Allele frequency-free inference of close familial relationships from genotypes or low depth sequencing data

Authors:

Ryan K Waples*, Anders Albrechtsen*[,1], Ida Moltke*[,1]

*Department of Biology, Section for Computational and RNA Biology, University of Copenhagen, 2200 Copenhagen N, Denmark,

[1]Corresponding authors: Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark. E-mail: albrecht@binf.ku.dk; and ida@binf.ku.dk

Abstract

Knowledge of how individuals are related is important in many areas of research and numerous methods for inferring pairwise relatedness from genetic data have been developed. However, the majority of these methods were not developed for situations where data is limited. Specifically, most methods rely on the availability of population allele frequencies, the relative genomic position of variants, and accurate genotype data. But in studies of non-model organisms or ancient human samples, such data is not always available. Motivated by this, we present a new method for pairwise relatedness inference, which requires neither allele frequency information nor information on genomic position. Furthermore, it can be applied to both genotype data and to low-depth sequencing data where genotypes cannot be accurately called. We evaluate it using data from SNP arrays and low-depth sequencing from a range of human populations and show that it can be used to infer close familial relationships with a similar accuracy as a widely used method that relies on population allele frequencies. Additionally, we show that our method is robust to SNP ascertainment, which is important for application to a diverse range of populations and species.

Introduction

The ability to infer the familial relationship between a pair of individuals from genetic data plays a key role in several research fields. In conservation biology, it is used to design breeding programs that minimize inbreeding (Kardos *et al.* 2015), in archeology it is helpful to understand burial patterns and other cultural traditions (Baca *et al.* 2012; Sikora *et al.* 2017), and in population and disease genetics it is often used to exclude relatives, because many analysis methods within those fields assume all sampled individuals are unrelated and violations of this assumption can lead to wrong conclusions (Balding 2006).

Numerous pairwise relatedness inference methods have been developed, e.g., Thompson (1975), Lee (2003), Purcell *et al.* (2007), Albrechtsen *et al.* (2009), Manichaikul *et al.* (2010), Stevens *et al.* (2011), Korneliussen and Moltke (2015), Conomos *et al.* (2016), Dou *et al.* (2017) and many are available in popular software packages, like PLINK (Purcell *et al.* 2007), and KING (Manichaikul *et al.* 2010). Most of these methods estimate either the three relatedness coefficients k_0, k_1 and k_2, or the kinship coefficient $\theta = \frac{k_1}{4} + \frac{k_2}{2}$ for each pair of diploid individuals, where k_0, k_1 and k_2 are the proportions of the genome where a pair of individuals share 0, 1, or 2 alleles identical by descent (IBD) (Thompson 2000). By definition, alleles are IBD when they are identical due to *recent* shared ancestry, but because alleles can also be identical due to older shared ancestry or recurrent mutations, IBD status cannot be directly observed. Therefore, the pairwise relatedness coefficients and the kinship coefficient have to be estimated, which can be done from patterns of observed genetic identity (identity by state; in short IBS) together with population allele frequencies. Once estimated, these relatedness statistics, k_0, k_1, k_2, and the kinship coefficient can be used to infer familial

relationships by comparison to the expectation of the statistics under different familial relationships (Hill and Weir 2011).

Although inference of relatedness is of wide interest, most existing methods are not immediately applicable in studies with limited data or genetic resources. First, most existing methods require that the allele frequencies of the source population are known. For most studies in modern humans this is not a problem. However, in studies of ancient humans or other species, accurate estimates of allele frequencies are not always available, due to limited number of samples.

Second, several existing methods examine loci in sliding windows or use hidden Markov models to leverage the non-independence of allele sharing along the genome between relatives, (e.g., Stevens *et al.* 2011; Monroy Kuhn *et al.* 2017). These methods are powerful, but require information about the genomic position of variable sites, and for non-model organisms high-quality reference genome assemblies are often not available.

Finally, nearly all existing methods --both frequency-based and others-- require genotype data. However, in sequencing studies, samples are often only sequenced to low depth due to cost and technical issues, which makes it infeasible to call genotypes accurately (Nielsen *et al.* 2012), precluding the use of these methods. There are a few methods that estimate relatedness from low depth data by utilizing genotype likelihoods (e.g., Korneliussen and Moltke 2015), or by using imputed genotype dosages (Dou *et al.* 2017). However, these methods function by leveraging access to hundreds or thousands of samples to estimate allele frequencies or perform accurate imputation from low depth sequencing data. They are not designed to apply to datasets with a low number of samples.

Hence, most existing methods to infer close familial relationships are not immediately applicable in studies where data is limited, including many studies of non-model organisms and ancient humans. One of the few exceptions is a simple but elegant test

83 for pairwise relatedness proposed in Lee (2003), that requires neither information on
84 allele frequencies nor the relative genomic position of the variable sites. This test relies
85 entirely on the relative frequency of different genotype combinations within the two
86 individuals and thus only requires genotype data from the two target individuals. While
87 useful, this test does not provide any means to distinguish between different types of
88 close familial relationships; it only provides a statistical test for relatedness vs the null
89 hypothesis of unrelated.

90 To our knowledge, there are only two published methods that can be used to distinguish
91 between different types of familial relationships when neither allele frequencies nor
92 information about the relative genomic position of sites is obtainable. The first of these
93 two methods consists of plotting the proportion of the genomic sites in which the two
94 individuals share both alleles IBS (which we will denote IBS2) versus the proportion of
95 the genomic sites in which they share zero alleles IBS (which we will denote IBS0). This
96 method was used in Rosenberg (2006), where it was applied to the Human Genome
97 Diversity Project (HGDP) dataset. In the resulting plot of the HGDP data (figure 1 of
98 Rosenberg (2006)), pairs of individuals with the same relationship category form
99 distinct clusters in such a way that it is possible to identify parent-offspring pairs, full
100 sibling pairs, and to a lesser extent more distant relationships such as half-
101 siblings/avuncular/grandparent-grandchild and first cousins.

102 The second existing method is based in part on the KING-robust kinship estimator
103 (Manichaikul *et al.* 2010), which was developed to overcome biases in kinship estimates
104 caused by cryptic population and family structure. More specifically, the KING-robust
105 estimator was developed to be robust to admixture, i.e. violations of the assumption that
106 the genotypes of the pair of individuals are sampled from a single homogenous
107 population, but in practice it has been shown to struggle when applied to admixed
108 samples (Thornton *et al.* 2012; Moltke and Albrechtsen 2014; Conomos *et al.* 2016).

However, while the KING-robust kinship estimator was not developed with the aim of providing a kinship estimator that works when allele frequencies within a single population are unknown, it is directly applicable to this scenario as long as samples are from the same homogenous population. The reason for this is that, like the test suggested in Lee (2003), it relies only on the genotype combinations within the two target individuals and does not require knowledge about of allele frequencies. Importantly, Manichaikul *et al.* (2010) show it is possible to infer if a pair of individuals are parent-offspring, full sibling, half siblings/avuncular/grandparent-grandchild, cousins, or unrelated by jointly considering the KING-robust estimator and the fraction of sites IBS0 with both these statistics estimated from SNP array data without using any information about allele frequencies. For example, see figure 3A in Manichaikul *et al.* (2010), a scatterplot of KING-robust kinship vs the fraction of sites IBS0.

However, both the existing methods discussed above have two key limitations. First, like most other pairwise relatedness estimation methods, they were developed for accurately called genotypes. For example, the KING software (Manichaikul *et al.* 2010) implementing the KING-robust kinship estimator requires accurate genotype data as input, which can be problematic for studies where only moderate or low-depth sequencing data is available. Second, both methods rely on estimates of the fraction of the genomic sites in which the pair of individuals share zero alleles identical by state, IBS0. This can be problematic because this fraction is highly sensitive to the SNP ascertainment procedure. This means that the results of the methods are platform-dependent and are likely to differ between different SNP arrays and especially between SNP array and sequencing datasets. And in turn, this among other things, means that it can be difficult to distinguish between full siblings and parent-offspring pairs using KING.

134 Motivated by the outlined limitations to the existing methods, we present a method for
135 relationship inference that - unlike most existing methods - relies neither on allele
136 frequencies nor on information about the relative position of the variant sites, and
137 which – unlike other frequency-free methods – 1) is applicable even to sequencing data
138 of so low depth that accurate genotypes cannot be called from it and 2) is robust to
139 ascertainment bias.

140 The method is inspired by previous methods; it uses the KING-robust kinship estimator
141 and a statistic R0, which is similar to the test statistic from the test for relatedness
142 suggested by Lee (2003). However, the method is new in two important way. First,
143 besides relying on the two statistics, R0 and KING-robust kinship, it also relies on a
144 third new statistic, R1. More specifically, the method consists of using two combinations
145 of these three statistics, R1-R0 and R1-KING-robust kinship, to infer relationships, and
146 it is this combination of statistics that makes the method robust to ascertainment bias.
147 In particular, the inclusion of R1 helps distinguish between full siblings and parent
148 offspring. Secondly, while the new method can straightforwardly be applied to accurate
149 genotype data, like other similar methods, we also present two ways to estimate the
150 three statistics directly from sequencing data that takes the uncertainty of genotypes
151 into account. This is what makes the method applicable to low-depth sequencing data as
152 well.

153 In the following we first fully describe the three statistics, R0, R1 and KING-robust
154 kinship, how they can be estimated and other methodological details. Next, using
155 publicly available SNP array data, we show that the new method provides similar
156 accuracy to the commonly used frequency-based method implemented in PLINK,
157 without utilizing any information on allele frequencies or genomic positions. Then using
158 sequence data from the 1000 Genomes, we show that the three statistics can be
159 estimated directly from sequencing data of low depth (~4x), here defined as depth

insufficient for genotype calling. Moreover, we show that the resulting estimates are useful and that this is not the case for estimates obtained from genotypes called from the same data. Next, using different subsets of the same data, we show that this new method, unlike previous similar methods, is robust to SNP ascertainment. Finally, we discuss some potential applications and limitations of the new method.

Methods and materials

The R0, R1, and KING-kinship statistics

The method for relationship inference we propose consists of first estimating three statistics called R0, R1 and KING-robust kinship from genetic data, then constructing two plots: R1 vs R0 and R1 vs KING-robust kinship, and finally interpreting these plots.

To define the three statistics, R0, R1 and the KING-robust kinship, we note that for a pair of diploid individuals, all diallelic sites fall into one of nine genotype combinations; all possible combinations of the two individuals carrying 0, 1, or 2 copies of a specific allele, e.g. the ancestral allele. For a pair of individuals, we can therefore fully characterize their genome-wide IBS-sharing pattern in nine counts or proportions denoted: A, B, C, D, E, F, G, H, and I (Figure 1A), similar to a two-dimensional site-frequency spectrum (SFS) across the two individuals. This is important to note because R0 and R1 statistics are defined as simple functions of a subset of these nine counts as shown in Figure 1B and C, and the KING-robust kinship statistic, originally defined by Manichaikul *et al.* (2010), can also be re-formulated as a function of these 9 values (Figure 1D).

The new method is motivated by several observations. First, the expected values of A-I vary depending on the familial relationship. Consequently, so do functions of these

183 counts, including R0, R1 and KING-robust kinship. Specifically, for two non-inbred
184 individuals from the same homogenous population the pairs (R1, R0) and (R1, KING-
185 robust kinship), both – just like (k_0, k_1, k_2) - have the property that the joint expectation
186 ranges for the five relationship categories parent-offspring (PO), full siblings (FS), half
187 siblings/avuncular/grandparent-grandchild (HS), first cousins (C1) and unrelated (UR)
188 do not overlap (Figure 2, for derivations see supplementary text) regardless of the
189 underlying allele frequency spectrum. Second, this method requires neither population
190 allele frequencies nor information about the relative position of the genomics sites
191 analyzed. The only information that is required to estimate the three statistics is genetic
192 data from the two individuals of interest. Third, we expect these statistics to be robust to
193 SNP ascertainment because they are ratios computed from sites that are variable within
194 the two samples and should thus be unaffected by the number of non-variable sites and
195 because the (unknown) underlying frequency spectrum should only have a limited effect
196 on these ratios. Finally, while the counts of the nine genotype combinations, A-I, and
197 thus R0, R1, and the KING-robust kinship estimator can be calculated from genotype
198 data, they can also be estimated directly from next-generation sequencing (NGS) data
199 based on the expected number of sites with each genotype combination. This makes the
200 approach appropriate even when the sequencing depth is too low for accurate genotype
201 calling (see below for methodological details).

202 Estimation from sequencing data

203 The nine genotype combinations, A-I, and thus R0, R1, and the KING-robust kinship for
204 a pair of individual, can be estimated directly from NGS data via the use of genotype
205 likelihoods calculated from the aligned sequencing reads. By using genotype likelihoods
206 it is possible to account for the genotype uncertainly inherent to low-depth NGS data.

207 We used two distinct, but related approaches to estimate these statistics from
208 sequencing data that both build on this idea.

209 The first approach, which we denote the IBS-based approach, considers all ten possible
210 genotypes at a diallelic site for each of the two individuals of interest and consists of a
211 maximum-likelihood (ML) estimation of the counts of each of the 100 (10 × 10) possible
212 genotype pairs (for details see supplementary text). To perform the ML estimation we
213 used an EM algorithm, which we have added to the `ANGSD` software package under the
214 name `IBS`. After obtaining the estimate of the counts of all the 100 possible genotype
215 pairs, we convert them into estimates of the genotype combinations A-I, by summing
216 over the counts that correspond to each combination. For example, the genotype pairs:
217 AA/AA, CC/CC, GG/GG and TT/TT all contribute to cells A or I of Figure 1. Counts
218 corresponding to genotype pairs with more than two different alleles (e.g. AC/AG) were
219 discarded. The advantage of this IBS-based approach is that it does not require
220 specification of a known allele at each site, and can thus be applied to nearly any
221 sequencing data set, even with low depth, without any prior information on the alleles at
222 each site.

223 The second approach, which we denote the SFS approach, consists of performing ML
224 estimation of the two-dimensional site-frequency spectrum (SFS) (Nielsen *et al.* 2012).
225 To find the ML estimate of the SFS, we used an expectation-maximization (EM) method
226 implemented in the `ANGSD` software package under the name `realSFS` (Korneliussen *et*
227 *al.* 2014). The SFS method requires one allele to be specified for each site e.g., the
228 ancestral, consensus or reference allele. The model underlying this method assumes that
229 all genotypes for each site have the possibility of containing this specified allele and up
230 to one other unspecified allele. To specify the alleles that do exist at each site, we used
231 the consensus sequences from the highest depth individual (NA19042) and restricted
232 our analysis to sites where the depth in this individual was at least three.

233 For both these genotype likelihood-based methods, we analyzed each chromosome
234 separately due to computational constraints. We used the original `GATK` genotype
235 likelihood model (Mckenna *et al.* 2010) with independent errors, as implemented in
236 `ANGSD`, to calculate the likelihood of sequencing data for each possible genotype. We also
237 tested the `samtools` genotype likelihood model (Li 2011) with a more complicated error
238 structure, but found that it produced worse results in this case (data not shown). Both
239 `IBS` and `realSFS` produces the expected counts for each value A-I and the values for
240 each chromosome were summed over to produce a genome-wide estimate. For example
241 command lines for each analysis, see supplemental text.

242 Confidence intervals

243 All of the above estimation methods treat each site as independent. This assumption
244 should not affect our expectation of each statistic (Wiuf 2006), but statistical non-
245 independence (e.g. linkage disequilibrium and IBD) does affect standard estimates of
246 uncertainty. To quantify the uncertainty, we therefore estimated confidence intervals for
247 all statistics using a block-jackknife procedure. Confidence intervals were calculated by
248 leaving each chromosome out (chromosome jackknife), which takes both the IBD and
249 the LD correlation into account. The weighted block jackknife variance estimator of
250 Busing *et al.* (1999) was used to estimate the variance from the distribution of estimates
251 for each statistic. The square root of this variance was interpreted as the standard error
252 in our estimate.

253 Application to real datasets

254 To assess the utility of the new method, we applied it to two different publicly available
255 datasets: SNP array data from seven HGDP populations (Rosenberg 2006), and

256 sequencing data from five related individuals from the Luhya in Webuye, Kenya (LWK)
257 population of the 1000 Genomes Project phase 3 (The Genomes Project 2015).

258 HGDP SNP array

259 The HGDP SNP array dataset was downloaded from ftp.cephb.fr/hgdp_supp1 (accessed
260 Jan. 13, 2017) and followed the quality control steps described in Rosenberg (2006) to
261 exclude mislabeled and duplicate samples. We selected seven populations from the
262 HGDP based on the presence of several close familial relationships; five non-African
263 populations: Surui, Pima, Karitiana, Maya, and Melanesian, and two African
264 populations: Mbuti Pygmies and Biaka Pygmies.

265 To ensure a fair comparison to the identity by descent inference method in PLINK, and
266 by proxy to other commonly used methods, we constructed data sets where these
267 methods have been shown to perform well. Specifically, we excluded individuals
268 showing obvious signs of admixture or inbreeding from the selected HGDP populations
269 as well as genotypes that may be inaccurate. Specifically, we ran `ADMIXTURE` (Alexander
270 *et al.* 2009) separately for each of the seven target population also including French,
271 and Han samples (as well as Yoruban samples the non-African target populations) and
272 excluded 16 individuals with >5% contribution from more than one ancestry
273 component. Then we estimated the inbreeding coefficient, F, for each of the remaining
274 individuals using PLINK and excluded two individuals with $F > 0.0625$. This left us with
275 a total of 142 individuals from the seven populations: Surui N=20, Pima = 20, Karitiana
276 N=21, Maya N=16, Melanesian N=19, Biaka Pygmies N=31 and Mbuti Pygmies N=15.
277 Finally, for each of the seven population samples we constructed a final set of accurate
278 genotypes by retaining genotypes from autosomal loci with genotyping rate (>99%),
279 MAF (>5%), HWE ($p > 10^{-4}$, with midp adjustment).

280 The R0, R1, and KING-robust kinship statistics for each of the 2902 within-population
281 pairs of individuals were then calculated from all sites where both individuals had non-
282 missing genotypes.

283 1000 Genomes sequencing data

284 To get sequencing data from several different relationship categories, we selected five
285 individuals from two families in the Phase 3 1000 Genomes (1000G) Luhya in Webuye,
286 Kenya (LWK) population: NA19027, NA19042, NA19313, NA19331, NA19334. Across
287 the five individuals, there is one pair of half-sibs (NA19027 & NA19042), and a separate
288 trio of related individuals with a pair of full siblings (NA19331 & NA19334), one parent-
289 offspring relationship (NA19313 & NA19331), and another unspecified second-degree
290 relationship (NA19313 & NA19334), possibly avuncular (The Genomes Project 2015).
291 These stated relationships leave six unrelated pairs among the five individuals.

292 For each pair of the five LWK individuals we estimated the R0, R1 and KING-robust
293 kinship statistics in five different ways: 1&2) by applying the two different sequencing-
294 based approaches described above to the 1000G aligned sequence files (~4x bam files),
295 3) by simple genotype counting based on the phased and high-quality curated genotypes
296 provided in the hg37 1000G VCF files, 4) genotype counting based on the subset of sites
297 in 3 that overlap with the Illumina 650Y sites for the HGDP data (to investigate
298 ascertainment, see below) and 5) by re-calling genotypes from the same 1000G bam files
299 and then simple counting from the called genotypes. In set 5 genotypes were called in a
300 basic manner meant to mimic data from a species with a reference genome but few
301 other genetic resources. Specifically, we used `samtools mpileup` (v1.3.1) to summarize
302 the reads overlapping each position, and `bcftools call` (v1.3.1) to assign the most
303 likely genotype at each position. We used mostly default settings; non-default flags to
304 `samtools` specified skipping indel positions. Non-default flags to `bcftools` specified

305 using the consensus caller and requesting only variant sites. For all sequence-based
306 analyses we only considered reads with a minimum phred-scaled quality score of 30 and
307 bases with minimum phred-scaled quality score of 20 and we restricted our analyses to
308 regions with a GEM 75mer mappability of 1 (Derrien *et al.* 2012). All of these methods
309 and filters can be applied to any reference-based sequencing study.

310 Assessing the effect of SNP ascertainment

311 To evaluate the effect of SNP ascertainment, we created a subset of the high-quality
312 genotype data from the five 1000 Genomes individuals. We selected the sites that
313 overlap with the Illumina 650Y array that was used for the HGDP and estimated our
314 three relatedness statistics. We compared the results from this subset of HGDP sites to
315 results for the full genotype dataset and also to the sequence-based analyses. For an
316 additional comparison, we also performed the same comparison for the methods
317 presented in Rosenberg (2006) and Manichaikul *et al.* (2010) by constructing
318 scatterplots by their methods.

319 Comparison to other methods

320 To get a categorization of relationships for the HGDP genotype dataset described above
321 based on a standard, commonly used allele frequency-based method, we first applied
322 the allele frequency based relatedness estimation program `PLINK` (v1.9) (Chang *et al.*
323 2015) to the individuals from each population separately to estimate the genome-wide
324 IBD fractions k_0, k_1, k_2, (denoted Z0, Z1 and Z2 in PLINK (Purcell *et al.* 2007)). Next, we
325 applied the relationship criteria proposed in Table 1 of Manichaikul *et al.* (2010) to the
326 relatedness coefficient estimates: the estimated k values were combined into an estimate
327 of the kinship coefficient $\theta = \frac{k_1}{4} + \frac{k_2}{2}$, and a relationship degree was assigned to each

pair of individuals based on comparing the estimated kinship coefficient to the criteria in the table. Parent-offspring and full siblings were differentiated based on k_2 values. This provided us with a categorization into five categories: parent offspring (PO), full siblings (FS), half-siblings / avuncular / grandparent-grandchild (HS), first cousin (C1), and unrelated (UR). To achieve additional resolution, we further divided the last category (UR) into two: distantly related (DR) and unrelated (UR). We did this by simply extending the logic behind the criteria proposed above. Specifically, we set the kinship threshold between DR and UR to $1/2^{13/2}$, which corresponds to including 2nd to 5th degree relatives in the DR category.

To estimate the statistics for identifying related individuals proposed by Rosenberg (Rosenberg 2006) and KING (Manichaikul *et al.* 2010) we note that they can all be calculated directly from the same nine proportions, A-I, as R1, R2 and King-robust kinship:

KING-kingship = (E – 2(C+G))/ (B+D+H+F+2E)

Fraction IBS0 = (C+G)/(A+B+C+D+E+F+G+H+I)

Fraction IBS2 = (A+E+I)/(A+B+C+D+E+F+G+H+I)

We used these formulas, because they allow for estimation of the statistics directly from sequence data in the same manner as for R0, R1 and KING-robust kinship. This allowed us to use sequence data to compare all the methods on a level playing field.

Results

To assess to how the new method works in practice, we applied it to real data from different data generation platforms and compared the results to those of other methods.

Application to SNP array data

We first applied the method to SNP array data from the Human Genome Diversity Project (HGDP) to see how the method works on data, on which standard, commonly used allele frequency-based methods like PLINK are known to perform well. More specifically, we applied it to accurately called genotypes from unadmixed and uninbred samples from seven populations originating from the HGDP. This resulted in the R1-R0 and R1-KING-robust kinship plots shown in Figure 3A and B (for population-specific plots see Figures S1 and S2). Because the true relationships for the pairs of individuals are not known, we colored each point in Figure 3A and B not according to the truth, but according to the relationship category inferred using a standard, commonly used allele frequency-based method (Figure 3C, Figure S3, for further details see Methods). Since there are least 15 individuals in each of the seven selected HGDP populations, the allele frequency estimates for these populations should be reasonably accurate even with some relatedness. Thus, although this classification and coloring scheme does not necessarily reveal the true relationships for all pairs (Hill and Weir 2011) it should provide a fair picture of how well our method performs in comparison to a standard, commonly used allele frequency-based method.

In general, points representing pairs of individuals that according to the allele frequency-based method have the same relationship, clearly cluster together on both the R1-R0 plot and the R1-KING-robust kinship plot (Figure 3) and near their expected ranges. More specifically, almost all pairs identified as parent-offspring (PO) by the frequency-based method are easy to identify as such in both the R1-R0 plot and the R1-KING-robust kinship plot, which is not the case when only a single statistic is used (see also Figures S1 and S2). The same is true for full siblings (FS). Points classified as half-siblings/avuncular/grandparent-grandchild (HS) or first cousins (C1) by the frequency-based method have a minimal overlap with each other and with less-related pairs (see

376 highlighted areas of Figure 3A and B). But notably the few pairs of individuals that are
377 difficult to classify are the same pairs as those that are edge cases for the allele
378 frequency-based method. This is apparent in an R1-R0 plot of the HGDP data
379 constructed excluding pairs that are closer than 0.01 to the kinship coefficient
380 thresholds that the frequency based method used when categorizing relationships
381 (Figure S4).

382 Application to sequencing data

383 To assess how well the new method works on sequencing data, we applied it to
384 sequencing data from five low-depth (~4x) human genomes from the 1000 Genomes
385 project. Among the five selected samples there is a parent-offspring pair, a pair of full
386 siblings, a pair of half siblings, an unspecified 2nd degree relationship (e.g. avuncular),
387 and the rest are unrelated. We estimated the R0, R1 and KING-robust kinship for each
388 pair in several ways. First, by using an IBS-based approach that estimates the
389 proportion all pairwise combinations of the 10 possible genotypes (Figure 4, "IBS").
390 Second, we estimated the two-dimensional site frequency spectrum (2D-SFS) of each
391 pair with a bi-allelic model and calculated R0, KING-robust kinship, and R1 based on
392 this spectrum (Figure 4, "realSFS"). Both these approaches base their estimates on
393 genotype likelihoods calculated from the sequencing read data, instead of on called
394 genotypes, and take the uncertainty of the underlying genotypes that is inherent to low-
395 depth sequencing data into account. The key difference between them is that the SFS-
396 based approach requires specification of an allele at each site, potentially excluding
397 studies with low-depth data from non-model species, whereas the IBS approach has no
398 such requirement (for methodological details see Methods and materials). The methods
399 also differ in how they deal with sites with more than two unique alleles. We include
400 both methods here to compare their performance on real data.

401 Additionally, we calculated the three statistics from the high-quality phased genotypes
402 for the same five individuals available from the 1000 Genomes Project Phase 3 (Figure
403 4, "1000G sites") to see how well the two genotype-likelihood-based approaches applied
404 to low-depth sequencing data perform compared to direct calculations from high-quality
405 genotype data for the same samples. In this comparison, results obtained by using the
406 genotype likelihood-based approaches applied to low-depth sequencing data are close to
407 those obtained from the high-quality genotypes for all of the pairs (Figure 4).
408 Chromosome block-jackknife estimates of uncertainty were generally small, and varied
409 by relationship type, with the pair of full siblings having the most uncertainty in R0, R1,
410 and KING-robust kinship. Notably, the IBS-based approach provided meaningful
411 results, even though it did not make use of any external allele information. With the IBS
412 method, unrelated individuals had a slight decrease in R0 and a slight increase in R1
413 and kinship, relative to the SFS method, making them appear slightly more related than
414 expected for unrelated individuals from a homogenous population. However, despite
415 this bias, it is still possible to correctly categorize of all the pairs to their relationship
416 category within the set of close familial relationships [PO, FS, HS, or UR], suggesting
417 that this method can be used when not enough information is available for the realSFS
418 method.

419 Finally, we also made R1-R0 plot and the R1-KING-robust kinship plots based on
420 genotypes that we obtained through a standard genotype calling procedure from the raw
421 read data. We did this to investigate if the genotype likelihood-based approaches are
422 necessary or one could just as well use genotypes called from the ~4x data. As expected,
423 genotype calling had a large negative effect on the outcome; in the resulting R1-R0 and
424 R1-KING-robust kinship plots the half siblings appear where first cousins are expected
425 and both the parent-offspring and full sibling pairs appear in the expected range of half
426 siblings (Figure 4, "called genotypes"). These results demonstrate the potential pitfalls

427 of basing relationship inferences, including R1-R0 and R1-KING-robust kinship plots,
428 on genotypes called from low-depth data. Notably, this is also the case for the methods
429 presented in (Rosenberg 2006) and (Manichaikul *et al.* 2010) (Figure 5A). Hence, for
430 ~4x sequencing data, calling genotypes without external information, such as an
431 imputation reference panel is not a good alternative to genotype likelihood-based
432 approaches, which means that software packages like KING cannot be used for this type
433 of data (as it requires input genotypes).

434 Assessing the effect of SNP ascertainment scheme

435 To assess the effect of SNP ascertainment on the new method, and thus the effect of the
436 data-generation platform or SNP filtering, we applied it to three different subsets of data
437 from the five 1000 Genomes individuals. The results for each of the three ascertainment
438 schemes (whole genomes, 1000 Genomes release sites, Illumina 650Y SNP array sites)
439 are similar (left four panels, Figure 4), showing SNP ascertainment does not have a large
440 effect.

441 For comparison, we performed the same assessment of the methods presented in
442 Rosenberg (Rosenberg 2006) and (Manichaikul *et al.* 2010) by constructing scatterplots
443 of the same type as those shown in their papers (Figure 5). This shows that both these
444 other methods are much more affected by ascertainment than the method proposed
445 here. In particular, the Rosenberg method is affected on both its x-axis (IBS0) and its y-
446 axis (IBS2), which means that the expected region of the plot for each relationship will
447 be different for different data sets (Figure 5A). The method presented in Manichaikul *et*
448 *al.* 2010 is affected by the SNP ascertainment mainly on its x-axis (IBS0, Figure 5B).
449 Therefore, the ascertainment mainly affects the ability to distinguish between parent-
450 offspring and full siblings, since the y-axis, which is only slightly affected by
451 ascertainment, is the kinship coefficient, which can be used to distinguish between most

452 close relationships except for parent-offspring and full siblings. The x-axis, IBS0, is
453 included in part to help make the distinction between PO and FS (Manichaikul *et al.*
454 2010), but this ability is clearly affected by SNP ascertainment (Figure 5B).

455 Discussion

456 We have presented a simple new method for inferring if and how two individuals are
457 related based solely on genetic data from the two target individuals. This new method
458 has several advantages compared to most existing relatedness inference methods, which
459 are almost all dependent on additional information. First, it does not require any
460 information about allele frequencies. Second, it does not require information about the
461 relative genomic positions of the sites included in the analysis. Third, although the
462 method can straightforwardly be applied to genotype data (where it boils down to
463 simple counting, addition, and division), we demonstrate two ways in which it can also
464 be applied directly to sequencing data via genotype likelihoods. And importantly, we
465 show that by using these, the method provides useful results when applied to 4x
466 sequencing data. In contrast, we also show that ours and similar methods like KING fail
467 to provide useful results if applied to genotypes called from the same ~4x data instead.
468 This implies that – unlike previous methods - this new method can be used even if all
469 you have is low-depth sequencing data from a few individuals from a species without a
470 reference genome. This is not likely to be the typical situation in studies of modern
471 humans, however it is much more likely to occur in studies of with non-model
472 organisms or ancient samples. Fourth and finally, we have shown that the new method
473 is robust to the SNP ascertainment present in common SNP array data sets.

Comparison to similar methods

The new method is based on plotting two statistics, R0 and KING-robust kinship against a third, new statistic R1. The R0 statistic is derived from the test statistic proposed by Lee (2003) to test for relatedness; the only differences are that the numerator and denominator are flipped and that E, the proportion of sites where both individuals are heterozygous, is included in the denominator in the statistic defined by Lee but absent in R0, making R0 similar to the PPC statistic in PLINK. The KING-robust kinship estimator was proposed by Manichaikul *et al.* (2010), here we extend it to estimation directly from sequencing data. Notably, our results suggest that this extension is important for low-depth sequencing data, because estimates based on genotypes called from low-depth sequencing data are very poor (rightmost panels of Figure 4 and 5). And importantly this extension can also be used for similar statistics and thus makes existing methods based on such statistics, like KING-robust kinship, more widely applicable.

However, this extension is not the only contribution of this study. Another key new contribution of this study is the fact that R0 and KING-robust kinship are combined with the new R1 statistic instead of with IBS0 (the proportion of sites where the two individuals share zero alleles IBS) as was previously proposed by Rosenberg (2006) and Manichaikul *et al.* (2010). As we have shown, the fraction of sites that are IBS0 or IBS2 is very sensitive SNP ascertainment, meaning that results are only comparable within each different ascertainment scheme. Across different ascertainment schemes, such as between SNP arrays and sequencing data, the method from Rosenberg (2006), where IBS0 is combined with IBS2, is difficult, if not impossible, to use for relatedness inference because the fraction of sites that are IBS2 and IBS0 varies so wildly. On the other hand, KING-robust kinship is still very useful, but it loses its ability to distinguish between parent-offspring and full siblings, as IBS0 was used for this. Due to this sensitivity to ascertainment samples cannot be analyzed in isolation, and must be placed

in the context of other samples with known relationships and the same ascertainment scheme. This requirement makes it difficult to apply these previous methods to ancient humans or other species with limited sample sizes.

In contrast, the ability to identify relatives based on expected values is maintained in the new method, regardless of ascertainment scheme due to the use of R1, instead of IBS0, which makes the new method robust to SNP ascertainment. Here the R1-R0 plot is especially useful, because parent-offspring pairs, tend to have an R0 estimate extremely close to 0, making them particularly easy to identify. The R1-KING-robust kinship, on the other hand, has the appealing aspect that the kinship axis has a biological interpretation, defined as the probability that two alleles sampled at random from two individuals are identical by descent. Hence the two plots types, R1-R0 and R1-KING-robust kinship, each have their advantages.

Additionally, it is worth noticing that R1 not only makes it easier to distinguish between parent-offspring and full siblings it can provide additional information across all relationship categories; both the new plot types, R1-R0 and R1-KING-robust kinship provide better separation of several of the different relationship categories than KING-robust kinship does alone (Figure S2).

Finally, it is worth noticing that the two plots types seem to work better than a range of other plots constructed from similar ratio statistics that we explored (Figure S5).

Limitations and applications

While the new method provides significant advantages over previous methods in situations with limited data, it does have some limitations. First, like most other relatedness inference methods, like PLINK, the proposed method assumes that the individuals are not inbred and that they originate from the same homogenous,

524 population. And like many other relationship inference methods, it is not necessarily
525 robust to violations of these assumptions. Previous studies have shown the effect of
526 admixture on relatedness inference is complex and can potentially lead to bias in either
527 direction depending on the circumstances, and this is true even for KING, which was
528 developed to be robust to admixture (Thornton *et al.* 2012; Moltke and Albrechtsen
529 2014; Conomos *et al.* 2016; Ramstetter *et al.* 2017). Specific methods have been
530 developed to correct for admixture when the allele frequencies in the admixing
531 populations are known (e.g. Thornton *et al.* 2012; Moltke and Albrechtsen 2014), or
532 enough samples are available (Conomos *et al.* 2016). But since these methods work by
533 exploiting knowledge about allele frequencies or access to many samples for their
534 correction; the pairwise R0, R1 and KING-robust kinship statistics cannot be corrected
535 in a similar manner. However, we note that Lee (2003) showed that the statistic he
536 proposed for testing for relatedness can also be used to detect if two unrelated samples
537 are not from the same homogeneous population. If this is the case, the Lee's statistic will
538 be significantly smaller than 2/3; and equivalently R0 will be significantly above 0.5,
539 which may be useful when interpreting R0, R1, KING-kinship plots in the presence of
540 admixture or population structure more generally. Regarding inbreeding, one potential
541 way to assess if one of the individuals is inbred is to compare heterozygosities across
542 individuals; non-inbred and non-admixed individuals from the same population should
543 have similar heterozygosity, so marked heterozygosity differences can be a warning
544 signal.

545 A second limitation is that the categorization of different relationship pairs is based on
546 visual inspection of plots and that in some cases such classification will be difficult
547 because the actual ranges of the different relationship in the plots are close to each other
548 and even overlap. However, this is to a large extent due to biological variation caused by
549 the randomness of recombination (Hill and Weir 2011). This limitation in classification

accuracy is therefore shared with the allele frequency-based inference methods, like the commonly used PLINK, as illustrated by Figure S5.

Despite these limitations, we believe that the results presented here suggest that the new method constitutes a helpful new tool for relatedness inference that is especially advantageous compared to other methods in studies with limited data. Identifying related individuals is a crucial step in nearly any genetic analysis and removing the requirements to specify allele frequencies and to have access to accurate genotypes has the potential allow the identification of relatives even in small studies of non-model species or ancient samples. These types of studies do not currently have many good options to address relatedness among their samples. An additional appeal of the IBS-based approach introduced here, is that because it considers all possible pairwise genotype categories, it can potentially help identify specific error types that are expected to occur in DNA from ancient samples, by allowing researchers to specifically identify transitions. Also, due to its non-reliance on allele frequencies the method is also potentially useful when relatedness is pervasive among many samples, complicating allele frequency estimation in the broader population (Mcpeek *et al.* 2004). Hence, the method will hopefully be useful to many.

Data Accessibility

The IBS method is available at:
http://www.popgen.dk/angsd/index.php/Genotype_Distribution

The data sets used are publicly available.

References

Albrechtsen, A., T. Sand Korneliussen, I. Moltke, T. van Overseem Hansen, F. C. Nielsen *et al.*, 2009 Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet Epidemiol 33**:** 266-274.

Alexander, D. H., J. Novembre and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19**:** 1655-1664.

Baca, M., K. Doan, M. Sobczyk, A. Stankovic and P. Weglenski, 2012 Ancient DNA reveals kinship burial patterns of a pre-Columbian Andean community. BMC Genet 13**:** 30.

Balding, D. J., 2006 A tutorial on statistical methods for population association studies. Nat Rev Genet 7**:** 781-791.

Busing, F. M. T. A., E. Meijer and R. Van Der Leeden, 1999 Delete-m jackknife for unequal m. Statistics and Computing 9**:** 3-8.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4**:** 7.

Conomos, M. P., A. P. Reiner, B. S. Weir and T. A. Thornton, 2016 Model-free Estimation of Recent Genetic Relatedness. Am J Hum Genet 98**:** 127-148.

Derrien, T., J. Estelle, S. Marco Sola, D. G. Knowles, E. Raineri *et al.*, 2012 Fast computation and applications of genome mappability. PLoS One 7**:** e30377.

Dou, J., B. Sun, X. Sim, J. D. Hughes, D. F. Reilly *et al.*, 2017 Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. PLoS Genet 13**:** e1007021.

Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res (Camb) 93**:** 47-64.

Kardos, M., G. Luikart and F. W. Allendorf, 2015 Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. Heredity (Edinb) 115**:** 63-72.

Korneliussen, T. S., A. Albrechtsen and R. Nielsen, 2014 ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15**:** 356.

Korneliussen, T. S., and I. Moltke, 2015 NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. Bioinformatics 31**:** 4009-4011.

Lee, W. C., 2003 Testing the genetic relation between two individuals using a panel of frequency-unknown single nucleotide polymorphisms. Annals of Human Genetics 67**:** 618-619.

609 Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association
610 mapping and population genetical parameter estimation from sequencing data.
611 Bioinformatics 27: 2987-2993.

612 Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale *et al.*, 2010 Robust
613 relationship inference in genome-wide association studies. Bioinformatics 26:
614 2867-2873.

615 McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome
616 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
617 sequencing data. Genome Res 20: 1297-1303.

618 McPeek, M. S., X. Wu and C. Ober, 2004 Best linear unbiased allele-frequency
619 estimation in complex pedigrees. Biometrics 60: 359-367.

620 Moltke, I., and A. Albrechtsen, 2014 RelateAdmix: a software tool for estimating
621 relatedness between admixed individuals. Bioinformatics 30: 1027-1028.

622 Monroy Kuhn, J. M., M. Jakobsson and T. Günther, 2017 Estimating genetic kin
623 relationships in prehistoric populations. bioRxiv.

624 Nielsen, R., T. Korneliussen, A. Albrechtsen, Y. Li and J. Wang, 2012 SNP calling,
625 genotype calling, and sample allele frequency estimation from New-Generation
626 Sequencing data. PLoS One 7: e37558.

627 Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira *et al.*, 2007 PLINK: a
628 tool set for whole-genome association and population-based linkage analyses. Am
629 J Hum Genet 81: 559-575.

630 Ramstetter, M. D., T. D. Dyer, D. M. Lehman, J. E. Curran, R. Duggirala *et al.*, 2017
631 Benchmarking Relatedness Inference Methods with Genome-Wide Data from
632 Thousands of Relatives. Genetics 207: 75-82.

633 Rosenberg, N. A., 2006 Standardized subsets of the HGDP-CEPH Human Genome
634 Diversity Cell Line Panel, accounting for atypical and duplicated samples and
635 pairs of close relatives. Ann Hum Genet 70: 841-847.

636 Sikora, M., A. Seguin-Orlando, V. C. Sousa, A. Albrechtsen, T. Korneliussen *et al.*, 2017
637 Ancient genomes show social and reproductive behavior of early Upper
638 Paleolithic foragers. Science 358: 659-662.

639 Stevens, E. L., G. Heckenberg, E. D. Roberson, J. D. Baugher, T. J. Downey *et al.*, 2011
640 Inference of relationships in population data using identity-by-descent and
641 identity-by-state. PLoS Genet 7: e1002287.

642 The Genomes Project, C., 2015 A global reference for human genetic variation. Nature
643 526: 68-74.

644 Thompson, E. A., 1975 The estimation of pairwise relationships. Ann Hum Genet 39:
645 173-188.

646 Thompson, E. A., 2000 Statistical inference from genetic data on pedigrees, pp. i-169 in
647 *NSF-CBMS regional conference series in probability and statistics*. JSTOR.

Thornton, T., H. Tang, T. J. Hoffmann, H. M. Ochs-Balcom, B. J. Caan *et al.*, 2012 Estimating kinship in admixed populations. Am J Hum Genet 91**:** 122-138.

Wiuf, C., 2006 Consistency of estimators of population scaled parameters using composite likelihood. Journal of Mathematical Biology 53**:** 821-841.

Author Contributions

AA, IM, and RKW designed the research. RKW performed research, with input from IM and AA. IM and RKM wrote the paper, with help from AA.

Figures for:

Allele frequency-free inference of close familial relationships from genotypes or low depth sequencing data

Authors:

Ryan K. Waples
Anders Albrechtsen
Ida Moltke

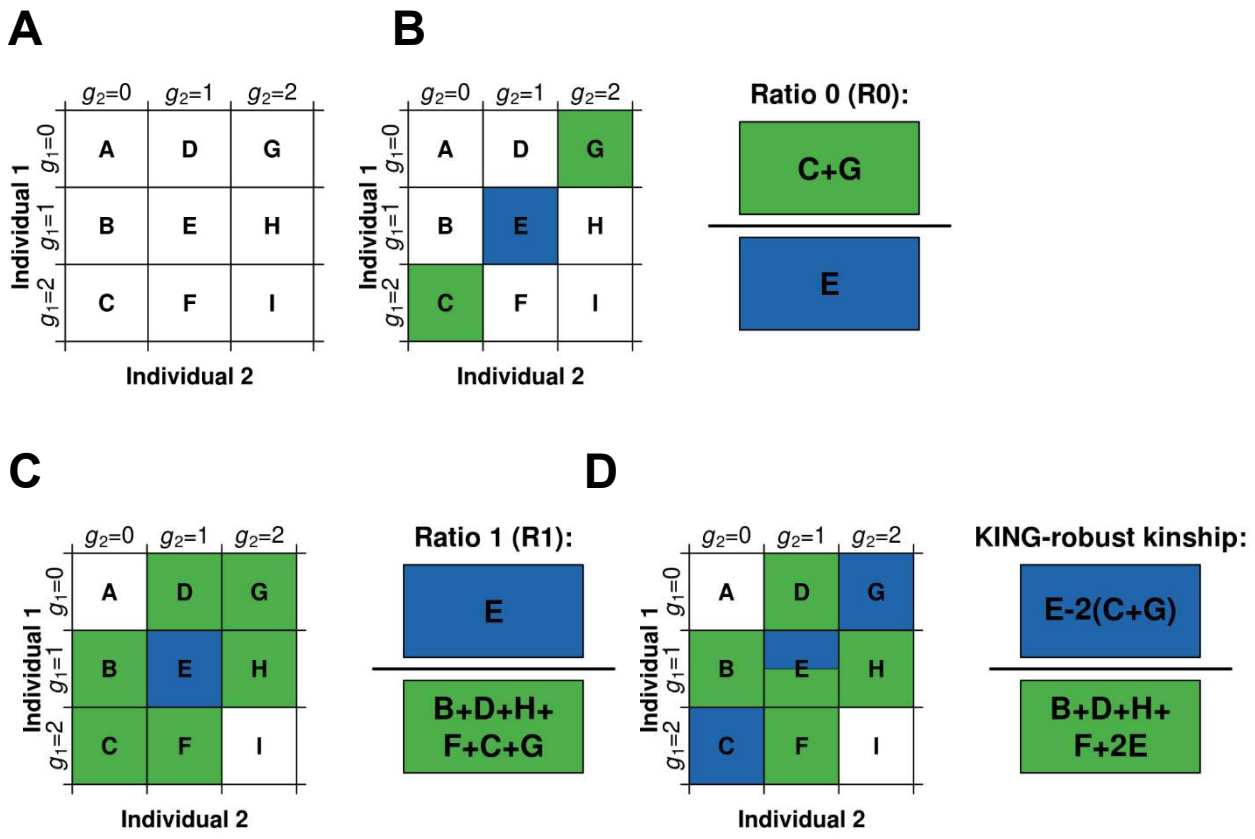

Figure 1. Overview of the R0, R1, and KING-robust kinship statistics
A Definition of the pairwise genotype categories A - I. Here g_1 and g_2 denotes the genotype for each of the two diploid individuals, 1 and 2, respectively. These genotypes are defined as the number of copies of a certain allele carried by 1 and 2, respectively. We assume diallelic variants such that g_1 and g_2 each has 3 possible values: 0, 1, and 2. For a pair of individuals, there are nine possible genotype combinations. We organize them into a 3x3 matrix and denote them with the letters from A to I. The values A – I can equivalently be either counts or proportions. **B** Definition of the R0 statistic based on the notation illustrated in A. **C** Definition of the R1 statistic based on the notation illustrated in A. **D** Definition of the KING-robust kinship estimator (Manichaikul *et al.* 2010), based on the notation illustrated in A.

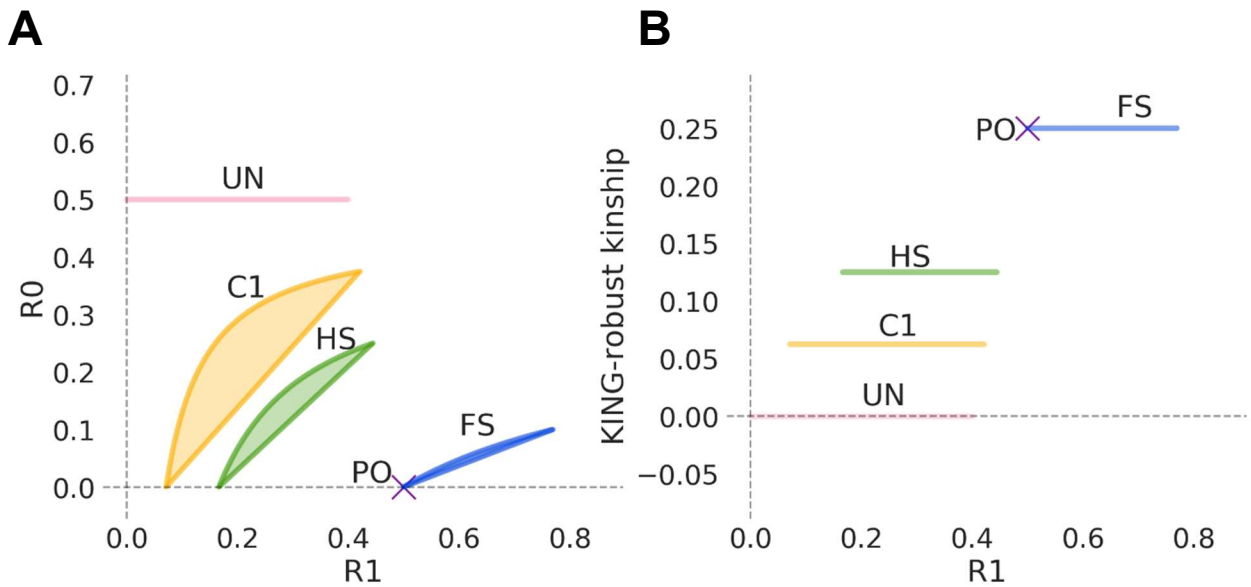

Figure 2. Joint expectation ranges for R1-R0 and for R1-KING robust kinship.

A Joint expectation ranges for R0 - R1 for five different relationship categories: Parent-offspring (PO), full siblings (FS), half siblings/avuncular/grandparent-grandchild (HS), first cousins (C1), and unrelated (UR). The expected value for PO is shown as a purple X. The expected value range for monozygotic twins (MZ) is not depicted because it is (R1=0, R0=∞). **B** Joint expectation ranges for KING-robust kinship and R1 jointly for six different relationship categories: Parent-offspring (PO), full siblings (FS), half siblings/avuncular/grandparent-grandchild (HS), first cousins (C1), and unrelated (UR). The expected pair of values for PO is shown as a purple X. The joint expectation ranges shown here are based on expected IBD sharing (i.e. values of k_0, k_1 and k_2) for each relationship category. True IBD sharing for related individuals has some variation due to biological randomness in recombination and is not likely to be exactly equal to their expected values. This can cause derived statistics to fall outside their expected ranges as well. For detail about how the expected ranges were derived see supplementary text.

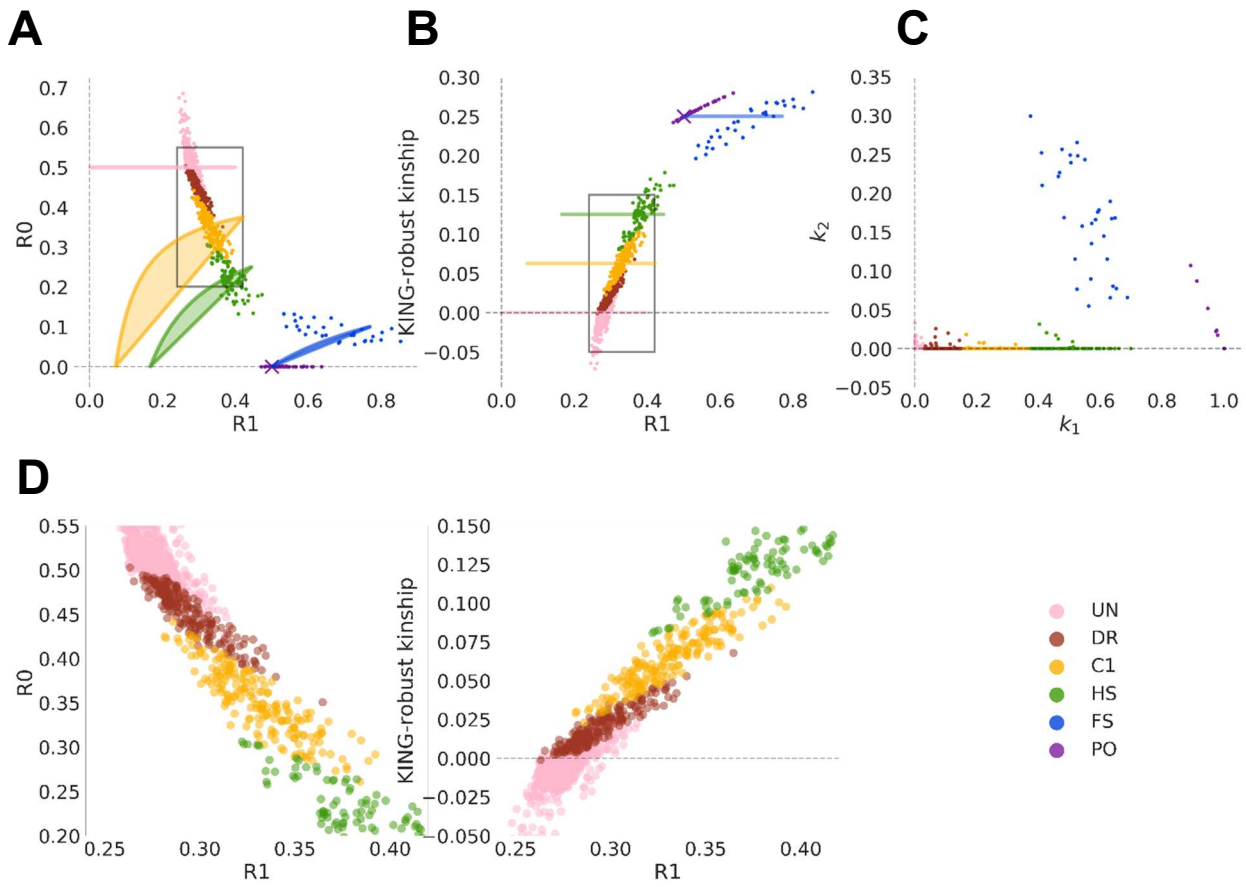

Figure 3. R1-R0 and R1-KING-robust kinship scatterplots for seven HGDP populations. Each point represents a pair of individuals and is colored according to the relationship category inferred using an allele frequency based approach. Shaded areas and lines show the expected ranges for specific relationship categories, as in Figure 2. Note that in addition to the relationship categories for Figure 2 there is an additional category here representing distantly related pairs (DR). **A** R1-R0 plot for all pairs of individuals within each population. **B** R1-KING-robust kinship plot for all pairs of individuals within each population. **C** Scatterplot of the two relatedness coefficients k_1 and k_2 for all pairs of individuals within each population estimated using the allele frequency-based approach implemented in PLINK. The coloring illustrates how these estimates were used to categorize the pairwise relationships, and thus determine the coloring in panel A, B, D and E. **D** Zoom into the outlined area in A. **E** Zoom into the outlined area in B.

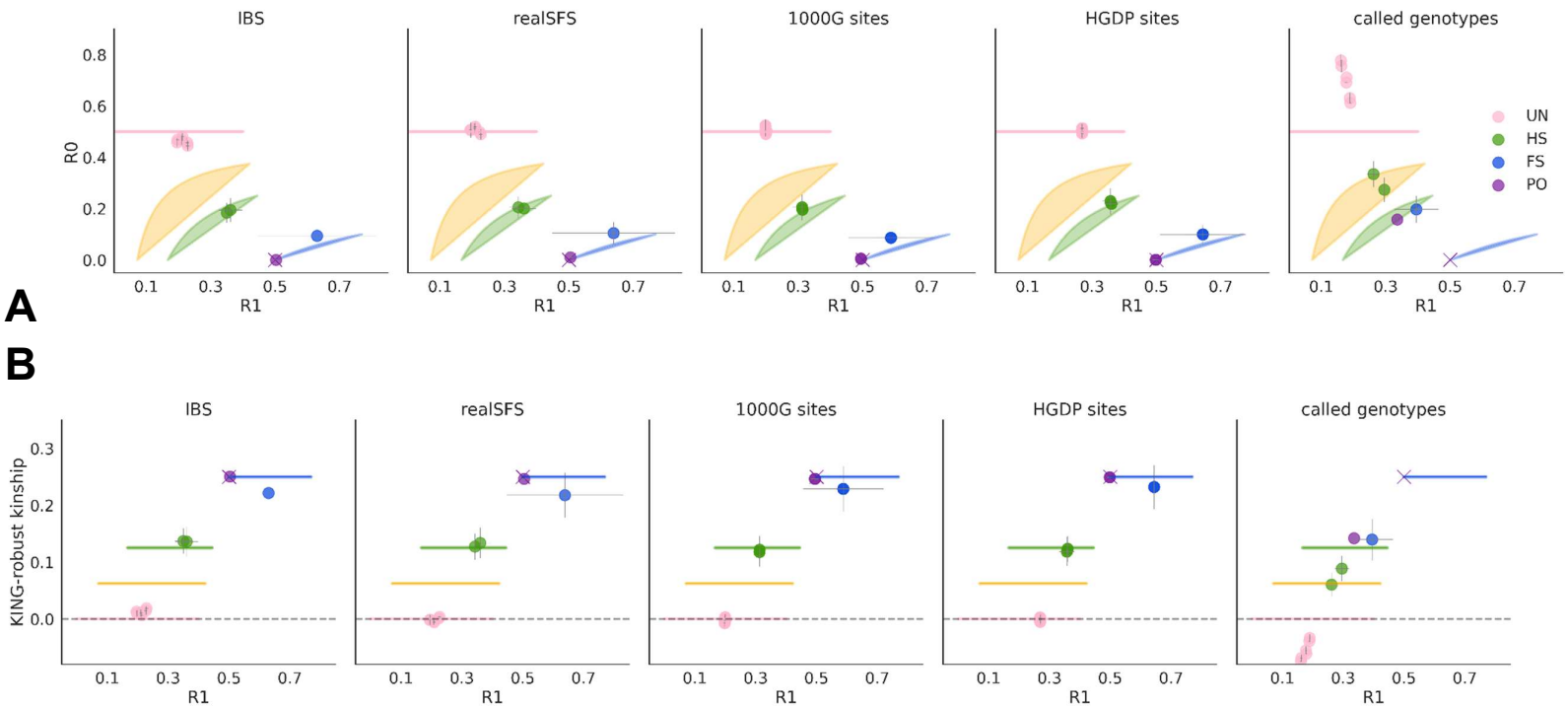

Figure 4. Relatedness plots for all pairs among 5 LWK individuals from the 1000G Project.
A R1-R0 plots for pairs of the same five LWK individuals for five different analysis methods: 1) IBS = reference-free estimation from (~4x) 1000G bam files, 2) realSFS = site-frequency spectrum based estimation from (~4x) 1000G bam files, 3) 1000G genotypes from the 1000G project, 4) HGDP sites=1000G genotypes at sites that overlap with the Illumina 650Y array used for the HGDP, and 5) called genotypes= de novo genotype calling from 1000G bam files. Shaded areas show the joint expected ranges across the two statistics. Points are colored by their true relationship status, as in Figure 3. Thin grey lines show the confidence intervals (+/- 2 SE) with a chromosome jackknife. **B** R1-KING-robust kinship plots for the same data sets, confidence intervals and expected ranges are constructed in the same way. The dashed gray line shows kinship of zero.

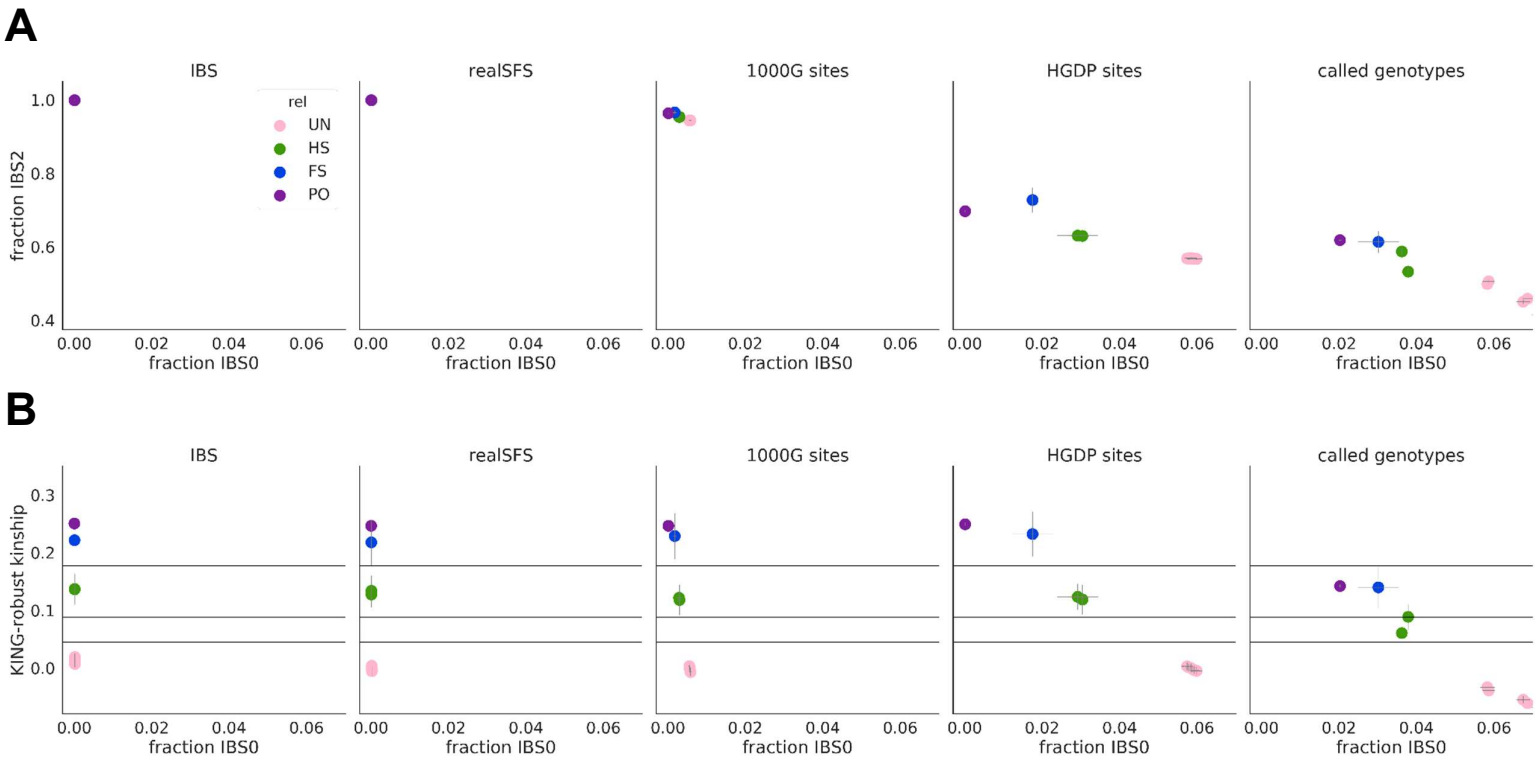

Figure 5. Plots of the two other frequency-free methods for different subsets of SNPs. A Results from applying the plotting approach from Rosenberg (2006) to pairs of the same five LWK individuals for five different analysis methods: 1) IBS = reference-free estimation from (~4x) 1000G bam files, 2) realSFS = site-frequency spectrum based estimation from (~4x) 1000G bam files, 3) 1000G genotypes from the 1000G project, 4) HGDP sites=1000G genotypes at sites that overlap with the Illumina 650Y array used for the HGDP, and 5) called genotypes= de novo genotype calling from (~4x) 1000G bam files. Pairs are colored by their true relationship status, as in Figure 3. Fraction IBS0/IBS2 are the overall fraction of sites that are IBS0/IBS2, respectively. Thin grey lines centered on each point show confidence intervals (+/- 2 SE) based on a chromosome jackknife. **B** Results from applying the KING-robust based approach to the same pairs of LWK individuals

using the same five different analysis methods as in A. The horizontal black lines show the kinship thresholds used to distinguish unrelated (UR), first-cousins (C1) half siblings (HS), full siblings (FS) and (PO) following (Manichaikul *et al.* 2010) from bottom to top, respectively. Like in A, thin grey lines centered on each point show confidence intervals (+/- 2 SE) based on a chromosome jackknife.