Supplementary Information

Common SNPs explain a large proportion of heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1*]

[1]*Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Queensland 4006, Australia*

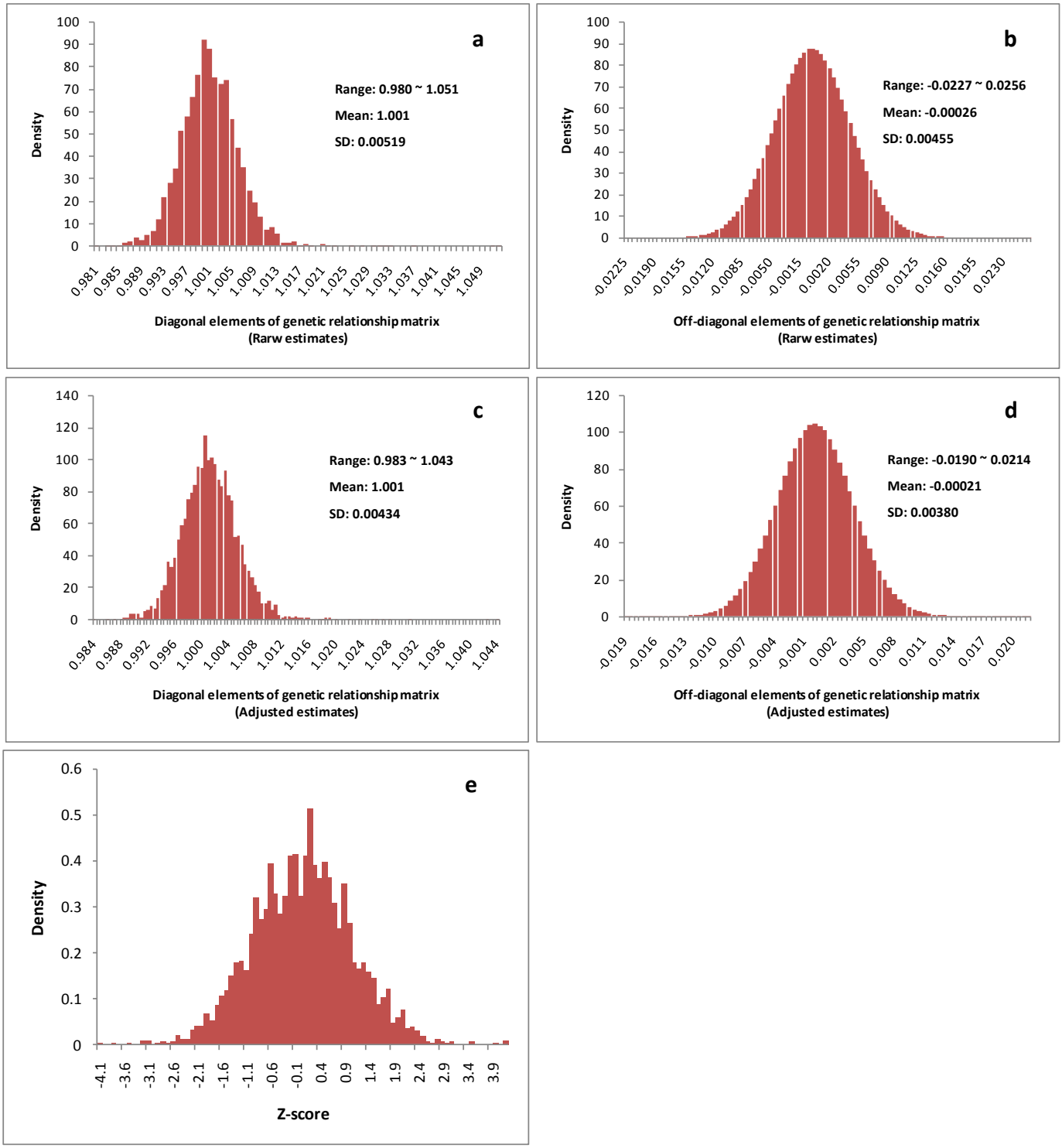
[2]*Department of Psychiatry, Washington University St. Louis, MO, USA*

[3]*Department of Food and Agricultural Systems, University of Melbourne, Parkville 3011, Australia*

[*] To whom correspondence should be addressed. Email: peter.visscher@qimr.edu.au

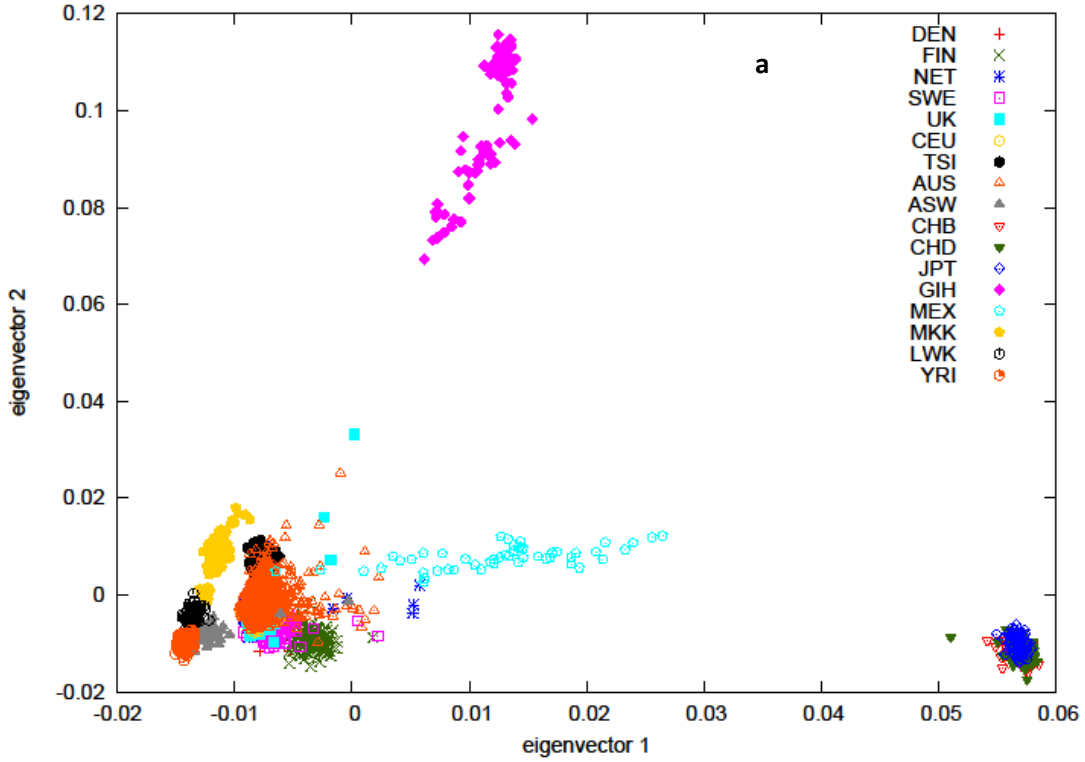
CONTENTS

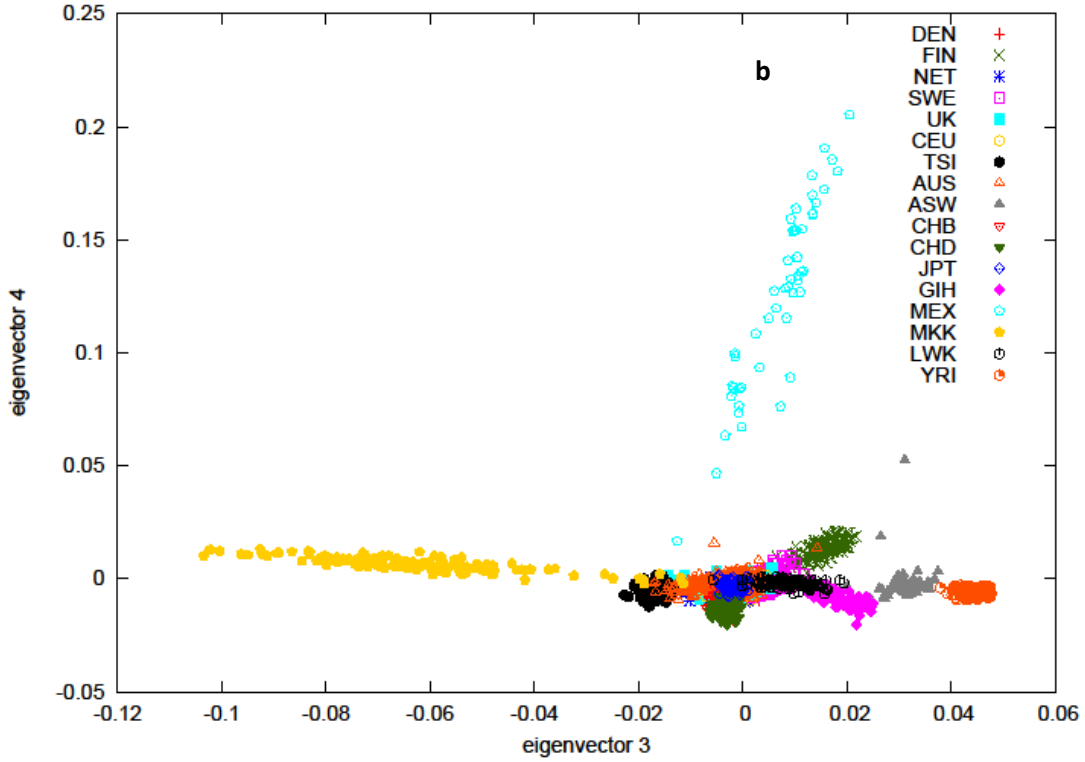
Supplementary Figures

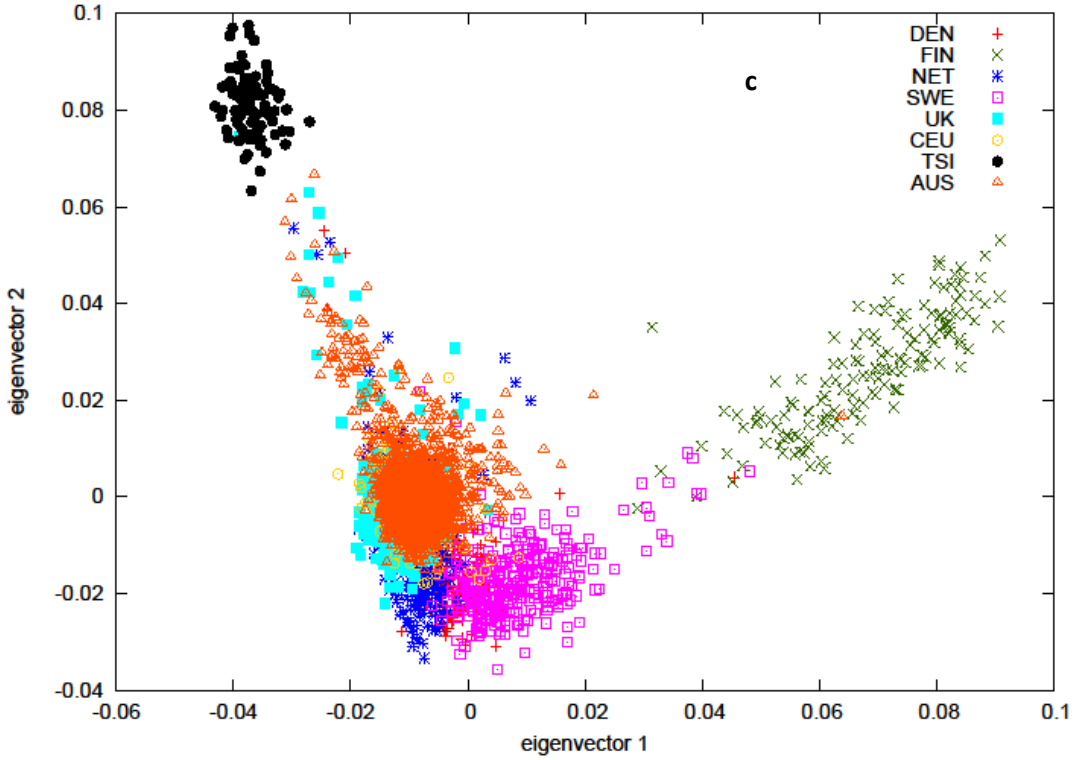


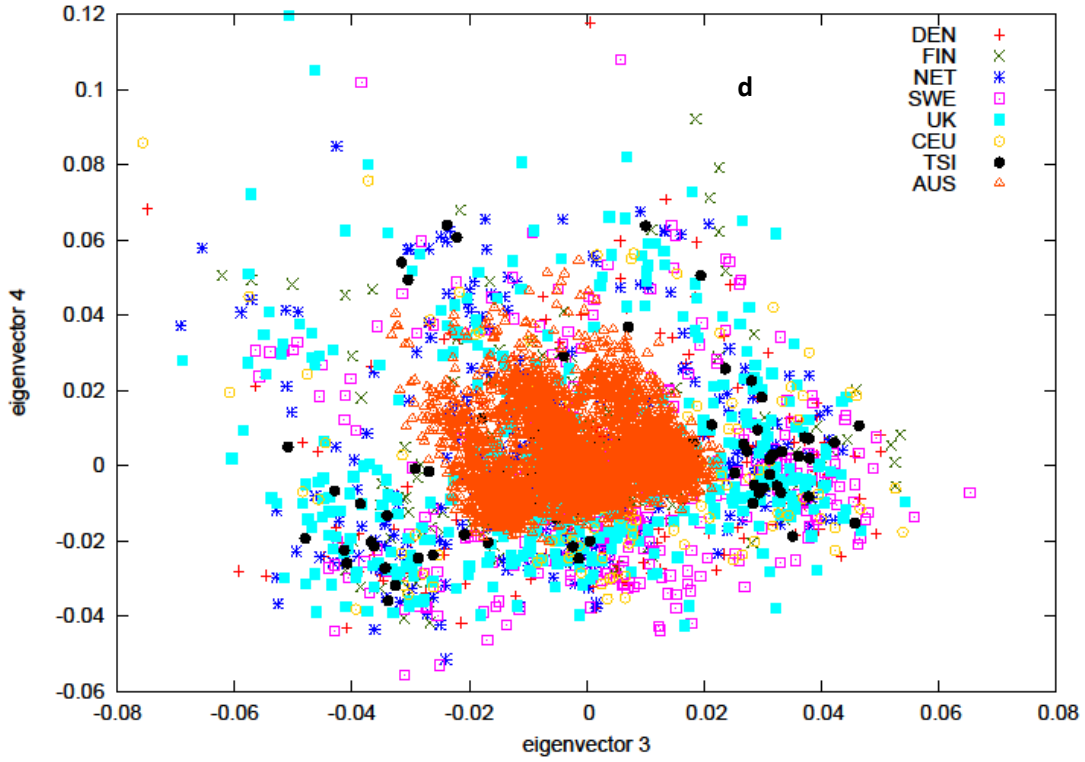
Supplementary Figure 1 Histograms of (**a**) the diagonal and (**b**) the off-diagonal elements of the raw estimates of the genetic relationship matrix, (**c**) the diagonal and (**d**) the off-diagonal

elements of the adjusted estimates of the genetic relationship matrix (assuming c = 0), and (**e**) the phenotypic values (z-scores).The genetic relationship matrix is estimated from 294,831 genome-wide SNPs genotyped on 3,925 individuals.


a
DEN
FIN
NET
SWE
UK
CEU
TSI
AUS
ASW
CHB
CHD
JPT
GIH
MEX
MKK
LWK
YRI
eigenvector 2
eigenvector 1
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.02
-0.01
0
0.01
0.02
0.03
0.04
0.05
0.06

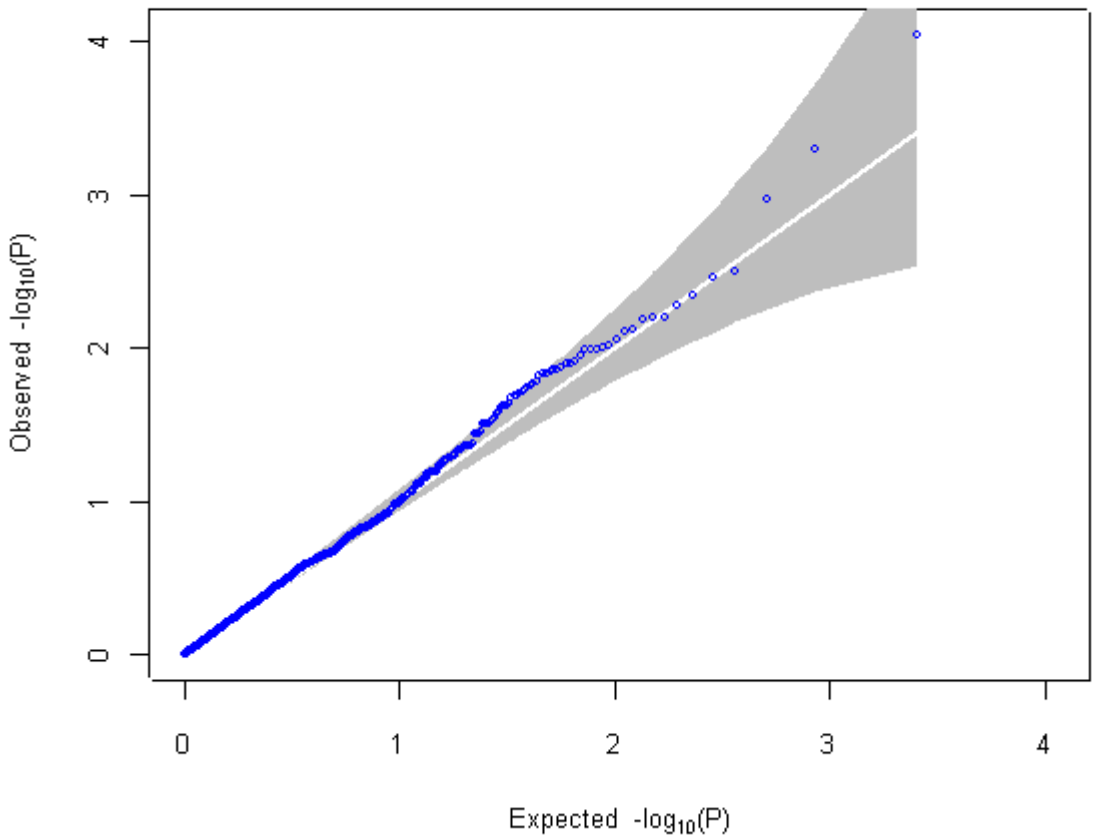

b
DEN
FIN
NET
SWE
UK
CEU
TSI
AUS
ASW
CHB
CHD
JPT
GIH
MEX
MKK
LWK
YRI
eigenvector 4
eigenvector 3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.12
-0.1
-0.08
-0.06
-0.04
-0.02
0
0.02
0.04
0.06


c
DEN
FIN
NET
SWE
UK
CEU
TSI
AUS
eigenvector 2
eigenvector 1
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08
0.1

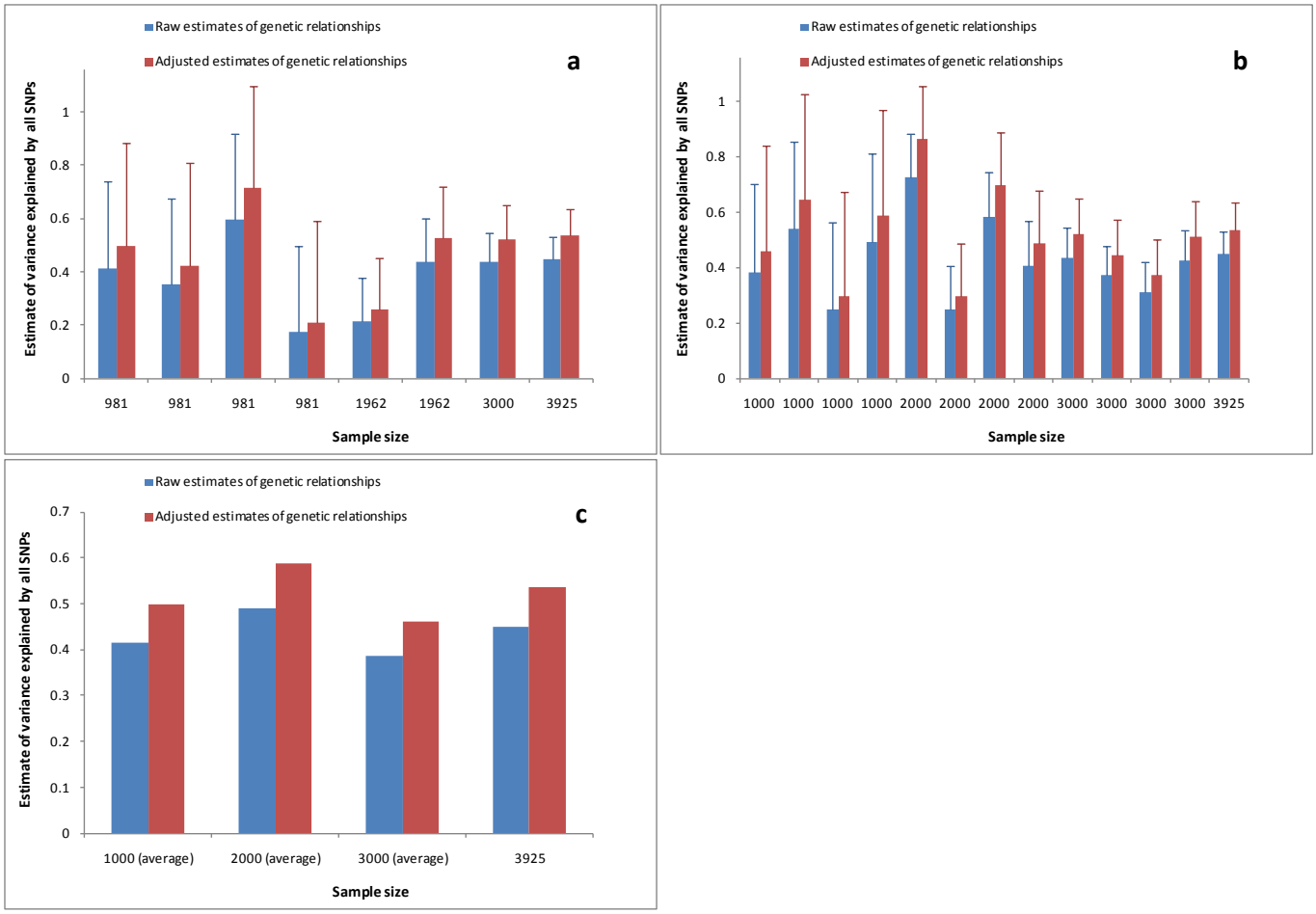

d
DEN
FIN
NET
SWE
UK
CEU
TSI
AUS
eigenvector 4
eigenvector 3
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08

Supplementary Figure 2 Principal component analysis (PCA) of ancestry. In order to identify and remove gross ethnic outliers, the Australian data was combined with 11 global populations from the Hapmap3 project and 5 additional Northern European populations from GenomEUtwin consortium[1]. Only unrelated people from Hapmap 3 were included while 34 previously identified European ethnic outliers from GenomEUtwin were excluded. Population codes and samples sizes are as follows: DEN-Denmark, *n*=161; FIN-Finland, *n*=149; NET- Netherlands, *n*=284; SWE-Sweden, n=302; UK-United Kingdom, n=433; ASW-African ancestry from Southwest USA, *n*=49; CEU-Utah residents with Northern and Western European ancestry from the CEPH collection, *n*=112; CHB-Han Chinese in Beijing, China, *n*=84; CHD-Chinese in Metropolitan Denver, Colorado, *n*=85; GIH-Gujarati Indians in Houston; Texas, *n*=88; JPT-Japanese in Tokyo, Japan, *n*=86; LWK-Luhya in Webuye, Kenya, *n*=90; MEX -Mexican ancestry in Los Angeles; California, *n*=50; MKK-Maasai in Kinyawa, Kenya, n=143; TSI-Tuscans, Italy, *n*=88; YRI- Yoruba in Ibadan, Nigeria *n*=113. PCA, implemented using the EIGENSOFT package[2,3], was thus conducted on a total of 2,317 individuals from 16 populations using ~225K SNPs autosomal SNPs that were genotyped in common between the Hapmap3, GenomEUtwin and present studies. The Australian individuals were not used in the generation of the PCs but were rather projected onto the resulting genetic space. (**a**) The major trend, Principal Component (eigenvector, PC) 1, tends to separate African from non-African population while PC2 separate East Asian from the others. The mean PC1 and PC2 scores of the European populations (DEN, FIN, NET, SWE, UK, CEU and TSI) were used as a reference point and any Australian more than 6 standard deviations from these along PC1 or 2 was deemed to be an ethnic outlier. The plots of (**a**) PC1 versus PC2 and (**b**) PC3 versus PC4 shows the Australian individuals retained for analysis (AUS, *n*=3,925) against the reference population set. These included individuals group with the cluster European samples. In order to show the genetic-geographic dispersion

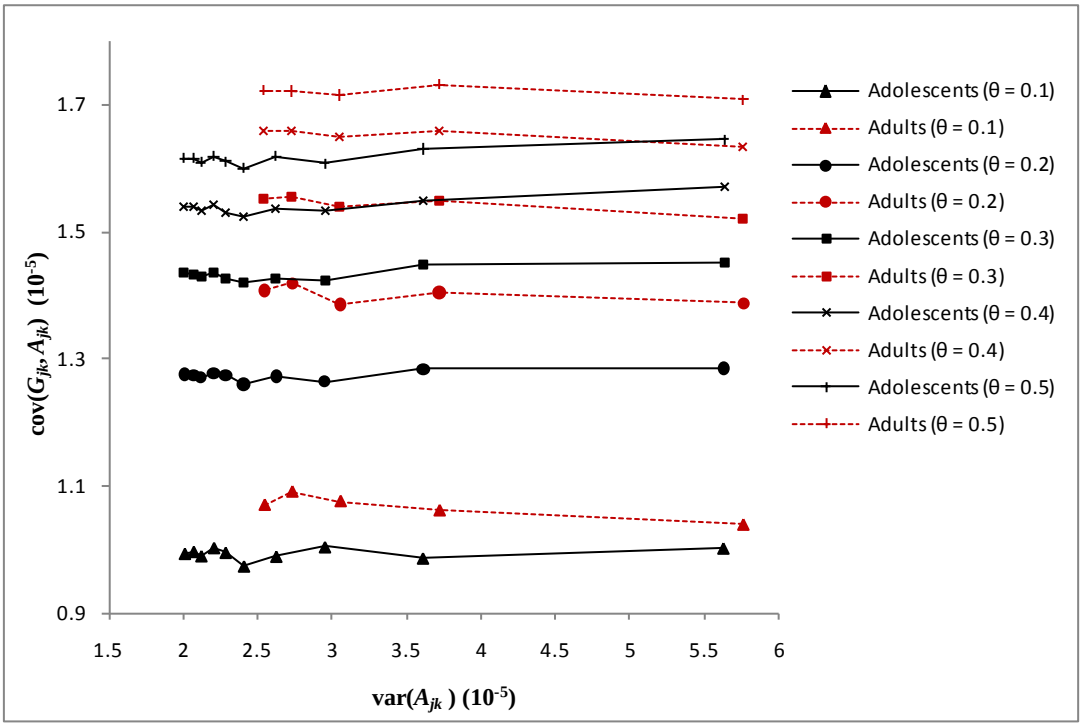
of the Australian individuals across Europe, the PCA was re-conducted using the European populations only (DEN, FIN, NET, SWE, UK, CEU and TSI). (**c**) The plot of PC1 versus PC2 from European-only PCA is consistent with the European North-South and East-West axes. The Australian individuals show close relationship to the UK population and do not show an apparent cline across Europe. (**d**) The plot of PC3 versus PC4 from European-only PCA cannot separate any population from the others.



Supplementary Figure 3 Quantile-quantile plot of the -log10 p-values from association analysis of ancestry informative markers (AIMs) within Europe on height (standardized z-scores after adjustment of age and sex). A set of 1,441 North-South Europeans AIMs is reported by Tian *et al.*[4], 1,286 of which that are in common with the genotyped SNPs in the present study were used in the association analysis. In general, the plotted values fall into the 95% confidence interval (gray area), suggesting that there is no significant inflation of the test statistic for the AIMs.



Supplementary Figure 4 Estimates of variance explained by all SNPs using subsets of individuals (**a**) by randomly splitting the 3,925 individuals into 4 groups (981 individuals in each group), into 2 groups (1,962 individuals in each group), and by taking a single random sample of 3,000 individuals; (**b**) by randomly sampling 1,000, 2,000 or 3,000 individuals with replacement from the 3,925 individuals for 4 replicates. The mean estimate of the 4 repeated samplings is present in (**c**). The error bar is the standard error of the estimate. In each case, the variance explained by all SNPs was estimated using both raw and adjusted estimates of relatedness.



Supplementary Figure 5 Plot of cov(G_{jk}, A_{jk}) against var(A_{jk}) for 10 subsets (50K, 100K, …, 500K) of randomly selected SNPs in the adolescent cohort and 5 subsets (50K, 100K, …, 250K) in the adult cohort. The SNPs in each subset were randomly and evenly split into two groups (see the second section in Online Methods for details). On the x-axis is the empirical variance of the estimated relatedness using the SNPs in the first group, and on the y-axis the covariance between the estimates of relatedness from the two groups of SNPs. Both y and x-axis values are scaled by 10^{-5}.

Supplementary Tables

Supplementary Table 1 Estimates of variance explained by all SNPs by fitting first 2, 4 and 10 principal components (eigenvectors, PCs) from European-only principal component analysis (**Supplementary Fig. 2**) as covariates in the REML analyses.

		Raw estimates of genetic relationships	[a] Adjusted estimates of genetic relationships				
			MAF ≤ 0.1	**MAF ≤ 0.2**	**MAF ≤ 0.3**	**MAF ≤ 0.4**	**MAF ≤ 0.5**
No PC	[b] h^2	0.449	0.836	0.668	0.600	0.563	0.537
	[c] s.e.	0.083	0.155	0.124	0.111	0.104	0.100
First 2 PCs	h^2	0.434	0.807	0.645	0.579	0.544	0.519
	s.e.	0.083	0.155	0.124	0.111	0.105	0.100
First 4 PCs	h^2	0.438	0.816	0.652	0.585	0.549	0.524
	s.e.	0.084	0.156	0.125	0.112	0.105	0.100
First 10 PCs	h^2	0.441	0.821	0.656	0.589	0.553	0.528
	s.e.	0.084	0.157	0.125	0.112	0.105	0.101

[a] Estimate of genetic relationship is adjusted for prediction error under the assumption that the relationship to be predicted is attributed to causal variants with MAF ≤ 0.1, 0.2, 0.3, 0.4 or 0.5;

[b] Estimate of variance explained by all SNPs and [c] its standard error.

Supplementary Table 2 Regression *p*-value of genetic relationships estimated from SNPs on one chromosome against those estimated from SNPs on another chromosome. None of the *p*-values is less than the threshold of 0.00022, which corresponds to an overall type-I error rate of 0.05 after a Bonferroni correction for multiple tests.

Chr. / Chr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	-	0.342	0.757	0.037	0.108	0.961	0.320	0.025	0.856	0.044	0.065	0.858	0.073	0.051	0.722	0.078	0.239	0.099	0.185	0.223	0.047	0.322
2	-	-	0.438	0.011	0.002	0.268	0.240	0.045	0.913	0.070	0.044	0.226	0.143	0.107	0.397	0.664	0.755	0.772	0.126	0.905	0.424	0.171
3	-	-	-	0.012	0.467	0.439	0.272	0.018	0.343	0.217	0.276	0.816	0.544	0.620	0.147	0.847	0.002	0.483	0.146	0.280	0.002	0.222
4	-	-	-	-	0.306	0.341	0.902	0.123	0.100	0.746	0.850	0.627	0.494	0.622	0.351	0.259	0.641	0.510	0.995	0.192	0.124	0.284
5	-	-	-	-	-	0.541	0.093	0.904	0.043	0.356	0.314	0.587	0.161	0.137	0.079	0.012	0.074	0.550	0.091	0.044	0.104	0.095
6	-	-	-	-	-	-	0.137	0.396	0.027	0.970	0.056	0.091	0.745	0.157	0.134	0.937	0.332	0.194	0.694	0.433	0.748	0.968
7	-	-	-	-	-	-	-	0.777	0.050	0.030	0.076	0.018	0.168	0.144	0.425	0.010	0.176	0.777	0.394	0.323	0.616	0.711
8	-	-	-	-	-	-	-	-	0.448	0.945	0.051	0.639	0.151	0.027	0.274	0.070	0.980	0.718	0.859	0.073	0.908	0.362
9	-	-	-	-	-	-	-	-	-	0.839	0.738	0.393	0.838	0.814	0.386	0.525	0.215	0.017	0.594	0.129	0.924	0.149
10	-	-	-	-	-	-	-	-	-	-	0.515	0.022	0.319	0.573	0.624	0.116	0.079	0.139	0.687	0.276	0.087	0.747
11	-	-	-	-	-	-	-	-	-	-	-	0.491	0.354	0.755	0.659	0.107	0.686	0.365	0.458	0.262	0.572	0.558
12	-	-	-	-	-	-	-	-	-	-	-	-	0.402	0.889	0.789	0.685	0.176	0.147	0.925	0.726	0.889	0.071
13	-	-	-	-	-	-	-	-	-	-	-	-	-	0.038	0.742	0.793	0.147	0.007	0.057	0.819	0.078	0.244
14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.497	0.219	0.966	0.814	0.340	0.445	0.997	0.888
15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.632	0.930	0.541	0.170	0.716	0.680	0.004
16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.064	0.920	0.015	0.822	0.046	0.500
17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.159	0.318	0.055	0.421	0.972
18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.565	0.311	0.434	0.466
19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.037	0.624	0.746
20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.717	0.461
21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.148
22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Supplementary Note

Simulation based on real genotype data

We simulated a quantitative trait based on the observed genotype data of 3,925 individuals and 294,831 SNPs. We performed the simulations in two scenarios: I) randomly sample *m* causal variants from all of the SNPs; II) randomly sample *m* causal variants from the SNPs with MAF ≤ 0.1. We generated the effects of causal variants (*u*) from a standard normal distribution, and calculated the genetic score of each individual by $g = \sum_i z_i u_i$, where *u* is the effect size and *z* is coded as 0, 1 or 2 for genotype *qq*, *Qq* or *QQ*. We generated residual non-genetic effects (*e*) from normal distribution with mean of 0 and variance of $\text{var}(g)(1/h^2 - 1)$, where var(*g*) is the empirical variance of genetic score, and h^2 is the heritability. We calculated the phenotypic value of each individual by $y = g + e$. We set the number of causal variants of 2000 and 3000, and h^2 of 50% and 80%.

With the simulated phenotype, we performed the following analyses:

1) Estimate h^2 using genetic relationships estimated from all of SNPs including the causal variants in both scenarios.
2) Estimate h^2 using genetic relationships estimated from SNPs with exclusion of causal variants in both scenarios.
3) Exclude the causal variants from analysis and adjust the estimates of genetic relationships using equation [9] with $c = 0$ in scenario I and $c = 6.2\times10^{-6}$ in scenario II.
4) Run the simulation with 30 replicates (randomizing the effects of causal variants in each replicate) and summarize the average h^2 over all of the simulation replicates.

Nature Genetics: doi:10.1038/ng.608

Correlated segregation of genetic and environmental factors?

One theoretical possibility is that the environment segregates proportionally to the degree of relatedness, so that we detect an environmental rather than genetic effect. This is extremely unlikely. Not only did we select individuals that had a mean genetic relationship that was low (< 0.025, but most pairs were closer to zero), but for distant relatives the proportional variation around the expected degree of relatedness increases[5], so that the realised relationships (which we use) may differ widely from the expected relationships based on pedigree (which may be correlated with environment).

Variation in identity for relative pairs *m* meioses apart was quantified from results given by Guo (1996)[6], Hill (1993)[7] and Visscher (2009)[5]. We assume that the autosomal genome-length is 35.78 Morgan[8], corresponding to ~3000 Mb. The Table below show the mean and variation in identity for relative pairs that m meioses apart (e.g. *m* = 2 is individual-grandparent).

m	Additive relationship (*a*)	Mean genome-length shared (Mb)	SD(*a*)	CV(*a*)
5	1 / 64 = 0.01563	93.8	0.0136	0.44
6	1 / 128 = 0.00781	46.9	0.0091	0.58
7	1 / 256 = 0.00391	23.4	0.0060	0.77
8	1 / 512 = 0.00195	11.7	0.0040	1.03
9	1 / 1024 = 0.00098	5.9	0.0027	1.36
10	1 / 2048 = 0.00049	2.9	0.0018	1.82
11	1 / 4096 = 0.00024	1.5	0.0012	2.44
12	1 / 8192 = 0.00012	0.7	0.0008	3.29

Our observed SD in identity is < 0.004 (**Supplementary Fig. 1d**), corresponding to relative pairs 8 and 9 meioses apart. These results show that the SD in realised relationship gets so large relative to the mean that for distant pairs of relatives there will be an overlap in realised relatedness across distinct groups of relatives based upon the expected values.

If genetic and environmental factors co-segregate then this confounding would also apply to SNP-trait associations in genome-wide association studies, yet the SNP-height

Nature Genetics: doi:10.1038/ng.608

associations that have been reported in the literature are well-replicated and not due to spurious associations.

The ancestry of the people in our sample is mainly from the British Isles (UK and Ireland), yet the measurements were taken in Australia. Therefore, for relatedness and environment to be correlated among pairs of individuals that are 8-9 meioses apart would imply that people who were ancestrally more related in the British Isles would have a more similar environment in Australia, which is unlikely.

Additive and non-additive models

The way that heritabilities are estimated in humans (and other species) is by fitting statistical models that capture the part of the resemblance between relatives that is caused by additive genetic effects[9]. Usually, genome-wide association studies are also analysed using an additive model, i.e. to capture additive genetic variation. Estimates of the heritability for height in humans suggest that ~80% of phenotypic variation is due to additive genetic effects and this is the benchmark when the problem of 'missing heritability' is discussed[10,11]. In our study, we seek to explain the conventional heritability of height, that is, the additive genetic variance as a proportion of the phenotypic variance. Thus an additive model for the SNPs is the correct choice. Gene-gene or gene-environment interactions or epigenetic effects are not directly relevant to our analysis and conclusions because they do not contribute to the estimate of the narrow sense heritability that we seek to explain.

References

1. McEvoy, B.P. et al. Geographical structure and differential natural selection among North European populations. *Genome Res.* **19**, 804-14 (2009).

Nature Genetics: doi:10.1038/ng.608

2. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

3. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-9 (2006).

4. Tian, C. et al. Analysis and Application of European Genetic Substructure Using 300 K SNP Information. *PLoS Genet.* **4**, e4 (2008).

5. Visscher, P.M. Whole genome approaches to quantitative genetics. *Genetica* **136**, 351-8 (2009).

6. Guo, S.W. Variation in genetic identity among relatives. *Hum. Hered.* **46**, 61-70 (1996).

7. Hill, W.G. Variation in genetic identity within kinships. *Heredity* **71**, 652-653 (1993).

8. Kong, X. et al. A combined linkage-physical map of the human genome. *Am. J. Hum. Genet.* **75**, 1143-8 (2004).

9. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era--concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255-66 (2008).

10. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21 (2008).

11. Manolio, T.A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).

Nature Genetics: doi:10.1038/ng.608