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

Phenotype	IBS matrix		BN matrix		Reference h^2	
	p-value ($\sigma_a^2 = 0$)	$\hat{h}^2_{IBS}$	p-value ($\sigma_a^2 = 0$)	$\hat{h}^2_{BN}$	**Kosrae** h^2	**Sardinia** h^2
CRP	1.7×10^{-2}	0.134	2.3×10^{-2}	0.116	0.245	0.296
TG	2.3×10^{-4}	0.178	2.4×10^{-3}	0.152	0.274	0.322
INS	8.3×10^{-4}	0.205	3.1×10^{-3}	0.152	N/A	0.260
DBP	4.7×10^{-4}	0.199	5.6×10^{-4}	0.167	0.289	0.186
BMI	3.9×10^{-6}	0.279	1.9×10^{-6}	0.242	0.473	0.426
GLU	4.2×10^{-5}	0.229	2.4×10^{-5}	0.197	0.188	0.362
HDL	5.5×10^{-11}	0.384	1.0×10^{-11}	0.324	0.391	0.486
SBP	2.7×10^{-8}	0.283	2.0×10^{-8}	0.233	0.243	0.253
LDL	1.4×10^{-17}	0.452	1.2×10^{-18}	0.384	0.414	0.425
HEIGHT	2.8×10^{-45}	0.738	2.5×10^{-48}	0.625	0.790	0.798

Supplementary Table 1: P-values for test of the null hypothesis $\sigma_a^2 = 0$ for all traits; pseudo-heritability estimates $h_a^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, and heritability estimates from Kosrae population[22] and Sardinia population[23]. A simple IBS matrix and Balding-Nichols (BN) matrix is used as estimates of relatedness.

Nature Genetics: doi:10.1038/ng.548

Phenotype	**Uncorr. vs EMMAX**	**Uncorr. vs ES100**	**ES100 vs EMMAX**	**Uncorr. λ**
CRP	0.891 (0.94)	0.635 (0.78)	0.660 (0.79)	1.007
TG	0.856 (0.92)	0.569 (0.72)	0.612 (0.76)	1.023
INS	0.826 (0.90)	0.535 (0.70)	0.603 (0.75)	1.029
DBP	0.843 (0.91)	0.607 (0.75)	0.646 (0.78)	1.031
BMI	0.790 (0.88)	0.544 (0.70)	0.607 (0.75)	1.031
GLU	0.775 (0.87)	0.528 (0.69)	0.604 (0.75)	1.045
HDL	0.693 (0.82)	0.500 (0.66)	0.576 (0.73)	1.052
SBP	0.684 (0.81)	0.481 (0.65)	0.597 (0.75)	1.066
LDL	0.624 (0.77)	0.474 (0.64)	0.587 (0.74)	1.098
HEIGHT	0.453 (0.62)	0.386 (0.55)	0.497 (0.66)	1.187

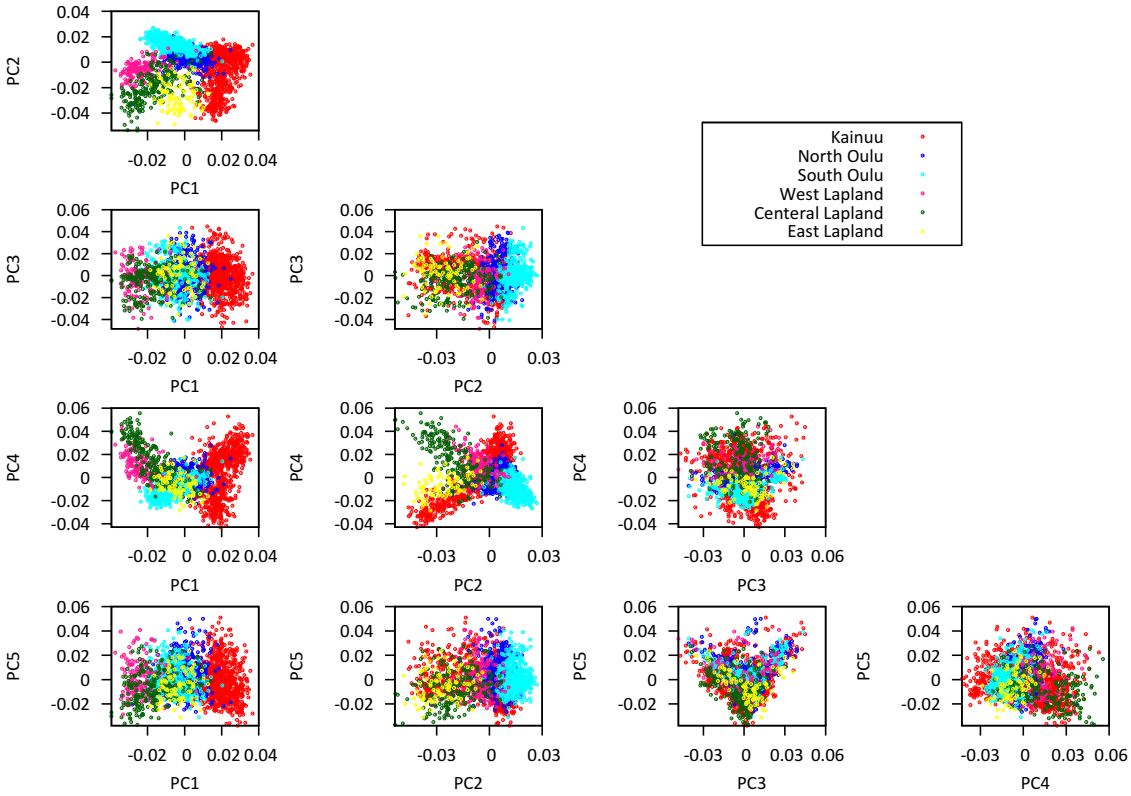
Supplementary Table 2: Comparison of top 2,000 hits obtained with uncorrected analysis, EIGENSOFT with 100 PCs (ES100), and EMMAX. The numbers in second to fourth column represents the proportion of shared SNPs between each pair of analysis, when selecting top 2,000 SNPs in each analysis. The values in parentheses are Cohen's kappa coefficients as a measure of the agreement between two tests. For clarity we have ordered the phenotypes with reference to their genomic control parameters and reported these as well in the last column.

Nature Genetics: doi:10.1038/ng.548

Phenotypes	Uncorrected	EMMAX-IBS	EMMAX-BN	Concordance
CRP	1.007	0.993	0.992	0.969 (0.98)
TG	1.023	1.002	1.000	0.969 (0.98)
INS	1.029	1.005	1.005	0.951 (0.97)
DBP	1.031	1.007	1.005	0.955 (0.98)
BMI	1.031	0.995	0.992	0.942 (0.97)
GLU	1.045	1.008	1.004	0.946 (0.97)
HDL	1.052	1.004	1.000	0.919 (0.96)
SBP	1.066	1.006	1.001	0.940 (0.97)
LDL	1.098	1.002	0.999	0.915 (0.96)
HEIGHT	1.187	1.003	0.994	0.838 (0.91)

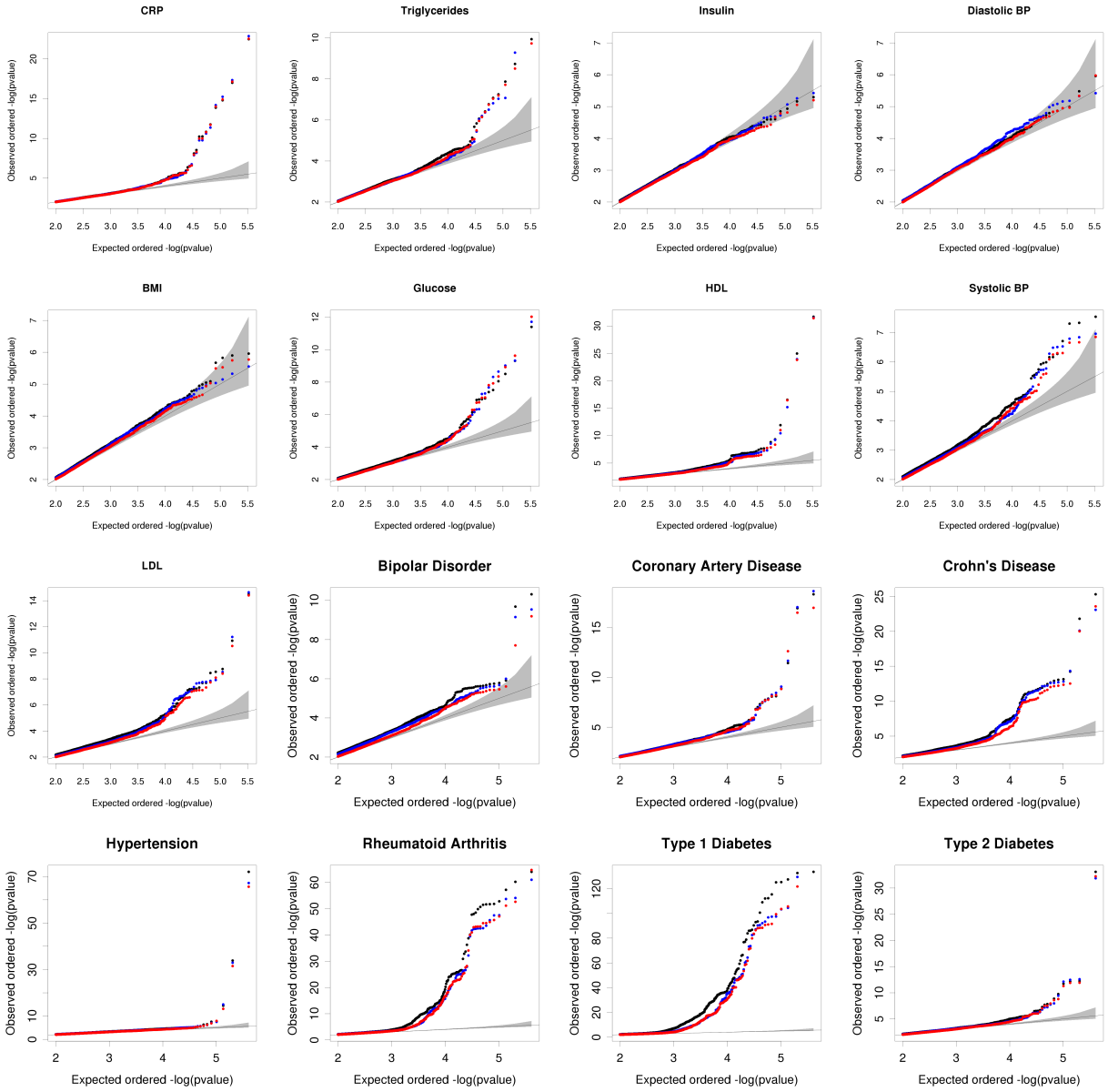
Supplementary Table 3: Comparison of genomic control inflation factors obtained with uncorrected analysis and EMMAX with IBS matrix and Balding-Nichols (BN) matrix. The "Concordance" column represents the proportion of shared SNP between top 2000 associations between EMMAX-IBS and EMMAX-BN method. The values in the parentheses are kappa statistic

Nature Genetics: doi:10.1038/ng.548



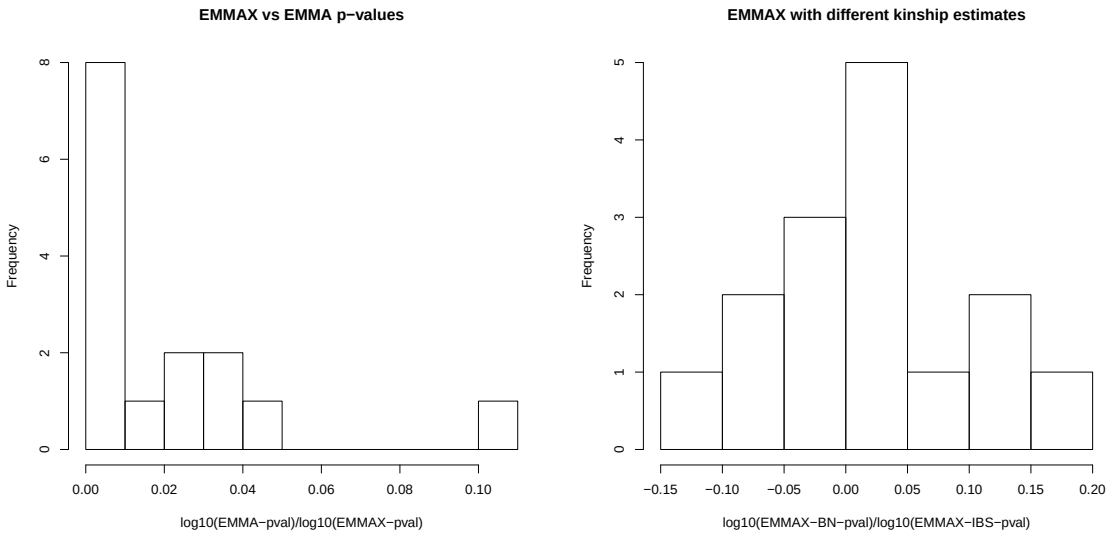
Supplementary Figure 1: Scatter plots of the first 5 principal components for individuals of known ancestry. The different linguistic/geographic subgroups are color-coded.

Nature Genetics: doi:10.1038/ng.548



Supplementary Figure 2: QQ-plots on the log10 scale of the association p-values obtained for nine traits according to three different models for 9 NFBC66 metabolic trais and 7 WTCCC disease phenotypes. In black, results from the unadjusted analysis; in blue results from the analysis conducted using 100 PC, and in red results from EMMAX.

Nature Genetics: doi:10.1038/ng.548

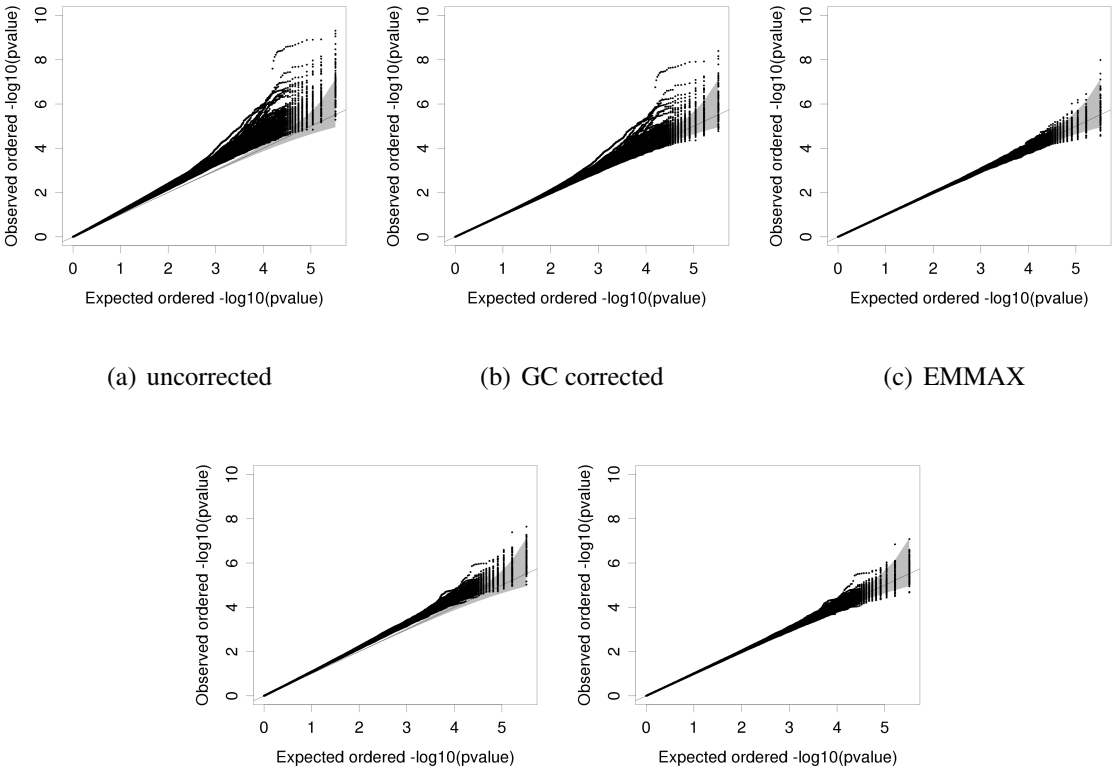


(a) EMMAX vs EMMA

(b) EMMAX-IBS vs EMMAX-BN

Supplementary Figure 3: Comparison of p-values obtained running EMMAX using IBS matrix with the corresponding value obtained using (a) the original EMMA and (b) EMMAX with Balding-Nichols (BN) matrix for the SNPs whose p-value under EMMAX was smaller than 7.2×10^{-8}.

Nature Genetics: doi:10.1038/ng.548

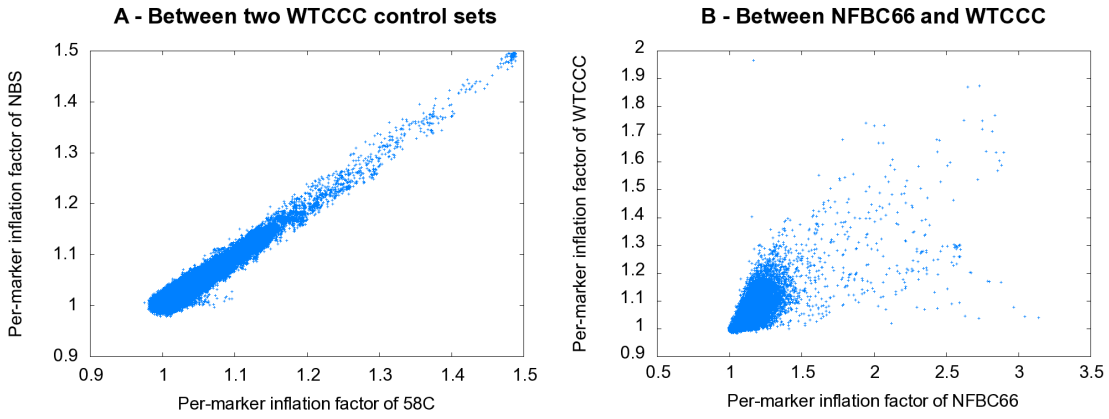


(a) uncorrected (b) GC corrected (c) EMMAX

(d) ES100 (e) ES100 + GC

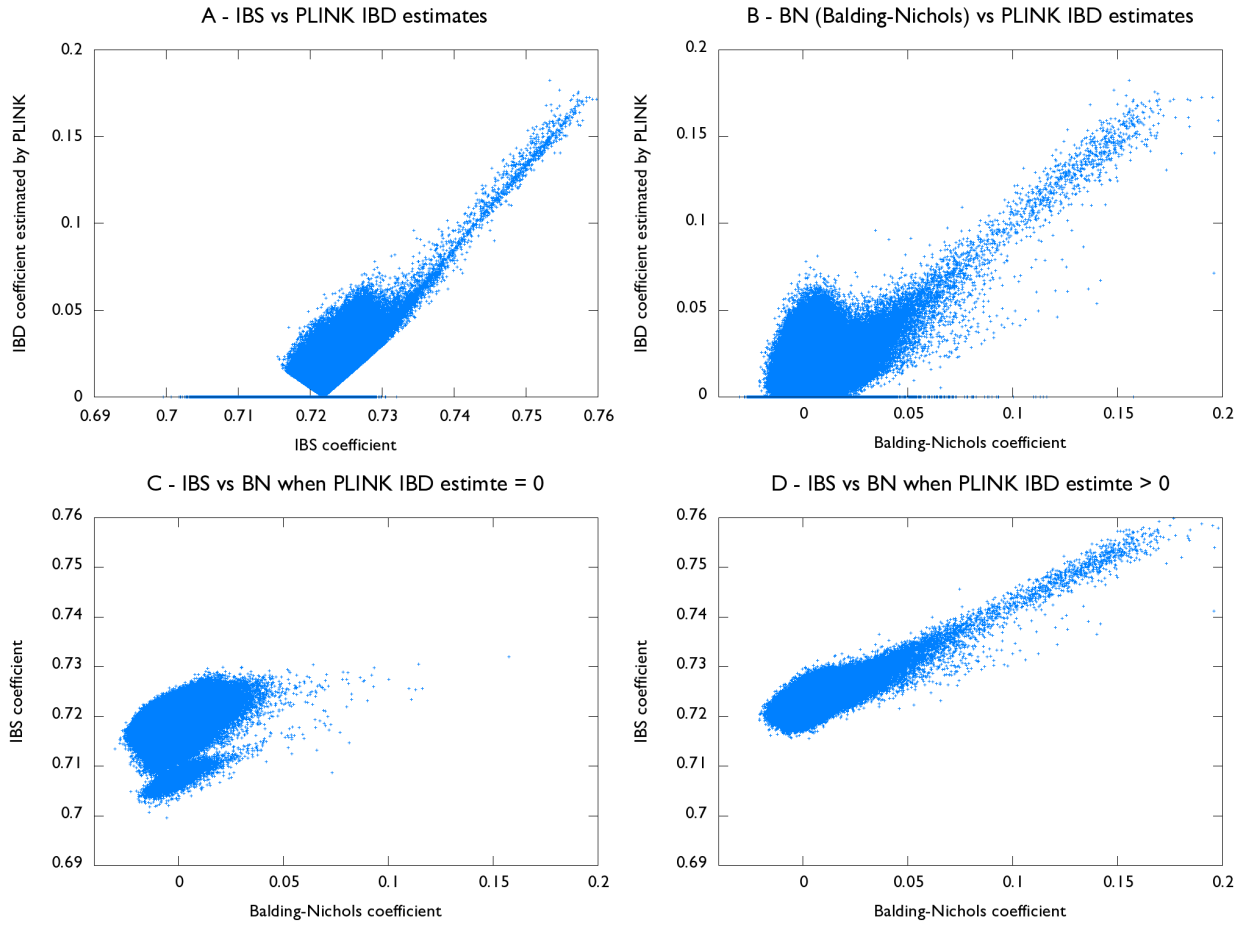
Supplementary Figure 4: QQ plots of 100 randomly generated phenotypes under the variance component model using a (a) uncorrected analysis, (b) genomic control adjustment, (c) EMMAX, (d) EIGENSOFT with 100 PCs, and (e) genomic control adjustment after applying EIGENSOFT with 100 PCs.

Nature Genetics: doi:10.1038/ng.548



Supplementary Figure 5: Concordance of per-marker inflation factor (A) between two different control sets (58C and NBS) in WTCCC data set, and (B) between NFBC66 samples and WTCCC control samples using the 50,298 overlapping markers.

Nature Genetics: doi:10.1038/ng.548



Supplementary Figure 6: Comparisons (A) between the IBS coefficients and IBD estimates computed by PLINK (B) between the Balding-Nichols (BN) coefficients and IBD estimates from PLINK, (C) between IBS and BN coefficients when IBD estimates are zero (D) IBS and BN coefficients when IBD estimates are positive.

Nature Genetics: doi:10.1038/ng.548

Supplementary Note

Estimation of relatedness from high-density markers

Unlike a traditional variance component model which uses IBD (identity by descent) coefficients estimated from the pedigree[1], our proposed method empirically estimate the genetic relatedness between the individuals from high-density markers. In model organism studies, Yu et al.[2] estimated kinship coefficients from multi-locus genotypes using method-of-moment estimators[3,4], and Zhao et al. and Kang et al.[5,6] demonstrated that using a haplotype-based IBS matrix or a simple IBS matrix more robustly corrects for the population structure resulting in a lower inflation factor than using the estimated IBD matrix from structured model organism samples. Zhao et al.[5] observe in Arabidopsis that while IBD is preferable to describe recent relatedness, IBS may be more apt to describe very distant relationships between individuals, that indeed blend into population level differences. Along these lines, Kang et. al[6] showed that IBS can precisely reflects the polygenic background under the assumption that each SNP is equally likely to contribute to the quantitative trait at a very small level. Several other methods[7–9] have been proposed to estimate IBD kinship coefficients or sample structure from multi-locus genotypes including the maximum-likelihood method implemented in PLINK software[10,11] and the PREST software[12]

The effectiveness of the empirically estimated pairwise relatedness in correcting for sample structure has not been comprehensively examined in a large-scale human association mapping studies, where the sample structure is much less heterogeneous than those among the strains of model organisms. For this reason, we compared three different empirical estimates of pairwise genetic relatedness from the NFBC66 samples. First is a simple IBS coefficient, and the second is a maximum-likelihood estimates (MLE) of IBD kinship coefficient[11] implemented in the PLINK[10] software. The third is the Balding-Nichols (BN) kinship coefficient[9].

The pairwise plots across these three methods suggest that the relatedness estimates computed by these methods are highly correlated with each other (Supplementary Figure 6). The MLE-based IBD estimates[11] shows a correlation of $r = 0.62$ with IBS coefficient, and $r = 0.48$ with

Nature Genetics: doi:10.1038/ng.548

BN coefficient. The MLE-based methods estimates 37% of the pairwise kinship coefficients to be positive, and those individual pairs show strong correlation of $r = 0.68$ between BN and IBS coefficients. Among the 63% of individual pairs where the MLE-based kinship coefficient are zero, a strong correlation of $r = 0.54$ is observed between the IBS and BN coefficients, suggesting that the unrelated individual pairs may still have different degrees of distant relatedness.

We applied either the simple IBS or the BN matrix as the surrogate of sample structure when applying EMMAX, and results with IBS matrix is reported unless specified or compared between the two methods. The MLE-based method does not guarantee that the estimated kinship matrix is positive semidefinite (all eigenvalues are non-negative), making it difficult to use in a variance component model. The EMMAX p-values across the two methods provide a very high concordance to each other (Supplementary Table 3 and Supplementary Figure 3B).

Methods for estimating marker specific inflation factors

Assuming that model (4) is true with $V = \text{Var}(\eta)$ and marker k has no effect on the phenotype, we define the inflation factor for marker k as the ratio between the expectation of the F statistics calculated from OLS for a model that includes k, to the expectation of the F statistics for the same model calculated from GLS. In fact, we do not compute this ratio explicitly, but simply provide an approximation. If one considers that as $n \longrightarrow \infty$, the expectation of the GLS F statistics under arbitrary V, as long as V is non singular, converges to 1; hence we simply need an approximation for the numerator of the ratio.

Specifically, let us assume, to simplify notation, that Y and X_k are centered to have zero sample mean so that $\hat{\beta}_0 = 0$ holds. In such a case, $V = \text{Var}(\eta)$ has to be centered to $V_C = PVP$ where $P = I - \mathbf{1}\mathbf{1}'/n$. In addition, for convenience purposes, we standardize X_k to satisfy $X_k^T X_k = n - 1$, where n is the number of individuals. Then the F-test statistic based on OLS[13] becomes

Nature Genetics: doi:10.1038/ng.548

$$F_{OLS} = \frac{((X_k'X_k)^{-1}X_k'Y)^2(X_k'X_k)(n-2)}{Y'(I - X_k(X_k'X_k)^{-1}X_k')Y} \quad (8)$$

$$= \frac{(X_k'Y)^2(n-2)}{nY'Y - (X_k'Y)^2}. \quad (9)$$

If $V = \sigma^2 I$, then F_{OLS} follows a F-distribution with $(1, n-2)$ degree of freedom. Then if n is large, F_{OLS} asymptotically converges to chi-square distribution with 1 degree of freedom. While the distribution of F_{OLS} is difficult to calculate when V has off-diagonal elements, the expected values of numerator and denominator in F_{OLS} are relatively easy to compute. The expectation of denominator becomes $n\text{Tr}(V_C) - X_k'V_CX_k$, and the expectation of numerator becomes $(n-2)X_k'V_CX_k$.

We can then take as operational definition of the marker specific inflation factor ζ_k at marker k,

$$\zeta_k = \frac{(n-2)X_k'V_CX_k}{(n-1)\text{Tr}(V_C) - (X_k'V_CX_k)} \quad (10)$$

$$\approx \frac{X_k'V_CX_k}{\text{Tr}(V_C)} \quad (11)$$

Note that when $V = \sigma^2 I$, then $\zeta_k = 1$ holds regardless of the values of X_k. Let $\hat{S}_C = P\hat{S}_N P$. When we take for V the specific form assumed in (7), we can further simplify the expression above:

$$\begin{aligned}
\zeta_k &= \frac{(n-2)X_k'(\sigma_a^2\hat{S}_C + \sigma_e^2 P)X_k}{(n-1)\text{Tr}(\sigma_a^2\hat{S}_C + \sigma_e^2 P) - (X_k'(\sigma_a^2\hat{S}_C + \sigma_e^2 P)X_k)} \\
&= \frac{\sigma_a^2(n-1)X_k'\hat{S}_CX_k + \sigma_e^2(n-1)(n-2)}{\sigma_a^2\left[(n-1)^2 - X_k'\hat{S}_CX_k\right] + \sigma_e^2(n-1)(n-2)} \\
&\approx \frac{\sigma_a^2 X_k'\hat{S}_CX_k/(n-1) + \sigma_e^2}{\sigma_a^2 + \sigma_e^2} \\
&= h_a^2 X_k'\hat{S}_CX_k/(n-1) + (1-h_a^2) \quad (12)
\end{aligned}$$

where $h_a^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$ is the pseudo-heritability.

We are now in the position to discuss the meaning and implication of the marker specific inflation factors we defined. The introduced marker-specific inflation factors essentially estimate

Nature Genetics: doi:10.1038/ng.548

the effects of the mis-specification of variance component by using OLS in the place of GLS. From expression (12) it is clear that the amount of inflation at any given marker depends on the level of correlation between the marker genotypes and the GLS variance-covariance matrix. This validates the common intuition that cryptic population structure may affect tests differently at different markers and it illustrates the reasons of such variability. Expression (12) also clarifies how the same level of sample structure will affect differently the association tests for different phenotypes. The inflation will be stronger the higher is the ratio of σ_a^2 to σ_e^2, while for a trait that does not follow the polygenic model $\sigma_a^2 = 0$, no amount of population structure will have any impact on the association tests. Finally, it is useful to recall that the inflation factors ζ_k, while marker specific, are calculated independently of the observed association between marker and phenotype, being based on expectations of test statistics under the null model.

More generally, if multiple confounding variables need to be accounted for in addition to the intercept under the null model, Equation (9) can be rewritten in a general form of F statistic to get the expectation of numerator and denominator. Such a procedure is asymptotically equivalent to centering an arbitrary variance component V to $V_C = (I - G(G'G)^{-1}G)V(I - G(G'G)^{-1}G)$, given a non-singular matrix of confounding variables G that includes the intercept. In this case, the SNP vector X_k also needs to be regressed out with respected to G, and $(n-2)$ in Equation (9) needs to be replaced with $(n-q-1)$, where q is the number of columns in G.

This method can also be extended for estimating the effect of mis-specified variance component or errors in the variance component estimation. Before running GLS, let $\hat{V} = \hat{\sigma_a}^2 \hat{S}_N + \hat{\sigma_e}^2 I$ be the estimated variance-covariance matrix when V is the true one. Assuming that Y and X_k are centered, the F test statistics for GLS is

$$F_{GLS} = \frac{((X_k'\hat{V}_C^{-1}X_k)^{-1}X_k'\hat{V}_C^{-1}Y)^2(X_k'\hat{V}_C^{-1}X_k)(n-2)}{Y'(\hat{V}_C^{-1} - \hat{V}_C^{-1}X_k(X_k'\hat{V}_C^{-1}X_k)^{-1}X_k'\hat{V}_C^{-1})Y} \quad (13)$$

$$= \frac{(X_k'\hat{V}_C^{-1}Y)^2(n-2)}{(X_k'\hat{V}_C^{-1}X_k)Y'\hat{V}_C^{-1}Y - (X_k'\hat{V}_C^{-1}Y)^2} \quad (14)$$

where $\hat{V}_C$ represents the centered matrix of $\hat{V}$. The ratio between expected numerator and denom-

Nature Genetics: doi:10.1038/ng.548

inator provides the inflation factor with mis-specified variance component.

$$\zeta_k = \frac{X_k' \hat{V}_C^{-1} V_C \hat{V}_C^{-1} X_k (n-2)}{(X_k \hat{V}_C^{-1} X_k)\mathrm{Tr}(\hat{V}_C^{-1}\hat{V}_C) - X_k' \hat{V}_C^{-1} V_C \hat{V}_C^{-1} X_k} \quad (15)$$

$$\approx \frac{(n-1) X_k' \hat{V}_C^{-1} V_C \hat{V}_C^{-1} X_k}{(X_k \hat{V}_C^{-1} X_k)\mathrm{Tr}(\hat{V}_C^{-1} V_C)} \quad (16)$$

Accounting for large effect sizes at some SNPs

The accuracy of EMMAX relies on the assumption that the effect of each SNP on the phenotype is negligible for the purpose of estimating σ_a^2 and σ_e^2 in model (7). This is a reasonable assumption for most of current human GWAS, because a majority of genome-wide significant signals reported so far explain only a small fraction of phenotypic variance[14]. For example, in a genome-wide study with 5,000 individuals, a genome-wide significance p-value of 7.2×10^{-8} corresponds to 0.58% of phenotypic variance explained. 10^{-10} corresponds to 0.84%, and 10^{-15} to 1.3%. A cumulative effect of several significant SNPs are still relatively small compared to the total genetic effects for most complex traits[14–17].

However, a number of phenotypes do not comply with the "negligible effect" assumption. There are many Mendelian traits where a single locus explains the total phenotypic variance almost completely. Among complex traits, several autoimmune diseases including Rheumatoid arthritis and Type I diabetes are largely explained by HLA alleles with relative risks 4 or greater[18,19], with extremely significant with p-values smaller than 10^{-50} or 10^{-100}, explaining 50% or even larger variance of these traits[20]. In such cases, where a number of SNPs explains a considerable portion of the phenotypic variance, the negligible effect assumption is ungrounded, and the strategy described so far impractical, because the variance parameter estimation can be substantially biased due to the large effect SNPs.

In fact, it is possible to use EMMAX even in this context, provided that one conditions on the effects of the strongly associated SNPs. Specifically, one can condition on the effects of the implicated SNPs by modeling them as fixed effects when estimating σ_a^2 and σ_e^2 in model (7). It is

Nature Genetics: doi:10.1038/ng.548

crucial, then, to decide on the effect of which SNPs one should condition upon. If we know *a priori* the identity of associated loci with strong effect, such as the MHC region in the above example, the choice will be obvious. Otherwise, we may condition on the effects of SNPs with highly significant p-values. It is important to use a very stringent significance threshold to avoid loss of power. In our analysis, we conditioned on the SNPs explaining more than 1% of phenotypic variance. In RA and T1D, 58 and 135 significant SNPs in MHC and PTPN2 region are conditioned on. Note that this conditioning procedure is really recommended only if (1) there are a few genomic loci largely explaining the phenotypic variance, and (2) significant over-dispersion or under-dispersion of test statistics is observed after applying EMMAX. It should be noted that it is also possible to account for the large effect SNPs in a more sophisticated way using regularization-based methods such as ridge regression or LASSO[21], instead of a simple threshold-based conditioning.

References

[1] Ober, C., Abney, M. & McPeek, M. S. The genetic dissection of complex traits in a founder population. *Am J Hum Genet* **69**, 1068–79 (2001).

[2] Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**, 203–8 (2006).

[3] Loiselle, B. A., Sork, V. L., Nason, J. & Graham, C. Spatial genetic structure of a tropical understory shrub, psychotria officinalis (rubiaceae). *American Journal of Botany* 1420–1425 (1995).

[4] Ritland, K. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetics Research* **67**, 175–185 (2009).

[5] Zhao, K. *et al.* An arabidopsis example of association mapping in structured samples. *PLoS Genet* **3**, e4 (2007).

Nature Genetics: doi:10.1038/ng.548

[6] Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–23 (2008).

[7] Bauman, L. E., Sinsheimer, J. S., Sobel, E. M. & Lange, K. Mixed effects models for quantitative trait loci mapping with inbred strains. *Genetics* **180**, 1743–61 (2008).

[8] Choi, Y., Wijsman, E. M. & Weir, B. S. Case-control association testing in the presence of unknown relationships. *Genet Epidemiol* **33**, 668–78 (2009).

[9] 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).

[10] Purcell, S. *et al.* Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–75 (2007).

[11] Milligan, B. G. Maximum-likelihood estimation of relatedness. *Genetics* **163**, 1153–67 (2003).

[12] McPeek, M. S. & Sun, L. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet* **66**, 1076–94 (2000).

[13] Ravishanker, N. & Dipak, D. *A first course in linear model theory* (CRC Press, 2002), illustrated edn.

[14] Manolio, T. A., Brooks, L. D. & Collins, F. S. A hapmap harvest of insights into the genetics of common disease. *J Clin Invest* **118**, 1590–605 (2008).

[15] Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–53 (2009).

[16] Lango, H. *et al.* Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes* **57**, 3129–35 (2008).

Nature Genetics: doi:10.1038/ng.548

[17] Bogardus, C. Missing heritability and gwas utility. *Obesity* **17**, 209–10 (2009).

[18] 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**, 1166–72 (2006).

[19] Clayton, D. G. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet* **5**, e1000540 (2009).

[20] WellcomeTrustCaseControlConsortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78 (2007).

[21] Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference* (Springer, 2004).

[22] 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).

[23] Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2**, e132 (2006).

Nature Genetics: doi:10.1038/ng.548