Genealogy Reconstruction

Methods and applications in cancer and wild populations

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt
von Diplom Informatiker *Markus Riester*
geboren am 31. März 1979 in Villingen-Schwenningen

Die Annahme der Dissertation haben empfohlen:

1. Professor Dr. Peter F. Stadler (Leipzig, Deutschland)
2. Professor Dr. David Bryant (Auckland, New Zealand)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 23. Juni 2010 mit dem Gesamtprädikat *magna cum laude*

Contents

Abstract

Genealogy reconstruction is widely used in biology when relationships among entities are studied. Phylogenies, or evolutionary trees, show the differences between species. They are of profound importance because they help to obtain better understandings of evolutionary processes. Pedigrees, or family trees, on the other hand visualize the relatedness between individuals in a population. The reconstruction of pedigrees and the inference of parentage in general is now a cornerstone in molecular ecology. Applications include the direct inference of gene flow, estimation of the effective population size and parameters describing the population's mating behaviour such as rates of inbreeding.

In the first part of this thesis, we construct genealogies of various types of cancer. Histopathological classification of human tumors relies in part on the degree of differentiation of the tumor sample. To date, there is no objective systematic method to categorize tumor subtypes by maturation. We introduce a novel algorithm to rank tumor subtypes according to the dissimilarity of their gene expression from that of stem cells and fully differentiated tissue, and thereby construct a phylogenetic tree of cancer. We validate our methodology with expression data of leukemia and liposarcoma subtypes and then apply it to a broader group of sarcomas and of breast cancer subtypes. This ranking of tumor subtypes resulting from the application of our methodology allows the identification of genes correlated with differentiation and may help to identify novel therapeutic targets. Our algorithm represents the first phylogeny-based tool to analyze the differentiation status of human tumors.

In contrast to asexually reproducing cancer cell populations, pedigrees of sexually reproducing populations cannot be represented by phylogenetic trees. Pedigrees are directed acyclic graphs (DAGs) and therefore resemble more phylogenetic networks where reticulate events are indicated by vertices with two incoming arcs. We present a software package for pedigree reconstruction in natural populations using co-dominant genomic markers such as microsatellites and single nucleotide polymorphism (SNPs) in the second part of the thesis. If available, the algorithm makes use of prior information such as known relationships (sub-pedigrees) or the age and sex of individuals. Statistical confidence is estimated by Markov chain Monte

Carlo (MCMC) sampling. The accuracy of the algorithm is demonstrated for simulated data as well as an empirical data set with known pedigree. The parentage inference is robust even in the presence of genotyping errors. We further demonstrate the accuracy of the algorithm on simulated clonal populations. We show that the joint estimation of parameters of interest such as the rate of self-fertilization or clonality is possible with high accuracy even with marker panels of moderate power. Classical methods can only assign a very limited number of statistically significant parentages in this case and would therefore fail. The method is implemented in a fast and easy to use open source software that scales to large datasets with many thousand individuals.

Acknowledgments

I want to thank Konstantin and Peter for their extraordinary support and for giving me the opportunity to work on this thesis. They *always* had time when I needed their help.

Next, I would like to thank Camille and Franziska for giving me the opportunity to work in crazy Manhattan.

I would also like to thank Petra for helping me filling out all these Dienstreise-, Reisekosten- and Urlaubsanträge.

Kudos to Jens for the administration of the computers and the coffee machine.

Of course I would like to thank my lovely office mates Clara and Jan. It was always a pleasure to work with them.

I would also like to thank my colleagues who contributed to the success of this thesis, especially Alex, Andrea, David, Florian, Gunnar, Lydia, Marc, Maribel, Mario, Nico, Philipp-Jens, Steps, Steve and Wolfgang.

Finally, I am indebted to my family for their patience and encouragement.

Financial support

This work was supported by the 6th Framework Programme of the European Union, project 043251 "EDEN".

1 Introduction

A pedigree, or family tree, is one of the best known graph structures in biology, even among non-scientists. Our own pedigree tells us who our ancestors and distant relatives are, or more generally where we actually come from. As an example, Fig. 1.1 illustrates the pedigree of the current Royal House of the United Kingdom. Pedigrees are also well-known to the broad public due to their importance in animal- and plant-breeding. The pedigree of a race horse, for instance, is always well documented and victorious stallions are put to stud after their retirement from racing. We say then the ensuing foals have a good pedigree.

For decades and even centuries, the main method for constructing pedigrees was simply the accumulation of observed parent-offspring and sibling relationships. Modern pedigree analysis started with the availability of methods that could determine DNA sequences of single individuals of a population. In the late eighties, the field of forensic was revolutionized by the discovery of what everybody now knows as DNA fingerprints: short tandem repeats (STRs). These are very short sequences of DNA (2-6 base pairs) that are repeated many times in a row within the human genome. The numbers of the repeats vary between individuals in a population. Since there are many of such STRs at different positions (loci) in the human genome, the probability that two individuals have the same combination of repeat numbers over all loci gets extremely small, when sufficiently many loci are considered. In 1997, the

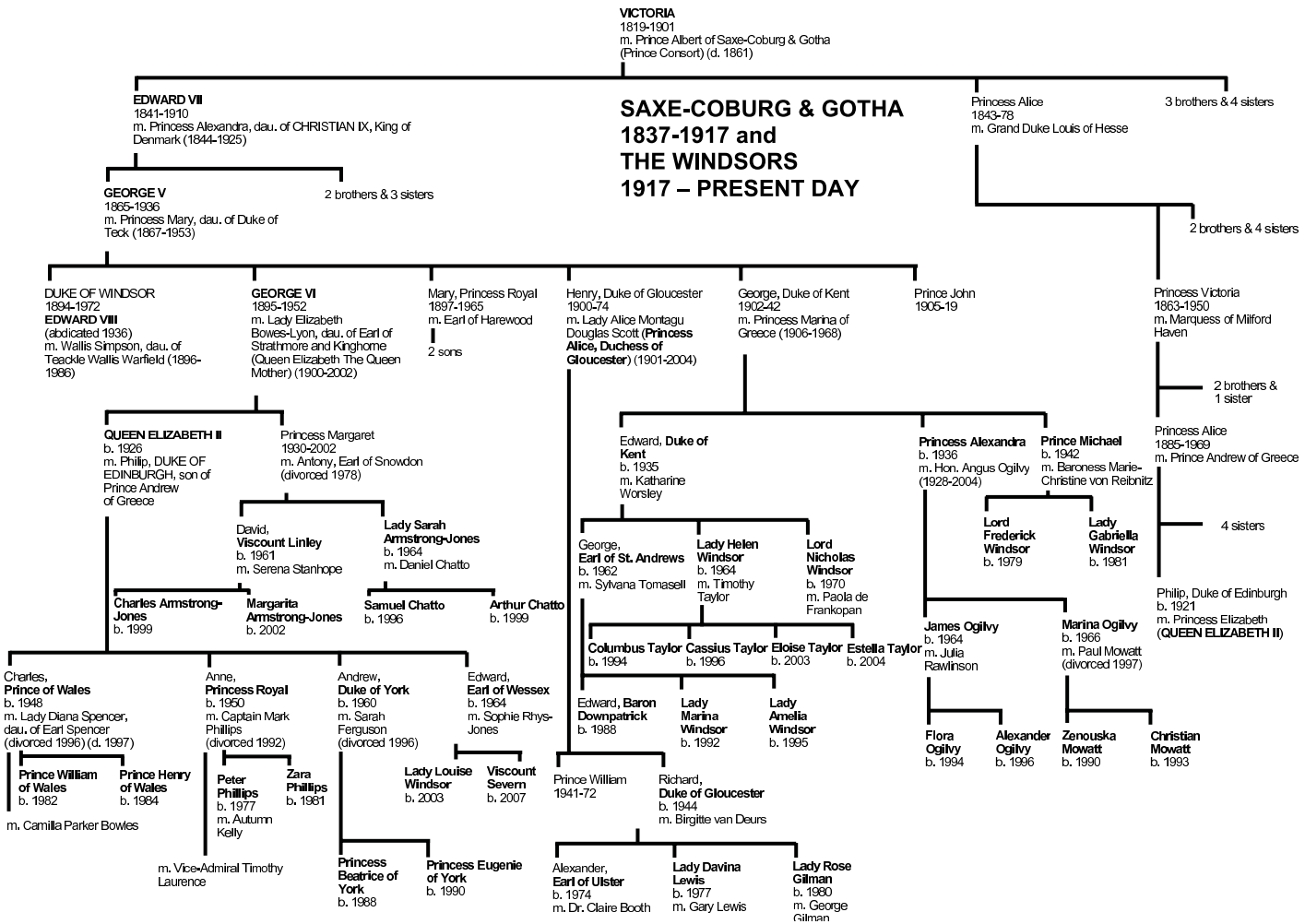


Figure 1.1. *A royal pedigree. The official pedigree of the House of Saxe-Coburg and Gotha and the House of Windsor.*
Source: `http://www.royal.gov.uk`

FBI introduced the Combined DNA Index System (CODIS), which is a set of 13 loci [Butler 2006]. The Interpol standard STR set is a subset of 7 of these loci [Ruitberg et al. 2001]. Thus only a tiny fraction of the whole DNA has to be sequenced to identify victims or criminals.

This technique soon showed its usefulness in a very related field: the paternity inference. Here the STR repeat numbers are determined from the child, the mother and the alleged father. As a child inherits one chromosome from the mother and one from the father, paternity is ruled out when the child shows repeat numbers different from the alleged parents. Although there are no reliable estimates of the number of paternity tests world wide, it is clear that this is now a billion dollar industry.

With the advances in modern sequencing techniques, the genotyping of individuals became cheaper and faster. It is now getting feasible to obtain DNA samples of thousands of individuals in a population. This makes pedigree reconstruction methods interesting also for natural populations. A population's pedigree bears witness of the mating behaviour of its individuals. For example, it allows a direct inference of inbreeding or fitness of particular individuals. In other words, it shows the recent past of a population with the highest possible resolution.

Phylogenies, or evolutionary trees, primarily show differences between species. Charles Darwin proved that all species are descendants from a common ancestor and he illustrated the ancient speciation events with a *tree of life*. New species can arise when populations get separated and then adept to new environments through mutations and selections. Another major force of diversification, especially in small isolated populations, is random genetic drift, i.e., when traits get lost by chance. Thus evolution is an ongoing process and it is rarely possible to obtain data from individuals of extinct populations or species. Therefore, the common ancestors, in graph terminology internal nodes, are reconstructed and not directly observed. Before the advent of DNA sequencing, phylogenies were reconstructed by means of phenotypic traits. In molecular phylogenetics, observed differences in DNA sequences of *homologous* genes, i.e., genes that are present in all of the species in the phylogeny, are now routinely used for the tree reconstructions.

Fundamental work in population genetics showed also importance in the field of cancer research. A population is here a group of N cells. Mutations of tumor suppressor genes may lead to cancer cells with different fitness values compared to the wild type cells. Population genetic models can help to understand cancer initiation and progression [Iwasa et al. 2005]. We will show in Chapter 2 that phylogenetic methods can help to understand relationships among cancer subtypes. Specialized cell types such as skin, blood or fat arise from different *progenitor* cells, which in turn arise from embryonic stem cells. Hence, as stem cells develop or *differentiate* to specialized somatic cells, it is possible to construct genealogies of cell

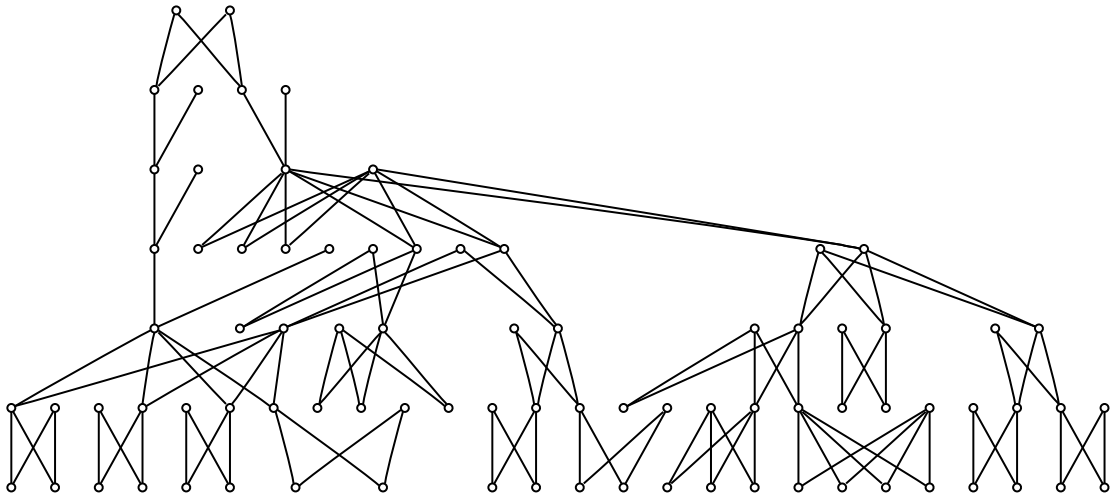


Figure 1.2. *The royal pedigree of Fig 1.1 in a graph representation. The vertices represent the individuals and the arcs the parent-offspring relationships. The direction of the arcs is from top to bottom.*

types. It is now apparent that disruption of normal differentiation is an important component of tumorigenesis. Fully differentiated somatic cells arise from stem cells, with changes in gene expression that can be experimentally determined. If cancers arise as the result of an abruption of the differentiation process, then poorly differentiated cancers would have a gene expression more similar to stem cells than to normal differentiated tissue, and well differentiated cancers would have a gene expression more similar to fully differentiated cells than to stem cells. In other words, tumor cells may leave the development path from stem cell to normal cell at some point. Genealogies of tumor subtypes thus have the potential to visualize the branching points from the differentiation course. In many cancers it can be observed that the earlier a tumor branched off this path, the poorer the prognosis. An important difference to species phylogenies is that it is possible to obtain data from stem cells being the common ancestor of all cancer and fully differentiated cells.

Chapter 2 is based on the following publication:

- Riester M.*, Stephan-Otto Attolini C.*, Downey R. J., Singer S. & Michor F. (2010). A differentiation-based phylogeny of cancer subtypes. *PLoS Computational Biology.* 6(5): e1000777. *Equal contribution.

Here we construct genealogies of various cancer types by means of *microarray* data. Microarrays are a rather inexpensive high-throughput technique for gene expression profiling of thousands of genes at the same time.

In contrast to asexually reproducing cancer cell populations, pedigrees of sexually reproducing populations cannot be represented by trees as individuals have two parents (see Fig 1.2). In graph terminology, pedigrees are directed acyclic graphs (DAGs) and therefore resemble more *phylogenetic networks* [*e.g.* Huson and Bryant 2006, Grünewald et al. 2007] where reticulate events such as hybridization, horizontal gene transfer or recombination might produce hybrid species. These are represented in a DAG with a node with two parents.

In Chapter 3, we will show how to reconstruct genealogies in natural populations by means of STR data. In particular, we will present a powerful new program `FRANz`. Chapter 3 is based on the following papers:

- Riester, M., Stadler, P. F. & Klemm K. (2008). `FRANz`: Fast reconstruction of wild pedigrees. *Lecture Notes in Informatics* P-136, 168 (Proceedings of the German Conference on Bioinformatics).
- Riester, M., Stadler, P. F. & Klemm K. (2009). `FRANz`: Reconstruction of wild multi-generation pedigrees. *Bioinformatics* **25**, 2134.
- Riester, M., Stadler, P. F. & Klemm K. (2010). Reconstruction of pedigrees in clonal plant populations. *Theoretical Population Biology.* In press.

2

Cancer Phylogenies

2.1 Introduction

Cancer research has traditionally focused on the identification of oncogenes and tumor suppressor genes, but in the last decades it has become increasingly apparent that disruption of normal differentiation is an important component of tumorigenesis. Lack of cellular maturation is now recognized as a hallmark of human cancers [Hanahan and Weinberg 2000], and the degree of differentiation of a tumor is important for diagnosis, prognosis, and treatment. Investigations of hematopoietic malignancies, for instance, have benefited considerably from an understanding of the differentiation hierarchy of hematopoietic cells. The identification of immunophenotypic markers and gene expression profiles correlated with maturation has enabled researchers to map the expansion of malignant cells to particular stages of hematopoietic differentiation [Bennett et al. 1976]. Such characterization has proven invaluable for diagnostic and prognostic purposes, and continues to provide clues for pharmacological interventions. Furthermore, the extent of differentiation indicated by the histologic subtype of liposarcoma is the most important determinant of the clinical outcome for this cancer type [Kooby et al. 2004, Singer et al. 2003, Dalal et al. 2006]. Nevertheless, attempts to categorize solid tumors have proven difficult due to an incomplete understanding of differentiation pathways

from stem cells into mesenchymal and epithelial tissues. The classifications undertaken so far have been based on *in vitro* measurements of genes expressed during the differentiation of stem cells into mature tissue; this data was then compared to expression profiles of different tumor subtypes to identify the maturation stages to which these subtypes correspond [Matushansky et al. 2008]. However, such approaches are not yet widely applicable since the prospective isolation of tissue-specific stem cells has been possible for only few tissue types, e.g. hematopoietic, mesenchymal, epithelial, and neural tissues (Minguell et al. [2001] and references therein). Similarly, *in vitro* methods of differentiation are available for only a few histologies [Beqqali et al. 2006]. Furthermore, the necessity of an array of growth factors for *in vitro* differentiation raises questions about the similarity of the *in vitro* model to in vivo processes. Often only a fraction of cells undergoes differentiation under *in vitro* conditions, and currently available methods do not allow isolation of those cells during the differentiation process from the bulk of unchanged cells.

An objective categorization of cancers according to maturity requires a methodology that does not depend on expression data obtained from *in vitro* models of differentiation. In this chapter, we develop a novel computational algorithm that assigns a degree of dissimilarity from stem cells to human cancer subtypes. Our methodology utilizes gene expression data of tumor subtypes to construct a phylogenetic tree based on genes differentially expressed among the subtypes, as well as gene expression data of stem cells and fully differentiated cells. The resulting phylogeny provides information about the maturation status of tumor subtypes and the relationship between them. The results of our algorithm are conceptually similar to the mapping of cellular expansion occurring during hematopoietic malignancies to the differentiation hierarchy of hematopoiesis. Our methodology allows classification of cancer subtypes according to their maturation status, to identify genes whose expression correlates with differentiation, and to discover candidate genes which are promising therapeutic targets. Our methodology is part of an increasing literature of mathematical and statistical investigations of cancer [Desper et al. 1999, von Heydebreck et al. 2004, Beerenwinkel et al. 2005, Newton 2002, Merlo et al. 2006, Michor et al. 2004].

This chapter is organized as follows. In Sec. 2.2 we give a short introduction into the phylogenetic methods used in this chapter. We further briefly explain microarrays. The methodology is presented in detail in Sec. 2.3. We then present our results in Sec. 2.4 and discuss them in Sec. 2.5.

2.2 Background

2.2.1 Phylogenetic Trees

Evolutionary relationships among species are typically visualized by phylogenetic trees. These are bifurcating trees where the leaves represent the species or taxa and the internal nodes the inferred common ancestors. Edge lengths can often be interpreted as time estimates.

Several methods for reconstruction of phylogenetic trees exist and can be classified in two major groups. The first group are the character-based methods which use the source data such as a multiple sequence alignment (MSA) directly. Maximum Parsimony for instance finds the tree that can explain the data with the smallest amount of mutations. The second group are the distance-based methods. In this chapter, we will focus on this group. Here, the source data is first used to estimate pairwise distances of the taxa. Various methods exist that try to find the tree that fits these distances best. We will explain the most important algorithms briefly in Sec. 2.3.2.

Bootstrap in Phylogeny

It is often desired to assess the uncertainty of a phylogenetic tree reconstruction. In Maximum Parsimony, Maximum Likelihood and distance-based methods, the *bootstrap* test [Efron 1979] is a widely used technique. This test empirically analyzes the variance of a parameter estimated out of a sample of size n by using only the n observations. Applied to phylogeny, the parameter of interest is the tree topology and the n observations are the characters, for example the columns in the MSA. In this test, it is assumed that the observed characters are drawn independently from the species' genomes. The test works by sampling n columns *with replacement*. The phylogeny is then reconstructed by means of this *bootstrap replicate*. This sampling is repeated m times; we thus have a forest of m trees. The observed *branches* in this forest are then counted. For example consider a phylogeny of hominoid primates (Fig. 2.1). Then one possible branch would be gorilla branching before human and chimpanzee. Some tree topologies might in contrast indicate that human branched before gorilla and chimpanzee. The *bootstrap value* of a branch is the percentage of trees having this particular branch. The tree that contains all branches with bootstrap values larger than 50% is called the *majority-rule consensus tree*. All trees we show in this chapter are *rooted* majority-rule consensus trees.

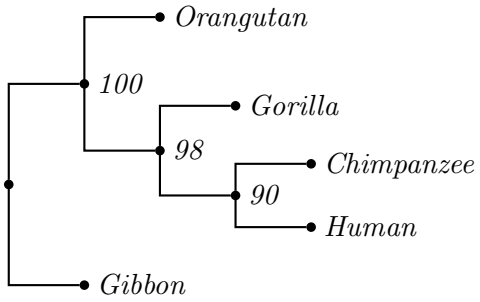


Figure 2.1. *Example phylogeny of hominoid primates. A phylogenetic tree visualizing the evolutionary relationships among human, chimpanzee, gorilla, orangutan and gibbon. The numbers represent bootstrap values. For example, the boostrap value of the branch* `(Chimpanzee, Human)|(Gorilla,Orangutan,Gibbon)` *is 90 and thus showing that it was observed in 90% of all topologies generated during the bootstrap test. The root of the tree was specified by including gibbon as an evolutionary outgroup.*

Rooting unrooted trees

Most tree reconstruction methods produce unrooted trees. A common way of rooting a tree is to include some taxon or taxa as evolutionary *outgroup*. These are taxa known to have branched before all other species in the phylogeny. Therefore the root, the common ancestor of all taxa, is located on the branch separating the outgroup and the species (Fig. 2.1). The outgroup should be related as closely as possible to the species. For example, gibbon is in the hominoid primate phylogeny a better outgroup than mouse and rat. If the outgroup is too distant, then it attaches almost randomly to the unrooted tree [Graham et al. 2002].

2.2.2 Microarrays

In eukaryotes, *genes* are stretches of DNA on the chromosomes which are located in the nucleus. Genes are *transcribed* into single stranded precursor messenger RNA (pre-mRNA). These pre-mRNAs consist of *exons* and *introns*. The latter are removed by a process called splicing. The ensuing mature mRNAs are then exported into the cytoplasm. Messenger RNAs transcribed from a protein-coding genes are then *translated* into polypeptide chains, the proteins. Small non-coding RNAs (ncRNAs) have shown to be important regulators of gene expression. Some classes of ncRNA such as siRNAs in prokaryotes can induce degradation of mRNAs, others, *e.g.* miRNAs in animals, can inhibit translation.

Microarrays are a rather inexpensive and fast high-throughput method for the measurement

of gene expression. We will focus here on the Affymetrix GeneChip® technology. The Human Genome U133 (HG-U133) GeneChip set consists of two micorarray chips which can measure the expression of approximately 39.000 genes. One important application in cancer research is the comparison of gene expression in cancer patients to that in a healthy control group. Statistical methods such as t-tests and variants of it are then used to find differentially expressed genes, i.e, genes that show a significantly different expression in cancer patients.

RNA isolated from tissue or cell line are reverse transcribed and labeled with biotin. The resulting cRNA sample is then put on the array. A GeneChip consists of *probes* that are complementary to cRNAs of well-substantiated genes. If some of these cRNAs are present in the sample, then they will hybridize to the probes. These hybridizations can be measured due to the biotin labeling. The measured probe intensity values are stored in Affymetrix *CEL files*. These files are then normalized together with all other CEL files considered in the study. This results in an *expression matrix* where the genes (or so called probe sets in the case of Affymetrix chips) are represented by the rows and the columns hold the normalized expression values of all experiments, *e.g.* of the patients and the control group.

2.3 Methods

Our algorithm uses gene expression data of tumor samples that have been pathologically classified into subtypes. The expression data is normalized and then analyzed for differentially expressed genes, i.e. those genes whose expression in samples from one tumor subtype differs from the expression in samples from at least one other subtype. We use these genes to compute the distances between all pairs of subtypes; the resulting distance matrix is then used to construct a phylogenetic tree. This construction is repeated several thousand times using different subsets of genes (of varying size) to estimate the statistical significance of the branches of the tree (Fig. 2.2).

2.3.1 Dataset compilation

We use gene expression data of sarcoma samples from [Singer et al. 2007, Barretina et al. 2010]. The gene expression was measured on Affymetrix U133a oligonucleotide arrays. The classification in [Singer et al. 2007] was performed using unsupervised hierarchical clustering and an SVM-based supervised classification method. To root the tree, we use expression data of 17 normal fat samples from the same study as well as expression data of 3 human embryonic stem cell lines (hESCs) and 3 hESC derived mesenchymal precursor lines, downloaded from

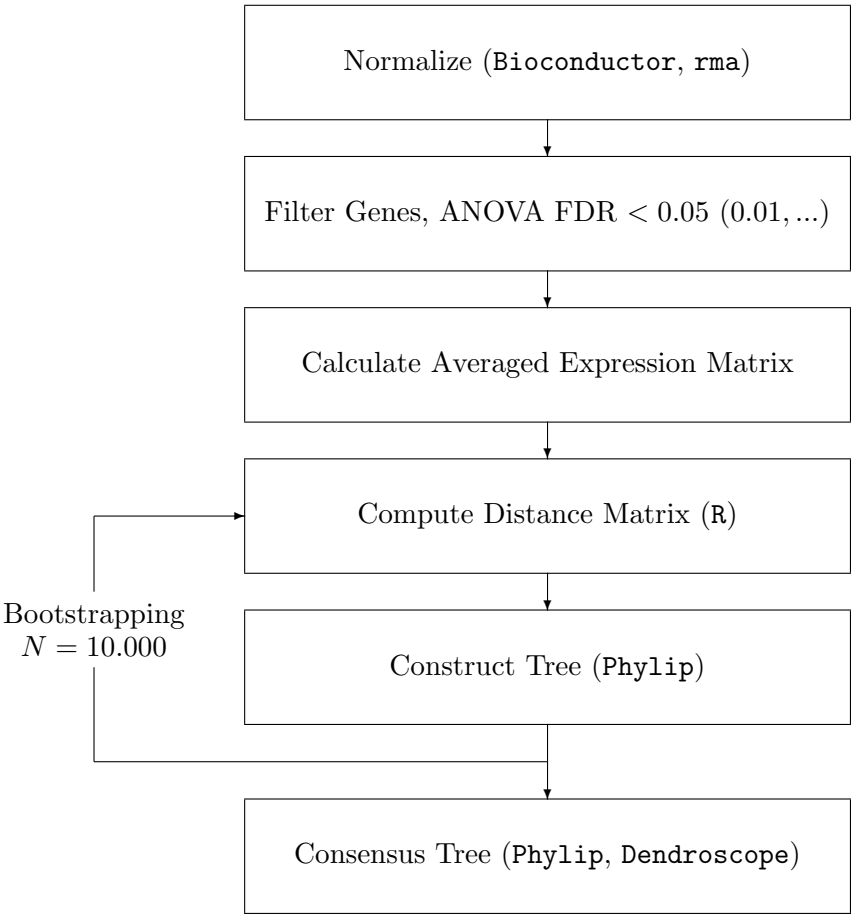


Figure 2.2. *Schematic outline of the methodology. The flow chart shows the main steps of the algorithm used to construct a phylogenetic tree of tumor subtypes. First, the data is normalized using the* `Bioconductor` *software. Then ANOVA is used to identify those genes that are differentially expressed in at least one tumor subtype; we use a False Discovery Rate (FDR) of less than 0.01. Afterwards, the expression of each differentially expressed gene is averaged across all samples of each subtype. Those average expression levels are then used to compute the distance matrix of the subtypes, which is in turn utilized to construct a phylogenetic tree using the* `Phylip` *or* `FastME` *software. To determine the consensus tree, the phylogenetic construction is repeated 10,000 times using different sets of differentially expressed genes (of varying number). The consensus tree produced with this bootstrapping approach is visualized with the* `Dendroscope` *software.*

the NCBI Gene Expression Omnibus (GEO) [Barrett et al. 2007] accession number GSE7332 [Barberi et al. 2007]. We use gene expression data of AML patient samples available within GEO (accession numbers GSE1159, GSE9476 [Stirewalt et al. 2008], GSE1729 [Gutiérrez et al. 2005], and GSE12417 [Metzeler et al. 2008]). The breast cancer dataset is also compiled from Microarray data published in GEO with dataset numbers GSE7390 [Desmedt et al. 2007], GSE2990 [Sotiriou et al. 2006], GSE3494 [Miller et al. 2005], and GSE9574 [Tripathi et al. 2008]. A problem of micrarray meta-analyses is that the different dataset sources may introduce a bias [Marot et al. 2009]. We therefore applied hierachical clustering to the compiled breast cancer dataset and did not observe a clustering according to the sources.

2.3.2 Statistical Methods and Analysis

Data preprocessing

The CEL files are normalized and summarized with the `rma` function of `Bioconductor` 2.2 [Gentleman et al. 2004, Irizarry et al. 2003, Bolstad et al. 2003]. For the phylogenetic tree construction and mainly as a strategy to remove potential noise from the data, we only consider genes that show significant differences in their expression profiles when comparing tumor subtypes. These differentially expressed genes are determined with a one-way analysis of variance (ANOVA). In addition, our `R` scripts support as alternative methods for finding differentially expressed genes the Welch approximation (`R` function `oneway.test(..., var.equal=FALSE)`) [Welch 1951] and the Kruskal-Wallis test (`kruskal.test()`) [Kruskal and Wallis 1952]. As default cutoff we choose Benjamini-Hochberg corrected p-values [Benjamini and Hochberg 1995] of 0.01. To obtain a differentiation baseline, we include expression data of normal fully differentiated tissue and, as an outgroup for the phylogenetic tree construction, the expression profile of tissue-specific stem cells. Pairwise distances of the cancer subtypes and baseline samples are computed with the Pearson Correlation Distance ($d = 1-p$) or the Euclidean Distance of the average group intensities.

Phylogenetic tree reconstruction methods

The phylogenetic trees are reconstructed with several distance-based methods. The `fitch` program includes the implementations of the Weighted Least Squares (WLS) [Fitch and Margoliash 1967] and Minimum Evolution (ME) [Rzhetsky and Nei 1993] methods, neighbor provides the Neighbor-Joining (NJ) [Saitou and Nei 1987] and UPGMA algorithms [Unweighted Pair Group Method with Arithmetic mean, Sokal and Michener 1958]. Both programs are available in the `Phylip` package version 3.67 [Felsenstein 1989]. WLS and ME are methods

designed to find the tree topology that fits the distance matrix best by optimization. The difference between these two algorithms is the optimization criterion. WLS minimizes the sum of squares error of the distances in the tree (d^T) compared to the ones in the distance matrix (δ):

$$\sum_{i,j} \frac{(\delta_{i,j} - d^T_{i,j})}{\delta^2_{i,j}}$$

The denominator thus weighs the deviations of δ from d^T for distantly related species less. As we often have very distant *in vitro* outgroups in the data, this is an important reason for us to choose WLS as the default tree reconstruction method. ME uses the same criterion to fit branch lengths to a given tree topology as WLS, but returns the topology with the smallest sum of branch lengths, not the one with the smallest sum of squares error. Another related method is Balanced Minimum Evolution (BME), implemented in the `FastME` program [Desper and Gascuel 2004]. Both WLS and BME have shown good performance on microarray data [Desper et al. 2004]. `FastME` is orders of magnitude faster than the `Phylip` implementation of WLS and thus suitable for very large datasets. It is further known to be very accurate compared to other tree reconstruction methods [Desper and Gascuel 2002; 2004, Bordewich et al. 2009]. NJ is another computationally very efficient distance-based tree reconstruction method and also popular because of its accuracy [*e.g.* Kuhner and Felsenstein 1994]. UPGMA [Sokal and Michener 1958] is a hierarchical clustering algorithm that works in a 'bottom-up' way: at the beginning, all elements form individual clusters which are consecutively combined until all elements are contained in only a single cluster. In each iteration, the pair with the smallest distance is combined into a higher-level cluster and the distance matrix is updated by calculating the distances to the newly formed cluster. The strength of the algorithm is twofold: it is computationally very efficient with both time and space complexity $\mathcal{O}(n^2)$ and it does not depend on the a priori selection of the number of clusters, in contrast to the k-means or SOMs algorithms.

Bootstrapping procedure

To assess the statistical significance of the phylogeny, the reconstruction is repeated 10,000 times with random subsets of the differentially expressed genes. We draw the bootstrap sample size n from the discrete uniform distribution on the interval $[50, N]$, where N is the total number of differentially expressed genes. Then n genes are sampled with replacement from the set of these N genes. We further bootstrap the tumor samples to incorporate the uncertainty of tumor classification. Therefore we sample for each tumor subtype n_i

experiments with replacement from the set of the n_i experiments of this subtype. Once a consensus tree is determined, it is rooted and visualized with `Dendroscope` version 2.2.2 [Huson et al. 2007].

Profile clustering

The phylogenetic tree explicitly specifies the differentiation order in the internal branch nodes. We then use the order of samples determined by the tree to calculate expression profile clusters with `mfuzz` [Futschik and Carlisle 2005], a fuzzy c-means `R` package commonly used for clustering profiles of time series. This algorithm is similar to the k-means algorithm and returns the probabilities that a gene belongs to particular expression profile cluster. As in the k-means algorithm, the number of expression profile clusters has to be set in advance and was set to 20 for the clustering of liposarcoma expression profiles in Fig. 2.7.

2.3.3 Comparison of our methodology to other clustering and dimension-reduction methods

Greedy ordering of subtypes

We use a naïve greedy algorithm in which subtypes are linearly ordered by their distance from hESC. The distance calculation and the bootstrapping are equivalent to the ones used by the phylogenetic tree reconstruction. Bootstrap values can be interpreted exactly as in the phylogenetic trees.

Self-Organizing Maps (SOMs)

SOMs [Kohonen 1990] are a type of unsupervised clustering algorithms that map high-dimensional data into a 2-dimensional grid – typically hexagonal or rectangular. The number of nodes in the grid must be set in advance, similarly to the k-means algorithm where the number of clusters is a predefined variable. The algorithm results in a two-dimensional map where similar data points tend to cluster together. SOMs are commonly applied to microarray data to cluster both genes [Tamayo et al. 1999] and tumors [Golub et al. 1999]. We calculate SOMs with the original implementation in the `SOM_PAK` version 3.1 [Kohonen et al. 1996] with the averaged group intensities of all differentially expressed genes (ANOVA FDR 0.01). We set the topology to hexagonal and choose the 'bubble' neighboring kernel.

Minimum Spanning Trees (MSTs)

MSTs are a well-established concept in graph theory. A spanning tree of a connected weighted graph G is an acyclic connected subgraph of G with the same set of vertices as G. A distance matrix can now be interpreted as a complete graph in which the edge weights correspond to the distances. The MST is the spanning tree that connects all vertices of G with the smallest sum of edge weights. MSTs have been shown to be useful for clustering and classification of microarray data [Xu et al. 2002]. For the MST calculation we use the `spantree` function of `R`, which is an implementation of Prim's algorithm [Prim 1957]. We apply this function to the Pearson distance matrix calculated again with all differentially expressed genes. A major disadvantage of this method is the lack of an established algorithm to find consensus MSTs for the resulting trees after bootstrapping, in contrast to phylogenetic trees where the availability of a wide range of methods and implementations makes it easy to summarize bootstrap results (e.g. [Margush and Mcmorris 1981, Holland et al. 2004]). Furthermore, there are no ancestral states (inner nodes) in an MST, as opposed to phylogenetic trees where subtypes are leaves in the tree and other nodes are created as ancestral states.

2.4 Results

2.4.1 Phylogenetic tree reconstruction method

We perform a systematic analysis of several methods and parameters used in our algorithm (see Methods for details). We find that combining ANOVA and Benjamini-Hochberg with a p-value of 0.01 gives good and robust results, while the Weighted Least Squares (WLS) tree reconstruction method works best when combined with the Pearson correlation matrix. Other combinations of methods give similar results and therefore should be tested in order to have an accurate understanding of a given dataset.

The phylogenetic tree resulting from this analysis contains information about the relation among subtypes as well as between subtypes and the root of the tree. The branching points represent the 'common ancestors' of the subtypes that are situated at the leaves of those branches. If the tree is rooted with expression data of a primitive cell type such as embryonic or tissue-specific stem cells, then the subtypes that are located more closely to the root correspond to types that are more similar to stem cells while the subtypes that are located farthest away from the root represent the most dissimilar types. The order of the branching points along the differentiation course can be interpreted as the ranking in dissimilarity of each of the subtypes to stem cells. The differences between stem cells and tumor subtypes are in

part caused by different differentiation status and in part by the abnormal cancer phenotype. In some situations, the order of the subtypes dictated by the tree is not unique, resulting from a non-fully balanced tree. For instance, more than one subtype can be mapped to exactly the same point in the ordering according to dissimilarity from stem cells. Furthermore, the two subtypes farthest away from the root share the same common ancestor and therefore cannot be distinguished in their level of dissimilarity. To resolve this conflict, expression data of a fully differentiated cell type can be included, which unambiguously defines the last branching point in the ranking. We validate our methodology with three datasets: (i) a dataset containing gene expression data of acute myeloid leukemia (AML) samples which are categorized according to the French-American-British (FAB) classification into classes that mirror maturation status [Bennett et al. 1976]; (ii) a dataset containing gene expression of breast cancer samples classified according to estrogen receptor status and Elston histological grade [Sotiriou et al. 2006, Desmedt et al. 2007, Miller et al. 2005]; and (iii) a dataset containing gene expression data of liposarcoma subtypes which have been analyzed for their differentiation status by comparing them to an *in vitro* differentiation time course [Matushansky et al. 2008].

Acute myeloid leukemia (AML) is a clonal disease characterized by the accumulation of myeloid progenitor cells in blood and bone marrow [Tenen 2003]. AML results from changes in transcription factor regulation that lead to a disruption of normal cellular differentiation. AML is classified into seven distinct subtypes depending on the morphology and differentiation status of tumor cells: dedifferentiated, myeloblastic, myeloblastic with maturation, promyelocytic, myelomonocytic, monocytic, and erythroleukemic AML. According to the FAB classification, these subtypes are denoted by M0, M1,..., and M6, respectively. Since AML is the result of alterations of the differentiation process, we validate our approach with a dataset of gene expression of AML patients.

Our leukemia dataset contains gene expression data of 362 AML patients and of 7 patients with unclassified Myelodysplastic Syndrome (MDS) (see Methods for details of dataset compilation) (Table 2.1). To root the AML tree, we use expression data of human embryonic stem cells (hESC); additionally, we include expression data of CD34+ hematopoietic cells from both peripheral blood (CD34 PB) and bone marrow (CD34 BM), human mesenchymal precursor cells (hESC MPC), as well as fully differentiated mononuclear cells from peripheral blood (PB) and bone marrow (BM). The surface glycophosphoprotein CD34 is expressed on undifferentiated hematopoietic stem and progenitor cells [Katz et al. 1985] and is widely used as a marker for less differentiated hematopoietic cells. We include these two subgroups as a further test of our methodology since their differentiation status is known. We use ANOVA to

Table 2.1. *French-American-British (FAB) classification of acute myeloid leukemia (AML) subtypes and numbers of samples. The table shows the names of subtypes as classified by FAB and the numbers of samples included in our study (see Fig. 2.3).*

FAB class	Name of subtype	Number of samples
M0	Dedifferentiated	14
M1	Myeloblastic	78
M2	Myeloblastic with maturation	78
M3	Promyelocytic	29
M4	Myelomonocytic	75
M5	Monocytic	78
M6	Erythroleukemic	10

identify those probe sets that are significantly differentially expressed in at least one subtype as compared to all other AML subtypes. The analysis identifies 11,105 probe sets that are differentially expressed among AML subtypes if a false discovery rate (FDR) [Benjamini and Hochberg 1995] of 0.01 is used. Use of this cutoff would lead us to expect 111 false positives. If we use the Holm correction method instead [Holm 1979], which controls the family wise error rate, then the number of differentially expressed probe sets decreases to 4,051 (with 0.01 expected false positives). The inclusion of less significantly differentially expressed genes is a potential source of noise; however, high cutoffs for significance discard genes that could be interesting for further analysis. The tradeoff between these two effects must be examined carefully to choose an appropriate cutoff. We decided to use a standard cutoff FDR of 0.01 because the tree topology remains stable for large gene sets, and also a larger number of potentially interesting genes are included which can be further filtered with other techniques.

The consensus phylogenetic tree based on this data is shown in Fig. 2.3. The order of the branching points of the subtypes coincides with the differentiation stages specified by the FAB classification: dedifferentiated AML (the M0 subtype) is located close to the stem cells while myelomonocytic (M4) and monocytic (M5) AML are located in the most distant leaves of the tree. The inner branching of the tree is also in accordance with the differentiation status suggested by the FAB classification (Table 2.1). The tree topology specifying the correct order of myeloblastic and promyelocytic maturation (M2 and M3), however, only has a moderate bootstrap value because the two subtypes are very similar in maturity. The branch leading to the erythroleukemic subtype (M6) is relatively unstable. This could be attributed to the small number of samples in this subtype or to a possible misclassification

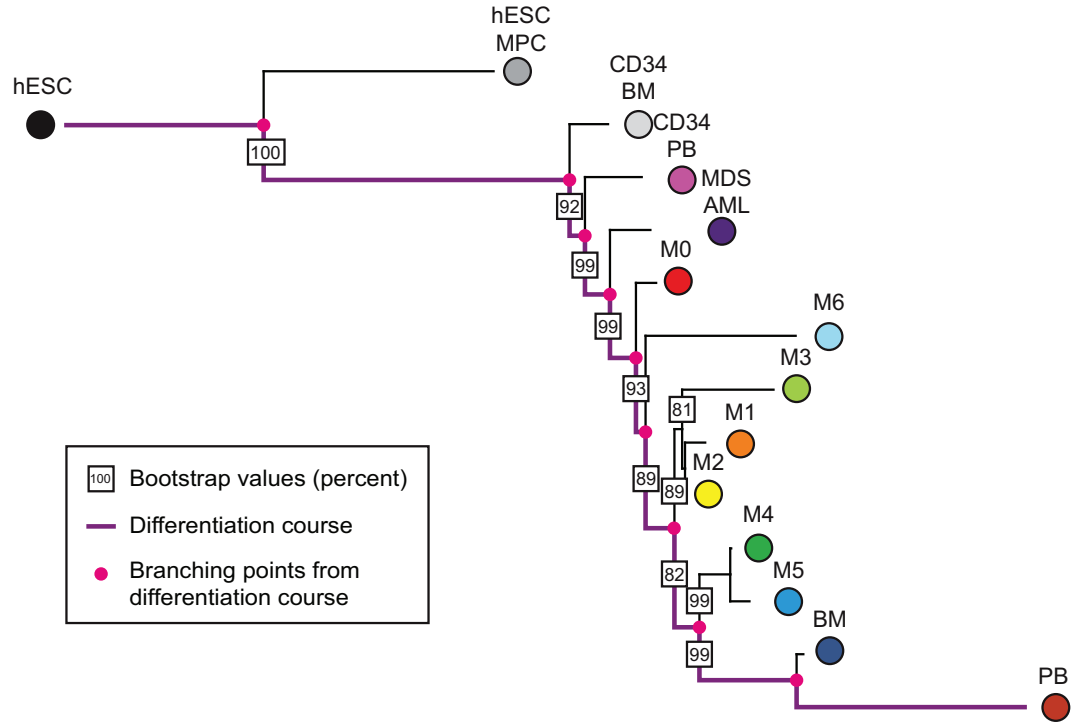


Figure 2.3. *A phylogeny of acute myeloid leukemia (AML) subtypes. According to the French-American-British (FAB) classification, AML samples are classified into seven different types according to their level of differentiation (see Table 1). Expression data from 362 AML patients and 7 Myelodysplastic Syndrome (MDS-AML) patients is used to construct a phylogeny of these leukemias. We include expression data of human embryonic stem cells (hESCs), CD34+ cells from bone marrow (CD34 BM) and peripheral blood (CD34 PB), and mononuclear cells from bone marrow (BM) and peripheral blood (PB). The differentiation pathway from hESCs to mononuclear cells from peripheral blood is represented in purple, and the common ancestors of subtypes are shown as pink dots. The bootstrap values of branches are indicated by boxed numbers, representing the percentage of bootstrapping trees containing this branch. The ranking of AML subtypes identified by the phylogenetic algorithm corresponds with the differentiation status indicated by the FAB classification. The M6 subtype, represented by only 10 samples in our dataset, has the least stable branch, leading to lower bootstrap values for those branches where it can alternatively be located.*

Table 2.2. *Breast cancer subgroups and numbers of samples. The table shows the names of the subgroups contained in the breast cancer dataset and the numbers of cancer samples as well as healthy tissue samples included in our study (see Fig. 2.4).*

Characterization of subgroup	Number of samples
Normal breast tissue (NB CA)	14
ER – Grade 3	76
ER – Grade 2	27
ER – Grade 1	3
ER + Grade 3	84
ER + Grade 2	179
ER + Grade 1	114

or erroneous diagnosis. Therefore, the position of this subtype in the tree is less certain than that of other subtypes; this uncertainty decreases the bootstrap values of the other branches at which this subtype can be located. All other branches in the tree are very stable under bootstrapping. Of central importance for the interpretation of the results is how well the tree captures the observed relationships in the data. A good measure of this fit is the average percent standard deviation of the distances between subtypes in the data compared to the ones in the tree. The Least Squares algorithm minimizes this score. For the Pearson correlation distance, the mean observed average percent deviation is 12.05%, which is a reasonable fit for this distance measure [Waddell and Kishino 2000]; hence our algorithm produces a phylogeny which accurately recapitulates the relationships seen in the data.

We also apply our algorithm to a breast cancer dataset in order to study the performance of our method using cancers with epithelial origin. The samples in our dataset were characterized by immunochemistry methods according to their estrogen receptor status (ER+ and ER-) and Elston histologic grade (G1, G2, and G3). We compile a total of 483 unique samples, among which we find all combinations of ER status and grade (Table 2.2). The raw data was analyzed as described in the methods section. We root the tree with human mesenchymal stem cells and also include samples of normal breast [Tripathi et al. 2008]. Results are shown in Fig. 2.4. We find 17,966 probes differentially expressed between the subgroups when using ANOVA with Benjamini-Hochberg correction and a cutoff value of 0.01. A negative ER status has been shown to correlate with poor prognosis [Osborne et al. 1980]. Consistent with this observation, our algorithm places ER-negative subgroups closer to stem cells, reflecting the more stem-like properties of these aggressive tumors, while ER+ tumors are placed closer

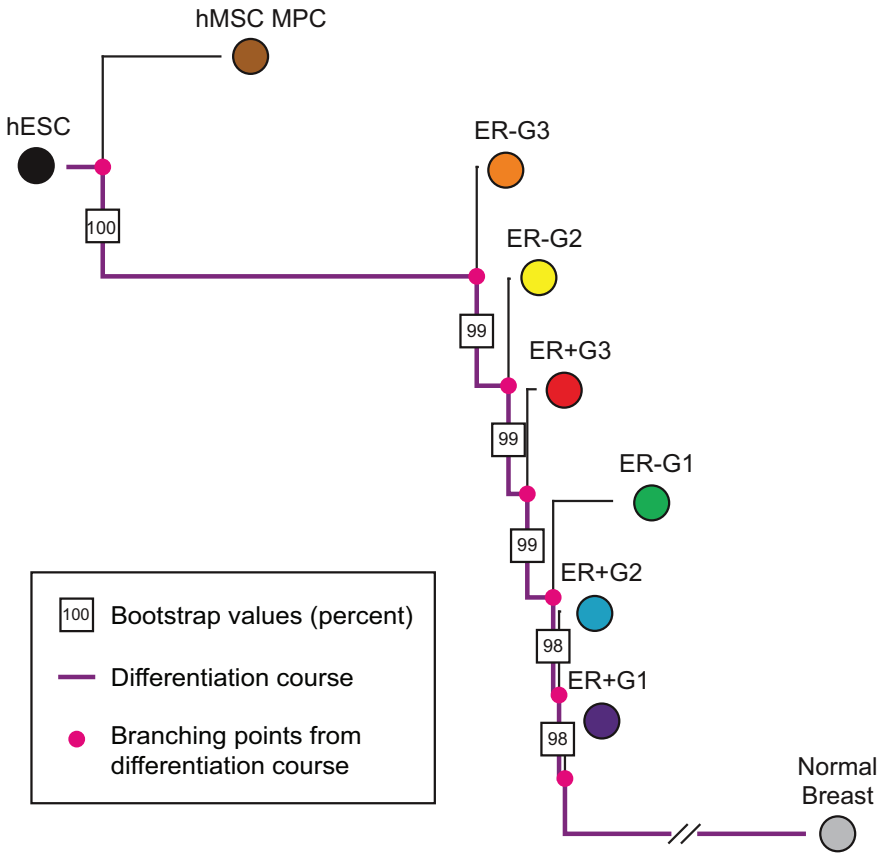


Figure 2.4. *A phylogeny of breast cancer subgroups. The figure shows the consensus tree of breast cancer subgroups. We use expression data of 483 breast cancer samples subdivided as shown in Table 2. The tree is rooted with expression data of human mesenchymal stem cells (hMSCs). We also include expression data of fully differentiated normal breast tissue. The differentiation pathway from hESC to fully differentiated breast tissue is indicated in purple, and the pink dots represent the common ancestors of (sets of) subgroups. The boxed numbers specify the bootstrap values of branches. The phylogeny ranks the breast cancer subtypes according to their dissimilarity from stem cells as ER- grade 3, ER- grade 2, ER+ grade 3, followed by ER- grade 1, ER+ grade 2 and ER+ grade 1.*

to the normal breast tissue samples. Tumor grades are ordered similarly, placing tumors of higher grade closer to stem cells. Most trees reconstructed with the different sets of genes have the same topology (bootstrap values close to 100%), reflecting a very robust topology. We conclude that our methodology is also able to accurately rank tumors of epithelial origin according to maturity.

Next we construct a phylogeny of liposarcoma subtypes. Liposarcoma is the most common type of soft tissue sarcoma accounting for about 20% of all tissue sarcomas [Mack 1995]. In 2008, 10,390 new cases of sarcoma were reported in the US [American Cancer Society 2008]. Surgery is the standard care for localized tumors but leads to worse prognoses in cases of locally advanced or disseminated disease [Singer et al. 2007]. Liposarcomas are classified into three biological types encompassing five subtypes: (i) well-differentiated/dedifferentiated, (ii) myxoid or round cell, and (iii) pleomorphic liposarcoma, based on morphological features and cytogenetic aberrations [Sandberg 2004]. Although the subtype is the main determinant of clinical outcome [Kooby et al. 2004, Singer et al. 2003, Nakayama et al. 2007, Barretina et al. 2010, Sekiya et al. 2004], liposarcomas of similar morphology can differ in response to treatment and in prognosis [Singer et al. 2007]. Microscopically well-differentiated liposarcoma is composed of relatively mature adipocytic proliferation showing significant variation in cell size and at least focal nuclear atypia. Histologically dedifferentiated liposarcoma is represented by the transition from well-differentiated liposarcoma to non-lipogenic sarcoma. Both well-differentiated and dedifferentiated liposarcomas contain characteristic ring or giant marker chromosomes with 12q14-15 amplification. Myxoid liposarcomas contain uniform round to oval shaped primitive non-lipogenic mesenchymal cells and a variable number of small signet-ring lipoblasts in a prominent myxoid stroma. Round cell tumors are characterized by solid sheets of primitive round cells with no intervening myxoid stroma. Pleomorphic liposacoma is a pleomorphic high grade sarcoma containing a variable number of pleomorphic lipoblasts.

Recently, progress has been made towards a classification of liposarcoma subtypes utilizing gene expression data. In 2007, a 142-gene predictor was identified which correctly distinguishes between liposarcoma subtypes and generates a set of differentiation-related genes that may contain candidate therapeutic targets [Singer et al. 2007]. In 2008, Matushansky et al. showed that the main liposarcoma subtypes can be ranked according to their differentiation status by comparing gene expression data of the tumor subtypes with the genes expressed during normal *in vitro* adipogenic differentiation [Matushansky et al. 2008]. The ranking generated by the latter approach is useful for validating our methodology.

Our liposarcoma dataset includes 180 surgical samples that have been pathologically classified

as 61 dedifferentiated, 52 well differentiated, 26 pleomorphic, 18 round cell, and 23 myxoid liposarcomas [Singer et al. 2007, Barretina et al. 2010]. Samples that were likely misclassified were filtered in previous studies, which is a pre-processing step critical for the outcome of the algorithm. For a FDR of the ANOVA filter of 0.01 after correction with the Benjamini-Hochberg method, we find 13,429 probe sets that are differentially expressed among the liposarcoma subtypes. Those sets are then used to construct an unrooted phylogenetic tree. To root the tree, we use expression data of mesenchymal stem cells [Graham et al. 2002] and fully differentiated adipocytes. The resulting consensus tree is shown in Fig. 4a. The tree topology is stable with bootstrap values larger than 85%. Based on the consensus tree, the subtypes can be ordered by increasing dissimilarity from stem cells as dedifferentiated, pleomorphic, myxoid/round-cell, and well-differentiated liposarcoma (Fig. 4a). This order coincides with experimental results based on the gene expression observed during *in vitro* differentiation published earlier (Fig. 4b) [Matushansky et al. 2008]. By setting the p-value threshold of the Holm correction to 0.01, we obtain 7,290 differentially expressed probe sets; these probe sets generate a tree topology that is identical to the case described above with bootstrap values larger than 91.5% (data not shown). When rooting with embryonic stem cells, the branching between embryonic stem cells and the rest of the tree is less stable since the expression of embryonic stem cells differs considerably from all other samples (data not shown). To increase the stability of the tree, it is preferable to root with an outgroup that is relatively closely related to the investigated samples (in this case, mesenchymal stem cells; see also Sec. 2.4.3) [Graham et al. 2002]. Again we test how well the tree fits the distance matrix and observe a mean average percent standard deviation of 11.3%, which has been reported to be a good fit for the Pearson correlation distance [Waddell and Kishino 2000]. Therefore, our methodology is also able to rank liposarcoma subtypes in the correct order according to their dissimilarity to stem cells.

Since our methodology correctly ranks leukemia, breast cancer, and liposarcoma samples according to their differentiation status, we now investigate a larger number of sarcoma subtypes to identify their relationship in maturity as well as candidate targets for therapeutic intervention. The sarcoma dataset includes the 180 liposarcomas discussed above as well as 36 myxofibrosarcomas, 5 pleomorphic malignant fibrous histiocytomas (MFH), 7 lipomas, and 23 leiomyosarcomas (Table 2.3). We use expression data of both mesenchymal stem cells and embryonic stem cells to root the tree. The consensus tree is shown in Fig. 2.5. Our methodology determines that leiomyosarcoma is closest in its differentiation status to stem cells, followed by MFH and myxofibrosarcoma, and finally the liposarcoma subtypes (ranked as determined above) and the benign subtype lipoma. The algorithm also clusters the subtypes according to tissue of origin, predicting that leiomyosarcoma branches before all other subtypes, and

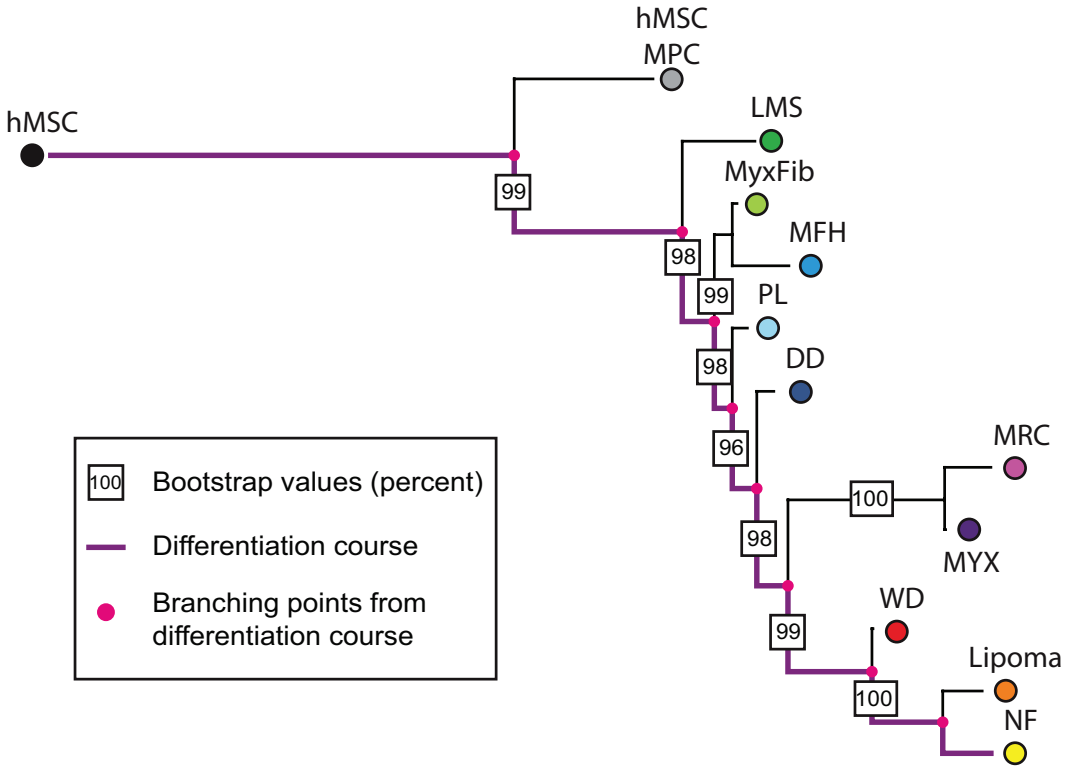


Figure 2.5. *A phylogeny of sarcoma subtypes. The figure shows the consensus tree of sarcoma subtypes. We use expression data of 251 sarcoma samples classified into the types shown in Table 2. The tree is rooted with expression data of human embryonic stem cells (hESCs). We also include expression data of human mesenchymal stem cells (hMSC) and of fully differentiated normal adipocytes. The differentiation pathway from hESC to fully differentiated adipocytes is indicated in purple, and the pink dots represent the common ancestors of (sets of) subtypes. The boxed numbers specify the bootstrap values of branches. The phylogeny ranks the sarcoma subtypes according to their dissimilarity from stem cells as leiomyosarcoma, malignant fibrous histiocytoma, myxofibrosarcoma, followed by the liposarcoma subtypes dedifferentiated liposarcoma, pleomorphic, myxoid/round-cell, and well-differentiated liposarcoma. Lipoma is identified as the subtype most dissimilar from stem cells.*

Table 2.3. *Sarcoma subtypes. The table shows the number of sarcoma subtypes included in our study (see Fig. 2.5)*

Tissue	Name of subtype	Number of samples
Fat	Dedifferentiated	61
	Pleomorphic	26
	Round-cell	18
	Myxoid	23
	Well-differentiated	52
	Lipoma	7
Smooth Muscle	Leiomyosarcoma	23
Fibrous Tissue	MFH	5
	Myxofibrosarcoma	36

that MFH and myxofibrosarcoma have a common ancestor; so do all liposarcoma subtypes and lipoma. Note that although pleomorphic liposarcomas and MFH/myxofibrosarcomas are very similar subtypes at the level of their genetic copy number aberrations [Barretina et al. 2010], our algorithm places them in different branches of the tree. This effect is a result of the phenotype-based nature of our method and is in accordance with the different tissues of origin of these subtypes. The tree has a very stable topology with bootstrap values larger than 0.90 except for the MFH subtype, which exhibits a lower bootstrap value of 0.60; this value is likely due to the small number of samples (5) available for this subtype. Note that with the current dataset, we cannot distinguish between the case in which the subtype located most closely to stem cells, leiomyosarcoma, is situated on the adipocytic differentiation path and the case in which leiomyosarcoma is alternatively located on a branch leading to fully differentiated tissue of another type. To resolve this ambiguity, gene expression data of fully differentiated tissue of all the types giving rise to sarcomas is needed.

We are interested in identifying genes that are related to adipogenesis, i.e. those genes that correlate with adipocyte differentiation. To identify such genes, we cluster our list of differentially expressed genes into a chosen number of groups depending on their expression pattern in sarcoma subtypes. When the subtypes are arranged according to their distance from stem cells (as indicated by the tree in Fig. 2.6a), the expression of some genes continuously increases from the less differentiated to the more differentiated subtypes, while the expression of other genes decreases or exhibits more complicated patterns (Fig. 2.7). We hypothesize that genes whose expression continuously increases or decreases are possibly related to gain

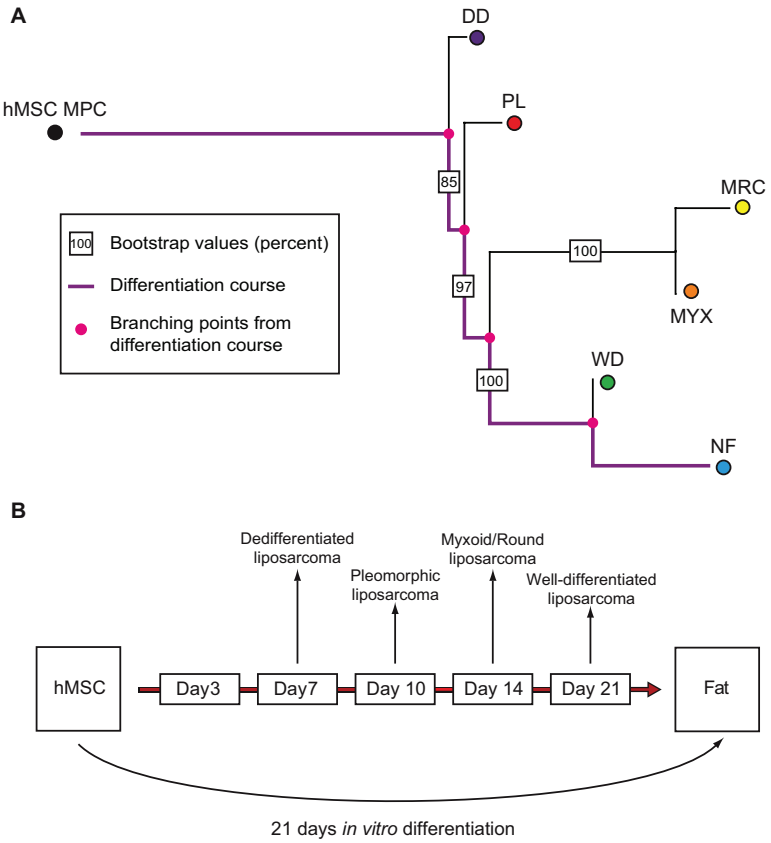


Figure 2.6. *A phylogeny of liposarcoma subtypes. (a) The figure shows the consensus tree of liposarcoma subtypes. The tree is rooted with expression data of human mesenchymal stem cells (hMSC), and expression data of normal fat cells is included as well. The differentiation pathway from hMSC to normal fat cells is represented in purple. The pink points represent common ancestors of (sets of) subtypes. The boxed numbers specify bootstrap values of branches. The tree indicates that dedifferentiated liposarcoma is most similar to stem cells, followed by pleomorphic, myxoid, round-cell, and finally well-differentiated liposarcoma. (b) The figure shows a schematic representation of the correlation of adipogenesis to liposarcoma differentiation. In Matushansky et al. [2008], human mesenchymal stem cells were differentiated in vitro to produce fat cells, and gene expression was measured for five different time points during the differentiation. The expression data of four different liposarcoma subtypes was then compared to the data obtained from the differentiation time course. This comparison identified dedifferentiated liposarcoma as the subtype most similar to stem cells, followed by pleomorphic, myxoid/round-cell, and well-differentiated liposarcoma. The correspondence between the results of our algorithm applied to gene expression datasets and these experimentally derived results serves as a validation of our methodology. Adapted from Matushansky et al. [2008].*

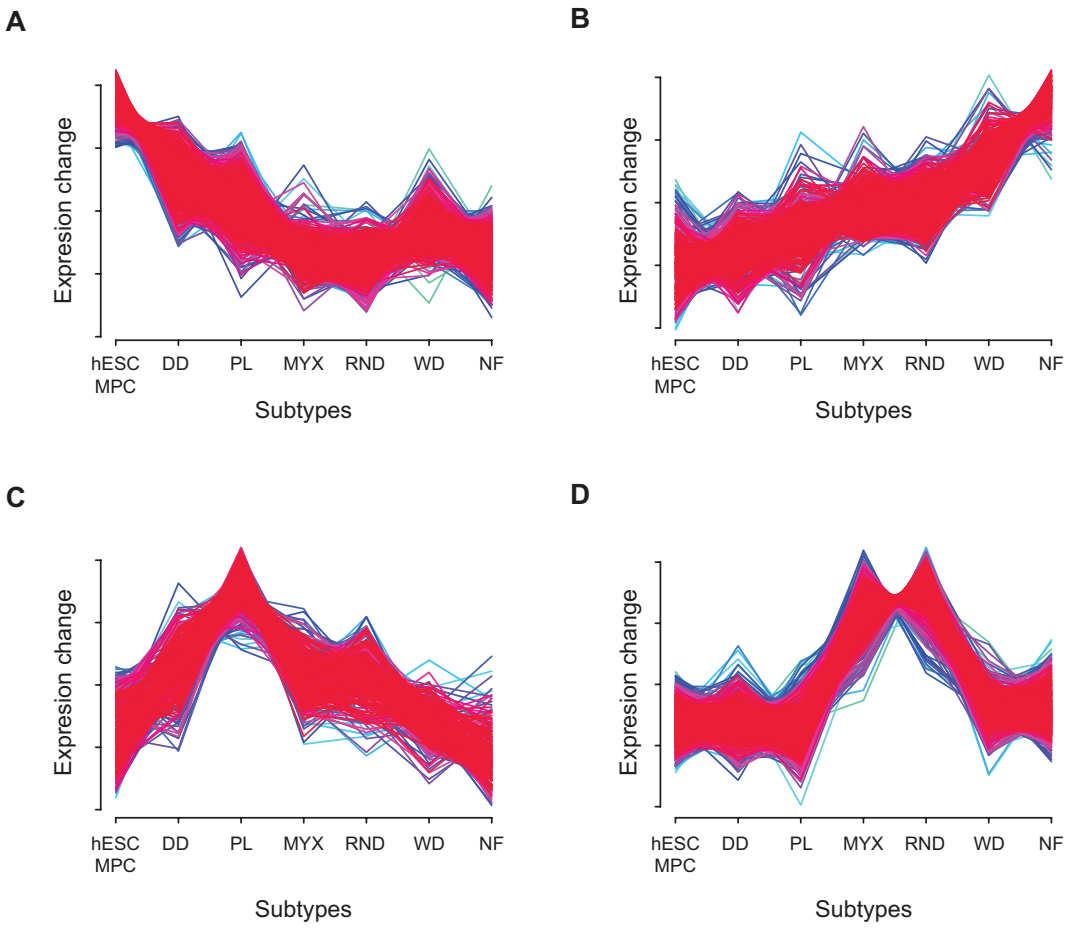


Figure 2.7. *Clusters of gene expression profiles. The figure shows four example groups of differentially expressed genes clustered according to their expression profiles (see Methods section for details on the clustering algorithm). On the horizontal axis, we show the liposarcoma subtypes ordered according to the ranking identified by the phylogenetic approach (see Fig. 2.6a) and in the vertical axis the corresponding standard normalized average expression values of the subtypes. We also include human embryonic stem cells (hESCs) and normal fat cells. The expression of some genes continuously decreases from less differentiated samples (hESC, dedifferentiated liposarcoma, ...) to more differentiated samples (..., well-differentiated liposarcoma, normal fat) (a), while the expression of other genes increases (b). Other genes are overexpressed in just a single liposarcoma subtype (c) or in a subset of subtypes (d). Those genes whose expression continuously increases or decreases are hypothesized to be related to adipogenesis (see Table 2.4).*

of the features of differentiation and loss of stem cell-associated functions, even though this association with maturation may not be causative. To test this hypothesis, we compare the genes whose expression increases or decreases along the order of subtypes to previously published lists of adipocytic differentiation-specific genes [Matushansky et al. 2008, Sekiya et al. 2004]. In these two studies, mesenchymal stem cells were differentiated *in vitro* into normal fat cells, and the expression profiles of cells were measured at multiple time points during the differentiation process. An investigation of genes whose expression levels changed statistically significantly along the differentiation time course led to the identification of 67 and 69 genes, respectively [Matushansky et al. 2008, Sekiya et al. 2004]. These genes are thought to be related to adipocytic differentiation.

We rank the genes whose expression increases or decreases along the liposarcoma subtypes (see Fig. 2.7 for example clusters) according to the fold change between their expression in human mesenchymal stem cell (hMSC) and in normal fat. Among the 11,105 probe sets obtained by the ANOVA filtering with FDR of 0.01 after Benjamini Hochberg correction, the top 25 genes in this ranking are listed in Table 4. About 64% of these genes coincide with the published lists [Matushansky et al. 2008, Sekiya et al. 2004]. These results suggest that our methodology is able to identify differentiation-related genes from the large number of differentially expressed genes. Additionally to the previously identified genes, our method identified other genes that have not been associated with adipocytic differentiation (Table 2.4). For instance, the protein phosphatase inhibitor 1 (PPP1R1A) is thought to be important in the control of glycogen metabolism and is primarily expressed in liver cells; the tyrosine kinase NTRK2 is part of a signaling pathway leading to neuronal differentiation, and the metabolism related enzyme system ACACB is exclusively expressed in adipocyte tissue.

2.4.2 Comparison of tree reconstruction methods to other algorithms

We compare the results obtained from phylogenetic tree reconstruction algorithms with other methods of data clustering and organization such as a simple greedy algorithm (in which subtypes are linearly ordered by their distance from hESC), self-organizing maps (SOMs), and minimum spanning trees (MSTs) (see the Methods section for details of the algorithms). When applying the greedy algorithm to our AML dataset, we find similar results to those produced by the tree reconstruction analysis (Fig. 2.8a). Although the correspondence between the results of this method and the reconstructed phylogenetic tree is very good, the former only contain information of a linear organization, as opposed to the richer information that can be extracted from the tree topology and branch lengths. An example of a self-organizing map (SOM) algorithm applied to the AML dataset is shown in Fig. 2.8b. Subtypes that are

Table 2.4. *Adipogenesis-related genes. The table shows 25 genes (represented by 28 probe sets) whose expression continuously increases or decreases from less differentiated to more differentiated samples as ranked in Fig. 2.7. The genes are ordered according to their fold change in expression between mesenchymal stem cells and normal fat cells. These genes are related to adipogenesis. About 64% of those genes have previously been reported in Matushansky et al. [2008] and Sekiya et al. [2004] (marked with [a] and [b], respectively).*

Gene Symbol	Gene Name	Fold Change
FABP4[ab]	fatty acid binding protein 4, adipocyte	352.1
LPL[ab]	lipoprotein lipase	164.3
ADH1B[ab]	alcohol dehydrogenase 1B (class I), beta polypeptide	150.1
HBA/B	hemoglobin	147.0
ADIPOQ[a]	adiponectin, C1Q and collagen domain containing	137.2
RBP4[ab]	retinol binding protein 4, plasma	104.0
GOS2[b]	G0/G1switch	85.6
FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	78.3
SORBS1[a]	sorbin and SH3 domain containing 1	72.0
PLIN[ab]	Perilipin	68.1
PRKAR2B[a]	PRKAR2B a protein kinase, cAMP-dependent, regulatory, type IIb	53.1
CHRDL1[a]	chordin-like 1	52.0
APOD[a]	apolipoprotein D	49.9
PPP1R1A	protein phosphatase 1, regulatory (inhibitor) subunit 1A	41.4
GHR	growth hormone receptor	41.4
AOC3[ab]	amine oxidase, copper containing 3 (vascular adhesion protein 1)	40.8
CLEC3B	C-type lectin domain family 3, member B	38.1
DPT[a]	dermatopontin	37.0
NTRK2	neurotrophic tyrosine kinase, receptor, type 2	36.5
PALMD	palmdelphin	34.1
ACACB	acetyl-Coenzyme A carboxylase beta	32.2
LEP[a]	leptin	28.8
VWF	von Willebrand factor	28.1
TIMP4[b]	TIMP metallopeptidase inhibitor 4	26.7
COL11A1[ab]	collagen, type XI, alpha 1	-11.7

known to be similar are mapped close together on the grid – e.g. human embryonic stem cells (hESC), mesenchymal stem cells (MSC), and samples with markers of poor differentiation (BMCD34 and CD34PB). Unfortunately, the overall organization of a SOM strongly depends on the shape and size of the grid, making it difficult to interpret the results in a robust and useful way for our purposes. Finally, we calculate a maximum spanning tree (MST) for the AML dataset (Fig. 2.8c). This algorithm accurately reproduces the reconstructed tree found with our original method, with the exception of mesenchymal stem cells being placed at the edge of the tree (instead of embryonic stem cells).

2.4.3 Systematic analysis of methods and parameters

We compare the different methodologies implemented in our algorithm for each step of the analysis in order to identify those methods and parameters that perform well in the analysis of our datasets. We apply our algorithm to all datasets using all combinations of the following methods and parameters: for finding differentially expressed genes: ANOVA, Kruskal-Wallis (KW) and Welch approximation (Welch); two methodologies for p-value correction: Benjamini-Hochberg (BH) and Holm; two p-value cutoffs: 0.01 and 0.05; five tree reconstruction and clustering algorithms: Weighted Least Squares (WLS), Minimum Evolution (ME), Neighbor-Joining (NJ), FastME, and Average Linkage (UPGMA); and two distance measures: Pearson correlation and Euclidean distance. The topologies found among the different combinations of parameters show that WLS, Pearson correlation, and BH with a cutoff value of 0.01 perform accurately in accordance with the AML, breast cancer, and liposarcoma datasets (*data not shown*).

Note that two main assumptions of the UPGMA algorithm are not fulfilled by cancer subtype data, namely: all species originate from a common ancestor and they all have evolved at the same pace. This issue explains why this method fails to reconstruct the right tree topologies; for example, in all sarcoma UPGMA topologies, some liposarcoma subtypes branch together with leiomyosarcoma, which is thought to arise from smooth muscle tissue.

It has been shown in previous studies that, in general, WLS performs better than NJ when trees have long external or internal branches [*e.g.* Bruno et al. 2000]. Note also that the use of Euclidean distance leads to less robust results than the use of Pearson correlation when trees with long branches are considered. For example, when the Euclidean distance method is applied to the liposarcoma data, the dedifferentiated and pleomorphic subtypes cluster together with the well-differentiated subtype and normal fat. The effect of long branches on the Euclidean distance method becomes even more pronounced when analyzing the sarcoma data; in this case, the least common topologies are observed only when the Euclidean distance

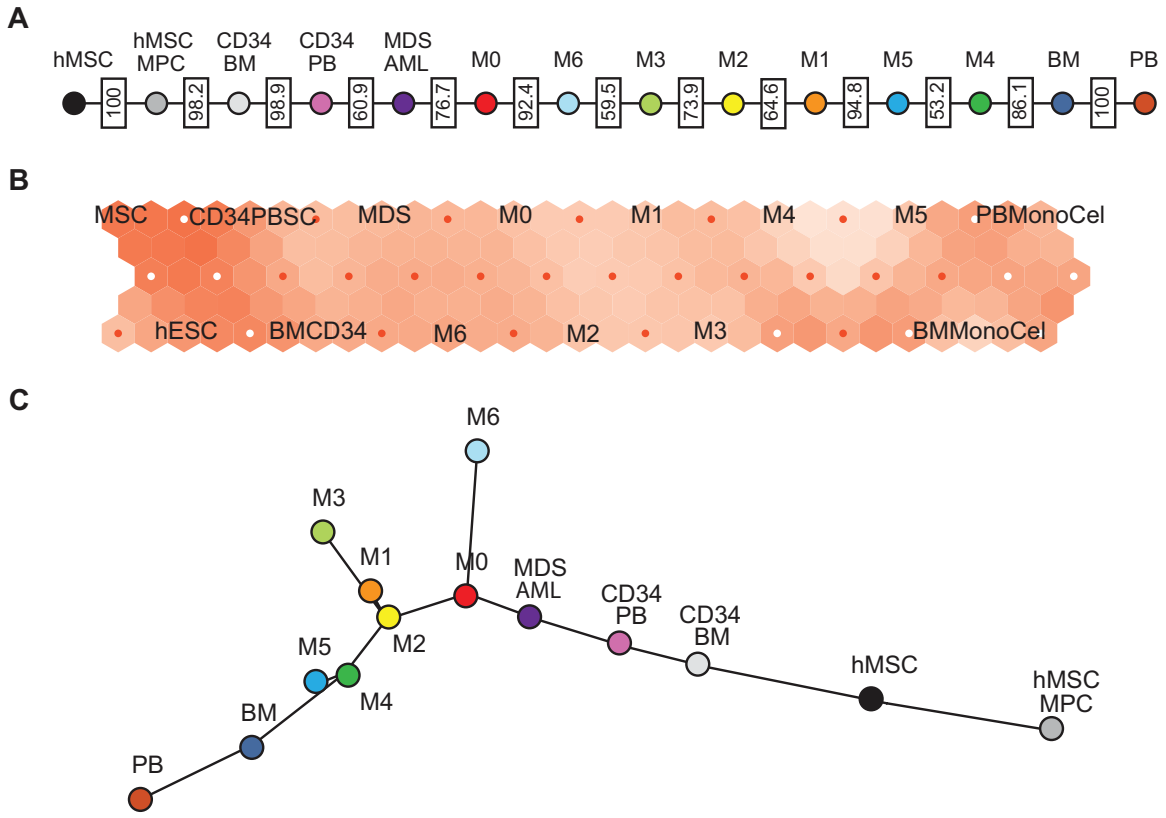


Figure 2.8. *Alternate distance based methods applied to acute myeloid leukemia (AML) data. (a) The figure shows the results of a simple algorithm that sorts the AML subtypes by their distance to hESC. The algorithm uses the same distances as the ones for the phylogenetic tree shown in Fig. 2.3. (b) Self-Organizing Maps. The AML subtypes are arranged on a hexagonal grid of 15 x 3 nodes. These nodes are visualized by the small red or white dots. The colors visualize the difference of neighboring nodes. For example, the light nodes surrounding M4 and M5 show that these subtypes are similar. MSC and CD34+ peripheral blood, however, show very different expression patterns despite the fact that they are ordered close together on the map. (c) Minimum Spanning Tree (MST) calculation of the Pearson correlation matrix of the AML dataset.*

method is used. If distant subgroups (i.e. hMSC and hMSC MPC) are removed from the analysis, then most parameter combinations including the Euclidean distance method favor the correct topology.

We do not observe a significant influence of the choice of the method on the identification of differentially expressed genes. More important for our data was the choice of the p-value cutoff. The results of our study suggest that BH with a cutoff of 0.01 is a good compromise, but we recommend investigating the effects of using different cutoff values.

In general, all tree reconstruction methods are very fast, especially since the number of different tumor subtypes in our analysis is typically limited. So it is possible to test many parameters in a reasonable time and we recommend doing so.

2.5 Discussion

We have presented a rational methodology to investigate the dissimilarity between cancer subtypes and stem cells. Our approach uses gene expression data of tumor samples which have been classified into histological subtypes as well as expression data of an 'evolutionary outgroup' such as embryonic stem cells, tissue-specific stem cells, and/or fully differentiated normal cells. The data of tumor subtypes is used to identify the genes that are differentially expressed among the subtypes, and those genes, together with data of the outgroup, allows construction of a phylogeny of cancers. Our algorithm estimates the statistical significance of the tree branches by bootstrapping, a repeated tree construction using a varying number of randomly chosen genes. The distance between the branching points of the tumor subtypes and the stem cells specifies their dissimilarity, which is caused in part by differences in maturity, and ranks the subtypes according to increasing differentiation. This ranking is then used to identify genes whose expression continuously changes depending on the degree of maturation.

Our methodology is validated by being able to correctly reproduce experimental results concerning the relationship in differentiation status of liposarcoma, breast cancer and AML subtypes [Bennett et al. 1976, Matushansky et al. 2008, Tenen 2003] and concerning genes related to adipocytic differentiation [Matushansky et al. 2008, Sekiya et al. 2004]. Our method is useful for identifying genes that are overexpressed in some tumor subtypes (Fig. 2.7c). For instance, genes whose expression is increased in a particular tumor type but not in normal tissue-specific stem cells and differentiated cells may represent candidates for targeted therapy, possibly with lessened side effects. Interestingly, some of the genes found to be differentially expressed in only one or a few liposarcoma subtypes can be targeted by currently available drugs. It will be an important next step to test those genes for a causal role in tumorigenesis.

In recent years, bioinformatic tools have been widely used to analyze the vast amount of data produced experimentally. In analyses of microarray data, simple algorithms for phylogenetic tree reconstruction, such as Average linkage (UPGMA) [Sokal and Michener 1958], produce rooted bifurcating trees and are routinely applied to visualize similarities in gene expression. The most prominent example for this type of analysis are heatmaps, a graphical representation of the clustered expression matrix where colors represent the measured gene intensities; a dendrogram is often added which shows the bifurcating tree best describing the differences in gene expression [Eisen et al. 1998]. Another important application of such algorithms is the clustering of tumor samples for improving or discovering subtype classification [*e.g.* Kapp et al. 2006]. Other more sophisticated tree reconstruction algorithms are only rarely applied to expression data [Waddell and Kishino 2000, Desper et al. 2000; 2004, Desper and Gascuel 2004, Planet et al. 2001, Uddin et al. 2004, Nugoli et al. 2003]. The 'molecular clock' assumption of UPGMA (specifying that changes occur at a constant rate, [Kimura 1968]) renders this algorithm inappropriate for our investigation. Other algorithms such as Maximum Parsimony, Neighbor-Joining (NJ) [Saitou and Nei 1987], or Least-Squares [Fitch and Margoliash 1967] enable us to root the tree and to estimate the differentiation status of tumor subtypes by a simple comparison of the lengths between the root of the tree and the branching points of the leaves. We do not use character-based methods such as Maximum Parsimony due to the necessity of artificially discretizing the continuous values of gene expression intensities.

The estimation of evolutionary distances between tumors from gene expression data is hindered by the fact that small differences in the biology of tumors may cause large differences in gene expression. Examples of such situations are given by genes which trigger the expression of cascades of other genes [Planet et al. 2001] and mutational events affecting the expression of several genes [Park et al. 2009]. Park et al. [2009] proposed the use of correction methods that estimate mutational distances from the observed expression distances. This approach represents an interesting new avenue to further explore in future work.

The phylogeny of tumor subtypes identified by our methodology cannot be used to reconstruct the evolutionary history of a single tumor sample. The fact that dedifferentiated liposarcomas, for example, branch earlier than well-differentiated liposarcomas is not to be taken as evidence that one subtype evolved into the other. Rather, it specifies the dissimilarity of the bulk of tumor cells between cancer subtypes from stem cells at the time of observation. Similarly, our methodology cannot be used to identify the cell of origin of a tumor type. Both the position of a subtype in a differentiation-based phylogeny and the similarity of a subtype to an *in vitro* differentiation time course provide information about the bulk of tumor cells only; to determine whether these cells are produced from tumor stem cells which arose from

cells of similar, earlier or more complete differentiation stages is outside the scope of this approach. Furthermore, the ability of a phylogenetic tree to reconstruct evolutionary trajectories when applied to genetic data rests on the assumption that the genetic material records the evolutionary history of the system. In the case of phenotypic information such as gene expression data, this assumption does not hold, and hence any information about the origin of the investigated cancer subtypes cannot be obtained.

The generality of our approach and the extensive availability of high-quality input datasets (e.g. GEO) makes this methodology a unique tool to investigate differentiation-related genes and the relationship in maturity of cancer subtypes. The use of data from patient samples reduces the problems encountered with *in vitro* studies regarding the reproducibility of the results in other systems and their significance to *in vivo* situations.

3

Wild Pedigrees

3.1 Introduction

Molecular markers such as highly polymorphic short tandem repeats (STRs), *e.g.* microsatellites [Queller et al. 1993], and more recently also diallelic single nucleotide polymorphisms (SNPs) [Glaubitz et al. 2003, Anderson and Garza 2006] are now routinely used to genotype individuals in natural populations. The reconstruction of genealogical relationships among diploid species by means of molecular markers has been an active field of research for more than three decades. A well-developed statistical theory of paternity inference has been published in series of articles by E.A. Thompson [e.g. Thompson 1976]. The study of parentage in natural populations was the topic of the pioneering papers by Meagher and Thompson [1986] and Marshall et al. [1998] and is recently reviewed in Blouin [2003], Jones and Ardren [2003], Pemberton [2008], Jones et al. [2009]. The pedigree structure of a sample of individuals is important for a wide range of ecological, evolutionary and forensic studies. Applications include genealogy reconstruction [e.g. for wine grape cultivars Vouillamoz and Grando 2006], the estimation of heritabilities in the wild [Thomas and Hill 2000], and victim identification [Lin et al. 2006].

In order to reconstruct the pedigree of a sample, the parents of each individual in the sample

need to be determined. If one has a large amount of genomic data, the task of identifying first degree relationships, i.e., parent-offspring and full-sibling relations, is trivial. Unfortunately, many datasets in natural populations do not contain enough information to unambiguously determine the parents. Another problem is that datasets often contain only a subset of a population. Thus, one or both parents of an observed individual may be missing from the dataset. Furthermore, many datasets are not free of errors.

Pedigree reconstruction has an especially long history in flowering plant populations, see e.g. Ellstrand and Marshall [1985] and Meagher and Thompson [1987]. It has been used mainly to find correlations between phenotypes and reproductive success, or to estimate pollen-mediated gene flow [Smouse and Meagher 1994, Burczyk et al. 1996, Smouse et al. 1999, Meagher et al. 2003, Wright and Meagher 2004]. To a lesser extent, parentage inference and related methods are used to estimate recent rates of self-fertilization (selfing) in a population [Ritland and Jain 1981, David et al. 2007, Wilson and Dawson 2007, Jarne and David 2008].

Pedigree reconstruction in clonal populations has received very little attention so far, although such an approach holds the promise to allow the direct inference of gene flow from a population's pedigree in particular in long-living clones with limited rate of sexual reproduction. It is a much harder problem than classical paternity or parentage inference for two main reasons: First, it is typically difficult to estimate the age of a clonal plant [Ally et al. 2008] so age data is often not available. Second, while it is normally easy to estimate the number of individuals (*ramets*), N_r, in a clonal plant population over the occupied space, it is typically very hard to estimate the number of different genotypes (*genets*), N_g. The genotype number, N_g, usually is a required input parameter in most software for the estimation of the statistical significance of a parentage. We will later demonstrate that both restrictions can be overcome at least in principle, however.

In this chapter, we will present a new software package `FRANz`. We will start with a short overview of exisiting tools and methods in Sec. 3.2. We then give a short review of the statistical frameworks typically used in parentage analyses in natural populations in Sec. 3.3. In Sec. 3.4 we explain our program `FRANz` in detail and in Sec. 3.5 we will apply our method to empirical and simulated data.

3.2 The molecular ecologist's tools of the trade

A molecular ecologist who wants to use parentage analyses has to choose between a vast amount of published tools. The reason therefore is simply that there are a lot different sampling schemes, i.e., the fraction of sampled offspring and maternal or paternal genotypes

in the population normally differs a lot between studies and tools are often specialized for a particular sampling scheme. In same cases for example, it is possible to get DNA data from almost all mating individuals in the population [*e.g.* Nielsen et al. 2001, Hadfield et al. 2006]. Sometimes mothers are known, for instance in plants when seeds are attached to their mothers. Or maybe it is known that groups of offspring are half- or full-sibs, for example in amphibian egg masses. Tools that support these kinds of *prior knowledge* will infer parentages with higher accuracy than tools that do not. Unfortunately, many niche tools lack basic features like a good handling of genotyping errors or mutations. Thus, many ecologists use standard tools and do not take advantage of their prior knowledge. On the other hand, if some tools assume a particular prior knowledge, *e.g.* paternity inference tools which assume knowledge of the maternal genotypes, then ecologists cannot use these tools in a straightforward way when this information is not available. Things look especially bleak for researchers who want to study the mating behaviour of polyploid or haplo-diploid species as almost no established tool supports non-diploid species.

In this section, we will give a very brief overview of the most important tools. For more comprehensive reviews, see Jones and Ardren [2003], Jones et al. [2009]. The flow chart in Fig. 3.1 gives an overview of the supported features of the tools presented in the following.

3.2.1 Sibship inference and parental reconstruction

Most programs support only datasets comprising one or two generations. The approach to partial pedigree reconstruction in one generation datasets are sibship algorithms. Here, genotype data is used to infer full-sib and half-sib relationships [Smith et al. 2001, Thomas and Hill 2002, Wang 2004b, Berger-Wolf et al. 2007]. The most popular tools are `COLONY` [Wang 2004b] and `PEDIGREE` [Smith et al. 2001]. Since version 2.0, `COLONY` can infer parentage jointly with sibships [Wang and Santure 2009].

The *parental reconstruction* technique is possible when family sizes are rather large ($> 8-10$ offspring), half- or even full-sib groups are known *a priori* and markers are highly polymorphic. Then parental genotypes can be reconstructed and thus the mating behaviour of unsampled individuals inferred. `GERUD` [Jones 2005] uses an exhaustive algorithm that tests all possible parental genotypes and guarantees to find the minimum number of parental genotypes to explain the data. The first version required that the maternal genotype is known, but this limitation is removed in the current version 2.0. The `PARENTAGE` [Emery et al. 2001] program uses a Markov chain Monte Carlo (MCMC, see Sec. 3.3.7) approach to sample parental genotypes. It might be the program of choice when markers are less informative [Jones et al. 2009].

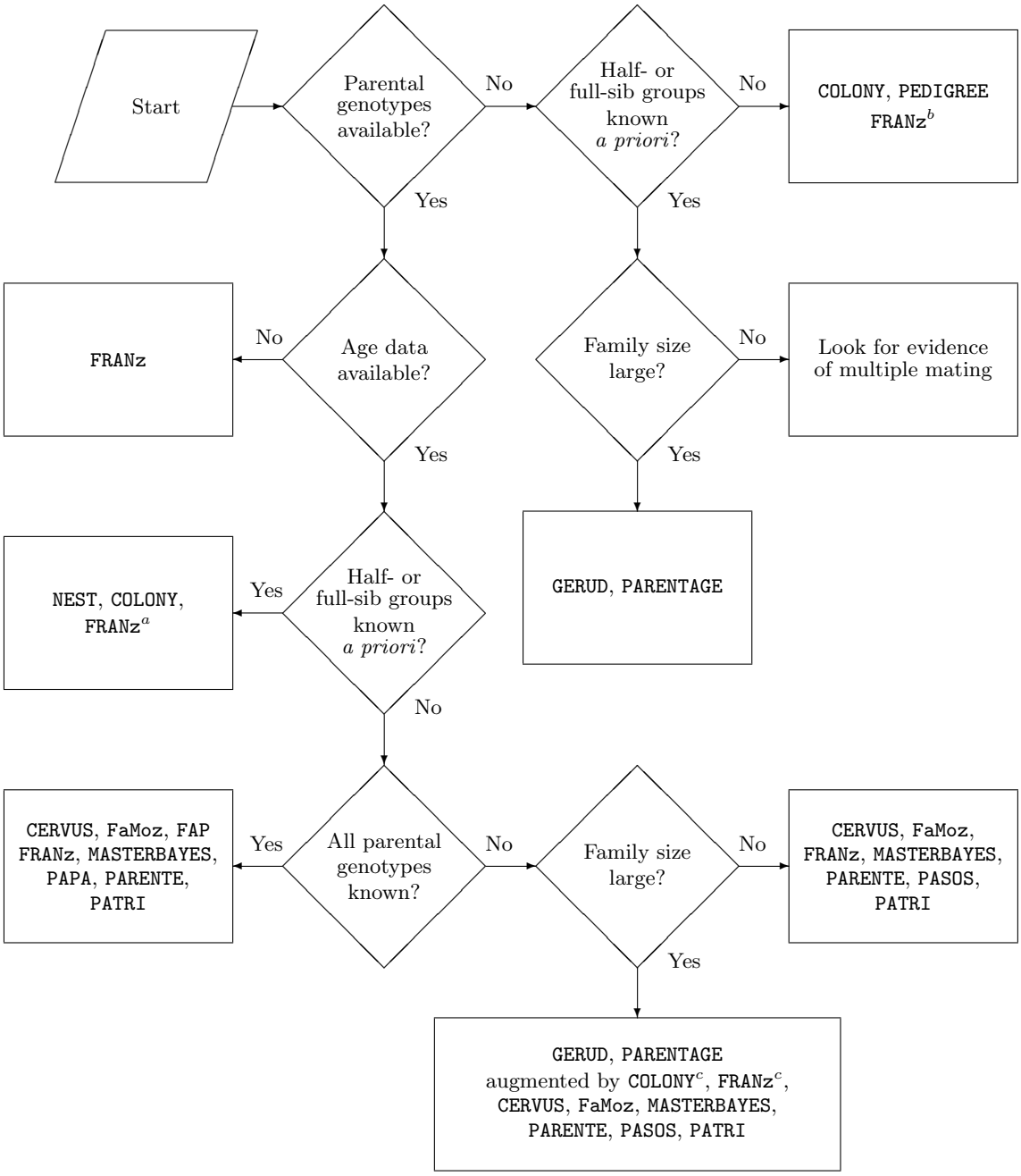


Figure 3.1. *A flow chart showing the available tools for various types of sampling schemes. See Sec. 3.2.1 and 3.2.2 for references and short introductions of the tools. We mainly focused here on tools that were recomended in a recent review [Jones et al. 2009]. This review introduces also a magnitude of other niche tools or prototype implementations. Adapted and extended from Jones et al. [2009].*

[a] only when one sex is monogamous; [b] calculates full-sib p-values; [c] can infer parent-offspring and full-sib relationships and should give in general better results

3.2.2 Parentage and paternity inference

Most classical parentage or paternity inference tools require age data; the absence of age data makes an *a priori* ordering of individuals in generations impossible. Such an ordering is assumed by traditional paternity inference software and also dramatically restricts the pedigree search space. So these programs typically take an offspring list, if known their mothers, and a list of candidate parents or fathers as input and output the most likely parentage for each offspring.

The most popular tool is `CERVUS` [Marshall et al. 1998, Kalinowski et al. 2007] which started as a paternity inference tool but supports now also parentage inference. `CERVUS` estimates the statistical significance of the parentages by simulation, which requires knowledge of the number of unsampled candidate parents and the genotyping error rate. Then it compares the observed likelihood distributions in the simulation (Sec. 3.3.4) with the likelihood scores in the data and *assigns* parentages if these scores exceed a given threshold. Only assigned parentages are considered in downstream analyses. Parentage inference is restricted to rather small datasets in the current version 3.0.3. Apart from the fact that `CERVUS` proved its robustness in hundreds of studies, one important reason for the popularity of this software is its graphical user interface and its comprehensive allele frequency analysis. Concerning features, very similar tools are `PASOS` [Duchesne et al. 2005] and `FaMoz` [Gerber et al. 2003]. The latter is one of the very few tools that support dominant markers, such as *amplified fragment length polymorphisms* (AFLPs). The counterpart of `PASOS` for closed systems, i.e, when all parental genotypes are known, is `PAPA` [Duchesne et al. 2002]. `FAP` [Taggart 2007] is another tool for closed systems.

An alternative to the simulation-based methods are tools that use Bayesian frameworks. `PARENTE` [Cercueil et al. 2002] was one of the first parentage inference tools. In contrast to `CERVUS` and `FaMoz`, it generates the lists of candidate parents internally by analyzing the specified years of birth and death and the age ranges in which individuals can reproduce. `PARENTE` assigns parentages by their posterior probabilities (we will explain this later in detail in Sec. 3.3.7 and 3.4.1). Another tool that uses a Bayesian framework is `PATRI` [Signorovitch and Nielsen 2002]. It only supports paternity inference. As two special features, it can estimate the male population size and can compare the reproductive success of groups of males. The `MASTERBAYES R` package [Hadfield et al. 2006] uses a similar *full probability* approach. Instead of using only assigned parentages, it estimates parameters of interest jointly with parentages. For example in Hadfield et al. [2006], the authors investigate how the distance d of the sampling locations between offspring and father affects the probability of paternity. They model the probability of paternity as an exponential function $\exp^{-d\lambda}$ and

estimate means and quantiles of the parameter λ in a MCMC approach simultaneously with the pedigree. `MASTERBAYES` was written for datasets where almost all candidate parents were sampled, but recent versions now support also incomplete sampling. It is one of the very few tools that still get regular updates. Finally, another tool with a Bayesian framework is `NEST` [Jones et al. 2007] which is an approach of parentage analysis for nest structured data.

3.2.3 Multigenerational pedigree reconstruction

Much less attention [e.g. in Almudevar 2003; 2007, Fernández and Toro 2006, Koch et al. 2008, Cowell 2009] compared to parentage and paternity inference or sibship algorithms has been given to multigenerational pedigrees in which the offspring and candidate parent sets are not necessarily non-overlapping. This is the case for example in the absence of age data. Then the ordering of genotypes into generations is not known a priori and has to be estimated from the genotype data only. Thus, at difference with parentage inference programs, this general case does not admit all possible parentage combinations as valid pedigrees. The task is therefore to find the parentage combinations that define the *maximum likelihood pedigree*. If the number of possible pedigrees is too large to enumerate, heuristics are necessary. So far, a flexible software package has not been available that allows the incorporation of prior information in addition to the genotypes and that is robust in the case of errors. We have implemented a package called `FRANz` which fills this gap and which we will introduce in detail in this chapter. As parentage inference and paternity inference are easier as age or even mothers are known, such a package has a much wider scope of application than the classical tools. Fig. 3.1 shows that most other tools only work with a very limited number of sampling schemes.

3.3 Background

3.3.1 Pedigrees

The core of our pedigree reconstruction tool `FRANz` is a probabilistic model calculating the likelihood of a given *pedigree*. Let us start with the necessary definitions.

A pedigree $\mathscr{P} = (V, A)$ is an acyclic digraph with vertex set V and arc set A, where the vertices represent the individuals and the arcs the parent-offspring relationships (Fig. 3.2). Thus V represents the set of all genotyped individuals in the sample. For an arc (u_i, v), we say that v is a *child* of u_i and u_i is a *parent* of v. The set of *parents* of v in $\mathscr{P}$ is denoted by $N^+(v) \subseteq V$; this set may contain two elements, $\{u_i, u_j\}$, one element, $\{u_i\}$, or none, $\emptyset$. In the latter case, v is called a *founder*. In selfing species, $u_i = u_j$ is allowed and $\mathscr{P}$ thus becomes a

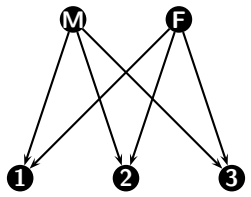


Figure 3.2. *A simple example pedigree $\mathscr{P} = (V, A)$ in a digraph representation. The vertex set V represents the individuals, here a family with a mother, a father and three offspring. The arc set A represent the parent-offspring relationships.*

multigraph. With $N^-(u) \subseteq V$ we denote the *offspring* of u.

3.3.2 Genotypes

The pedigree in Fig. 3.3 shows the *genotypes* of the believed remains of the Romanov family [Gill et al. 1994, Coble et al. 2009]. Short tandem repeat (STR) data is available from four different *loci*. A locus is a fixed position in a genome and genetic markers such as STRs are loci that have a variable content in a population. The different possible contents are called *alleles*. SNPs are mostly diallelic which means that one part of the population has for example the nucleotide Cytosine and the other part a Thymine at a particular SNP locus. So we have the alleles T and C at this locus. STRs in contrast are highly polymorphic with many more alleles. They are short (2-6 base pair) repeated DNA motifs, *e.g.* $[GT]_n$ and the alleles are the observed n, the repeat numbers. Highly polymorphic loci typically show high mutation rates and are *neutral*, i.e., mutations [*e.g.* Levinson and Gutman 1987] have no influence on the fitness of the individual. In *diploid* species, every individual has two homologous copies of each chromosome, typically one inherited from the mother and one from the father. As a consequence, we can observe two alleles at a given locus (Fig. 3.3). If both alleles are equal, we call the locus *homozygous*, if they are different *heterozygous*.

For a given individual i, we denote in the following an observed single-locus genotype by g_i and its multi-locus genotype by G_i.

3.3.3 Mendelian segregation probability

The Mendelian segregation probability is the probability that an offspring of F_i and M_j has the genotype G_O. In the pedigree in Fig. 3.3 for example, both parents have at the first locus the same genotype 15.16 and as an offspring randomly inherits one allele from each parent, the possible offspring genotypes are 15.15, 15.16, 16.15 and 16.16. As the haplotypes are almost always unknown, i.e., it is not known which allele was inherited from the mother and

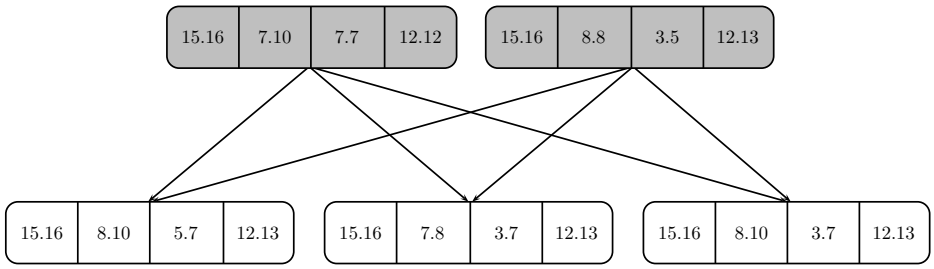


Figure 3.3. *The example pedigree of Fig. 3.2 showing parental (gray) and offspring (white) alleles at four different loci. The data are STR genotypes of skeletons found in Ekaterinburg, Russia and are believed to be the remains of the Romanov family [Gill et al. 1994, Coble et al. 2009].*

which from the father, the order is ignored and the probability of the heterozygous genotype 15.16 is therefore 0.5. The well-known general equation for triples (offspring and two putative parents) is:

$$\delta(a_{o1}.a_{o2}, a_{p1}.a_{p2}) = \begin{cases} 1 & \text{if } a_{o1}.a_{o2} = a_{p1}.a_{p2} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{T}(a_{o1}.a_{o2}|a_{m1}.a_{m2}, a_{f1}.a_{f2}) = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}\delta(a_{o1}.a_{o2}, a_{mi}.a_{fj}) \tag{3.1}$$

For multi-locus genotypes, we assume throughout this chapter that the loci are *unlinked*, i.e., that the loci are inherited independently during meiosis. Multi-locus segregation probabilities are thus calculated by multiplication of all n_L single-locus probabilities:

$$\mathrm{T}(G_o|G_m, G_f) = \prod_{i}^{n_L}\mathrm{T}(g_{oi}|g_{mi}, g_{fi}) \tag{3.2}$$

We can also calculate the segregation probabilities for offspring-parent pairs. By calculating the population allele frequencies and by assuming Hardy-Weinberg equilibrium, we have [Meagher and Thompson 1986]:

$$\begin{aligned} \mathrm{T}(a_i.a_j|a_i.a_i) &= \Pr(a_j), \\ \mathrm{T}(a_i.a_i|a_i.a_j) &= 0.5\Pr(a_i), \\ \mathrm{T}(a_i.a_j|a_i.a_k) &= 0.5\Pr(a_j), \\ \mathrm{T}(a_i.a_j|a_i.a_j) &= 0.5\left(\Pr(a_i) + \Pr(a_j)\right). \end{aligned} \tag{3.3}$$

Here, $\Pr(a_i)$ denotes the allele frequency of the allele a_i. Explicit equations for $\mathrm{T}(\cdot)$ that tolerate genotyping errors are derived in Sec. 3.3.5.

3.3.4 LOD Scores

Consider a triplet of individuals (*A, B, C*) with single locus genotypes g_A, g_B and g_C. In likelihood-based paternity analyses, one compares the likelihood of the hypothesis (H_1) that the three individuals are offspring, mother and father, with the likelihood of the alternative hypothesis (H_2) that the three individuals are unrelated. This comparison is usually expressed as a log-ratio, the difference in log-likelihood (LOD) score [e.g. Meagher and Thompson 1986]:

$$\mathrm{LOD}(g_A, g_B, g_C) = \log \frac{\Pr(g_A, g_B, g_C|H_1)}{\Pr(g_A, g_B, g_C|H_2)} = \log \frac{\mathrm{T}(g_A|g_B, g_C)}{\Pr(g_A)} \tag{3.4}$$

The likelihood of (H_2) is the probability of observing the three genotypes when randomly drawn from a population in Hardy-Weinberg equilibrium. For diploid heterozygotes, the probability of a genotype with the alleles a_1 and a_2 and with the allele frequencies p and q is $\Pr(a_1, a_2) = 2pq$; for homozygotes, we have $\Pr(a_1, a_1) = p^2$. The Mendelian segregation probability is again denoted by $\mathrm{T}(\cdot)$.

If one parent, typically the mother is known, we have [Meagher and Thompson 1986]:

$$\mathrm{LOD}(g_A, g_B, g_C) = \log \frac{\mathrm{T}(g_A|g_B, g_C)}{\mathrm{T}(g_A|g_B)} \tag{3.5}$$

LOD scores in Maximum Likelihood pedigree reconstruction are important because only parentages with positive LOD score need to be considered, because adding the arcs of a parentage with negative LOD score would decrease the pedigree likelihood.

In parentage and paternity inference, LOD scores are also often used of for assessing the confidence of the parentage with the largest LOD score. Marshall et al. [1998] use ΔLOD as test statistic, which is the difference of the LOD scores between the two most likely parentages. The critical value of this test statistic is obtained by simulation. Gerber et al. [2003] use the LOD scores as test statistic.

3.3.5 Genotyping Errors

If a single-locus genotype of an offspring does not share one allele with each of the candidate parents, we call this a *mismatch*. In true parent(s)-offspring pairs or triples, we will observe mismatches only in case of genotyping error or in case of somatic mutations [Bonin et al. 2004]. The latter case is even for markers with relatively high mutation rates such as microsatellites

rather rare, but even high quality datasets contain genotyping errors. Genotyping errors occur when the genotype determined by molecular analysis does not correspond to the real genotype. For instance, a common type of genotyping error in microsatellite datasets are null alleles, which are often the result of a mutation in the primer annealing site. Thus it is unwise to exclude a parent immediately when observing such a mismatch.

The model implemented in `FRANz` defines an error to be the replacement of the true genotype at a particular locus in an individual with a random genotype. This leads to a modification of the expressions for the LOD score [Kalinowski et al. 2007].

In the following, we present the likelihood formulas for paternity or parentage inference of the typing error model described in Kalinowski et al. [2007]. We corrected some typos in the original version presented in the appendix of Kalinowski et al. [2007] and also simplified the formulas where possible. $L(H_1)$ is the likelihood of the hypothesis H_1 that the alleged parent is the true parent; the alternative hypothesis H_2 is that the alleged parent is unrelated. We follow the notation of the original paper instead of ours: the single locus genotypes of mother, alleged father and offspring are denoted with g_m, g_a and g_o (corresponding to g_f, g_m and g_o in our notation). $\Pr(g)$ is again the probability of observing the genotype g in a population in Hardy-Weinberg equilibrium (for heterozygotes $\Pr(a_1.a_2) = 2pq$; for homozygotes, we have $\Pr(a_1.a_1) = p^2$ where p and q are the allele frequencies of a_1 and a_2, respectively). The estimated typing error rate is the probability that one or both alleles of an genotype are not correctly observed and is denoted as ϵ. Finally, $\mathrm{T}(\cdot)$ denotes again the Mendelian segregation probability. For details see Marshall et al. [1998], Kalinowski et al. [2007]. The likelihoods for paternity when the mother is unknown are:

$$\begin{aligned} L(H_1) &= \Pr(g_a)\{(1-\epsilon)^2 \mathrm{T}(g_o|g_a) + \epsilon(1-\epsilon)2\Pr(g_o) + \epsilon^2 \Pr(g_o)\} \\ &= \Pr(g_a)\{(1-\epsilon)^2 \mathrm{T}(g_o|g_a) + \epsilon(2-\epsilon)\Pr(g_0)\} \\ L(H_2) &= \Pr(g_a)\{\Pr(g_o)\} \end{aligned} \tag{3.6}$$

In the curly bracket of the likelihood of H_1, the terms consider the three possible cases that both, one or no genotypes are correctly observed.

The likelihoods for paternity and maternity jointly are

$$\begin{aligned} L(H_1) = \Pr(g_m)\Pr(g_a)\{&(1-\epsilon)^3 \mathrm{T}(g_o|g_m, g_a) + \epsilon(1-\epsilon)^2 [\mathrm{T}(g_o|g_m) + \mathrm{T}(g_o|g_a) + \Pr(g_o)] \\ &+ \epsilon^2(3-2\epsilon)\Pr(g_o)\} \\ L(H_2) = \Pr(g_m)\Pr(g_a)\{&\Pr(g_o)\} \end{aligned} \tag{3.7}$$

The likelihood of the alternative hypothesis H_2 for paternity when the mother is known is:

$$L(H_2) = \Pr(g_m)\Pr(g_a)\{(1-\epsilon)^3 \mathrm{T}(g_o|g_m) + \epsilon(1-\epsilon)^2[\mathrm{T}(g_o|g_m) + 2\Pr(g_0)] + \epsilon^2(3-2\epsilon)\Pr(g_o)\} \tag{3.8}$$

$L(H_1)$ is the same as for the parentage inference case (Eq. 3.7).

If some parent-offspring relationships are known, then a simple formula for estimation of typing error rate e_l at locus l is given in Marshall et al. [1998]:

$$e_l \approx \frac{1}{2P_l} \cdot \frac{m_l}{M_l} \tag{3.9}$$

where P_l is the exclusion probability [Jamieson and Taylor 1997, see Sec. 3.4.7] at locus l, m_l the number of mismatches in M_l comparisons. If the number of comparisons is too low or if the error rate is constant across n loci, a better estimate is:

$$e \approx \frac{1}{n}\sum_{l}^{n} e_l \tag{3.10}$$

3.3.6 IBD coefficients

For each pair of individuals, we can calculate the probability that the two have a particular relationship $\mathbb{R}$: unrelated $\mathbb{U}$, parent-offspring $\mathbb{PO}$, full-sib $\mathbb{FS}$, half-sib $\mathbb{HS}$, etc. The usual way of calculating the likelihoods $\Pr(g_A.g_B|\mathbb{R})$ uses the so-called *IBD (identical by descent) coefficients* k_0, k_1 and k_2. Alleles are identical by descent if they are identical and are segregated from a recent common ancestor. A child, for example, shares with each parent exactly one allele that is identical by descent ($k_1 = 1$); monozygotic twins share two ($k_2 = 1$) whereas unrelated individuals share no alleles ($k_0 = 1$) identical by descent. For full-sibs, it is easy to show that the probability that they share one allele identical by descent is 0.5 and that they share no or two is in both cases 0.25 (so $k_0 = 0.25$, $k_1 = 0.5$ and $k_2 = 0.25$). Given the allele frequencies, the probabilities that the genotype pair $g_A.g_B$ shares 0, 1 or 2 alleles identical by descent, P_0, P_1 and P_2, are then calculated and are inserted in the final IBD likelihood formula [for details, see e.g. Blouin 2003]:

$$\Pr(g_A.g_B|\mathbb{R}) = k_0P_0 + k_1P_1 + k_2P_2 \quad (k_0 + k_1 + k_2 = 1) \tag{3.11}$$

For unlinked loci, the logarithms of the IBD scores are again additive over the loci.

3.3.7 Bayesian MCMC

Bayes' Theorem

The Bayes' Theorem is best explained with an example. Assume that a casino uses five different sets of dice. Now four of these sets are fair and one of them is unfair. Let the probability of throwing a 6 with an unfair dice be 1/3. The dealer randomly picks one set and then the player throws $(6, 6)$. By applying Bayes' Theorem, one can easily calculate the probabilty that the player got the unfair set. The Theorem states that:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{3.12}$$

Here, A and B are two *dependent* events (*e.g.* A = *player got an unfair set of dice* and B = *throwing* $(6, 6)$). $\Pr(A)$ is often called the *prior probability* or just the *prior* of A and is in the casino example the probability that the dealer picks the unfair set, which is 1/5. $\Pr(B)$ is called the *marginal probability* of B and is here the probability of throwing $(6, 6)$ with both the fair and the unfair set. $\Pr(A|B)$, the conditional probability of A given B, is called the *posterior probability* of A. The posterior probability in this example is the probability that the player got an unfair set after seeing his throw $(6, 6)$:

$$\begin{aligned}\Pr(\text{unfair}|6,6) &= \frac{\Pr(6,6|\text{unfair})\Pr(\text{unfair})}{\Pr(6,6)} \\ &= \frac{(1/3)^2 \cdot 1/5}{(1/3)^2 \cdot 1/5 + (1/6)^2 \cdot 4/5} \\ &= 0.5\end{aligned}$$

Thus, $\Pr(\text{unfair}|6, 6)$ and $\Pr(\text{fair}|6, 6)$ are equal in this example, the experiment gives no hint whether the player got the fair or unfair set.

So typically we have some observed data D and a discrete set of n hypotheses H where a hypothesis is often a set of r parameters $H_i = (\beta_1, ..., \beta_r)$. Then by applying Bayes' Theorem one can calculate the posterior of a particular hypothesis H_i:

$$\Pr(H_i|D) = \frac{\Pr(D|H_i)\Pr(H_i)}{\sum_j^n \Pr(D|H_j)\Pr(H_j)} \tag{3.13}$$

So all is needed is $\Pr(D|H_i)$, the *likelihood function* which calculates the likelihood of H_i given the observation D. However, the number of hypotheses is often very large. For example assume we want to reconstruct a phylogeny of m species from which we have DNA data D. Further assume we have a likelihood function that calculates the likelihood of a tree $\mathscr{T}_i$, $\Pr(D|\mathscr{T}_i)$. Now we want to calculate the posterior probability of a particular tree.

However, the number of trees, n, we need to consider in the denominator grows extremely fast [Felsenstein 2003]:

$$n = \frac{(2m-5)!}{2^{m-3}(m-2)!}$$

For ten species, we already have to calculate the likelihood of over 2 million different trees, for 13 species there are more than 13 billion trees. Thus an exhaustive enumeration and evaluation of all hypotheses is computationally intractable in most cases of interest.

Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a technique that allows indirect sampling from the posterior distribution $\Pr(H_i|D)$. It does so by generating a sample path $H_1, ..., H_t$ where each parameter $H_j = (\beta_{1j}, ...\beta_{rj})$ has the posterior distribution $\Pr(H_i|D)$. Bayesian MCMC thus uses the posterior distribution to estimate means and quantiles of the parameters β. Although a complete overview of MCMC is beyond the scope of this thesis, we will later use and explain the two most common MCMC algorithms, the Metropolis-Hastings method and Gibbs sampling. For an introduction to MCMC, see for example Fishman [2005].

3.4 Methods

In the following, we will describe the main steps `FRANz` does to reconstruct the pedigree and how this software estimates the statistical significance of the procedure (see Fig. 3.4).

3.4.1 Likelihood Model

Using Bayes' Theorem, we calculate the posterior probability that the female F_i and the male M_j are the parents of O,

$$\Pr(G_{Fi}, G_{Mj}|G_O, G_F, G_M, A, N_m, N_f) = \frac{\mathrm{T}(G_O|G_{Fi}, G_{Mj}) \Pr(G_{Fi}, G_{Mj})}{\Pr(G_O)} \tag{3.14}$$

where G_O, G_F, and G_M are the offspring, candidate maternal, and paternal genotypes, A the population allele frequencies, N_m the total number of breeding males (alleged fathers) in the population. Correspondingly, N_f is the total number of candidate mothers. The symbol $\mathrm{T}(\cdot)$ denotes again the Mendelian segregation probability, $\Pr(G_{Fi}, G_{Mj})$ is the prior probability of F_i and M_j being parents of O and $\Pr(G_O)$ is the marginal probability of the offspring genotype.

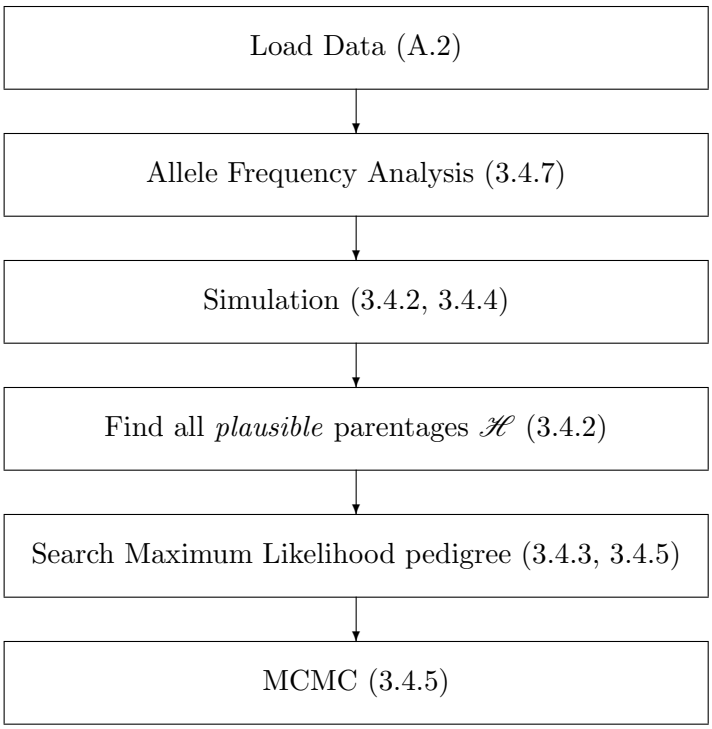


Figure 3.4. *Schematic outline of the methodology. The flow chart shows the main steps of the* `FRANz` *algorithm. First, the input files as described in Appendix A.2 are parsed. Then the allele frequencies are analyzed (see Sec. 3.4.7). In simulations, mismatch (Sec. 3.4.2) and IBD distributions (Sec. 3.4.2) are generated. In the next step, for every offspring* `FRANz` *finds all plausible parentages. These are all parentages that do not exceed the mismatch and posterior probability thresholds (Sec. 3.4.2). Then,* `FRANz` *searches for the Maximum Likelihood pedigree, either by an exact algorithm or by Simulated Annealing (Sec. 3.4.3, 3.4.5). Finally we do an MCMC sampling to estimate the statistical significance of the parentages (Sec. 3.4.5).*

By assuming equal priors, it can be computed as

$$\Pr(G_O) = \frac{1}{N_m}\left[\sum_{k}^{n_m} \mathrm{T}(G_O|G_{Fi}, G_{Mk}) + (N_m - n_m)\ \mathrm{T}(G_O|G_{Fi}, A)\right] \tag{3.15}$$

provided the mother F_i is already known [Nielsen et al. 2001]. The first term is the sum of segregation probabilities of all n_m sampled candidate paternal genotypes, the second term accounts for the unsampled ones. This second term is the probability, given the population's allele frequencies, that a random genotype is the true father, weighted by the number of unsampled candidates. For the case that the mother is unknown, we may write

$$\begin{aligned}\Pr(G_O) = \frac{1}{N_f N_m}\Bigg[&\sum_{k}^{n_m}\sum_{l}^{n_f} \mathrm{T}(G_O|G_{Mk}, G_{Fl}) \\ &+ (N_m - n_m)\sum_{l}^{n_f} \mathrm{T}(G_O|G_{Fl}, A) + (N_f - n_f)\sum_{k}^{n_m} \mathrm{T}(G_O|G_{Mk}, A) \\ &+ (N_m - n_m)(N_f - n_f)\Pr(G_O|A)\Bigg]\end{aligned} \tag{3.16}$$

with n_f denoting the number of sampled female candidates. Here, the first term again collects the segregation probabilities of all sampled male and female genotypes. The following two sums are the cases that either the true father or the true mother are unsampled. The last term accounts for the case that both parents are unsampled, which is the probability of observing the offspring genotype in a population with allele frequencies A, weighted by the number of unsampled pairs.

In monoecious plant populations, where all individuals can mate with each other and with themselves, we finally have:

$$\begin{aligned}\Pr(G_O) = \frac{2}{N_g(N_g+1)}\Bigg[&\sum_{k=1}^{n_g}\sum_{l=k}^{n_g} \mathrm{T}(G_O|G_k, G_l) \\ &+ (N_g - n_g)\sum_{k}^{n_g} \mathrm{T}(G_O|G_k, A) \\ &+ \frac{(N_g - n_g)(N_g - n_g + 1)}{2}\Pr(G_O|A)\Bigg]\end{aligned} \tag{3.17}$$

with n_g denoting the number of sampled genets and G_k the multi-locus genotype of k-th genet; $G_k \neq G_O$. This assumes equal prior probabilities for selfed and outcrossed parentages.

3.4.2 Efficient Likelihood Calculation

With a marker panel of sufficient power for parentage inference, most of the multi-locus segregation probabilities $\mathrm{T}(\cdot)$ that appear in the marginal probabilities $\Pr(G_O)$ (Eq. 3.15,

3.16, and 3.17) will be 0 in the absence of typing errors and mutations. This is because the probability that an unrelated individual has a genotype that fits without mismatches to an offspring genotype according the Mendelian laws is very low. In fact, these *exclusion probabilities* are a common measure of the power of the marker panel (see Sec. 3.4.7). Thus only parentages without mismatching loci need to be considered, which reduces the pedigree search space and the parentage posterior calculation. The first is important for the mixing time of the MCMC sampling, the latter can be a significant speedup especially when A, N and G are variables (see section 3.4.5).

When tolerating typing errors, however, all parentages have a non-zero probability. An exact computation, therefore, needs to take into account all pairs and triples. However, parentages with many mismatches will have a very small posterior probability and can be ignored. Our implementation therefore uses simulations to generate mismatch distributions for parent(s)-offspring/unrelated relationships in order to determine an appropriate mismatch cutoffs.

For an offspring v, we denote the set of all plausible parents according to this cutoff by $\mathscr{H}_v$. It includes in particular also the cases that none or only one of the parents are sampled. Note that $\mathscr{H}_v \subset V \times V \cup V \cup \{\varnothing\}$. Apart from the number of mismatches, also *prior information* such as sex, age, and known mothers restrict $\mathscr{H}_v$. The posterior probability of a parentage x of offspring i can be expressed in the form

$$\pi_i(x) = \frac{\Pr(O_i|x)}{\sum\limits_{y_j \in \mathscr{H}_i} \Pr(O_i|y_j)}\,. \tag{3.18}$$

With $\Pr(O_i|x)$ denoting the probability of parentage x as shown in Eq. 3.15 to 3.17. For Eq. 3.16 we have for example:

$$\begin{aligned}
\Pr(O_i|\{F_i, M_j\}) &= \mathrm{T}(G_{Oi}|G_{Fi}, G_{Mj})/(N_f N_m) \\
\Pr(O_i|\{F_i\}) &= \frac{(N_m - n_m)}{(N_f N_m)} \mathrm{T}(G_{Oi}|G_{Fi}, A) \\
\Pr(O_i|\{\varnothing\}) &= \frac{(N_m - n_m)(N_f - n_f)}{(N_f N_m)} \Pr(G_{Oi}|A)
\end{aligned}$$

Eq. 3.18 is thus an approximation of Eq. 3.14 because only plausible parentages are considered. In a second filter step, all parentages with negative LOD score [*e.g.* Meagher and Thompson 1986] are ignored. These are all parentages which would decrease the pedigree likelihood if the corresponding arcs would be added to the pedigree. If N is estimated jointly with the pedigree, then these two filter steps can introduce a bias. We thus store the sum of the probabilities of all in the second step filtered offspring-candidate mother and offspring-candidate father pairs and all offspring-candidate parents triples. This sum is then added

to the denominator of Eq. 3.18 and the two pair sums are weighted according Eq. 3.15 to 3.17. The log-likelihood of a pedigree $\mathscr{P}$ is now the sum over all log-transformed parentage posterior probabilities:

$$\mathrm{LL}(\mathscr{P}) \quad = \quad \sum_{i \in V} \log \pi_i(N^+(i)) \tag{3.19}$$

In clonal populations, we can include the number of ramets for every genet in the posterior calculation:

$$\pi_i(x) = \frac{\Pr(O_i|x) n_r(x)}{\sum\limits_{y_j \in \mathscr{H}_i} \Pr(O_i|y_j) n_r(y_j)} \tag{3.20}$$

where $n_r(x)$ is the sum of the number of ramets of the parents in parentage x, and $n_r(x) = 1$ if $x = \{\varnothing\}$. This prior increases the likelihood of parentages with frequently observed genets.

3.4.3 Maximum Likelihood Pedigree

For each individual v, we have to choose one parentage $N^+(v) \in \mathscr{H}_v$. If age data is not or only partially available, then not all such combinations of parents are possible, because this may introduce directed cycles into the pedigree. $\mathscr{T}$ denotes the set of all *valid pedigrees*. If no parentages are known a priori, then $\mathscr{T}$ is never empty. In this case, $\mathscr{H}$ contains for every individual the empty set $\{\varnothing\}$ and thus the empty pedigree without arcs is a valid pedigree. Further, every directed cycle can be elimated by removing arcs.

If parentages are known, then the minimal pedigree that only includes the corresponding arcs is always a valid pedigree. If the minimal pedigree is cyclic, then the priors are wrong, because a cycle in a pedigree means an individual is its own ancestor.

We now consider the problem of finding the pedigree that maximizes the likelihood:

$$\max_{\mathscr{P} \in \mathscr{T}} \mathrm{LL}(\mathscr{P}) = \sum_{i \in V} \log \pi(N^+(i)) \tag{3.21}$$

This problem is similar to learning a Bayesian network from data [Cooper and Herskovits 1992]. The biological restrictions make the problem less complex, however. First, as an individual has only two parents, the maximal indegree of a node in pedigree is 2 and second, the number of considered incoming arcs, here $|\mathscr{H}v|$, is typically restricted. The maximium number of elements in $\mathscr{H}$ is when selfing is allowed:

$$(n-1) + \frac{n(n-1)}{2} + 1$$

This is because all remaining $(n-1)$ individuals could be parents of the offspring v, without prior knowledge we assume that they can all mate with each other and with themselves and

finally, we also consider the case that both parents are unsampled. Without selfing, we have a maximum number of:

$$\frac{n(n-1)}{2} + 1$$

However, as we have already shown in the previous section, with a powerful marker panel or prior knowledge, many parentages can be ignored and the average number of possible parentages per offspring is therefore much smaller. Nevertheless, the number of pedigrees still grows very fast with n. `FRANz` guarantees to find the correct maximum likelihood pedigree for datasets with about less than 26 individuals by using the currently state-of-the art algorithm of Silander and Myllymäki [2006] described in Cowell [2009] for the pedigree reconstruction problem. It has a runtime of $\mathcal{O}(n^3 2^n)$ and a memory usage of $\mathcal{O}(2^n)$. For larger datasets, heuristics are necessary [Almudevar 2003].

3.4.4 Full siblings

As described in detail in Thompson and Meagher [1987], if we cannot exclude two full-sibs, v_i and v_j, as parent and offspring, they in general give a higher likelihood than do true parents. Thus, for highly probable full-sibs, a reasonable strategy is to use only the intersection of the valid parent combinations: $\mathscr{H}_i = \mathscr{H}_j = \mathscr{H}_i \cap \mathscr{H}_j$. A problem with this approach is that false positives may result in an exclusion of the true parents in the pedigree reconstruction. In `FRANz` it is possible in input known full-sib groups (see Appendix A.2.2). This information is available for example in nest structured data when it is known that both parents are monogamous.

Unfortunately, no sibship software package was able to assign full-sib p-values to pairs of individuals. In contrast, they try to partition the dataset in full-sib and half-sib groups, typically by maximum likelihood. We have implemented this feature in `FRANz` and it is available over the `--fullsibtest` command line flag.

We test the null hypothesis that a pair is a full-sib against the alternative hypotheses that they are unrelated, parent-offspring or half-sib. The p-values are calculated by simulation. Given the population allele frequencies and the expected typing error rate, which are either estimated using the sample itself or provided by the user, we generate individuals with known relationships (again $\mathbb{FS}, \mathbb{HS}, \mathbb{PO}, \mathbb{U}$) to determine following distributions:

$$\begin{aligned} \Delta_u &= \log \Pr(G_i.G_j|\mathbb{FS}) - \log \Pr(G_i.G_j|\mathbb{U}) \\ \Delta_{po} &= \log \Pr(G_i.G_j|\mathbb{FS}) - \log \Pr(G_i.G_j|\mathbb{PO}) \\ \Delta_{hs} &= \log \Pr(G_i.G_j|\mathbb{FS}) - \log \Pr(G_i.G_j|\mathbb{HS}) \end{aligned} \tag{3.22}$$

For example Δ_{po} is generated for full-sibs and parent-offspring pairs to estimate the statistical significance of an observed positive Δ_{po} value. So Δ_{po} should be always positive for full-sibs and always negative for parent-offspring pairs. A p-value of 0.05 can be interpreted as 5% of all pairs with a value larger than this delta value, say 1.4, were parent-offspring pairs in the simulation, and not full-sibs despite the fact that 1.4 is positive. As another example, assume a delta value of 0 that has a corresponding p-value of 0.1. Then we would make in 10% of all comparisons an error if we would just look at the sign of Δ_{po}. Note that $\mathbb{HS}$ are all second degree relationships (half-sib, grandparent-grandoffspring and avuncular), which can be considered by weighting in `FRANz` in the p-value calculation.

If two individuals have a common parent pair in $\mathscr{H}$, this is an additional hint that v_i and v_j are full-sibs. Modelling this in the p-value calculation is difficult, we could use however a less conservative critical α value in this case. As default values for α, we use 0.001 and 0.05, respectively. The observed p-values are adjusted for multiple-testing [Holm 1979, Benjamini and Hochberg 1995].

`FRANz` also performs a sanity check of the full-sib assignments. Consider three individuals A, B and C. If (A, B) and (B, C) are full-sibs with significant p-value, then `FRANz` will test (A, C). If (A, C) are unlikely full-sibs, then (A, B) or (B, C) is a false positive, if they are likely, which means the p-value of (A, C) is close to the threshold, then (A, B, C) are probably full-sibs. It is recommended that the user tries to manually resolve these unsure assignments by careful inspection of the p-values.

3.4.5 Algorithm

If age data is not or only partially available, then an ordering of individuals in generations is not possible. Thus not all combinations of parentages may represent a valid pedigree of the sample as some of these combinations may introduce directed cycles into the pedigree. In such a "cyclic pedigree", some individuals would be their own ancestors. The MCMC and Simulated Annealing (SA) procedures now sample valid, cycle-free, pedigrees from the pedigree posterior distribution [Almudevar 2007]. We will later use these sampled pedigrees to estimate parameters and to estimate the statistical significance of a parentage. `FRANz` also

supports the joint estimation of the population's allele frequencies, the number of unsampled candidate parents or missing data imputation (see below in section 3.4.6) in which cases we also need MCMC or SA. In these algorithms, one computes the likelihood of a given pedigree $\mathscr{P}_{i-1}$, randomly generates a new pedigree $\mathscr{P}_i$ and then accepts the change if either $\mathscr{P}_i$ has a higher likelihood or with probability $\exp([\mathrm{LL}(\mathscr{P}_i) - \mathrm{LL}(\mathscr{P}_{i-1})]/T)$. More precisely, in the following $\mathscr{P}_{i-1,j}(y)$ denotes a pedigree $\mathscr{P}_{i-1}$ where the parentage of offspring O_j has been changed from x to y; $x, y \in \mathscr{H}_j, x \neq y$.

We set $i = 1$ and repeat the following steps until convergence (SA) or until i is greater than a given maximum number of iterations (MCMC).

Pedigree Change Step:

We select a random offspring genotype O_j and select a new parentage for O_j from the proposal function $\{q(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j}(y))\}$ which is defined as

$$q(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j}(y)) = \begin{cases} \frac{\pi_j(y)}{1-\pi_j(x)} & \text{if } y \neq x \\ 0 & \text{if } y = x \end{cases} \tag{3.23}$$

This function thus selects a new parentage according their posterior probabilities. We then accept this change with the following probability:

$$\alpha(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j}(y)) = \begin{cases} 0 & \text{if } \mathscr{P}_{i-1,j}(y) \text{ cyclic} \\ \exp\Big(\Big[\mathrm{LL}(\mathscr{P}_{i-1,j}(y)) - \mathrm{LL}(\mathscr{P}_{i-1}) + \\ \quad \log \frac{q(\mathscr{P}_{i-1,j}(y), \mathscr{P}_{i-1})}{q(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j}(y))} \Big]/T \Big) & \text{otherwise} \end{cases} \tag{3.24}$$

With (partially) missing age data, to ensure irreducibility of the MCMC sampler, it is necessary to perform swap steps in which the direction of a random arc (j, k) of the pedigree is reversed [Koch et al. 2008]. Note that age data implies the direction of an pedigree arc, so a swap step would always return an invalid pedigree in the case that age data is available. We can therefore write down the probability to perform such a swap change in the following form

$$\Pr(\text{swap}) = \begin{cases} \frac{|A|}{|A|+n_o} & \text{if age data missing} \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

Otherwise we change the parentage of a random offspring as described above. With n_o we denote the number of sampled individuals in the offspring generation(s) ($n_o = n_g$ in the absence of age data) and $|A|$ is the number of arcs in the pedigree. In a swap change, the

parentages of two individuals j and k are changed and this change is accepted with probability

$$\alpha(\mathscr{P}_{i-1}, \mathscr{P}_{i-1,j,k}(y,z)) = \begin{cases} 0 & \text{if } \mathscr{P}_{i-1,j,k}(y,z) \text{ cyclic} \\ 0 & \text{if } y \notin \mathscr{H}_j \vee z \notin \mathscr{H}_k \\ \exp([\mathrm{LL}(\mathscr{P}_{i-1,j,k}(y,z)) - \mathrm{LL}(\mathscr{P}_{i-1})]/T) & \\ \quad \text{otherwise} & \end{cases} \tag{3.26}$$

The second case is necessary because a swap might generate an invalid parentage, for instance one with too many mismatches. In selfing parentages, both arcs are swapped as otherwise a swap would always produce a cycle.

Estimation of the size of the unsampled population:

If the number of unsampled candidates is not known within reasonable accuracy, it is possible to estimate this number together with with the pedigree, either by sampling N every n_o steps from a uniform distribution in the interval $[n, N_{max}]$ (where N_{max} is specified by the user) or by treating N as a latent variable, estimated again every n_o steps from the indegree distribution of the pedigree. In the latter case we utilize the fact that the pedigree structure itself contains information about the sampling rate in the ratio of the number vertices with indegree 1 and with indegree 2, d_1 and d_2:

$$r = \frac{1}{\frac{d_1}{2d_2} + 1} \quad \text{and} \quad N \approx \frac{n}{r} \cdot x \quad \text{for} \quad x \geq r\,. \tag{3.27}$$

For larger samples, setting $x = 1$ should give a good point estimate of N when we assume that r and x are constant across sampled generations and are the same for both sexes. To increase the accuracy, we sample x from a flat distribution $[r, x_{max}]$. A value of 4 for x_{max} showed a very robust performance in our tests.

Temperature schedule:

In standard MCMC sampling, the temperature T is always kept at the constant value 1. To speedup mixing of the sampler, `FRANz` supports Metropolis-Coupled Markov Chain Monte Carlo (MCMCMC) [Geyer 1991] where k chains are run in parallel. The temperature of the i-th chain is set to

$$T_i = \frac{1}{1 + t(i-1)} \tag{3.28}$$

where t is the heating parameter with default of 0.2. These are the temperature values used in the `MrBayes` software [Huelsenbeck and Ronquist 2001, Altekar et al. 2004]. Then, the

states of the chains i and j are swapped and accepted with probability:

$$\alpha\left((\mathscr{P}_i, T_i), (\mathscr{P}_j, T_j)\right) = \exp\left(\frac{\mathrm{LL}(\mathscr{P}_i) - \mathrm{LL}(\mathscr{P}_j)}{T_j} + \frac{\mathrm{LL}(\mathscr{P}_j) - \mathrm{LL}(\mathscr{P}_i)}{T_i}\right). \tag{3.29}$$

Pedigrees are only sampled from the first, unheated chain. The higher the temperature, the faster the chains moves through the pedigree space, but also the lower the probability that the states of chains are swapped.

In the SA optimization, we use the temperature schedule as described in Aarts and Korst [1989], which has been shown to be efficient in pedigree reconstruction [Almudevar 2003]. In short, the SA optimization starts by determining a temperature that yields in an acceptance probability close to 1, commonly 0.95. A simple algorithm for this purpose performs a random walk in the pedigree space and stores the observed likelihood scores. Then the minimum temperature that would have accepted 95% of the transitions is used as initial SA temperature T_0. The temperature is then lowered every βN iterations, where N is the neighborhood size, which is the sum of the number of plausible parentages over all individuals, and β a parameter with default of 3. The new temperature $T_{i,j+1}$ of the i-th chain is calculated with the following equation:

$$T_{i,j+1} = \frac{T_{i,j}}{1 + \frac{T_{i,j}\log(1+i\delta)}{3\sigma_{T_{i,j}}}} \tag{3.30}$$

where δ the increment parameter with default 0.1 and $\sigma_{T_{i,j}}$ the observed variance of the process in the previous βN iterations. The reason for the staggered heating of the parallel chains is that good solutions are found very fast. The advantage is twofold: First, the chains with small i still perform a very thorough search, but don't find new maximum likelihood solutions too often which have to be stored. Second, if the pedigree search space is too large for some given β and δ, the first chains might not converge within a reasonable number of iterations. Convergence is inferred when the difference of the likelihood means in n_ϵ consecutive temperature changes is smaller than a given ϵ (defaults are $n_\epsilon = 3$, $\epsilon = 0.001\sigma_{T_{i,0}}$).

The MCMCMC and the parallel SA procedures require a compiler that supports `OpenMP` Version 2.0 or higher. All main loops are parallelized and all libraries are thread-safe. `FRANz` uses a thread-safe Mersenne Twister (DCMT Version 0.6.1, [Matsumoto and Nishimura 2000]) for obtaining random numbers.

3.4.6 Missing Values

A common problem in most datasets is that genotyping failed for a significant amount of loci. The problem of dealing with missing data has seen remarkably few attention in parentage analysis. `FRANz` offers two options for dealing with missing data. The first is imputation by

a single-site Gibbs sampler. Here, the alleles of a random genotype g_{ij} with unobserved data of individual i at locus j are sampled proportional to the product of all affected segregation probabilities of the pedigree:

$$\Pr(g_{ij}|\mathscr{P}, A) = \mathrm{T}(G_i|N^+(i), A) \prod_{o \in N^-(i)} \mathrm{T}(G_o|N^+(o), A) \tag{3.31}$$

So the segregation probability of i and of all offspring of i need to be updated. The reason for sampling conditional to this product instead of the pedigree likelihood is that it is more sensitive to changes of a single allele. This Gibbs sampling is available in `FRANz` and can be selected with the `--gibbsmissing` command line flag. An example output is shown in Appendix Fig. A.3.

A well-known problem of this approach is that the irreducibility condition is only guaranteed for diallelic loci [Thompson 2000]. This can be circumvented with non-zero segregation probabilities or for example with MCMCMC [Geyer 1991] samplers [Sheehan and Thomas 1993, Cannings and Sheehan 2002]. The implemented error model ensures the first with non-zero typing error rates and the latter is also available in our implementation. We make such a Gibbs sampling step after n_o pedigree changes.

The second option for dealing with missing data is to include only observed alleles in the likelihood calculations. Noteworthy, this is the standard method employed by most other parentage or paternity inference tools. `FRANz` also supports partially observed genotypes. In general, if one allele of a single locus genotype is missing, then all alleles are considered and we have $\Pr(?) = 1$, where the question mark codes a missing allele. The genotype probabilities are thus:

$$\Pr(?.?) = 1,\ \Pr(a_i.?) = \Pr(a_i) \tag{3.32}$$

For parent-offspring pairs, we have with both parental alleles missing no additional information and thus have the genotype probability:

$$\mathrm{T}(a_i.a_j|?.?) = \Pr(a_i.a_j) \tag{3.33}$$

With one offspring allele missing we have:

$$\delta(a_o, a_p) = \begin{cases} 1 & \text{if } a_o = a_p \vee a_o = ? \\ \Pr(a_o) & \text{if } a_p = ? \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{T}(a_i.?|a_j.a_k) = 0.5 \Pr(a_i) + 0.25 \left[\delta(a_i, a_j) + \delta(a_i, a_k)\right] \tag{3.34}$$

Finally, with one parental allele missing, we can write down the Mendelian segregation probability as:

$$\begin{aligned} \mathrm{T}(a_i.a_i|a_j.?) &= 0.5\left[\delta(a_i,a_j)\Pr(a_i)+\Pr(a_i.a_i)\right] \\ \mathrm{T}(a_i.a_j|a_k.?) &= 0.5\left[\delta(a_i,a_k)\Pr(a_j)+\delta(a_j,a_k)\Pr(a_i)\right]+0.5\Pr(a_i.a_j) \end{aligned} \tag{3.35}$$

For parents-offspring triples, when both maternal or paternal alleles missing, we again use the equations for pairs as there is no additional information. If one offspring allele is missing, we have:

$$\mathrm{T}(a_i.?|a_{j1}.a_{j2},a_{j3}.a_{j4}) = \frac{1}{4}\sum_{k=1}^{4}\delta(a_i,a_{jk}) \tag{3.36}$$

And finally, if one maternal and/or paternal allele is missing, we have:

$$\delta(a_{o1}.a_{o2},a_{p1}.a_{p2}) = \begin{cases} 1 & \text{if } a_{o1}.a_{o2}=a_{p1}.a_{p2} \\ \Pr(a_{o1}.a_{o2}) & \text{if } a_{p1}=?\wedge a_{p2}=? \\ \Pr(a_{o1}) & \text{if } a_{o2}=a_{p1}\wedge a_{p2}=? \\ \Pr(a_{o2}) & \text{if } a_{o1}=a_{p1}\wedge a_{p2}=? \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{T}(a_{o1}.a_{o2}|a_{m1}.a_{m2},a_{f1}.a_{f2}) = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}\delta(a_{o1}.a_{o2},a_{mi}.a_{fj}) \tag{3.37}$$

3.4.7 Allele frequencies

The population allele frequencies are often unknown. If the sample size is large and family sizes are small, it is reasonable to assume that individuals are unrelated and then to use all genotypes for the estimation. Maximum likelihood estimates of the frequency q of allele a are then calculated as $\hat{q} = y/n$ where y is the number of a alleles and n the total number of typed alleles. Standard errors (SEs) for the estimates are given by the binomial formula $\sqrt{\hat{q}(1-\hat{q})/n}$ [Ceppellini et al. 1955]. This is the default method for allele frequency estimation in `FRANz`.

If not, if individuals in the population are related, then this strategy will overestimate the frequency of rare alleles in large families and underestimate the frequency of common alleles. This is because offspring genotypes are derived by chance from the parental genotypes and can provide no further information concerning allele frequencies [Ceppellini et al. 1955]. `FRANz` therefore optionally updates the allele frequencies during SA optimization or MCMC sampling by counting the alleles in the founder generation only. This is computationally extensive, but it is not necessary to update after every change of the pedigree [Thomas and Hill 2000]. So a simple algorithm for this new Maximum Likelihood estimator ignores nodes in the pedigree

with indegree 2. For nodes with indegree 1, only offspring alleles that are not in the parental genotype are counted. Nodes with indegree 0 are included normally. To avoid that rare alleles get lost by this procedure, all observed alleles are counted at least once. This method is available in `FRANz` over the `--updatefreqs` command line flag.

It is also possible in `FRANz` to provide allele frequencies as input file (see Appendix A.2.3). This is useful for example when pedigree reconstruction is only applied to a small subset of sampled individuals. Then the complete dataset can be used to estimate the frequencies.

Allele frequencies are required for the pedigree likelihood calculation, but they provide important additional information about the data. For example, one important question when applying parentage inference methods on marker data is whether the marker panel has sufficient power. The higher this power is, the more statistically significant parentages can be assigned. A standard measure of this power are the exclusion probabilities [Jamieson and Taylor 1997, Wang 2007]. These are the probabilities that a random male in a population has a genotype compatible to a offspring-mother pair. `FRANz` outputs these probabilities and variants for different scenarios of it (i.e., when the mother is unknown or n offspring fullsibs are genotyped). A related measure is the probability of identity, the probability that two random genotypes are equal, which is commonly used in forensics. A third measure of the power of the marker panel is the polymorphic information content (PIC) [Botstein et al. 1980] which is often used in linkage analysis. `FRANz` outputs all these probabilities for each locus and for the complete marker panel, assuming that loci are unlinked.

If an exclusion probability is high enough for a successful application of parentage inference methods in a particular population mainly depends on the population size. The larger the population is, the more loci are needed. In fact, there is linear relationship between the log of the population size and the number of loci needed to attain a certain level of uncertainity [Jones 2003]. Furthermore, exclusion probabilities assume unrelatedness among candidate parents. If it is likely that relatives (*e.g.* grand-parents, aunts and uncles) are present in the sample, then more loci are needed to attain the desired level of uncertainity [*e.g.* Ford and Williamson 2010]. There is however still a linear relationship between the log of the population size and the number of required loci [Jones 2003].

The likelihood model assumes that the population is under Hardy-Weinberg equilibrium (HWE) and allele frequencies are also used to test for this. `FRANz` includes a thread-safe version of the original implementation of Guo and Thompson [1992] for the exact test of Hardy-Weinberg proportion for multiple alleles. For diallelic loci, the standard exact test is performed [Hartl and Clark 2007].

Finally, a common problem in many datasets are null alleles (allelic dropout) which is when the genotyping of one of the two alleles of a particular locus failed [Bonin et al. 2004]. For example consider that the true genotype is $a_1.a_2$ and genotyping of a_2 failed, maybe due to a mutation in the PCR primer site, then the observed genotype will be $a_1.a_1$. `FRANz` outputs the estimates of the EM-Algorithm of Kalinowski and Taper [2006]. If parent-offspring relationships are known, we include observed homozygote mismatches in the likelihood calculation by assuming that the reason for these mismatches are null alleles.

On computers with multiple CPUs, `FRANz` distributes the analysis loci-wise among the available CPUs. Examples of the `FRANz` allele frequency analysis output are shown in Appendix Fig. A.1 and A.2.

3.4.8 Rates of Self-fertilization

Given a pedigree $\mathscr{P}$, we can estimate the selfing rate r_s over the number of observed self-fertilizations S in $\mathscr{P}$:

$$r_s = \frac{2S}{\sum\limits_{i \in V} |N^+(i)|} \tag{3.38}$$

The normalization according the indegrees takes account for the fact that the probability of observing an outcrossed parentage is twice as high observing a selfed one, as in the latter case there is only one instead of two parents.

3.4.9 Rates of Clonality

Estimating the rate at which a population reproduces clonally is notoriously difficult [de Meeûs and Balloux 2004]. We assume in the following that we can estimate N_r, the total number of ramets in the population, within reasonable accuracy for example over the occupied space. We further assume a population in which N_r is constant across generations. In every generation, a ramet reproduces either clonally with rate c or sexually with rate $(1-c)$. A random ramet is killed to keep N_r constant if necessary. Then we use pedigree reconstruction to estimate the total number of genets, N_g. We then estimate the rate of clonal reproduction with the following equations.

By $h(s)$ we denote the expectation value of the number of genets with exactly s ramets, for integer $s \in \{0, \ldots, N_r\}$. We have $h(0) = 0$ by definition. Let us establish conditions for the stationary values of h. In a birth event, $h(s)$ changes by

$$\Delta^{\mathrm{b}}(s) = -c\frac{s}{N_r}h(s) + c\frac{s-1}{N_r}h(s-1) \tag{3.39}$$

for all $s \in \{2, \ldots, N_r\}$, whereas $h(1)$ is changed by

$$\Delta^{\mathrm{b}}(1) = (1-c) - \frac{c}{N_r} h(1) \ . \tag{3.40}$$

A death event causes a change

$$\Delta^{\mathrm{d}}(s) = -\frac{s}{N_r} h(s) + \frac{s+1}{N_r} h(s+1) \ . \tag{3.41}$$

for all $s \in \{1, \ldots, N\}$. At constant population size N_r, birth and death events occur equally often. Therefore

$$\Delta^{\mathrm{b}}(s) + \Delta^{\mathrm{d}}(s) = 0 \tag{3.42}$$

must be fulfilled in equilibrium for all $s\{1, \ldots, N_r\}$. We obtain the set of equations

$$\begin{aligned} h(0) &= 0 \\ h(1) &= N_r(1-c) \\ h(2) &= \frac{1}{2}\left[(1+c)h(1) - N_r(1-c)\right] \\ h(s+1) &= \frac{1}{s+1}\left[s(1+c)h(s) - c(s-1)h(s-1)\right] \end{aligned} \tag{3.43}$$

the last equation being valid for all $s \in \{2, \ldots, N_r\}$. Taken together, this is a second order linear difference equation for a given c and N_r. We solve it numerically and obtain the expected number of genotypes as

$$N_g = \sum_{s=1}^{N_r} h(s) \ . \tag{3.44}$$

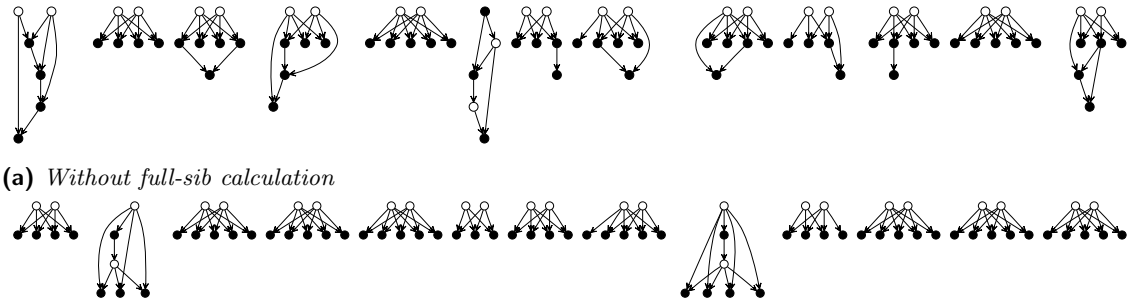
For the inverse problem with given N_g, we obtain c by means of nested intervals.

3.5 Results

To test our algorithm and implementation, we apply it to one empirical and several simulated datasets with known genealogies.

3.5.1 Real Microsatellite Data

Our first dataset is a microsatellite dataset of the black tiger shrimp *Penaeus monodon* [Jerry et al. 2006]. The true pedigree is known from direct observation. The dataset consists of 13 families with a total number of 85 individuals (of which 59 offspring), genotyped at seven highly polymorphic loci. For ten individuals, alleles are missing at one locus. The error rate is very low, with only one observed mismatch. Fig. 3.5 shows the best pedigrees with and



(a) *Without full-sib calculation*

(b) *With full-sib calculation*

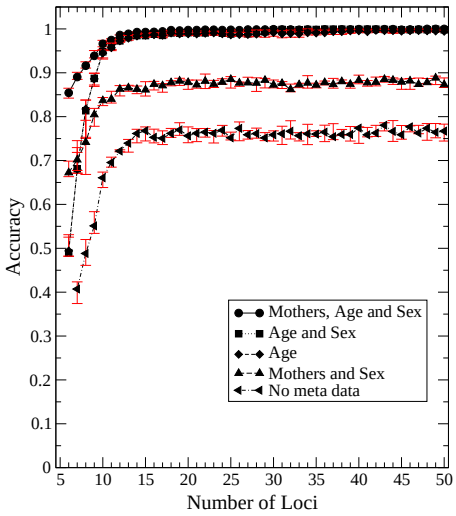
Figure 3.5. *Reconstructed Penaeus monodon pedigree (Section 3.5.1). The white vertices are the parental genotypes, black the offspring genotypes.*

without full-sib calculation. Full-sibs tend to have higher parentage likelihoods, but large full-sib groups greatly enhance the performance of our algorithm such that the accuracy of the reconstructed pedigree increases from 82.8 to 97.1 percent. A recent publication [Berger-Wolf et al. 2007] listed an accuracy rate of several sibling reconstruction methods ranging from 67.8 to 78.0 percent on the same dataset. Classic parentage inference programs such as `CERVUS` [Marshall et al. 1998], where the absence of age data violates main assumptions, assign statistical significant parentages to the parental genotypes even when the correct parameters (sampling rate, fraction of relatives in the candidate parents) are provided.

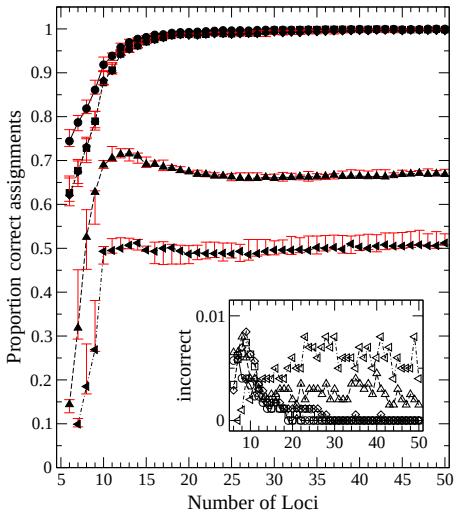
3.5.2 Simulated Human Population

We artificially generate human population datasets as follows. A population of 100 unrelated founders is created by drawing genotypes independently with allele frequencies of 64 human microsatellites [Jin et al. 2000]. Then we let individuals die, mate or marry according to rates extracted from the statistics of the German population [Federal Statistical Office 2007]. As mating partners or husbands, we only allow unrelated individuals. Married couples only mate with each other. We stop when the desired number of individuals is reached. In order to simulate typing errors, we replace the true allele with a random one. Null alleles are simulated in heterozygote genotypes by replacing the null allele with the other allele ($a_i.a_n$ becomes $a_i.a_i$). Homozygote genotypes are marked as missing.

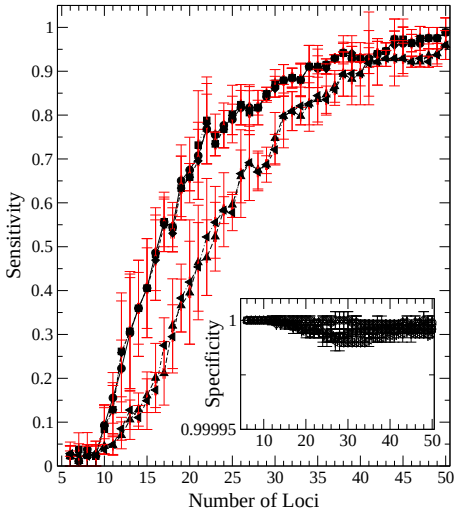
We analyze the accuracy of the Maximum Likelihood pedigree, found by Simulated Annealing or during MCMCMC, as a function of the number of available loci, see Fig. 3.6a. In all cases where the accuracy is below 1, the optimal pedigree from our algorithm has an even larger likelihood than the true one. Thus without exceptions, our algorithm finds a pedigree with at least the log-likelihood of the true pedigree (data not shown). The plots show that the



(a) *The accuracy of the Maximum Likelihood Pedigree.*



(b) *The proportion of incorrect (unfilled symbols) and correct parentages with a posterior probability* > 0.95.



(c) *The sensitivity and specificity of the sibling calculation.*

Figure 3.6. *These plots visualize the results of the reconstruction of simulated human pedigrees (Section 3.5.2). The various measurements are plotted as a function of the number of loci. The values are the median of ten randomly generated pedigrees of size 1000, reconstructed with different combinations of available prior knowledge. The error bars indicate the first and third quartile. The dataset has a sampling rate of 0.5 (1000 of 2000 individuals sampled) and has an overall typing error rate of* 0.01. *In addition, the first locus comprises one null allele* ($p_n = 0.05$). *The pedigree depth ranges from 5 to 9 and the mean number of sampled candidate parents is* 82. N_{max} *was largely overestimated set to 1000.*

reconstruction is robust even when the upper limit of the total number of breeding individuals per generation in the population N_{max} was largely overestimated (164 *vs.* 1000).

Age data is clearly the most informative prior knowledge. Knowledge about the sex rarely helps to exclude a false parentage mainly because mothers are sampled like all individuals with a rate of 0.5 and sex requires candidate parent pairs for exclusion. Thus the knowledge of the sex does not resolve the difficult cases where the true parents are unsampled but a close relative (e.g. aunt or uncle) is sampled.

In Nielsen et al. [2001], a parentage was assigned when the posterior probability was higher than 0.95. These are parentages that are observed in at least 95% of all sampled pedigrees. Fig. 3.6b visualizes the proportion of correct and incorrect assignments. In almost all cases, the proportion of wrongly assigned parentages was smaller than 0.01. These parentages are mainly the difficult cases mentioned above or false positives of the sibling calculation, whose sensitivity and specificity is plotted in Fig. 3.6c. Without age data, the direction of a large fraction of parent-offspring arcs cannot be determined, which explains the plateaus in the plots. These parentages are easily identified by their posterior probability which is typically near 0.5.

3.5.3 Simulated Clonal Plant Population

Growing Population

We next simulate data of clonal populations. We use the model of a growing population, where individuals do not die once they reproduced sexually. The data is simulated with allele frequencies with 8 alleles per locus. We use random ("broken stick") frequencies with a maximum frequency of 0.5. In every year, a ramet reproduces either sexually or clonally. For sexual reproduction we assume a fixed rate of self-fertilization or outcrossing with a random ramet. With a probability of 0.01, we replace one allele of a single-locus genotype with a random one to simulate genotyping errors.

We choose relatively high rates of selfing (0.1) and clonality (0.9). This results in an extremely difficult test dataset as individuals are closely related because old plants grow fast and thus mate very often. Exclusion probabilities [Jamieson and Taylor 1997, Wang 2007], which assume unrelatedness among candidate parents, thus overestimate the power of a marker suite. The exclusion probabilities of the simulated datasets are shown in Fig. 3.7.

For the sampling rates, we choose a relatively high one of 0.01 of (1.000 of about 100.000 ramets) and a more realistic one of 0.001 (350 of about 350.000). We generate 10 datasets

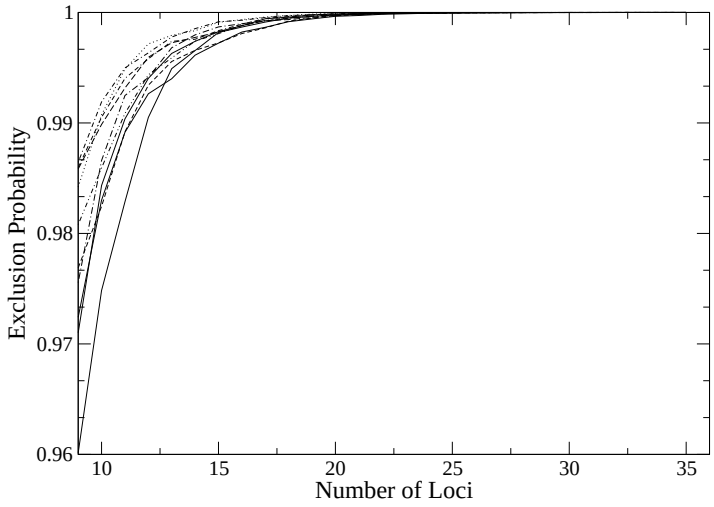


Figure 3.7. *The exclusion probabilities [Jamieson and Taylor 1997] of 10 randomly generated datasets (3.5.3). In our data, the average probability that a random individual has a genotype that is compatible with an offspring genotype is* $< 1 \times 10^{-5}$ *for more than 25 loci.*

for each of the two sampling rates.

The accuracy of the pedigree reconstructions are shown in Fig. 3.8a as in the simulated human population. The incorporation of the number of sampled ramets per genet (using Eq. 3.20 instead of Eq. 3.18) improves the reconstruction significantly. A reason for that improvement is that the number of ramets is an approximation of the age of genets; without age data it is sometimes not possible to identify parent and offspring in a parent-offspring pair, which is also the reason why the accuracy does not reach 100%. The plot also shows that the sampling rate has surprisingly little influence on the accuracy, the amount of available genomic information is the crucial factor here.

Our implementation `FRANz` is highly efficient and is able to sample millions of large pedigrees in a minute, see Table 3.1 for a benchmark. The Metropolis-Hastings acceptance rates are for 12 loci close to the optimal acceptance rate of 0.234 [Roberts et al. 1997]. With higher power of the marker panel, the acceptance rate gets lower because the likelihood differences between correct and wrong parentages increases very fast and thus the Markov chain reaches local optima very soon. For the same reason, swap steps are more likely to result in invalid parentages with increasing amount of genomic information.

Fig. 3.8b shows the fraction of correct and incorrect parentages with a pedigree posterior probability of > 0.95. Although these probabilities are not exactly comparable to the thresholds of classical, simulation based paternity inference tools such as `CERVUS` [Marshall et al.

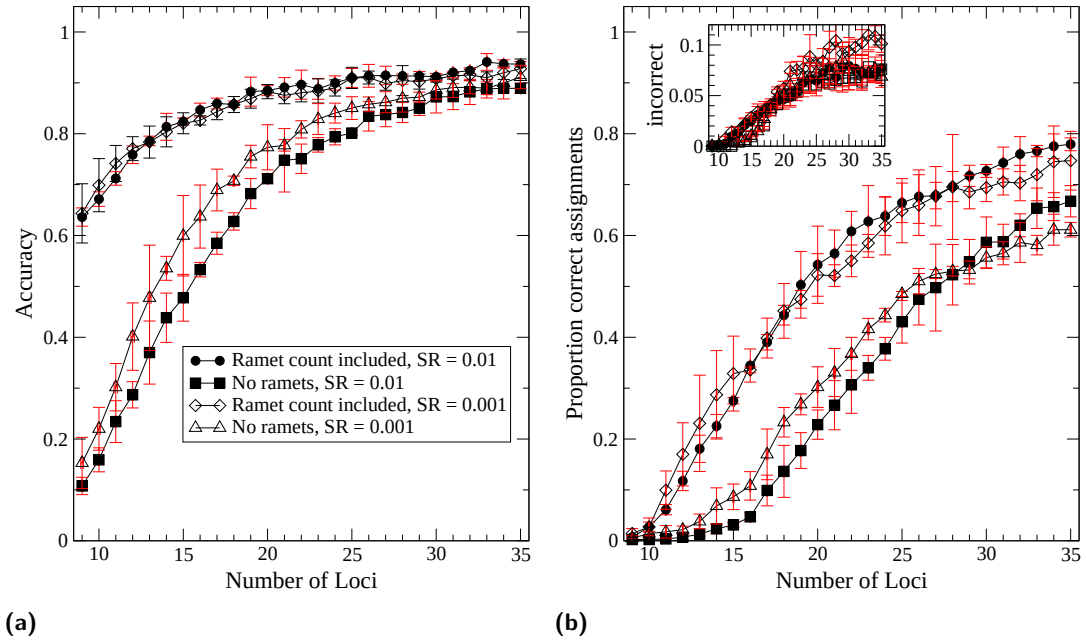


Figure 3.8. *The accuracy of the reconstructed Maximum Likelihood pedigrees of simulated clonal populations is plotted in Fig. 3.8a as a function of the number of loci. The values are the median accuracy of ten randomly generated pedigrees of size 10.000 genets with a sampling rate (SR) of 0.01 (filled symbols) and 35.000 genets with a sampling rate of 0.001 (unfilled symbols). The simulated datasets are reconstructed once with the standard parentage posterior probabilities (Eq. 3.18) and once with the number of ramets included as priors (Eq. 3.20). The error bars indicate the first and third quartile. Fig. 3.8b lists the proportion of correct and incorrect parentages with a posterior probability* > 0.95.

1998] due to the very different approaches, the plot nevertheless shows that even with powerful marker panels, i.e., when there is a high amount of genomic information in the data, we can only assign relatively small numbers of parentages, which is especially a problem in low sampling rate datasets. It is therefore crucial to use an approach that uses the complete data for the estimation of parameters of interest, not only highly significant parentages. The high rate of incorrect assignments, especially in the 0.001 dataset, is explained by the large violation of the assumption that candidate parents are unrelated.

We then use the MCMC sampled pedigrees to estimate the parameters of interest. As an example, we plot in Fig. 3.9 the estimated rates of self-fertilization (Eq. 3.38) for both sampling rates (0.01 and 0.001, using Eq. 3.20) as a function of the number of loci. As to expect, the accuracy of the test dataset with high sampling rate (Fig. 3.9a) is higher than the one with lower sampling rate (Fig. 3.9b) because the number of observed parentages is much higher.

Table 3.1. *This table lists the performance of* `FRANz` *for simulated clonal datasets with 1000 ramets (about 430 genets). The values are the average over ten simulated datasets. In all cases, 3.500.000 x 8 (CPUs) pedigrees were evaluated.*
Intel® Xeon® CPU, 2.33GHz, 8 cores, 16GB RAM

Loci	Acceptance Rate	Runtime	
		sec.	Pedigrees/sec.
9	0.329	1455	19250
12	0.210	825	33946
15	0.139	363	77135
30	0.030	84	332016

Table 3.2. *Comparison of the estimated rates of self-fertilization. This table lists the means and standard deviations of the selfing rate estimates from the RMES software [David et al. 2007] and the present approach for simulated datasets (Sec. 3.5.3) with a true selfing rate of 0.1.*

S.R.[a]	Loci	`RMES`		`FRANz`	
0.01	9	0.154	±0.044	0.100	±0.012
0.01	35	0.176	±0.026	0.096	±0.023
0.001	9	0.148	±0.055	0.099	±0.027
0.001	35	0.175	±0.025	0.103	±0.034

[a]S.R. sampling rate (ramets)

The estimates are fairly independent of the number of loci and already accurate with very low amount of genomic information. Other parameters such as male fertilities [Morgan and Conner 2001] could be calculated analogously.

We also compare our selfing rate estimates with the Maximum Likelihood estimates of the RMES software [David et al. 2007] in Table 3.2. RMES estimates selfing rates over observed multi-locus heterozygosity deficiencies and does not require parent-offspring relationships. These allele frequency approaches are therefore in principle capable of estimating *long-term* selfing rates but are inherently less robust with respect to violations of the assumptions. `FRANz` on the other hand, provides quite accurate estimates of the *recent* selfing rates for the 10 datasets.

This pedigree reconstruction approach works in the model of a growing population even with

very low sampling rates extremely well. This is because old founder plants are sampled with probability close to one and these plants have many offspring. This is the reason why we observe enough parentages for reliable parameter estimation.

Constant Population Size

To show that the present approach is able to estimate the sampling rate of the genets, we next simulate data under the model of population with a constant number of ramets N_r. We start with a founder generation of 100 unrelated genets with 9 loci and use the same allele frequencies as before. Then in every generation, all ramets reproduce again either sexually or clonally. If sexually, then again with a selfing rate of 0.1. Outcrossing happens with a random living ramet or with an migrated ramet. Here we use a migration rate of 0.01. If after such a reproduction event there are more than N_r ramets, one random living ramet is killed. We stop the simulation after the birth of the 20000-th genet. Then we sample $n_r = 500$ living ramets. We generate again 10 datasets for every parameter combination. Here we vary the rate of clonality (0.5, 0.8, 0.9 and 0.95) and N_r (4.000 and 10.000). N_g is then estimated over the indegree distributions of the MCMC sampled pedigrees [Riester et al. 2009]. The rate of clonality is then estimated with the model described in Sec. 3.4.9. N_r is assumed to be known *a priori*.

In Table 3.3 we present the results of the pedigree reconstruction of the datasets with constant population size (Sec. 3.5.3). Our approach significantly underestimates the true N_g. This is partly explained by the fact that the probability of sampling old and big plants, i.e., ones with many ramets is higher than sampling small genets. And as old plants have in general more offspring than young ones, we observe more parentages as we would expect by assuming equal sampling probabilities of parents for all individuals, which we do. More observed parentages result in a higher sampling rate and therefore a smaller N_g. Our model (Sec. 3.4.9) also slightly underestimates the true N_g. Nevertheless, with relatively high sampling rates of 0.125, we can observe fairly accurate estimates. As the number of genets increases with decreasing rate of clonal reproduction, we observe less parentages if these rates are low. This explains the high variances in the datasets with a clonal rate of 0.5. At clonal rates smaller than 0.9, we see that a sampling rate of 0.05 is not high enough for a reliable parameter estimation: the selfing rates are in these cases also significantly underestimated.

Table 3.3. *Estimated rates of N_g, clonality and self-fertilizations. This table lists means and standard deviations of the in* **FRANz** *estimated N_g for ten simulated datasets (Sec. 3.5.3) for each of the 8 parameter combinations. It further lists the true N_g and the ones estimated of our model (Sec. 3.4.9). Then the estimated rates of clonality are presented. Finally, the mean and standard deviations of estimated rates of self-fertilization are shown (with a true selfing rate of 0.1).*

S.R.[a]	N_g					Rate of Clonality			Selfing Rate	
	FRANz		True		Model	FRANz		True	FRANz	
0.125	2574.95	±614.77	2834.00	±20.18	2772.59	0.541	±0.19	0.5	0.079	±0.04
0.05	5832.13	±3508.20	6955.67	±43.36	6931.47	0.611	±0.32	0.5	0.037	±0.03
0.125	1460.90	±228.39	1711.20	±18.88	1609.44	0.826	±0.05	0.8	0.082	±0.02
0.05	2447.02	±522.00	4000.80	±31.02	4023.59	0.905	±0.03	0.8	0.068	±0.02
0.125	876.26	±80.41	1121.90	±18.89	1023.37	0.920	±0.01	0.9	0.095	±0.02
0.05	1814.28	±243.40	2792.70	±37.12	2558.43	0.939	±0.01	0.9	0.066	±0.03
0.125	555.72	±45.25	720.10	±21.70	630.68	0.958	±0.01	0.95	0.094	±0.03
0.05	1218.60	±121.70	1847.80	±31.42	1576.70	0.965	±0.01	0.95	0.081	±0.03

[a]S.R. sampling rate (ramets)

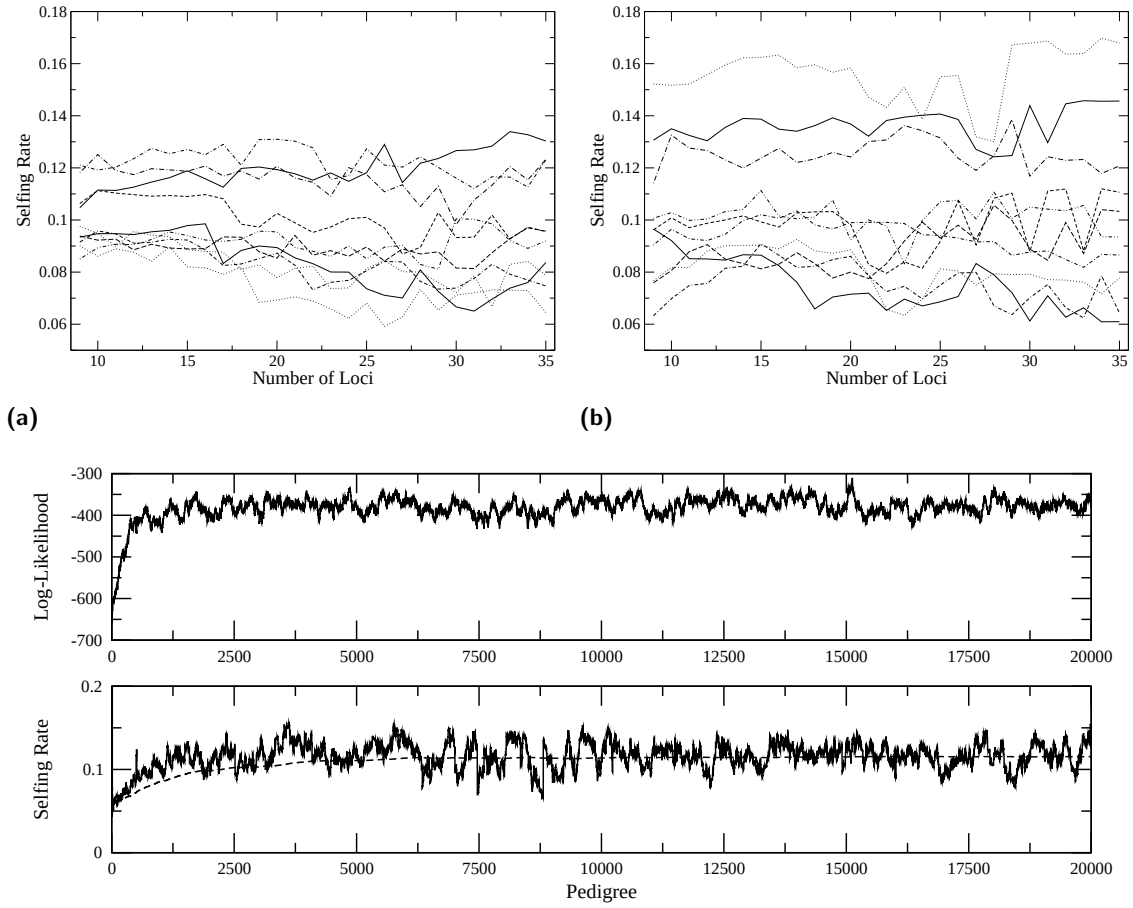


Figure 3.9. *The estimated rates of self-fertilization of the simulated datasets. A line visualizes the selfing rate of one of the 10 datasets for different amounts of genomic information. Fig. 3.9a are the rates from the* 0.01 *sampling rate datasets, Fig. 3.9b from the* 0.001 *one. In Fig. 3.9c we show a trace plot of the MCMCMC sampling of one simulated dataset (9 loci, sampling rate* 0.01*, 8 parallel chains). The dashed curve visualizes the mean of the selfing rate.*

3.6 Discussion

We have presented a new algorithm for the multigenerational pedigree reconstruction problem. The publicly available implementation is written in the C programming language and is platform-independent. We have demonstrated the accuracy and good MCMC(MC) mixing properties of the implementation on simulated and empirical data. Our efficient likelihood calculations allows parentage analysis on huge datasets with thousands of individuals and these genealogies are typically reconstructed in a few minutes. Our implementation is flexible in incorporating additional data like age, sex, sampling locations, sub-pedigrees and allele frequencies and works with a wide range of sampling schemes (Fig. 3.1). This was suggested in Almudevar [2003] but not previously implemented in a publicly available software package. The reconstruction of large and deep pedigrees is highly accurate with only 10-15 polymorphic microsatellite loci. Our approach is to our knowledge the first one that combines paternity inference and sibship reconstruction. As in contrast to most other newly developed parentage analysis tools, it is not just a prototype implementation. Table 3.4 shows that `FRANz` is fairly feature-complete.

In Almudevar [2003], some remaining challenges in the pedigree reconstruction problem were listed. These are the assumption that founders are unrelated, a better estimation of allele frequencies, linkage, support for typing errors or mutation, and estimation of the error of the reconstruction procedure. `FRANz` makes significant progress in the latter two tasks by combining the error model described in Kalinowski et al. [2007] with an MCMC sampling.

The error model, however, was criticized in the literature because of its simplicity. Other programs explicitly model special kinds of errors, for example null alleles and sample the true genotypes with an individual-by-individual Gibbs sampling [Wang 2004b, Hadfield et al. 2006]. For multigenerational pedigrees, one has to sample over the family to ensure irreducibility of the Markov chain [Sheehan 2000]. For large pedigrees, this becomes very fast computationally infeasible and the gain is questionable. Extending the likelihood formulas in [Kalinowski et al. 2007] to model null alleles, however, could be a valuable extension if they occur at higher rates. Now, `FRANz` estimates the null allele frequency [Kalinowski and Taper 2006] and warns the user when null alleles are likely to be present in the data.

It is possible in `FRANz` to estimate allele frequencies simultaneously with the pedigree. A possible extension in future versions of `FRANz` could be support of sub-populations with different allele frequencies. As only very few tools support the estimation of allele frequencies, little is currently known about the accuracy gain and further research is necessary here.

Extensions of the LOD scores for linked loci when the linkage phase is known are proposed

Table 3.4. *Feature comparison of several parentage analysis tools. Adapted and extended from Jones et al. [2009]. PM, paternity/maternity; PP, parent pair allocation; MG, multigenerational; PR, parental reconstruction; SR, sibship reconstruction; IC, Ability to assign statistical confidence for particular parent-offspring pairs; EC, Ability to assess the expected confidence in assignments on an experiment-wide basis; EP, exclusion probabilities; EP ($\mathbb{FS}$), exclusion probabilities for n full-sibs; FP, full probability parentage analysis; DM, dominant markers; AF, Allele frequency estimation; HWE, Hardy-Weinberg equilibrium test.*

	Available functions													Error accommodation	
	PM	PP	MG	PR	SR	IC	EC	EP	EP ($\mathbb{FS}$)	FP	DM	AF	HWE	Null Alleles	Error/Mut
CERVUS 3.0	X	X				X	X	X					chi square	Moderate	Good
COLONY 2.0	X	X		X	X	X					X	X		Good	Good
FaMoz	X	X				X	X	X			X			Poor	Good
FAP 3.6		X					X							None	Moderate
FRANz	X	X	X		X	X		X	X	X		X	exact	Moderate	Good
GERUD				X			X	X						None	None
MASTERBAYES	X	X				X				X	X	X		Good	Good
NEST		X								X				None	Poor
PAPA 2.0	X	X					X							None	Good
PARENTAGE				X		X								None	Moderate
PARENTE	X	X				X								Poor	Good
PASOS 1.0	X	X					X							None	Good
PATRI	X					X				X				None	None
PEDIGREE 2.2				X	X									Poor	Poor

in Devlin et al. [1988]. If the linkage phase and recombination rates are known with high accuracy, the incorporation of this prior information can significantly enhance the performance of the parentage assignments [Devlin et al. 1988]. However, in most cases the linkage phase is unknown and has to be estimated jointly. Loose linkage of a small fraction of markers should not seriously bias multilocus likelihood calculations [Meagher 1991]. Tightly linked loci in contrast, such as neighboring SNPs, can be combined and treated as one single *pseudolocus*. In general, linked loci are less informative than unlinked ones and therefore the calculated LOD scores are too large. The best advice now is probably to avoid medium linked loci [Jones and Ardren 2003].

The framework we have presented in this chapter may easily be extended to incorporate prior knowledge in the likelihood calculation [Neff et al. 2001]. Currently, prior knowledge is only used to reduce the search space and, in the case of clonal populations, to incorporate the number of sampled ramets per genet. For parentages, sampling locations and behavioural data have been successfully used to increase the parentage assignments in Hadfield et al. [2006]. However, it should be noted that it is often very difficult to specify a good model here. For example drops mating success exponential or linear with distance between sampling locations? Wrong assumptions will seriously compromise the parentage assignments. In doubt and if possible, one should consider acquiring genomic information from 1-2 additional loci to increase the assignments to a reasonable level without the risk of compromising the results with unsure priors.

Priors about the pedigree structure (the expected inbreeding rates, number of offspring, etc.) might further improve the performance [Sheehan and Egeland 2007]. Information of this kind is oftentimes unknown *a priori*, however. In fact, these are parameters that one typically would like to infer from the reconstructed pedigrees.

Our implementation currently only allows co-dominant markers. In Gerber et al. [2000], the original LOD scores for co-dominant markers [Meagher and Thompson 1986] were modified for dominant markers. Statistics for estimating pairwise relationships with dominant markers were proposed e.g. in Wang [2004a].

Our incorporation of full-sib probabilities is a reaction to the concern expressed in Meagher and Thompson [1986] that non-excluded full-sibs of the offspring have on average a higher LOD score than the true father. To keep the pedigree likelihood function simple and efficient to calculate, we use only highly significant full-sibs to reduce the pedigree space. It seems possible to include more siblings than just the highly significant ones into the pedigree likelihood calculation without the risk of excluding the true parents. Since such "local" factors in the pedigree likelihood are also not very computationally intensive, we plan to explore this

avenue in future work.

We have also presented a novel likelihood model for the reconstruction of pedigrees in monoecious clonal plant populations. We have shown that the joint estimation of parameters of interest such as the rate of self-fertilization is possible with high accuracy even with marker panels of moderate power. Classical methods can only assign a very limited number of statistically significant parentages in this case and would therefore fail, especially if sampling rates are low which is still a problem in most parentage studies. We have also shown that our likelihood model is surprisingly robust for violations of assumptions such as unrelatedness of candidate parents and constant effective population size.

As mating success drops off with distance between mates, several authors suggested likelihood models that include the sampling location of the genotypes [*e.g.* Burczyk et al. 1996, Smouse et al. 1999, Hadfield et al. 2006]. In principle, it is possible to add the corresponding prior probability distributions in our model. To calculate the distance between two clones A and B, δ_{AB}, one can use the locations of all sampled ramets as an approximation for the real distance between the (maybe unsampled) mating ramets. An obvious strategy for the calculation of δ_{AB} would be the average distance between all sampled ramets of A and B. Another possibility would be to use the minimum distance.

It should be noted that other methods for the estimation of recent selfing rates exist which do not necessarily require that parental genotypes are sampled. For example if it is possible to obtain progeny arrays, the known family structure in the data can be used to reconstruct maternal genotypes. Selfing rates are then estimated by comparison of maternal with offspring genotypes [*e.g.* Jarne and David 2008, for a review]. If neither such a family structure nor parental genotypes are known, then reconstruction of the genotypes of the previous generations might be possible by MCMC sampling [Wilson and Dawson 2007]. However, this assumes that the model used in MCMC sampling fits the population under investigation.

We assumed in this chapter that all ramets have the same genotype. However, especially in long-living plant populations with high rates of clonality, somatic mutations may lead to clones with different genotypes. In this case it could be necessary to extend the model to allow multiple genotypes per genet and include them in the segregation probability calculation (see Sec. 3.3.5). `FRANz` supports partially genotyped loci where only one of the two alleles are known and this feature could be used in these cases to mark an observed mutation as unknown without loosing much information.

With the rapid progress and decay of cost in high-throughput sequencing techniques, it is just a matter of time until there are whole genomes of complete populations available. Large

amounts of SNP data with high quality genetic maps will be therefore available, at least for some model organisms. The identification of parents with such an amount of data is a trivial task and the methods are well known [Boehnke and Cox 1997]. The extraordinarily efficient MCMC(MC) pedigree sampling procedure in `FRANz` could be then extended to allow gaps in the pedigrees, i.e., also infer second degree relationships. A very challenging question is also how many unobserved generations we can reconstruct back in time (see Steel and Hein [2006] and Thatte and Steel [2007] for first results). As we cannot expect an elegant solution to this problem, MCMC heuristics are promising tools for throwing some light on a population's immediate past.

4

Conclusions

In this thesis, we developed means for genealogical reconstruction, in the first part of cancer subtypes and in the second part of individuals in natural populations. In Chapter 2, we successfully applied distance-based phylogenetic tree reconstruction methods to microarray data of sarcoma, acute myeloid leukemia (AML) and breast cancer patients. We validated our methodology on experimental data with known genealogy. With this method, we were further able to find adipogenesis-related genes without the use of *in vitro* methods of differentiation, which are available for only a few histologies [Beqqali et al. 2006].

In Chapter 3, we presented a flexible new software package for pedigree reconstruction and parentage inference in natural populations called `FRANz`. It is the first tool that can estimate the effective population size even for populations where the age of individuals is not easily observable, for instance in clonal plant populations. Our research shows that pedigree reconstruction can accurately infer parameters describing the population's mating behaviour, such as rates of self-fertilization and clonality. This is possible even with marker panels of only moderate power, which is the case for many existing datasets. Additionally, we showed that the present approach scales very well with increasing number of individuals. This is important because sequencing becomes cheaper and faster and thus genotyping of many thousand individuals in a population feasible. Furthermore, we developed the first method for the

calculation of p-values for pairs of individuals being full-siblings. In contrast to many other related tools, `FRANz` is fairly feature complete. Only few major features are left for future versions, *e.g.*, support for sub-populations with different allele frequencies or a better error model that handles null alleles. A graphical user-interface would probably also help to reach a wide distribution of the software among molecular ecologists. Especially as we could show how powerful analyses of the MCMC sampled pedigrees are, it should be possible for users to estimate their parameters of interest easily.

In the next years, the number and extend of single cell analyses in cancer research are expected to increase significantly. The pedigree of cells of the tumor will give information about mutation rates and relative fitness values, very similar to our pedigrees of natural clonal plant populations. It seems straightforward to extend our likelihood models for strictly asexually reproducing clonal populations. The arcs between genets in the pedigree then visualize mutations instead of matings.

A

FRANz

A.1 Availability

An open source implementation of FRANz is available at http://www.bioinf.uni-leipzig.de/Software/FRANz. For a simple interaction and comparison with other tools, we provide a user-friendly Web 2.0 input file generator (see Appendix A.4) on the FRANz website. Furthermore, it is now possible to convert FRANz input files into several other formats (currently supported are CERVUS [Marshall et al. 1998, Kalinowski et al. 2007], PARENTE [Cercueil et al. 2002], GENEPOP [Roussett 2008], and RMES [David et al. 2007]).

A.2 Input files

A.2.1 Main input file

The main input file lists all individuals and their genotypes as well as ecological data such as years of birth and death, sex and for clonal organisms, the number of sampled ramets:

```
1 3 / SIMPSONS
LID1
```

```
LID2
LID3
7 Springfield
Grampa     1 1920 ? M 110/100 200/208 ?/?
Homer      1 1950 ? M 110/170 200/210 300/302
Bart       1 1982 ? M 110/120 200/212 302/304
Lisa       1 1980 ? F 140/170 200/218 302/306
Maggie     1 1988 ? F 110/140 210/212 300/304
Marge      1 1952 ? F 120/140 212/218 ?/306
Flanders   1 ? ? ? 150/160 214/220 300/?
```

The first line in this example file,

```
1 3 / SIMPSONS
```

says the dataset includes one sampling location and three loci. The alleles of diploid genotypes are separated by a slash (/), and the dataset title is "SIMPSONS". Optionally, loci ids can be provided, one id per line. The next line is then for the first (and in this case the only) sampling location:

```
7 Springfield
```

This means 7 genotypes in sampling location "Springfield". These genotypes have to be stored in the following format:

```
Grampa     1 1920 ? M 110/100 200/208 ?/?
```

The first ten characters are a description of the genotype or individual. Then, the next number is how often this genotype was observed. This is meant for clonal organisms where it is the number of sampled ramets. The 1920 is year of birth of Grampa, ? his year of death (unknown), M his sex (F for females and ? if unknown). The rest of the line is reserved for the 3 diploid loci. Missing data is again coded with a ?.

A.2.2 Known relationships

Known parent-offspring relationships are defined in FRANz with a pedigree infile (`--pedigreein <filename>`):

```
7
    Grampa
     Homer
      Bart
```

```
     Lisa
   Maggie
    Marge
 Flanders
    Marge      Bart
    Marge      Lisa
    Marge    Maggie
```

The first line is the number n of individuals, the next n lines are the exactly ten characters long names or descriptions of the individuals. They must be identical to the ones in the genotype file. Then, the remaining lines are the pedigree arcs in the format

```
   parent     child
```

Known fullsib relationships are defined with `--fullsibin <filename>`. If it is known that some individuals are either fullsibs or halfsibs, one can specify this with the `--halfsibin <filename>` command line argument. This is useful for example in nest structered data when one or both sexes are monogamous. The file format is in both cases the same:

```
1
3
      Bart
      Lisa
   Maggie
```

The first line is the number of sibling groups, the 3 is the number of siblings in the first group and the following 3 lines contain the ids of the individuals as in the pedigree infile.

A.2.3 Allele frequencies

Allele frequencies are either be estimated from the data or provided by the user with the `--freqin <filename>` command line parameter:

```
3
7 100 170
100 0.071429
110 0.285714
120 0.142857
140 0.214286
150 0.071429
160 0.071429
170 0.142857
```

```
7 200 220
200 0.285714
208 0.071429
210 0.142857
212 0.214286
214 0.071429
218 0.142857
220 0.071429
4 300 306
300 0.300000
302 0.300000
304 0.200000
306 0.200000
```

The first line is the number of loci (3 in this example). The second line is for the first locus and says that there are 7 different alleles in range 100 to 170. The next 7 lines are the alleles with their frequency, separated by space.

A.2.4 Sampling locations

The sampling locations are provided either as pairwise distances (`--geofile`) or coordinates (`--coordfile`). In both cases, the order of the locations must be the same as the one in the genotype file.

```
3
AcquaAzz1 0.000 0.000 1030.116
AcquaAzz2 0.000 0.000 1030.116
Addaia    1030.116 1030.116 0.000

3
AcquaAzz  36.43 15.09
AcquaAzz2 36.43 15.09
Addaia    40.016 4.207
```

A.3 Output files

```
*** Locus ZUXP4.82 ***

+--------+--------+------------+------------+----------------------+--------------------------+
| Allele |  Count | Heterozyg. |   Homozyg. |  Frequency         SE | Frequency (Null Alleles) |
+--------+--------+------------+------------+----------------------+--------------------------+
|    153 |     14 |         14 |          0 |    0.08235    0.02108 |                  0.08235 |
|    155 |      9 |          9 |          0 |    0.05294    0.01717 |                  0.05294 |
|    159 |      7 |          7 |          0 |    0.04118    0.01524 |                  0.04118 |
|    162 |     86 |         50 |         18 |    0.50588    0.03835 |                  0.50588 |
|    164 |     48 |         34 |          7 |    0.28235    0.03452 |                  0.28235 |
|    169 |      4 |          4 |          0 |    0.02353    0.01163 |                  0.02353 |
|    186 |      2 |          2 |          0 |    0.01176    0.00827 |                  0.01176 |
+--------+--------+------------+------------+----------------------+--------------------------+

Observed Heterozygosity     : 0.7059
Expected Heterozygosity     : 0.6562
Hardy-Weinberg Test
   p-Value                  : 0.2784
   Standard Error           : 0.0088
Estimated Null Allele Freq. : 0.0000
Estimated Genotyping Failure: 0.0000 (0 untyped)
Polymorphic Inform. Content : 0.6035

Exclusion probability when 1 to 6 fullsibs are genotyped
  First Parent              : 0.2465613  0.3323548  0.4314385  0.4779754  0.4870763  0.4878616
  Second Parent             : 0.4142009  0.5289062  0.6202651  0.6592185  0.6677287  0.6685116
  Parent Pair               : 0.6009220  0.7255527  0.8085812  0.8373807  0.8421883  0.8425784

Probability of identity
  2 unrelated individuals   : 0.1715275
  Siblings                  : 0.4662469

Sibship exclusion probability
  3 unrelated individuals   : 0.2262139
  4 unrelated individuals   : 0.4645058
```

Figure A.1. *FRANz output of the allele frequency analysis. This is the detailed output of one locus of the Penaeus monodon dataset [Jerry et al. 2006] dataset described in Sec. 3.5.1.*

```
*** Summary Statistics ***

Locus       Alleles    Min    Max      N   Hobs   Hexp    PIC  EX 1P  EX 2P  EX PP     ID  IDsib  P_NULL  HWE PV  HWE SE
DTLP109          21    268    384     85  0.941  0.931  0.921  0.740  0.850  0.963  0.011  0.290  0.0000  0.0000  0.0000
DTLP103          19    363    620     85  0.882  0.881  0.864  0.608  0.756  0.912  0.027  0.319  0.0000  0.0584  0.0062
DTLP136          14    404    485     85  0.788  0.873  0.854  0.582  0.737  0.897  0.032  0.324  0.0211  0.0000  0.0000
ZUXP4.82          7    153    186     85  0.706  0.656  0.603  0.247  0.414  0.601  0.172  0.466  0.0000  0.2784  0.0088
DTLP110a         26    387    598     85  0.953  0.941  0.932  0.773  0.871  0.972  0.008  0.284  0.0000  0.0000  0.0000
DTLP402a         10    151    283     75  0.707  0.772  0.742  0.398  0.580  0.778  0.080  0.386  0.0398  0.1966  0.0093
DTLP313a         18    521    592     85  0.882  0.903  0.890  0.664  0.798  0.937  0.019  0.306  0.0163  0.0000  0.0000

Average number of alleles   :  16.429 (+- 6.554)
Average observed heterozyg. :   0.837 (+- 0.104)
Average expected heterozyg. :   0.851 (+- 0.102)
Average PIC                 :   0.829 (+- 0.118)

Cumulative exclusion probability when 1 to 7 fullsibs are genotyped
  First Parent              : 0.9985224  0.9998560  0.9999961  0.9999996  0.9999997  0.9999997  0.9999997
  Second Parent             : 0.9999384  0.9999972  0.9999999  1.0000000  1.0000000  1.0000000  1.0000000
  Parent Pair               : 0.9999999  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000

Cumulative probability of identity
  2 unrelated individuals   : 0.0000000
  Siblings                  : 0.0004689

Cumulative sibship exclusion probability
  3 unrelated individuals   : 0.9999442
  4 unrelated individuals   : 1.0000000
```

Figure A.2. *FRANz output of the summary of the allele frequency analysis. The data is again the Penaeus monodon dataset.*

```
*** Locus DTLP402a ***

Genotype        151     203     207     231     235     259     263     267     279     283
313          0.1420  0.1270  0.1279  0.1085 *0.3282  0.0604  0.0426  0.0325  0.0215  0.0095
313          0.0111  0.0177  0.0330  0.0426 *0.2794  0.1070  0.1072  0.1217  0.1412  0.1391
222          0.1451  0.1343  0.1197 *0.1091  0.3375  0.0517  0.0418  0.0350  0.0167  0.0090
222          0.0101  0.0197  0.0301  0.0392 *0.2952  0.1011  0.1117  0.1250  0.1294  0.1386
1810         0.1812  0.1644  0.1519 *0.1265  0.1086  0.0904  0.0738  0.0495  0.0385  0.0152
1810         0.0167  0.0384  0.0586  0.0731  0.0925  0.1095  0.1324  0.1362 *0.1590  0.1835
1812         0.1819 *0.1686  0.1451  0.1305  0.1024  0.0880  0.0724  0.0546  0.0363  0.0201
1812         0.0176  0.0368  0.0577  0.0763  0.0918 *0.1061  0.1300  0.1469  0.1579  0.1790
1822         0.1061  0.1016  0.1081 *0.1016  0.0604  0.1112  0.1008  0.1182  0.0999  0.0921
1828         0.1753  0.1606  0.1473  0.1275 *0.1099  0.0935  0.0747  0.0545  0.0375  0.0192
1828         0.0174  0.0357  0.0541  0.0734  0.0862  0.1082  0.1280 *0.1459  0.1645  0.1866
2018         0.0000 *0.4728  0.0050  0.0053  0.4899  0.0001  0.0055  0.0167  0.0040  0.0007
227          0.0038  0.0454  0.0218  0.0187  0.1241  0.0056  0.0168 *0.7455  0.0136  0.0047
2116        *0.4795  0.0046  0.0014  0.0018  0.5001  0.0002  0.0030  0.0065  0.0026  0.0003
2122         0.4956  0.0045  0.0017  0.0027 *0.4874  0.0002  0.0016  0.0041  0.0012  0.0009
```

Figure A.3. *FRANz output of the missing value Gibbs sampler. This output lists all genotypes with missing data on a particular locus of the Penaeus monodon dataset (Sec. 3.5.1). The numbers represent the fraction of MCMC sampled pedigrees with the alleles. The star marks the alleles of the Maximum Likelihood pedigree.*

```
Offspring,Loci Typed,Parent 1,Loci Typed,Parent 2,Loci Typed,LOD,Posterior,Common Loci Typed,
   Mismatches,n_f,n_m,Pair LOD Parent 1,Pair LOD Parent 2
292,7,,,,,0.000000E+00,1.0000,7,0,12,12,,
218,7,,,,,0.000000E+00,1.0000,7,0,12,12,,
1590,7,218,7,292,7,1.969629E+01,0.9972,7,0,42,42,7.941741E+00,8.982472E+00
1592,7,218,7,292,7,1.998774E+01,0.7098,7,0,42,42,6.388146E+00,9.094778E+00
1594,7,218,7,292,7,2.039001E+01,0.8495,7,0,42,42,6.995817E+00,9.390651E+00
1596,7,218,7,292,7,1.749679E+01,0.9824,7,0,42,42,7.477002E+00,8.450023E+00
288,7,,,,,0.000000E+00,1.0000,7,0,12,12,,
215,7,,,,,0.000000E+00,1.0000,7,0,12,12,,
1612,7,215,7,288,7,1.884399E+01,0.9839,7,0,42,42,8.802218E+00,6.061792E+00
1614,7,215,7,288,7,1.908858E+01,0.8233,7,0,42,42,8.620373E+00,6.475110E+00
1616,7,1618,7,1614,7,2.188524E+01,0.6490,7,0,42,42,9.379043E+00,1.181620E+01
1618,7,215,7,288,7,1.881211E+01,0.9797,7,0,42,42,1.029590E+01,4.392019E+00
1622,7,215,7,288,7,1.627626E+01,0.8404,7,0,42,42,7.593334E+00,5.246640E+00
```

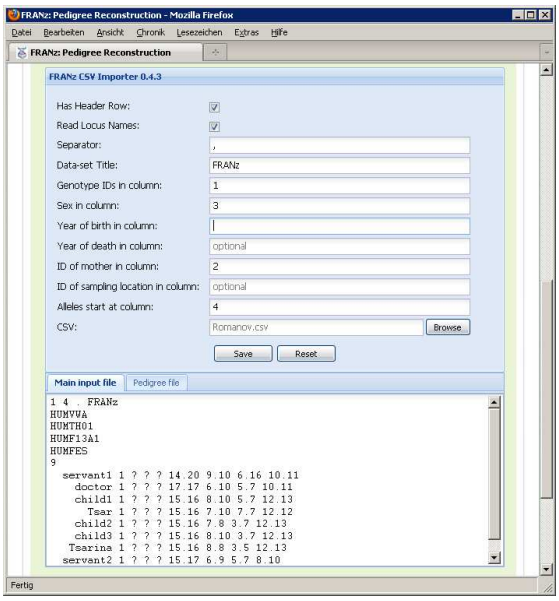
Figure A.4. *FRANz output of the parentages. This output lists the LOD scores (Sec. 3.3.4), posterior probabilities and additional information such as the number of typed loci and mismatches. It is formatted as CSV file to be read in Excel for example. Here we show the parentages of two families again of the Penaeus monodon dataset.*

A.4 Web 2.0 Interface

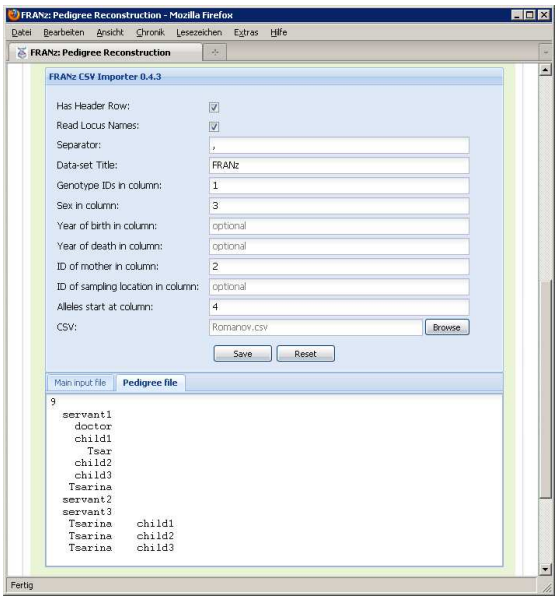
A user-friendly interface for FRANz is available on the FRANz website. Data formatted as CSV files can be uploaded and input files (A.2.1 and A.2.2) are then generated. In Fig. A.5, we show how to convert the Romanov data [Gill et al. 1994] into FRANz input files.

```
Name,Mother,Sex,HUMVWAa,HUMVWAb,HUMTH01a,HUMTH01b,
   HUMF13A1a,HUMF13A1b,HUMFESa,HUMFESb
servant1,,?,14,20,9,10,6,16,10,11
doctor,,?,17,17,6,10,5,7,10,11
child1,Tsarina,?,15,16,8,10,5,7,12,13
Tsar,,?,15,16,7,10,7,7,12,12
child2,Tsarina,?,15,16,7,8,3,7,12,13
child3,Tsarina,?,15,16,8,10,3,7,12,13
Tsarina,,?,15,16,8,8,3,5,12,13
servant2,,?,15,17,6,9,5,7,8,10
servant3,,?,16,17,6,6,6,7,11,12
```

(a)



(b)



(c)

Figure A.5. *FRANz Web 2.0 Interface. This figure demonstrates how to convert CSV formatted data, here in A.5a from the Romanov family [Gill et al. 1994] into FRANz input files. The user specifies the file and which columns contain the supported ecological information (sex, years of birth and death, known mothers) in A.5b. If some mothers are known, the resulting pedigree infile is also generated (A.5c).*

List of Figures

List of Tables

List of Abbreviations

AML	Acute Myeloid Leukemia
ANOVA	Analysis of Variance
BH	Benjamini-Hochberg
BME	Balanced Minimum Evolution
DAG	Directed Acyclic Graph
FAB	French-American-British
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
hESC	human Embryonic Stem Cell
hMSC	human Mesenchymal Stem Cell
HWE	Hardy-Weinberg Equilibrium
IBD	Identical by Descent
KW	Kruskal-Wallis
LOD	Difference in Log-likelihood
MCMC	Markov chain Monte Carlo
MCMCMC	Metropolis-Coupled Markov Chain Monte Carlo
MDS	Myelodysplastic Syndrome
ME	Minimum Evolution
MFH	Malignant Fibrous Histiocytoma
MSA	Multiple Sequence Alignment
MST	Minimum Spanning Tree
NJ	Neighbor-Joining
PIC	Polymorphic Information Content
SA	Simulated Annealing
SE	Standard Error
SNP	Single Nucleotide Polymorphism
SOM	Self-organizing Map

SR	Sampling Rate
STR	Short Tandem Repeat
UPGMA	Unweighted Pair Group Method with Arithmetic mean
WLS	Weighted Least Squares

Bibliography

Aarts, E. and J. Korst (1989). *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. Chichester: Wiley.

Ally, D., K. Ritland, and S. P. Otto (2008, Nov). Can clone size serve as a proxy for clone age? An exploration using microsatellite divergence in *Populus tremuloides*. *Mol Ecol 17*(22), 4897–4911.

Almudevar, A. (2003, Mar). A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theor Popul Biol 63*, 63–75.

Almudevar, A. (2007, Mar). A graphical approach to relatedness inference. *Theor Popul Biol 71*(2), 213–229.

Altekar, G., S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist (2004, Feb). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics 20*, 407–415.

American Cancer Society (2008). *Cancer Facts & Figures 2008*. American Cancer Society.

Anderson, E. C. and J. C. Garza (2006, Apr). The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics 172*(4), 2567–2582.

Barberi, T., M. Bradbury, Z. Dincer, G. Panagiotakos, N. D. Socci, and L. Studer (2007, May). Derivation of engraftable skeletal myoblasts from human embryonic stem cells. *Nat Med 13*(5), 642–648.

Barretina, J., B. S. Taylor, A. H. Ramos, M. Lagos-Quintana, S. Banerji, P. DeCarolis, K. Shah, N. D. Socci, B. A. Weir, A. Ho, D. Y. Chiang, B. Reva, C. Mermel, G. Getz, Y. Antipin, R. Beroukhim, J. E. Major, C. Hatton, R. Nicoletti, M. Hanna, T. Sharpe, T. Fennell, K. Cibulskis, R. C. Onofrio, T. Saito, N. N. Shukla, C. Lau, S. Nelander, S. Silver, C. Sougnez, A. Viale, W. Winckler, R. G. Maki, L. A. Garraway, A. Lash,

H. Greulich, D. Root, W. R. Sellers, G. K. Schwartz, C. R. Antonescu, E. S. Lander, H. E. Varmus, M. Ladanyi, C. Sander, M. Meyerson, and S. Singer (2010). Subtype-specific genomic alterations define new targets for soft tissue sarcoma therapy. In revision.

Barrett, T., D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar (2007, Jan). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res 35*(Database issue), 760–765.

Beerenwinkel, N., J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer (2005, Jul-Aug). Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol 12*(6), 584–598.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*, 289–300.

Bennett, J. M., D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, and C. Sultan (1976, Aug). Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br J Haematol 33*(4), 451–458.

Beqqali, A., J. Kloots, D. Ward-van Oostwaard, C. Mummery, and R. Passier (2006, Aug). Genome-wide transcriptional profiling of human embryonic stem cells differentiating to cardiomyocytes. *Stem Cells 24*(8), 1956–1967.

Berger-Wolf, T. Y., S. I. Sheikh, B. DasGupta, M. V. Ashley, I. C. Caballero, W. Chaovalitwongse, and S. L. Putrevu (2007, Jul). Reconstructing sibling relationships in wild populations. *Bioinformatics 23*, 49–56.

Blouin, M. S. (2003, Oct). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution 18*(10), 503–511.

Boehnke, M. and N. J. Cox (1997, Aug). Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet. 61*, 423–429.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed (2003, Jan). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics 19*(2), 185–193.

Bonin, A., E. Bellemain, P. Bronken Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet (2004, Nov). How to track and assess genotyping errors in population genetics studies. *Mol. Ecol. 13*, 3261–3273.

Bordewich, M., O. Gascuel, K. T. Huber, and V. Moulton (2009, Jan-Mar). Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. *IEEE/ACM Trans Comput Biol Bioinform 6*(1), 110–117.

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis (1980, May). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet 32*(3), 314–331.

Bruno, W. J., N. D. Socci, and A. L. Halpern (2000, Jan). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol Evol 17*(1), 189–197.

Burczyk, J., W. T. Adams, and J. Y. Shimizu (1996). Mating patterns and pollen dispersal in a natural knobcone pine (*Pinus attenuate Lemmon.*) stand. *Heredity 77*, 251–260.

Butler, J. M. (2006, Mar). Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci 51*(2), 253–265.

Cannings, C. and N. A. Sheehan (2002, Oct). On a misconception about irreducibility of the single-site gibbs sampler in a pedigree application. *Genetics 162*(2), 993–996.

Ceppellini, R., M. Siniscalco, and C. A. Smith (1955, Oct). The estimation of gene frequencies in a random-mating population. *Ann Hum Genet 20*(2), 97–115.

Cercueil, A., E. Bellemain, and S. Manel (2002, Nov-Dec). PARENTE: computer program for parentage analysis. *J Hered 93*(6), 458–459.

Coble, M. D., O. M. Loreille, M. J. Wadhams, S. M. Edson, K. Maynard, C. E. Meyer, H. Niederstätter, C. Berger, B. Berger, A. B. Falsetti, P. Gill, W. Parson, and L. N. Finelli (2009, Mar). Mystery solved: the identification of the two missing romanov children using DNA analysis. *PLoS One 4*(3), e4838.

Cooper, G. F. and E. Herskovits (1992, Oct). A bayesian method for the induction of probabilistic networks from data. *Machine Learning 09*(4), 309–347.

Cowell, R. G. (2009, Dec). Efficient maximum likelihood pedigree reconstruction. *Theor Popul Biol 76*(4), 285–291.

Dalal, K. M., M. W. Kattan, C. R. Antonescu, M. F. Brennan, and S. Singer (2006, Sep). Subtype specific prognostic nomogram for patients with primary liposarcoma of the retroperitoneum, extremity, or trunk. *Ann Surg 244*(3), 381–391.

David, P., B. Pujol, F. Viard, V. Castella, and J. Goudet (2007, Jun). Reliable selfing rate estimates from imperfect population genetic data. *Mol Ecol 16*(12), 2474–2487.

de Meeûs, T. and F. Balloux (2004, Dec). Clonal reproduction and linkage disequilibrium in diploids: a simulation study. *Infect Genet Evol 4*(4), 345–351.

Desmedt, C., F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse, C. Sotiriou, and TRANSBIG Consortium (2007, Jun). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res 13*(11), 3207–3214.

Desper, R. and O. Gascuel (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol 9*(5), 687–705.

Desper, R. and O. Gascuel (2004, Mar). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol Biol Evol 21*(3), 587–598.

Desper, R., F. Jiang, O.-P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer (1999). Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol 6*(1), 37–51.

Desper, R., F. Jiang, O.-P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer (2000). Distance-based reconstruction of tree models for oncogenesis. *J Comput Biol 7*(6), 789–803.

Desper, R., J. Khan, and A. A. Schäffer (2004, Jun). Tumor classification using phylogenetic methods on expression data. *J Theor Biol 228*(4), 477–496.

Devlin, B., K. Roeder, and N. Ellstrand (1988, Sep). Fractional paternity assignment: theoretical development and comparison to other methods. *TAG Theoretical and Applied Genetics 76*(3), 369–380.

Duchesne, P., T. Castric, and L. Bernatchez (2005, Sep). PASOS (parental allocation of singles in open systems): a computer program for individual parental allocation with missing parents. *Molecular Ecology Notes 5*(3), 701–704.

Duchesne, P., M. H. Godbout, and L. Bernatchez (2002). PAPA (package for the analysis of parental allocation): a computer program for simulated and real parental allocation. *Molecular Ecology Notes 2*(2), 191–193.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998, Dec). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A 95*(25), 14863–14868.

Ellstrand, N. C. and D. L. Marshall (1985). Interpopulation gene flow in *Raphanus sativus*. *American Naturalist 126*, 606–616.

Emery, A. M., I. J. Wilson, S. Craig, P. R. Boyle, and L. R. Noble (2001, May). Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Mol Ecol 10*(5), 1265–1278.

Federal Statistical Office (2007). *Statistical Yearbook 2007 For the Federal Republic of Germany*. Wiesbaden: Federal Statistical Office. ISBN: 978-3-8246-0803-4.

Felsenstein, J. (1989). PHYLIP (phylogeny inference package) version 3.2. *Cladistics 5*, 164–166.

Felsenstein, J. (2003, Sep). *Inferring Phylogenies* (2 ed.). Sinauer Associates.

Fernández, J. and M. A. Toro (2006, May). A new method to estimate relatedness from molecular markers. *Mol. Ecol. 15*, 1657–1667.

Fishman, G. S. (2005, Sep). *A First Course in Monte Carlo*. Brooks Cole.

Fitch, W. M. and E. Margoliash (1967, Jan). Construction of phylogenetic trees. *Science 155*(760), 279–284.

Ford, M. J. and K. S. Williamson (2010, Jan-Feb). The aunt and uncle effect revisited–the effect of biased parentage assignment on fitness estimation in a supplemented salmon population. *J Hered 101*(1), 33–41.

Futschik, M. E. and B. Carlisle (2005, Aug). Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol 3*(4), 965–988.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol 5*(10), R80+.

Gerber, S., P. Chabrier, and A. Kremer (2003). FaMoz: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. *Molecular Ecology Notes 3*(3), 479–481.

Gerber, S., S. Mariette, R. Streiff, C. Bodenes, and A. Kremer (2000). Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. *Molecular Ecology 9*(8), 1037–1048.

Geyer, C. J. (1991). Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of 23rd Symposium on the Interface Interface Foundation*, pp. 156–163. Fairfax Station.

Gill, P., P. L. Ivanov, C. Kimpton, R. Piercy, N. Benson, G. Tully, I. Evett, E. Hagelberg, and K. Sullivan (1994, Feb). Identification of the remains of the romanov family by DNA analysis. *Nat Genet 6*(2), 130–135.

Glaubitz, J. C., O. E. Rhodes, and J. A. Dewoody (2003, Apr). Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Mol Ecol 12*(4), 1039–1047.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999, Oct). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*(5439), 531–537.

Graham, S. W., R. G. Olmstead, and S. C. Barrett (2002, Oct). Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Mol Biol Evol 19*(10), 1769–1781.

Grünewald, S., K. Forslund, A. Dress, and V. Moulton (2007, Feb). QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. *Mol Biol Evol 24*(2), 532–538.

Guo, S. W. and E. A. Thompson (1992, Jun). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics 48*, 361–372.

Gutiérrez, N. C., R. López-Pérez, J. M. Hernández, I. Isidro, B. González, M. Delgado, E. Fermiñán, J. L. García, L. Vázquez, M. González, and J. F. San Miguel (2005, Mar). Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia. *Leukemia 19*(3), 402–409.

Hadfield, J. D., D. S. Richardson, and T. Burke (2006, Oct). Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Mol. Ecol. 15*, 3715–3730.

Hanahan, D. and R. A. Weinberg (2000, Jan). The hallmarks of cancer. *Cell 100*(1), 57–70.

Hartl, D. L. and A. G. Clark (2007). *Principles of Population Genetics, Fourth Edition*. Sinauer Associates, Inc.

Holland, B. R., K. T. Huber, V. Moulton, and P. J. Lockhart (2004, Jul). Using consensus networks to visualize contradictory evidence for species phylogeny. *Mol Biol Evol 21*(7), 1459–1461.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Huelsenbeck, J. P. and F. Ronquist (2001, Aug). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics 17*, 754–755.

Huson, D. H. and D. Bryant (2006, Feb). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol 23*(2), 254–267.

Huson, D. H., D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics 8*, 460–460.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed (2003, Apr). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*(2), 249–264.

Iwasa, Y., F. Michor, N. L. Komarova, and M. A. Nowak (2005, Mar). Population genetics of tumor suppressor genes. *J Theor Biol 233*(1), 15–23.

Jamieson, A. and S. C. Taylor (1997, Dec). Comparisons of three probability formulae for parentage exclusion. *Anim. Genet. 28*, 397–400.

Jarne, P. and P. David (2008, Apr). Quantifying inbreeding in natural populations of hermaphroditic organisms. *Heredity 100*(4), 431–439.

Jerry, D. R., B. S. Evansa, M. Kenwayb, and K. Wilson (2006, May). Development of a microsatellite DNA parentage marker suite for black tiger shrimp penaeus monodon. *Aquaculture 255*(1-4), 542–547.

Jin, L., M. L. Baskett, L. L. Cavalli-Sforza, L. A. Zhivotovsky, M. W. Feldman, and N. A. Rosenberg (2000, Mar). Microsatellite evolution in modern humans: a comparison of two data sets from the same populations. *Ann. Hum. Genet. 64*, 117–134.

Jones, A. G. (2005). GERUD 2.0: a computer program for the reconstruction of parental genotypes from half-sib progeny arrays with known or unknown parents. *Mol. Ecol. Notes 5*, 708–711.

Jones, A. G. and W. R. Ardren (2003, Oct). Methods of parentage analysis in natural populations. *Mol. Ecol. 12*, 2511–2523.

Jones, A. G., C. M. Small, K. A. Paczolt, and N. L. Ratterman (2009). A practical guide to methods of parentage analysis. *Mol. Ecol. Resources 10*(1), 6–13.

Jones, B. (2003, Sep). Balancing population size and genetic information in parentage analysis studies. *Biometrics 59*, 694–700.

Jones, B., G. D. Grossman, D. C. Walsh, B. A. Porter, J. C. Avise, and A. C. Fiumera (2007, Aug). Estimating differential reproductive success from nests of related individuals, with application to a study of the mottled sculpin, cottus bairdi. *Genetics 176*(4), 2427–2439.

Kalinowski, S. T. and M. L. Taper (2006). Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics 7*, 991–995.

Kalinowski, S. T., M. L. Taper, and T. C. Marshall (2007, Mar). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol. 16*, 1099–1106.

Kapp, A. V., S. S. Jeffrey, A. Langerod, A. L. Borresen-Dale, W. Han, D. Y. Noh, I. R. Bukholm, M. Nicolau, P. O. Brown, and R. Tibshirani (2006). Discovery and validation of breast cancer subtypes. *BMC Genomics 7*, 231–231.

Katz, F. E., R. Tindle, D. R. Sutherland, and M. F. Greaves (1985). Identification of a membrane glycoprotein associated with haemopoietic progenitor cells. *Leuk Res 9*(2), 191–198.

Kimura, M. (1968, Feb). Evolutionary rate at the molecular level. *Nature 217*(5129), 624–626.

Koch, M., J. D. Hadfield, K. M. Sefc, and C. Sturmbauer (2008, Oct). Pedigree reconstruction in wild cichlid fish populations. *Mol Ecol 17*(20), 4500–4511.

Kohonen, T. (1990, Sep). The self-organizing map. *Proceedings of the IEEE 78*(9), 1464–1480.

Kohonen, T., J. Hynninen, J. Kangas, and J. Laaksonen (1996). SOM PAK: The self-organizing map program package. Technical Report TKK-F-A31.

Kooby, D. A., C. R. Antonescu, M. F. Brennan, and S. Singer (2004, Jan). Atypical lipomatous tumor/well-differentiated liposarcoma of the extremity and trunk wall: importance of histological subtype with treatment recommendations. *Ann Surg Oncol 11*(1), 78–84.

Kruskal, W. H. and W. A. Wallis (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association 47*, 583–621.

Kuhner, M. K. and J. Felsenstein (1994, May). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol 11*(3), 459–468.

Levinson, G. and G. A. Gutman (1987, May). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol. 4*, 203–221.

Lin, T. H., E. W. Myers, and E. P. Xing (2006, Jul). Interpreting anonymous DNA samples from mass disasters–probabilistic forensic inference using genetic markers. *Bioinformatics 22*, 298–306.

Mack, T. M. (1995, Jan). Sarcomas and other malignancies of soft tissue, retroperitoneum, peritoneum, pleura, heart, mediastinum, and spleen. *Cancer 75*(1 Suppl), 211–244.

Margush, T. and F. R. Mcmorris (1981). Consensus n-trees. *Bulletin of Mathematical Biology 43*, 239–244.

Marot, G., J. L. Foulley, C. D. Mayer, and F. Jaffrézic (2009, Oct). Moderated effect size and p-value combinations for microarray meta-analyses. *Bioinformatics 25*(20), 2692–2699.

Marshall, T. C., J. Slate, L. E. Kruuk, and J. M. Pemberton (1998, May). Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol. 7*, 639–655.

Matsumoto, M. and T. Nishimura (2000). Dynamic creation of pseudorandom number generators. In *Monte Carlo and Quasi-Monte Carlo Methods 1998: Proceedings of a Conference, Held at Claremont Graduate University, Claremont, California, USA*, pp. 56–69. Springer.

Matushansky, I., E. Hernando, N. D. Socci, T. Matos, J. Mills, M. A. Edgar, G. K. Schwartz, S. Singer, C. Cordon-Cardo, and R. G. Maki (2008, Apr). A developmental model of sarcomagenesis defines a differentiation-based classification for liposarcomas. *Am J Pathol 172*(4), 1069–1080.

Meagher, T. R. (1991). Analysis of paternity within a natural population of *Chamaelirium luteum*. II. Patterns of male reproductive success. *The American Naturalist 137*(6), 738–752.

Meagher, T. R., F. C. Belanger, and P. R. Day (2003, Jun). Using empirical data to model transgene dispersal. *Philos Trans R Soc Lond B Biol Sci 358*(1434), 1157–1162.

Meagher, T. R. and E. A. Thompson (1986, Feb). The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology 29*(1), 87–106.

Meagher, T. R. and E. A. Thompson (1987). Analysis of parentage for naturally established seedlings of *Chamaelirium Luteum* (Liliaceae). *Ecology 68*(4), 803–812.

Merlo, L. M., J. W. Pepper, B. J. Reid, and C. C. Maley (2006, Dec). Cancer as an evolutionary and ecological process. *Nat Rev Cancer 6*(12), 924–935.

Metzeler, K. H., M. Hummel, C. D. Bloomfield, K. Spiekermann, J. Braess, M. C. Sauerland, A. Heinecke, M. Radmacher, G. Marcucci, S. P. Whitman, K. Maharry, P. Paschka, R. A. Larson, W. E. Berdel, T. Büchner, B. Wörmann, U. Mansmann, W. Hiddemann, S. K. Bohlander, C. Buske, Cancer and Leukemia Group B, and German AML Cooperative Group (2008, Nov). An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood 112*(10), 4193–4201.

Michor, F., Y. Iwasa, and M. A. Nowak (2004, Mar). Dynamics of cancer progression. *Nat Rev Cancer 4*(3), 197–205.

Miller, L. D., J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh (2005, Sep). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A 102*(38), 13550–13555.

Minguell, J. J., A. Erices, and P. Conget (2001, Jun). Mesenchymal stem cells. *Exp Biol Med (Maywood) 226*(6), 507–520.

Morgan, M. T. and J. K. Conner (2001, Feb). Using genetic markers to directly estimate male selection gradients. *Evolution 55*(2), 272–281.

Nakayama, R., T. Nemoto, H. Takahashi, T. Ohta, A. Kawai, K. Seki, T. Yoshida, Y. Toyama, H. Ichikawa, and T. Hasegawa (2007, Jul). Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma. *Mod Pathol 20*(7), 749–759.

Neff, B. D., J. Repka, and M. R. Gross (2001, Jun). A bayesian framework for parentage analysis: the value of genetic and other biological data. *Theor Popul Biol 59*, 315–331.

Newton, M. A. (2002, Dec). Discovering combinations of genomic aberrations associated with cancer. *Journal of the American Statistical Association 97*, 931–942.

Nielsen, R., D. K. Mattila, P. J. Clapham, and P. J. Palsbøll (2001, Apr). Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. *Genetics 157*, 1673–1682.

Nugoli, M., P. Chuchana, J. Vendrell, B. Orsetti, L. Ursule, C. Nguyen, D. Birnbaum, E. J. Douzery, P. Cohen, and C. Theillet (2003, Apr). Genetic variability in MCF-7 sublines: evidence of rapid genomic and RNA expression profile modifications. *BMC Cancer 3*, 13–13.

Osborne, C. K., M. G. Yochmowitz, W. A. Knight, and W. L. McGuire (1980, Dec). The value of estrogen and progesterone receptors in the treatment of breast cancer. *Cancer 46*(12 Suppl), 2884–2888.

Park, Y., S. Shackney, and R. Schwartz (2009, Apr-Jun). Network-based inference of cancer progression from microarray data. *IEEE/ACM Trans Comput Biol Bioinform 6*(2), 200–212.

Pemberton, J. M. (2008, Mar). Wild pedigrees: the way forward. *Proc. Biol. Sci. 275*, 613–621.

Planet, P. J., R. DeSalle, M. Siddall, T. Bael, I. N. Sarkar, and S. E. Stanley (2001, Jul). Systematic analysis of DNA microarray data: ordering and interpreting patterns of gene expression. *Genome Res 11*(7), 1149–1155.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technology Journal 36*, 1389–1401.

Queller, D. C., J. E. Strassmann, and C. R. Hughes (1993). Microsatellites and kinship. *Trends in Ecology & Evolution 8*(8), 285 – 288.

Riester, M., P. F. Stadler, and K. Klemm (2009, Aug). FRANz: Reconstruction of wild multi-generation pedigrees. *Bioinformatics 25*(16), 2134–2139.

Ritland, K. and S. Jain (1981). A model for the estimation of outcrossing rate and gene frequencies using n independent loci. *Heredity 47*(1), 35–52.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Probab 7*(1), 110–120.

Roussett, F. (2008). GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources 8*(1), 103–106.

Ruitberg, C. M., D. J. Reeder, and J. M. Butler (2001, Jan). STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res 29*(1), 320–322.

Rzhetsky, A. and M. Nei (1993, Sep). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol Biol Evol 10*(5), 1073–1095.

Saitou, N. and M. Nei (1987, Jul). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol 4*(4), 406–425.

Sandberg, A. A. (2004, Nov). Updates on the cytogenetics and molecular genetics of bone and soft tissue tumors: liposarcoma. *Cancer Genet Cytogenet 155*(1), 1–24.

Sekiya, I., B. L. Larson, J. T. Vuoristo, J. G. Cui, and D. J. Prockop (2004, Feb). Adipogenic differentiation of human adult stem cells from bone marrow stroma (MSCs). *J Bone Miner Res 19*(2), 256–264.

Sheehan, N. A. (2000). On the application of markov chain monte carlo methods to genetic analyses on complex pedigrees. *International Statistical Review 68*, 83–110.

Sheehan, N. A. and T. Egeland (2007, Jul). Structured incorporation of prior information in relationship identification problems. *Ann. Hum. Genet. 71*, 501–518.

Sheehan, N. A. and A. Thomas (1993, Mar). On the irreducibility of a markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics 49*(1), 163–175.

Signorovitch, J. and R. Nielsen (2002, Feb). PATRI-paternity inference using genetic data. *Bioinformatics 18*(2), 341–342.

Silander, T. and P. Myllymäki (2006). A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia, pp. 445–452. AUAI Press.

Singer, S., C. R. Antonescu, E. Riedel, and M. F. Brennan (2003, Sep). Histologic subtype and margin of resection predict pattern of recurrence and survival for retroperitoneal liposarcoma. *Ann Surg 238*(3), 358–370.

Singer, S., N. D. Socci, G. Ambrosini, E. Sambol, P. Decarolis, Y. Wu, R. O'Connor, R. Maki, A. Viale, C. Sander, G. K. Schwartz, and C. R. Antonescu (2007, Jul). Gene expression profiling of liposarcoma identifies distinct biological types/subtypes and potential therapeutic targets in well-differentiated and dedifferentiated liposarcoma. *Cancer Res 67*(14), 6626–6636.

Smith, B. R., C. M. Herbinger, and H. R. Merry (2001, Jul). Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics 158*(3), 1329–1338.

Smouse, P. E. and T. R. Meagher (1994, Jan). Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) Gray (Liliaceae). *Genetics 136*(1), 313–322.

Smouse, P. E., T. R. Meagher, and C. J. Kobak (1999). Parentage analysis in *Chamaelirium luteum* (L.) Gray (Liliaceae): why do some males have higher reproductive contributions? *Journal of Evolutionary Biology 12*(6), 1069–1077.

Sokal, R. R. and C. D. Michener (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin 38*, 1409–1438.

Sotiriou, C., P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi (2006, Feb). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst 98*(4), 262–272.

Steel, M. and J. Hein (2006, Jun). Reconstructing pedigrees: a combinatorial perspective. *J. Theor. Biol. 240*, 360–367.

Stirewalt, D. L., S. Meshinchi, K. J. Kopecky, W. Fan, E. L. Pogosova-Agadjanyan, J. H. Engel, M. R. Cronk, K. S. Dorcy, A. R. McQuary, D. Hockenbery, B. Wood, S. Heimfeld, and J. P. Radich (2008, Jan). Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer 47*(1), 8–20.

Taggart, J. G. (2007). FAP: an exclusion-based parental assignment program with enhanced predictive functions. *Molecular Ecology Notes 7*(3), 412–415.

Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub (1999, Mar). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A 96*(6), 2907–2912.

Tenen, D. G. (2003, Feb). Disruption of differentiation in human cancer: AML shows the way. *Nat Rev Cancer 3*(2), 89–101.

Thatte, B. D. and M. Steel (2007, Dec). Reconstructing pedigrees: A stochastic perspective. *J. Theor. Biol. 251*(3), 440–449.

Thomas, S. C. and W. G. Hill (2000, Aug). Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics 155*, 1961–1972.

Thomas, S. C. and W. G. Hill (2002, Jun). Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genet. Res. 79*, 227–234.

Thompson, E. A. (1976). Inference of genealogical structure. *Social Science Information 15*(2-3), 477–526.

Thompson, E. A. (2000). *Statistical inference from genetic data on pedigrees*. IMS, Beachwood, OH: NSF-CBMS Regional Conference Series in Probability and Statistics.

Thompson, E. A. and T. R. Meagher (1987, Sep). Parental and sib likelihoods in genealogy reconstruction. *Biometrics 43*, 585–600.

Tripathi, A., C. King, A. de la Morenas, V. K. Perry, B. Burke, G. A. Antoine, E. F. Hirsch, M. Kavanah, J. Mendez, M. Stone, N. P. Gerry, M. E. Lenburg, and C. L. Rosenberg (2008, Apr). Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *Int J Cancer 122*(7), 1557–1566.

Uddin, M., D. E. Wildman, G. Liu, W. Xu, R. M. Johnson, P. R. Hof, G. Kapatos, L. I. Grossman, and M. Goodman (2004, Mar). Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proc Natl Acad Sci U S A 101*(9), 2957–2962.

von Heydebreck, A., B. Gunawan, and L. Füzesi (2004, Oct). Maximum likelihood estimation of oncogenetic tree models. *Biostatistics 5*(4), 545–556.

Vouillamoz, J. F. and M. S. Grando (2006, Aug). Genealogy of wine grape cultivars: "Pinot" is related to "Syrah". *Heredity 97*, 102–110.

Waddell, P. J. and H. Kishino (2000). Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Inform Ser Workshop Genome Inform 11*, 129–140.

Wang, J. (2004a, Oct). Estimating pairwise relatedness from dominant genetic markers. *Mol. Ecol. 13*, 3169–3178.

Wang, J. (2004b, Apr). Sibship reconstruction from genetic data with typing errors. *Genetics 166*, 1963–1979.

Wang, J. (2007, Aug). Parentage and sibship exclusions: higher statistical power with more family members. *Heredity 99*(2), 205–217.

Wang, J. and A. W. Santure (2009, Apr). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics 181*(4), 1579–1594.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika 38*, 330–336.

Wilson, I. J. and K. J. Dawson (2007, Nov). A markov chain monte carlo strategy for sampling from the joint posterior distribution of pedigrees and population parameters under a fisher-wright model with partial selfing. *Theor Popul Biol 72*(3), 436–458.

Wright, J. W. and T. R. Meagher (2004, Mar). Selection on floral characters in natural spanish populations of *Silene latifolia*. *J Evol Biol 17*(2), 382–395.

Xu, Y., V. Olman, and D. Xu (2002, Apr). Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics 18*(4), 536–545.

Curriculum vitae

Name:	Markus Riester
Date of birth:	March 31, 1979
Place of birth:	Villingen-Schwenningen, Germany

Education:

Since January 2007	Scientific assistant (PhD student) at the University of Leipzig (Germany), Department of Computer Science, Bioinformatics Group
August 2000 - December 2006	Study of Bioinformatics at the University of Tübingen (Germany); Degree: Diploma
September 1999 - July 2000	Mandatory Civilian Service
July 1999	Abitur Otto-Hahn-Gymnasium Tuttlingen (Diploma qualifying for university admission or matriculation)

Publications

- Riester, M., Stadler, P. F. & Klemm, K. (2010). Reconstruction of pedigrees in clonal plant populations. *Theoretical Population Biology.* In press.
- Riester, M.*, Stephan-Otto Attolini, C.*, Downey, R. J., Singer, S. & Michor, F. (2010). A differentiation-based phylogeny of cancer subtypes. *PLoS Computational Biology.* 6(5): e1000777. *Equal contribution.
- Riester, M., Stadler, P. F. & Klemm K. (2009). `FRANz`: Reconstruction of wild multi-generation pedigrees. *Bioinformatics* **25**, 2134.
- Riester, M., Stadler, P. F. & Klemm K. (2008). `FRANz`: Fast reconstruction of wild pedigrees. *Lecture Notes in Informatics* P-136, 168 (Proceedings of the German Conference on Bioinformatics).

Conference Talks

- `FRANz`: Fast reconstruction of wild pedigrees. *German Conference on Bioinformatics 2008.*

Contributions

- Tanzer, A., Riester, M., Hertel, J., Bermudez-Santana, C.I., Gorodkin, J., Hofacker, I.L. & Stadler, P.F. (2010). Evolutionary genomics of microRNAs and their relatives. *In Evolutionary Genomics and Systems Biology, edited by Gustavo Caetano-Anollés.* Wiley-Blackwell, Hoboken, pp 295-327.
- Dreyer, C., Hoffmann, M., Lanz, C., Willing, E.M., Riester, M., Warthmann, N., Sprecher, A., Tripathi, N., Henz, S.R. & Weigel, D. (2007) ESTs and EST-linked polymorphisms for genetic mapping and phylogenetic reconstruction in the guppy, *Poecilia reticulata*. *BMC Genomics.* **8**, 269.
- Schwab, R., Ossowski, S., Riester, M., Warthmann, N. & Weigel, D. (2006). Highly specific gene silencing by artificial microRNAs in Arabidopsis. *Plant Cell* **18** , 1121-33.
- Schwab, R., Palatnik, J.F., Riester, M., Schommer, C., Schmid, M. & Weigel D (2005). Specific effects of microRNAs on the plant transcriptome. *Dev Cell* **8** , 517-27.

Scientific Cooperations

- Camille Stephan-Otto Attolini & Franziska Michor, MSKCC New York City; Cancer Phylogenies
- Filipe Alberto & Ester Serrão, CIMAR Faro; Accepted grant for a project called "The sexual and asexual balance of a marine clonal plant: the case of the seagrass Cymodocea nodosa in the Canary Islands" (January 1th 2010- December 31th 2012).

(Programming) Languages

- ANSI C, C++, Perl, Java, LaTeX, Postscript, HTML, PHP, Javascript
- German, English, French

Eigenständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder Dienstleistungen als solche gekennzeichnet.

Leipzig, den 7. April 2010

Markus Riester