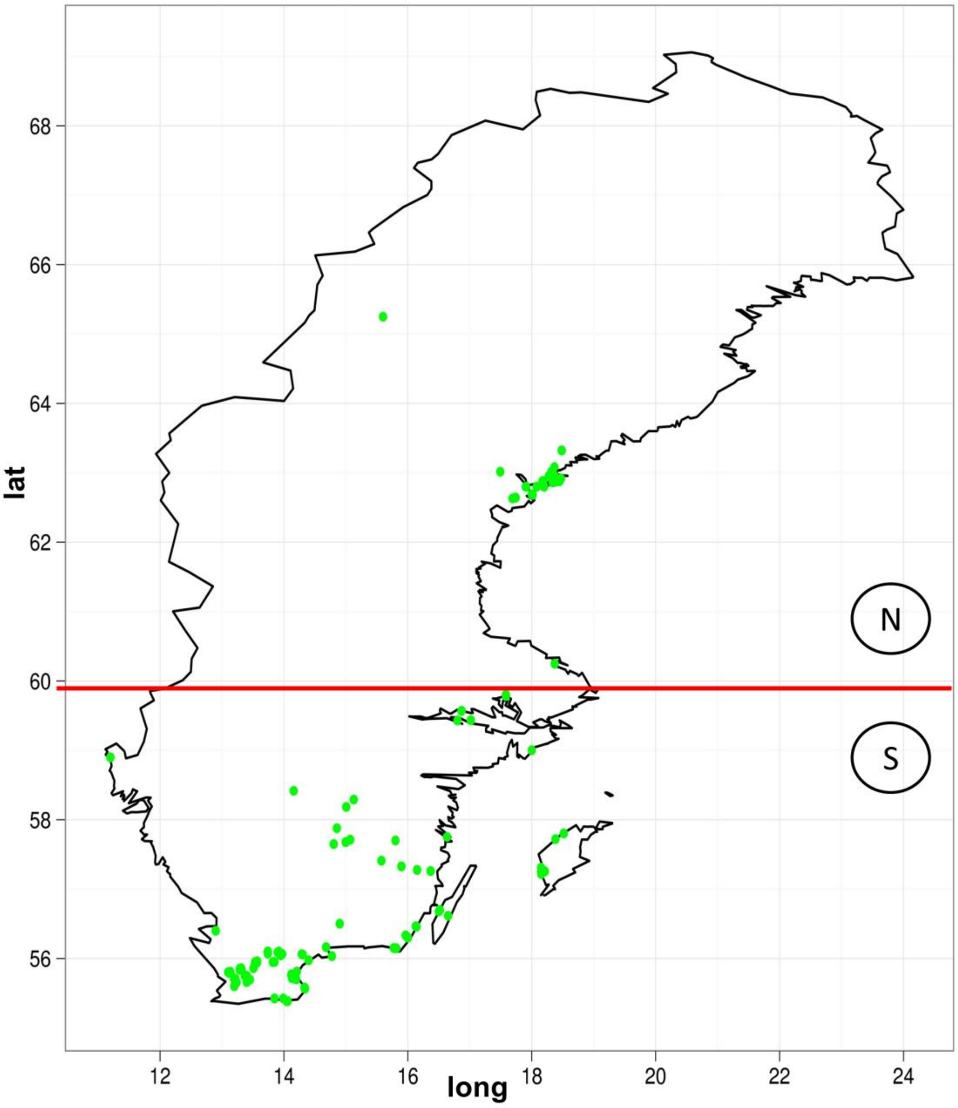
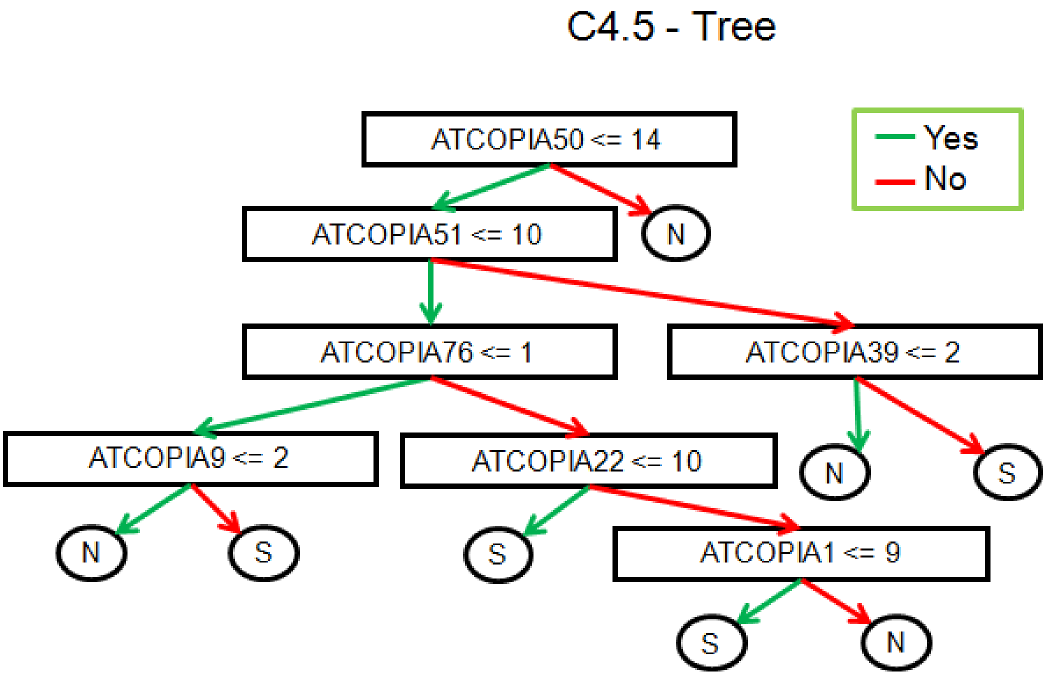


Figure 7. Result of the C4.5 algorithm for a classification of north *versus* south individuals with respect to their TE-family copy number. 92.5% of the individuals are correctly classified at the 10× cross-validation.



We performed genome-wide association studies (GWAS) using the 4 million SNPs from the sequences as genotype, and each of the 18 TE superfamilies copy number as phenotype. The question for this analysis is, how much of the variation in TE copy numbers could be explained by the genotype. We used a mixed model [31] to control population structure and Bonferroni correction to control an inflated significance level due to multiple-test issues. Two of these GWAS with many significant SNPs are shown in Figure 8. As expected, there are many significant SNPs located in TEs themselves and unfortunately nearly none in (well-annotated) genes. An exception is one significant SNP in the auxin response factor-12 gene (AT1G34310) for the copy number of RathE3.

It is remarkable that most of the significant SNPs for a superfamily are located in another superfamily. It is not clear whether this could be a problem of a too-high similarity between the superfamilies or a non-optimal separation. However, if one of these issues is causing the effect, we should have observed a symmetrical relationship between the pair of superfamilies: if SNPs associated with superfamily A are located in superfamily B, then we should also observe SNPs associated with superfamily B located in superfamily A. However, what we observed is an asymmetric hierarchy (Figure 9): it is never the case that if one superfamily has significant SNPs in another, that this is also present in the reverse case. It would be interesting to investigate the biology of this observation.

Figure 8. Manhattan plot of logged *p*-values of association between the SNPs and the TE copy number. The chromosomes are sequential in different colors. The upper plot uses the DNA TE-superfamily as phenotype, the lower the TE-superfamily SADHU. The Bonferroni threshold is 2.5×10^{-7}.

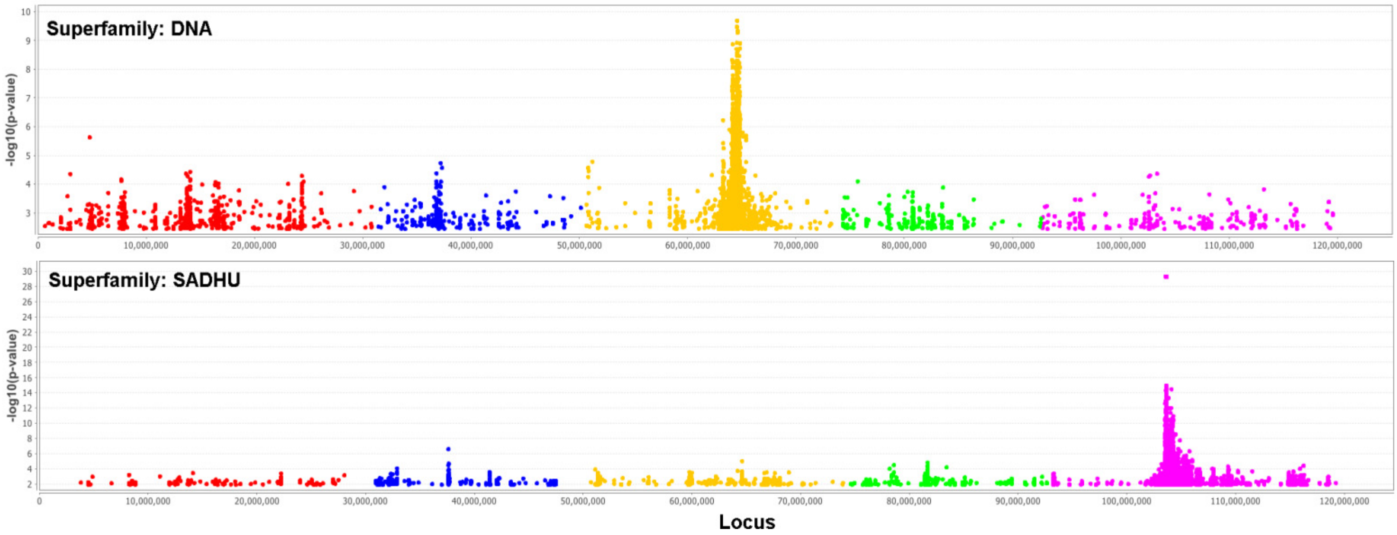
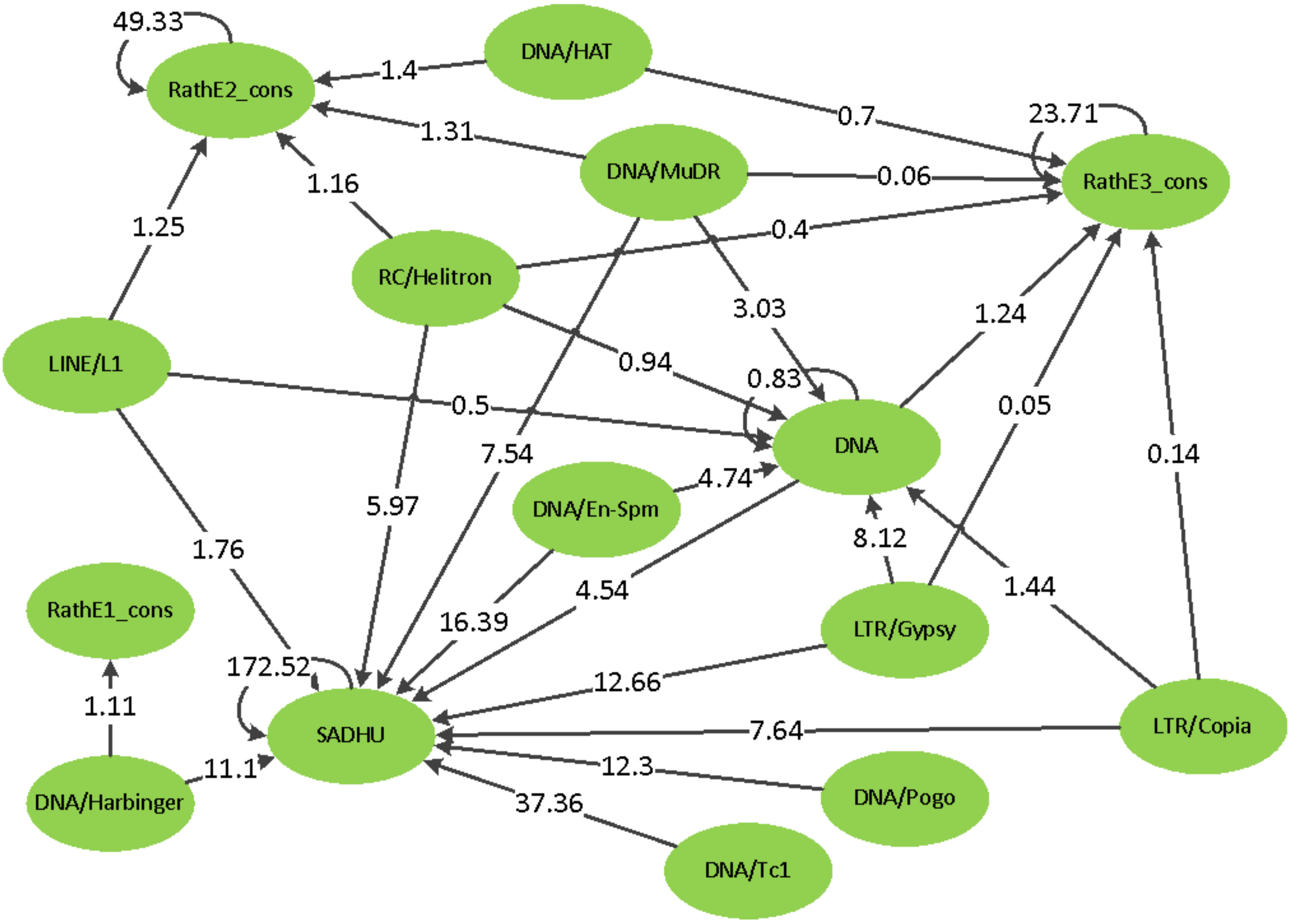


Figure 9. The SNP to copy number hierarchy from GWAS. The arrows indicate that the SNPs located in the superfamily on the blunt side of the arrow are significantly associated with the copy number of the superfamily on the side of the arrowhead. The number within the arrow is the number of SNPs normalized by the total length of TEs in the corresponding superfamily. There were no cases of arrows traveling in both directions.



3. Methods

TE-Locate assumes that the user has paired-end reads. Before running TE-Locate, the read pairs are aligned with any aligner producing a BAM/SAM file (e.g., BWA [23], Smalt [32], or Segemehl [33]). With the previously prepared annotation, TE-Locate calls the TE as shown in Figure 1. TE-Locate will identify and collect all mate-pairs that have one end mapped inside a TE and the other end mapped with good quality to any region outside all TEs. By clustering all the evidential reads, the new copy of TE will then be reported. To leverage the population sharing that is crucial for structural variant callings [34], the tool is written to act on all individuals in the population at once. In this manner, individuals with very low coverage at a particular region can take advantage of other individuals when there is a genuine event also called by other good coverage individuals.

The results are reported in two files: one is a CSV file in which the have-or-have-not information for all individuals and all events is provided. In a separate information file, TE-Locate also provides a summary of more detailed event information (features of the TE, the number of supporting reads, *etc.*) An example output is shown in Table 1; the columns are explained in detail in Table 2.

Table 1. Example output of TE-Locate.

chr	loc	len	event_type_ref	non_ref_counts	anc_status	read_pair_support	<unused>...	call_method	Orientation	#pPairs	#iPairs	new/old
1	5421	7679	TE+DNA/MuDR/DNA/MuDR	5	N	15	|	PairEndTE	inverse	4	11	new
1	16726	3890	TE+RC/Helitron/RC/Helitron	171	N	900	|	PairEndTE	uncertain			old
1	20843	1292	TE+RC/Helitron/RC/Helitron	3	N	63	|	PairEndTE	inverse	20	43	new
1	11897	79	TE+LTR/Copia/LTR/Copia	55	N	69	|	PairEndTE	uncertain			old
1	22277	1736	TE+DNA/MuDR/DNA/MuDR	7	N	15	|	PairEndTE	inverse	6	9	new
1	42355	10046	TE+RC/Helitron/RC/Helitron	4	N	11	|	PairEndTE	parallel	10	1	new
1	42210	4671	TE+DNA/MuDR/DNA/MuDR	5	N	11	|	PairEndTE	inverse	1	10	new
1	50968	651	TE+LTR/Gypsy/LTR/Gypsy	6	N	10	|	PairEndTE	parallel	9	1	new
1	52425	382	TE+LTR/Copia/LTR/Copia	2	N	26	|	PairEndTE	inverse	1	25	new
1	70064	4814	TE+LTR/Copia/LTR/Copia	1	N	19	|	PairEndTE	inverse	0	19	new
1	71152	799	TE+LTR/Copia/LTR/Copia	1	N	31	|	PairEndTE	parallel	31	0	new
1	55676	900	TE+DNA/HAT/DNA/HAT	174	N	2133	|	PairEndTE	uncertain			old
1	77569	831	TE+RC/Helitron/RC/Helitron	178	N	1661	|	PairEndTE	uncertain			old
1	76844	656	TE+LINE/L1/LINE/L1	75	N	753	|	PairEndTE	uncertain			old
1	84679	12225	TE+LTR/Gypsy/LTR/Gypsy	7	N	12	|	PairEndTE	parallel	10	2	new
1	91443	7263	TE+LTR/Gypsy/LTR/Gypsy	6	N	13	|	PairEndTE	parallel	11	2	new
1	116237	2941	TE+LTR/Copia/LTR/Copia	1	N	57	|	PairEndTE	parallel	47	10	new
1	129878	5185	TE+LTR/Copia/LTR/Copia	4	N	23	|	PairEndTE	parallel	23	0	new
1	154331	87	TE+LINE/L1/LINE/L1	89	N	138	|	PairEndTE	uncertain			old
1	192934	593	TE+RC/Helitron/RC/Helitron	177	N	1915	|	PairEndTE	uncertain			old

Table 2. Description of the TE-Locate output.

Column	Description
chr	Locus
loc	
len	The length of the corresponding reference event.
event_type_ref	The class of this event annotated (resp. the item/TE)
non_ref_counts	The number of individuals sharing this event.
anc_status	Unused
read_pair_support	The total number of all supporting read pairs of all individuals.
bp_range1	Unused...
bp_range2	
four_gamete_left	
four_gamete_right	
call_method	For TE-Locate, here is written 'PairEndTE', used if merged with other data in this format.
Orientation	'parallel', 'inverse' or 'uncertain': The orientation according to the reference sequence.
#pPairs	The number of read pairs supporting parallel orientation. Not used if the orientation is 'uncertain'.
#iPairs	The number of read pairs supporting inverse orientation. Not used if the orientation is 'uncertain'.
new/old	'new' or 'old'. 'old' if the item is called at the locus in the reference; 'new' otherwise. Note that at higher hierarchical levels, all locations of this item are meant, e.g., any Copia called at a Copia locus in the reference is called 'old' as the item's name is the only distinction.

In the real data analysis presented in this paper, the reference sequence and the TE annotations are taken from TAIR [19] in .fasta and .gff formats respectively. The *Arabidopsis thaliana* lines are sequenced by Illumina GAII as well as by HiSeq 2000 with paired-end reads 2 × 76 bp or 2 × 100 bp. The coverage ranges from 10× to 70×. More details of the dataset will be published soon and can be downloaded from the 1,001 genomes project public website [22].

The hierarchical levels of TE families are from the Gypsy Database—GyDB [35] (Figure 2). The hierarchical level should be high enough to ensure that no very similar sequences are present at different items, but low enough to have a good resolution. Most of the demonstration analysis uses the superfamily and family level.

4. Discussion and Conclusions

TE-Locate is a flexible tool to call known sequences of a reference in new individuals. This is particularly interesting for TEs. The theoretical computational complexity is $O(n*log(n))$, where n is the number of reads. In practice, we observed that the implementation is sufficiently efficient, at least for our deeply-sequenced *Arabidopsis* lines. In our real data, TE-Locate needed much less computational time than the initial preprocessing of the data (mapping reads, *etc.*). Although the implementation is not parallelized, no GPGPU (General-purpose computing on graphics processing units) is used and the code is written in Perl and Java.

The current initial release of TE-Locate runs fast and its algorithm is rather straightforward. Many extensions are possible. One immediate extension is to include indel callings from various sources, perhaps also combined with graphs from *de-novo* assembly. We could also count negative support (=contradicting read pairs) and evaluate the optimal set in contradictory cases. Finally, it may be beneficial to combine with split read alignments [36] and/or develop an efficient aligner for this [37].

Not all the possible extensions will necessarily have a positive effect, at least if the thresholds for trade-offs are not chosen carefully. An example would be the trade-off between negative and positive support and the weight of split-reads against read pairs. The computational complexity will likely increase, especially if it is to find an optimal set or combination.

TE-Locate is a nice complement to other tools [38] for a similar purpose. T-lex [39] uses single split reads and only checks whether the reference loci are present or not; REPET [40], RECON [41], and TESeeker [42] call new TE sequences without leveraging existing annotations; TE-HMM [43] analyzes genomes itself to discover TEs without using read-level information. Also, all above-mentioned tools do not take advantage of paired-end information, which is not ideal for most ongoing NGS projects in which the paired-end reads will be generated. Various indel calling tools [44] are also beneficial to TE analysis, since TEs can also be considered merely as ordinary indels. The program is freely available online [45].

Acknowledgments

We are grateful to Ashley Farlow and Magnus Nordborg for discussions, and Thomas Friese for proof-reading. This work is supported by the Austrian Academy of Sciences.

References and Notes

1. Castillo-Davis, C.I. The evolution of noncoding DNA: How much junk, how much func? *Trends Genet.* **2005**, *21*, 533–536.
2. McClintock, B. *The Discovery and Characterization of Transposable Elements: The Collected Papers of Barbara McClintock*; Garland Publishing, Inc.: New York, NY, USA, 1987.
3. Nowacki, M.; Higgins, B.P.; Maquilan, G.M.; Swart, E.C.; Doak, T.G.; Landweber, L.F. A functional role for transposases in a large eukaryotic genome. *Science* **2009**, *324*, 935–938.
4. Tenaillon, M.I.; Hollister, J.D.; Gaut, B.S. A triptych of the evolution of plant transposable elements. *Trends Plant Sci.* **2010**, *15*, 471–478.
5. Hollister, J.D.; Gaut, B.S. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* **2009**, *19*, 1419–1428.
6. Kazazian, H.H., Jr. Mobile elements and disease. *Curr. Opin. Genet. Dev.* **1998**, *8*, 343–350.
7. Kazazian, H.H., Jr. Mobile elements: Drivers of genome evolution. *Science* **2004**, *303*, 1626–1632.

8. Bourque, G.; Leong, B.; Vega, V.B.; Chen, X.; Lee, Y.L.; Srinivasan, K.G.; Chew, J.L.; Ruan, Y.; Wei, C.L.; Ng, H.H.; *et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **2008**, *18*, 1752–1762.
9. Lippman, Z.; Gendrel, A.V.; Black, M.; Vaughn, M.W.; Dedhia, N.; McCombie, W.R.; Lavine, K.; Mittal, V.; May, B.; Kasschau, K.D.; *et al.* Role of transposable elements in heterochromatin and epigenetic control. *Nature* **2004**, *430*, 471–476.
10. Cordaux, R.; Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **2009**, *10*, 691–703.
11. Belancio, V.P.; Hedges, D.J.; Deininger, P. Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Res.* **2008**, *18*, 343–358.
12. Gottlieb, B.; Beitel, L.K.; Alvarado, C.; Trifiro, M.A. Selection and mutation in the "new" genetics: An emerging hypothesis. *Hum. Genet.* **2010**, *127*, 491–501.
13. Gupta, S.; Gallavotti, A.; Stryker, G.A.; Schmidt, R.J.; Lal, S.K. A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes. *Plant Mol. Biol.* **2005**, *57*, 115–127.
14. Jiang, N.; Bao, Z.; Zhang, X.; Eddy, S.R.; Wessler, S.R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **2004**, *431*, 569–573.
15. Kordis, D. Transposable elements in reptilian and avian (sauropsida) genomes. *Cytogenet. Genome Res.* **2009**, *127*, 94–111.
16. Lai, J.; Li, Y.; Messing, J.; Dooner, H.K. Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 9068–9073.
17. Schroder, A.R.; Shinn, P.; Chen, H.; Berry, C.; Ecker, J.R.; Bushman, F. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **2002**, *110*, 521–529.
18. Conconi, A.; Sogo, J.M.; Ryan, C.A. Ribosomal gene clusters are uniquely proportioned between open and closed chromatin structures in both tomato leaf cells and exponentially growing suspension cultures. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 5256–5260.
19. Lamesch, P.; Dreher, K.; Swarbreck, D.; Sasidharan, R.; Reiser, L.; Huala, E. Using the Arabidopsis information resource (TAIR) to find information about Arabidopsis genes. *Curr. Protoc. Bioinformatics* **2010**, *Chapter 1*, Unit1 11.
20. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079.
21. Weigel, D.; Mott, R. The 1001 genomes project for Arabidopsis thaliana. *Genome Biol.* **2009**, *10*, 107.
22. The 1001 Genomes Project Website. Available online: http://www.1001genomes.org (accessed on 1 July 2012).
23. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **2010**, *26*, 589–595.
24. Chen, K.; Wallis, J.W.; McLellan, M.D.; Larson, D.E.; Kalicki, J.M.; Pohl, C.S.; McGrath, S.D.; Wendl, M.C.; Zhang, Q.; Locke, D.P.; *et al.* BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **2009**, *6*, 677–681.

25. Long, Q.; Rabanal, F.A.; Meng, D.; Huber, C.D.; Farlow, A.; Platzer, A.; Zhang, Q.; Vilhjálmsson, B.J.; Korte, A.; Nizhynska, V.; *et al.* Massive genomic variation and strong selection in Swedish Arabidopsis thaliana. Gregor Mendel Institute, Vienna, Austria. Unpublished work, 2012.
26. Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I.H. Data mining in bioinformatics using Weka. *Bioinformatics* **2004**, *20*, 2479–2481.
27. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: Burlington, MA, USA, 1993.
28. Platt, J.C. A fast algorithm for training support vector machines. Technical Report for Microsoft Research, Redmond, WA, USA, 21 April 1998. MSR-TR-98-14.
29. Turner, A.K.; Delacruz, F.; Grinsted, J. Temperature sensitivity of transposition of class-Ii transposons. *J. Gen. Microbiol.* **1990**, *136*, 65–67.
30. Paquin, C.E.; Williamson, V.M. Temperature effects on the rate of ty transposition. *Science* **1984**, *226*, 53–55.
31. Kang, H.M.; Sul, J.H.; Service, S.K.; Zaitlen, N.A.; Kong, S.Y.; Freimer, N.B.; Sabatti, C.; Eskin, E. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **2010**, *42*, 348–354.
32. Ponstingl, H. *SMALT*; Wellcome Trust Sanger Institute: Cambridge, UK, 2011.
33. Hoffmann, S.; Otto, C.; Kurtz, S.; Sharma, C.M.; Khaitovich, P.; Vogel, J.; Stadler, P.F.; Hackermuller, J. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.* **2009**, *5*, e1000502.
34. Handsaker, R.E.; Korn, J.M.; Nemesh, J.; McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **2011**, *43*, 269–276.
35. Llorens, C.; Futami, R.; Covelli, L.; Dominguez-Escriba, L.; Viu, J.M.; Tamarit, D.; Aguilar-Rodriguez, J.; Vicente-Ripolles, M.; Fuster, G.; Bernet, G.P.; *et al.* The Gypsy Database (GyDB) of mobile genetic elements: Release 2.0. *Nucleic Acids Res.* **2011**, *39*, D70–D74.
36. Ye, K.; Schulz, M.H.; Long, Q.; Apweiler, R.; Ning, Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **2009**, *25*, 2865–2871.
37. Abyzov, A.; Gerstein, M. AGE: Defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* **2011**, *27*, 595–603.
38. Bergman, C.M.; Quesneville, H. Discovering and detecting transposable elements in genome sequences. *Brief Bioinform.* **2007**, *8*, 382–392.
39. Fiston-Lavier, A.S.; Carrigan, M.; Petrov, D.A.; Gonzalez, J. T-lex: A program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Res.* **2011**, *39*, e36.
40. Flutre, T.; Inizan, O.; Hoede, C.; Quesneville, H. REPET: Pipelines for the identification and annotation of transposable elements in genomic sequences. In Proceedings of the Plant & Animal Genome (PAG) XVIII Conference, San Diego, CA, USA, 9–13 January 2010.
41. Bao, Z.; Eddy, S.R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **2002**, *12*, 1269–1276.

42. Kennedy, R.C.; Unger, M.F.; Christley, S.; Collins, F.H.; Madey, G.R. An automated homology-based approach for identifying transposable elements. *BMC Bioinformatics* **2011**, *12*, 130.
43. Andrieu, O.; Fiston, A.S.; Anxolabehere, D.; Quesneville, H. Detection of transposable elements by their compositional bias. *BMC Bioinformatics* **2004**, *5*, doi:10.1186/1471-2105-5-94.
44. Medvedev, P.; Stanciu, M.; Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **2009**, *6*, S13–S20.
45. TE-Locate Website. Available online: http://zendto.gmi.oeaw.ac.at/pickup.php?claimID=Y3tZVfN5xipYyBDN&claimPasscode=NArXMbTjmkorWjSM&emailAddr=te_locate%40gmx.at (accessed on 1 July 2012, will be long-term maintained).

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).

8.5 Results – Discussion

One conclusion for the method in the manuscript of the last section is that the method is mainly dependent on the aligner. If the alignment is perfect, the TE-locate calls are perfect. The validation/simulation is therefore more a test of the aligner than of the calling method.

The tool itself was programmed at the time when the paper of section 6.6 was written, which is why the data for demonstration therein is the data of the paper on the Swedish samples [36].

Because the tool's paper is an application paper, it contains several good starting points for analyses, but unfortunately, at least at our place, these are not followed, as functional validations would require a great deal of wet lab work.

9. Conclusions & future directions

9.1 …of biological big data in general

The conclusions of this thesis will be as broad as the sweep it has made of the field. From the point of view of data: the amount of data will continue to increase. Due to increasing masses of data, the ratio between those in the wet lab and those analyzing will change further. My rough guess would be an equilibrium of 1:1, as it is already in some labs.

The high-throughput data sources will increase in amount and number, and with new superlatives. However, it is difficult to guess the comparative to 'high-throughput' and 'next-generation DNA sequencing'. More and more different data sources make it to a high-throughput version, like bisulfite sequencing [161] and proteome data [162]. As these sources are not independent from each other and have likely more interplay than is currently known, combined analyses make sense; but they will provide new challenges, as missing data on different levels, different biases from different sources, different and larger p >> n and of course simply by the sheer amount of data. This also leads to the temptation to linger longer in explorative data analysis, as only few people can counterproof, and naturally more effort is put into new things than into validation.

When data is increasing more rapidly than it is being analyzed or even controlled in any sense, and for other reasons too, new ideas remain cheap in biology [163, 164]. The positive aspects of having many ideas are that it is easy to find topics without stepping on other's toes, that there is no danger that the field is running out of work, and that there is always enough material to produce a new paper. The negative aspects of an abundance of ideas are that the temptation exists to jump to the next topic before the present one is done or at least resulted in a paper, that ideas are discarded regardless of their quality, and that topics which are, or would be, the basis for other ideas are not addressed if they seem boring. The latter sounds rather unscientific, but often the temptation is too strong to make something very new, even if it is known (but hopefully not proven) that the underlying source has certain weaknesses which are not analyzed.

The organization of people in data generation and data analysis will remain tricky: these two areas have a quite different profile, so that almost none can be up-to-date, even in a specialized topic, in both fields. For years, a considerable effort has been spent on closing the gap between the wet lab and data analysis. This is done with far more statistics and bioinformatics for the 'normal' study of biology, especially in Vienna, where bioinformatics institutes are located in several universities related to biology instead of bioinformatics concentrated at one university. Certainly the gap is smaller now than during the race for the first human genome [165]; yet it is moving slowly compared with the possibilities, and there are several pitfalls:

- Who is leading? - If data generation and data analysis are in the same group, one side usually leads, which drives often to the following situations: if the data generating side leads, the data is supposed to prove or provide a certain result -> until the raw data becomes freely available and analyzed further, it might be highly biased; if the data analyzing side leads, they tend to make imaginary constructions.
- When data generation and analysis are done at the same place and there is no leading issue, it comes usually to more iterations; in this case two dangers are likely: the hand-over of data is not clean, and when a considerable effort has gone into data generation, the analysis is much less likely to judge stringently.

- The problems, tasks and questions fall under several different categories at the same time, but, as always in academia, every area, field and institute has its claim and usually avoids going into other areas. The following example areas are presented in this thesis, which almost never exist together in one group: very new wet lab methods, data analysis, algorithm development, implementation, and efficient programming. A group having more than two of these areas together usually splits or works quite inefficiently.
- When projects are larger, a professional project management is needed, but unfortunately, the level of project management in biology is quite low.
- Although the current masses of data did not arrive yesterday, it has only quite recently dawned on some minds that large amounts of data also need a certain effort in terms of IT-infrastructure, organization and manpower to deal with it.
- Outsourcing - What is sometimes effective in the economy might be effective for science: recently, a few more 'areas', services and tasks have been outsourced directly or indirectly to companies or service facilities. This certainly makes sense if a method has become more or less standard and simply needs to be scaled up or made more efficient, but there is also the temptation to outsource everything that is not in the main focus of the group: data analyzing groups order data and data generating groups try to 'buy' the analysis. The question then arises: if all the parts are somehow given away, what is then left for (academic) science?

From the economic perspective, molecular biology/genetics is still highly dependent on public spending. For certain reasons, there is no real 'killer application' [166] in this area. In a literal sense, killer application sounds alarming here, but in the defined sense it perhaps needs to grow in importance and become less dependent on public funding, as computer science is now.

9.2 …of the work presented in the thesis

Because of the increase of data, new methods and general progress, the content of this work corresponds rather to steps forward than to final answers to the big questions. The outlook and future directions for the papers presented in the thesis:

9.2.1 Article: Characterization of protein-interaction networks in tumors (section 3.3)

This paper is quite extensive and self-contained. The number and the implementation of the graph measures are comprehensive. More graph measures were implemented and tried than mentioned in the paper, but no results came from them. Of course, more methods can be implemented, as was done later, and it certainly makes sense to redo this analysis with new combinations of gene lists and protein interaction networks. Nevertheless, the idea itself - the use of graph measures to obtain properties of gene lists, where no simpler properties are known for these genes - is sold well.

9.2.2 Article: The Occurrence-in-subtuple problem (section 4.4)

This paper is even more closed than the previous paper, as it is exactly one full derivation for a specific problem. No other usage than the one described in the paper has arisen since the publication. One effect was that this method partly prevented another paper be written; showing that one putative regulatory network is not as solid as first thought.

9.2.3 Article: Metabolic profiling reveals key metabolic features of renal cell carcinoma (section 5.3)

Since the time of this paper, machine learning is today a little more present in medicine. Nevertheless, this paper can continue to act as a template for similar analysis in the future. The following changes can occur:

- if many more metabolites can be safely distinguished, it may fall into the p>>n class of problems and the type of methods have to change to something similar as for microarrays and/or full genome data
- if many more samples and more classes/labels of them arrive, these might be suitable for more sophisticated machine learning methods; there likely already exist methods in another scientific field designed for such data combinations
- if the number of samples grow extensively and are available as raw data (we intend human data here), it might be possible to solve the classification problems analytically

One related challenge here is the distribution of data as the largest source is from humans. This kind of data is mainly generated in hospitals, but is usually not shared freely. It is not trivial to find a trade-off between making all analyses possible and privacy.

9.2.4 Article: Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden (section 6.6)

The future directions of this scientific thread are quite clear, because this paper is one of the 1001 Genomes Project and this project is to date not completed. This paper is only about the Swedish subset, the full data of the project is being collected and at least one paper will be written for the full set. Recently generated related data sets also lurk around, for instance expression data at different temperatures, bisulfite sequencing data, additional phenotypes, and some others. All these data are more or less the prototype of big data piles waiting for a great deal of analyses time. Possible future directions: extensive GWASes, causality analysis, functional validation of findings, pattern search in several variations, focusing on subsets as for example TEs, certain genes, specific regulatory patterns, comparisons with lab generated data, more dimension reduction, using as standard data set for new methods, and so on. Unfortunately, ~more validation is also still on the list, which means validation is rather lagging behind.

9.2.5 Article: Visualization of SNPs with t-SNE (section 7.3)

As mentioned in section 7.4 many more dimension reduction methods remain to be compared. Another open question is whether specific properties exist in NGS data with respect to dimension reduction compared with other high-dimensional data sources, or if the dimension reduction methods are general enough for all data sources.

9.2.6 Article: TE-Locate (section 8.4)

There are three tracks for follow-ups. Firstly, as in the last paragraph of the paper, there are several tools for calling TEs which focus on different aspects of calling. Only a few of these different aspects are mutually exclusive, so combining as many as possible in one tool would outperform each of the individual tools.

Secondly, as the task is highly dependent on the aligner, it would make sense to start there. Split-read-alignment is not finally resolved, and the focus on the highly similar sequences of TEs is an additional challenge.

Thirdly, functional validation would be helpful to develop some of the initiated storylines into full stories, for example for Figure 9 in the paper. One hypothesis for Figure 9 is, that this is the functional hierarchy of TE superfamilies, that the arrows indicating superfamilies needing others to be functional.

10. Bibliography

1. Bruns, H., *Computer sucht Gen*, in *c't*. 2001. p. 216.
2. Carr, G., *Biology 2.0*, in *The Economist*. 2010: print/online.
3. Duncan, D.E., *A DNA Tower of Babel*, in *MIT Technology Review*. 2011: online.
4. di Bernardo, D., et al., *Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks.* Nat Biotechnol, 2005. **23**(3): p. 377-83.
5. Kling, J., *Careers in systems biology. Working the systems.* Science, 2006. **311**(5765): p. 1305-6.
6. Sauer, U., M. Heinemann, and N. Zamboni, *Genetics. Getting closer to the whole picture.* Science, 2007. **316**(5824): p. 550-1.
7. Ideker, T., T. Galitski, and L. Hood, *A new approach to decoding life: systems biology.* Annu Rev Genomics Hum Genet, 2001. **2**: p. 343-72.
8. **consortium, g. *1001 Genomes - A Catalog of Arabidopsis thaliana Genetic Variation*. 12.07.2014]; Available from: http://1001genomes.org/.**
9. Wingender, E., et al., *TRANSFAC: an integrated system for gene expression regulation.* Nucleic Acids Res, 2000. **28**(1): p. 316-9.
10. Mosher, D.S., et al., *A mutation in the myostatin gene increases muscle mass and enhances racing performance in heterozygote dogs.* PLoS Genet, 2007. **3**(5): p. e79.
11. Neto, E.C., et al., *Modeling causality for pairs of phenotypes in system genetics.* Genetics, 2013. **193**(3): p. 1003-13.
12. Semsarian, C. *Whole genomes and personalised medicine: where are we heading?* 2012; Available from: http://chrissemsarian.wordpress.com/category/genome/.
13. Wikipedia. *Signal-to-noise ratio -- Wikipedia, The Free Encyclopedia*. 2014.
14. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
15. Jaluria, P., et al., *A perspective on microarrays: current applications, pitfalls, and potential uses.* Microb Cell Fact, 2007. **6**: p. 4.
16. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
17. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761 Suppl): p. C47-52.
18. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.
19. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989. **340**(6230): p. 245-6.
20. Brown, K.R. and I. Jurisica, *Online predicted human interaction database.* Bioinformatics, 2005. **21**(9): p. 2076-82.
21. Aparicio, O., et al., *Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo.* Curr Protoc Mol Biol, 2005. **Chapter 21**: p. Unit 21 3.
22. Collas, P., *The current state of chromatin immunoprecipitation.* Mol Biotechnol, 2010. **45**(1): p. 87-100.
23. Farag, M.A., et al., *Metabolomics reveals novel pathways and differential mechanistic and elicitor-specific responses in phenylpropanoid and isoflavonoid biosynthesis in Medicago truncatula cell cultures.* Plant Physiol, 2008. **146**(2): p. 387-402.
24. Rios-Estepa, R. and B.M. Lange, *Experimental and mathematical approaches to modeling plant metabolic networks.* Phytochemistry, 2007. **68**(16-18): p. 2351-74.
25. Wishart, D.S., et al., *HMDB 3.0--The Human Metabolome Database in 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D801-7.
26. Wishart, D.S., et al., *HMDB: the Human Metabolome Database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D521-6.
27. Wikipedia. *Gas chromatography -- Wikipedia, The Free Encyclopedia*. 2014.

28. Ltd, K.T. *Converging Annular Time-of-flight Mass Spectrometer*. 2005; Available from: http://www.kore.co.uk/ms-200_principles.htm.
29. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
30. Staden, R., *A strategy of DNA sequencing employing computer programs.* Nucleic Acids Res, 1979. **6**(7): p. 2601-10.
31. Shen, R., et al., *High-throughput SNP genotyping on universal bead arrays.* Mutat Res, 2005. **573**(1-2): p. 70-82.
32. Wikipedia. *DNA sequencing -- Wikipedia, The Free Encyclopedia*. 2014.
33. bioinformatics.org. *Which computer language are you most interested in learning (next) for bioinformatics R&D?* 2014 [cited 2014 22.06.2014]; Available from: http://www.bioinformatics.org/poll/index.php?dispid=16&vo=16.
34. Fourment, M. and M.R. Gillings, *A comparison of common programming languages used in bioinformatics.* BMC Bioinformatics, 2008. **9**: p. 82.
35. Yang, X., et al., *HTQC: a fast quality control toolkit for Illumina sequencing data.* BMC Bioinformatics, 2013. **14**: p. 33.
36. **Long, Q., et al., *Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden.* Nat Genet, 2013. 45(8): p. 884-90.**
37. Benjamini, Y. and T.P. Speed, *Summarizing and correcting the GC content bias in high-throughput sequencing.* Nucleic Acids Res, 2012. **40**(10): p. e72.
38. Renaud, G., et al., *freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers.* Bioinformatics, 2013. **29**(9): p. 1208-9.
39. Mole, B., *The gene sequencing future is here.* Science News, 2014.
40. Li, H. and N. Homer, *A survey of sequence alignment algorithms for next-generation sequencing.* Brief Bioinform, 2010. **11**(5): p. 473-83.
41. Hernandez, D., et al., *De novo finished 2.8 Mbp Staphylococcus aureus genome assembly from 100 bp short and long range paired-end reads.* Bioinformatics, 2014. **30**(1): p. 40-9.
42. Utturkar, S.M., et al., *Evaluation and validation of de novo and hybrid assembly techniques to derive high quality genome sequences.* Bioinformatics, 2014.
43. Mascher, M., et al., *Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ).* Plant J, 2013. **76**(4): p. 718-27.
44. Ngan Nguyen, G.H., Daniel R. Zerbino, Brian Raney, Dent Earl, Joel Armstrong, David Haussler, Benedict Paten, *Building a Pangenome Reference for a Population.* Lecture Notes in Computer Science, 2014. **8394**: p. 207-221.
45. Berger, S.L., et al., *An operational definition of epigenetics.* Genes Dev, 2009. **23**(7): p. 781-3.
46. Broman, K.W., *The genomes of recombinant inbred lines.* Genetics, 2005. **169**(2): p. 1133-46.
47. Wikipedia. *Phred quality score -- Wikipedia, The Free Encyclopedia*. 2014.
48. Cornish-Bowden, A., *Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.* Nucleic Acids Res, 1985. **13**(9): p. 3021-30.
49. Pan, W., et al., *DNA polymerase preference determines PCR priming efficiency.* BMC Biotechnol, 2014. **14**: p. 10.
50. van Dijk, E.L., Y. Jaszczyszyn, and C. Thermes, *Library preparation methods for next-generation sequencing: tone down the bias.* Exp Cell Res, 2014. **322**(1): p. 12-20.
51. Poptsova, M.S., et al., *Non-random DNA fragmentation in next-generation sequencing.* Sci Rep, 2014. **4**: p. 4532.
52. Li, L., et al., *Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version).* eprint arXiv:cond-mat/0501169, 2005.
53. Shannon, C.E., *A Mathematical Theory of Communication.* Bell System Technical Journal. **27**(3): p. 379–423.
54. Smit, A., Hubley, R & Green, P. *RepeatMasker Open-3.0*. 1996-2010; Available from: http://www.repeatmasker.org.

55. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *Fourteenth International Joint Conference on Artificial Intelligence*. 1995.
56. Wikipedia. *Big O notation -- Wikipedia, The Free Encyclopedia*. 2014.
57. Wikipedia. *Graph theory -- Wikipedia, The Free Encyclopedia*. 2014.
58. Adar, E., *GUESS: a language and interface for graph exploration*, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2006, ACM: Montr&#233;al, Qu&#233;bec, Canada. p. 791-800.
59. Kobourov, S.G., *Spring Embedders and Force Directed Graph Drawing Algorithms.* ArXiv e-prints, 2012.
60. Wikipedia. *Combinatorics -- Wikipedia, The Free Encyclopedia*. 2014.
61. Mitchell, T., *Machine Learning*. 1997: McGraw Hill.
62. Bishop, C.M., *Pattern Recognition and Machine Learning*. 2006: Springer.
63. **Platzer, A. *Machine Learning - Overview*. 2012; Available from: http://sourceforge.net/projects/machine-learning2012/files/.**
64. Hume, A. and D. Sunday, *Fast string searching.* Software: Practice and Experience, 1991. **21**(11): p. 1221-1248.
65. Melichar, B., Jan Holub, and J. Polcar, *Text Searching Algorithms*. 2005: stringology.org.
66. Knuth, D., *The Art of Computer Programming*. 1968: Addison-Wesley.
67. Sedgewick, R., *Algorithms in C++*. 1998: Addison-Wesley.
68. Taubenfeld, G., *Synchronization Algorithms and Concurrent Programming*. 2006: Pearson / Prentice Hall.
69. Sanders, J., *CUDA by Example: An Introduction to General-Purpose GPU Programming*. 2010: Addison Wesley.
70. TIOBE. *TIOBE Index*. 2014; Available from: http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html.
71. Carbonnelle, P., *PYPL PopularitY of Programming Language index.* 2014.
72. Montmollin, G.d. *Language Popularity Index*. 2013; Available from: http://lang-index.sourceforge.net/.
73. Gouy, I. *The Computer Language Benchmarks Game*. 2014; Available from: http://benchmarksgame.alioth.debian.org/.
74. Stellman, A.G., Jennifer, *Applied Software Project Management*. 2005: O'Reilly Media.
75. Wikipedia. *Database -- Wikipedia, The Free Encyclopedia*. 2014.
76. Charette, R.N., *Why Software Fails*, in *IEEE spectrum*. 2005.
77. Wilson, G., et al., *Best practices for scientific computing.* PLoS Biol, 2014. **12**(1): p. e1001745.
78. Cavero, J., B. Vela, and P. Cáceres, *Computer science research: more production, less productivity.* Scientometrics, 2014. **98**(3): p. 2103-2111.
79. Curral, L.A., Forrester, R. H. Dawson, J. F. and West, M. A., *It's What You Do And The Way That You Do It: Team Task, Team Size, and Innovation-Related Group Processes.* European Journal of Work & Organizational Psychology, 2001. **10**(2): p. 187-204.
80. Yang, W., et al., *Nonlinear effects of group size on collective action and resource outcomes.* Proc Natl Acad Sci U S A, 2013. **110**(27): p. 10916-21.
81. Wikipedia. *Hard link -- Wikipedia, The Free Encyclopedia*. 2014.
82. Tido Eger, C.E.a.P.J.C. *The role of design freeze in product development*. in *ICED 05, the 15th International Conference on Engineering Design*. 2005. Melbourne, Australia.
83. Greshake, B., et al., *openSNP--a crowdsourced web resource for personal genomics.* PLoS One, 2014. **9**(3): p. e89204.
84. Haury, A.C., P. Gestraud, and J.P. Vert, *The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures.* PLoS One, 2011. **6**(12): p. e28210.
85. Venet, D., J.E. Dumont, and V. Detours, *Most random gene expression signatures are significantly associated with breast cancer outcome.* PLoS computational biology, 2011. **7**(10): p. e1002240.

86. Wilkinson, D.M. and B.A. Huberman, *A method for finding communities of related genes.* Proc Natl Acad Sci U S A, 2004. **101 Suppl 1**: p. 5241-8.
87. Zhu, D. and Z.S. Qin, *Structural comparison of metabolic networks in selected single cell organisms.* BMC Bioinformatics, 2005. **6**: p. 8.
88. Sen, T.Z., A. Kloczkowski, and R.L. Jernigan, *Functional clustering of yeast proteins from the protein-protein interaction network.* BMC Bioinformatics, 2006. **7**: p. 355.
89. Wuchty, S., *Scale-free behavior in protein domain networks.* Mol Biol Evol, 2001. **18**(9): p. 1694-702.
90. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
91. **Platzer, A., *Numerische Deskriptoren zur Charakterisierung von Proteinnetzwerken*, in *A490 - Molekulare Biologie*. 2006, University of Vienna: ÖNB Hauptabt. Heldenplatz.**
92. Rhodes, D.R., et al., *ONCOMINE: a cancer microarray database and integrated data-mining platform.* Neoplasia, 2004. **6**(1): p. 1-6.
93. Katzgraber, H.G., K. Janzen, and C.K. Thomas, *Boolean decision problems with competing interactions on scale-free networks: critical thermodynamics.* Phys Rev E Stat Nonlin Soft Matter Phys, 2012. **86**(3 Pt 1): p. 031116.
94. **Platzer, A., et al., *Characterization of protein-interaction networks in tumors.* BMC Bioinformatics, 2007. 8: p. 224.**
95. Leclerc, R.D., *Survival of the sparsest: robust gene networks are parsimonious.* Mol Syst Biol, 2008. **4**: p. 213.
96. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.
97. Li, X.Y., et al., *Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm.* PLoS Biol, 2008. **6**(2): p. e27.
98. Gene Ontology, C., *The Gene Ontology in 2010: extensions and refinements.* Nucleic Acids Res, 2010. **38**(Database issue): p. D331-5.
99. **Platzer, A., *The Occurrence-in-subtuple problem.* ArXiv e-prints, 2008.**
100. Fisher, R.A., *The Use of Multiple Measurements in Taxonomic Problems.* Annals of Eugenics, 1936. **7**(2): p. 179–188.
101. Rosenblatt, F., *The perceptron : a probabilistic model for information storage and organization in the brain.* Psychological Reviews, 1958: p. 386-408.
102. Abdi, H., *Neural Networks*. 1999: SAGE.
103. Breiman, L., *Statistical Modeling: The Two Cultures.* Statist. Sci., 2001. **16**(3): p. 199-231.
104. Mark Hall, E.F., Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, 2009. **11**(1).
105. **Catchpole, G., et al., *Metabolic profiling reveals key metabolic features of renal cell carcinoma.* J Cell Mol Med, 2011. 15(1): p. 109-18.**
106. Yoav Freund, L.M. *The Alternating Decision Tree Algorithm*. in *16th International Conference on Machine Learning*. 1999.
107. Patro, R., S.M. Mount, and C. Kingsford, *Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.* Nat Biotechnol, 2014. **32**(5): p. 462-4.
108. Zhang, Z. and W. Wang, *RNA-Skim: a rapid method for RNA-Seq quantification at transcript level.* Bioinformatics, 2014. **30**(12): p. i283-i292.
109. Zhao, M., et al., *SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications.* PLoS One, 2013. **8**(12): p. e82138.
110. Beach, T.E. *Types of Computers*. 2004 10.7.2014]; Available from: http://www.unm.edu/~tbeach/terms/types.html.
111. *Classes of computers*. 2013 10.7.2014]; Available from: http://en.wikipedia.org/wiki/Category:Classes_of_computers.
112. APA, *Neuer Supercomputer "Mendel" für Biologen in Wien*, in *Der Standard*. 2013: web.

113. **Platzer, A. *Analysis of whole-genome sequence data*. 2014; Available from: http://sourceforge.net/projects/whole-genome-analysis-part1/files/.**
114. Pabinger, S., et al., *A survey of tools for variant analysis of next-generation genome sequencing data.* Brief Bioinform, 2014. **15**(2): p. 256-78.
115. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
116. Li, H., *Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly.* Bioinformatics, 2012. **28**(14): p. 1838-44.
117. El-Metwally, S., et al., *Next-generation sequence assembly: four stages of data processing and computational challenges.* PLoS Comput Biol, 2013. **9**(12): p. e1003345.
118. Clayton, D.G., et al., *Population structure, differential bias and genomic control in a large-scale, case-control association study.* Nat Genet, 2005. **37**(11): p. 1243-6.
119. Freedman, M.L., et al., *Assessing the impact of population stratification on genetic association studies.* Nat Genet, 2004. **36**(4): p. 388-93.
120. Marchini, J., et al., *The effects of human population structure on large genetic association studies.* Nat Genet, 2004. **36**(5): p. 512-7.
121. Han, E., J.S. Sinsheimer, and J. Novembre, *Characterizing bias in population genetic inferences from low-coverage sequencing data.* Mol Biol Evol, 2014. **31**(3): p. 723-35.
122. **Alexander Platzer, D.M. *Programs - Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden*. 2013; Available from: http://downloads.gmi.oeaw.ac.at/downloads/nordborg/data-release-swedish-lines/programs.**
123. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol, 2010. **34**(8): p. 816-34.
124. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
125. **Platzer, A., V. Nizhynska, and Q. Long, *TE-Locate: A Tool to Locate and Group Transposable Element Occurrences Using Paired-End Next-Generation Sequencing Data.* Biology, 2012. 1(2): p. 395-410.**
126. K. Beyer, J.G., R. Ramakrishnan, U. Shaft. *When is "Nearest Neighbor" Meaningful?* in *ICDT'99*. 1999. Springer.
127. M. E. Houle, H.-P.K., P. Kröger, E. Schubert, A. Zimek. *Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?* in *21st International Conference on Scientific and Statistical Database Management (SSDBM)*. 2010.
128. RE, B., *Adaptive Control Processes*. 1961, Princeton, NJ: Princeton University Press.
129. L.J.P. van der Maaten, E.O.P., and H.J. van den Herik, *Dimensionality Reduction: A Comparative Review*, in *Tilburg University Technical Report*. 2009, Tilburg University.
130. Kullback, S.L., R.A, *On Information and Sufficiency.* Annals of Mathematical Statistics, 1951. **22**(1): p. 79–86.
131. Ali, S.M.S., S. D., *A general class of coefficients of divergence of one distribution from another.* Journal of the Royal Statistical Society, Series B, 1966. **28**(1): p. 131–142.
132. Onclinx, V., et al. *Dimensionality reduction by rank preservation*. in *Neural Networks (IJCNN), The 2010 International Joint Conference on*. 2010. IEEE.
133. Wikipedia. *Differential entropy -- Wikipedia, The Free Encyclopedia*. 2014.
134. Dunn, J.C., *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.* 1973.
135. Rousseeuw, P.J., *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of computational and applied mathematics, 1987. **20**: p. 53-65.
136. Van der Maaten, L. and G. Hinton, *Visualizing data using t-SNE.* Journal of Machine Learning Research, 2008. **9**(2579-2605): p. 85.
137. **Platzer, A., *Visualization of SNPs with t-SNE.* PLoS One, 2013. 8(2): p. e56883.**

138. Begun, D.J., et al., *Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans.* PLoS Biol, 2007. **5**(11): p. e310.
139. Tenenbaum, J.B., V. de Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction.* Science, 2000. **290**(5500): p. 2319-23.
140. Roweis, S.T. and L.K. Saul, *Nonlinear dimensionality reduction by locally linear embedding.* Science, 2000. **290**(5500): p. 2323-6.
141. **Alexander Platzer, K.F., Jasmin Music, Michael Neumann. *Visualisierung von genomischen Daten*. 2013 [cited 2014 18.07.2014]; Available from: http://sourceforge.net/projects/dimensionreductiondemo/files/.**
142. Castillo-Davis, C.I., *The evolution of noncoding DNA: how much junk, how much func?* Trends Genet, 2005. **21**(10): p. 533-6.
143. McClintock, B., *The Discovery and Characterization of Transposable Elements: The Collected Papers of Barbara McClintock*. 1987, New York, NY: Garland Publishing, Inc.
144. Nowacki, M., et al., *A functional role for transposases in a large eukaryotic genome.* Science, 2009. **324**(5929): p. 935-8.
145. Hollister, J.D. and B.S. Gaut, *Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression.* Genome Res, 2009. **19**(8): p. 1419-28.
146. Tenaillon, M.I., J.D. Hollister, and B.S. Gaut, *A triptych of the evolution of plant transposable elements.* Trends Plant Sci, 2010. **15**(8): p. 471-8.
147. Kazazian, H.H., Jr., *Mobile elements and disease.* Curr Opin Genet Dev, 1998. **8**(3): p. 343-50.
148. Kazazian, H.H., Jr., *Mobile elements: drivers of genome evolution.* Science, 2004. **303**(5664): p. 1626-32.
149. Lippman, Z., et al., *Role of transposable elements in heterochromatin and epigenetic control.* Nature, 2004. **430**(6998): p. 471-6.
150. Bourque, G., et al., *Evolution of the mammalian transcription factor binding repertoire via transposable elements.* Genome Res, 2008. **18**(11): p. 1752-62.
151. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
152. Belancio, V.P., D.J. Hedges, and P. Deininger, *Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health.* Genome Res, 2008. **18**(3): p. 343-58.
153. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.
154. Yang, J., et al., *Advantages and pitfalls in the application of mixed-model association methods.* Nat Genet, 2014. **46**(2): p. 100-6.
155. Li, G. and H. Zhu, *Genetic Studies: The Linear Mixed Models in Genome-wide Association Studies.* Open Bioinformatics Journal, 2013. **7**(1): p. 27-33.
156. Jallow, M., et al., *Genome-wide and fine-resolution association analysis of malaria in West Africa.* Nat Genet, 2009. **41**(6): p. 657-65.
157. Bush, W.S. and J.H. Moore, *Chapter 11: Genome-wide association studies.* PLoS Comput Biol, 2012. **8**(12): p. e1002822.
158. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.
159. Johnson, A.D. and C.J. O'Donnell, *An open access database of genome-wide association results.* BMC Med Genet, 2009. **10**: p. 6.
160. Johnson, D.S., *A theoretician's guide to the experimental analysis of algorithms.* Data structures, near neighbor searches, and methodology: fifth and sixth DIMACS implementation challenges, 2002. **59**: p. 215-250.
161. Frommer, M., et al., *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.* Proc Natl Acad Sci U S A, 1992. **89**(5): p. 1827-31.
162. Kim, M.S., et al., *A draft map of the human proteome.* Nature, 2014. **509**(7502): p. 575-81.

163. Siegel, V., *Where credit is due.* Disease models & mechanisms, 2008. **1**(4-5): p. 187-191.
164. Rose, A., *Hackathons: Proof That Ideas Are Cheap and Implementation Is Expensive*, in *Dr. Dobb's*. online.
165. Abbott, A., *Human genome at ten: The human race.* Nature, 2010. **464**: p. 668-669.
166. Wikipedia. *Killer application -- Wikipedia, The Free Encyclopedia*. 2014.

11. Appendix

11.1 Curriculum Vitae with complete list of publications

Name: Alexander Platzer

Address: Kornhäuselgasse 3/2/15, 1200 Vienna, Austria

email: AlexanderP@gmx.at

March 1981	born (Vienna, Austria)
1987 – 1991	elementary school, 2243 Matzen, Lower Austria
1991 – 1995	grammar school, 2243 Matzen, Lower Austria
1995 – 2000	Higher Technical School (HTL), department D (data processing and organisation), 1220 Vienna
2000 – 2001	company Evis (EVolution of Intelligent Systems, development of algorithms), 1040 Vienna
2001 – 2004	Molecular Biology at University of Vienna, first part
2004 – 2006	Molecular Biology at University of Vienna, second part
Febr. – Oct. 2006	Master thesis
2007-2008	Max Planck Institute of Molecular Plant Physiology in Potsdam (Department Bioinformatics), Bioinformatician
2009 – 2011	ftw - Telecommunications Research Center Vienna, Scientist/Engineer for data analysis
From Sept. 2011	Gregor Mendel Institute of Molecular Plant Biology (GMI), Bioinformatician

Teaching experience

Machine learning - 2012

http://sourceforge.net/projects/machine-learning2012/files/
at VBC (Campus Vienna BioCenter)

Analysis of Whole-Genome Sequence Data - 2014

http://sourceforge.net/projects/whole-genome-analysis-part1/files/
at University of Neuchâtel

Practical Course on Next Generation Sequencing - 2014

http://sourceforge.net/projects/de-novo-assembly-lecture/files/

at Vetmeduni Vienna

Publications

Long Q*, Rabanal FA*, Meng D*, Huber CD*, Farlow A*, **Platzer A**, Zhang Q, Vilhjalmsson BJ, Korte A, Nizhynska V, Voronin V, Korte P, Sedman L, Mandakova T, Lysak MA, Seren U, Hellmann I, Nordborg M. 2013. Massive genomic variation and strong selection in Swedish *Arabidopsis thaliana*. Nature Genetics 45, 884–890

Platzer A. 2013. Visualization of SNPs with t-SNE. PLoS ONE 8(2): e56883. doi:10.1371/journal.pone.0056883

Platzer Alexander, Nizhynska Viktoria, Long Quan. 2012. TE-Locate: A Tool to Locate and Group Transposable Element Occurrences Using Paired-End Next-Generation Sequencing Data. Biology 1, no. 2: 395-410.

Hossfeld T, Biedermann S, Schatz R, **Platzer A**, Egger S, Fiedler, M: The memory effect and its implications on Web QoE modeling; Teletraffic Congress (ITC), 2011 23rd International, 103 - 110

Schatz R, Egger S, **Platzer A** 2011. Poor, Good Enough or Even Better? Bridging the Gap between Acceptability and QoE of Mobile Broadband Data Services. ICC'11 CQRM

Catchpole G*, **Platzer A***, Weikert C, Kempkensteffen C, Johannsen M, Krause H, Jung K, Miller K, Willmitzer L, Selbig J, Weikert S. 2009. Metabolic profiling reveals key metabolic features of renal cell carcinoma. J Cell Mol Med **15**(1):109-18

Platzer A. 2008. The Occurrence-in-subtuple problem. arXiv:0811.4192v1 [math.CO]

Platzer A, Perco P, Lukas A, Mayer B. 2007. Characterization of protein interaction networks in tumors. BMC Bioinformatics **8**(1):224.

11.2 List of figures

Figure 1. Diagram of a gas chromatograph *10*
Figure 2. A TOF-MS detector *10*
Figure 3. Illumina NGS method *12*
Figure 4. Segments of DNA, resp. their names in a fragment for read-pair sequencing *12*
Figure 5. The amount of sequencing *14*
Figure 6. Computational complexity *17*
Figure 7. Example of a small graph *19*
Figure 8. Example of a large graph *19*
Figure 9. Example scale-free network *24*
Figure 10. Decision Tree Model (ADTree) for RCC *59*
Figure 11. Pipeline for the full population of samples in section 6.6 .... *66*
Figure 12. LAE-finder – Inversion – Event support *109*
Figure 13. LAE-finder - Inversion – Negative support. *110*
Figure 14. LAE-finder - Translocation – Event support *110*
Figure 15. LAE-finder – Translocation – Finding the best breakpoint *111*
Figure 16. LAE-finder – Translocation – Negative support *111*
Figure 17. Combining masks of samples *112*
Figure 18. Full-uniqueness property *113*
Figure 19. Sequence realignment *114*
Figure 20. Using read-pairs for inferring insertions or deletions *115*
Figure 21. Example coverage for a likely duplication in the middle. *115*
Figure 22. Split reads *116*

11.3 List of tables

Table 1. Overview of DNA sequencing technologies *11*
Table 2. Family of Bachmann–Landau notations / asymptotic notations *18*

12. Acknowledgements

For making this work possible and generally for my scientific habit, I want to thank

- Company *EVIS*, where I saw first the pure data analysis and how it is organized; especially *Helmut Mach* as head and *Günther Brunthaler* as leading computer scientist. The company was wound up one year after I left them for study, but I can clearly say it was not because of the analyses.
- Company *emergentec*, where I was doing my diploma thesis, which led later to my first great publication; especially *Bernd Mayer* and *Paul Perco* as my first and secondary supervisor
- *Erich Pils*, one of the true master computer scientists I know and a great and tough teacher
- the *Erasmus Programme*, which made my semester abroad possible in an organized way, even when it was only in Germany
- *Rene Pilz* for having the highest and lowest level discussions about source code
- *Alessandro D'Alconzo* for many funny, but still solid scientific discussions
- *Tobias Hoßfeld* for a really efficient collaboration
- *Quan Long* for many good discussions, support and papers
- *Thomas Friese* for proof-reading and for highly improving some of my complex long sentences
- my supervisor *Irina Druzhinina* for making this work possible and streamlined
- my family and my friends