Moment Estimators of Relatedness From Low-Depth Whole-Genome Sequencing Data

Anthony Francis Herzig (anthony.herzig@inserm.fr)
Inserm, Univ Brest, EFS, UMR 1078, GGB

Marina Ciullo
Institute of Genetics and Biophysics A. Buzzati-Traverso - CNR and IRCCS Neuromed

Anne-Louise Leutenegger
Inserm, Université de Paris, UMR 1141

Hervé Perdry
CESP Inserm U1018, Université Paris-Saclay, Villejuif

FranceGenRef Consortium
LABEX GENMED, Centre National de Recherche en Génomique Humaine

Research Article

Keywords: Kinship, fraternity coefficient, low-depth, sequencing data, genotype likelihoods, moment estimators

Posted Date: January 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1109592/v1

Moment estimators of relatedness from low-depth whole-genome sequencing data

Herzig AF[1], Ciullo M[2,3], FranceGenRef Consortium[4], Leutenegger A-L[5], Perdry H[6]

1. Inserm, Univ Brest, EFS, UMR 1078, GGB, F-29200 Brest, France
2. Institute of Genetics and Biophysics A. Buzzati-Traverso - CNR, Naples, Italy
3. IRCCS Neuromed, Pozzilli, Isernia, Italy
4. LABEX GENMED, Centre National de Recherche en Génomique Humaine, Evry, Paris
5. Inserm, Université de Paris, UMR 1141, NeuroDiderot, F-75019 Paris, France
6. CESP Inserm U1018, Université Paris-Saclay, Villejuif, France

Corresponding Author:

Anthony Francis Herzig

anthony.herzig@inserm.fr

+33298017361

Inserm UMR 1078, 22 Avenue Camille Desmoulins, 29238 Brest, France

Abstract

Background. Estimating relatedness is an important step for many genetic study designs. A variety of methods for estimating coefficients of pairwise relatedness from genotype data have been proposed. Both the kinship coefficient φ and the fraternity coefficient ψ for all pairs of individuals are of interest. However, when dealing with low-depth sequencing or imputation data, individual level genotypes cannot be confidently called. To ignore such uncertainty is known to result in biased estimates. Accordingly, methods have recently been developed to estimate kinship from uncertain genotypes.

Results. We present new method-of-moment estimators of both the coefficients φ and ψ calculated directly from genotype likelihoods. We have simulated low-depth genetic data for a sample of individuals with extensive relatedness by using the complex pedigree of the known genetic isolates of Cilento in South Italy. Through this simulation, we explore the behaviour of our estimators, demonstrate their properties, and show advantages over alternative methods. A demonstration of our method is given for a sample of 150 French individuals with down-sampled sequencing data.

Conclusions. We find that our method can provide accurate relatedness estimates whilst holding advantages over existing methods in terms of robustness, independence from external software, and required computation time. The method presented in this paper is referred to LowKi (**Low**-depth **Ki**nship) and has been made available in an R package (https://github.com/genostats/LowKi).

Keywords: Kinship, fraternity coefficient, low-depth, sequencing data, genotype likelihoods, moment estimators.

Background

Accurate estimates of genetic relatedness between individual organisms are essential for a wide range of study designs and analyses strategies currently at play in plant, animal, or human genetics. These coefficients that describe the similarity and extent of shared origin between genomes may currently be estimated in a large variety of ways and a multitude of methods have been proposed. One's data characteristics and envisaged analyses will dictate the most appropriate method to be used. For overviews of the current options for relatedness estimation, and its utility, we point the reader to (1–4) and references therein.

In recent years the cost of whole-genome sequencing (WGS) has continued to tumble. Accordingly, more and more study designs have emerged that require large sample sizes to power their analyses. The depth of sequencing carried out over a large sample will have a significant effect on a researcher's budget. Whilst the accuracy of genotyping is highly dependent on the depth (5), there are often more advantages to being able to sequence a large number of individuals but at a low depth. Recent high profile association studies using this approach include (6) and (7). Indeed, low-depth sequencing data was used in many of cohorts participating in the Haplotype Reference Consortium panel (8). Furthermore, shallow sequencing is often unavoidable in the expanding field of ancient DNA, where the possibilities of sequencing DNA from remains of long deceased organisms (9) are being widely explored. Whilst technological advances allow for greater and greater accuracy in this field, in some circumstances, sequencing to a high depth may simply not be feasible due to the paucity of available genetic material. Another area where genetic material of high quality might be difficult to ascertain is in the study of wild animal populations where DNA is collected from more challenging sources such as hair, feathers, egg membranes or similar (10).

In adaptation to this recent trend of low-depth sequencing studies, a number of methods have been proposed to estimate relatedness coefficients from such datasets. The specificity of these methods is that they work upon genotype likelihoods or posterior probabilities, thus incorporating the uncertainty of genotype calls. These include Hidden Markov Model (HMM) based methods, maximum likelihood expectation based methods, and method-of-moment estimates. The former two approaches can be computationally heavy while moment-based-estimators present a quick and simple alternative. However, the loss of information entailed by analysing genotype likelihoods as a proxy for true genotypes will lead to biased estimates of relatedness which methods using moment-based-estimators need to account for.

Moment-based methods have so far only been developed for estimation of the kinship coefficient and the one software that performs this estimation requires an intermediate step from an existing HMM method. We propose here LowKi, a method to directly estimate genetic relatedness matrices from genotype likelihoods in a single step, which has now been incorporated into the genetic data management and analysis R-package 'Gaston' (11). LowKi calculates moment estimates of relatedness in the form of a genetic relatedness matrices (GRMs) with suitable adjustments for the genotype uncertainty that is present with low-depth WGS data. For a pair of individuals i and i', LowKi provides estimates of $\varphi^{ii'}$, the kinship coefficient of individuals i and i'. This is the probability that a pair of randomly drawn alleles from individual i and i' at the same locus will be in a state of Identity-by-Descent (IBD). LowKi also provides a moment-estimate of $\psi^{ii'}$ the fraternity coefficient which is the proportion of the genome for which individuals i and i' share two pairs of alleles at the same locus (IBD=2).

First, 'naïve' moment estimators were defined by an approximation of the construction of the classical moment estimators used for genotype data; but based on individual genotypes

likelihoods. These estimators, which are referred to as 'unadjusted estimates' in this study, are biased. Simulation studies show that as the average read depth decreases, the bias increased. It is indeed intuitive that additional uncertainty or 'fuzziness' in the genotype likelihoods gives a stronger downward bias in a moment-estimate. This makes sense when considering that the additional fuzziness represents an increasing lack of information about the genotypes as random error contributions to the genotype likelihoods (occurring independently between individuals) become more and more prevalent. We found that it sufficed to fit regression models between point-wise moment estimators and a summary statistic of the genotype likelihood fuzziness to obtain accurate estimates (denominated in the text as 'adjusted estimates').

To assess our approach, we have analysed both simulated and real data. Firstly, we used a simulation dataset which constitutes 1,444 simulated whole-genome sequences derived from the complex pedigree structure of the genetic isolates of Cilento (12–14). This simulation dataset was first produced to assess phasing and imputation methods (15) before being used as a tool to explore heritability estimation (16). Here we overlay a second layer of data simulation to convert our simulated sequencing data into low-depth sequencing data. To complement our simulation analysis, we also apply our models to a real dataset of 150 individuals from the FranceGenRef WGS panel (Labex GENMED http://www.genmed.fr/). These individuals have been sequenced to a depth of 30-40× so we down-sampled individual bam files to create a dataset representative of WGS data at a depth of 2.5×. Our aim was to show that we can recover relatedness matrices similar to GRMs calculated on high quality genotypes from low-depth data in an expedient manner. We compared our approach to two existing methods: SEEKIN (v1.01) (17) and NGSRelateV2 (v2) (18,19). We show that our

estimates are accurate and are less time consuming to compute than those provided by alternative software.

Results

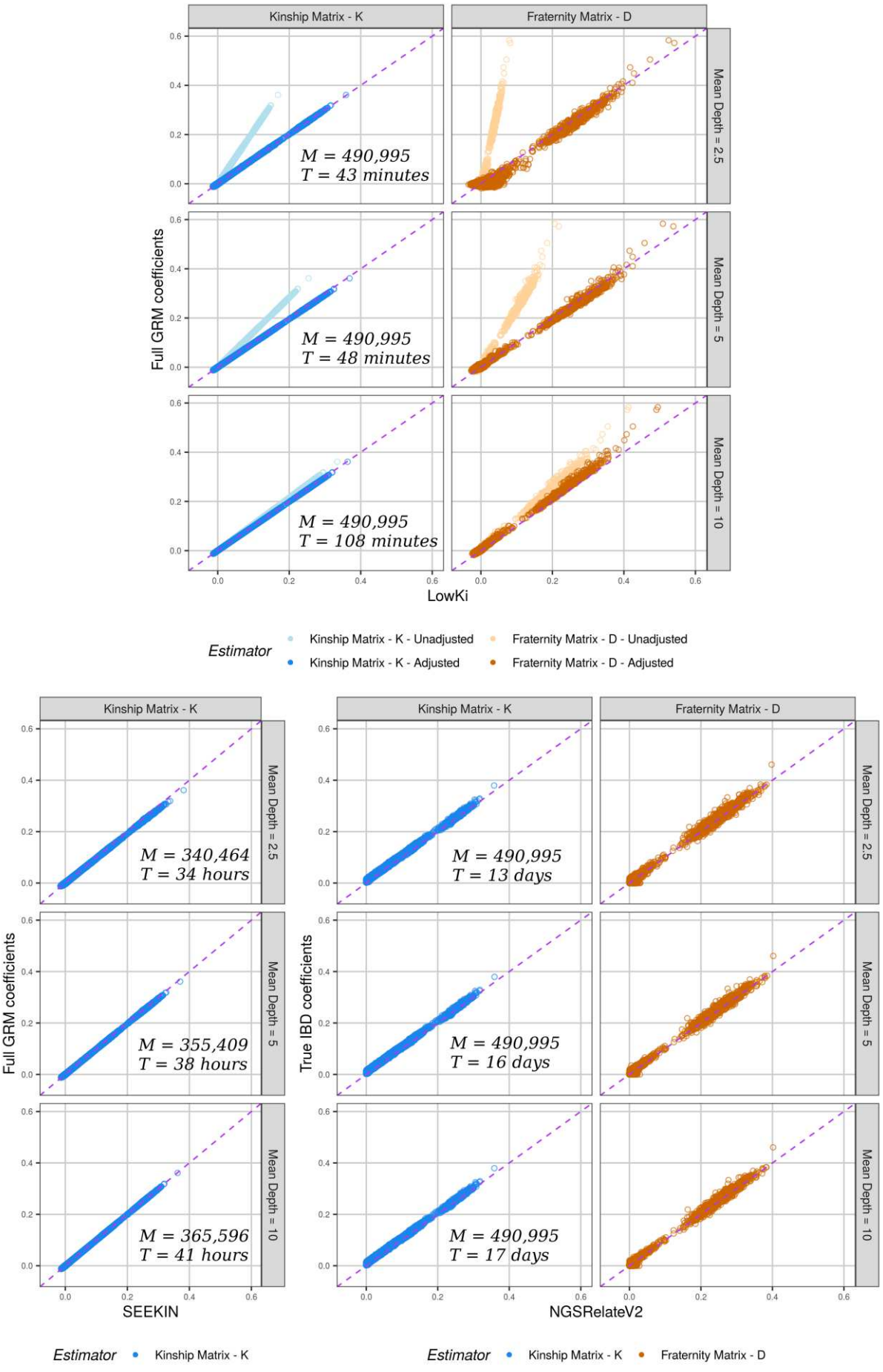
Relatedness estimation from genotype likelihoods in CilentoSim

Our primary simulation dataset (here denoted as 'CilentoSim' and described in the Methods) comprises 1,444 individuals and 490,995 genetic variants across the 22 autosomal chromosomes (see Methods). We established that this variant set was appropriate for the calculation of a GRM as this set captured the known pedigree structure of Cilento. This is seen by comparing the kinship and fraternity GRMs (calculated from the simulated genotypes) to the true IBD sharing matrices calculated based on records of all haplotype mosaics created in the simulation (see Methods) (Supplementary Figure 1). For Kinship, the GRM gives a very precise estimate of the exact simulated IBD-sharing fractions. For fraternity, the GRM estimates are highly correlated with the simulated IBD-sharing despite a lower precision compared to kinship.

We artificially reduced the depth of our simulated sequencing data by drawing random alleles from each simulated individual genotype to a specified depth and then replacing in our simulation the true genotype with three genotype likelihoods. The method used here is based on the simulation proposed by Kim et al. (20), uses a simplified version of the genotype likelihood model of GATK (21–23), and is described fully in the Methods. We used this additional layer of simulation to give new datasets with average read depths of 2.5×, 5×, and 10×.

We applied our method alongside SEEKIN and NGSRelateV2. In Figure 1a, the off-diagonal elements of our estimated GRM matrices are compared to the the 'Full GRM' estimates from

complete simulated genotypes. As a moment estimator, SEEKIN is also compared to the 'Full GRM' but NGSRelateV2 (a maximum likelihood estimator) is compared to the simulated IBD sharing probabilities (which are similar but not identical to the Full GRM estimates, see Supplementary Figure 1) to give a fairest assessment of the three methods. LowKi is able to recover the structure of the full kinship and fraternity matrices ('Adjusted Estimates'; dark blue and dark brown). Included in Figure 1a are also the 'Unadjusted Estimates' (light blue and light brown) from our model that are downwardly biased by a multiplicative factor. This demonstrates the efficiency of the adjustment procedure which has been sketched in the Introduction and is described in detail in the Methods.


Kinship Matrix - K
Fraternity Matrix - D
Mean Depth = 2.5
Mean Depth = 5
Mean Depth = 10
M = 490,995
T = 43 minutes
M = 490,995
T = 48 minutes
M = 490,995
T = 108 minutes
Full GRM coefficients
LowKi
Estimator
Kinship Matrix - K - Unadjusted
Kinship Matrix - K - Adjusted
Fraternity Matrix - D - Unadjusted
Fraternity Matrix - D - Adjusted
M = 340,464
T = 34 hours
M = 355,409
T = 38 hours
M = 365,596
T = 41 hours
M = 490,995
T = 13 days
M = 490,995
T = 16 days
M = 490,995
T = 17 days
SEEKIN
True IBD coefficients
NGSRelateV2

142

Figure 1a

LowKi estimates for kinship and fraternity for CilentoSim. Off-diagonal elements of the estimated kinship and fraternity matrices against the true simulated IBD sharing coefficient in CilentoSim at three different simulated mean read depths (2.5×, 5×, and 10×). Lighter colours represent the unadjusted estimated from our method and the darker colours give the final recalibrated estimates. The number of variants (M) and the time (T) required for the calculation of the two matrices are overlaid on the figure.

Figure 1b-c

Corresponding estimates from SEEKIN (kinship only) and NGSRelateV2.

Comparison to existing software on simulated data

In Figure 1b and 1c, relatedness estimates given by two alternative algorithms, SEEKIN and NGSRelateV2, respectively, along with running times for these programs. SEEKIN only produces an estimate for the kinship matrix and indeed uses in part a similar moment-estimator to our method presented here. SEEKIN gave very accurate kinship estimates. The key specificities of SEEKIN involve an intermediate step of the imputation software BEAGLE (v4.1) (24), the leveraging of an external reference panel (here the 1000 Genomes Project phase 3 haplotype reference panel (25) was used) and a re-weighting based on the imputation quality of variants in the summation that forms each GRM entry. As the initial step of BEAGLE cannot be avoided, we include the runtime of BEAGLE into the run time of SEEKIN. For low-depth data, running BEAGLE is very time consuming. We followed the recommendations for using BEAGLE as described by the authors of SEEKIN. The use of BEAGLE will change the data in two particular ways: firstly, the uncertainty present in the initial data will be largely removed as BEAGLE will effectively take the prior information given to it in the form of genotype likelihoods and add more precision based on similarities between pairs of individuals in the sample or between individuals in the sample and the external panel of reference haplotypes using the same haplotype clustering HMM machinery as is applied in BEAGLE's haplotype phasing and genotype imputation methods. Secondly, running BEAGLE is likely to require the

removal of some variants. For example, on our simulated dataset with genotype likelihoods created using a mean depth of 2.5×, for our initial dataset of 490,995, BEAGLE returns information for only 340,464 variants. In Supplementary Figure 2, the difference in precision is displayed between the initial genotype likelihoods supplied to BEAGLE (for a random selection of 25,000 variants) and the posterior genotype probabilities. This demonstrates the importance of the use of BEAGLE to the SEEKIN method.

A different approach is proposed by NGSRelateV2 which directly estimates relatedness parameters through maximum likelihood estimation. Indeed, for the fraternity matrix, the true IBD coefficients were estimated more precisely with NGSRelateV2 than with a GRM on simulated genotypes (a comparison can be made between Figure 1 and Supplementary Figure 1). This software also produces additional information as it gives estimates for all nine condensed identity-by-descent states. NGSRelateV2 gave very accurate estimates for both kinship and fraternity in the CilentoSimu analysis though did require extensive amounts of running time on default settings. NGSRelateV2 is multithreaded and uses four threads as a default; we did not alter this default setting. Using more than the four default threads would give a big increase in speed but this software will remain computationally expensive for large sample sizes.

Testing with real data

We also applied our method, SEEKIN, and NGSRelateV2 to a set of real genotypes. 150 individuals with WGS data were made available to us from the FranceGenRef panel; of which all individuals are not closely-related except for two pairs of siblings. We down-sampled this dataset from 30-40× to 2.5× in order to create realistic low-depth WGS data. The estimates of the Kinship and Fraternity matrix entries for these two sibling pairs are given in Table 1.

Moment-estimators on down-sampled data can were compared against moment-estimators on the original 30-40× data with one exception: the estimates from NGSRelateV2 on down-sampled data were compared to the estimates of NGSRelateV2 applied the original 30-40× data. This was because with a sample size of 150, the fraternity coefficients in particular may be imprecisely estimated even on the original data and so it would not necessarily be fair to benchmark NGSRelateV2 against GRM estimates.

Table 1		WGS data at 30-40×		Down-sampled WGS data at ~ 2.5×				
		Gaston Full GRM	NGS-RelateV2	LowKi (Unadjusted)*	LowKi	BEAGLE+ LowKi	BEAGLE+ SEEKIN	NGS-RelateV2‡
Sibling Pair 1	$\hat{\varphi}$	0.258	0.268	0.122	0.262	0.254	0.268	0.256
	$\hat{\psi}$	0.294	0.379	0.059	0.192	0.237	†	0.465
Sibling Pair 2	$\hat{\varphi}$	0.216	0.228	0.100	0.215	0.214	0.226	0.216
	$\hat{\psi}$	0.196	0.270	0.035	0.113	0.150	†	0.377
M (number of variants)				1,009,181	1,009,181	949,075	917,715	1,051,789
Time				4 minutes	4 minutes	15 hours **	15 hours **	6 hours

$\hat{\varphi}$: estimate of the two siblings' unobserved kinship coefficient φ
$\hat{\psi}$: estimate of the two siblings' unobserved fraternity coefficient ψ
* We also give the unadjusted values from LowKi, similar to in Figure 1a
** The time required for these estimators is almost entirely due to BEAGLE, both LowKi and SEEKIN require only a few minutes
† SEEKIN does not provide estimates of ψ
‡ NGSRelateV2 results on down-sampled data should be compared to results from the same software but applied to the original data that had not been down-sampled

All methods were able to distinguish the two pairs as being closely related compared to all other pairs (Supplementary Figure 3). All methods produced accurate estimates for the kinship coefficient but the fraternity coefficient proved difficult for all methods to estimate accurately. One intuitive solution is to use the same approach as SEEKIN and first performed

imputation with BEAGLE; this allowed us to significantly improve our estimates of the fraternity coefficients between the two sibling pairs (Supplementary Figure 3, panel (c)). This allowed us to have improved estimates for the fraternity matrix. In every case, LowKi underestimates both fraternity coefficients; however, NGSRelateV2 gave overestimations. Even without knowing the true fraternity coefficients, an overestimation is apparent as NGSRelateV2 returned fraternity coefficients outside of the expected range of likely coefficients between siblings in an outbred population (approximately between 0.1 and 0.4) (26). LowKi's estimates remain within this range. When running NGSRelateV2 on data without down-sampling, we found significantly lower fraternity coefficients for the two sibling pairs (Table 1, Supplementary Figure 3).

Discussion

It is intuitive that with data of the huge breadth of the whole human genome, even when the quality of sequencing data is extremely low, relatedness between individuals should still be captured. Existing methods have either involved maximum likelihood estimation or moment-estimators of relatedness coefficients. The former estimators carry a high computation burden and require a modelling of the mechanism that links true genotypes to genotype likelihoods. The latter moment estimators have a lower computational burden, but will however suffer from bias. This can either be dealt with using an intermediate imputation algorithm to improve the data as is the case of SEEKIN that requires BEAGLE; or by attempting to explicitly account for the bias as in the method we have developed here.

By estimating orthogonal components for additive and non-additive genotypic effects, we constructed a moment estimator for the kinship the fraternity coefficient from low-depth data. Such a moment-estimator has never been provided for fraternity by an alternative

software. Estimation of fraternity is important for classifying relatives and also for exploring the effects of non-additive genetic effects (16). Our moment-estimators for fraternity were sufficient to distinguish pairs of siblings in the analysis of FranceGenRef where even NGSRelateV2 was unable to give perfect estimates of fraternity from data down-sampled to 2.5×. For the fraternity matrix our re-adjusted estimators were not as accurate as for kinship, though it was clear that fraternity coefficients are harder to estimate not always perfectly estimated with moment-estimators even when using genotype data. The NGSRelateV2 model expects genotype likelihoods to accurately reflect the probabilities of the unobserved value of the hidden genotype. This may explain why it gave very good estimates in the simulation based on Cilento (where data was created with a naïve version of the GATK models and was generated from simulated genotypes) but performed less well in the more complex real data scenario.

To correct for bias in the estimates of LowKi, we introduced an innovative simulation extrapolation type approach that could account for unseen error models and the inherent reduction in variance typical to low-depth genetic data. These adjustment methods proposed should be robust to different sources of bias arising from genotype uncertainty coming from different types of bioinformatics pipeline. This could give more flexibility than likelihood based methods such as lcMLkin (27), NGSRelateV2, as well as similar methods proposed for estimating inbreeding coefficients (28). Note that we did not test lcMLkin here following its assessment in the publication presenting SEEKIN (17). The adjustment technique developed for LowKi could be harnessed in other areas of research involving shallow sequencing data.

We showed that the alternative to such bias-correction, using an external imputation algorithm, could also lead to lengthy run times and a reliance on the accuracy of the external

algorithm. Notwithstanding such observations, the methodology of SEEKIN that uses the intermediate step of BEAGLE is clearly often highly effective and can also be used in conjunction with LowKi. Indeed, BEAGLE probably worked exceptionally well in the case of our CilentoSim dataset due to the many pairs of very closely related individuals in the sample. However, in the circumstance where only a small sample size is present or when an appropriate reference panel cannot be ascertained (both might be the case in studies of ancient DNA or small isolated populations for examples), it is beneficial to have a method for proceeding directly to relatedness estimates from genotype likelihood data.

Conclusion

LowKi was effective at computing very accurate GRMs in a large sample of individuals with a full spectrum of IBD-sharing between pairs of related individuals in a detailed simulation study. We complemented this analysis by assessing an example of real low-depth genetic data from FranceGenRef, where our re-adjusted relatedness coefficient estimates were able to quickly and accurately identify the pairs of siblings in the sample. By analysing real data, we have illustrated that our estimators perform well outside of the idealised setting of a simulations. Real data will harbour phenomena such as allele-balance bias (29) or region-specific sequencing error rates (30) so it was important to verify our estimators on an example of true sequencing data.

When comparing to existing methods, LowKi does not require the use of intermediate software such as BEAGLE and requires the least computation time. The innovative adjustment method applied in LowKi gives flexibility to the method to account for different possible sources of bias. The LowKi methods proposed here for estimating relatedness have been made available at https://github.com/genostats/LowKi and work in conjunction with the existing R-

package Gaston. This represents a fast and accurate standalone option for computing kinship and fraternity coefficients from low-depth sequencing data.

Methods

Throughout, the index $i \in 1, \ldots, N$ will denote individuals (with two different individuals denoted as i and i') and $j \in 1, \ldots, M$ will indicate bi-allelic genetic variants. Individual level genotype data are denoted as G^{ij} which takes values in $\{0,1,2\}$ for the three possible genotypes AA, Aa, and aa, respectively.

Simulation of low-depth data

Existing simulated WGS data for 1,444 individuals based on the pedigree of the Cilento isolates was our starting point (16). These simulated individuals were constructed as mosaics of haplotype chunks sourced from the UK10K imputation panel. The formation of mosaic haplotypes from the UK10K imputation reference panel (31) has been described in two previous studies (15,16). Through gene-dropping, the individuals share chunks in accordance with the known pedigree of Cilento by means of gene-dropping (32) on to the pedigree. By recording the source of each chunk (within the UK10K), we have knowledge of the exact IBD-sharing probabilities in the simulated population. For this study, we added an additional layer of simulation to translate simulated genotypes into simulated genotype-likelihoods typical of low-depth WGS data. We also only retained 490,995 variants by first selecting those with a minor allele frequency above 5% and then by performing pruning on linkage disequilibrium with Gaston.

In the Cilento cohort, there are 19 individuals with whole-genome sequencing data. These individuals were sequenced to an average depth of 50-60×. From this dataset, we took a list of per-variant mean read depths and scaled each entry so that the global mean read depth

would either be 10, 5, or 2.5. These lists became the lists of mean depths for each variant for our simulation. For each individual level genotype G^{ij} and for an assigned average read depth d_j for the position, we draw three sets of reads to represent the number of reads carrying the reference allele A, the minor allele a, and error reads that carry a base that matches neither A or a. The size of these three groups are denoted as R_A, R_a, and R_ε. We model the occurrence of reads with Poisson distributions and thus draw R_A as Poisson with parameter $\rho_A^{ij} d_j$, R_a as Poisson with parameter $\rho_a^{ij} d_j$, and R_ε as Poisson with parameter $\rho_\varepsilon^{ij} d_j$. These parameters have values depending on the true genotypes G^{ij} and the error rate ε_j at the position as shown in Table 2.

Table 2	$G^{ij}=0$ Genotype AA	$G^{ij}=1$ Genotype Aa	$G^{ij}=2$ Genotype aa
ρ_A^{ij}	$1-\varepsilon_j$	$\frac{1}{2}(1-\varepsilon_j)+\frac{1}{6}\varepsilon_j$	$\frac{1}{3}\varepsilon_j$
ρ_a^{ij}	$\frac{1}{3}\varepsilon_j$	$\frac{1}{2}(1-\varepsilon_j)+\frac{1}{6}\varepsilon_j$	$1-\varepsilon_j$
ρ_ε^{ij}	$\frac{2}{3}\varepsilon_j$	$\frac{2}{3}\varepsilon_j$	$\frac{2}{3}\varepsilon_j$

The values of ε_j were drawn randomly as 10^{-u_j} with u_j drawn uniformly between 2 and 3. In any case where $R_A = R_a = 0$, we set the all three genotype likelihoods to missing. In order to compute genotype likelihoods, we apply a flat prior and binomial likelihoods as used in the simplest interpretation of the GATK calling algorithm. This leads to the likelihood of the observed reads occurring given the true genotypes as proportional to $\left(\rho_A^{ij}\right)^{R_A} \times \left(\rho_a^{ij}\right)^{R_a} \times \left(\rho_\varepsilon^{ij}\right)^{R_\varepsilon}$.

Moment estimators of relatedness from low-depth

In order to define our new moment-estimators for relatedness matrices, we give first a brief introduction and explanation of notations and theory. Here, the concepts of additive and non-additive components are being borrowed from the literature of quantitative genetics and in

particular the polygenic models first proposed by RA Fisher (33) where the genetic effects of each variant can be split into two orthogonal components. The first being the additive contribution, describing the effect that increases linearly with the number of minor alleles in the genotype, and the second being the non-additive contribution which describe the deviations away from the additive model caused by interactions between the two alleles and a single locus as is observed for example in recessive or dominant models.

Genotype-based GRM estimates for kinship and fraternity matrices (denoted as K and D, respectively) can be defined as follows:

$$K_{ii'} = \frac{1}{M}\sum_{j=1}^{M} X_A^{ij} \times X_A^{i'j} \quad \& \quad D_{ii'} = \frac{1}{M}\sum_{j=1}^{M} X_D^{ij} \times X_D^{i'j}$$

Where X_A^{ij} and X_D^{ij} are the classical additive and non-additive components of the individual level genotypes G^{ij} which are defined as follows:

$$X_A^{ij} = \alpha_0^j \, 1_{\{G^{ij}=0\}} + \alpha_1^j \, 1_{\{G^{ij}=1\}} + \alpha_2^j \, 1_{\{G^{ij}=2\}}$$

$$X_D^{ij} = \delta_0^j \, 1_{\{G^{ij}=0\}} + \delta_1^j \, 1_{\{G^{ij}=1\}} + \delta_2^j \, 1_{\{G^{ij}=2\}}$$

where

$$\alpha_k^j = \frac{k-2q_j}{\sqrt{2p_jq_j}}, (k=0,1,2) \text{ and } \delta_0^j = \frac{q_j}{p_j}, \delta_1^j = -1, \delta_2^j = \frac{p_j}{q_j},$$

q_j being the minor allele frequency of variant j and $p_j = 1 - q_j$. Alternative notations are presented in (34) and (35) but give the same moment-estimators. The values of $(\alpha_0^j, \alpha_1^j, \alpha_2^j)$ are obtained through standardisation of $G^{\cdot j}$, interpreted as a random variable (the SNP index j is fixed, the sample is constituted of the values G^{ij} for $i = 1, \ldots, N$) : its expected value is $2q_j$ and its standard deviation is $\sqrt{2p_jq_j}$ (assuming Hardy-Weinberg proportions). The resulting random variable $X_A^{\cdot j}$ has expected value 0 and variance 1. The values of $(\delta_0^j, \delta_1^j, \delta_2^j)$ can then

be determined by imposing three constraints on the resulting variable $X_D^{\cdot j}$: $E\left(X_D^{\cdot j}\right) = 0$, $\mathrm{var}(X_D^{\cdot j}) = 1$, and $E\left(X_A^{\cdot j}\, X_D^{\cdot j}\right) = 0$ (the two variables are independent – or 'orthogonal' – in the sample).

It is well established that under the circumstances of correct Hardy-Weinberg proportions in the population and of having in-hand the correct value of the minor allele frequencies, $K_{ii'}$ will be an unbiased estimator of $2\varphi^{ii'}$ and $D_{ii'}$ will be an unbiased estimator of $\psi^{ii'}$. Such genetic relatedness matrices were first introduced in (36,37) for kinship and in (38) for heritability and have been repurposed for many other uses.

Such moment estimators necessitate allele frequency information. For low-depth sequencing data, it is possible to estimate allele frequencies directly from genotype probabilities. This is however problematic as the additional uncertainty in the data will characteristically lead to increased estimates of allele frequencies as well as potential perturbations to Hardy-Weinberg proportions. This can be observed in Supplementary Figure 4 where we compared observed minor alleles frequencies and heterozygosity statistics from the original simulated genotypes of CilentoSim against those estimated from genotype likelihoods at a depth of 2.5×. The perturbation to allele frequencies is difficult to avoid, but the issue of potential Hardy-Weinberg deviations may be circumvented by defining our additive and non-additive components on estimated genotype frequencies (rather than allele) in order to correctly achieve orthogonality. The derivations that we give here are equivalent to those found in Vitezica et al. (35).

Across the sample, we estimate genotype probabilities by averaging across all genotype probabilities in the sample. First, individual genotype likelihood data (typically available on a log-scale) in the form GL_{AA}^{ij}, GL_{Aa}^{ij}, and GL_{aa}^{ij} are rescaled to genotype probabilities P_{AA}^{ij},

P_{Aa}^{ij}, and P_{aa}^{ij}. Then we estimate genotype frequencies in the sample as: $\bar{P}_{AA}^{j} = \frac{1}{N}\sum_{i=1}^{N} P_{AA}^{ij}$, $\bar{P}_{Aa}^{j} = \frac{1}{N}\sum_{i=1}^{N} P_{Aa}^{ij}$, and $\bar{P}_{aa}^{j} = \frac{1}{N}\sum_{i=1}^{N} P_{aa}^{ij}$.

The additive and dominant component are defined as

$$\tilde{X}_{A}^{ij} = \tilde{\alpha}_{0}^{j}\, P_{AA}^{ij} + \tilde{\alpha}_{1}^{j}\, P_{Aa}^{ij} + \tilde{\alpha}_{2}^{j}\, P_{aa}^{ij}$$

$$\tilde{X}_{D}^{ij} = \tilde{\delta}_{0}^{j}\, P_{AA}^{ij} + \tilde{\delta}_{1}^{j}\, P_{Aa}^{ij} + \tilde{\delta}_{2}^{j}\, P_{aa}^{ij}$$

As previously, the values of the triplet $(\tilde{\alpha}_{0}^{j}, \tilde{\alpha}_{1}^{j}, \tilde{\alpha}_{2}^{j})$ are obtained by standardizing the vector with $(0, 1, 2)$ using the observed mean and variance of the expected minor allele count (or genotype dosage) $\tilde{G}^{\cdot j}$ which is constituted of the values $\tilde{G}^{ij}$ for $i = 1, \ldots, N$, where $\tilde{G}^{ij} = P_{Aa}^{ij} + 2P_{aa}^{ij}$. The values of $(\tilde{\delta}_{0}^{j}, \tilde{\delta}_{1}^{j}, \tilde{\delta}_{2}^{j})$ are derived from the constraints $E(\tilde{X}_{D}^{\cdot j}) = 0$, $\mathrm{var}(\tilde{X}_{D}^{\cdot j}) = 1$, and $E(\tilde{X}_{A}^{\cdot j}\, \tilde{X}_{D}^{\cdot j}) = 0$, where, as before, expected values are computed across the sample (j is fixed and i goes from 1 to N). We obtain

$$(\tilde{\alpha}_{0}^{j}, \tilde{\alpha}_{1}^{j}, \tilde{\alpha}_{2}^{j}) = \left(\bar{P}_{Aa}^{j} + 4\bar{P}_{aa}^{j}\bar{P}_{AA}^{j} - {\bar{P}_{Aa}^{j}}^{2}\right)^{-\frac{1}{2}} \times \left(-\bar{P}_{Aa}^{j} - 2\bar{P}_{aa}^{j}, 1 - \bar{P}_{Aa}^{j} - 2\bar{P}_{aa}^{j}, 2 - \bar{P}_{Aa}^{j} - 2\bar{P}_{aa}^{j}\right)$$

and

$$(\tilde{\delta}_{0}^{j}, \tilde{\delta}_{1}^{j}, \tilde{\delta}_{2}^{j}) = \left(\bar{P}_{aa}^{j} + 4\frac{\bar{P}_{aa}^{j}\bar{P}_{AA}^{j}}{\bar{P}_{Aa}^{j}} + \bar{P}_{AA}^{j}\right)^{-\frac{1}{2}} \times \left(\sqrt{\frac{\bar{P}_{aa}^{j}}{\bar{P}_{AA}^{j}}}, -2\sqrt{\frac{\bar{P}_{aa}^{j}\bar{P}_{AA}^{j}}{{\bar{P}_{Aa}^{j}}^{2}}}, \sqrt{\frac{\bar{P}_{AA}^{j}}{\bar{P}_{aa}^{j}}}\right).$$

Finally, the GRM matrices using genotype likelihoods are computed as

$$\widetilde{K}_{ii'} = \frac{1}{M}\sum_{j=1}^{M} \tilde{X}_{A}^{ij} \times \tilde{X}_{A}^{i'j} \text{ and } \widetilde{D}_{ii'} = \frac{1}{M}\sum_{j=1}^{M} \tilde{X}_{D}^{ij} \times \tilde{X}_{D}^{i'j}.$$

Correcting the bias

Our initial simulation results indicated a clear relationship between the average depth and the biases in the estimates of both off-diagonal and diagonal elements in the GRMs. Indeed, the bias observed appeared similar to the bias that occurs when hard-called genotypes (setting the genotypes to the most probable genotype) are used for estimating GRMs as reported by Dou et al. (17). For a given average read depth, our simulation results suggest that $E\left[\widetilde{K}_{ii'}\right] = 2\beta_1\varphi_{ii'}$ and $E\left[\widetilde{D}_{ii'}\right] = \beta_2\psi_{ii'}$ for some unknown constants β_1 and β_2.

Each off-diagonal element of matrices $\widetilde{K}$ and $\widetilde{D}$ are themselves averages over many point estimated from each genetic variant. These point-wise come from genetic variants with differing read depths and hence we should expect some variants to be giving greater or lesser biased point-wise estimations. When the depth is low, the three genotype probabilities tend to become less certain, we move further away from a tuple of probabilities such as $(1,0,0)$ (which represents a certain genotype of AA) towards a tuple such as $\left(\frac{1}{2},\frac{1}{2},0\right)$ or even $\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$ where there is no certainty as to what the true genotype may be. This uncertainty or 'fuzziness' of the data can be summarised by the variance of the genotype (here thought of as a random variable taking values in {0,1,2} occurring at probabilities P_{AA}^{ij}, P_{Aa}^{ij}, and P_{aa}^{ij}, respectively. We denote this measure as $v^{ij} \coloneqq P_{Aa}^{ij}\left(1 - P_{Aa}^{ij}\right) + 4P_{aa}^{ij}\left(1 - P_{aa}^{ij}\right) - 4P_{Aa}^{ij}P_{aa}^{ij}$. To demonstrate this, we simulated repeatedly low-depth data for a single variant shared (with IBD status at random) by two siblings; varying values of the average depth and minor allele frequency for the variant. Pairs of siblings are expected to share at least one haplotype IBD for 50% of their genome and to share both haplotypes IBD for 25%. By varying the depth, we could see the change in the expected bias (Supplementary Figure 5) suggesting clearly that additional uncertainty or 'fuzziness' in the genotype likelihoods gives a stronger downward bias in a GRM moment-estimate. In Supplementary Figure 5, the average point-wise estimates

of kinship are plotted against the varying values of $v^{ii'j} \coloneqq \frac{1}{2}\left(v^{ij} + v^{i'j}\right)$. Different mean values of $v^{ii'j}$ came from simulating read data with depths varying between 2× and 25×. Here, we observed broadly linear relationships, with the slope changing slightly depending on the minor allele frequency of the variant. We can also see that as $v^{ii'j}$ tends to zero, the multiplicative bias in our estimate tends to one; and thus the estimator becomes unbiased. This suggests that if we can have a model for this relationship between bias and the fuzziness of each variant, it should be possible to gain an estimation of the unbiased value of the relatedness coefficients between i and i'. Hence we use an idea similar to simulation extrapolation (39) though rather than artificially adding more noise to our data, we simple take advantage to the different levels of noise between different SNPs and extrapolate what our relatedness estimators would be with zero noise.

Our point wise estimates for the two matrices are written as $\widetilde{K}^{j}_{ii'}$ and $\widetilde{D}^{j}_{ii'}$ and we use linear regression to perform what was found to be the most appropriate medialisation which was the following:

$$E\left[K^{j}_{ii'}\right] = (z_1 + z_2\varphi_{ii'})\left(1 + z_3\left(v^{ij} + v^{i'j}\right) + z_4 v^{ij} v^{i'j}\right)$$

$$E\left[D^{j}_{ii'}\right] = (u_1 + u_2\psi_{ii'})\left(1 + u_3\left(v^{ij} + v^{i'j}\right) + u_4 v^{ij} v^{i'j}\right)$$

Here, $\varphi_{ii'}$ and $\psi_{ii'}$ are the kinship and fraternity coefficients between i and i'. The model represents the intuition that when the fuzziness (v^{ij} and $v^{i'j}$) is null, the pointwise estimators should have expected values of the 'true' pointwise estimator from full WGS data; though we allow for the expectation to be linear in the 'true' estimator by introducing quantities z_1 and z_2 for kinship and u_1 and u_2 for fraternity. Indeed, all quantities z_{1-4} and u_{1-4} are nuisance

parameters that allow a flexible modelling of potential biases that could be created by studying low-depth data.

Using this model, regressing values of $\widetilde{K}_{ii'}^{j}$, or $\widetilde{K}_{ii'}^{j}$, across values of j against corresponding values of v^{ij} and $v^{i'j}$ leads to estimates of $\varphi_{ii'}$ and $\psi_{ii'}$ from the intercepts of the linear regression models. Our adjustment procedure circumvents the nuisance parameters by firstly performing the aforementioned regression on the diagonal elements of the matrices $\widetilde{K}$ and $\widetilde{D}$ $(i = i')$ with the knowledge that $2\varphi_{ii}$ and ψ_{ii} should be equal to 1. Then in a second step, we regress the mean unadjusted estimates ($\widetilde{K}_{ii'}$ or $\widetilde{D}_{ii'}$) against the intercepts from the aforementioned linear regression models that compared $\widetilde{K}_{ii}^{j}$, or $\widetilde{K}_{ii'}^{j}$, with v^{ij} and $v^{i'j}$. These steps combined provide new regression-based estimates of ϕ and φ.

This adjustment procedure carries a reasonable computational burden, so we simply apply it to a subset of pairs which are chosen to represent a good range of relatedness estimates ($\hat{\varphi}$ or $\hat{\psi}$) among the unadjusted estimates in the sample calculated by LowKi. The adjustment proceeds by comparing the outcome of the unadjusted estimates and the regression-based estimates (described here) in order to calculate the appropriate multiplicative biases β_1 and β_2 in order to finally adjust the initial unadjusted estimates of LowKi.

Testing existing software

To run SEEKIN (v1.01), we first applied BEAGLE (v4.1). BEAGLE was given reference haplotypes from the 1000 Genomes project (Phase 3) and was run in windows of 750 variants with buffers of 250 variants. We found that BEAGLE required very long runtimes, hence we set the parameter 'modelscale' equal to 3 which the authors of BEAGLE suggested in the software's manual as an appropriate setting to increase both speed and accuracy when applying BEAGLE

to genotype likelihood data. Otherwise, both NGSRelateV2 (v2) and SEEKIN were run with the default recommended parameters.

Testing on real data

In order to test our method on a real dataset, we were given access to 150 individuals from FranceGenRef and down-sampled their individual bam files to an average of 2.5× coverage. The FranceGenRef panel comprises 856 individuals from the population of France and combines individuals from the GAZEL cohort (www.gazel.inserm.fr/en), from the PREGO cohort (www.vacarme-project.org), and 50 blood donors from the Finistere region. The down-sampling was achieved by simply counting the number of reads in the original bam-files, and randomly sampling the appropriate proportion of these reads given that full bam files correspond to average read depth of 35×. This set of 150 individuals contains two sibling pairs who have an expected kinship of 0.25 and expected fraternity coefficient of 0.25. All other pairs are expected to have kinship and fraternity coefficients very close to zero. There may be residual population structure in the sample as individuals of FranceGenRef come from different regions in France; a country with substantial fine-scale population structure (40). Down-sampling and calling were performed with samtools (v0.1.19) (41), Sambamba (v0.7.1) (42), and GATK HaplotyeCaller (v3.7) which provides the genotype likelihoods that we supplied to LowKi as well as NGSRelateV2, and BEAGLE followed by SEEKIN.

We observed that LowKi's estimators were improved if variants with a very small observed expected minor allele frequency were removed from the calculation and such a filter has been added. Specifically, the quantity $\bar{P}_{Aa}^{j} + 2\bar{P}_{aa}^{j}$ should be in the range 0.05 to 1.95. In the example of the 150 individuals of FranceGenRef, 1,009,181 variants out of a possible 1,051,789 were used in the calculation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

LowKi is freely available at https://github.com/genostats/LowKi and implemented in R. Instructions for download and implementation as well as example datasets are also provided at this location. The package contains a small simulated dataset allowing to test the method.

Contact for applications for access to genetic data from the Cilento isolates: Marina Ciullo (marina.ciullo@igb.cnr.it).

Contacts for applications for access to simulation datasets based on the pedigree structure of the Cilento Isolates: Marina Ciullo (marina.ciullo@igb.cnr.it), Anne-Louise Leutenegger (anne-louise.leutenegger@inserm.fr) and Anthony F. Herzig (anthony.herzig@inserm.fr).

Data from the FranceGenRef panel will be submitted to the French Centralized Data Center of the France Medicine Genomic Plan that is under construction. Enquiries for the use of this data can be addressed to GENMED LABEX (http://www.genmed.fr/index.php/en/contact).

Competing interests

The authors declare that they have no competing interests.

Funding

485 This work was supported by LABEX GENMED funded as part of “Investissement d’avenir”
486 program managed by Agence Nationale pour la Recherche (grant number ANR-10-LABX-
487 0013), and by the French regional council of Pays-de-le-Loire (VaCaRMe project). This work
488 was also supported by the POPGEN project as part of the Plan Médecine Génomique 2025
489 (FMG2025/POPGEN) and by Inserm cross cutting project GOLD.

490 **Authors’ contributions**

491 AFH and HP designed the method and wrote the R package. AFH was the main writer of the
492 manuscript. MC gave access to the Cilento data. A-LL participated significantly to the
493 manuscript elaboration. All authors read and approved the final manuscript.

494 **Acknowledgments**

495 We would like to thank kindly all of the participants of the Vallo di Diano project and the
496 Cilento cohort. We acknowledge the Center of Biological Resources (CHU Nantes, Hotel Dieu,
497 CRB, Nantes, F-44093, France), the Dijon CRB, the CEPH and the Genomics and Bioinformatics
498 Core Facility of Nantes (GenoBiRD, Biogenouest). We thank the FranceGenRef Consortium for
499 giving us the opportunity to work with the FranceGenRef sequencing data for this project. We
500 are also very grateful to Emmanuelle Génin for her valuable advice and insights regarding this
501 project.

502

References

1. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nature Reviews Genetics. 2015;16(1):33–44.

2. Thompson EA. Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. Genetics. 2013;194(2):301.

3. Weir BS, Anderson AD, Hepler AB. Genetic relatedness analysis: modern data and new challenges. Nat Rev Genet. 2006;7(10):771–80.

4. Goudet J, Kay T, Weir BS. How to estimate kinship. Molecular Ecology. 2018;27(20):4121–35.

5. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121–32.

6. Gilly A, Ritchie GR, Southam L, Farmaki A-E, Tsafantakis E, Dedoussis G, et al. Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. Hum Mol Genet. 2016;25(11):2360–5.

7. Converge Consortium, Cai N, Bigdeli TB, Kretzschmar W, Li Y, Liang J, et al. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 2015;523(7562):588–91.

8. the Haplotype Reference Consortium, McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics. 2016;48(10):1279–83.

9. Hofreiter M, Paijmans JLA, Goodchild H, Speller CF, Barlow A, Fortes GG, et al. The future of ancient DNA: Technical advances and conceptual shifts. Bioessays. 2015;37(3):284–93.

10. Städele V, Vigilant L. Strategies for determining kinship in wild populations using genetic data. Ecology and Evolution. 2016;6(17):6107–20.

11. Perdry H, Dandine-Rolland C, Banddyopadhyay D, Kettner L. Gaston: Genetic data handling (QC, GRM, LD, PCA) & linear mixed models. CRAN. 2018;https://cran.r-project.org/web/packages/gaston/index.html.

12. Colonna V, Nutile T, Astore M, Guardiola O, Antoniol G, Ciullo M, et al. Campora: a young genetic isolate in South Italy. Hum Hered. 2007;64(2):123–35.

13. Colonna V, Nutile T, Ferrucci RR, Fardella G, Aversano M, Barbujani G, et al. Comparing population structure as inferred from genealogical versus genetic information. Eur J Hum Genet. 2009;17(12):1635–41.

14. Nutile T, Ruggiero D, Herzig AF, Tirozzi A, Nappo S, Sorice R, et al. Whole-Exome Sequencing in the Isolated Populations of Cilento from South Italy. Scientific Reports. 2019;9(1).

15. Herzig AF, Nutile T, Babron M-C, Ciullo M, Bellenguez C, Leutenegger A-L. Strategies for phasing and imputation in a population isolate. Genetic Epidemiology. 2018;42(2).

16. Herzig AF, Nutile T, Ruggiero D, Ciullo M, Perdry H, Leutenegger A-L. Detecting the dominance component of heritability in isolated and outbred human populations. Scientific Reports. 2018;8(1).

17. Dou J, Sun B, Sim X, Hughes JD, Reilly DF, Tai ES, et al. Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. PLOS Genetics. 2017;13(9):e1007021.

18. Hanghøj K, Moltke I, Andersen PA, Manica A, Korneliussen TS. Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. Gigascience. 2019;8(5).

19. Korneliussen TS, Moltke I. NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. Bioinformatics. 2015;31(24):4009–11.

20. Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, et al. Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics. 2011;12:231.

21. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.

22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.

23. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1-33.

24. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet. 2016;98(1):116–26.

25. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

26. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, et al. Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLOS Genetics. 2006 Mar 24;2(3):e41.

27. Lipatov M, Sanjeev K, Patro R, Veeramah KR. Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data. bioRxiv. 2015;023374.

28. Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R. Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. Genome Res. 2013;23(11):1852–61.

29. Muyas F, Bosio M, Puig A, Susak H, Domènech L, Escaramis G, et al. Allele balance bias identifies systematic genotyping errors and false disease associations. Hum Mutat. 2018/11/23 ed. 2019;40(1):115–26.

30. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843–51.

31. The UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526:82.

32. Wijsman EM, Rothstein JH, Thompson EA. Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. Am J Hum Genet. 2006;79(5):846–58.

33. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of The Royal Society of Edinburgh. 1919;52(2):399–433.

34. Zhu Z, Bakshi A, Vinkhuyzen AAE, Hemani G, Lee SH, Nolte IM, et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. Am J Hum Genet. 2015;96(3):377–85.

35. Vitezica ZG, Legarra A, Toro MA, Varona L. Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. Genetics. 2017;206(3):1297–307.

36. VanRaden PM. Genomic measures of relationship and inbreeding. Interbull Annual Meeting Proceedings. 2007;(37):33–33.

37. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.

38. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. 2010;42(7):565–9.

39. Cook JR, Stefanski LA. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. Journal of the American Statistical Association. 1994;89(428):1314–28.

40. Saint Pierre A, Giemza J, Alves I, Karakachoff M, Gaudin M, Amouyel P, et al. The genetic history of France. European Journal of Human Genetics. 2020;28(7):853–65.

41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

42. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31(12):2032–4.

Figure legends

Figure 1a

LowKi estimates for kinship and fraternity for CilentoSim. Off-diagonal elements of the estimated kinship and fraternity matrices against the true simulated IBD sharing coefficient in CilentoSim at three different simulated mean read depths (2.5×, 5×, and 10×). Lighter colours represent the unadjusted estimated from our method and the darker colours give the final recalibrated estimates. The number of variants (M) and the time (T) required for the calculation of the two matrices are overlaid on the figure.

Figure 1b-c

Corresponding estimates from SEEKIN (kinship only) and NGSRelateV2.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- LowDepthAFH20211123sup.docx