FamLink2 - A comprehensive tool for likelihood computations in pedigrees analyses involving linked DNA markers accounting for genotype

Citation for the original published paper (version of record):
Kling, D., Mostad, P., Tillmar, A. (2025). FamLink2 - A comprehensive tool for likelihood computations in pedigrees analyses involving
linked DNA markers accounting for genotype uncertainties. Forensic Science International: Genetics, 74. http://dx.doi.org/10.1016/j.fsigen.2024.103150

(article starts on next page)

Forensic Science International: Genetics 74 (2025) 103150

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

FamLink2 – A comprehensive tool for likelihood computations in pedigrees analyses involving linked DNA markers accounting for genotype uncertainties

Daniel Kling [a,b,c,*], Petter Mostad [d], Andreas Tillmar [b,e]

[a] Department of Forensic Sciences, Oslo University Hospital, Pb. 4950 Nydalen, Oslo NO-0424, Norway
[b] Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden
[c] Biostatistics (BIAS), Norwegian University of Life Sciences, Aas, Norway
[d] Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Göteborg, Sweden
[e] Department of Biomedical and Clinical Sciences, Faculty of Health Sciences, Linköping University, Linköping, Sweden

ARTICLE INFO

Keywords:
Genetic linkage
Kinship
Genotype likelihoods
Low coverage
Sequencing

ABSTRACT

There is an increasing demand for software that can handle an arbitrary number of linked markers in forensic genetics; primarily with application to inference of relationships and direct matching but also in applications such as ancestry inference and mixture interpretation. With the emergence of sequencing technologies, denser sets of SNP markers are generated and analyzed. Additionally, sequence data of low quality and quantity DNA generate uncertainty about the underlying true genotype. We provide an efficient implementation of a general model for pedigree likelihood computations with genetic marker data using a three-layered approach. The top and first layer is the population model where allele frequencies and population substructure are accounted for. The second layer is the inheritance model which efficiently handles linked markers using an IBD model. The third and bottom layer is the observational level where we model the likelihood of the true genotype given underlying reads as well as parameters for errors. We exemplify the utility of our implementation as well as provide validation according to guidelines established by the ISFG using a combination of two published SNP panels. We demonstrate that computations are feasible for panels encompassing 10,000 markers and we argue that, due to the properties of the underlying algorithm, extending the number of markers will result in a linear increase in computation time. In addition we study the impact of parameters used in our model and suggest some guidelines pertaining to their values. The results demonstrate that a probabilistic model for low coverage sequence read data is needed instead of relying on an a threshold based genotype and applying our general model for inference of relationships on a real case can be superior, i.e. higher information content, to other methods relying on either fixed genotypes with low quality sequence data or simple pair wise relationship tests. In summary, the implementation, FamLink2 (freely available at https://famlink.se), can jointly handle genetic linkage, genotype uncertainty and population substructure for an arbitrary pedigree with data for any number of individuals. Whereas the current study will focus on calculations disregarding mutations, FamLink2 has the ability to model mutations for certain built-in pedigrees.

1. Introduction

Progress in forensic genetics has made several expanded panels of SNP markers available to the community, primarily aimed at massively parallel sequencing platforms [1–4]. Tillmar et al. recently published a joint effort to unify the panel of genetic markers used in forensic genetics [1] without focusing on a particular library preparation protocol. The FORCE panel encompasses a total of 5422 markers with 3931 autosomal SNPs particularly suitable for kinship applications. The markers have been carefully selected to a) avoid linkage disequilibrium between alleles, b) minimize population dependency of allele frequencies, i.e. low diversity across continents and c) maximize kinship information content which for bi-allelic SNP markers is achieved when the minor allele frequency approaches 0.5. Secondly, the commercial ForenSeq

* Corresponding author at: Department of Forensic Sciences, Oslo University Hospital, Pb. 4950 Nydalen, Oslo NO-0424, Norway.
E-mail address: daniel.kling@rmv.se (D. Kling).

https://doi.org/10.1016/j.fsigen.2024.103150
Received 21 March 2024; Received in revised form 16 August 2024; Accepted 20 September 2024
Available online 24 September 2024

Table 1
Illustrative example of the binary/threshold based genotype calling approach and the genotype likelihood approach. Data is given for a single SNP marker with alleles A/C in the population. Note the notation used in the table is simplified for this example. Pr(Data|G) corresponds to the conditional probability of the read data given a particular genotype, G.

Example	#Reads with A	#Reads with C	Binary/Threshold based*	Genotype likelihood**
1	12	11	AC	Pr(Data|AA)=0.01 Pr(Data|CC)=0.008 Pr(Data|AC)=1
2	11	0	No call	Pr(Data|AA)=1 Pr(Data|CC)=0.001 Pr(Data|AC)=0.009
3	22	0	AA	Pr(Data|AA)=1 Pr(Data|CC)=0.0001 Pr(Data|AC)=0.0009
4	4	3	No call	Pr(Data|AA)=0.06 Pr(Data|CC)=0.04 Pr(Data|AC)=1

* Analytical threshold (AT)=10 and stochastic threshold (ST)=20.
** Note, the likelihood is scaled such that the most probable genotype has a likelihood of 1.

Kintelligence panel [3], encompasses 10,230 SNPs, with its main application in inference of relationships, it aims to supplement dense SNP microarray and whole genome sequencing data in genealogy contexts. Moreover, Gorden et al. [4] published two different panels including roughly 25,000 and 95,000 SNP markers with the aim to resolve distant kinship from samples of low quality.

In contrast to traditional forensic DNA data from capillary electrophoresis (CE) analysis, sequence data generates individual reads which contain meta information such as sequence quality (per base) and mapping quality (per read). Both these parameters feed into the genotype calling procedure. Ultimately, genotypes can be called based on the number of reads covering a region of interest, which is comparable to relative fluorescent units (RFU) in CE applications, but also taking the afore-mentioned quality measures into account. For instance, GATK [5, 6] and bcftools [7] are two different tools commonly used in genetic applications to call the most likely genotype based on sequence data. Tillmar et al. [8] proposed a binary model to call genotypes based on thresholds for the number observed reads per variant allele, the variant allele frequency distribution and sequencing quality. Such exact binary calling of genotypes is well suited for high coverage and quality sequence data where the true underlying genotype is well represented in the reads, but

when quality or coverage is low, say below 5X, may be mistaken in its calls with a small probability so that some errors in the profile are introduced, and when it does not make a call, it ignores the (sometimes considerable) information that the reads give (see Table 1 for an illustrative example).

Instead of calling genotypes based on read thresholds, a probabilistic genotyping model can be constructed to use a likelihood distribution for all possible genotypes for a given genetic marker in the biostatistical calculations. The tool *ngsRelate* from Korneliussen et al. [9] implements a model to compute measures of relatedness based on sequence read data using a maximization likelihood approach, but lacks a model for genetic linkage and in addition relies on pair wise tests between individuals. Merlin [10], on the other hand, provides efficient likelihood calculations for general pedigrees with linked markers with a simple model for genotyping errors, but does not model sequence data.

In this paper we evaluate a new model for likelihood calculations with linked markers while simultaneously modelling genotype likelihoods. Mostad et al. [11] recently published an efficient IBD model leveraging pedigree symmetries to speed up likelihood calculations. In addition, *op cit.* describes a model for using sequencing read data to model likelihood for genotypes combining the power of the aforementioned methods. The aim with the present study is to present a general and efficient implementation of the model to compute likelihoods for relationship applications and forensic match statistics. Mostad et al. [11] provided an implementation in R, we expand the implementation to a user-friendly GUI in C++ with various additional functionality, described later. We provide ways of validating the results as well as a real case example of when our model is crucial to solving a case. The model for sequence read data is described for SNP data but can be adopted for STR markers and the IBD model for inheritance makes no distinction as to what type of markers are used. Besides kinship and forensic matching applications, our observational model is useful also in other forensic areas, e.g. ancestry inference and phenotype predictions accounting for genotype uncertainty. We follow the recommendations set forth by Coble et al. for validating bio-statistical software in forensic

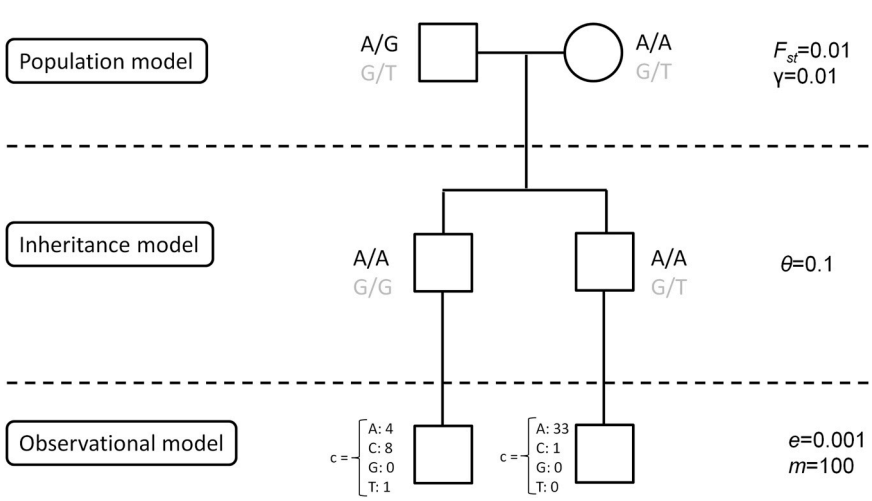

Fig. 1. Illustration of our model for likelihood computations starting at the top with our population model with parameters for subpopulation structure (F_{st}) and γ (frequency for alleles not in the population database) as inputs. In the middle our inheritance model with recombination rates (θ) as input which can be different for males or females and the bottom layer, our model for observations where e (sequencing and mapping errors) and m (PCR imbalance) are used to create genotype likelihoods. The top two layers illustrate data for two SNP markers while the bottom layer illustrates read data for a single marker.

genetics [12].

2. Material and methods

The below sections are divided as follows; First we briefly describe our model for computations, details are provided in Mostad et al. [11]. Secondly, we outline the data used to test and validate our implementation. Thirdly, we present a study to evaluate the impact of model parameters and finally we present authentic cases where previous efforts have been insufficient to resolve questions about the alleged relationship. In the description that follows, we will refer to low coverage next generation sequencing (lcNGS) data without exactly defining the criteria. We generally use this abbreviation when coverage data is <10X.

2.1. Model and implementation

Conceptually, computation of the probability of observed data given a particular pedigree can be subdivided into three parts: First, computation of the probability of all possible genotypes for the founder alleles in the pedigree, i.e., the maternal or paternal alleles not inherited from another person in the pedigree. Such computations are done according to a population model specified below. Secondly, for all possible founder genotypes, we need to compute the probability of the possible observable (un-phased) genotypes of the tested persons. We call the specification of how this is done the inheritance model. Finally, we must compute the probability of the actual observed data given all possible genotypes of the tested persons, using a specification in an observation model. Multiplying these three probabilities together and summing over all possible founding genotypes and tested genotypes yields the probability we seek.

Clearly, the sum has too many terms to be computed by direct summation, so instead we compute it by applying a version of the Lander-Green algorithm [13], see Mostad et al. [11] for details. Below we focus on describing the three layers used in computations and illustrated in Fig. 1.

The population model assumes independence of the alleles at different loci along the chromosome, i.e. linkage equilibrium. For each locus, the vector of counts of nucleotides A, C, G, or T appearing among the founder alleles is modelled with a Dirichlet-Multinomial distribution, see Mostad et al. for details. Briefly, we use a population fixation (kinship) parameter F_{st} together with a parameter γ specifying the probability, for any allele, that it is randomly selected from A, C, G, or T instead of selected according to given population frequencies (f_A, f_C, f_G, f_T). Specifically, we use a Dirichlet distribution where the i'th parameter ($i = A, C, G, T$) is $(\frac{1}{F_{st}} - 1)(f_i(1-\gamma) + \frac{\gamma}{4})$. The Dirichlet-Multinomial corresponds to a Polya urn model where the initial total number of balls is $(\frac{1}{F_{st}} - 1)$. Marginalizing to a model for count k of alleles of type i and count $s-k$ of other types, we get a Beta-Binomial model with parameters α and β where $\alpha = (\frac{1}{F_{st}} - 1)(f_i(1-\gamma) + \frac{\gamma}{4})$ and $\alpha + \beta = (\frac{1}{F_{st}} - 1)$. In this model, given s and k, the probability of drawing another allele of type i is $(k + \alpha)/(s + \alpha + \beta)$, which works out to $\frac{k + (\frac{1}{F_{st}} - 1)(f_i(1-\gamma) + \frac{\gamma}{4})}{s + \frac{1}{F_{st}} - 1} = \frac{F_{st}k + (1-F_{st})(f_i(1-\gamma) + \frac{\gamma}{4})}{1 + F_{st}(s-1)}$. From the perspective of pseudo-counts, we see that, from the total size of $\frac{1}{F_{st}} - 1$ of the database, $(\frac{1}{F_{st}} - 1)(1-\gamma)$ consists of frequency counts, while $\left(\frac{1}{F_{st}} - 1\right)\gamma$ consists of pseudo-counts, with a quarter of this of each type.

The inheritance model specifies probabilities for the possible unphased genotypes of tested persons given the genotypes of pedigree founder alleles. The computation uses Mendelian inheritance patterns at each locus and assumes there are no mutations. Crucially, inheritance patterns are correlated between adjacent loci, using crossover probabilities computed from given genetic map data, in our example given as a sex-average. Details on the inheritance model are provided in Mostad et al. [11].

Finally, the observational model describes, at each locus independently, a probability distribution for lcNGS data given a specific genotype (g_1, g_2) of a tested person, where each g_i is one of A, C, G, or T. We describe this model by describing how to simulate from it using several steps: First, for each of m DNA templates that end up founding PCR amplicons, where m is a model parameter, it is randomly chosen whether the template is based on g_1 or g_2. The identity of each recorded read in the lcNGS data is then determined as follows: With a probability e, where e is a model parameter, it is chosen uniformly at random from A, C, G, T. Otherwise it is equal to g_1 with a probability $\frac{k}{m}$ and g_2 with probability $\frac{m-k}{m}$, where k is the number of DNA templates based on g_1.

Note how the parameter e is broadly related to the drop-in rate, so that, if $e = 0$, no drop-ins will originate as noise in the lcNGS data. However, if $\gamma > 0$, data could still contain alleles not observed in the frequency database used, but such data could then indicate a true, but until now unobserved, genotype. Similarly, m is related to the drop-out rate. When m is large (say $m = 1000$), $\frac{k}{m} \approx \frac{1}{2}$, and a drop-out can only occur, in our observation model, when the total number of reads is so low that they all can happen to be based on the same of the two alleles g_1 or g_2. When m is smaller (say $m = 10$), $\frac{k}{m}$ can be further away from $\frac{1}{2}$, increasing the imbalance in the sampling and thus the chance of drop-outs. If m is quite small (say $m = 3$), there is a considerable chance that $\frac{k}{m} = 0$ or $\frac{k}{m} = 1$, in which cases there will be a drop-out (for a heterozygote genotype) no matter how many reads are recorded for the single allele. Finally, we note that our model for genotype likelihoods is defined using two parameters (e and m) whereas there in reality are other parameters that could be relevant.

The algorithm alluded to above and fully described in Mostad et al. [11], is implemented in freely available R code (https://familias.name/lcNGS/) which can be used for research purposes. We present an efficient C++ implementation in the software FamLink2 with several tools for forensic purposes. The cores of the R code and C++ calculations are almost identical. However, since the R code is intended for research purposes and as a proof of the original model description it will be updated less frequently. The current study will focus on the implementation in FamLink2.

2.2. Data

2.2.1. Genetic markers

We used marker data from the FORCE panel [1] using 3931 autosomal SNP markers developed for kinship applications and from the KIntelligence panel consisting of a selection of 9618 autosomal SNP markers [3]. The latter was only used for performance testing while the FORCE panel was used for several other purposes, described later. Genetic positions were extracted from Rutger's repository [14] or alternatively interpolated for markers with missing data. We used an allele frequency database based on individuals with European ancestry (CEU, GBR, TSI and IBS) in the 1000 G project [15], which is henceforth referred to as NFE (Non Finnish Europeans).

2.2.2. Import of data

We first tested rudimentary functionality of FamLink2. To test that the software correctly imports genetic map and population frequency data we use data sets available through https://famlink.se/f_validation.html for the FORCE panel. We first import data into FamLink2 and subsequently export the data to files again. An R-script tests that identical input/output is obtained. Similarly, to test that genotypes are imported correctly, genotype data from the fictive samples (also available through https://famlink.se/f_validation.html) is imported into a case and subsequently exported back to file. We confirm that identical input/output is reported. The procedure is tested for the following input formats; VCF, CLC, Familias-like, Genemapper, DTC-like, vertical data and horizontal data. The CLC and VCF file formats are currently the only

Table 2
Overview of the experiments performed on the impact of the parameters in the observation model. D=Depth, NFE=Non-Finnish European allele freqencies.

Experiment	Genetic markers	Population data	Inheritance data	Sampled read data	Parameters used in calculations	#Simulations
Observational model	Single marker	None	None	*D*=1X–30X; *e*=[0.1, 0.01, 0.001,0]; *m*=[2,3,4,5,10, 100]	Same as for sampling	10000
Pilot study	FORCE	NFE	First cousins	D=1X,5X,10X; *e*=[0.1, 0.05, 0.01]; *m*= [5,10,20]	*e*=[0.001, 0.01, 0.05, 0.1, 0.2] and *m*=[5,10,25,100]	1000
Study of the parameter *e*	FORCE	NFE	First cousins	D=1X–5X,10X,20X,30X; *e*=[0.001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2] and *m*=10	Same as for sampling	100
Study of the parameter *m*	FORCE	NFE	First cousins	D=1X–5X,10X,20X,30X; *e*=[0.001, 0.01, 0.1]; *m*=[5, 10, 15, 20, 25, 50, 100]	Same as for sampling	100

Table 3
Description of parameters used to sample read data. Read data is sampled from a dicretized Gamma distribution based on the actual genotypes with expectation (Depth) and standard deviation (SD) where *e* relates to the dropin error and *m* to PCR imbalance.

Combination	Depth	SD	*e*	*m*
1	1X	0.5X	0.1	5
2	5X	1X	0.05	10
3	10X	2X	0.01	20

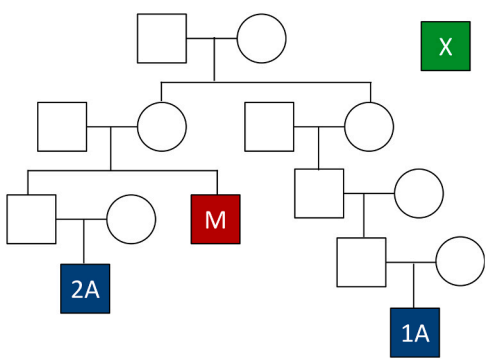

Fig. 2. Pedigree for the authentic test case. SNP data was available for human remains presumably originating from M, together with SNP data from reference individuals 1 A (1C2R to M), 2 A (nephew to M) and X (unrelated to M, 1 A and 2 A).

2.2.3. Validation cases

2.2.3.1. Study of the inheritance model. Next we generated 100 sets of genotypes for the kinship markers in the FORCE panel [1] where we simulate data for pair wise sets of individuals assuming the relationships full siblings, half siblings, first cousins, first cousins once removed as well as second cousins. Data is generated through simulations accounting for linkage while mutations and subpopulations structure is disregarded. For each simulated pair of individuals we compute the LR in FamLink2 using unrelated as the alternative hypotheses as well as in Merlin [10] expecting identical output when our model for observations is turned off. As mean of continuous validation, FamLink2 allows computations in Merlin to be performed simultaneously with our method.

2.2.3.2. Study of the observation model. To verify the output from our observation model we used real sequence data from a dilution series (10 ng, 1 ng, 0.3 ng, 0.1 ng and 0.06 ng input DNA) of a control DNA (2800 M, Promega[1]) with known golden standard genotypes. Data was generated using the methods described in Staadig et el [16]. The bioinformatic analysis was conducted in CLC Genomic Workbench (Qiagen) outputting the results as a text file with four rows per marker, each line signifying a base (A,C,G,T) and its read count. Additionally, the read output for the 1 ng sample was down-sampled to 5000 total reads. The resulting BAM-files were processed in ANGSD to generate genotype likelihoods [17]. In detail we used the call *angsd –GL 2 –doGlf 2 –i [Sample].bam –out [Sample]_likelihood –rf force_snps_hg38.bed* to generate a list of log likelihoods for each kinship marker in the FORCE panel using the GATK likelihood method (-GL 2) and outputting all 10 possible genotype likelihoods (-doGlf 2), see McKenna et al. [6]. The GATK model resembles our parameterization and should give similar output. We used the output from CLC to generate genotype likelihoods in FamLink2 with *m* ranging from low values 1–10 to high 100 as well as 20 different values on *e* ranging from 0.00001 to 0.2.

Table 4
Time consumption for likelihood calculations using the method and implementation outlined in this paper. Calculations are performed on a 12 core 2.1 Ghz CPU with 16 Gb RAM for varying degree of genetic markers (bi-allelic data) for some relationships (listed in first column). Single core means that the calculations have only utilized a single of the available cores. 1000 simulations were performed whenever *simulations* is included in the table header.

Relation	FORCE single core	FORCE multi core	FORCE simulations multi core	KIntelligence single core	KIntelligence multi core	KIntelligence simulations multi core
Full siblings	<1 sec	<1 sec	5 min 50 sec	3 sec	2 sec	13 min 10 sec
Half siblings	<1 sec	<1 sec	7 min 31 sec	3 sec	2 sec	17 min 7 sec
First cousins	<1 sec	<1 sec	7 min 20 sec	4 sec	2 sec	14 min 23 sec
First cousins once removed	2 sec	1 sec	14 min 9 sec	5 sec	3 sec	26 min 32 sec
Second cousins	6 sec	2 sec	32 min 40 sec	13 sec	6 sec	1 h 11 min
Second cousins once removed	19 sec	7 sec	1 h 41 min	50 sec	17 sec	4 h 6 min

ones that allow read counts and/or genotype likelihoods to be imported. We randomly generate read counts using a discretized Gamma distribution with 100X as the expectation and 10X as standard deviation for each of the observed alleles in the genotype. For a detailed description of the input format we refer to the software or the manual. Note that import/export of data should always be internally validated for new software.

[1] https://se.promega.com/products/forensic-dna-analysis-ce/str-amplification/2800m-control-dna/?catNum=DD7101

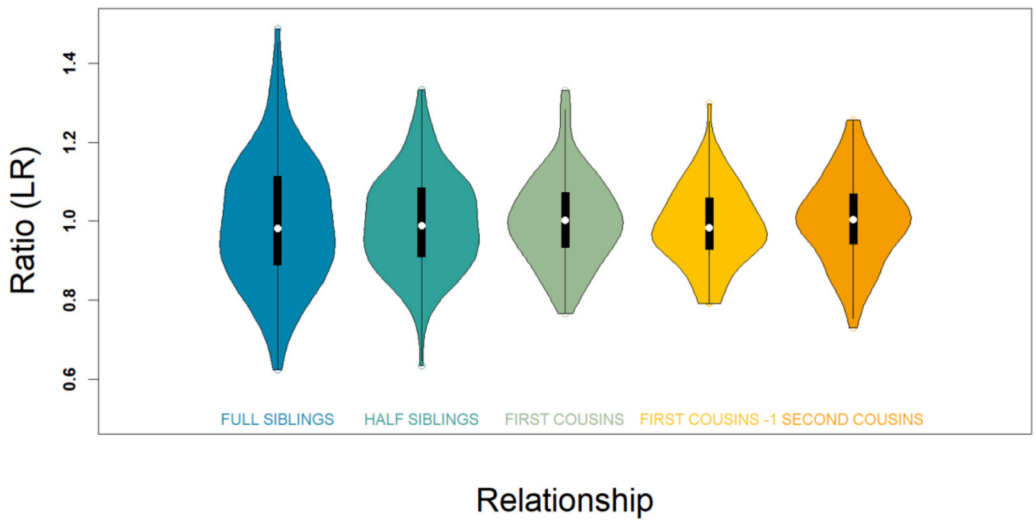

Fig. 3. Results from validation cases (N=100), where relatives have been simulated in FamLink2 and the resulting genotypes analyzed in Merlin and FamLink2 respectively. The violin plot illustrates the ratio between the two LRs obtained in each software.

Table 5
Summary of genotype callings based on the observation model described by Mostad et al. We provide the results based on the combination of *e/m* that minimizes the differences which may be different for each sample. Note, even though different callings (maximum likelihood estimates) is produced both models still provide a non-zero likelihood for all genotypes. The column "Match with reference (%)" indicate results from comparisons with golden standard reference genotypes.

Sample	Concordant callings (%)	Discordant callings (%)	Match with reference (%)	*e/m* used	Average depth (X)
10 ng	100	0	99.847	*e*=0, *m*=2	560.6
1 ng	100	0	99.822	*e*=0, *m*=2	119.9
0.3 ng	99.975	0.025	99.745	*e*=0.021, *m*=7	75.8
0.1 ng	99.439	0.561	98.751	*e*=0.032, *m*=7	45
0.06 ng	98.448	1.552	88.205	*e*=0.021, *m*=5	27.2
Downsampled	70.567	29.433	64.901	*e*=0.179, *m*=2	0.9

Table 6
log10LRs for direct matching comparisons between reference sample 2800 M genotypes and read data from a dilution series of 2800 M, and a downsampled dataset.

Samples	Coverage	log10LR					
Reference vs	Mean X	*m*=50, *e*=0.001	*m*=50, *e*=0.01	*m*=10, *e*=0.001	*m*=10, *e*=0.01	*m*=2, *e*=0.001	*m*=2, *e*=0.01
10 ng	560.6	3819	3862	3803	3852	3137	3137
1 ng	119.9	3853	3862	3819	3854	3136	3137
0.3 ng	75.8	3840	3849	3776	3845	3132	3128
0.1 ng	45	3635	3574	3657	3729	3064	3058
0.06 ng	27.2	787	567.9	2104	2162	2441	2419
Downsampled	0.9	976.7	969.9	976.7	969.2	961.5	948.4

2.2.3.3. Direct matching. Finally, we perform direct-matching where we compute the LR in FamLink2 for all pairs of combinations of the samples alluded to previously. We model genotype likelihoods in the search to accommodate the low read data and explore *m*=2, 10, 50 and *e*=0.001, 0.01.

2.3. Impact of model parameters

We performed a series of evaluation of the observation model, summarized in Table xx and described in detail below.

2.3.1. Observation model

The model described in 2.1 and detailed in Mostad et al. [11], requires specification of some parameters. In particular, we explore the impact of the parameters used in the observation model, namely *e* (a float in the range 0–1) related to sequencing and mapping errors and *m* (an integer larger than 0) related to the PCR allelic imbalance. To understand the model, we start at its core by studying the genotype likelihood matrix where each element of the symmetric matrix (4×4) describes the likelihood of the read data given each possible genotype and where we assume a SNP marker with four possible alleles. To this end we sample read data for 10,000 heterozygote genotypes and 10,000 homozygote genotypes separately. Note that the exact genotypes are irrelevant as we are studying the read data only. We sample reads using *e*=[0.1, 0.01, 0.001], *m*=[2,3,4,5,10, 100] and average depth ranging from 1X–30X. We subsequently analyze the data with the same sets of *e* and *m*. We summarize the data and explore the likelihood of the true genotype for each combination of parameters described previously.

2.3.1.1. Combining pedigree and observation data. Next, we simulate genotypes using our population model and inheritance model for 1000 pairs of first cousins as well as 1000 pairs of unrelated pairs. We use allele frequencies described in Section 2.2.1 with marker data from the FORCE panel and F_{st}=0 to simulate founder genotypes. The simulations subsequently use gene-dropping [18] and the genetic map described in Section 2.2.1 to sample non-founder alleles. Note, the simulations merely output genotypes whereas the next steps will simulate sequence read data. To mimic a realistic scenario, we sample low quality read data

Table 7
Illustration of log10 LR in the authentic case described in the main text. Different methods for calculations are described in the first column with settings used on *e* and *m*. The table includes the binary calling method in Tillmar et al. for comparison [1]. Relationship hypotheses are indicated in parenthesis encompassing second cousins once removed (2C1R), first cousins twice removed (1C2R) and Uncle/nephew.

Computation method	01 A vs 02 A (2C1R vs Unrel)	01 A vs M (1C2R vs Unrel)	02 A vs M (Uncle vs Unrel)	01 A vs X (2C1R vs Unrel)	02 A vs X (2C1R vs Unrel)	M vs X (2C1R vs Unrel)	Full pedigree
Tillmar et al. 2021	2.58	−0.04	−0.17	0.00	0.00	0.00	-
FamLink2: *m*=50; *e*=0.001	2.43	4.21	38.93	−0.58	−0.06	0.62	40.32
FamLink2: *m*=50; *e*=0.01	2.69	4.28	39.37	−0.62	−0.08	0.60	41.56
FamLink2: *m*=10; *e*=0.001	2.33	4.28	39.42	−0.52	−0.03	0.61	49.29
FamLink2: *m*=10; *e*=0.01	2.44	4.30	39.27	−0.60	−0.06	0.59	48.92
FamLink2: *m*=5; *e*=0.001	2.33	4.33	39.33	−0.56	−0.05	0.61	49.91
FamLink2: *m*=5; *e*=0.01	2.47	4.36	39.56	−0.60	−0.07	0.59	49.42
FamLink2: *m*=2; *e*=0.001	2.68	4.52	34.90	−0.60	−0.08	0.61	49.25
FamLink2: *m*=2; *e*=0.01	2.70	4.64	34.77	−0.63	−0.08	0.59	48.39

described in the legend with different settings used on e and m. The figure includes the binary calling method in Tillmar et al. for comparison [1].

for one of the samples and use high quality genotypes for the second sample. To this end, we sample sequence read data using a Gamma distribution with a combination of settings and then analyze each set of read data, i.e. study the impact on the final LR, using a combination of values for *e* and *m*. Since the number of combination with which data can be simulated and subsequently analyzed is in theory infinite we first perform a pilot study where three levels of read depth and choices of *e* and *m* are explored, see Table 3. In addition to using our model for genotype likelihoods we also call the genotypes for each simulated data set using a model described in Mostad et al. [11].

For analysis we study all combinations of *e*=[0.001, 0.01, 0.05, 0.1, 0.2] and *m*=[5,10,25,100]. In total, we sample sequence data and perform $3 \times 5 \times 4 = 60$ likelihood calculations for each set of simulated genotype data.

Next we use the results from the pilot study and refine our simulation to explore data from a selection of 100 simulated genotypes where we sample read data using depths 1–5, 10,20 and 30X and *e*=[0.001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2] and *m*=10. That is we focus on studying the impact of *e* for different depths first. We note that 100 simulations is a comparatively low number which is due to the computational effort required for each likelihood computation and the large number of comparisons being performed. For each set we analyze the data, i.e. perform likelihood computations for the same set of *e* as data is simulated from. Similar to the first tests we also call the low coverage genotypes using a model described in Mostad et al. [11]. In total 8×8 x $(1+8) = 576$ calculations are performed for each set of simulated genotypes.

In the final set of tests we explore the effect of the *m* parameter. Again we restrict our analysis to 100 simulated genotypes and simulate read data using depths 1X–5X, 10X,20X and 30X with *e*=[0.001, 0.01, 0.1] and *m*=[5, 10, 15, 20, 25, 50, 100]. For analysis we used the same set of *e* and *m* as used in the simulations and in addition call the low coverage genotypes using a model described in Mostad et al. [11]. In total, $8 \times 3 \times 7$ x $(3 \times 7 + 1) = 3696$ likelihood calculations are performed for each simulated set of genotypes.

2.4. Authentic cases

In Tillmar et al. [1] the FORCE SNP panel was developed and tested on six different case scenarios with authentic SNP data from both low and high quality samples. One of these cases, which in *op cit.* resulted in inconclusive relationship estimates, was chosen as a test case for our implementation. The case consisted of SNP data from human remains (M) originating from World War II (WWII) and associated SNP data from two reference individuals, one nephew (2 A) and one first cousin twice removed (1 A). For comparison, we also added SNP data from an individual (X) unrelated to M, 1 A and 2 A. For the comparisons with this X individual, we tested a 1C2R vs Unrelated case scenario. In Tillmar et al., the comparisons, and LR calculations, were performed with a threshold based approach for the genotype calling. Applying a 10X threshold for coverage, only 15 SNPs met this criteria resulting in LRs with an inconclusive decision when compared to the references, see Tillmar et al. for details on the thresholds used in decisions. However, roughly 4000 SNPs were covered to at least 1X, making it a suitable case to test the genotype likelihood model described in this study and to, potentially, increase the possibility for relationship matching with low coverage SNP data. We used the methods described previously and the implementation in FamLink2 to compute likelihood ratios using the FORCE kinship SNPs and European SNP allele frequencies. All calculations were performed in a pair wise fashion between M, 1 A, 2 A and X (see Fig. 2) with various settings for values for *m* (2, 5, 10, 50) and *e* (0.001 and 0.01). Finally, we computed the full pedigree LR using the same range of *m*/*e*.

3. Results

3.1. Implementation and performance

Mostad et al. [11] presented an implementation of the algorithm using open-source R scripts, available at https://familias.name/lcNGS/. In this paper we expand the implementation to FamLink2 [20] which was first released in 2012 for pairs of linked markers. The expansion of FamLink2 allows for general calculations for linked markers where subpopulation correction is accounted for in the model for population allele frequencies and genotype likelihoods based on sequence read counts. The software can use the observation model presented in Mostad et al. and alluded to in the Methods section, but also allows genotype likelihoods from external sources, e.g. from imputed data or genotype variant callers. In addition, our inheritance model is general in the sense that any pedigree is allowed with any number of typed individuals. FamLink2 further implements a variety of user-friendly functionality such as blind searches, a simplified DVI module as well as inference of biogeographic ancestry. The latest version is freely available at https://famlink.se.

We did performance testing of our implementation using a 12 logical core 2.1 Ghz CPU workstation with 16 Gb of RAM. The implementation

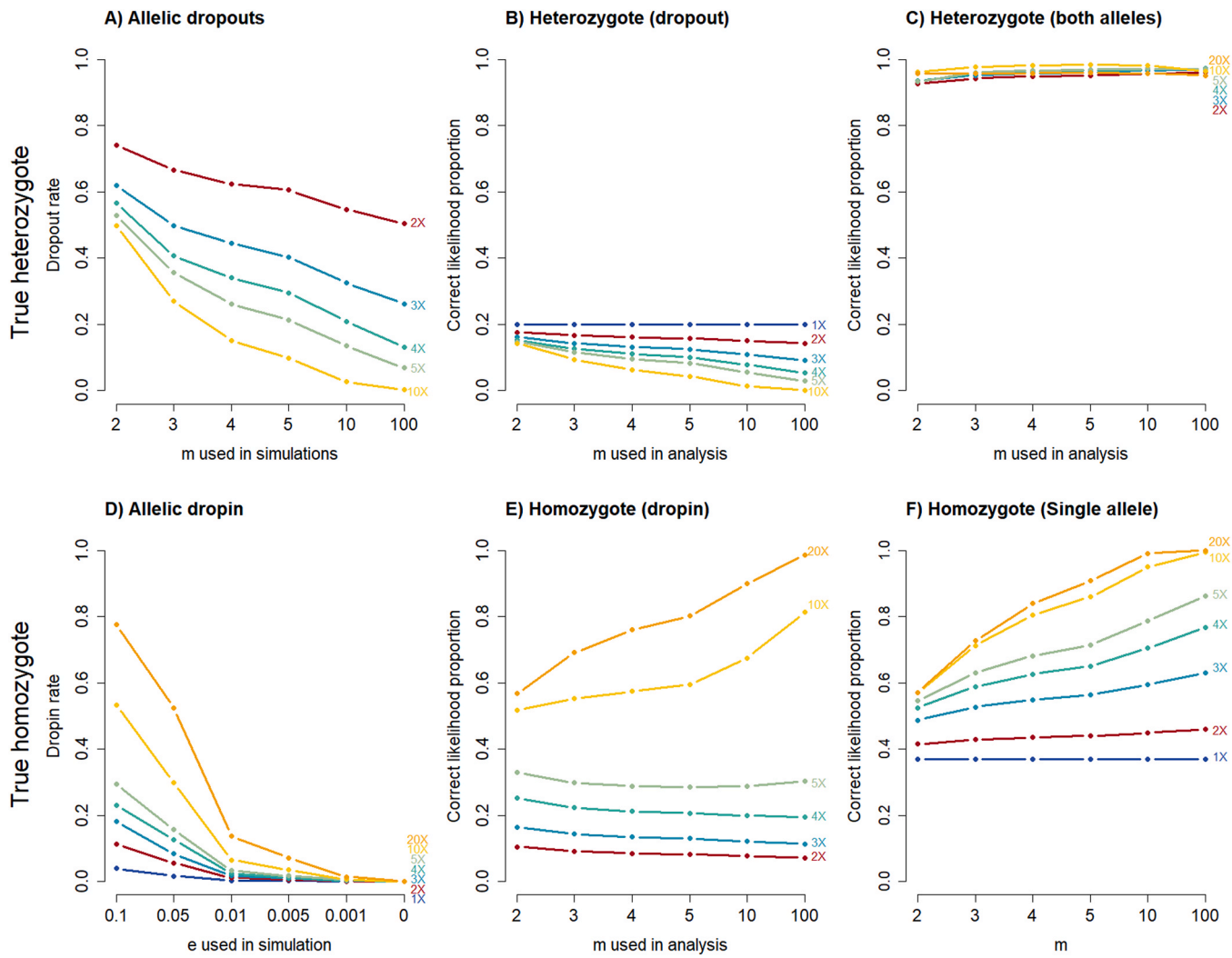

Fig. 4. Summary of the results from extensive sampling (n=10000) of read data using different combinations of depths, *e* and *m*. Details given in main text. A) Allelic dropout rate when data is sampled with *e*=0.01 at different depths (legend) and *m* (x-axis), signified by a complete allelic dropout, i.e. no reads, for one of two alleles in a true heterozygote genotype. Note that at 2X and *m*=2, dropout occur in 75 % of all samples whereas for 10X and m=100 almost no complete allelic dropouts is observed. B) Correct likelihood proportion for the true heterzygote genotype based on a subset of data where a complete allelic dropout has occurred, data is sampled with *e*=0.01 and at *m*=2 and various depths (legend). C). Correct likelihood proportion for the true heterzygote genotype based on a subset of data when both true alleles are observed (no dropin), data is sampled with *e*=0.01 and at *m*=2 and various depths (legend). D) Allelic dropin rate, signified by the occurrence of two (or more) alleles in a true homozygote genotype. Note that at 20X and *e*=0.1, dropin occur in roughly 80 % of all samples whereas for 2X very few dropins are observed B) Correct likelihood proportion for the true homozygote genotype based on a subset of data where a allelic dropin has occurred, data is sampled with *e*=0.01 and various depths (legend). F). Correct likelihood proportion for the true homozygote genotype based on a subset of data when no dropins are observed, data is sampled with *e*=0.01 and various depths (legend).

can leverage a maximum of 22 computational cores/threads, with parallelization across chromosomes. We note that the software benefits from more CPU cores (up to 22) with higher individual clock speed whereas the memory requirement is low, typically 2–4 Gb is sufficient. We used constructed data based on the markers in the FORCE panel [1] and the KIntelligence panel [3] for some standard pair wise relationships, described in Table 4. In addition, we performed repeated calculation on simulated data 1000 times for the same relationships, using our multi core setup. Details on the data and simulations are reported in Section 2.2. We note that for the pair wise relationships described in Table 4, computation times are very low, typically less than 10 seconds except for second cousins once removed, but acknowledge that extended pedigrees involving several typed individuals may require considerably longer time to finish (Data not shown). Leveraging pedigree symmetries as described in Mostad et al. [11] can greatly reduce computation speed for such extended pedigrees in future releases of FamLink2.

3.2. Validation

All files used in the validations (except files relating to the authentic case) are available following links from https://famlink.se/f_validation.html. The files can be used to test new version of FamLink2 with expected results given for each file. As means of validating the inheritance model, we simulated genotype data for a range of relationships and computed the likelihood and likelihood ratios (assuming linkage) in Merlin [10] and FamLink2. As illustrated in Fig. 3, the ratio between the two LRs is small, with a mean centered close to 1 for all relationships. The small deviations detected in some cases are possibly explained by the use of different mapping functions, i.e. to convert cM values to recombination rates, or numerical issues.

Next we performed a comparative study where the observation model was applied on a selection of real samples with varying DNA quantity/quality used in the sequence library preparation. The output (either a text file with read counts for each nucleotide at a given position

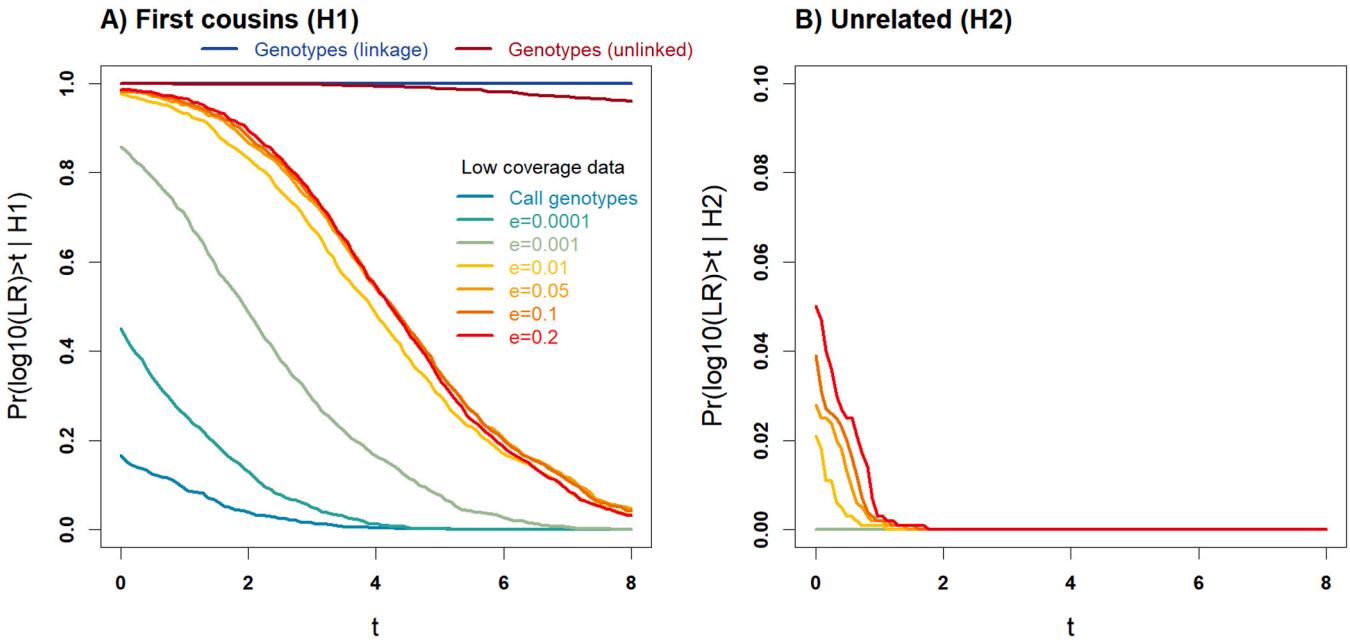

Fig. 5. Illustration of exceedance probabilities (inclusion power). Data is based on 1000 simulated genotypes for A) First cousins (H1) and B) Unrelated individuals (H2). Briefly, genotype data is simulated according to our population and inheritance model. Likelihood ratios (LRs) are then computed with true genotypes (dark blue and red lines). Low coverage sequence read data is subsequently generated based on expected depth 1X and standard deviation equal to 1 with e=0.1 and m=5 (see text for details). LRs are then calculated with m=5 and different values on e (see legend). All LRs compare the data given first cousins versus unrelated.

or a BAM-file) was analyzed in FamLink2 as well as in ANGSD generating genotype likelihoods for each genetic marker. Since the comparison is not entirely fair, i.e. ANGSD works on individual reads with assigned quality parameters and phred scores whereas FamLink2 works on read count summary data and overall error parameters, we summarized the number of genetic markers where the two models produced identical and different genotypes, based on the maximum likelihood estimate. We also compared the genotypes against the reference data for the control sample. A summary of the results is illustrated in Table 5 where we note a high concordance between the model in ANGSD and FamLink2. Note that whereas the table presents numbers pertaining to the concordance rate, the full model assigns likelihoods to all genotypes and the outcome from the comparison, although illustrating a high concordance rate, is therefore not entirely relevant to the implementation. Nonetheless, we expect the models to assign similar maximum likelihood estimates for genotypes given the input read count and error rates.

Finally, we performed an all-against-all search using the blind search function in FamLink2 where we compute LR for genetic identity versus the likelihood that the two samples are from unrelated individuals. The results are illustrated in Table 6 and show that the LRs in favor for the samples originating from the same individual are all extremely high, even for the sample with mean coverage below 1X, for all tested values of m and e.

3.3. Authentic case

The results in Tillmar et al. [1] were inconclusive for the authentic case, where a 10X stringent threshold was used to drop markers from genotype calling. In contrast, we applied our model for genotype likelihoods allowing down to 1X markers to be included. We conducted pair wise comparison, results illustrated in Table 7, although a complete pedigree analysis would have been possible in FamLink2. The comparison of individuals denoted 1 A vs M (alleged first cousins twice removed) gave results indicating evidence for a 1C2R relationship (LR> 10^4). Comparison of 2 A vs M (alleged uncle/nephews) gave results indicating evidence for a uncle/nephew relationship (LR> 10^{38}). Finally, comparison of 1 A vs 2 A (known second cousins once removed) now resulted in similar LRs as for the 10X threshold approach $10^2 - 10^3$. The majority of the LRs obtained for the comparisons with an unrelated individual, X, were in the order of 1, except for the comparison between 2 A and X when the value of m was set to 20, for which the LR was calculated to around 300 for a 2C1R vs unrelated comparison. Finally, we computed the full pedigree LR, where both relatives (01 A and 02 A) were included, with high LRs for all selections of m/e, but with limited added information compare to only using the alleged uncle as the only reference.

3.4. Impact of model parameters

3.4.1. Observation model

We summarize the results using, what refer to as, correct likelihood proportions. Briefly this proportion is calculated as the likelihood for the true genotype, given our parameters e/m, divided by the sum of all possible genotype likelihoods (10 in total assuming that all four bases (A/C/G/T) are possible). We note that the interpretation of these result are purely in the sense of providing an understanding of the observation model and the interplay with the complete likelihood model (i.e. with the population and inheritance model) is complex and explored later. For a detailed mathematical evaluation of the properties of our observational model we refer to Supplementary Data. First, we make some theoretical notes, i) when sampling data for a homozygote genotype at depth (D) 1X, the probability to observe an allele different then the true allele is 3e/4, which is typically small, ii) when sampling data for a homozygote genotype with D>1X the probability to observe at least one dropin is $1\text{-}(1-3e/4)^D$ where D is the depth, which for high depths indicates a high probability for at least one dropin, iii) when sampling data for a heterozygote genotype at D=1X the probability to observe a complete allelic dropout is 1, that is only a single true allele is observed, and iv) when sampling data for a heterozygote genotype at D>1X the probability to observe a complete allelic droput is $\sum_{k=0}^{m}\binom{m}{k}\left(\frac{(1-e)k}{m}+e/4\right)^D\left(\frac{1}{2}\right)^m$, see details in Supplementary data. The results are visualized in Fig. 4 and suggest that, i) when the true

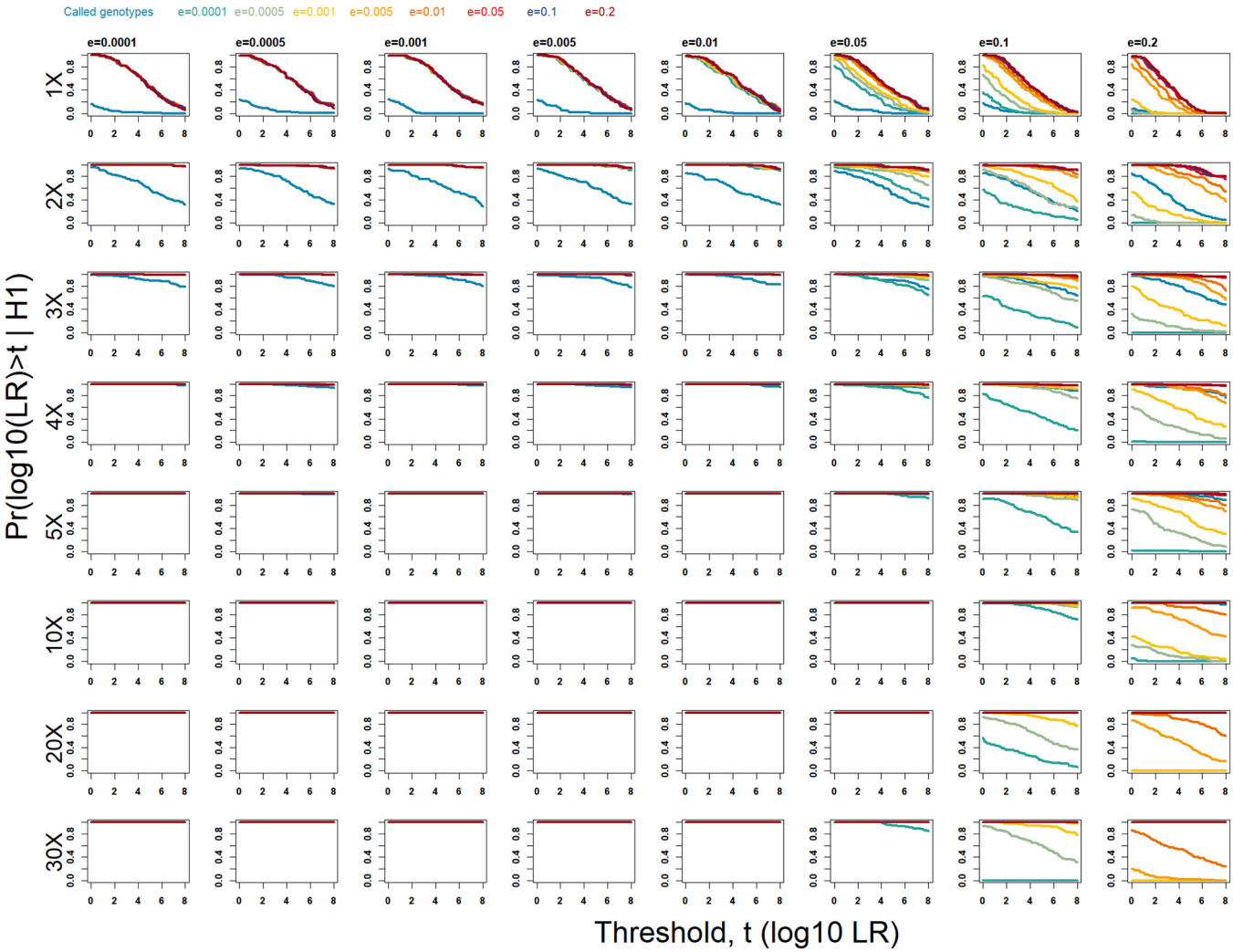

Fig. 6. Exceedance plots with data from 100 simulated genotypes from two first cousins. Details on the simulations is given in the main text. Each row represents data sampled with expected depth (^{1}X–^{30}X) and standard deviation 1X and m=10. The e parameter is varied in the sampling of sequence read data (each column of the figure) and subsequently analyzed with different choices of e (see top legend). In addition, genotypes are called for each simulated sequence data based on a method described in main text.

genotype is a heterozygote and a complete allelic dropout is observed, i.e. only reads for a single allele is observed, (see Fig. 4A for rate and Fig. 4B for results), lower choices of *m* will increase the likelihood proportion for the true heterozygote, but never exceed 0.2, ii) when the true genotype is a heterozygote and both alleles is observed (see Fig. 4C), lower choices of *m* will decrease the likelihood proportion for the true heterozygote, but for higher depths, say >10X, the impact of *m* is small, iii) when the true genotype is a homozygote and one (or more) alleles have dropped in (see Figure 9 D and E), increasing m will increase the likelihood proportion for the true homozygote, in particular for high depths (>=10X) and iv) when the true genotype is a homozygote and only the single (true) allele is observed (see Fig. 4F), increasing *m* also increases the likelihood proportion for the true homozygote genotype whereas *m* has little or no impact when coverage is low (1–2X)

3.4.2. Pilot study

Next, we studied the impact of *e* and *m* in the calculation of pedigree likelihoods (i.e. genotype likelihoods combined with inheritance and population data). We used 1000 simulated genotypes for a pair of first cousins as well as 1000 genotypes from two unrelated individuals and sampled read data with three different depths (average coverage), standard deviation and values on *e* and *m*, see Table 2. We summarize the results by means of exceedance plots, see Fig. 5 for an illustrative example, where we compute the probability that the LR will exceed a given threshold. For reference purposes we include the LR computed using complete genotype data (with and without a model for linkage). Fig. 5 illustrates the importance of a model for genotype likelihoods, in particular for low coverage data (1X) whereas the impact is less evident for higher coverage (5X) with our choice of parameters, see Supplementary Figure 5. Fig. 5A further illustrates that even though the true positive rate (for a given threshold) increases when the *e* parameter approaches its true (simulated) value, the false positive rate also increases, see Fig. 5B. However, we note that if the LR threshold is 1000 (equivalent to a log10 LR of 3), the false positive rate is zero for all choices of *e* used in the analysis.

3.4.3. Study of the parameter e

Next we focus on extending the conditions when sampling sequence read data with fixed value on *m=10* but varying *e*, both when simulating data and when analyzing the data. An excerpt of the results is visualized in Fig. 6 when genotypes are based on data from two first cousins. The results indicate that, i) a high value of *e* always results in high LRs in favour of first cousins, regardless of other conditions, ii) calling of genotypes fails, i.e. LRs are deflated, for high error rates, e.g. *e*>=0.1 and iii) for very low depths (e.g. below 5X), increasingly complete locus dropouts will cause the LR to decrease regardless of the value of *e*. When

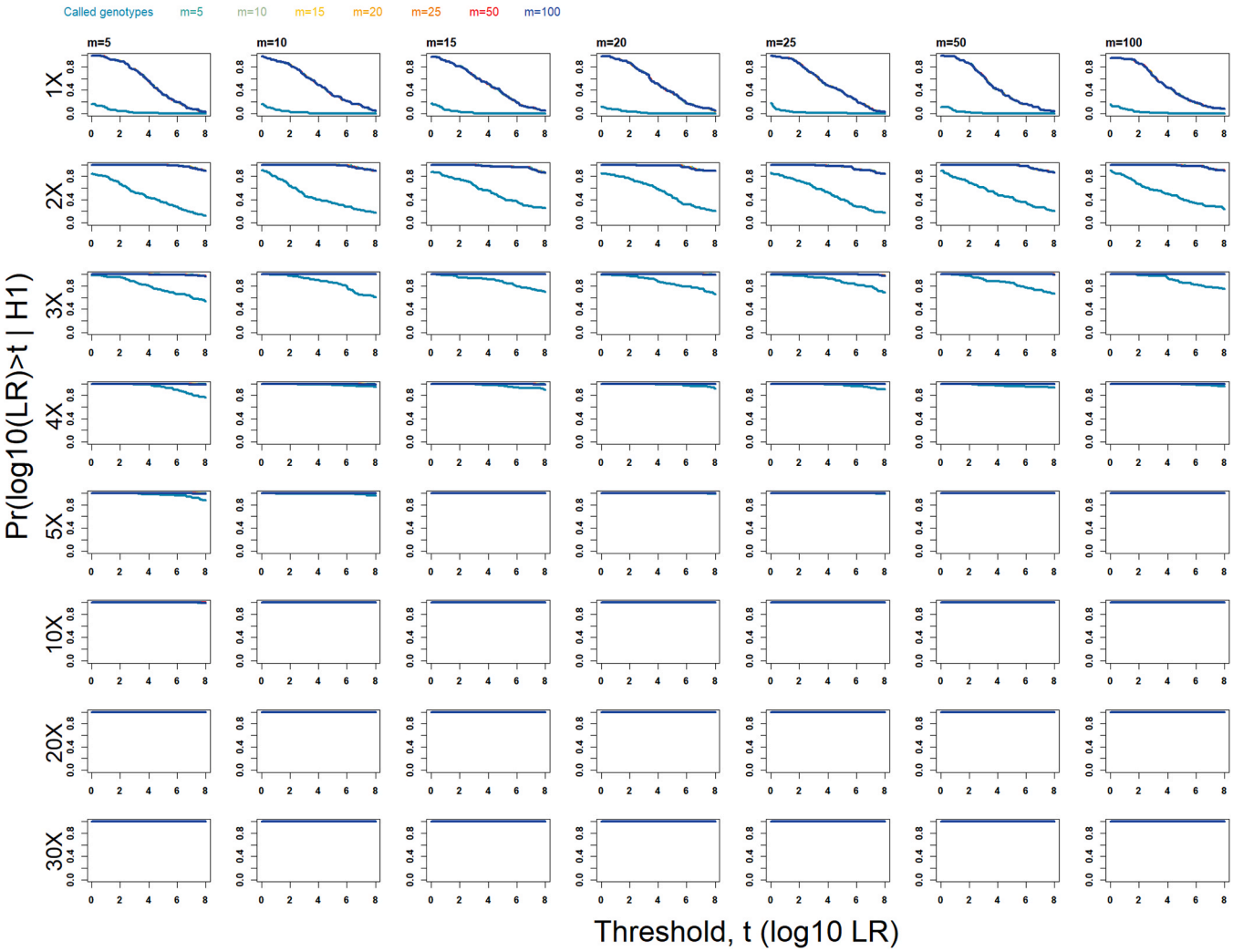

Fig. 7. Exceedance plots with data from 100 simulated genotypes from two first cousins. Details on the simulations is given in the main text. Each row represents data sampled using different depths (1X–30X) with standard deviation 0.5)X and e=0.1. The m parameter is varied in the simulation of sequence read data (each column of the figure) and subsequently analyzed with different choices of m (see top legend). In addition, genotypes are called for each simulated sequence data based on a method described in main text.

the genotypes are based on two unrelated individuals, Supplementary Figure 6,we note that i) for low to medium depths (e.g. >4X), the results will always favour unrelated, regardless of choices of e and ii) for very low depths (e.g. <5X), locus dropouts and to some extent the value of e will falsely inflate the LR.

3.4.4. *Study of the parameter m*

Finally, we performed extended sampling of read data varying both e and m. The results indicate that the choice of m has limited impact on the results in this example, see Fig. 7 In fact the results obtained with varying m used in the analysis cannot easily be visually distinguished in Fig. 7 regardless of the value of m used in the simulations.

4. Discussion

The model described in Mostad et al. [11] allows for general likelihood computations combining low and high coverage sequencing data as well as determined genotypes in questions of relationship inference. The model essentially combines the inheritance model implemented in Merlin [10] expanding the population model to handle subpopulation correction, with a model for sequencing data, similar to what is described in Korneliussen et al. [9]. Our observation model can be further tuned to include parameters such as mapping quality and average sequence base quality which is the focus of current research. Since most forensic sequencing applications deal with read count data and potentially quality parameters for instance as part of a vcf-like file, we have refrained from implementing a model where individual mapped reads (e.g. BAM file) can be used as input. For such cases we refer to commonly available callers such as GATK [5], bcftools [7] or ansgd [17]. All of the afore-mentioned work on BAM-files to produce a genotype likelihood and a vcf-file. The vcf-file with likelihoods attached to each genetic marker can subsequently be used as input for FamLink2. In our implementation we study the output generated in CLC Genomic Workbench using the “Identify known mutations” tool where individual counts for each allele at a given positions can be obtained.

We performed a sensitivity analysis where two parameters of our observation model are studied, i.e. e which is used to address sequencing and mapping errors, i.e. dropins, and m which is used to address allelic imbalance or dropout of alleles. We studied both the immediate impact on the genotype likelihood using sequence data sampled with different conditions; see Fig. 4, as well as the interplay with the joint likelihood for the pedigree, Figs. 5–7. The results illustrate a complex interaction where it appears the value of m has limited impact on the joint likelihood in our sensitivity analysis. In summary, we recommend a comparatively high value on e (say 0.01) and low value of m (say 5) to account for allelic dropouts as well as sequencing/mapping errors. In addition to the

data presented in this validation study, we believe a more in-depth study is necessary to evaluate and develop our model for genotype likelihoods. This study should focus on realistic forensic grade samples to elude what parameters adequately describe the data to further expand the recommendations for end-users. For instance individual read quality metrics, e.g. mapping quality and base quality scores should be explored in addition to read counts only. Moreover, FamLink2 currently assigns the same values on e/m for all samples whereas the observation model allows sample and even marker specific numbers to be assigned which will also likely be implemented in future version of the software.

Our population model does not currently handle linkage disequilibrium (LD) which we believe is a minor limitation as most forensic panels are constructed to avoid this, see for instance Tillmar et al. as well as Gordon et al. However, the model can be expanded using a similar approach as Abecasis et al. [21] where superloci, i.e. clusters of closely located markers, are constructed.

One parameter not explored in the present study, but with potential impact in the likelihood model is the γ parameter of our population model. In particular with high dropin error rates (i.e. high values on e) random, unobserved, alleles without a frequency in the population will occur in the data and could inflate the LRs if it appears in the sequence data for two samples involved in the pedigree. We suggest a comparatively high value of γ (say 0.05) to partly mitigate such issues, but further studies should explore the interplay between this population parameter and our observation model.

We did not specifically study the impact of accounting for linkage, or not, a subject that has been extensively studied previously, see for instance Tillmar et al. [1] for the markers in the FORCE panel and Kling et al. [19]. These, and other studies, have noted the impact of not accounting for linkage with expanded marker panels necessitating a model for recombinations.

A limitation of the current implementation is that male and female genetic maps are not differentiated, i.e. the inheritance vectors need a sex-average value to be used in each transmission. It is well know that male have a lower rate of recombination across most parts of the genome [Ref Behrer et al.], however we argue that few implementations accounts for this difference and that information about male/female transitions is not always available for pedigrees.

Future work will also focus on implementing further improvements relating to pedigree symmetries, partly described in the Supplementary material of Mostad et al. [11] which could greatly improve computational speed for extended pedigrees with several typed individuals, where we acknowledge that FamLink2 currently is time consuming.

CRediT authorship contribution statement

Daniel Kling: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Petter Mostad:** Writing – original draft, Methodology, Conceptualization. **Andreas Tillmar:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Competing Interest

None.

Acknowledgements

We would like to acknowledge the comments from two anonymous reviewers that have greatly improved both the quality as well as the correctness of the manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2024.103150.

References

[1] A. Tillmar, et al., The FORCE Panel: an all-in-one SNP marker set for confirming investigative genetic genealogy leads and for general forensic applications, Genes 12 (12) (2021) 1968.
[2] C. Phillips, et al., A compilation of tri-allelic SNPs from 1000 genomes and use of the most polymorphic loci for a large-scale human identification panel, Forensic Sci. Int.: Genet. 46 (2020) 102232.
[3] J. Snedecor, et al., Fast and accurate kinship estimation using sparse SNPs in relatively large database searches, Forensic Sci. Int.: Genet. 61 (2022) 102769.
[4] E.M. Gorden, et al., Extended kinship analysis of historical remains using SNP capture, Forensic Sci. Int. Genet. 57 (2022) 102636.
[5] M.A. DePristo, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (5) (2011) 491–498.
[6] A. McKenna, et al., The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (9) (2010) 1297–1303.
[7] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics 27 (21) (2011) 2987–2993.
[8] A. Tillmar, et al., Whole-genome sequencing of human remains to enable genealogy DNA database searches–a case report, Forensic Sci. Int.: Genet. 46 (2020) 102233.
[9] T.S. Korneliussen, I. Moltke, NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data, Bioinformatics 31 (24) (2015) 4009–4011.
[10] G.R. Abecasis, et al., Merlin–rapid analysis of dense genetic maps using sparse gene flow trees, Nat. Genet. 30 (1) (2002) 97–101.
[11] P. Mostad, A. Tillmar, D. Kling, Improved computations for relationship inference using low-coverage sequencing data, BMC Bioinforma. 24 (1) (2023) 90.
[12] M.D. Coble, et al., DNA commission of the international society for forensic genetics: recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, Forensic Sci. Int.: Genet. 25 (2016) 191–197.
[13] E.S. Lander, P. Green, Construction of multilocus genetic linkage maps in humans, Proc. Natl. Acad. Sci. USA 84 (8) (1987) 2363–2367.
[14] T.C. Matise, et al., A second-generation combined linkage–physical map of the human genome, Genome Res. 17 (12) (2007) 1783–1786.
[15] G.P. Consortium, A global reference for human genetic variation, Nature 526 (7571) (2015) 68–74.
[16] A. Staadig, J. Hedman, A. Tillmar, Applying unique molecular indices with an extensive all-in-one forensic snp panel for improved genotype accuracy and sensitivity, Genes 14 (4) (2023) 818.
[17] T.S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: analysis of next generation sequencing data, BMC Bioinforma. 15 (1) (2014) 1–13.
[18] J.W. MacCluer, et al., Pedigree analysis by computer simulation, Zoo. Biol. 5 (2) (1986) 147–160.
[19] D. Kling, On the use of dense sets of SNP markers and their potential in relationship inference, Forensic Sci. Int.: Genet. 39 (2019) 19–31.
[20] D. Kling, T. Egeland, A.O. Tillmar, FamLink - a user friendly software for linkage calculations in family genetics, Forensic Sci. Int.: Genet. 6 (5) (2012) 616–620.
[21] G.R. Abecasis, J.E. Wigginton, Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers, Am. J. Hum. Genet. 77 (5) (2005) 754–767.