Bioinformatics Advance Access published August 26, 2016

A likelihood ratio based method to predict exact pedigrees for complex families from next-generation sequencing data

Verena Heinrich [1,2,*], Tom Kamphans [3], Stefan Mundlos [1,2], Peter N. Robinson [2], Peter M. Krawitz [2]

[1]Max Planck Institute for Molecular Genetics, Ihnestraße 63-73 14195 Berlin
[2]Institute for Medical Genetics and Human Genetics, Charité Universitätsmedizin Berlin, Augustenburger Platz 1 13353 Berlin
[3]Smart Algos, Germany

Associate Editor: Prof. Alfonso Valencia

ABSTRACT

Motivation: Next generation sequencing (NGS) technology considerably changed the way we screen for pathogenic mutations in rare Mendelian disorders. However, the identification of the disease-causing mutation amongst thousands of variants of partly unknown relevance is still challenging and efficient techniques that reduce the genomic search space play a decisive role. Often segregation- or linkage analysis are used to prioritize candidates, however, these approaches require correct information about the degree of relationship among the sequenced samples. For quality assurance an automated control of pedigree structures and sample assignment is therefore highly desirable in order to detect label mix-ups that might otherwise corrupt downstream analysis.

Results: We developed an algorithm based on likelihood ratios that discriminates between different classes of relationship for an arbitrary number of genotyped samples. By identifying the most likely class we are able to reconstruct entire pedigrees iteratively, even for highly consanguineous families. We tested our approach on exome data of different sequencing studies and achieved high precision for all pedigree predictions. By analyzing the precision for varying degrees of relatedness or inbreeding we could show that a prediction is robust down to magnitudes of a few hundred loci.

Availability: A java standalone application that computes the relationships between multiple samples as well as a Rscript that visualizes the pedigree information is available for download as well as a web service at www.gene-talk.de.

Contact: heinrich@molgen.mpg.de

1 INTRODUCTION

In recent years high-throughput sequencing approaches have been successfully applied to identify thousands of novel pathogenic mutations in genetic disorders. However, due to the high variability of the human genome, there are several hundred variants of unclear significance in each individual and the analysis of such high dimensional data is still challenging. In order to reduce the search space, related individuals are usually sequenced for filtering purposes or linkage analysis, when studying unknown disorders. Even in routine diagnostics sequencing of additional family members is common practice, whenever the disorder is highly heterogeneous and *de novo* mutations are the most promising candidates (Veltman and Brunner, 2012).

However, these approaches rely on correct pedigree information and thus there is a great need for robust and easily manageable methods for checking the relationship between samples. So far it is common practice to infer relatedness with different kinds of genetic markers and a comprehensive overview of the existing methods is given by Blouin (2003) and Pemberton (2008).

Most of the approaches that are based on likelihood ratios, LR, test different pre-defined relationship models and have already been thoroughly discussed by others (Brenner (1997), Marshall *et al.* (1998), Aoki *et al.* (2001)). In our work we study the effectiveness of pedigree predictions when also rare markers are used, that become only accessible by direct sequencing.

In general the accuracy of the prediction increases with the information content of the available markers. Microsatellites are very variable in length and are thus highly informative. With 20 to 30 unlinked microsatellites it is usually possible to achieve accurate predictions about the relationship between samples. In contrast, SNVs are mostly biallelic and in exomes their mean heterozygosity is only around 0.3. On the other hand, large numbers of SNVs are detected in high-throughput sequencing projects and we will study in this work how the accuracy of pedigree predictions scales with the number of such markers.

All models that have been developed for the prediction of relationships also have to account for genotyping errors. Microsatellites are more difficult to genotype than SNVs (Pompanon *et al.*, 2005). Additionally, the intrinsically higher rates of mutations can introduce biases in analyses that rely on identity by state, IBS (Hardy *et al.*, 2003). Due to these technical challenges the evaluation of microsatellite data requires a high human expertise to avoid false parentage exclusions as discussed in Dakin and Avise

*to whom correspondence should be addressed

(2004) and it is also more difficult to automate these methods computationally.

In most approaches that are used to predict relationships, marker sets are defined prior to screening the samples. In order to maximize the probability that a marker is informative, SNPs or SSRs with a high heterozygosity are therefore usually chosen. However, certainly bi-allelic markers with a high population frequency may also be IBS by pure chance.

With whole exomes and other types of reference-guided resequencing data, there is no need for choosing marker loci $a\ priori$, as variant calls are plentiful. Instead we can simply consider all positions where the genotype of at least one of the samples differs from the reference sequence. The potential of such rare SNVs for pedigree prediction has not been studied so far.

Most existing methods that work with genome data rely on haplotype reconstruction and can certainly achieve high precision (He *et al.*, 2013). Unfortunately, for exome data and other enrichment based NGS data sets, the derivation of long haplotypes is often not possible.

We analyzed systematically likelihood-based approaches to reconstruct entire pedigrees that also take into account rare markers and we found that especially rare variants are well suited for assessing second-order relations. By this means we achieve high precision in pedigree predictions and present a tool that can easily be integrated in existing analysis pipelines to ensure quality and avoid sample mix-ups.

2 APPROACH

We developed a method to reconstruct and visualize complex pedigree structures for any given number of individuals that is based on likelihood ratios of different models of relationship types for any combination of two samples (dyad). For any possible dyad we test the following models of kinship:

0 unrelated

1 technical/biological replicates (or identical twins)

2 full siblings

3 parent-child

4 second-order relationship

The different degrees of relationship reflect the expected values for the proportion of the genome that is shared between two individuals. This can also be expressed by the coefficient of relationship which defines the probability that an allele at an arbitrary locus in the genome originates from the same common ancestor ((Wright, 1922)). For instance, we would expect that a parent and a child as well as siblings share 50% of their genetic material and thus the same coefficient of relationships. In second order relationships as e.g. in a grandparent-grandchild, or uncle-nephew dyad only 25% of the alleles are identical by descent, IBD, in an outbred population. On the contrary, in highly inbred populations and consanguineous marriages the expected coefficients of relationship of 0.5 and 0.25 may deviate upwards substantially. In the following we will also use the coefficient of relationship to discriminate between different degrees of kinship and refer to *parent-child* and *full siblings* as first-order relationship and classify half-siblings, grandparent-grandchild and aunt/uncle-nephew/niece as *second-order* relationships. Additionally we are defining the model *technical/biological replicates* to identify samples that have a relationship coefficient close to 1, which is either the case if the same sample has been processed twice or if two samples are from identical twins. For each possible dyad all these models are tested and compared to a null hypothesis that assumes the individuals are unrelated, which is explained in more detail in the next section. The computation of the probabilities of all hypotheses requires knowledge about the expected allele frequencies in a population. In this work, if an allele was observed at least once in the 1000 genomes project (Altshuler *et al.* (2010), The 1000 Genomes Project Consortium (2012)) we use the respective frequency. The most likely relatedness class is then identified by the largest likelihood ratio among all hypotheses for one dyad.We provide a tool for our algorithm that works on genotype data presented in VCF format ((Danecek *et al.*, 2011)) and that generates an intermediate output that we refer to as extended ped format: for each individual the most likely relationships to the remaining individuals are recorded. In addition to mother and father, also children and sibling relationships are listed (see Supplemental Methods). An R Script takes the predicted classes of relationship as input and visualizes the result in a pedigree graph. We tested our method on family exome data from the 1000 genomes project (referred to as 1KG data in the following), as well as on in-house data. For the systematic analysis of performance at different degrees of inbreeding as well as increasing degrees of error rates we also simulated samples based on sequencing data from real individuals. The java application, as well as the visualizing R script, is integrated in GeneTalk ((Kamphans *et al.*, 2013)), and also available as a standalone application at *www.gene-talk.de/vcf2ped*. We analyzed the performance of our approach for a decreasing number of exomic and genomic markers. For exomic markers that are mostly unlinked we achieved high accuracy for all relationships (first and second order) with a minimum of 10,000 markers. The same accuracy can be achieved with a comparable number of genomic markers when randomly distributed over the genome (data not shown).

3 MATERIALS AND METHODS

Data processing and simulation of technical replicates

The implementation was tested on publicly available data of the 1KG project as well as on data that had been sequenced in-house over the past years with the approval of the ethical Board of the Charité. All in-house samples were sequenced on an Illumina HiSeq 2000 and the resulting reads were mapped to the hg19 reference genome using BWA-MEM (Li, 2013). Multi-sample variant calling was performed on 39 families with different population backgrounds, downloaded from the ftp server of the 1KG project, as well as on in-house families, including three families which show a high degree of consanguinity within the pedigree, with GATK (Mckenna *et al.*, 2010) (version 2.7-2) using standard Best Practices recommendations (DePristo *et al.*, 2011). To guarantee a fair comparison between all analysed sequencing samples we restricted all variant calls to the pilot 3 consensus exonic target region defined by the 1KG project. We defined a minimum coverage of 5 reads per base in a homozygous state and a minimum of 5 reads per allele for heterozygous variants.

All individuals of the *1KG* project are subdivided into single families of $2, 3$ or 4 family members, which are connected via different degrees of relationship (*siblings*, *one parent - one child*, *trio: two parents - one child*,

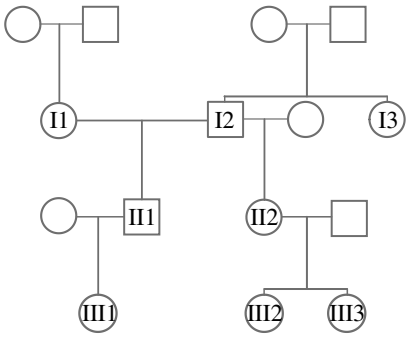

Fig. 1. Family structure of individuals that were sequenced in-house including three generations with different degrees of relationship (first-order and second-order). Circles (female samples) and rectangles (male samples) with labels indicate that genotype data was available.

quartet: two parents - two children and *second-order*). in-house families can be grouped in *trios*, *quartets* and families ranging over three generations (see Figure 1), including second-order relationships. Additionally we simulated technical replicates of one individual of the 1KG project (*NA06986*) by randomly reducing the coverage of the original alignment yielding a mean per-base coverage of $313, 140, 120, 80, 50$ and 30 reads.

Likelihood ratio analysis

Likelihood analysis is used to evaluate the goodness of fit of a hypothesis H_i relative to another hypothesis $H_j, i \neq j$, (e.g. the *null*-hypothesis H_0), depending on the same underlying data D. The likelihood ratio can be expressed as follows:

$$LR(H_i, H_j|D) = \frac{\Pr(D|H_i)}{\Pr(D|H_j)}$$

For every two individuals (x_1 and x_2) we consider five different hypotheses $H_i, i \in I = \{0, 1, 2, 3, 4\}$:

H_0: Sample x_1 and x_2 are unrelated

H_1: Sample x_1 and x_2 are technical/biological replicates

H_2: Sample x_1 and x_2 are full siblings

H_3: Sample x_1 is a parent of sample x_2

H_4: Sample x_1 and x_2 have a second-order relationship

As already introduced and extended in previous works (including Thomas *et al.* (2002), Aoki *et al.* (2001), Marshall *et al.* (1998) and Brenner (1997)) we estimate the probability of the combination of two genotypes of two individuals with the additional use of population data (Table 1, see supplemental material for a detailed derivation of the formulas). In this study we calculated allele frequencies per position from 2535 unrelated individuals which were recently updated by the KG project ((Altshuler *et al.*, 2010)). The likelihood ratios for the combinations of genotypes ($gt \in \{a_n, a_m\}, a_{n,m} \in \{A, C, G, T\}$) for all variable positions, $k \in K$, can be combined for unlinked marker loci. Due to the small probabilities it is computationally more efficient to work with logarithms of the likelihood ratios:

$$\frac{1}{K} log_{10} LR(H_i, H_j|D) = \frac{1}{K} log_{10} \prod_{k \in K} LR_k(H_i, H_j|gt_{x_1}(k), gt_{x_2}(k))$$

$$= \frac{1}{K} \sum_{k \in K} log_{10} \frac{Pr(gt_{x_1}(k), gt_{x_2}(k)|H_i)}{Pr(gt_{x_1}(k), gt_{x_2}(k)|H_j)}$$

For some relatedness classes, the probability of having one genotype in one sample and another genotype in a second sample would be zero, assuming perfect data quality and no *de novo* mutations. With these prerequisites it is, for instance, not possible that a parent has genotype a_1a_1 while the child has genotype a_2a_2 at the same position. This could yield infinitely large likelihood ratios. In this case we use an additional parameter $e = 0.001$, accounting for sequencing errors and *de novo* mutations.

Table 1. Likelihood ratios for all subject-query genotype combinations gt for different hypotheses. The frequencies, f_n refer to the allele frequency of allele a_n and are pre-calculated using data from the 1KG project and we assume $\sum_n f_n = 1$. Combinations of genotypes that do not occur for certain relationships ('Mendelian error') could still be observed due to e.g. erroneous genotyping ($e = 0.001$).

gt_{x_1} gt_{x_2}	$H_1 \sim H_0$	$H_2 \sim H_0$	$H_3 \sim H_0$	$H_4 \sim H_0$	$H_3 \sim H_2$
$a_1a_1\,a_1a_1$	$\frac{1}{f_1^2}$	$\frac{(f_1+1)^2}{4f_1^2}$	$\frac{1}{f_1}$	$\frac{f_1+1}{2f_1}$	$\frac{4f_1}{(f_1+1)^2}$
$a_1a_1\,a_1a_2$	$\frac{e}{2f_1^3f_2}$	$\frac{f_1+1}{4f_1}$	$\frac{1}{2f_1}$	$\frac{2f_1+1}{4f_1}$	$\frac{2}{f_1+1}$
$a_1a_1\,a_2a_2$	$f_1^2f_2^2$	$\frac{1}{4}$	$\frac{e}{f_1^2f_2^2}$	$\frac{1}{2}$	$\frac{4e}{f_1^2f_2^2}$
$a_1a_2\,a_1a_2$	$\frac{1}{2f_1f_2}$	$\frac{1+f_1f_2}{4f_1f_2}$	$\frac{1}{4f_1f_2}$	$\frac{4f_1f_2+f_1+f_2}{8f_1f_2}$	$\frac{1}{1+f_1f_2}$
$a_1a_1\,a_2a_3$	$\frac{e}{4f_1^2f_2f_3}$	$\frac{1}{4}$	$\frac{e}{4f_1^2f_2f_3}$	$\frac{1}{2}$	$\frac{e}{f_1^2f_2f_3}$
$a_1a_2\,a_1a_3$	$\frac{e}{8f_1^2f_2f_3}$	$\frac{2f_1+1}{8f_1}$	$\frac{1}{8f_1}$	$\frac{4f_1+1}{8f_1}$	$\frac{1}{1+2f_1}$
$a_1a_2\,a_3a_4$	$\frac{e}{8f_1f_2f_3f_4}$	$\frac{1}{4}$	$\frac{e}{8f_1f_2f_3f_4}$	$\frac{1}{2}$	$\frac{e}{2f_1f_2f_3f_4}$

Infering relatedness classes from log likelihood ratios

For each dyad we computed the LRs for all variable positions and hypotheses H_i versus the null hypothesis (not related). As the number of variable positions might differ between dyads we used the mean LR per position for a better comparison (Fig. 2). The violin plots visualize the distributions of the LRs for different hypothesis comparisons for unrelated dyads. LR tests for most of the related individuals (circles) yield positive values and the correct hypothesis maximizes the likelihood ratio, $max\{LR(H_i, H_0)\}_{i \geq 0}$ For large sample sizes the likelihood ratio test statistics approximate a χ^2 distribution and may also be used for a probabilistic interpretation (Wilks's theorem).

In families of more than two members and even highly complex sub-structures, such as shown in Fig. 1, the whole pedigree is reconstructed from the predicted relationships of dyads.

All relationship hypothesis are not directed. This means, when a the parent-child relationship is detected in a dyad we don't know who is the father and who is the son. However, the directionality can be clarified by the relationsships of additional family members. The sex of each individual is determined by the ratio of heterozygous variants versus all variants on the X chromosome and is also used in the pedigree reconstruction.

4 RESULTS

Separation efficiency of log likelihood ratios

We tested different hypotheses on related individuals from the 1KG data (Fig. 2) and used these data to infer correct relatedness states

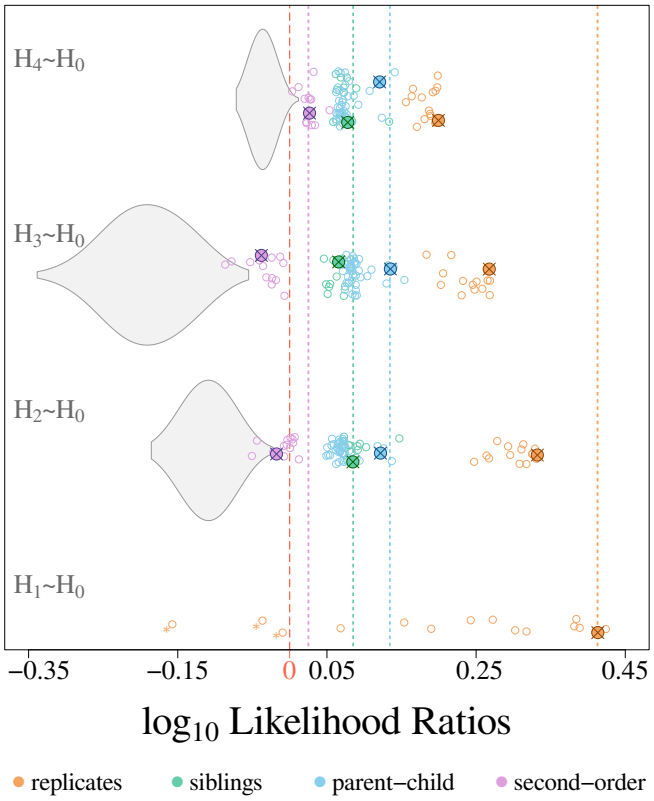

Fig. 2. Comparison of the mean likelihood ratio LR per position for different relationships between pairs of individuals (dyads). The likelihood for all relationship models H_i versus the null hypothesis (not related) were computed for all dyads from the 1000 genomes project and additional family data. LRs of all dyads that are not related are depicted as background distribution (gray violin plots) whereas dyads with a relationship are illustrated as circles. The true relationship is color-coded and maximazes the LR for the correct hypothesis. Exemplarily this is shown for a dyad of each relationship (crosses). However, misclassifications can occur especially in replicate identification for low quality samples with a high genotyping error rate (orange asterix).

to families with different substructures (Fig. 2). The comparison of likelihood ratios of most models show a segmentation of the dyads corresponding to their kinship coefficients. For instance, in Fig. 2 we see for the comparison H_3, H_0 technical replicates (H_1) with a kinship coefficient close to 1 at the right end, dyads of siblings (H_2) or parent-child (H_3) in the middle and second-order relationships (H_4) at the left side. Dyads that are not related are visualized as a gray background distribution (violin plots). An interesting exception is $LR(H_3, H_2)$, where two models with the same expected kinship coefficient (0.5) are compared. This likelihood ratio is well suited for discriminating between parent-child and sibling relationships.

Since the probability that a parent and a child share the same allele is the same as in siblings (50%), it is hard to distinguish between these two relationship states and lead to a low variance in likelihood ratios, as seen in the likelihood ratios for $LR(H_3, H_2)$ in Fig. 2. A similar kinship coefficient is also the reason why the LR values for parent-child and full siblings are nearly on the same level for the comparisons $LR(H_3, H_0)$ and $LR(H_2, H_0)$. This leads to the overall rule: the discriminatory power to distinguish between different hypotheses gets lower if the underlying assumed relationship types become more similar.

Especially with increasing error rates the identification of identical twins/technical replicates becomes more difficult as the expected kinship coefficient of 1 drops. Although all simulated replicates form a distinct cluster in the comparison of $LR(H_1, H_0)$ in Fig. 2, there are three dyads ($*$) that yield a $LR < 0$. These false classifications can be avoided if the error rate e is adjusted appropriately.

Precision depends on the number of markers and heterozygosity

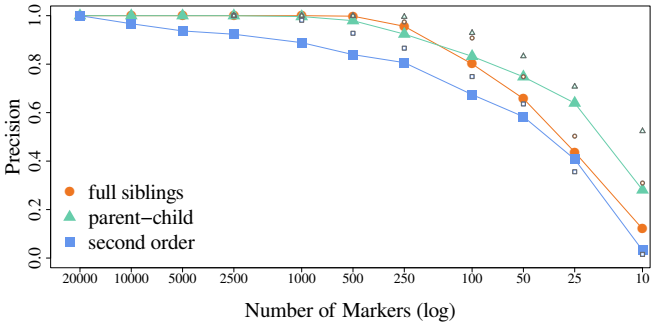

Fig. 3. The precision of the prediction of all relationships decreases with the number of available markers and their heterozygosity. The number of markers was either randomly reduced or restricted to a subset of highly informative markers. The positive predictive values for parent-child as well as for full sibling relationships start to drop when using less than $\sim$ 500 markers. At a comparable number of markers, second-order relationships are more difficult to distinguish from unrelated controls. The mean heterozygosity of biallelic variants in an exome is 0.3 and represents the information content of a marker. A higher precision can be achieved when choosing markers with a heterozygosity above this average (unfilled shapes).

We analyzed the effect of the number of variants on the precision of the classification process by reducing the amount of markers.

Besides restricting to smaller randomly chosen subsets we also studied the performance for subsets of loci with high heterozygosity. Such subsets are more similar to the markers used in SNP-arrays and allow a comparison with existing tools. Heterozygosity is defined as $h = 1 - \sum_n f_n^2$ and relates to the information content of a marker. The average heterozygosity of a SNV in our exome data was 0.3. In contrast most existing tools use polymorphic sites with considerably larger h. The positive predictive value (precision) of parent-child and full sibling comparisons starts to drop when using less than 500 randomly selected markers (Fig. 3). The correct assignment of second-order relationships requires substantially more markers for a comparable precision. However, the performance for the second-order classifications is good for whole exome sequencing data comprising several thousands of variants. Only a reduction to fewer loci, as encountered in gene panels, will lead to an excess of false positive predictions and therefore to lower precision values. Especially the number of unrelated dyads that are erroneously classified as second-order increases.

We hypothesized that increasing the heterozygosity of these small marker sets might improve the precision. When choosing only loci with a heterozygosity above average, the precision for all relationship models increased (unfilled shapes in Fig. 3). This is in good agreement with the results from other studies that achieved a high precision for predictions that were based on markers with high information content (Epstein *et al.*, 2000). The most polymorphic site accessible via exome sequencing is the human leukocyte antigen (HLA) cluster on chromosome 6 with more than 7300 different alleles known up to date (Robinson, et al., 2016). Recently, bioinformatics tools became available that allow HLA typing from exome data and can be used additionally to rule out relatedness in questionable dyads (Szolek, et al., 2014). However, a general comparison between the predictive power of multiple iallelic markers such as most SNVs and a single polymorphic marker such as the HLA shows that the discriminatory power is limited: Assuming equal likelihoods for all alleles, the information content, $IC = -\sum_n f_n log_{10} f_n$, of 13 unlinked biallelic markers $(n = 2)$ would be comparable to the IC of the HLA locus $(n = 7300)$.

The contributions of the different genotype combinations to the log-likelihood ratios also depend on the heterozygosity. For instance in $LR(H_3, H_0)$, for common variants the major contribution will come from the genotype combinations $a_1a_2 \sim a_1a_2$ and $a_1a_1 \sim a_2a_2$. If H_0 is the true hypothesis the later, $a_1a_1 \sim a_2a_2$, will push $LR(H_3, H_0)$ below zero (Fig. S1d). However, for smaller f_1, the choice of e will influence this result (Fig. S1a). On the other hand, if H_3 is true then the ratio $1/2f_1$ (Table 1 for $a_1a_1 \sim a_1a_2$) will shift $LR(H_3, H_0)$ to positive values. The contribution of $a_1a_1 \sim a_1a_2$ to $LR(H_3, H_0)$ is positive for frequencies $f_1 < 0.5$ and thus in favor of H_3. This combination will more likely occur for small values of f_1 and an disproportionate high occurence of this term will indicate relatedness.

Interestingly, we found most misclassified second-order relationships within the same sub-populations, after randomly reducing the number of variants. That also emphasizes the importance of using rare low-frequency variants, that are specific within families rather than using polymorphisms that are more suitable to discriminate between sub-populations.

Directionality of parent-child relationships

The directionality of a parent-child connection can be solved by considering additional relationships if more family members are available ((Thomas and Hill, 2000)). Fig. 4a) exemplifies a case where both, mother and father, are present in the available family tree (referred to as *trio*). The clear assignment of *mother - child* and *father - child* can be resolved by the additional information that both parents are not related. In another example, (Fig. 4b)), the directionality between parent and child can be resolved by the additional knowledge, that two of the three available samples in a pedigree (*II1* and *II2*) have a sibling relationship. A more detailed description for the pedigree reconstruction, that is based on this rule set, can be found in the supplemental material and is also encoded in our tool.

From a theoretical point of view, we wondered whether it is possible to infer the directionality of a parent-child relationship, if only a dyad is given (Fig. 4 c)). There is one scenario in which the directionality can be resolved by data of the dyad itself. If both parents have a different ethnic background the offspring will show two sets of population specific SNPs (Fig. 4 d). Furthermore the heterozygosity of the offspring will be higher than of each of the parents, which is illustrated in Fig. 4 e), where we simulated an offspring of two 1KG individuals from distinguishable populations (*CHS* and *YRI*). The frequency for many polymorphisms differs depending on the population background. In an offspring from two different ethnic backgrounds there will be a movement towards the joint mean, similar to an increasing entropy in thermodynamics when two isolated systems are mixed (Fig. 4 f).

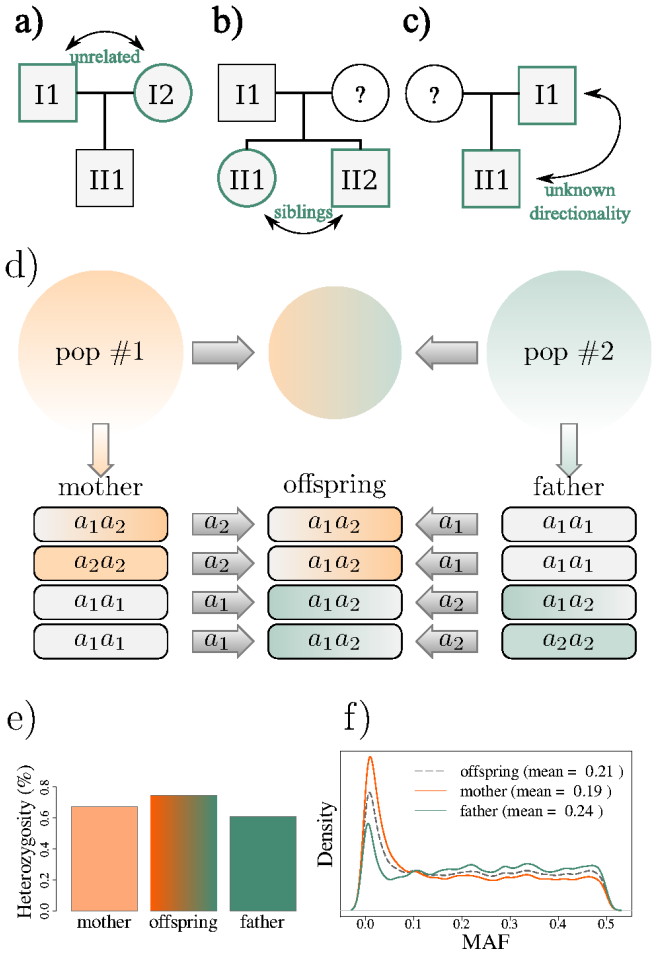

Fig. 4. The parent-child directionality can be resolved with at least three available samples, either due to the knowledge that both parents are unrelated (a) or with additional information about siblings (b). If just two sequenced samples are available in the analysis, the directionality can not be resolved (c). Panel d) shows an illustrated example of two different populations (#1 and #2) with two exemplarily chosen individuals (*mother* and *father*). The shading colors (orange and green) are illustrating the population specific variants, that do not appear in the other group respectively. Ideally, an offspring of two members of these distinguishable groups would share half of the population specific variants of the mother and half of the father. A simulated offspring of two individuals with different ethnic backgrounds, YRI (orange) and CHS (green), has an increased proportion of heterozygous variants (e) and the mean minor allele frequency (*MAF*) is located between the MAF of the parents (f).

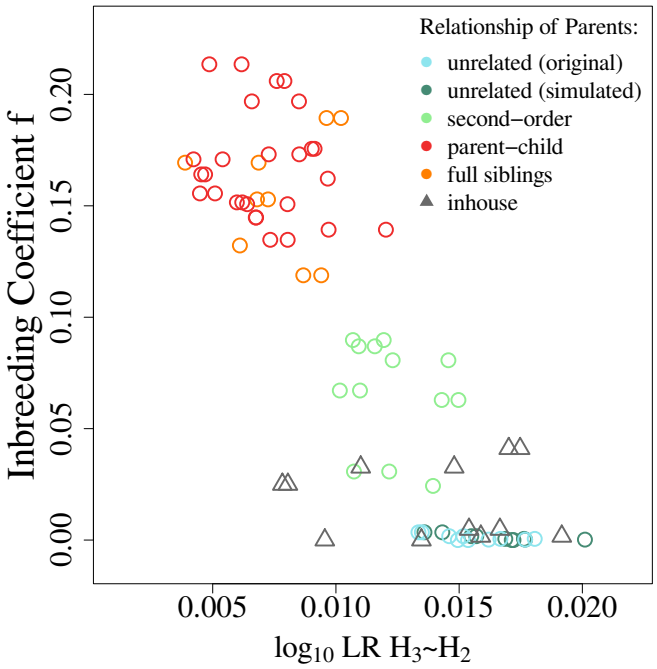

Fig. 5. Classification of relationships in highly consanguineous families. In addition to real offspring from related parents, we also simulated offspring for 1KG dyads with different degrees of relatedness. Individual inbreeding coefficients were calculated for all offspring and log likelihood ratios for all parent-child pairs. For offspring with a higher inbreeding coefficient log likelihood ratios of parent-child versus siblings approach decrease. However, for exome data a correct classification of all parent-child pairs is still possible.

Influence of inbred structures within complex families

In highly consanguineous families the prediction of the exact relationships becomes more difficult as there is a stronger deviation from the null model that is based on the allele frequencies of an outbred population.

In addition to exomes from highly consanguineous families, we investigated the influence of inbred structures by simulating offspring from related 1KG dyads. For each related pair, we randomly chose one allele from each of both assumed parents to create a new genotype at each autosomal position. We used the individual inbreeding coefficients, f, for each offspring, to quantify the extent of consanguinity.

The inbreeding coefficcient is defined as the ratio of the genome that is IBD (Wright, 1922) and was computed as described by (Gazal *et al.*, 2014). As expected, the simulated offspring with closely related parents show higher individual inbreeding coefficients. These higher inbreeding factors correlate with lower likelihood ratios between relationship models of the same order, such as parent-child versus siblings (Fig. 5).

When using entire exome data sets a correct classification was still possible for all tested cases, even for families with a high degree of consanguinity.

5 DISCUSSION

In this work we analyzed the performance of reconstructing family structures with genotype data that become available in NGS sequencing studies, including rare variants. We were able to derive pedigrees with high precision for publicly available samples of families, as well as for in-house data, even for as little as a few hundred markers. Our method is based on comparisons between likelihood ratios for different kinship classes and - as expected - the closer the coefficients of relationship were between two models, the more difficult it became to differentiate between them. The most challenging discrimination is between unrelated individuals in a highly inbred population and samples that are related by second-order relationship. So far highly polymorphic marker loci such as microsatellites have been used in paternity testing and to detect distant relationships. However, ancestry can also effectively be identified with the use of rare and family specific single nucleotide variants. We observed the biggest variance in likelihood ratios for all dyads for the comparison of the hypothesis H_3, H_0 (parent-child $\sim$ unrelated). One drawback of our approach is that different quality levels of the data are not considered. We simulated technical replicates with a high error rate by decreasing the coverage per base in the alignments and studied the influence on the classification process. Some replicates with a large difference in coverage and quality were misclassified as full siblings instead of technical replicates, as visualized in Fig. 2 (violet dots for $LR(H_1, H_2)$). With decreasing coverage, the error rates - especially for heterozygous calls - increase. However, the site-specific error probabilities that are reported by standard genotype callers are usually vastly underestimated for rare variants. Allegedly small error rates pose a problem especially in the identification of replicates. Another approach to estimate genotyping errors on exome variant lists, that also considers allele frequencies from population studies, reported a mean genotyping error rate of e=0.001 as more accurate for most current exomes (Heinrich *et al.*, 2013).This error rate also yielded the best performance in our tests and is therefore suggested as a default setting. The consequences of an unsuitable choice of this parameter on the likelihood ratios is discussed in more detail in the supporting information.

In summary we have shown that genotype data from NGS studies can also be used to deduce pedigree information with high precision. Our approach doesnt require any additional generation of data and will thus be easy to integrate into existing analysis pipelines. This will help to identify sample mix-ups at an early stage and improve the overall quality in the diagnostic procedure.

ACKNOWLEDGEMENT

Funding: This work was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG KR 3985/1-1) to P.M.K.

REFERENCES

Altshuler, D., Lander, E., and Ambrogio, L. (2010). A map of human genome variation from population scale sequencing. *Nature*, **476**(7319), 1061–1073.

Aoki, Y., Nakayama, Y., Saigusa, K., Nata, M., and Hashiyada, M. (2001). Comparison of the likelihood ratio and identity-by-state scoring methods for analyzing sib-pair test cases: A study using computer simulation. *Tohoku J. Exp. Med.*, **194**, 241–250.

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

Brenner, C. H. (1997). Symbolic kinship program. *Genetics*, **145**, 535–542.

Dakin, E. E. and Avise, J. C. (2004). Microsatellite null alleles in parentage analysis. *Heredity*, **93**(5), 504–509.

Danecek, P., Auton, A., Abecasis, G., Albers, C. a., Banks, E., DePristo, M. a., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, **27**(15), 2156–2158.

DePristo, M. a., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, a. a., del Angel, G., Rivas, M. a., Hanna, M., McKenna, a., Fennell, T. J., Kernytsky, a. M., Sivachenko, a. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next- generation DNA sequencing data. *Nat Genet*, **43**(5), 491–498.

Epstein, M. P., Duren, W. L., and Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. *American journal of human genetics*, **67**(5), 1219–1231.

Gazal, S., Sahbatou3, M., Babron, M.-C., Génin, E., and Leutenegger, A.-L. (2014). FSuite: exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics*, **23**(10), 1289–1291.

Hardy, O. J., Hardy, O. J., Charbonnel, N., and Charbonnel, N. (2003). Microsatellite Allele Sizes: A Simple Test to Assess Their Significance on Genetic Differentiation. *Computer*, **1482**(April), 1467–1482.

He, D., Wang, Z., Han, B., Parida, L., and Eskin, E. (2013). IPED: Inheritance Path-based Pedigree Reconstruction Algorithm Using Genotype Data. *Journal of Computational Biology*, **20**(10), 780–791.

Heinrich, V., Kamphans, T., Stange, J., Parkhomchuk, D., Hecht, J., Dickhaus, T., Robinson, P. N., and Krawitz, P. M. (2013). Estimating exome genotyping accuracy by comparing to data from large scale sequencing projects. *Genome Medicine*, **5(7)**, 69.

Kamphans, T., Sabri, P., Zhu, N., Heinrich, V., Mundlos, S., Robinson, P. N., Parkhomchuk, D., and Krawitz, P. M. (2013). Filtering for Compound Heterozygous Sequence Variants in Non-Consanguineous Pedigrees. *PLoS ONE*, **8**(8), 1–6.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v1 [q-bio.GN]*.

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

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and Depristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. pages 1297–1303.

Pemberton, J. M. (2008). Wild pedigrees: the way forward. *Proceedings. Biological sciences / The Royal Society*, **275**(1635), 613–621.

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature reviews. Genetics*, **6**(11), 847–859.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **135**(V), 0–9.

Thomas, S. C. and Hill, W. G. (2000). Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, **155**(4), 1961–1972.

Thomas, S. C., Coltman, D. W., and Pemberton, J. M. (2002). The use of marker-based relationship information to estimatethe heritability of body weight in a natural population: a cautionary tale. *Journal of Evolutionary Biology*, **15**, 92–99.

Veltman, J. a. and Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics*, **13**(8), 565–575.

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, **56**, 330–338.