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

For submission to Genetics as an Investigation

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 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), and many are available in popular software packages, e.g., 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 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 or 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 instead of genotypes (e.g., Korneliussen and Moltke 2015), however, these methods still rely on accurate allele frequency estimates.

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

To our knowledge, there are only two published methods that can be used to distinguish between different types of familial relationships when neither allele frequencies nor information about the relative genomic position of sites is obtainable. The first of these two methods is to plot the proportion of the genomic sites in which the two individuals share both alleles IBS (which we denote IBS2) versus the proportion of the genomic sites in which they share zero alleles IBS (which we denote IBS0). This method was used

in Rosenberg (2006), where it was applied to the Human Genome Diversity Project (HGDP) dataset. In the resulting plot of the HGDP data (figure 1 of Rosenberg (2006)), pairs of individuals with the same relationship category form distinct clusters in such a way that it is possible to identify parent-offspring pairs, full sibling pairs, and to a lesser extent more distant relationships such as half-siblings/avuncular/grandparent-grandchild and first cousins.

The second existing method is based in part on the KING-robust kinship estimator (Manichaikul *et al.* 2010), which was developed to overcome biases in kinship estimates caused by cryptic population and family structure. More specifically, the KING-robust estimator was developed to be robust to violations of the assumption that the genotypes of the pair of individuals are sampled from a single homogenous population. It was not developed with the aim of providing a kinship estimator that works when allele frequencies within a single population are unknown. However, it is directly applicable to this scenario because, like the test suggested in Lee (2003), it relies only on the genotype combinations within the two target individuals and does not require specification of allele frequencies. Importantly, Manichaikul *et al.* (2010) show it is possible to use SNP array data 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. For example, see figure 3A in Manichaikul *et al.* (2010), a scatterplot of KING-robust kinship vs the fraction of sites IBS0.

Notably, both 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 genotype data as input, which can be problematic for studies where only medium 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

123 likely to differ between different SNP arrays and especially between SNP array and
124 sequencing datasets.

125 Motivated by the outlined limitations to the existing methods, we present a method for
126 relationship inference among non-inbred individuals from the same homogeneous
127 population that relies neither on allele frequencies nor on information about the relative
128 position of the variant sites. The method is inspired by previous methods; it uses the
129 KING-robust kinship estimator and a statistic R0, which is similar to the test statistic
130 from the test for relatedness suggested by Lee (2003). However, it also uses a third
131 statistic, R1, which helps distinguish between full siblings and parent-offspring. Very
132 briefly, the method consists of using two combinations of these three statistics, R1-R0
133 and R1-KING-robust kinship, to infer relationships and compared to previous similar
134 methods it has the advantage that it is robust to SNP ascertainment. Also, we present
135 two ways to estimate the three statistics directly from sequencing data in the absence of
136 called genotypes, which makes the method applicable not only to genotype data, like
137 previous similar methods, but also to low-depth sequencing data.

138 In the following we first fully describe the three statistics, R0, R1 and KING-robust
139 kinship, how they can be estimated and other methodological details. Next, using
140 publicly available SNP array data, we show that the new method provides similar
141 accuracy to the frequency-based method implemented in PLINK, without utilizing any
142 information on allele frequencies or genomic positions. Then using sequence data from
143 the 1000 Genomes, we show that these statistics can be estimated directly from
144 sequencing data of moderate depth (~4x) and that utilizing the sequencing data directly
145 is markedly better than calling genotypes and subsequently estimating R0, R1, and
146 kinship from the called genotypes. Also, using different subsets of the same data, we
147 show that this new method, unlike previous similar methods, is robust to SNP
148 ascertainment. Finally, we discuss some potential applications and limitations.

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, R0) and (R1, 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 counts, including R0, R1 and KING-robust kinship. Specifically, for two non-inbred individuals from the same homogenous population [R1, R0] and [R1, KING-robust kinship], both have the property that the joint expectation ranges for the five relationship categories parent-offspring (PO), full siblings (FS), half siblings/avuncular/grandparent-grandchild (HS), first cousins (C1) and unrelated (UR) do not overlap (Figure 2, for derivations see supplementary text) regardless of the underlying allele frequency spectrum. Second, this method requires neither population allele frequencies nor information about the relative position of the genomics sites analyzed. The only information that is required to estimate the three statistics is genetic data from the two individuals of interest. Third, we expect these statistics to be robust to

SNP ascertainment because they are ratios computed from sites that are variable within the two samples and should thus be unaffected by the number of non-variable sites and because the (unknown) underlying frequency spectrum should only have a limited effect on these ratios. Finally, while the counts of the nine genotype combinations, A-I, and thus R0, R1, and the KING-robust kinship estimator can be calculated from genotype data, they can also be estimated directly from next-generation sequencing (NGS) data based on the expected number of sites with each genotype combination. This makes the approach appropriate even when the sequencing depth is too low for accurate genotype calling (see below for methodological details).

Estimation from sequencing data

The nine genotype combinations, A-I, and thus R0, R1, and the KING-robust kinship for a pair of individual, can be estimated directly from NGS data via the use of genotype likelihoods calculated from the aligned sequencing reads. By using genotype likelihoods it is possible to account for the genotype uncertainly inherent to low-depth NGS data. We used two distinct, but related approaches to estimate these statistics from sequencing data that both build on this idea.

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

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

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

Confidence intervals

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

234 for each statistic. The square root of this variance was interpreted as the standard error
235 in our estimate.

236 Application to real datasets

237 To assess the utility of the new method, we applied it to two different publicly available
238 datasets: SNP array data from seven HGDP populations (Rosenberg 2006), and
239 sequencing data from five related individuals from the Luhya in Webuye, Kenya (LWK)
240 population of the 1000 Genomes Project phase 3 (The Genomes Project 2015).

241 HGDP SNP array

242 The HGDP SNP array dataset was downloaded from `ftp.cephb.fr/hgdp_supp1` (accessed
243 Jan. 13, 2017). Prior to any filters, the collated HGDP dataset had 1043 individuals,
244 644,258 variants, and a genotyping rate of 0.999.

245 Following Rosenberg (2006), we performed basic quality control to exclude error-prone
246 or duplicated samples. Specifically, based on the HGDP subsets described in:
247 `web.stanford.edu/group/rosenberglab/supplements/hgdprelSupp.pdf` (last accessed Jan. 13,
248 2017), we excluded subset 3: "found by (Rosenberg *et al.* 2002) to be extremely atypical
249 and potentially mislabeled". We also removed the duplicate samples listed in subsets 6
250 and 7 and retained subset 9 containing all remaining samples.

251 Next, we selected populations from the HGDP based on two criteria: 1) at least 15
252 unadmixed individuals per population, and 2) presence of several close familial
253 relationship among the individuals. Based on these criteria, we considered samples from
254 seven target populations: five non-African populations: Surui, Pima, Karitiana, Maya,
255 and Melanesian, and two African populations: Mbuti Pygmies and Biaka Pygmies.

256 We excluded individuals showing obvious signs of admixture or inbreeding from the
257 selected HGDP populations. To quantify the admixture proportions of each individual
258 we ran `ADMIXTURE` (Alexander *et al.* 2009) separately for each of the seven target
259 population, while also including samples from several reference populations. In the
260 analyses of each of the non-African target populations we included French, Han, and

261 Yoruban individuals and selected K=4, and in the analyses of each of the African target
262 populations we included French and Han individuals only and selected K=3. To select a
263 set of autosomal loci for `ADMIXTURE`, filters for genotyping rate (>99%), MAF (>5%),
264 HWE ($p > 10^{-10}$ with midp adjustment), and LD (indep-pairwise 50 5 0.5) were applied
265 to each set of selected individuals. Twenty replicate runs of `ADMIXTURE` were performed
266 for each of the seven target populations; all replicates gave rise to likelihoods within 3
267 log likelihood units of each other, suggesting convergence was reached. The results from
268 the replicate with the highest likelihood were selected and individuals in the target
269 population with >5% contribution from more than one ancestry component were
270 excluded. Admixture analyses excluded 16 individuals: 1 Karitiana, 1 Surui, 5 Pima, and
271 9 Maya. The remaining individuals were split by population and we constructed a final
272 set of loci by retaining autosomal loci with genotyping rate (>99%), MAF (>5%), HWE
273 ($p > 10^{-4}$, with midp adjustment), from the original HGDP data (no LD filter). Finally,
274 estimates of the inbreeding coefficient (F) of each individual were calculated within each
275 population and individuals with $F > 0.0625$ were excluded, excluding two Karitiana.
276 This left us with a total of 142 remaining individuals across the six populations: Biaka
277 Pygmies N=31, Karitiana N=21, Maya N=16, Mbuti Pygmies N=15, Melanesian N=19,
278 Pima = 20, Surui N=20. This produced 2902 within-population pairs of individuals that
279 were evaluated for relationship status. The R0, R1, and KING-robust kinship statistics
280 for each pair were calculated from all sites where both individuals had non-missing
281 genotypes.

282 1000 Genomes sequencing data

283 We selected five individuals from two families in the Phase 3 1000 Genomes (1000G)
284 Luhya in Webuye, Kenya (LWK) population: NA19027, NA19042, NA19313, NA19331,
285 NA19334. Across the five individuals, there is one pair of half-sibs (NA19027 &
286 NA19042), and a separate trio of related individuals with a pair of full siblings (NA19331
287 & NA19334), one parent-offspring relationship (NA19313 & NA19331), and another
288 unspecified second-degree relationship (NA19313 & NA19334), possibly avuncular (The

289 Genomes Project 2015). These stated relationships leave six unrelated pairs among the
290 five individuals.

291 For each pair of the five LWK individuals we estimated the R0, R1 and KING-robust
292 kinship statistics in four different ways: 1) by simple genotype counting based on the
293 phased and high-quality curated genotypes provided in the hg37 1000G VCF files, 2&3)
294 by applying the two different sequencing-based approaches described above to the
295 1000G aligned sequence files (bam files) and 4) by re-calling genotypes from the same
296 1000G bam files and then simple counting from the called genotypes. In the fourth
297 estimation method, genotypes were called in a basic manner meant to mimic data from
298 a species with a reference genome but few other genetic resources. Specifically, we used
299 `samtools mpileup` (v1.3.1) to summarize the reads overlapping each position, and
300 `bcftools call` (v1.3.1) to assign the most likely genotype at each position. We used
301 mostly default settings; non-default flags to `samtools` specified skipping indel positions.
302 Non-default flags to `bcftools` specified using the consensus caller and requesting only
303 variant sites. For all sequence-based analyses we only considered reads with a minimum
304 phred-scaled quality score of 30 and bases with minimum phred-scaled quality score of
305 20 and we restricted our analyses to regions with a GEM 75mer mappability of 1
306 (Derrien *et al*. 2012).

307 Assessing the effect of SNP ascertainment

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

Comparison to other methods

To get an allele frequency-based categorization of relationships for the HGDP genotype dataset described above, we first applied the commonly used the allele frequency based relatedness inference program `PLINK` (v1.9) (Chang *et al.* 2015) to the individuals from each population separately to estimate the genome-wide IBD fractions k_0, k_1, k_2, (Z0, Z1 and Z2 in PLINK (Purcell *et al.* 2007). Next, we applied the relationship criteria proposed in Table 1 of Manichaikul *et al.* (2010) to the relatedness coefficient estimates: the estimated k values were combined into an estimate 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 worked on this type of data. More specifically, we applied it to data from seven populations from the HGDP after removing samples with signs of admixture or inbreeding. This resulted in the R1-R0 and R1-KING-robust kinship plots shown in Figure 3 A and B (for population-specific plots see Figure S1 and S2). Because the true relationships for the pairs of individuals is not known, we colored each point not according to the truth, but according to the relationship category inferred using a common 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 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). Also, they are located fairly close to 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 highlighted areas of Figure 3A and B). But notably the

few pairs of individuals that are difficult to classify are the same pairs as those that are edge cases for the allele frequency-based method. This is apparent in an R1-R0 plot of the HGDP data constructed excluding pairs that are closer than 0.01 to the kinship coefficient thresholds that the frequency based method used when categorizing relationships (Figure S4).

Application to sequencing data

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

Additionally, we calculated the three statistics from the high-quality phased genotypes for the same five individuals available from the 1000 Genomes Project Phase 3 (Figure 4) to see how well the two genotype-likelihood-based approaches applied to low-depth sequencing data perform compared to direct calculations from high-quality genotype data for the same samples. In this comparison, results obtained by using the genotype

401 likelihood-based approaches applied to low depth sequencing data are close to those
402 obtained from the high-quality genotypes for many of the pairs (Figure 4). Chromosome
403 block-jackknife estimates of uncertainty were generally small, and varied by relationship
404 type, with the pair of full siblings having the most uncertainty in R0, R1, and KING-
405 robust kinship. Notably, the IBS-based approach provided meaningful results, even
406 though it did not make use of any external allele information. With the IBS method,
407 unrelated individuals had a slight decrease in R0 and a slight increase in R1 and kinship,
408 relative to the SFS method, making them appear slightly more related than expected for
409 unrelated individuals from a homogenous population. However, despite this bias, it is
410 still possible to correctly categorize of all the pairs to their relationship category within
411 the set of close familial relationships [PO, FS, HS, or UR]. Suggesting that this method
412 can be used when not enough information is available for the realSFS method.

413 Finally, we also made R1-R0 plot and the R1-KING-robust kinship plots based on
414 genotypes that we obtained through a standard genotype calling procedure from the raw
415 read data. We did this to investigate if the genotype likelihood-based approaches are
416 necessary or one could just as well use genotypes called from the 4x data. Genotype
417 calling had a large negative effect on the outcome; in the resulting R1-R0 and R1-KING-
418 robust kinship plots the half siblings appear where first cousins are expected and both
419 the parent-offspring and full sibling pairs appear in the expected range of half siblings
420 (Figure 4). These results demonstrate the potential pitfalls of basing relationship
421 inferences, including R1-R0 and R1-KING-robust kinship plots, on genotypes called
422 from low-depth data. Notably, this is also the case for the methods presented in
423 (Rosenberg 2006) and (Manichaikul *et al.* 2010) (Figure 5A). Hence, for 4x sequencing
424 data, called genotypes are not a good alternative to genotype likelihood-based
425 approaches.

426 Assessing the effect of SNP ascertainment scheme

427 To assess the effect of SNP ascertainment on the new method, and thus the effect of the
428 data-generation platform or SNP filtering, we applied it to three different subsets of data
429 from the five 1000 Genomes individuals. The results for each of the three ascertainment

schemes (Illumina 650Y SNP array sites, 1000 Genomes release sites and whole genomes) are similar (Figure 4E and F versus Figure 4B and C, respectively), suggesting SNP ascertainment does not have a large effect.

For comparison, we performed the same assessment of the methods presented in Rosenberg (Rosenberg 2006) and (Manichaikul *et al.* 2010) by constructing scatterplots of the same type as those shown in their papers (Figure 5). This shows that both these other methods are much more affected by ascertainment than the method proposed here. The Rosenberg method is affected on both its x-axis (IBS0) and its y-axis (IBS2), which means that the expected region of the plot for each relationship will be different for different data sets. The method presented in Manichaikul *et al.* 2010 is affected by the SNP ascertainment mainly on its x-axis (IBS0). In this case, the ascertainment mainly affects the ability to distinguish between parent-offspring and full siblings, since the y-axis, which is only slightly affected by ascertainment, is the kinship coefficient, which can be used to distinguish between most close relationships except for parent-offspring and full siblings. The x-axis, IBS0, is included in part to help make the distinction between PO and FS (Manichaikul *et al.* 2010), but this ability is clearly affected by SNP ascertainment (Figure 5B).

Discussion

We have presented a simple new method for inferring if and how two individuals are related based solely on genetic data from the two target individuals. This new method has several advantages compared to most existing relatedness inference methods, which are almost all dependent on additional information. First, it does not require any information about allele frequencies. Second, it does not require information about the relative genomic positions of the sites included in the analysis. Third, although the method is straightforward when applied to genotype data (where it boils down to simple counting, addition, and division), we demonstrate two ways to utilize sequencing data via genotype likelihoods and we show that 4x sequencing data that is insufficient for high-quality genotype calling can provide useful results. This implies that this new method can be used even if all you have is low-depth sequencing data from a few

459 individuals from a species without a reference genome. This is not likely to be the typical
460 situation in studies of modern humans, however it is much more likely to occur in
461 studies of with non-model organisms or ancient samples. Fourth and finally, we have
462 shown that the new method is robust to the SNP ascertainment present in common SNP
463 array data sets.

464 Comparison to similar methods

465 The new method is based on plotting two statistics, R0 and KING-robust kinship against
466 a third, new statistic R1. The R0 statistic is derived from the test statistic proposed by
467 Lee (2003) to test for relatedness; the only differences are that the numerator and
468 denominator are flipped and that E, the proportion of sites where both individuals are
469 heterozygous, is included in the denominator in the statistic defined by Lee but absent
470 in R0. The KING-robust kinship estimator was proposed by Manichaikul *et al.* (2010),
471 here we simply extend it to estimation directly from sequencing data. Notably, our
472 results suggest that this extension is important for low-depth sequencing data, because
473 estimates based on genotypes called from low-depth sequencing data are not very good.
474 And importantly this extension can also be used for similar statistics and thus makes
475 existing methods based on such statistics, like KING-robust kinship, more widely
476 applicable.

477 However, this extension is not the only contribution of this study. Another key new
478 contribution of this study is the fact that R0 and KING-robust kinship are combined
479 with the new R1 statistic instead of with IBS0 (the proportion of sites where the two
480 individuals share zero alleles IBS) as was previously proposed by Rosenberg (2006) and
481 Manichaikul *et al.* (2010). As we have shown, the fraction of sites that are IBS0 or IBS2
482 is very sensitive SNP ascertainment, meaning that results are only comparable within
483 each different ascertainment scheme. Across different ascertainment schemes, such as
484 between SNP arrays and sequencing data, the method from Rosenberg (2006), where
485 IBS0 is combined with IBS2, is difficult, if not impossible, to use for relatedness
486 inference because the fraction of sites that are IBS2 and IBS0 varies so wildly. On the
487 other hand, KING-robust kinship is still very useful, but it loses its ability to distinguish

488 between parent-offspring and full siblings, as IBS0 was used for this. Due to this
489 sensitivity to ascertainment samples cannot be analyzed in isolation, and must be placed
490 in the context of other samples with known relationships and the same ascertainment
491 scheme. This requirement makes it difficult to apply these previous methods to ancient
492 humans or other species with limited sample sizes.

493 In contrast, the ability to identify relatives based on expected values is maintained in the
494 new method, regardless of ascertainment scheme due to the use of R1, instead of IBS0,
495 which makes the new method robust to SNP ascertainment. Here the R1-R0 plot is
496 especially useful, because parent-offspring pairs, tend to have an R0 estimate extremely
497 close to 0, making them particularly easy to identify. Additionally, it is worth noticing
498 that R1 not only makes it easier to distinguish between parent-offspring and full
499 siblings, both the new plot types, R1-R0 and R1-KING-robust kinship provide better
500 separation of several of the different relationship categories than KING-robust kinship
501 does alone (Figure S2). One appealing aspect of KING-robust kinship is that kinship
502 axis has a biological interpretation, defined as the probability that two alleles sampled at
503 random from two individuals are identical by descent. Hence the two plots types, R1-R0
504 and R1-KING-robust kinship, each have their advantages.

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

507 Limitations and applications

508 While the new method provides significant advantages over previous methods in
509 situations with limited data, it does have some limitations. First, like most other
510 relatedness inference methods the proposed method assumes that the individuals are
511 not inbred and that they originate from the same homogenous, population. And like
512 many other relationship inference methods, it is not necessarily robust to violations of
513 these assumptions. Previous studies have shown the effect of admixture on relatedness
514 inference is complex and can potentially lead to bias in either direction depending on
515 the circumstances (Thornton *et al*. 2012; Moltke and Albrechtsen 2014; Conomos *et al*.
516 2016; Ramstetter *et al*. 2017). Specific methods have been developed to correct for

517 admixture when the allele frequencies in the admixing populations are known (e.g.
518 Thornton *et al.* 2012; Moltke and Albrechtsen 2014), or enough samples are available
519 (Conomos *et al.* 2016). However, since these methods work by exploiting knowledge
520 about allele frequencies or access to large sample sizes for their correction, the R0, R1
521 and KING-robust kinship statistics can not be corrected in a similar manner. Instead we
522 suggest trying to detect inbreeding and admixture before performing relatedness
523 analyses and to proceed with caution when it is present. One potential way to identify
524 admixture or inbreeding is to compare the heterozygosities for the pair of individuals;
525 two non-inbred and non-admixed individuals from the same population should have
526 similar heterozygosity, so heterozygosity differences can be a warning signal.

527 We can also flip these relatedness statistics around and ask if the pair of individuals
528 appear to be *less* related than we expect from two individuals sampled from a single
529 population. In this vein, Lee (2003) showed that the statistic he proposed can also be
530 used to detect if two unrelated samples are not from the same homogeneous population.
531 If this is the case, the Lee's statistic will be significantly larger than 2/3; equivalently, R0
532 will be significantly above 0.5 and KING-robust kinship will be negative.

533 A second limitation is that the categorization of different relationship pairs is based on
534 visual inspection of plots and that in some cases such classification will be difficult
535 because the ranges of the different relationship in the plots are close to each other and
536 even overlap. However, this is to a large extent due to biological variation caused by the
537 randomness of recombination (Hill and Weir 2011). This limitation in classification
538 accuracy is therefore shared with the allele frequency based inference methods, as
539 illustrated by Figure S5.

540 Despite these limitations, we believe that the results presented here suggest that the new
541 method constitutes a helpful new tool for relatedness inference that is especially well-
542 suited for situations with limited data. Identifying related individuals is a crucial step in
543 nearly any genetic analysis and removing the requirements to specify allele frequencies
544 and to call genotypes has the potential allow the identification of relatives even in small
545 studies of non-model species or ancient samples. These types of studies do not currently
546 have many good options to address relatedness among their samples. An additional

547 appeal of the IBS-based approach introduced here, is that because it considers all
548 possible pairwise genotype categories, it can potentially help identify specific error types
549 that are expected to occur in DNA from ancient samples, by allowing researchers to
550 specifically identify transitions. Also, due to its non-reliance on allele frequencies the
551 method is also potentially useful when relatedness is pervasive among many samples,
552 complicating allele frequency estimation in the broader population (Mcpeek *et al.*
553 2004). Hence, the method will hopefully be useful to many.

554

555 Code availability:

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

558

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.

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.

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

596 Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale *et al.*, 2010 Robust
597 relationship inference in genome-wide association studies. Bioinformatics 26:
598 2867-2873.

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

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

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

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

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

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

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

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

620 Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002
621 Genetic structure of human populations. Science 298: 2381-2385.

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

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

628 The Genomes Project, C., 2015 A global reference for human genetic variation. Nature
629 526: 68-74.

630 Thompson, E. A., 1975 The estimation of pairwise relationships. Ann Hum Genet 39:
631 173-188.

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

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

636 Wiuf, C., 2006 Consistency of estimators of population scaled parameters using
637 composite likelihood. Journal of Mathematical Biology 53: 821-841.
638

Figures for:

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

For submission to Genetics

Authors:

Ryan K. Waples
Anders Albrechtsen
Ida Moltke

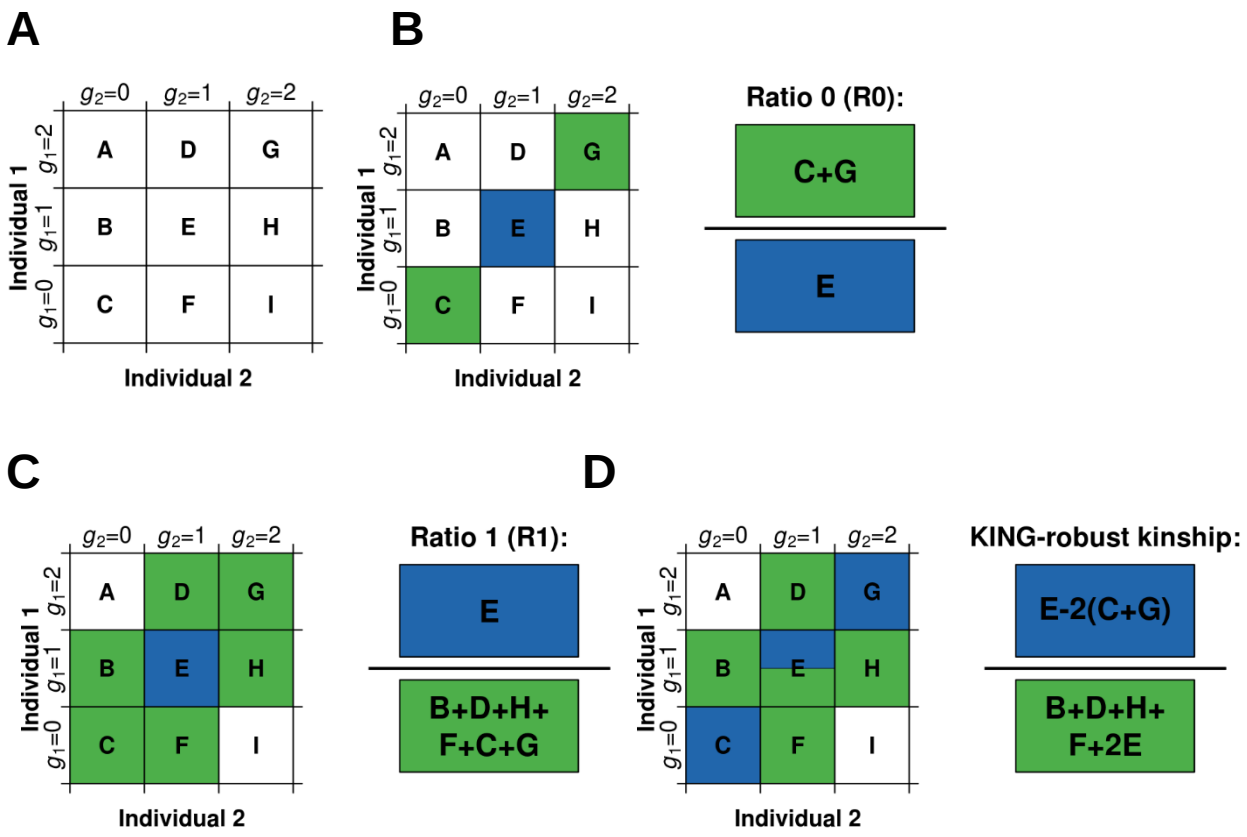

Figure 1. Overview of the R0, R1, and King-robust kinship statistics
A Definition of the pairwise genotype categories A - I. Here g_1 and g_2 denotes the genotype for each of the two diploid individuals, 1 and 2. These genotypes are defined as the number of copies of a certain allele carried by 1 and 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.

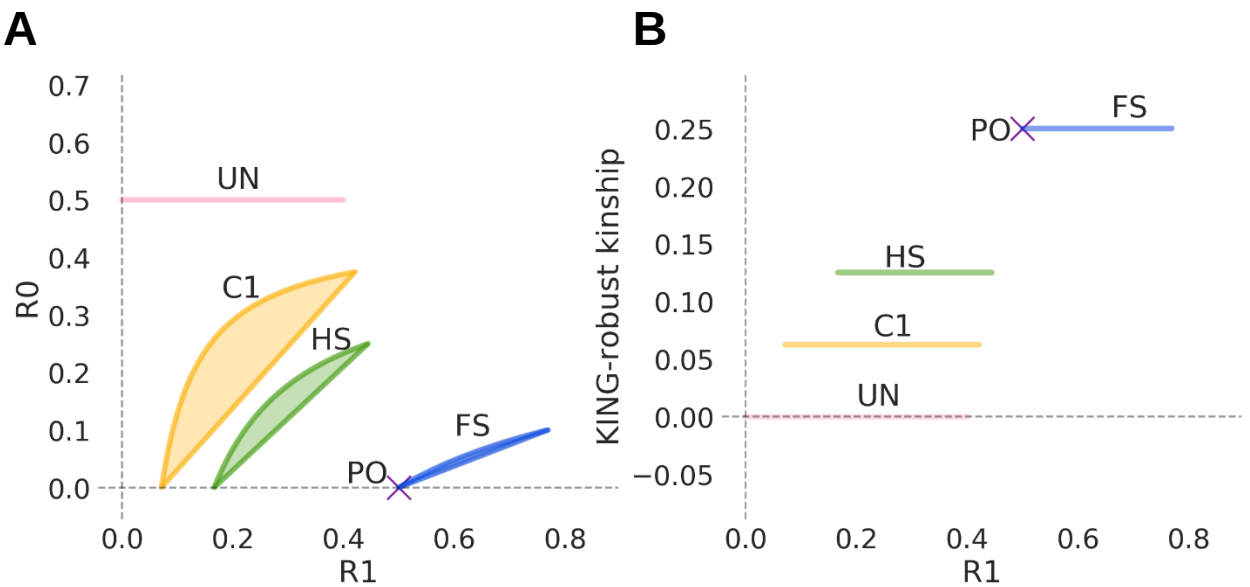

Figure 2. Joint expectation ranges for R1-R0 and for R1-KING robust kinship.

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 value for PO is shown as a purple X.

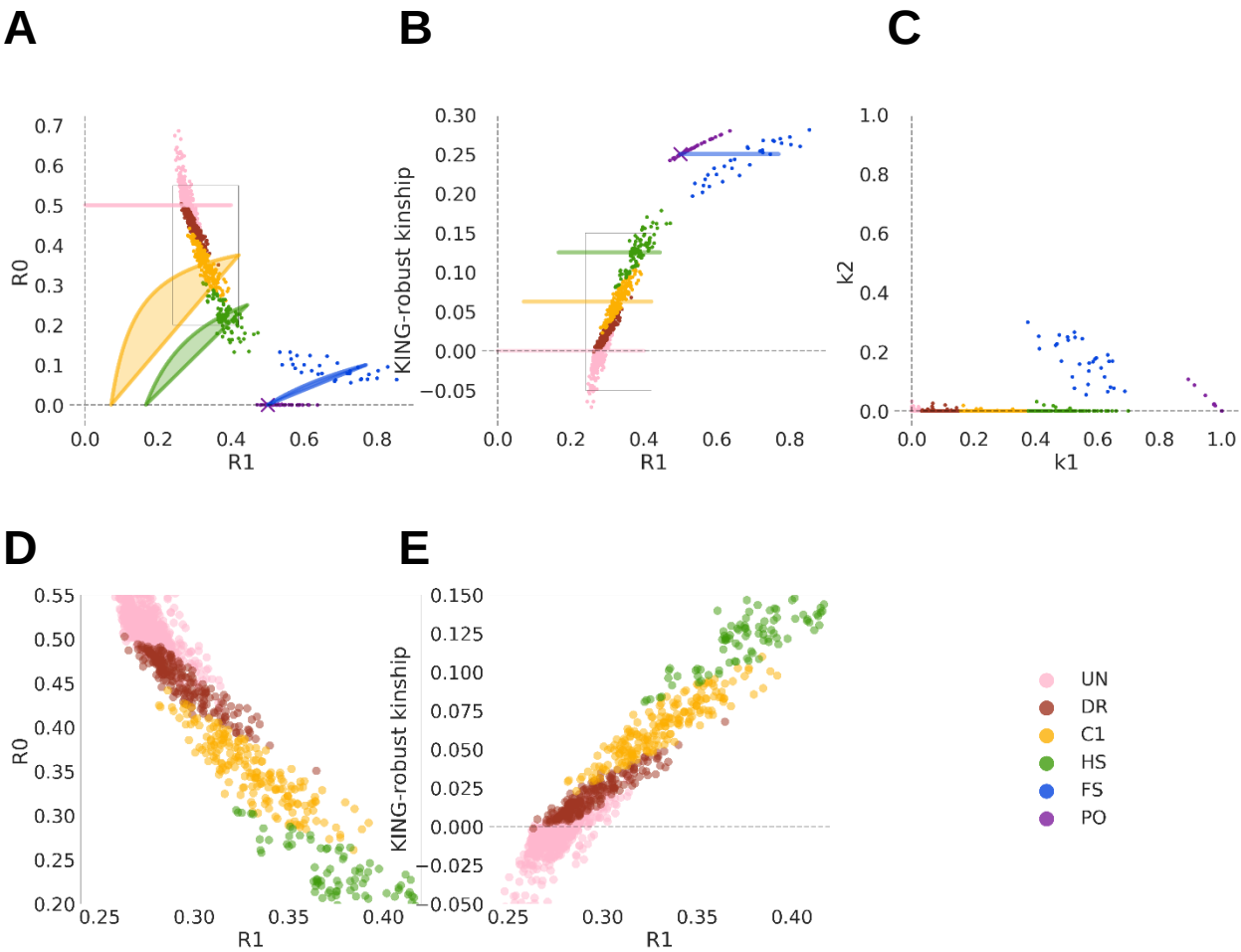

Figure 3.
R1-R0 and R1-KING-robust kinship scatterplots for seven HGDP populations. Each point represents a pair of individuals and is colored according to the relationship category inferred using an allele frequency based approach. Shaded areas and lines show the expected ranges for specific relationship categories, as in Figure 2. Note that in addition to the relationship categories for Figure 1 there 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 k1 and k2 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** Zoom into the outlined area in A. **E** Zoom into the outlined area in B.

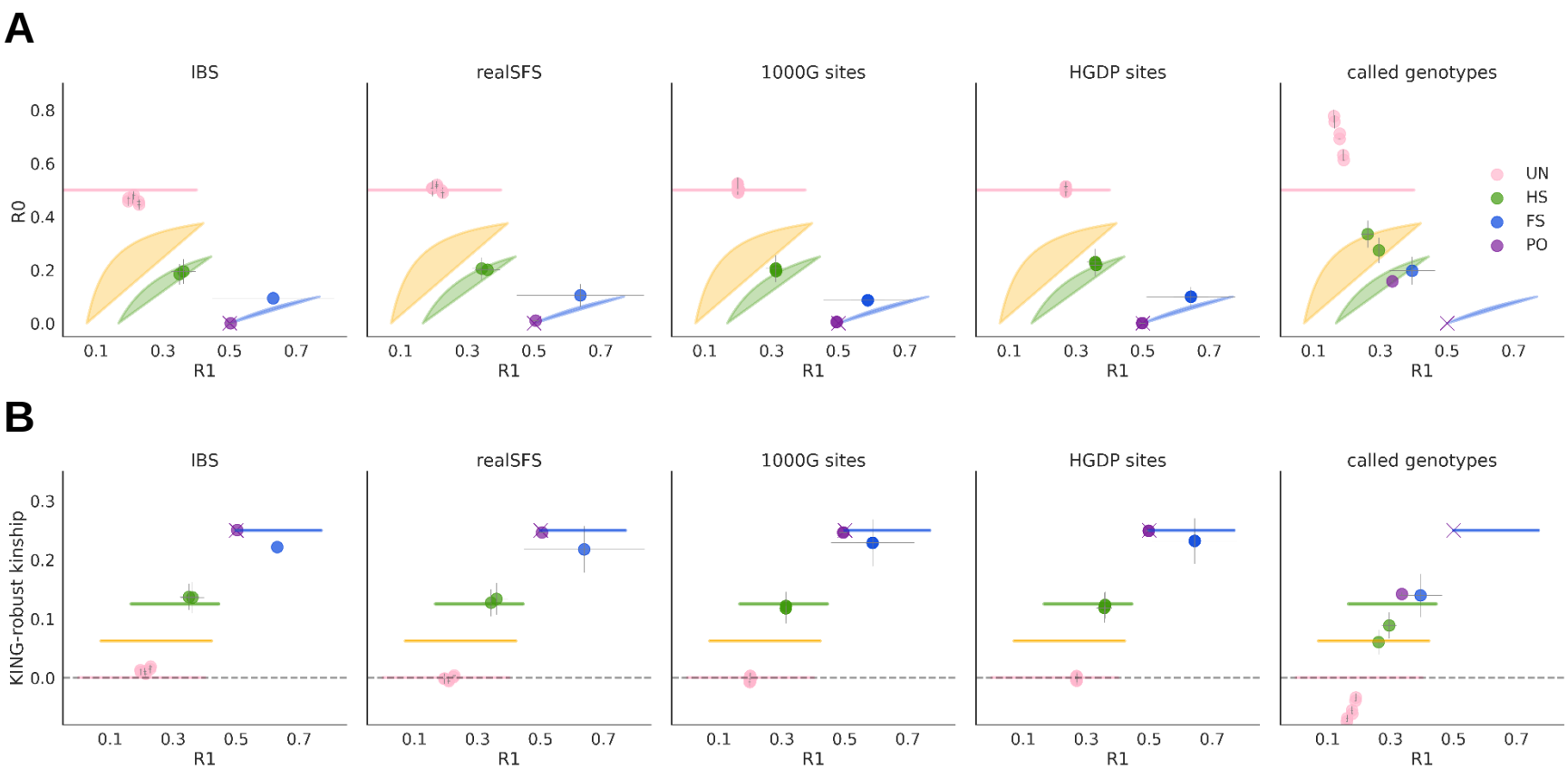

Figure 4.

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.

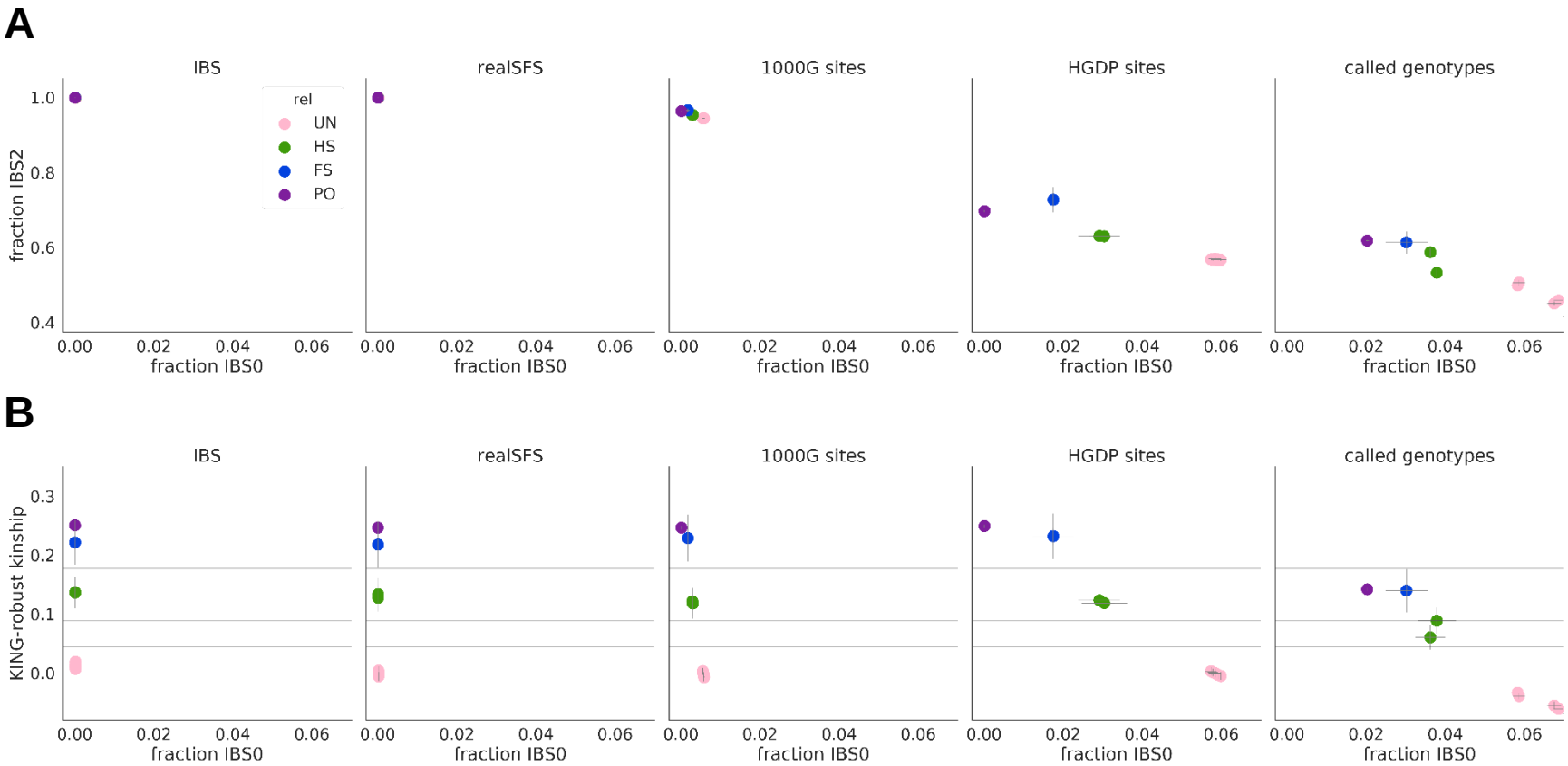

Figure 5. Plots of the two other frequency-free methods for different subsets of SNPs.
A Results from applying the plotting approach from Rosenberg (2006) to pairs of the same five LWK individuals for five different analysis methods: 1) IBS = reference-free estimation from (~4x) 1000G bam files, 2) realSFS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