Robust relationship inference in genome-wide association studies

Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale and Wei-Min Chen

1 Center for Public Health Genomics, 2 Department of Public Health Sciences, Division of Biostatistics and Epidemiology, University of Virginia, Charlottesville, VA, 3 Department of Otolaryngology, University of Minnesota, Minneapolis, MN, 4 Department of Medicine and 5 Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA, USA

ABSTRACT

Genome-wide association studies (GWASs) have been widely used to map loci contributing to variation in complex traits and risk of diseases in humans. Accurate specification of familial relationships is crucial for family-based GWAS, as well as in population-based GWAS with unknown (or unrecognized) family structure. The family structure in a GWAS should be routinely investigated using the SNP data prior to the analysis of population structure or phenotype. Existing algorithms for relationship inference have a major weakness of estimating allele frequencies at each SNP from the entire sample, under a strong assumption of homogeneous population structure. This assumption is often untenable.

Results: Here, we present a rapid algorithm for relationship inference using high-throughput genotype data typical of GWAS that allows the presence of unknown population substructure. The relationship of any pair of individuals can be precisely inferred by robust estimation of their kinship coefficient, independent of sample composition or population structure (sample invariance). We present simulation experiments to demonstrate that the algorithm has sufficient power to provide reliable inference on millions of unrelated pairs and thousands of relative pairs (up to 3rd-degree relationships). Application of our robust algorithm to HapMap and GWAS datasets demonstrates that it performs properly even under extreme population stratification, while algorithms assuming equilibrium (HWE) assumptions. Estimation of allele frequencies at each SNP according to Hardy–Weinberg Equilibrium (HWE) assumption is critical for the performance of family-based GWA, as well as in population-based data with unknown family structure. High-throughput genotyping performed in a GWAS presents new opportunities for pedigree error detection using millions of SNPs to assess the degree of relationship between a pair of individuals. With these opportunities come the challenges of accounting for linkage disequilibrium among typed markers, while managing computational resources to analyze the large amount of genotype data. Compared to linkage studies, association studies also require consideration of population substructure, misreported race and ethnicity and unreported familial relationships among samples recruited as unrelated individuals.

One well-developed approach for relationship inference in linkage studies offers fully parametric methods for sib pairs (Boehnke and Cox, 1997) and extensions to general pedigrees (McPeek and Sun, 2000) using hidden Markov models (HMM) to calculate multipoint marker probabilities, incorporated into a likelihood framework to assess evidence in support of particular pair-wise relationships. In considering full multipoint marker probabilities, computational demands increase with the number of markers genotyped, making analysis of GWAS SNPs for all pairs of individuals prohibitive. A simple method, known as GRR (Graphical Representation of Relationship errors; Abecasis et al., 2001), uses clustering of readily available non-parametric estimates for mean and standard deviation (SD) of identical by state (IBS) statistics at a series of markers for each pair of relatives. GRR identifies outliers of clusters as relationship errors. Performance of the clustering algorithm used to classify relative pairs depends on the panel of genetic markers, the underlying allele frequencies of genetic markers for different individuals, and the number of individuals genotyped. If certain pairs of individuals do not cluster—either due to limitations in sample size or due to the different underlying allele frequencies between different pairs (e.g. in the presence of population structure) —GRR fails to detect the pedigree errors. One efficient implementation of relationship inference in GWAS data is available in a widely used software package, PLINK (Purcell et al., 2007). The identical-by-descent (IBD) statistics between each pair of individuals are estimated using the average of IBS and the estimation of sample-level allele frequencies at each SNP according to Hardy–Weinberg Equilibrium (HWE) assumptions.

All popular algorithms for relationship inference depend on reliable estimates of allele frequencies at each SNP, assuming a homogeneous population without stratification (Abecasis et al., 2001; Boehnke and Cox, 1997; Lynch and Ritland, 1999; McPeek et al., 2001). Here, we present a robust algorithm for relationship inference in GWAS datasets that allows the presence of unknown population substructure. The family structure in a GWAS should be routinely investigated using the SNP data prior to the analysis of population structure or phenotype. Existing algorithms for relationship inference have a major weakness of estimating allele frequencies at each SNP from the entire sample, under a strong assumption of homogeneous population structure. This assumption is often untenable.

Results: Here, we present a rapid algorithm for relationship inference using high-throughput genotype data typical of GWAS that allows the presence of unknown population substructure. The relationship of any pair of individuals can be precisely inferred by robust estimation of their kinship coefficient, independent of sample composition or population structure (sample invariance). We present simulation experiments to demonstrate that the algorithm has sufficient power to provide reliable inference on millions of unrelated pairs and thousands of relative pairs (up to 3rd-degree relationships). Application of our robust algorithm to HapMap and GWAS datasets demonstrates that it performs properly even under extreme population stratification, while algorithms assuming equilibrium (HWE) assumptions fail to detect the pedigree errors. One efficient implementation of relationship inference in GWAS data is available in a widely used software package, PLINK (Purcell et al., 2007). The identical-by-descent (IBD) statistics between each pair of individuals are estimated using the average of IBS and the estimation of sample-level allele frequencies at each SNP according to Hardy–Weinberg Equilibrium (HWE) assumptions.
and Sun, 2000; Purcell et al., 2007). Recent GWAS analytic advances for association mapping have incorporated the presence of unknown family and population structure (Choi et al., 2009; Kang et al., 2010; Thornton and McPeek, 2010; Yang et al., 2010; Zhang et al., 2010); however, algorithms to estimate family relationships remain based on the assumption of population homogeneity. In samples with undetected population substructure, this strong assumption of population homogeneity leads to biased results, systematically inflating the degree of relatedness among individuals of the same racial group.

Current approaches to relationship and population structure inference are somewhat circular. The relationship inference relies on correct specification of a homogeneous subpopulation (Purcell et al., 2007), while the detection of population structure relies on the correct identification of unrelated individuals (Zhu et al., 2008). In addition to the non-robustness to the population structure, existing approaches do not apply to small datasets, e.g. for comparison of a single pair of individuals, or relationship inference on a single pedigree.

We present a novel framework for relationship inference, Kinship-based Inference for Genome-wide association studies (KING), together with a rapid algorithm for relationship inference appropriate for use on samples with thousands of individuals genotyped at millions of SNPs from autosomes, consistent with a scale typically achieved in a GWAS. Within this framework we present two methods: (i) KING-homo, derived under the assumption of population homogeneity and (ii) KING-robust that provides robust two methods: (i) KING-homo, derived under the assumption of population homogeneity leads to biased results, and (ii) KING-robust that provides robustness to the population structure, existing approaches do not apply to small datasets, e.g. for comparison of a single pair of individuals, or relationship inference on a single pedigree.

We propose an alternative framework for relationship inference, Kinship-based Inference for Genome-wide association studies (KING), together with a rapid algorithm for relationship inference appropriate for use on samples with thousands of individuals genotyped at millions of SNPs from autosomes, consistent with a scale typically achieved in a GWAS. Within this framework we present two methods: (i) KING-homo, derived under the assumption of population homogeneity and (ii) KING-robust that provides robustness to the population structure, existing approaches do not apply to small datasets, e.g. for comparison of a single pair of individuals, or relationship inference on a single pedigree.

2 METHODS

Consider two individuals, indexed by i and j. Let $\phi_{ij}$ denote the kinship coefficient, defined as the probability that two alleles sampled at random from two individuals are identical by descent, and $\tau_{ij}$ and $\pi_{ij}$ denote the probability that the two individuals share zero, one and two alleles identical by descent, respectively. Table 1 lists values of $\phi_{ij}$ and $\tau_{ij}$ for relative pairs, including monozygotic twins, parent-offspring pairs, sib pairs, 2nd-degree relative pairs (such as half-sibs, avuncular pairs and grandparent-grandchild pairs), 3rd-degree relative pairs (such as first cousins), and unrelated pairs. Note that the kinship coefficient is a function of IBD-sharing statistics with relationship $\hat{\phi}_{ij} = \pi_{ij}/2 + \tau_{ij}$. Inference criteria presented in Table 1 are derived using powers of 2, with the basis that this is the natural scale of the kinship and zero IBD-sharing statistics. In Section 3, we see that these inference criteria work well in practice.

2.1 Relationship inference in a homogeneous population

We first review existing methods that allow relationship inference under the assumption of a homogeneous population. Assume $p$ is the frequency of a reference allele (with label A) at a SNP, and the number of alleles IBS between individuals i and j is $IBS_{ij}$. Since only $IBD_{ij} = 0$ (not $IBD_{ij} = 1$ or 2) can result in $IBS_{ij} = 0$ (i.e. the pair of individuals has genotypes AA and aa), the expected proportion of SNPs with zero IBS can be specified assuming

| Relationship | $\phi_{ij}$ | Inference criteria | $\tau_{ij}$ | Inference criteria |
|--------------|-------------|--------------------|-------------|--------------------|
| Monozygotic twin | 1.0 | $\phi_{ij}$ > 0.5 | 0 | $\phi_{ij}$ > 0.1 |
| Parent-offspring | 0.5 | $\tau_{ij}$ = 0 | 0 | $\phi_{ij}$ > 0.1 |
| Full sib | 0.25 | $\tau_{ij}$ = 0.5 | 0 | (0.1, 0.365) |
| 2nd Degree | 0.125 | $\tau_{ij}$ = 1 (0.365, 1 - $\tau_{ij}$) | 0 | $\phi_{ij}$ > 0.1 |
| 3rd Degree | 0.0625 | $\tau_{ij}$ = 1 (0.365, 1 - $\tau_{ij}$) | 0 | $\phi_{ij}$ > 0.1 |
| Unrelated | 0 | $\tau_{ij}$ = 0 | 1 | $\phi_{ij}$ > 0.1 |

HWE:

$$Pr(IBS_{ij} = 0) = Pr(\text{AA, aa} | IBD_{ij} = 0) = 2p^2(1 - p^2)\pi_{AA}$$

This leads to the estimator

$$\hat{\phi}_{ij} = \frac{\sum_{m=0}^{\infty} \tau_{ij}(1 - \pi_{AA})^m - \sum_{m=0}^{\infty} \pi_{AA}(1 - \pi_{AA})^m}{\sum_{m=0}^{\infty} \hat{\pi}_{AA}(1 - \hat{\pi}_{AA})^m}$$

Note $\hat{\pi}_{AA}$, $\hat{\pi}_{Aa}$ and $\hat{\pi}_{aa}$ are the total number of individuals with genotype AA, Aa and aa, respectively, at the m-th SNP. The remaining two IBD statistics can be estimated based on $\hat{\pi}_{AA}, \hat{\pi}_{Aa}$ and $\hat{\pi}_{aa}$ (Purcell et al., 2007). Since the sum of the three IBD statistics is unity, only two IBD statistics are needed to infer the relationship.

We propose an alternative framework to estimate the kinship coefficient between a pair of individuals. Suppose the frequency of a reference allele is $p$ at a SNP for both individuals. The genotype score, defined by the number of the reference allele for individuals i, is $X_{ij}$. We model genetic distance between a pair of individuals in terms of their kinship coefficient (derived under the assumption of HWE in the Supplementary Material) as

$$E(X_{ij} | X_{ij} = i) = 4p(1 - p)\pi_{AA}$$

Let $\hat{H}_i/M_i$ be a consistent estimator of $\sum_{l=0}^{\infty} \tau_{il}(1 - \pi_{AA})/M_i$ where $M_i$ is the total number of non-missing markers for the pair of individuals. Now, we can estimate the kinship coefficient as

$$\hat{\phi}_{ij} = \left(\frac{\sum\left(\pi_{AA} - \pi_{AA}^2\right)\tau_{ij}}{4M_i}\right)^{1/2}$$

Note only markers with genotype data for both individuals i and j are used in calculation of $\hat{\phi}_{ij}$. When the sample of individuals is homogeneous, $\hat{\phi}_{ij}$ can be estimated by the observed allele frequency $\hat{\pi}_{AA}$ in (3). The plug-in estimator

$$\hat{H}_i/M_i = \sum_{l=0}^{\infty} 2\pi_{Aa}(1 - \phi_{il})/M_i$$

is consistent for $\sum_{l=0}^{\infty} 2\pi_{Aa}(1 - \phi_{il})/M_i$ and it follows that the estimator $\hat{\phi}_{ij}$ based on (5) and (6) is consistent for $\phi_{ij}$. We name the estimating method in Equations (5) and (6) as KING-homo. Together with the IBD estimator (2), all relationships presented in Table 1 can be determined uniquely. Note estimation of $\pi_{AA}$ and $\pi_{Aa}$ can be derived easily according to equations $\pi_{AA} = 2 - 2\hat{\phi}_{01} - 4\hat{\phi}$ and $\pi_{Aa} = 4\hat{\phi} - \pi_{AA} - 1$. 2868
2.2 Analytical framework for efficient computation

We propose a general approach for computationally efficient relationship inference as follows. First, we derive an identity (details in the Supplementary Material) to represent the genetic distance between a pair of individuals in terms of their shared genotype counts

\[(X_i^0 - X_j^0)^2 = 4M_{AA} - 2M_{Aa} + I_{AA} + I_{Aa} \]

where \(I_{AA}\) and \(I_{Aa}\) indicate whether the i-th individual is heterozygous, whether both individuals are heterozygous, and whether the two individuals have different homozygotes, respectively. Now, we rewrite Equation (5) in terms of genotype counts

\[\phi_{ij} = \frac{N_{AA} - 2N_{Aa}}{2N_{ij}} \left(1 - \frac{N_{i}^{(0)} + N_{j}^{(0)}}{4N_{ij}}\right) \]

where \(N_{AA}, N_{Aa}\) and \(N_{ij}^{(0)}\) are the total numbers of SNPs at which both individuals are of the same genotype, and the total number of heterozygotes for the i-th and j-th individual, respectively, excluding those SNPs with missing genotypes in either individual of the pair.

When each genotype is stored in two bits, \(N_{ij}^{(0)}\) can be computed using only bit operations (i.e. AND, OR, XOR and NOT), eliminating multiplication and division during the process of scanning the genome. For KING-homo, further computational savings is achieved by pre-calculating \(H_{ij}\) at all SNPs prior to an identity (details in the Supplementary Material), and then using to reflect the set of observed genotypes used in analysis of each pair of individuals.

2.3 Robust relationship inference in the presence of population substructure

A key assumption underlying KING-homo and other existing methods (e.g. relatedness estimation of parent-offspring pairs) is that a pair of individuals is both related and from the same population. In most datasets, relative pairs are sampled from the same population, and pairs from different populations are unrelated. In both situations, the robust estimator given in Equation (9) is a consistent estimator for either the kinship coefficient or a measure of population heterogeneity. It is possible that a pair of individuals is both related and from different populations, e.g. one or both individuals are mixed, in which case the robust estimator is no longer a consistent estimator of the kinship coefficient. In this scenario, we have observed that this impact is rather small for the specification of relatives up to the 3rd-degree.

The only assumption required for our robust estimator of kinship coefficient (Equation 9) is HWE among SNPs with the same underlying allele frequencies. In practice, there is small proportion of individuals deviating from the HWE, due to reasons such as genotyping errors, recent admixture in a mixed population, or removing Mendelian errors in families. When the violation of HWE is not too large, the robust estimator (Equation 9) can over-estimate the kinship coefficient. In order to guard against potential estimation inflation due to departure from individual-level HWE, we consider the smaller of the observed heterozygosity rates, min \((N_{AA}/M_{AA}, N_{Aa}/M_{Aa})\), as an alternative to \(E(2(1-P))\). Without loss of generality, suppose the i-th individual has lower heterozygosity than the j-th individual. Then, the robust estimator is

\[\phi_{ij} = \frac{1}{4} \sum_{l} \left(\frac{N_{l}^{(0)} - N_{l}^{(0)}}{N_{l}^{(0)}} \right)^2 = \frac{N_{AA} - 2N_{Aa}}{2N_{ij}} - \frac{1}{2} \frac{N_{i}^{(0)} + N_{j}^{(0)}}{4N_{ij}} \]

The estimator above is no larger than the estimator in Equation (9), and both estimators are bounded above by 0.5.

We use estimator (Equation 9) for within-family relationship checking and estimator (Equation 11) for between-family relationship checking, naming this combined approach KING-robust. Using KING-robust, individuals of different ethnicities are less likely to be misspecified as relative pairs. All relationships in Table 1, except for the two types of 1st-degree relationships, can be uniquely specified through the kinship estimates provided by KING-robust. To further distinguish parent-offspring from full-sib pairs, we examine the observed BHS making use of the fact that BHS between a parent–offspring pair is always 1 or 2 at any SNP in the absence of genotyping errors. More advanced inference of pedigree structure can be carried out by simultaneously using the information from multiple pair-wise relationships.

The HWE assumption as in Equation (8) also allows estimation of the variance of allele frequencies in each individual

\[\text{Var}(P) = E(P^2) - E(P)^2 = P(AA) - \frac{1}{4} E(X)^2\]
Together with estimation of the allele frequency mean in each individual

\[ E(P) = \frac{1}{2} E(X) = \frac{1}{2} E(X) \]  

the population structure in a GWAS dataset can be resolved, even in the presence of unspecified family structure. Thus, our approach (Equations 12 and 13) provides a useful tool for population structure analysis in the context of a family-based GWAS.

3 RESULTS

3.1 Resolution of relationship inference varies with genotyping density

We performed simulations to demonstrate the resolution of kinship coefficient estimation using high-throughput genotype data. We simulated 1000 three-generation pedigrees that contained 1st, 2nd and 3rd-degree relative pairs. The detailed algorithms of simulating pedigrees were shown previously (Chen and Deng, 2001; Chen et al., 2009). SNPs from 22 autosomes with varying densities (50k, 150k and 500k) were simulated, with minor allele frequencies ranging (randomly) from 0.1 to 0.5.

We first examined the distribution of actual or realized IBD sharing (versus the estimated distribution) between relative pairs that is defined as half of the actual proportion of the genome that is shared IBD between the pair of relatives. The actual IBD sharing between a pair of relatives varies around its expectation except parent–offspring and monozygotic twin pairs (Visscher et al., 2008). This estimator is expected to provide an upper bound on precision for estimators of IBD-sharing statistics based on the same set of SNP data. Note that the realized IBD sharing of unrelated pairs is a constant, zero. We examined the distribution of the estimated kinship coefficients using the robust estimator (Equation 9). The distributions of realized IBD sharing with 150k SNPs, and estimated kinship coefficients with 150k SNPs, 5k SNPs and 500k SNPs are shown in Figure 1.

With 150k independent SNPs, the distribution of the realized IBD sharing and estimated kinship coefficients of relative pairs is rather similar, showing our robust kinship coefficient estimation achieves optimal power to classify relative pairs. However, even with the best possible estimation (e.g., the true value without estimation), pair-wise relationship misspecification can be still observed in GWAS data, especially for relative pairs of 3rd-degree and more distant. In all simulations, there is no distribution overlap between unrelated pairs and 3rd-degree relatives, and there is slight overlap between 3rd- and 2nd-degree relatives. The SDs for the estimated kinship coefficient of the 1st-, 2nd- and 3rd-degree relative pairs are 0.021, 0.017 and 0.012, respectively, in a GWAS with 150K SNPs. The SD for the estimated kinship coefficient of unrelated pairs is 0.002 (with 150K SNPs) or much less with denser SNPs. With a denser SNP panel, distributions between unrelated and related pairs are more separate, but its impact on the distribution for closely related pairs (up to 3rd-degree) is limited. In a linkage dataset with ~5k SNPs, only closely related pairs (up to 2nd-degree) and unrelated pairs can be estimated reliably, and there is noticeable overlap of distributions between 3rd-degree and unrelated pairs. In linkage datasets, there could be millions of unrelated pairs and, therefore, it is not feasible to correctly distinguish 3rd-degree relatives from unrelated pairs. However, a linkage dataset is still valuable for detection of 1st- and 2nd-degree relative pairs among millions of unrelated pairs, frequently ignored in current analysis of linkage data. Dense SNP data were also simulated for over one million unrelated pairs (data not shown), and the robust estimate of the kinship coefficient never exceeded 0.022.

3.2 Robust relationship inference in the presence of population stratification

We illustrate our robust relationship inference through application to data from the 269 HapMap (International HapMap Consortium, 2005). The HapMap data used in this study consisted of 30 CEU trios, 30 YRI trios, 45 CHB samples and 44 JPT samples. Each individual is genotyped at ~3 million SNPs in the consensus Phase II HapMap data with an average genotype missing rate 1.5% (note that ~20% of SNPs are not polymorphic in each population). Potential pedigree errors can be viewed easily through graphical displays, in which the inferred kinship coefficients are plotted against the estimated probability of zero IBD (or proportion of zero IBS). Algorithms assuming a homogeneous population perform poorly to estimate the kinship coefficients (Fig. 2C–F), systematically inflating the degree of relatedness among individuals of the same racial group. The kinship coefficients (KING-homo) of unrelated CEU pairs within-families are estimated to be ~0.044, so they are all incorrectly inferred to be 3rd-degree relative pairs.

Estimation of between-family pairs is much worse. Many unrelated pairs between families are misspecified as 2nd-degree relatives using both algorithms (Fig. 2D and F), and a large proportion of unrelated pairs are misspecified as 1st-degree relatives in PLINK (Fig. 2F). In contrast, KING-robust gives clean results even in the presence of population stratification (Fig. 2A and B), with kinship coefficient estimates consistent with those from the stratified data (data not shown). All algorithms identify relatedness across three pairs of YRI trios; the closest relationships in these three pairs of families are 1st-, 2nd- and 3rd-degree relatives, respectively.

We compare the performance of algorithms in KING to identify the population structure with the PCA algorithm (Price et al., 2006; Zhu et al., 2008). Figure 3 demonstrates three clear clusters in the
Robust relationship inference in GWAS

3.3 Robust relationship inference in a real GWAS

We further used the KING algorithms to screen pedigree errors in a GWAS of otitis media (Daly et al., 2004). This dataset includes 602 individuals from 143 families in which each is genotyped at 350K SNPs. The majority of individuals are Caucasian, one family of size four is Asian, one family of size two is Native American, and a few families have mixed ethnicity. Overall, we detect a higher degree of relatedness through analysis of genotype data compared to the relationships formally reported for the study. We detected 14 relationship errors within a family that are due to mis specification of one individual, and two disconnected families that are related (data not shown).

After fixing these two sets of errors, we display the inferred relationships in Figure 4A,B. We also applied the KING-homo and PLINK, both of which assume a homogeneous population (Fig. 4C–F). For this relatively homogeneous dataset, different algorithms give similar results for the majority of pairs. By all three algorithms, 14 pairs of individuals from three unrelated sibships are estimated to be 3rd or 4th-degree relatives, and 2 out of 33 formally reported 3rd-degree relatives are misspecified as 4th-degree relatives (due to the limited power). Note the two pairs of 4th-degree relatives are correctly specified. In addition, PLINK reports six additional unrelated pairs as related (kinship $> 0.022$), while both KING algorithms clearly separate related pairs from unrelated pairs. KING-homo overestimates the kinship coefficient of a pair of unrelated Asian parents (Fig. 4C), which is expected given that Asians have a lower heterozygosity than other individuals (Note that $2\hat{H}_{ij} > N_A^{(i)} + N_A^{(j)}$ implies the kinship estimate in Equation 5 is larger than the one in Equation 9). The population structure of this data is shown in an allele frequency plot (Supplementary Fig. 1B) as well as in principal component plots (Supplementary Fig. 1C and D).

3.4 Computational efficiency: minutes rather than days

We compared the analysis time between the KING algorithms (both the robust algorithm and the one that assumes population homogeneity) with the algorithm implemented in PLINK, in the...
Computation time of two software implementations to estimate kinship coefficients in three sets of GWAS SNP data

Table 2.

| Index | No. of SNPs | No. of samples | No. of pairs | KING | PLINK |
|-------|-------------|---------------|--------------|------|-------|
| 1     | 3 079 857   | 269           | 36 046       | 2 min | 2 h 9 min |
| 2     | 324 748     | 602           | 180 901      | 1 min | 1 h 13 min |
| 3     | 549 338     | 2 454         | 3 009 832    | 25 min | 28 h 30 min |

The computation time refers to the time to estimate kinship coefficients for all pairs of individuals, excluding overhead costs such as the time to load data into the computer memory. The two KING implementations (the robust algorithm and the algorithm assuming homogeneous samples) took a similar amount of computational time. This computation time can be estimated reliably as the analysis time for the entire data minus the analysis time for only the within-family data. The unit of computation time is in minutes. All computation was performed on an Intel Xeon with 3.20 GHz processor.

above two datasets as well as an additional GWAS dataset consisting of 2450 individuals (Table 2).

In all three datasets that we examined, the computational time of the KING implementation is in minutes; in contrast, it took hours to days to analyze the same datasets using PLINK on the same workstation. The computational saving of our implementation over PLINK is over 60-fold, and this result can be generalized to larger datasets. This computational efficiency makes our implementation particularly attractive for the analysis of large GWAS datasets that exceed 10 000 individuals (hours of computation time in KING compared to a projected >1 month computation time using other software), making it feasible to perform the millions of pairwise comparisons necessary for a comprehensive between-family analysis.

Note the reported computation time in Table 2 excludes time to load the data, which grows linearly as the sample size increases, in contrast to the exponential increase of analysis time. The time to load the data in KING was <30 s for all three scenarios when binary format genotypes were used as the input of the KING implementation, and was comparable to other software implementations (21 m, 6 m and 40 m, respectively) when the MERLIN (Abecasis et al., 2002) format genotypes were used as the input.

4 DISCUSSION

We have proposed a robust algorithm to infer relationships using high-density genotype data from a GWAS. Our approach to relationship inference incorporates simple estimates for key genetic parameters, reported with high precision due to the large number of SNPs typed by current high-throughput panels. The framework underlying the KING approach to relationship inference centers on modeling genetic distance between a pair of individuals as a function of their allele frequencies and kinship coefficient. In studies with homogeneous populations and relatively large sample sizes, allele frequencies at all SNPs can be estimated accurately from the given data, and used to inform the estimate of allelic heterogeneity needed to calculate the kinship coefficient in KING-homo. Under population stratification, a single set of allele frequencies for the given SNP panel is not appropriate for examination of the entire dataset, motivating our use of the robust estimator in KING-robust.

As demonstrated by our power analysis and application to the otitis media data, our approach based on estimation of the kinship-coefficient between any pair of individuals is sufficient to classify relative pairs as monozygotic twins, parent-offspring pairs, full sibs, 2nd- or 3rd-degree relatives. Unlike approaches that assume a homogeneous population, our robust approach classifies relative pairs correctly even under extreme population stratification seen in the pooled HapMap data. Our relationship inference is not impacted by the linkage disequilibrium structure among adjacent SNPs according to the large sample theory, and as demonstrated in the two GWAS analyses.

The robust algorithm in KING performs pair-wise relationship inference using only information from the two individuals under comparison. The inference is invariant to inclusion of any additional samples and to use of different SNP panels, producing reliable results using genotypes from GWAS or from studies of rare variants alone. The sample size of the data can be as small as two, and the analysis can be performed rapidly for a single pedigree or pair of individuals, with a wide range of applications, including forensic DNA analysis and paternity/maternity testing (assuming the current forensics technology transitions to high-density SNP genotyping). The ability to perform between-family relationship inference robust to population structure also allows population structure analysis.

Fig. 4. Relationship checking in OM GWAS data. (A), (C) and (E) are within-family relationship checking using three algorithms, and (B), (D) and (F) are between-family relationship checking using three algorithms. Negative kinship coefficient estimates are truncated to 0.
without the worry of spurious principal components produced by undetected family structure. Ultimately, the combination of robust inference and rapid computation can be applied toward automated pedigree reconstruction and association mapping in the absence of any pre-specified pedigree or population structure (Chen and Abecasis, 2007; Chen et al., 2009; Chei et al., 2009; Kang et al., 2010, Thornton and McPeek, 2010; Zhang et al., 2010).

The KING algorithms (robust and homo) for relationship inference have been implemented in a user-friendly software package. KING is able to process large-scale GWAS data consisting of thousands of individuals (a few minutes to check all pair-wise relationships for millions of pairs of individuals). Tools to detect population structure in the presence of genetic relatedness, including a modified PCA algorithm (Zhu et al., 2008) and allele frequency statistics (Equations 12 and 13) have also been implemented in KING to facilitate the analysis of GWAS data. Future toolsets include relationship inference between two groups of individuals (rather than two individuals), clustering samples in families and reconstructing pedigrees, automatic pedigree error fixing, robust PCA structure analysis and genome-wide association analysis in the presence of unknown genetic relatedness in the sample.

ACKNOWLEDGEMENTS

We thank Gonçalo Abecasis for sharing C++ source code for the KING implementation, Xuanlin Hou for assistance in preparing the Otitis Media GWAS data, and three anonymous reviewers for valuable input that improved the article.

Funding: Research grant DC003166 (to K.D., in part).

Conflict of Interest: none declared.

REFERENCES

Abecasis, G.R. et al. (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet., 30, 97–101.

Boichik, M. and Cox, N.J. (1997) Accurate inference of relationships in sub-pair linkage studies. Am. J. Hum. Genet., 61, 423–429.

Chen, W.M. and Abecasis, G.R. (2007) Family-based association tests for genomewide association scans. Am. J. Hum. Genet., 81, 913–926.

Cheng, W.M. and Deng, H.W. (2001) A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. Genet. Epidemiol., 21, 53–67.

Chen, W.M. et al. (2009) A generalized family-based association test for dichotomous traits, Am. J. Hum. Genet., 85, 364–376.

Choi, Y. et al. (2009) Case-control association testing in the presence of unknown relationships. Genet. Epidemiol., 33, 668–678.

Daly, K.A. et al. (2004) Chronic and recurrent otitis media: a genome scan for susceptibility loci. Am. J. Hum. Genet., 75, 988–997.

International HapMap Consortium (2005) A haplotype map of the human genome. Nature, 437, 1299–1320.

Kang, H.M. et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nat. Genet., 42, 348–354.

Lynch, M. and Ritland, K. (1999) Estimation of pairwise relatedness with molecular markers. Genetics, 152, 1755–1766.

McPeek, M.S. and Sun, L. (2000) Statistical tests for detection of mispecified relationships by use of genome-screen data. Am. J. Hum. Genet., 66, 1076–1094.

Purcell, S. et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet., 38, 904–909.

Purcell, S. et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet., 81, 559–575.

Thornton, T. and McPeek, M.S. (2010) ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am. J. Hum. Genet., 86, 172–184.

Visscher, P.M. et al. (2008) H inferability in the genomics era—concepts and misconceptions. Nat Rev Genet., 9, 255–266.

Yang, J. et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat. Genet., 42, 565–569.

Zhang, Z. et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat. Genet., 42, 355–360.

Zhu, X. et al. (2008) A unified association analysis approach for family and unrelated samples correcting for stratification. Am. J. Hum. Genet., 82, 352–365.