nature genetics

Variance component model to account for sample structure in genome-wide association studies

Hyun Min Kang[1,2,8], Jae Hoon Sul[3,8], Susan K Service[4], Noah A Zaitlen[5], Sit-yee Kong[4], Nelson B Freimer[4], Chiara Sabatti[6] & Eleazar Eskin[3,7]

Although genome-wide association studies (GWASs) have identified numerous loci associated with complex traits, imprecise modeling of the genetic relatedness within study samples may cause substantial inflation of test statistics and possibly spurious associations. Variance component approaches, such as efficient mixed-model association (EMMA), can correct for a wide range of sample structures by explicitly accounting for pairwise relatedness between individuals, using high-density markers to model the phenotype distribution; but such approaches are computationally impractical. We report here a variance component approach implemented in publicly available software, EMMA eXpedited (EMMAX), that reduces the computational time for analyzing large GWAS data sets from years to hours. We apply this method to two human GWAS data sets, performing association analysis for ten quantitative traits from the Northern Finland Birth Cohort and seven common diseases from the Wellcome Trust Case Control Consortium. We find that EMMAX outperforms both principal component analysis and genomic control in correcting for sample structure.

GWASs may utilize either case-control cohorts to test for associations with diseases or population cohorts to identify associations with quantitative traits. In both cases, it is assumed that the cohorts consist of unrelated individuals that share the same population background, although this may not hold in practice for cohorts used in many current GWASs. The presence of related individuals within a study sample results in sample structure, a term that encompasses population stratification and hidden relatedness. Population stratification refers to the inclusion of individuals from different populations within the same study sample. Hidden relatedness refers to the presence of unknown genetic relationships between individuals within the study sample[1,2]. The effects of sample structure present in cohorts used for genetic association studies have been well documented and identified as a cause for some spurious associations[3,4].

Although limiting study samples entirely to unrelated individuals may be difficult or impossible, genotype data provides valuable information on the sample structure that can inform genetic association analysis. For example, the STRUCTURE software[5] uses genotype data to partition the sample into subpopulations within which there is no sample structure and subsequently carries out association tests within the identified subpopulations. To eliminate the effects of hidden relatedness, one can estimate the proportion of genes identical by descent (IBD) between any pair of individuals in the sample and exclude from the analysis those individuals that appear closely related[1,6]. Population stratification and hidden relatedness, however, constitute just two extreme manifestations of sample structure, and methods are needed to correct for other forms of sample structure. In the genomic control approach[7,8], which has been widely adopted, the distribution of test statistics from the single-marker analysis is used to estimate the inflation factor, λ, with which the test statistics are subsequently rescaled, constraining the risk of false positives. The EIGENSTRAT software[9,10] uses principal components analysis (PCA) to detect and describe sample structure and has been widely used in GWASs. Some principal components may represent broad differences across individuals within a given data set, effectively capturing a few major axes of population structure, but it is unclear how to interpret the rest of the principal components as surrogates of sample structure[11,12]. Currently, association studies typically use a combination of these strategies, first identifying close relatives to remove them from analysis, then correcting for broad sample structure using principal components or spatial information and finally correcting for the residual inflation with genomic control[6,13,14].

If we knew the complete genealogy of the population, we could, in principle, apply a variance component method to model the effects of the genetic relationships on the phenotypes; this approach would be similar in spirit to the classical polygenic model[15] directly applied to association mapping[16]. The variance component would capture the complex mixture of both population stratification and hidden relatedness that directly results from the genealogy and would correct for these relationships during the mapping. Although the exact genetic

[1]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA. [2]Center for Computational Medicine and Bioinformatics, The University of Michigan Medical School, Ann Arbor, Michigan, USA. [3]Computer Science Department, University of California, Los Angeles, California, USA. [4]Center for Neurobehavioral Genetics, University of California, Los Angeles, California, USA. [5]Department of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [6]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, USA. [7]Department of Human Genetics, University of California, Los Angeles, California, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to C.S. (sabatti@stanford.edu) or E.E. (eeskin@cs.ucla.edu).

Received 23 July 2009; accepted 9 February 2010; published online 7 March 2010; doi:10.1038/ng.548

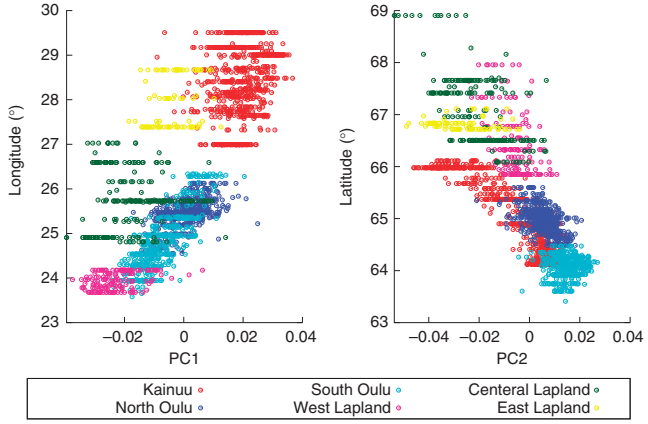

Figure 1 Scatter plots of the first two principal components against latitude and longitude. Only individuals of known ancestry are included in the plot. Latitude and longitude are defined as the average latitude and longitude of the parents' birthplaces. Colors indicate linguistic or geographic subgroups.

relationships between individuals in the samples are unknown, we could take advantage of the high-density genotype information to empirically estimate the level of relatedness between reportedly unrelated individuals.

We report here an approach for correcting for sample structure within GWASs, based on a linear mixed model (also sometimes referred to as a mixed linear model) with an empirically estimated relatedness matrix to model the correlation between phenotypes of sample subjects. Similar variance component approaches have been used successfully in animal models[17–19]. However, applying even an efficient implementation of a variance component approach, such as EMMA (ref. 19), is computationally intractable for data sets consisting of thousands of individuals, owing to the heavy computational burden in the estimation of variance parameters. Capitalizing on the characteristics of complex traits in humans, we make a few simplifying assumptions that allow us to markedly increase the speed of computations, making our approach readily applicable to GWASs with tens of thousands of individuals assayed at hundreds of thousands of SNPs. For most genetic association studies in humans, because the effect of any given locus on the trait is very small[20], we need to estimate the variance parameters only once for each data set, and we can globally apply them to each marker. Our computational improvements reduce the running time for the analysis of a typical GWAS data set using a variance component model from years to hours. The advantage of the variance component approach is that the empirical relatedness matrix encodes a wide range of sample structures, including both hidden relatedness and population stratification. Principal component–based methods, in contrast, by estimating major axes of the pairwise genetic similarity matrix, capture some, but not all, of the sample structure, as we show below.

We evaluate our method using two human GWAS data sets, from the 1966 Northern Finland Birth Cohort (NFBC66)[13,21] and the Wellcome Trust Case Control Consortium (WTCCC)[6]. The NFBC66 is based on a founder population, which is expected to minimize genetic heterogeneity, increasing the chances of mapping genes underlying traits of interest[22]. This is an ideal sample to evaluate our method because a detailed study[23] of this data set has revealed the presence of substantial population structure that could influence the results of genetic association studies. In addition, we apply our method to the case-control studies for seven common complex diseases conducted by the WTCCC[6]. In both data sets, our method consistently outperforms both genomic control and principal component analysis. We term our method EMMA eXpedited (EMMAX) because it builds on the previous approach EMMA (ref. 19) and markedly reduces the computational cost.

RESULTS

Revisiting principal component analysis in NFBC66

To more closely examine the extent of sample structure within the NFBC66, we used PCA of the genotype covariance matrix[9] and multidimensional scaling analysis (MDS) of the identity-by-state (IBS) matrix from NFBC66 samples. The first two coordinates identified by MDS are known to correlate well with the geographical location of the linguistic groups[13]. The first two principal components in the current sample correlate well with latitude and longitude of parental birthplaces for the subset of individuals with known ancestry (**Fig. 1**). Indeed, we noted that PCA of genotypes and classical MDS of the IBS matrix lead to very similar results. There is a correlation coefficient of 0.9993 between the first components from PCA and MDS and a correlation coefficient of 0.9978 between the second components. The first five principal components separate to varying degrees the linguistic and geographic subgroups comprising northern Finland (**Supplementary Fig. 1**), consistent with the previous analysis using MDS[13]. Despite the clear correlation between geographical regions of origin and the first two principal components, clustering analyses of the IBS matrix using PLINK software or hierarchical clustering in R did not identify separate subgroups.

Association analysis

Performing a simple uncorrected association test for each of the nine phenotypes originally examined in ref. 13, we made the following estimates of the genomic control parameters λ: body mass index, 1.031; C-reactive protein (CRP), 1.007; diastolic blood pressure, 1.031; glucose, 1.045; high-density lipoprotein (HDL), 1.052; insulin plasma levels, 1.029; low-density lipoprotein, 1.098; systolic blood pressure, 1.066; triglyceride, 1.023. These values are all higher than the ones obtained previously with a smaller sample size[13] and are substantially higher than what one would expect in a sample with no structure. In addition, the height phenotype, which was not analyzed in the previous study[13], has a λ value of 1.187. For reference, note that a conservative estimate of the 95% confidence interval of the inflation factor is between 0.992 and 1.008, assuming independence between the markers.

Table 1 Comparison of genomic control inflation factors obtained with different models

	Genomic control inflation factor			
Phenotype	Uncorrected	IBD < 0.1	ES100	EMMAX
CRP	1.007	1.007	1.019	0.993
TG	1.023	1.010	1.019	1.002
INS	1.029	1.022	1.013	1.005
DBP	1.031	1.019	1.028	1.007
BMI	1.031	1.024	1.016	0.995
GLU	1.045	1.033	1.030	1.008
HDL	1.052	1.056	1.036	1.004
SBP	1.066	1.056	1.021	1.006
LDL	1.098	1.089	1.040	1.002
Height	1.187	1.151	1.074	1.003

ES100, EIGENSOFT correcting for 100 principal components; IBD < 0.1, uncorrected analysis after excluding 611 individuals whose PLINK's IBD estimates with another individual is greater than 0.1; phenotype abbreviations are CRP, C-reactive protein; TG, triglyceride; INS, insulin plasma levels; DBP, diastolic blood pressure; BMI, body mass index; GLU, glucose; HDL, high-density lipoprotein; SBP, systolic blood pressure; LDL, low density lipoprotein.

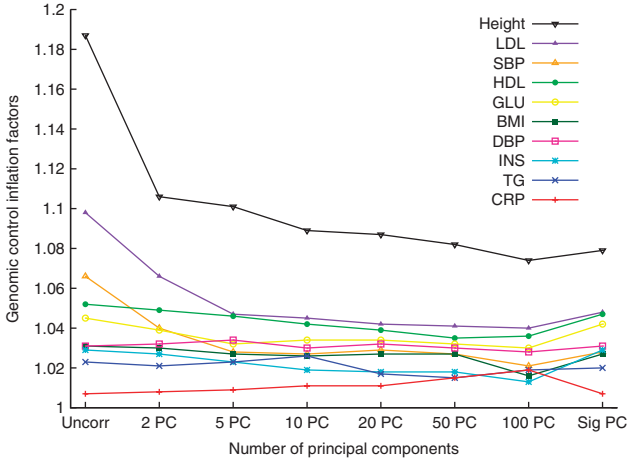

Figure 2 The genomic control parameters for ten traits change with the number of principal components used for adjustment. Sig PC, significant principal components, includes the principal components (PC) that have a *t*-test *P* value < 0.005 as predictors for each of the phenotypes. LDL, low density lipoprotein; SBP, systolic blood pressure; HDL, high-density lipoprotein; GLU, glucose; BMI, body mass index; DBP, diastolic blood pressure; INS, insulin plasma levels; TG, triglyceride; CRP, C-reactive protein.

As hidden relatedness is a possible cause of inflated genomic control parameters, we reanalyzed the data after excluding a larger number of possibly related subjects (a genome-wide IBD estimate of >10% was used as a cutoff with PLINK software, excluding an additional 611 individuals). This resulted in a slight reduction of λ for some phenotypes (**Table 1**).

As suggested in ref. 9, we explored the effect of including a variable number of principal components in the association tests. Although including two or five principal components are included has a considerable effect on the λ values, further augmenting the number of principal components does not substantially decrease the genomic control parameter (**Fig. 2**). It is often suggested that only principal components having predictive power for the phenotype should be included in the regression[11]. We identified principal components for each phenotype that have a *t*-test $P < 0.005$ as predictors; the results of their inclusion in the association tests are reported in **Figure 2**.

Correcting for sample structure

We analyzed the ten NFBC66 phenotypes with EMMAX using a three-step procedure (see Online Methods). First, we computed a pairwise relatedness matrix from high-density markers, which we used to represent the sample structure. Second, we estimated the contribution of the sample structure to the phenotype using a variance component model, resulting in an estimated covariance matrix of phenotypes that models the effect of genetic relatedness on the phenotypes. Third, we applied a generalized least square (GLS) *F*-test[24], or alternatively a score test[25], at each marker to detect associations accounting for the sample structure using the covariance matrix.

The second step also provides us with the fraction of phenotypic variance explained by the empirically estimated relatedness matrix. We call this fraction pseudoheritability because it resembles the heritability estimated from a pedigree[26], although this is not directly interchangeable with heritability of the trait because the estimated pairwise relatedness does not correspond exactly to the kinship coefficients. Nonetheless, the pseudoheritability estimates are concordant with the previous heritability estimates from a large family based study of Kosrae and Sardinian populations[27,28]. Different methods for estimating the pairwise relatedness provide slightly different but highly correlated estimates of pseudoheritability across the ten traits. (**Supplementary Table 1**).

Using the estimated covariance matrix, we proceeded with the GLS *F*-test to test the effect of each marker on the phenotype and then applied genomic control to quantify the amount of residual inflation. The genomic control λ parameters we obtained with EMMAX are much lower than those obtained using either standard association methods or regression analysis including 100 principal components (**Table 1**). **Figure 3** and **Supplementary Figure 2** illustrate the results using quantile-quantile plots of the *P* value distributions from these three tests. Only one of the ten phenotypes showed λ values within the 95% confidence interval of 0.992–1.008 with uncorrected or principal component analysis, whereas all of them fell in the confidence interval with EMMAX.

Unlike genomic control, the EMMAX model alters the ranking of SNPs by their association statistics. This is especially important as many recent GWAS follow-up and multistage design studies take the approach of genotyping all SNPs exceeding some predefined threshold[29–31]. We examined the extent to which the adoption of the EMMAX model changes the SNP rankings in comparison to the uncorrected and principal component analyses. We took the top *k* markers from the results of EMMAX, the uncorrected method, and regression including 100 principal components (as implemented in EIGENSOFT software), for *k* between 10 and 5,000. For each of these sets, we calculated the number of SNPs shared between the lists and the fraction of these shared SNPs relative to the number of unique SNPs in each pair of lists. Although many of the top SNPs reported by each method overlap, a considerable number of highly ranked SNPs differ between the methods (**Fig. 4** and **Supplementary Table 2**). In general, EMMAX results are similar to uncorrected analysis when the inflation of test statistics is small, but they become more similar to the PCA as the inflation increases. Notably, the PCA consistently shows larger departures from the uncorrected analysis than EMMAX does across all ten phenotypes. For example, when the overdispersion

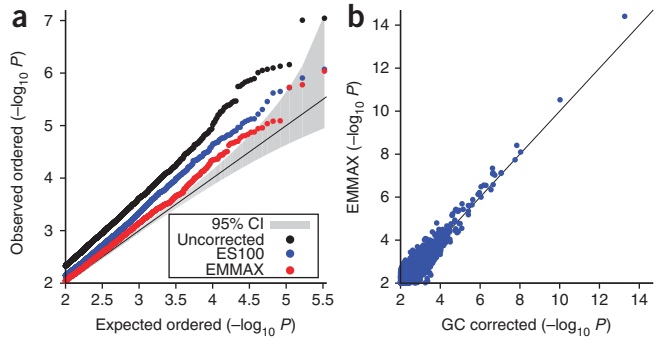

Figure 3 Comparison of *P* value distributions across different methods with NFBC66 data. (**a**) Quantile-quantile plot of the height phenotype, which shows the largest inflation of test statistics, before application of genomic control. The shadowed region represents a conservative 95% confidence interval (CI) computed from the beta distribution assuming independence markers. ES100 indicates EIGENSOFT correcting for 100 principal components. (**b**) Comparison of LDL association *P* values between uncorrected and EMMAX analysis after application of genomic control in a logarithmic scale.

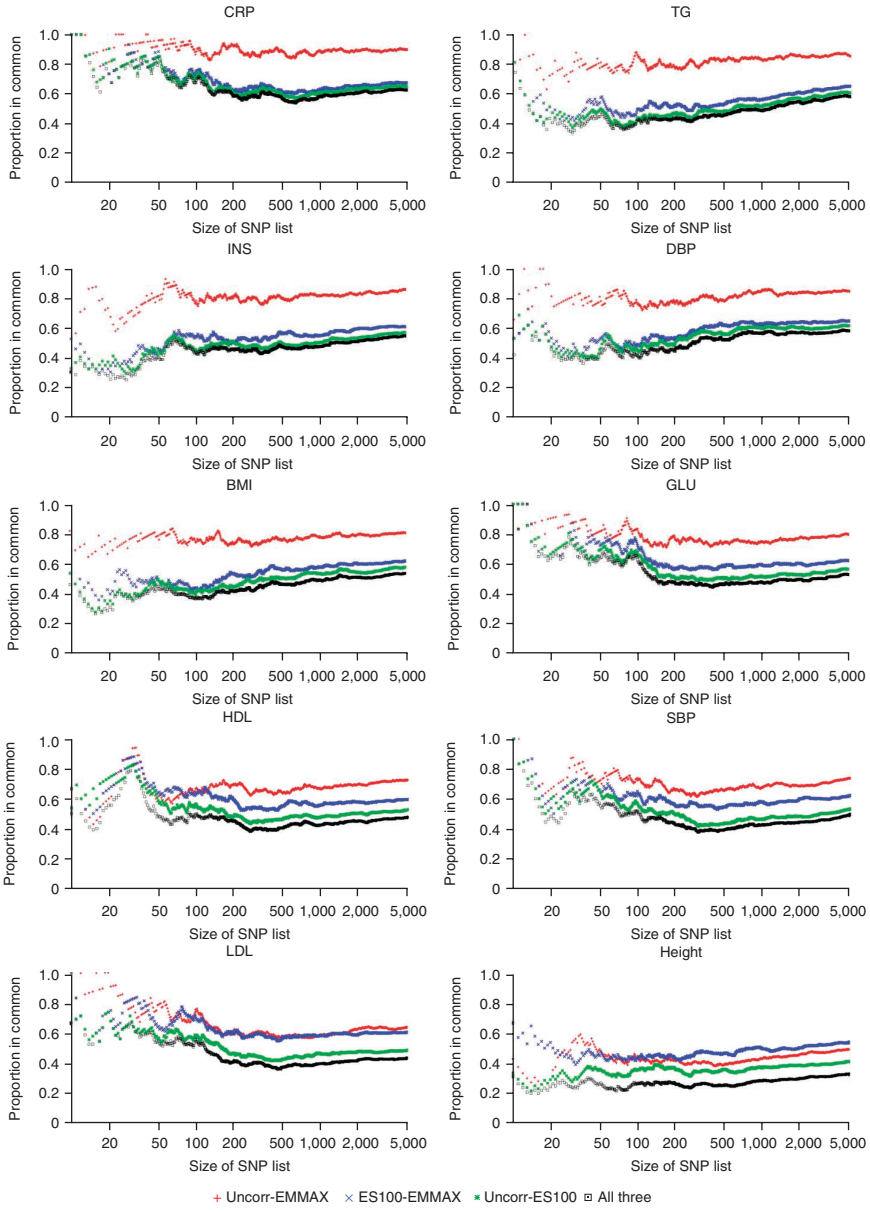

Figure 4 Rank concordance comparison of strongly associated SNPs between different methods. The ten NFBC66 phenotypes (abbreviated as in **Fig. 2**) are ordered by their genomic control inflation factors. Rank concordance is presented as CAT plots[45]. The proportion of SNPs shared between sets of the top k SNPs for different methods are shown for $10 \leq k \leq 5000$. Pairs of sets being compared are indicated in key at bottom; for example, Uncorr-EMMAX, comparison of uncorrected set and EMMAX set. ES100 indicates EIGENSOFT correcting for 100 principal components.

of test statistics was negligible, such as in the CRP phenotype, only 66% of the top 2,000 hits were concordant between the principal component and the uncorrected analysis, whereas 89% were concordant between EMMAX and the uncorrected analysis.

EMMAX prevents the overdispersion of test statistics using a statistical model that explicitly takes into account sample structure, rather than correcting the overdispersed test statistics caused by not taking into account genetic relatedness in the statistical model. Consequently, EMMAX can also prevent the overcorrection that would remove true positive associations. We identified 15 genome-wide significant loci with at least one of the uncorrected, 100 principal components–corrected, or EMMAX, analyses after genomic control at the suggested P threshold[32] of 7.2×10^{-8} across the ten phenotypes (**Table 2**). In 13 of the 15 loci, EMMAX P values become smaller than the uncorrected analysis. The two-sided binomial P value of the observed asymmetry is 9.8×10^{-4} if two methods have the same statistical power. With the 100 principal components–corrected analysis, 10 of the 15 loci show smaller P values than the uncorrected analysis (binomial P value of 0.12). Although 12 of the 15 loci are found by all methods to be genome-wide significant at $P < 7.2 \times 10^{-8}$, two known loci[33], APOB (with triglyceride) and HNF4A (with HDL), pass the threshold only with EMMAX. In contrast, the locus NR1H3 (with HDL), which is genome-wide significant only with uncorrected analysis, turns out to be the only locus whose association has not yet been replicated by an independent study among the 15 loci.

Because EMMAX estimates the variance parameters under the null hypothesis, one may suspect that it is underpowered compared to the full mixed model, which estimates the variance parameters under the alternative hypothesis. This is comparable to the difference between the score statistic and the efficient score statistic[25,34,35]. As most genetic variants associated to date with human complex traits are estimated to explain only a small fraction of phenotypic variance[20], the difference between the two approaches will be negligible in most cases. To assess the seriousness of this concern, we ran the original EMMA, which uses a full mixed effect model, on the 15 peak SNPs and compared the resulting P values to those estimated with EMMAX using GLS. Overall, as expected, the P values from the full mixed effect model tended to be smaller than the P value from the GLS model, but the magnitude of the difference was very small (**Supplementary Fig. 3a**). However, the running times for EMMA were substantially longer. Because the original EMMA re-estimates the variance parameters at each marker, given the size of the NFBC data set, it took more than 10 min of CPU time per marker on an Intel Xeon 3-GHz processor, even with an efficient C implementation of EMMA. A simple extrapolation suggests that it would take more than 6 years of CPU time to analyze a single GWAS data set using EMMA, taking a full mixed model approach. The total computational time using EMMAX for this data was 6.6 h in a single CPU, and the procedure could easily be parallelized to speed it up further.

Application to Wellcome Trust Case Control Consortium data

We also applied our method to the WTCCC data set consisting of case-control studies for seven common diseases[6]. To analyze case-control phenotypes, we applied a linear model to the binary phenotypes, in the spirit of Armitage's test (see Online Methods). We performed association testing over the seven disease phenotypes using EMMAX, EIGENSTRAT and uncorrected analysis. The values we observed for inflation factors λ were very similar to those in the original study,

Table 2 Fifteen peak associated SNPs with genome-wide significance

Trait	rsID	Chr	Base position[a]	Closest gene	*P* value		
					Uncorrected + GC	ES100 + GC	EMMAX + GC
HDL	rs3764261	16	55550825	*CETP*	7.0×10^{-31}	3.8×10^{-31}	**3.7×10^{-32}**
CRP	rs2794520	1	157945440	*CRP*	4.8×10^{-23}	3.6×10^{-23}	**3.0×10^{-23}**
LDL	rs646776	1	109620053	*CELSR2*	5.4×10^{-14}	7.7×10^{-15}	**3.8×10^{-15}**
CRP	rs2650000	12	119873345	*LEF1*	2.1×10^{-12}	7.0×10^{-12}	**1.9×10^{-12}**
HDL	rs1532085	15	56470658	*LIPC*	**4.3×10^{-12}**	7.9×10^{-11}	1.0×10^{-11}
GLU	rs560887	2	169471394	*G6PC2*	1.1×10^{-11}	4.1×10^{-12}	**3.1×10^{-12}**
LDL	rs693	2	21085700	*APOB*	9.6×10^{-11}	**1.5×10^{-11}**	2.8×10^{-11}
TG	rs1260326	2	27584444	*GCKR*	1.9×10^{-10}	**5.9×10^{-11}**	1.8×10^{-10}
HDL	rs255049	16	66570972	*LCAT*	3.9×10^{-9}	**1.2×10^{-9}**	1.4×10^{-8}
LDL	rs11668477	19	11056030	*LDLR*	1.4×10^{-8}	3.2×10^{-8}	**4.1×10^{-9}**
GLU	rs2971671	7	44177862	*GCK*	1.8×10^{-8}	**1.7×10^{-9}**	1.6×10^{-8}
HDL	rs7120118	11	47242866	*NR1H3*[b]	**4.8×10^{-8}**	*6.6×10^{-5}*	*1.1×10^{-6}*
TG	rs10096633	8	19875201	*LPL*	2.0×10^{-8}	**1.1×10^{-8}**	1.9×10^{-8}
TG	rs673548	2	21091049	*APOB*	*8.0×10^{-8}*	*1.2×10^{-7}*	**6.4×10^{-8}**
HDL	rs1800961	20	42475778	*HNF4A*	*1.5×10^{-7}*	*9.5×10^{-8}*	**1.8×10^{-8}**

These SNPs had *P* values below the suggested[32] genome-wide significance threshold of 7.2×10^{8} in the uncorrected, the 100 principal components–corrected (ES100) or the EMMAX analysis after genomic control (+GC). Traits are HDL, high-density lipoprotein; CRP, C-reactive protein; LDL, low density lipoprotein; GLU, glucose; TG, triglyceride. rsID, reference SNP ID assigned by dbSNP; Chr, chromosome; boldface indicates the strongest *P* values across the three methods; italics indicate *P* values that did not surpass the significance threshold.
[a]Positions are based on National Center for Biotechnology Information build 36.1. [b]*NR1H3* is the locus whose association with HDL that has not yet been replicated by other independent studies.

in which the test statistics were uncorrected: bipolar disease, 1.11; coronary artery disease, 1.06; Crohn's disease, 1.10; hypertension, 1.06; rheumatoid arthritis, 1.03; type 1 diabetes, 1.04; and type 2 diabetes, 1.07. Consistent with our observations over the NFBC66 data, correcting for 100 principal components only partially reduced the inflation factors (**Table 3** and **Supplementary Fig. 2**). When EMMAX was applied, the estimated inflation factors were below the upper bound of the confidence interval, suggesting that none of the phenotypes show significant inflation of test statistics.

However, we noticed that two of the phenotypes, rheumatoid arthritis and type 1 diabetes, show significant deflation of test statistics beyond the 95% confidence interval ($\lambda = 0.965$ for rheumatoid arthritis, $\lambda = 0.946$ for type 1 diabetes). This is not unexpected, considering that a substantial fraction of the phenotypic variance in these autoimmune diseases is explained by the HLA loci, leading to inaccurate estimation of variance parameters under the null hypothesis when the HLA effect is not accounted for. In fact, the set of genome-wide significant SNPs ($P < 7.2 \times 10^{-8}$; ref. 32) in this region account for 47% and 60% of the phenotypic variance of rheumatoid arthritis and type 1 diabetes, respectively[6]. We re-estimated the variance parameters by conditioning on the 57 and 134 SNPs within the extended human MHC region[36] that explain more than 1% of phenotypic variance of rheumatoid arthritis and type 1 diabetes, respectively (as described in the **Supplementary Note**). As a result, the genomic control λ increased to 0.989 for rheumatoid arthritis and 0.991 for type 1 diabetes. We performed this conditioning procedure only for estimating variance parameters and not in the SNP association test so that the *P* values would be consistent with the unconditioned analysis. Conditioning on the SNPs with such a strong effect may further improve the power to identify novel loci. A more sophisticated conditional analysis—for example, one including haplotype effects or epistatic interactions into covariates—may also better account for the strong effects in the autoimmune diseases[37].

Marker-specific inflation factors

Under certain conditions, one can expect the variance of the test statistics to be inflated by a constant across the genome[7,8]. A formal model of hidden relatedness based on the coalescent theory[1] also suggests a constant inflation across the genome when the sample structure is entirely due to hidden relatedness[7]. However, for a more complex genealogical relationship among individuals, it is not clear how the inflation of test statistics will behave.

Using the same variance component framework, we developed a method to estimate the marker-specific inflation of test statistics using the correlation between each marker and the empirically estimated kinship matrix (described in the **Supplementary Note**). These estimates are concordant with the genome-wide genomic control inflation factor on average but show substantial differences across the SNPs (**Fig. 5a**). In the height phenotype, for example, the estimated marker-specific inflation factors have a mean of 1.107, s.d. of 0.090 and median value of 1.093. In light of this, we explored the relationship between marker-specific inflation factors and the overdispersion of test statistics with the uncorrected analysis. The distribution of height association *P* values for SNPs with inflation factors <1.05 shows a less marked departure from uniform distribution than does the distribution for SNPs with inflation factors >1.20 (**Fig. 5b**,**c**). Considering that SNPs with a higher inflation factor were identified without consideration of their possible association with the phenotype, it is reasonable to conclude that this excess of small *P* values reflects overdispersion of test statistics.

These results underscore how correcting the test statistics using a single inflation factor may be inappropriate, possibly reducing power and not sufficiently controlling for false positives. To further demonstrate this point, we ran a simple simulation using the variance component model on which EMMAX is based. Although simulating data under this model puts our method at an advantage, and the approach is therefore less suited for comparison to other models, it does demonstrate that under some circumstances uniformly deflating *P* values may be inappropriate. We randomly simulated 100 sets of phenotypes solely from the sample structure with no SNP effects and examined the quantile-quantile plots across different methods. Although the inflation for most of the SNPs is corrected by genomic

Table 3 Comparison of genomic control inflation factor obtained with different models in seven WTCCC phenotypes

Phenotype	Genomic control inflation factor		
	Uncorrected	ES100	EMMAX
BD	1.105	1.071	0.998
CAD	1.063	1.048	1.006
CD	1.098	1.055	1.000
HT	1.055	1.051	0.997
RA	1.028	1.031	0.965 (0.989[a])
T1D	1.043	1.028	0.946 (0.991[a])
T2D	1.065	1.042	0.996

ES100, EIGENSOFT correcting for 100 principal components; BD, bipolar disorder; CAD, coronary artery disease; CD, Crohn's disease; HT, hypertension; RA, rheumatoid arthritis; T1D, type 1 diabetes; T2D, type 2 diabetes.
[a]The variance component parameters (σ^2_a and σ^2_e) are estimated by conditioning on the large-sized SNP effects explaining 1% or more phenotypic variance.

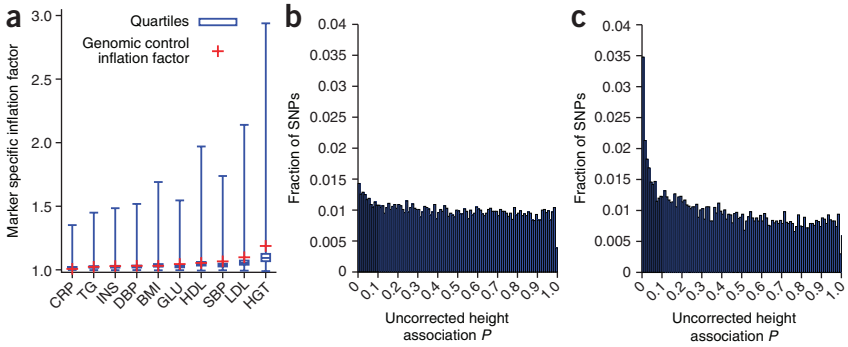

Figure 5 Distribution of the marker-specific inflation factors from NFBC66 data sets. (**a**) Box plots of the marker-specific inflation factors across ten phenotypes, in addition to the genomic control inflation factor for each phenotype. Abbreviations are as in **Figure 2**. (**b**,**c**) Distributions of *P* values of the height phenotype association when the estimated per-marker inflation factors are less than 1.05 (35,988 SNPs; **b**) and when they are greater than 1.2 (15,874 SNPs; **c**).

control as expected, we observed substantial fluctuations of the test statistics at the tail of the distribution (**Supplementary Fig. 4a**,**b**).

More than 25% of the phenotypes showed inflation or deflation beyond the 95% confidence interval. This is because the SNPs with higher per-marker inflation are not sufficiently corrected by the constant genomic control inflation factor. In contrast, EMMAX results in *P* values close to the expected distribution (**Supplementary Fig. 4c**).

The finding that marker-specific inflation factors vary substantially across the genome has notable implications for meta-analyses and multistage analyses. Such studies typically combine the test statistics after correcting for potential inflation using genomic control[30,31,38]. The disadvantages of using the same global correction rather than a marker-specific one can become more serious when this step is done repeatedly. To better understand these effects in the context of meta-analysis, we first compared the marker-specific inflation factors between the two WTCCC control groups, collected from essentially the same population. We observed a very strong correlation (r = 0.95; **Supplementary Fig. 5a**). We further compared the inflation factors across different populations and different genotyping platforms using the NFBC66 samples and WTCCC control samples. We observed a strong correlation of r = 0.70 (**Supplementary Fig. 5b**), suggesting that the marker-specific inflation factors can be correlated across the multiple data sets used in meta-analysis or multistage analysis owing to the shared genetic history. If this is the case, the standard approach that corrects with genomic control before merging the *P* values from different studies may lead to further inaccuracies: tests at some markers would be excessively, or not sufficiently, deflated multiple times, resulting in an accumulation of errors.

DISCUSSION

We report here the development of the EMMAX program, taking an expedited mixed linear model approach to correct for sample structure within human GWASs. We demonstrate its effectiveness with the analysis of two human GWAS data sets, including quantitative as well as disease traits. The proposed approach differs substantially from genomic control in that it accounts for inflation owing to population structure in a marker-specific manner, resulting in a modified ranking of association results. Accounting for marker-specific effects can reduce both false positives and false negatives. We discuss this issue in more detail in the **Supplementary Note**.

There are several other methods that take into account pedigree-based or empirically estimated kinship matrices in the statistical test[39–42]. One of the key differences between these methods and the mixed model methods, including EMMAX, is that the mixed model methods have a procedure for estimating the contribution of the kinship matrix to the phenotypes, whereas the other methods do not. Estimating the phenotypic variance contributed by the sample structure enabled us to avoid undercorrection or overcorrection of the sample structure in the NFBC66 and WTCCC data sets.

The effective application of our method depends on an appropriate estimate of the variance parameters. The IBS or Balding-Nichols matrix[43] appears to be better than IBD estimates at capturing the long-distance relationships that result in variations at the population level. However, when the structure of the sample at hand is better described in terms of fairly recent hidden relatedness, methods based on the estimation of IBD may have an advantage. In principle, our approach is also suitable for association testing in a data set including individuals from a heterogeneous population with admixed background. In such cases, it is important to consider SNP ascertainment bias in estimating the degree of relatedness between individuals. Because many SNP probes in genotyping arrays are selected from European populations, the marker-based pairwise distance between two individuals may appear to be larger between unrelated European samples than between unrelated individuals from other populations. To resolve the resulting ascertainment bias, each SNP may be differently weighted when the IBS similarity matrix is computed. A general framework has been presented[19] for computing the similarity matrix with a different weight for each marker. Different weighting schemes can also be used to account for heterogeneous distribution of effect size from each marker or each genomic region.

Besides the choice of the kinship matrix, the estimation of variance parameters is also a crucial part of the EMMAX approach. In our analysis of the NFBC data, we show that estimating these parameters under the null hypothesis does not lead to appreciable bias in the association *P* values. The example of rheumatoid arthritis and type 1 diabetes in the WTCCC data set, in contrast, reveals the difficulties encountered by EMMAX when there are SNPs explaining a large fraction of phenotypic variance. In such cases, we show that estimating variance parameters conditionally on the SNPs with stronger effects alleviates these concerns.

Finally, whereas the analysis presented here relies on decomposing the variance into two terms, a genetic relatedness component and a component representing residual effects, future studies may need to account for additional variance components to more precisely model the heterogeneous phenotypic variance. In expression quantitative trait loci mapping, for example, one may want to add additional variance components to account for technical bias[44]. When multiple variance components are involved, one would need to make use of algorithms such as PROC MIXED implemented in SAS, as EMMA is developed for two variance components only; this would increase the running time of the first step of our procedure. However, because the same variance components estimated from the null hypothesis would be used across the genome-wide markers, the overall computational time should still be acceptable.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the NFBC66 team for access to phenotype and genotype data used in the analyses presented here. The genotype data were generated at the Broad Institute with support from National Heart, Lung, and Blood Institute grant 6R01HL087679-03. We thank D. Clayton for reading through the manuscript and for providing important suggestions. We acknowledge the WTCCC for allowing us to use their data set. H.M.K., N.A.Z., J.H.S. and E.E. are supported by National Science Foundation grants 0513612, 0731455 and 0729049, and National Institutes of Health (NIH) grants 1K25HL080079 and U01-DA024417. N.A.Z. is supported by the Microsoft Research Fellowship. H.M.K. is supported by the Samsung Scholarship, National Human Genome Research Institute grant HG00521401, National Institute for Mental Health grant NH084698 and GlaxoSmithKline. C.S. is partially supported by NIH grants GM053275-14, HL087679-01, P30 1MH083268, 5PL1NS062410-03, 5UL1DE019580-03 and 5RL1MH083268-03. N.B.F. and S.K.S. are supported by NIH grants HL087679-03, 5PL1NS062410-03, 5UL1DE019580-03 and 5RL1MH083268-03. This research was supported in part by the University of California, Los Angeles subcontract of contract N01-ES-45530 from the National Toxicology Program and National Institute of Environmental Health Sciences to Perlegen Sciences.

AUTHOR CONTRIBUTIONS

H.M.K., J.H.S., C.S. and E.E. designed the methods and experiments; H.M.K., J.H.S., S.K.S., S.-y.K., N.B.F., C.S. and E.E. jointly analyzed the NFBC66 data set; H.M.K., J.H.S., N.A.Z., C.S. and E.E. jointly analyzed the WTCCC data set; H.M.K., J.H.S., S.K.S., N.B.F., C.S. and E.E. wrote the manuscript; all authors contributed their critical reviews of the manuscript during its preparation.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Voight, B.F. & Pritchard, J.K. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet.* **1**, e32 (2005).
2. Weir, B.S., Anderson, A.D. & Hepler, A.B. Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* **7**, 771–780 (2006).
3. Newman, D.L., Abney, M., McPeek, M.S., Ober, C. & Cox, N.J. The importance of genealogy in determining genetic associations with complex traits. *Am. J. Hum. Genet.* **69**, 1146–1148 (2001).
4. Helgason, A., Yngvadttir, B., Hrafnkelsson, B., Gulcher, J. & Stefnsson, K. An Icelandic example of the impact of population structure on association studies. *Nat. Genet.* **37**, 90–95 (2005).
5. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
6. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
7. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
8. Bacanu, S.A., Devlin, B. & Roeder, K. Association studies for quantitative traits in structured populations. *Genet. Epidemiol.* **22**, 78–93 (2002).
9. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
10. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
11. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40**, 646–649 (2008).
12. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
13. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35–46 (2009).
14. Cho, Y.S. *et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* **41**, 527–534 (2009).
15. Fisher, S.R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
16. Ober, C., Abney, M. & McPeek, M.S. The genetic dissection of complex traits in a founder population. *Am. J. Hum. Genet.* **69**, 1068–1079 (2001).
17. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
18. Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3**, e4 (2007).
19. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
20. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
21. Rantakallio, P. Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr. Scand.* **193** (suppl.) 1–71 (1969).
22. Varilo, T. & Peltonen, L. Isolates and their potential use in complex gene mapping efforts. *Curr. Opin. Genet. Dev.* **14**, 316–323 (2004).
23. Jakkula, E. *et al.* The genome-wide patterns of variation expose significant substructure in a founder population. *Am. J. Hum. Genet.* **83**, 787–794 (2008).
24. Kariya, T. & Kurata, H. *Generalized Least Squares* (John Wiley & Sons, 2004).
25. Chen, W.M. & Abecasis, G.R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81**, 913–926 (2007).
26. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, Massachusetts, 1998).
27. Lowe, J.K. *et al.* Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae. *PLoS Genet.* **5**, e1000365 (2009).
28. Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* **2**, e132 (2006).
29. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
30. Thomas, G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.* **41**, 579–584 (2009).
31. Ahmed, S. *et al.* Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat. Genet.* **41**, 585–590 (2009).
32. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
33. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.* **40**, 189–197 (2008).
34. Hinkley, D.V. *Theoretical Statistics* (CRC Press, Boca Raton, 1979).
35. Whittemore, A.S. & Tu, I.P. Simple, robust linkage tests for affected sibs. *Am. J. Hum. Genet.* **62**, 1228–1242 (1998).
36. de Bakker, P.I.W. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).
37. Nejentsev, S. *et al.* Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature* **450**, 887–892 (2007).
38. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40**, 638–645 (2008).
39. Thornton, T. & McPeek, M.S. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am. J. Hum. Genet.* **81**, 321–337 (2007).
40. Guan, W., Liang, L., Boehnke, M. & Abecasis, G.R. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet. Epidemiol.* **33**, 508–517 (2009).
41. Choi, Y., Wijsman, E.M. & Weir, B.S. Case-control association testing in the presence of unknown relationships. *Genet. Epidemiol.* **33**, 668–678 (2009).
42. Rakovski, C.S. & Stram, D.O. A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors. *PLoS One* **4**, e5825 (2009).
43. Balding, D.J. & Nichols, R.A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12 (1995).
44. Kang, H.M., Ye, C. & Eskin, E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* **180**, 1909–1925 (2008).
45. Irizarry, R.A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350 (2005).

ONLINE METHODS

Variance component model. We consider here the simplest form of Fisher's[15] polygenic model. Let $Z_{i,j}$ be the contribution of factor j to person i; then we assume that the phenotype y_i can be modeled as

$$y_i = \sum_{j=1}^{J} Z_{i,j} + \varepsilon_i \quad \mathrm{E}(\varepsilon_i) = 0 \quad \mathrm{Cov}(\varepsilon_{i1}, \varepsilon_{i2}) = 0 \quad \text{if } i_1 \neq i_2 \tag{1}$$

with ε_i being a random variable representing environmental effects on the phenotype. In equation (1) and throughout the paper, we include only variables accounting for the genetic factors, and all genetic factors contribute additively. This is purely a convenient assumption to simplify notation, and nongenetic factors can be modeled as additional regressors with a straightforward extension. Epistatic loci can be incorporated by including additional interaction terms in equation (1) to model a diverse set of possible types of interactions[46,47].

Let the vector $Y = \{y_1, \ldots, y_n\}$ contain the phenotypes of the individuals computed from a pedigree. Assuming that the environmental components are uncorrelated, the variance covariance structure of Y depends on the number of genes shared among subjects. In absence of dominance effects, we have

$$\mathrm{Var}(Y) = 2\sigma_a^2 \Phi + \sigma_e^2 I \tag{2}$$

where Φ is the matrix of kinship coefficients between each pair of individuals in the pedigree, and I is an identity matrix[48]. Let σ_a^2 represent the parameter for additive genetic variance and σ_e^2 represent the parameter for random environmental variance. An analysis of variance with random effects leads to the estimates of σ_a^2 and σ_e^2, and in turn to the evaluation of heritability $\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$[48].

In linkage studies, this decomposition of variance is carried one step further. By tracking the transmission of marker genes in the vicinity of locus k, one can calculate the conditional kinship coefficients (Φ_k, probabilities that two genes sampled from two individuals at locus k are IBD) and decompose the variance $\mathrm{Var}(Y)$ to emphasize the contribution of the kth locus

$$\mathrm{Var}(Y) = 2\sigma_{ak}^2 \Phi_k + 2\Phi\sigma_a^2 + \sigma_e^2 I$$

To investigate the contribution of locus k to the phenotype, one tests the null hypothesis that $\sigma_{ak}^2 = 0$. The values of the variance parameters are estimated with maximum likelihood procedures[49].

In association studies, using a much denser set of genotypes, we aim to associate the phenotypes directly to the alleles at marker loci; in other words, our goal is to estimate fixed effects. Assuming additive effects only, equation (1) can be translated to the following regression framework:

$$y_i = \beta_0 + \sum_{k=1}^{M} \beta_k X_{ik} + \varepsilon_i \tag{3}$$

with $\mathrm{Var}(\epsilon) = \sigma_e^2 I$, and X_k being the individuals' minor allele counts at locus $k \in \{1,2,L,M\}$ (for simplicity, we assume all markers are biallelic). Our goal is to identify which elements in the $M \times 1$ vector β are different from 0.

Whereas model (3) is fundamentally a multivariate one, association studies are typically carried out by testing the hypothesis H_0: $\beta_k = 0$ for each of the M loci, one locus at a time, on the basis of the model

$$y_i = \beta_0 + \beta_k X_{ik} + \eta_{i\bar{k}} \tag{4}$$

where β_k is the effect size of marker k, and the error term $\eta_{i\bar{k}} = \sum_{s \neq k} \beta_s X_{is} + \varepsilon_i$. With respect to equation (3), equation (4) is mis-specified if $\eta_{i\bar{k}}$ values are assumed to be independently and identically distributed (i.i.d.): relevant regressors are omitted; in other words, we ignore the polygenic background of the trait.

The appropriate statistical methods to estimate β_k in equation (4) depends on the nature of the sample. If the n individuals are related with a known degree of relatedness, the variance covariance of $\eta_{i\bar{k}}$ in equation (4) can be represented approximately as in equation (2). That is, the effect of the genotype at locus k can be modeled as a main effect, whereas the relationships among all individuals are taken into account by means of variance components of random polygenic effects[16]. This model is sometimes referred to as an instance of a mixed effect model[17].

If the n individuals are unrelated and there is no dependence across the genotypes, so that the $\eta_{i\bar{k}}$ values are i.i.d., a simple linear regression would make appropriate inference. However, these conditions are not easily met. First, because of linkage disequilibrium, X_k values corresponding to markers with close genomic position are correlated. Moreover, neither the homogeneity of population background nor the level of relatedness are easily controlled in the sampling stage. If the n individuals in the sample belong to distinct populations or are (albeit distantly) related, one can expect a substantial correlation between the rows and columns of X. This translates to bias in the estimate of β_k from equation (4), and the distribution of $\hat{\beta}_k$, a best unbiased linear estimation of β_k, is different from what is assumed in standard linear regression (that is, the $\eta_{i\bar{k}}$ values in equation (4) are not i.i.d.).

Using dense, genome-wide genotype data, it has become possible to estimate the degree of relationship or kinship matrix between independently ascertained subjects[50–55] in the absence of genealogical information. With an estimated kinship matrix one can, in principle, use variance component techniques in linear mixed models (as in ref. 16) to analyze population samples. If many SNPs are involved in a trait and the contribution of each SNP to the total trait variance is almost negligible, as appears to be the case for human quantitative traits[20,56], the variance components for $\eta_{i\bar{k}}$ in equation (4) can be approximated to $\eta_i = \sum_{s=1}^{M} \beta_k X_{ik} + \varepsilon_i$ and may not need to be estimated separately for each SNP. Instead, one might estimate the values σ_a^2 and σ_e^2 from a variance decomposition model as in equation (2), keep them fixed and then estimate the parameter β_k in equation (4) using a GLS procedure.

Application to quantitative traits. We used the following procedure to analyze human population samples in association studies for quantitative traits. Let n be the sample size, p the total number of genotyped SNPs and Y the vector of observed phenotypes.

1. Use the genotype data to calculate the $n \times n$ matrix $\hat{S}$ pairwise genetic relatedness between individuals, such as IBS or Balding-Nichols matrix, and normalize $\hat{S}$ to have sample variance 1 using a Gower's centered matrix[57].

$$\hat{S}_N = \frac{(n-1)\hat{S}}{\mathrm{Tr}(P\hat{S}P)} \tag{5}$$

where $P = I - \mathbf{1}\mathbf{1}'/n$ and $\mathbf{1}$ is a vector of ones. It should be noted that $\hat{S}$ can be substituted to various other pairwise relatedness matrices estimated from the genotypes[9,50–55], as long as the matrix is positive-semidefinite.

2. Use a variance component model to estimate the restricted maximum likelihood parameters (or alternatively, maximum likelihood parameter) of σ_a^2 and σ_e^2 in

$$\mathrm{Var}(Y) = \sigma_a^2 \hat{S}_N + \sigma_e^2 I \tag{6}$$

Test the hypothesis H_0: $\sigma_a^2 = 0$. If the null hypothesis is rejected, proceed to step 3; otherwise, use ordinary least squares to estimate the coefficients of each of the SNPs genotyped.

3. For each marker, use GLS F-test[24], or alternatively a score test, to estimate the effects β_k and test the hypothesis $\beta_k \neq 0$ in the following model:

$$y_i = \beta_0 + \beta_k X_{ik} + \eta_i \quad \mathrm{Var}(\eta) = V \propto \hat{\sigma}_a^2 \hat{S}_N + \hat{\sigma}_e^2 I \tag{7}$$

The above model can be easily extended to have additional confounding variables by substituting β_0 for a multicolumn matrix containing the confounding variables, such as sex and age. Note that these additional confounding variables should be included in the procedure of restricted maximum likelihood estimation of the variance component parameters. Multilocus models can be incorporated by including additional interaction terms[46,47]. For the variance component estimation procedure in step 2, we use EMMA[19]. We term our method EMMA eXpedited (EMMAX) because it markedly reduces the computational cost compared to the original EMMA by avoiding the repetitive variance component estimation procedure for each single marker. We investigated the effects of this simplification on the data sets we analyzed.

Application to case control data sets. Although EMMAX was developed with quantitative traits in mind, it can also be adapted to the analysis of case-control data sets. As the case-control phenotypes do not follow a normal distribution, applying a generalized linear mixed model using logit or probit link function is preferable to a linear mixed model. However, the computational

cost of a generalized linear model with a correlated variance component is much higher, and currently available algorithms cannot handle thousands of individuals simultaneously[58].

When the hypothesis of an additive model appears reasonable, the Armitage trend test[59] can be used to test for the presence of a genetic effect (see, for example, ref. 7, and note the equivalence of an Armitage test to a score test in logistic regression for H_0: $\beta = 0$ (ref. 60)). The Armitage test can be described as testing the significance of the slope coefficient in a linear regression of a 0–1 variable representing case-control status on the additively coded genotypes. Armitage[59] suggested using a χ^2_1 test that is slightly different from the square of a standard t-test in linear regression. The statistic proposed by Armitage is $\chi^2_0 = \beta / \mathrm{var}(\beta)$, but instead of estimating the variance of the error terms using the residuals from the regression, we estimate it using the variance of the response variable. Therefore, χ^2_0 is equal to the square of the correlation between the response and the genotype variables, multiplied by the number of samples.

Despite this suggestion, Armitage indicated that the standard t statistic may be preferable, especially to construct confidence intervals. Therefore, it seems that one can carry out tests in the spirit of Armitage simply using a standard linear regression framework with a 0–1 quantitative response variable representing the case-control status. Adopting this approach, we were immediately able to translate the problem to the methodology suggested for quantitative traits.

Genotype and phenotype data. We analyzed two data sets: one that contains measurements of quantitative traits (NFBC66), and one for binary disease traits (WTCCC). Genotype data were available for 5,546 Finnish subjects from NFBC66 (ref. 13), all with genotyping completeness >95%. We excluded subjects from further analysis because they had withdrawn consent (15), had discrepancies between reported sex and sex determined from the X chromosome (14), were sample duplications (2), were too related to another subject (77), had more than 5% missing genotypes (1) or had no phenotype data (111), leaving 5,326 subjects for analysis. For the relatedness criterion, we identified all pairs of subjects with probability of IBD >20% and included one subject from each such pair in further analyses. In most cases, the subject with the most nonmissing phenotype data was chosen for analysis. If the two subjects had an equal amount of missing phenotype data, we used the subject with the most nonmissing genotype data.

Using these 5,326 subjects, we examined the 368,177 SNP markers for Hardy-Weinberg equilibrium (exact test), genotyping completeness and minor allele frequency. Markers were excluded for more than two discordant genotype calls between different methods (4,711), Hardy-Weinberg equilibrium $P < 10^{-4}$ (5,260), genotyping completeness <95% (2,535) and minor allele frequency <1% (27,002), leaving 331,475 markers for analysis (some SNP markers failed quality checks on more than one criterion). We adjusted the nine phenotypes used in the original data for sex, pregnancy status and use of oral contraceptives, as described[13], and adjusted height for sex only.

The NFBC66 database contains information on the birth locations of subjects and their parents, which can be used to derive ancestry information. Ref. 13 describes how six distinct linguistic and geographical groups can be identified in the northern provinces of Finland. Given the patterns of internal migrations and their variation over time, we can assign individuals in NFBC66 to one of these groups when both parents were born in a municipality within the same group. Approximately 50% of the sample can be assigned this way, and these individuals were used to compare the results of population stratification analysis based on genotypes.

We also obtained the genotypes of the WTCCC subjects collected for a GWAS of seven common diseases[6]. We applied the same quality-control criteria as suggested in the original paper. We also excluded the SNPs that the original studies excluded in their analysis. We considered a total of 404,862 SNPs after the quality control across 2,938 shared controls and 13,241 case individuals across seven diseases.

Additionally, it appears that using the simple identity-by-state (IBS) between individuals, rather than the more laboriously constructed kinship coefficients, may be sufficient, and in some cases more appropriate, to model the dependency in the sample. We investigate this assumption further in the **Supplementary Note**, **Supplementary Table 3** and **Supplementary Figures 3b** and **6**.

URL. The EMMAX software is available at http://genetics.cs.ucla.edu/emmax.

46. Marchini, J., Donnelly, P. & Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37**, 413–417 (2005).
47. Evans, D.M., Marchini, J., Morris, A.P. & Cardon, L.R. Two-stage two-locus models in genome-wide association. *PLoS Genet.* **2**, e157 (2006).
48. Falconer, D.S. & Mackay, T.F.C. *Introduction to Quantitative Genetics* 4th edn. (Longman, 1996).
49. Lange, K. *Mathematical and Statistical Methods for Genetic Analysis* (Springer, 2002).
50. Lynch, M. & Ritland, K. Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753–1766 (1999).
51. Epstein, M.P., Duren, W.L. & Boehnke, M. Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet.* **67**, 1219–1231 (2000).
52. Thomas, S.C. & Hill, W.G. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* **155**, 1961–1972 (2000).
53. Ritland, K. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet. Res.* **67**, 175–185 (2009).
54. McPeek, M.S. & Sun, L. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.* **66**, 1076–1094 (2000).
55. Milligan, B.G. Maximum-likelihood estimation of relatedness. *Genetics* **163**, 1153–1167 (2003).
56. Maher, B. Personal genomes: the case of the missing heritability. *Nature* **456**, 18–21 (2008).
57. McArdle, B.H. & Anderson, M.J. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**, 290–297 (2001).
58. McCulloch, C.E. *Generalized Linear Mixed Models* (Institute of Mathematical Statistics, Alexandria, Virginia, and American Statistical Association, Beachwood, Ohio, 2003).
59. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386 (1955).
60. Agresti, A. & Wiley, J. *Categorical Data Analysis* (Wiley, New York, 1990).

doi:10.1038/ng.548