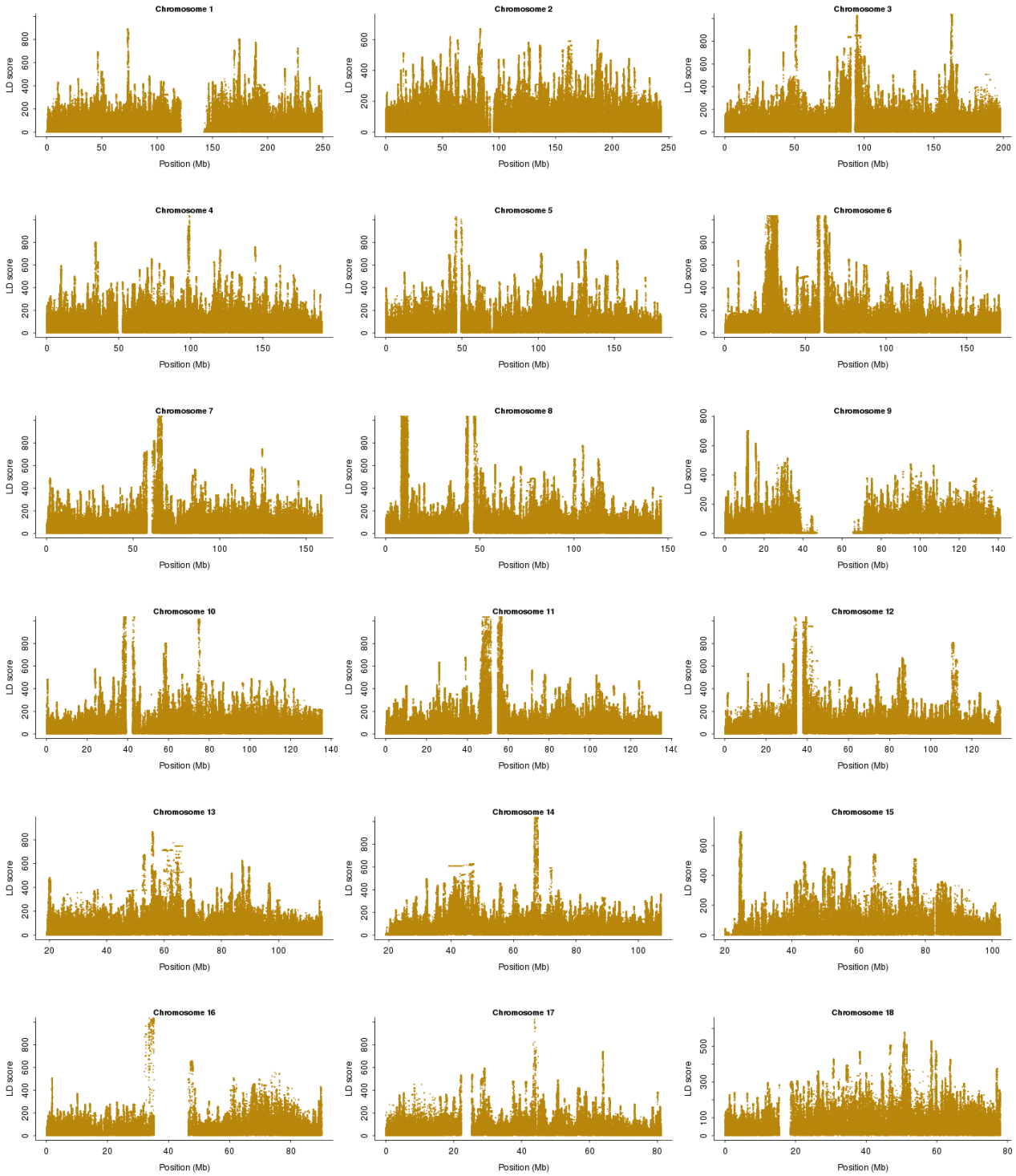
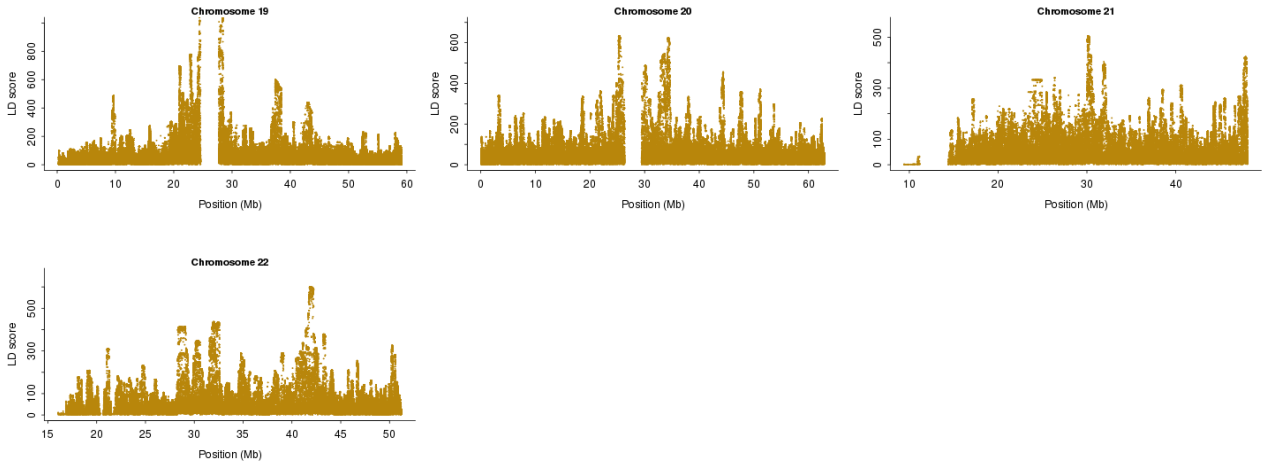
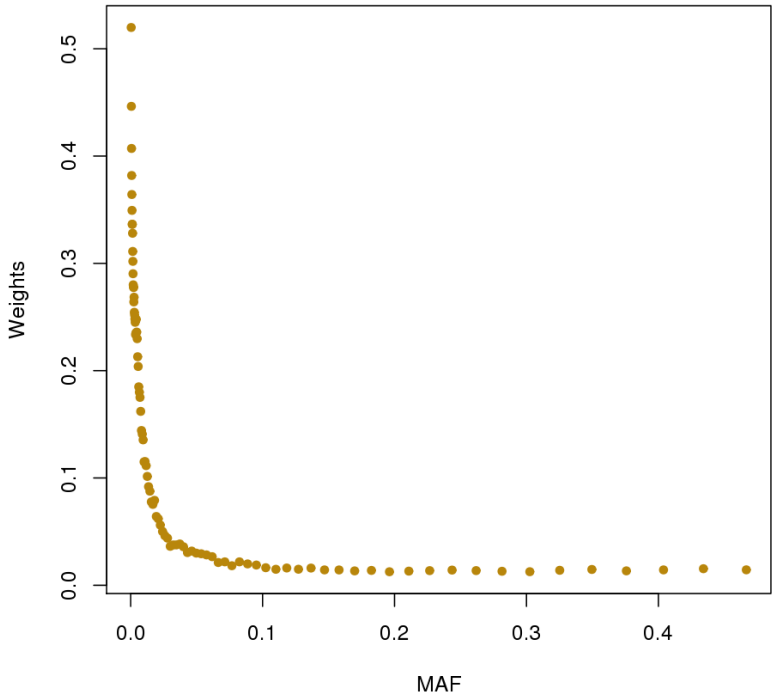
Supplementary Figures

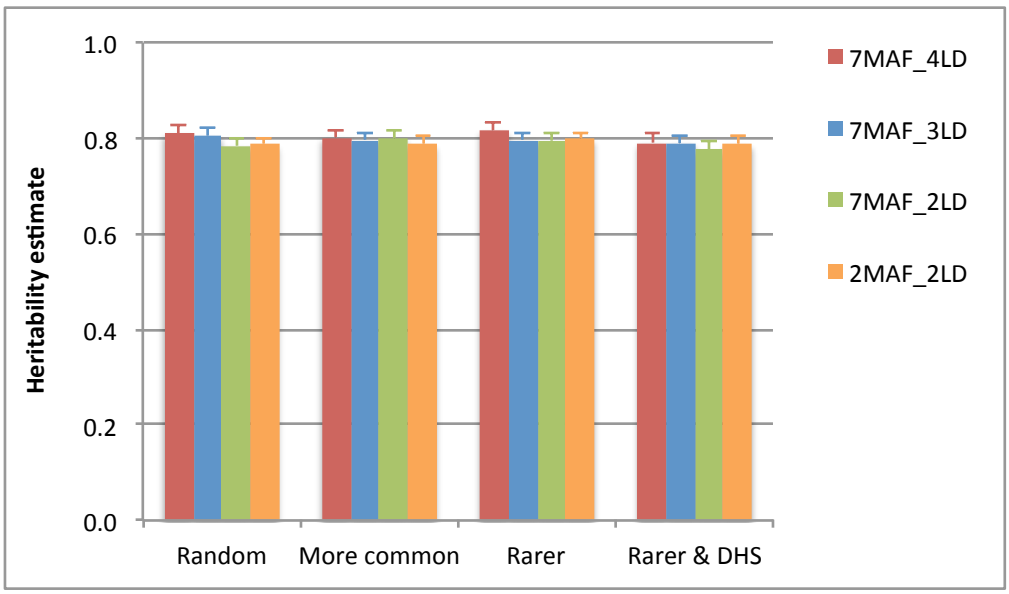




Supplementary Figure 1 Region-specific LD heterogeneity of the genome. LD score of each variant is defined as the sum of LD r^2 between the target variant and all variants (including the target variant) within ±10Mb distance. For better graphic display, the y-axis is truncated at 1,000.

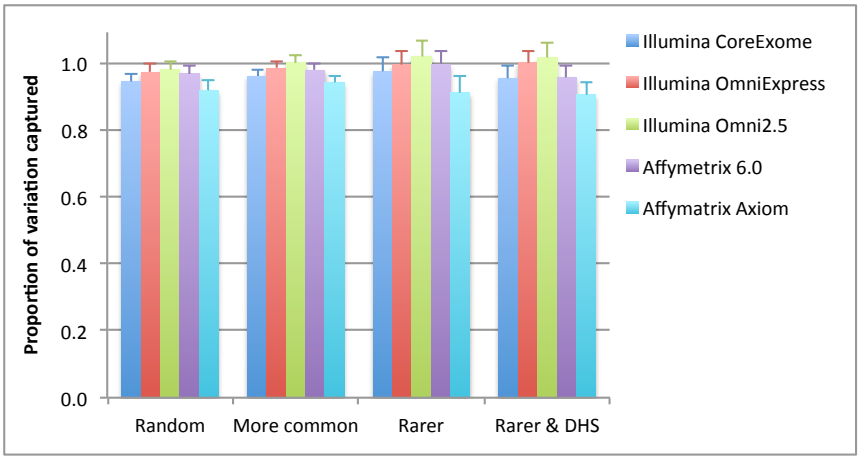


Supplementary Figure 2 LDAK weights vs. MAF for the 233,588 variants on chromosome 22 in the UK10K-WGS data. For better graphical presentation, shown are the mean weights in 100 MAF bins.

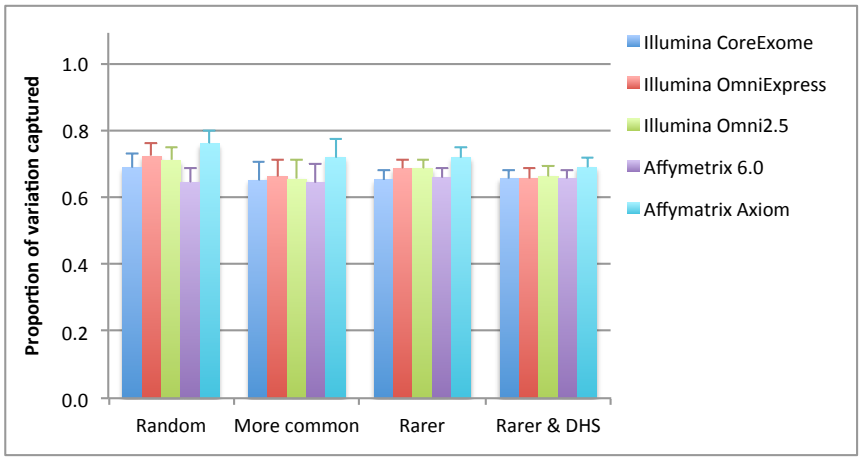


Supplementary Figure 3 GREML-LDMS estimate of h^2_{WGS} using sequence variants and simulated phenotype based on the UK10K-WGS data. Each column represents the mean estimate from 200 simulations. Error bar is the s.e. of the mean estimate. The true h^2 parameter is 0.8 for the simulated traits (see Online Methods for the 4 simulation scenarios). Variants are stratified based on the distribution of the segment-based mean LD score (see Online Methods for details). 4LD: first, second, third and forth quartiles. 3LD: first, (second + third), and forth quartiles. 2LD: (first + second), and (third + forth) quartiles. In each LD group, variants are further stratified by MAF. 7MAF: variants are stratified into 7 MAF groups, i.e. MAF $\leq$ 0.001, 0.001 < MAF $\leq$ 0.01, 0.01 < MAF $\leq$ 0.1, 0.1 < MAF $\leq$ 0.2, 0.2 < MAF $\leq$ 0.3, 0.3 < MAF $\leq$ 0.4 and 0.4 < MAF $\leq$ 0.5. 2MAF: variants are stratified into 2 MAF groups, i.e. MAF $\leq$ 0.01 and 0.01 < MAF $\leq$ 0.5.

(a) Common variants (MAF ≥ 0.01)

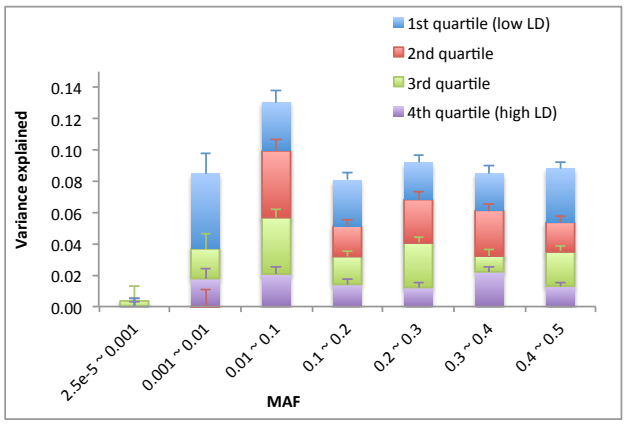


(b) Rare variants (MAF < 0.01)

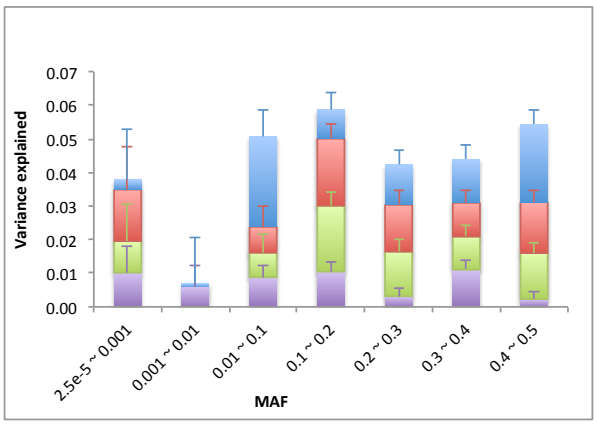


Supplementary Figure 4 Proportion of variation at sequence variants captured by 1KGP imputation in the UK10K-WGS data. Details of the 4 simulation scenarios can be found in Online Methods. The estimates are from GREML-LDMS analysis (4 LD groups with each LD group being further stratified into 7 MAF groups, 28 groups in total) using all the 1KGP-imputed variants after QC (without filtering variants for IMPUTE-INFO). The proportion of variation at sequence variants captured by imputation (i.e. multi-variant tagging) is defined as the estimate of phenotypic variance explained by 1KGP-imputed variants summed over all the relevant groups (i.e. 20 groups for common variants and 8 groups for rare variants in the GREML-LDMS analysis) divided by true simulation parameter (i.e. variance explained by causal variants). Plotted value is the mean estimate from 200 simulations. Error bar is the s.e. of the mean estimate. Panel (a): common variants. Panel (b): rare variants. The multi-variant tagging averaged across SNP arrays and simulation scenarios is ~0.97 for common variants and ~0.68 for rare variants.

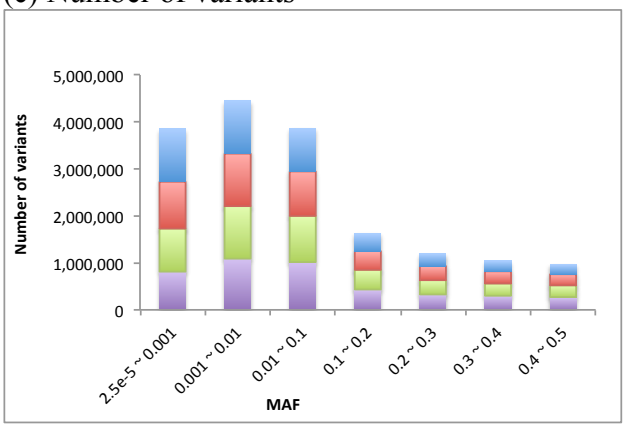
(a) Height



(b) BMI

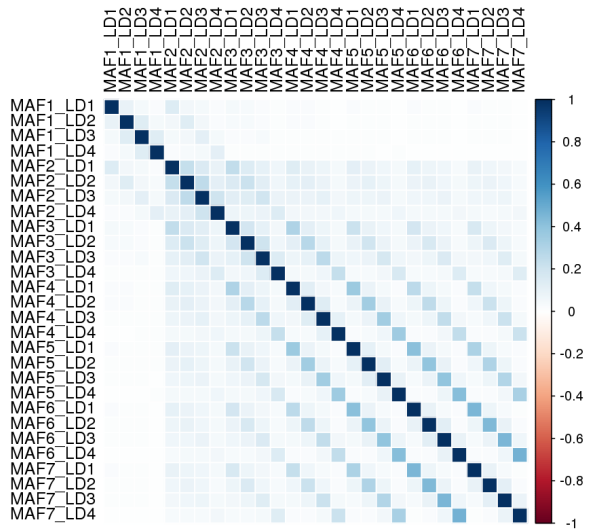


(c) Number of variants

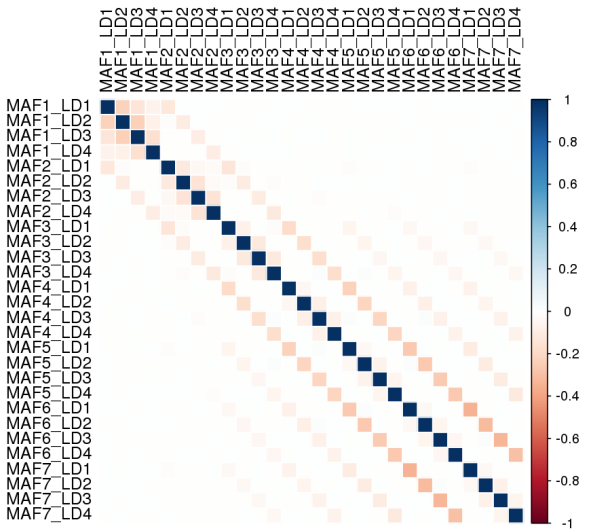


Supplementary Figure 5 Estimate of variance explained by 1KGP-imputed variants stratified by MAF and LD for height and BMI. The estimates of variance explained are from the GREML-LDMS analyses of fitting all the 28 genetic components simultaneously in the combined data from 7 GWAS cohorts (44,126 unrelated individuals and 17M variants). Segmental LD score increases from the 1st to 4th quartiles (See Online Methods for the LD stratified approach). Error bar represents the standard error. For better display, negative estimates are not shown on the figure, which are available in **Supplementary Table 3**. Panel (c) shows the number of variants in each of the 28 groups.

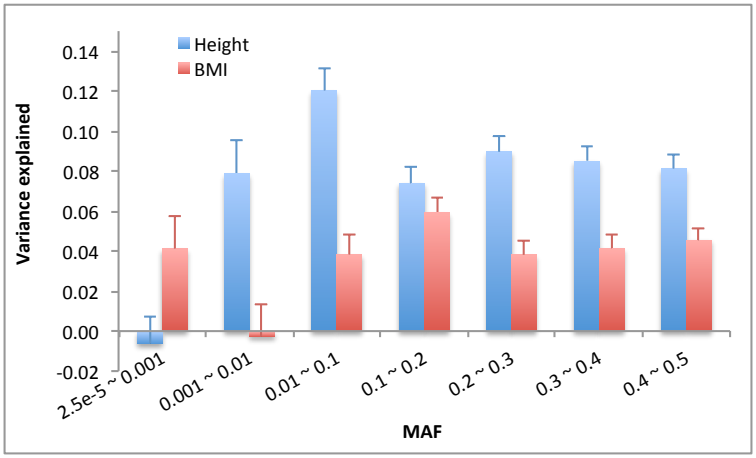
(a) Correlation of GRM



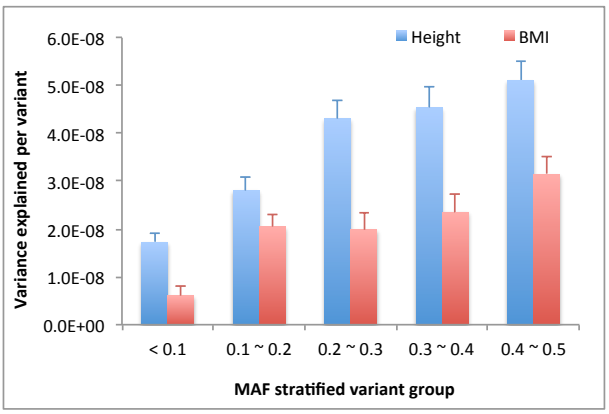
(b) Correlation of the estimate of genetic variance



Supplementary Figure 6 Correlation of GRM (or estimate of variance component) between each pair of the variant groups for the GREML-LDMS analysis in the combined GWAS data set. There are ~17M 1KGP-imputed variants stratified by MAF and segment-based LD score into 28 groups (Online Methods). MAF1: MAF $\leq$ 0.001; MAF2: 0.001 < MAF $\leq$ 0.01, MAF3: 0.01 < MAF $\leq$ 0.1; MAF4: 0.1 < MAF $\leq$ 0.2; MAF5: 0.2 < MAF $\leq$ 0.3; MAF6: 0.3 < MAF $\leq$ 0.4; MAF7: 0.4 < MAF $\leq$ 0.5. LD1, LD2, LD3 and LD4 are the four quartiles of the segment based LD score distribution (from low to high), respectively. Panel (a) shows the correlation of the off-diagonal elements of GRM between each pair of groups. Panels (b) shows the correlation of the estimate of genetic variance ($\hat{\sigma}_v^2$) between each pair of groups for height. The correlation of $\hat{\sigma}_v^2$ is calculated as $\mathrm{cov}(\hat{\sigma}^2_{v(i)},\hat{\sigma}^2_{v(j)}) / \sqrt{\mathrm{var}(\hat{\sigma}^2_{v(i)})\,\mathrm{var}\,\hat{\sigma}^2_{v(j)})}$. The GRMs of different groups are positively correlated, in particular for those in the same LD group (e.g. a maximum correlation of 0.471 between groups MAF6_LD4 and MAF7_LD4), which leads to negative correlation of $\hat{\sigma}_v^2$. The correlation of $\hat{\sigma}_v^2$ for BMI is almost identical to that for height and is therefore not shown.

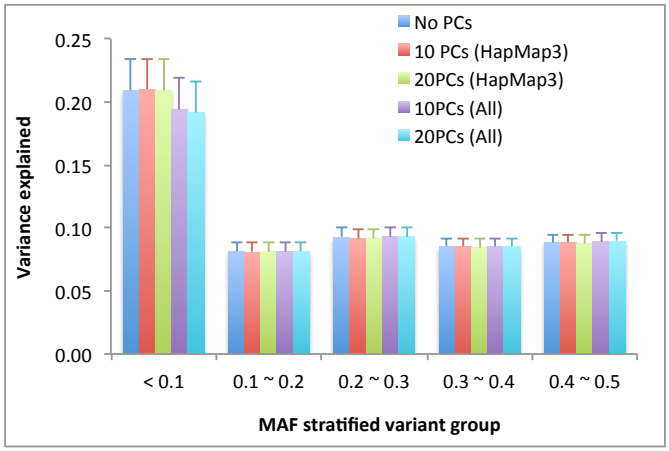


Supplementary Figure 7 Estimate of variance explained by 1KGP-imputed variants stratified by MAF for height and BMI. The estimates of variance explained are from the GREML-MS analyses of fitting all the 7 genetic components simultaneously in the combined data set (44,126 unrelated individuals and 17M variants). Error bar is the standard error. This figure shows that rare variants (MAF $\leq$ 0.01) do explain a significant amount of variance for both height (variants with 0.001 < MAF $\leq$ 0.01) and BMI (variants with MAF $\leq$ 0.001), consistent with the results from GREML-LDMS (**Supplementary Table 3 and Supplementary Fig. 5**).

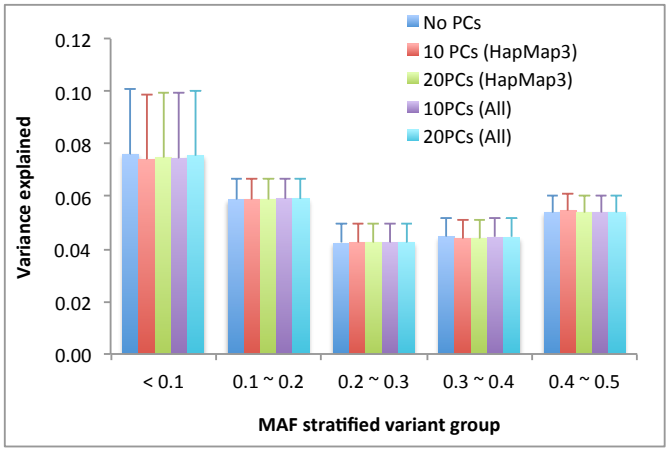


Supplementary Figure 8 Variance explained per variant. The estimate of variance explained per variant is calculated as the $\hat{h}^2_{1\text{KGP}(i)} / m_i$ where $\hat{h}^2_{1\text{KGP}(i)}$ is the estimate of variance explained by the variants in *i*-th MAF group (shown on x-axis) from the GREML-LDMS analysis in the combined set data, and m_i is the number of variants in the *i*-th MAF group.

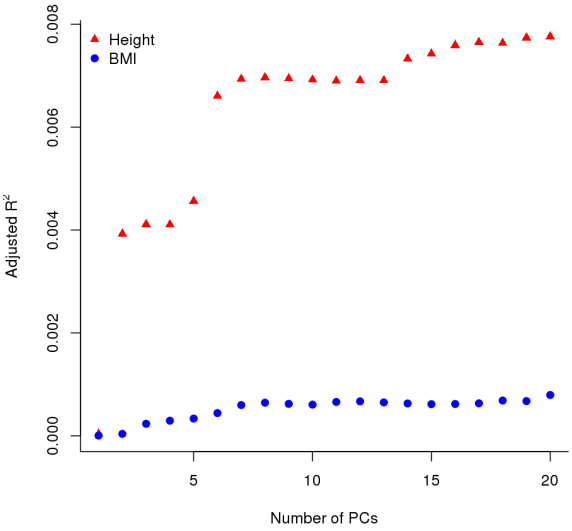
(a) GREML-LDMS estimate of h^2_{1KGP} (height)



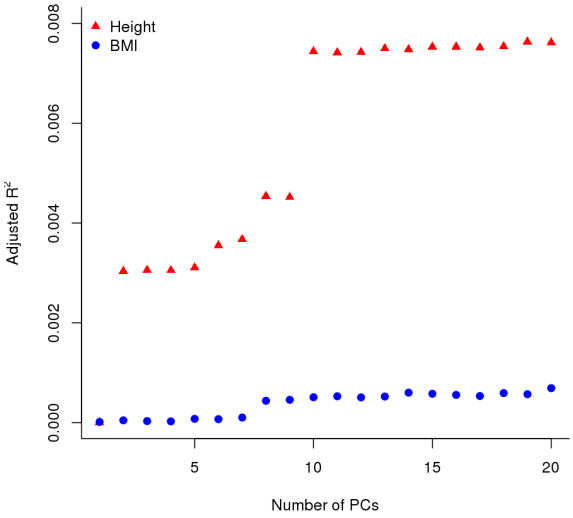
(b) GREML-LDMS estimate of h^2_{1KGP} (BMI)



(c) Adjusted R^2-adj for the first PCs (HapMap3)



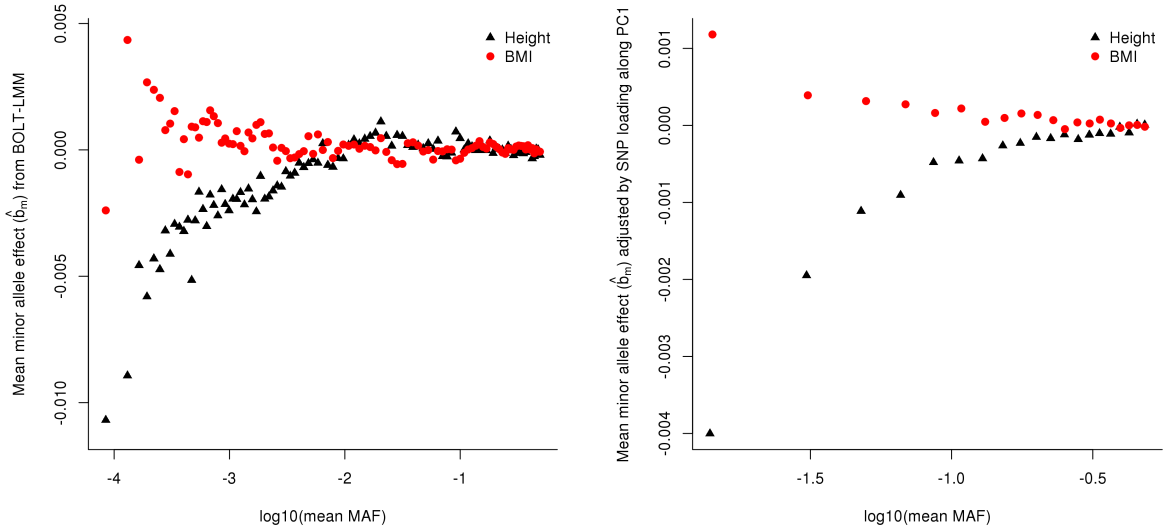
(d) Adjusted R^2 for the first PCs (All)



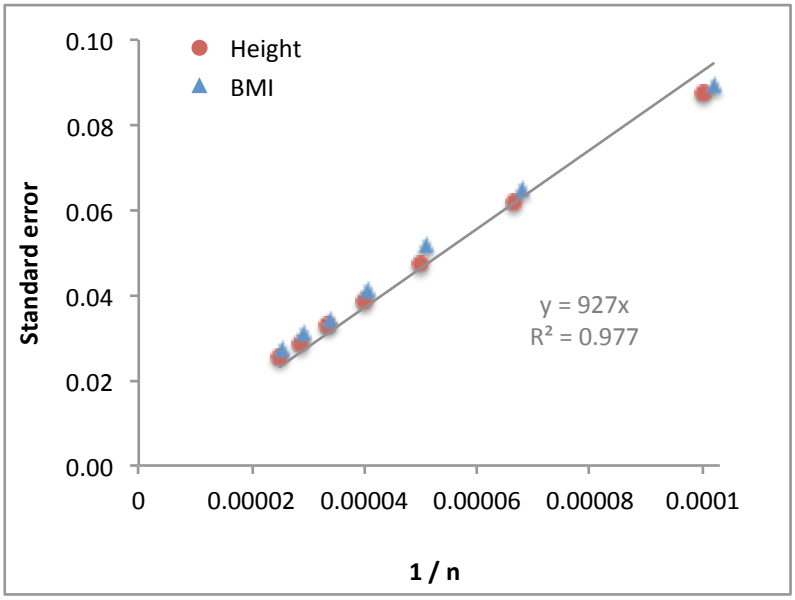
Supplementary Figure 9 Variance explained by the first PCs for height and BMI and the GREML-LDMS estimates of h^2_{1KGP} with different number of PCs estimated from different sets of variants.

Shown are the results from the analyses in the combined data set. Shown in panels (a) and (b) are the results from the GREML-LDMS analyses fitting different numbers of PCs as fixed covariates for (a) height and (b) BMI. 10 PCs (HapMap3): GREML-LDMS analysis fitting the first 10 PCs estimated from the common variants on HapMap3. 20 PCs (All): GREML-LDMS analysis fitting the first 20 PCs estimated from all the variants. In panels (c) and (d), plotted are cumulative variance explained by the first x PCs, which are the adjusted multiple regression R^2 of height (or BMI) phenotype on the first x PCs calculated from using (c) all common variants on HapMap3 or (d) all the variants.

(a) (b)

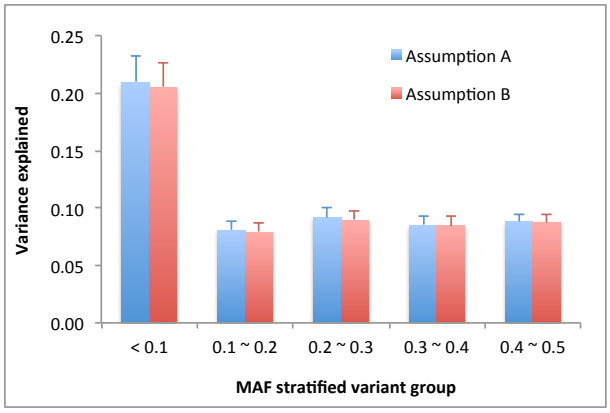


Supplementary Figure 10 Evidence that height- and BMI-associated genetic variants being under natural selection is not driven by population stratification. We re-ran the analyses as presented in **Fig. 4c** and **4d** using slightly different methods. In panel (b), we performed mixed linear model based association analyses using BOLT-LMM[1], association tests of ~17M 1KGP-imputed variants (each fitted as a fixed effect) with ~1.2M common variants on the HapMap3 fitted as random effects to control for population stratification. In panel (b), we calculated the loading of each SNP along first PC (PC computed from all ~17M 1KGP-imputed variants) following the method described in Galinsky et al.[2], and adjusted the minor allele effect (b_m) by SNP loading using linear regression. The results are almost identical as those shown in **Fig. 4c** and **4d**.

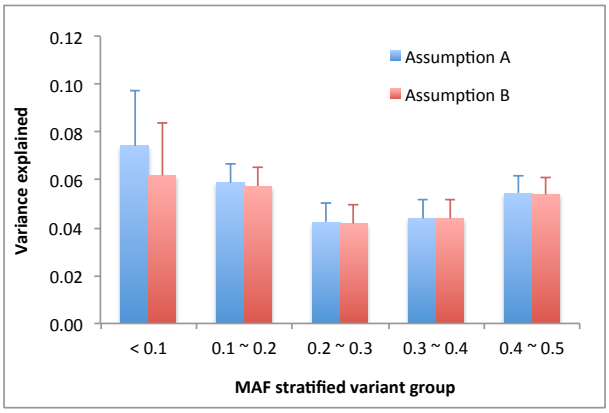


Supplementary Figure 11 Standard error (s.e.) of the estimate from GREML-LDMS is approximately inversely proportional to sample size (n). We randomly sampled the 10,000 to 40,000 samples by steps of 5,000 from the combined GWAS data set (n = 44,126 unrelated individuals), and repeated the GREML-LDMS analyses for height and BMI using 1KGP-imputed variants as described in Online Methods. The result shows that the s.e. of GREML-LDMS estimate of h^2_g (variance explained by all sequence or imputed sequence variants) is approximately 927 / n.

(a) Height

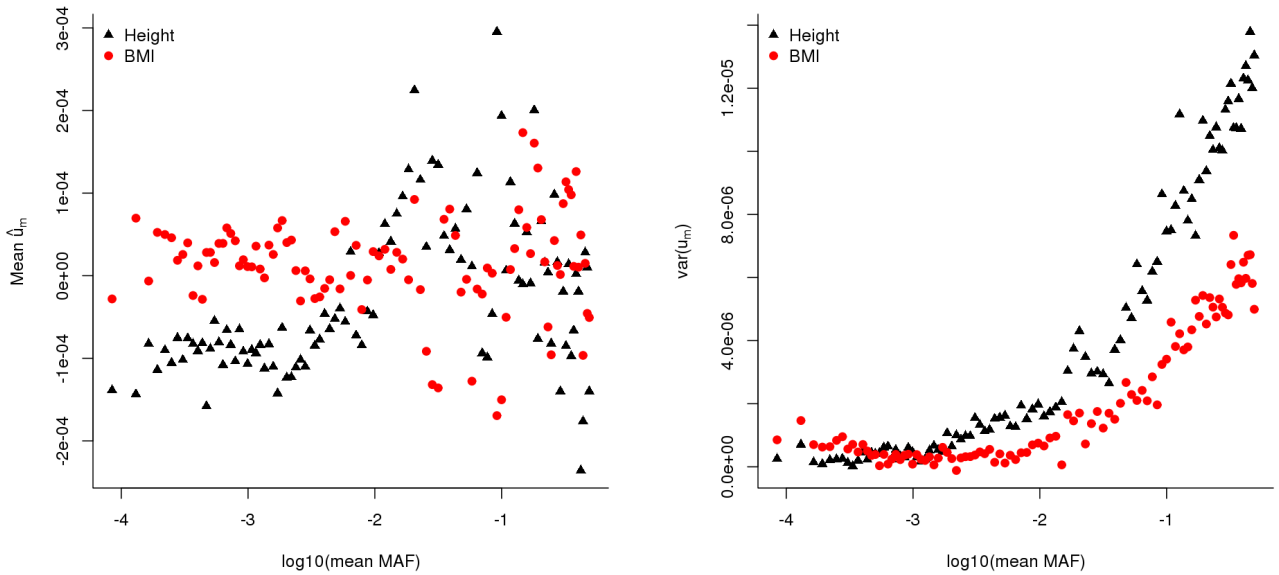


(b) BMI

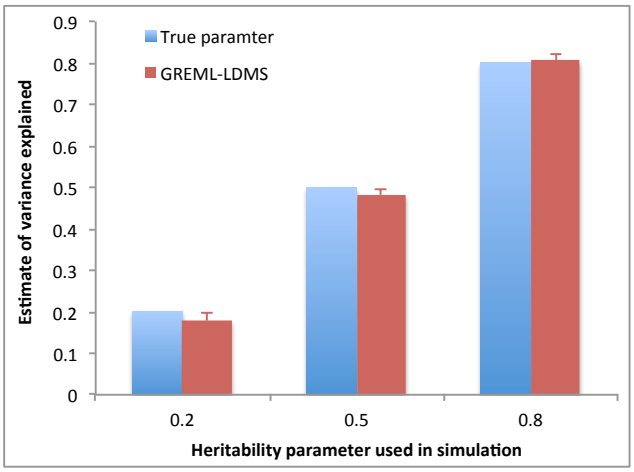


Supplementary Figure 12 Estimates of variance explained by 1KGP-imputed variants stratified by MAF. The results are from the GREML-LDMS analyses of fitting 28 genetic components in the combined GWAS data set of 44,126 unrelated individuals for (a) height and (b) BMI. Shown is the sum of estimates in each of the 5 MAF groups. Error bar is the standard error. Assumption A: assuming a normal distribution of the effect sizes corresponding the standardised genotype variables (default assumption for the GREML methods). Assumption B: assuming a normal distribution of the allelic substitution effects. For height, the log likelihood (logL) of the model is -20189.2 (AIC = 40458.4 and BIC = 40564.0) under Assumption A, and is 20189.2 (AIC = 40458.3 and BIC = 40563.9) under Assumption B. For BMI, the logL is -21234.5 (AIC = 42548.9 and BIC = 42654.4) under Assumption A, and -21235.7 (AIC = 42551.3 and BIC = 42656.8) under Assumption B. These results suggest that the two models fit almost equally well with the data for both height and BMI.

(a) (b)



(c)

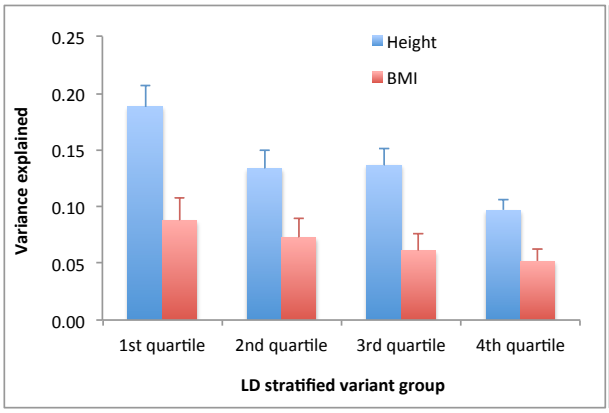


Supplementary Figure 13 Simulations with effect sizes of causal variants sampled from the distribution estimated from real data for height. It is shown in **Fig. 4c** that mean $\hat{b}_m$ is correlated with log10(mean MAF) across 100 MAF bins for both height and BMI, where each MAF bin is the 1% quartile of MAF distribution, and b_m is defined as the effect size of the minor allele of a variant. Using the same data as in **Fig. 4c**, we show in panel (a) that mean $\hat{u}_m$ is also correlated log10(mean MAF) for height (r = 0.44, P_{permu} = 1.3e-5) but not for BMI (r = -0.03, P_{permu} = 0.76), where $u_m = b_m\sqrt{2p(1-p)}$ with p being the MAF and P_{permu} is calculated from 1 million permutations. The relationship between

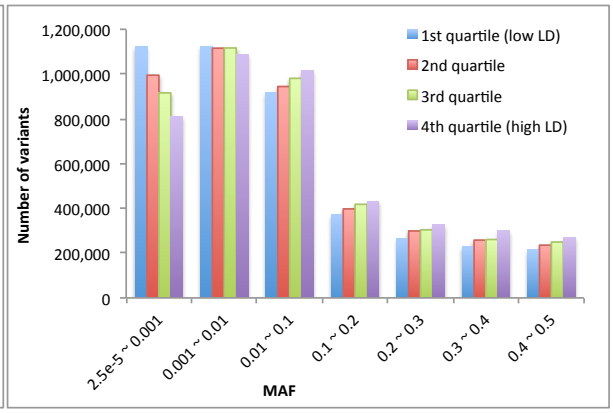
u_m and MAF violates the assumption of the GREML methods. We therefore performed simulations (based on the UK10K-WGS data) to test the robustness of GREML-LDMS to such a correlation. The variance of u_m, as shown on the y-axis of panel (b), is calculated as $\text{var}(u_m) = \text{var}(\hat{u}_m) - SE^2(\hat{u}_m)$. The simulation strategy largely follows that in Online Methods with an important difference that we simulated u_m from a normal distribution with mean from that shown on the y-axis of panel (a) and variance from that of panel (b) for height according to its MAF. For the ease of stimulation, we used the absolute value of var(u_m) if the estimate of var(u_m) is negative. Shown on panel (c) are the mean estimates from 200 simulations. Error bar is the s.e. of the mean estimate.

Nature Genetics: doi:10.1038/ng.3390

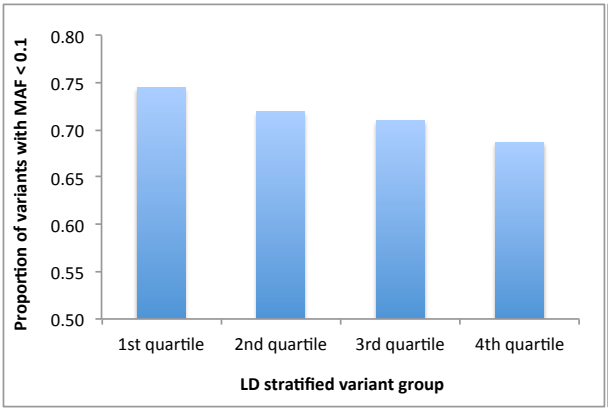
(a) Original estimate



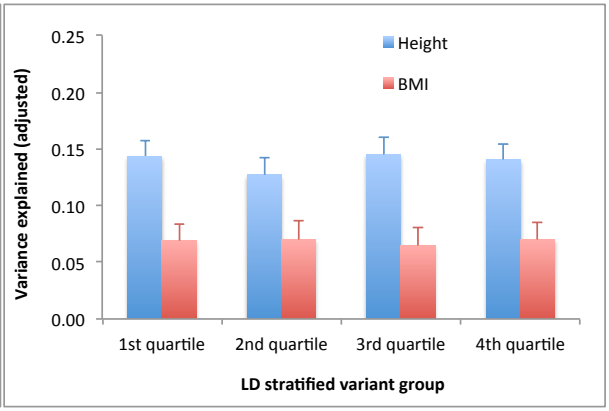
(b) Number of variants in each group



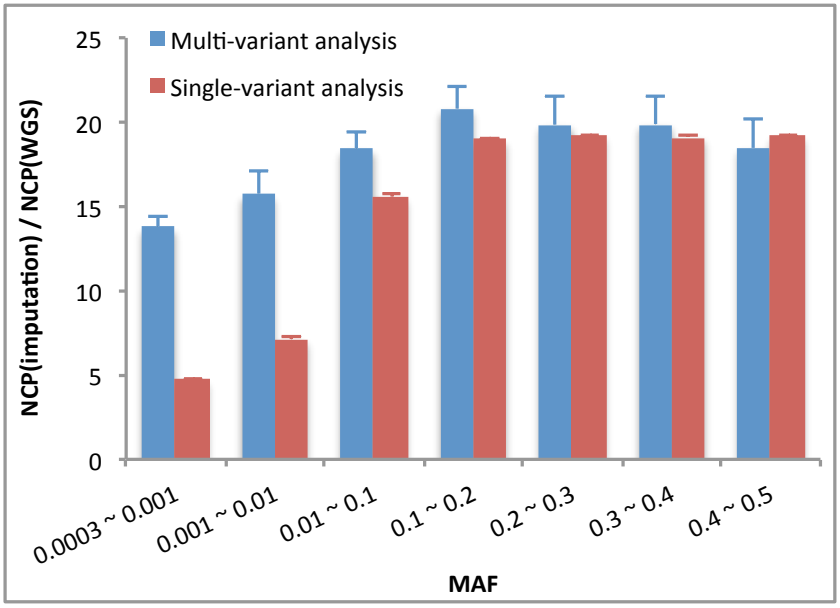
(c) Proportion of variants with MAF < 0.1



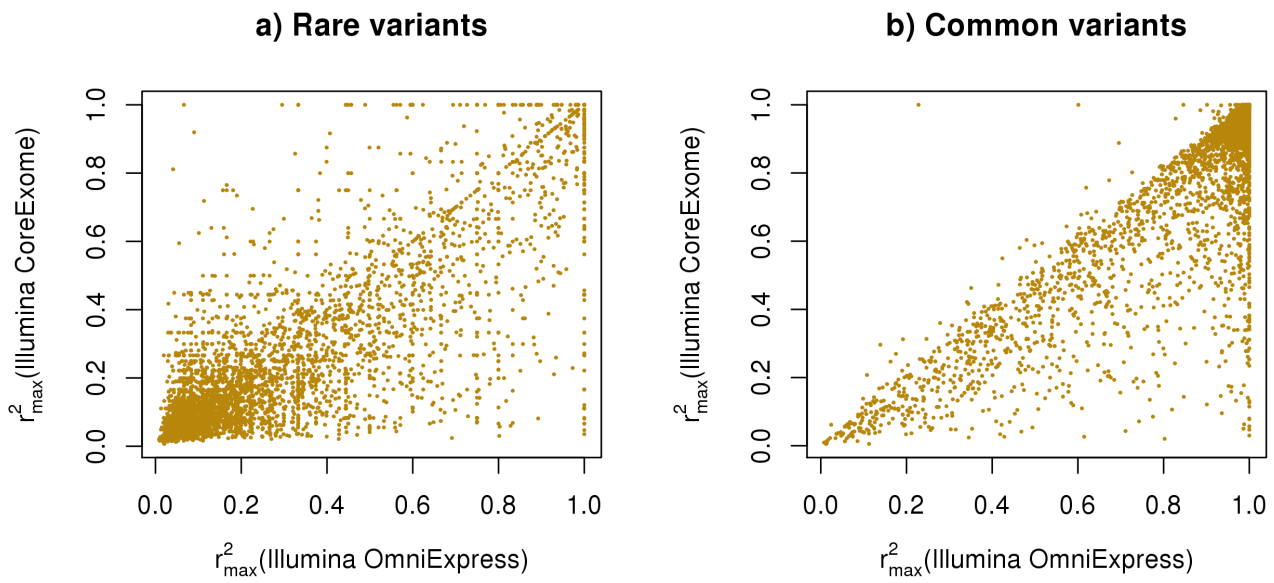
(d) Adjusted estimate



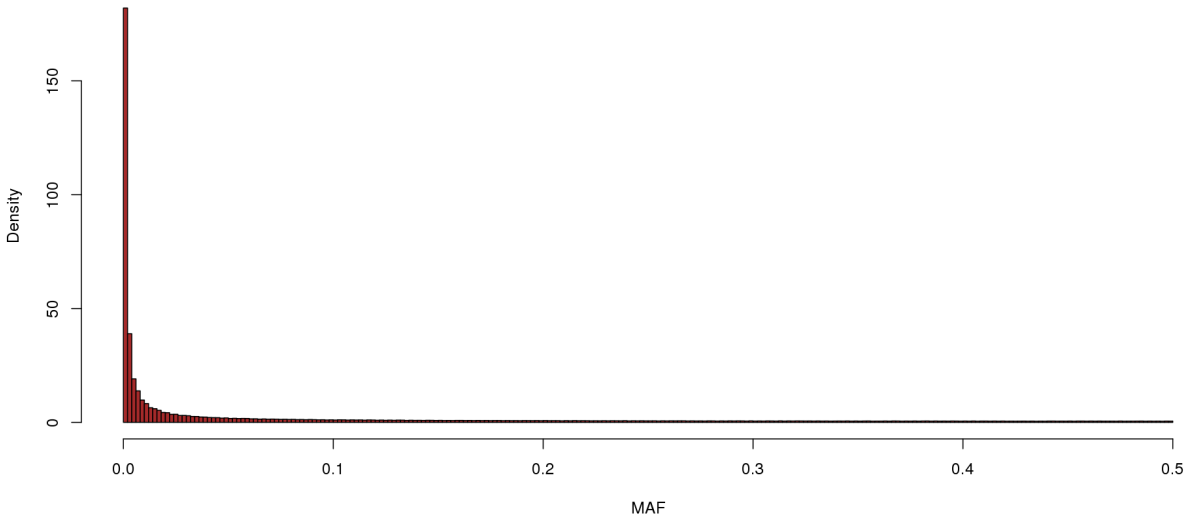
Supplementary Figure 14 Adjusting the estimate of h^2_{1KGP} in each LD group by the proportion of low-MAF variants in the group. The estimates are from the GREML-LDMS analyses (28 genetic components) in the combined data set for height and BMI. It is shown in panel (a) that variants in lower LD regions tend to explain a larger proportion of variance, for height in particular. However, there is also an enrichment of low-MAF variants in regions with lower LD as shown in panels (b) and (c). We therefore adjusted $\hat{h}^2_{1KGP}$ (and its s.e.) by the proportion of variants with MAF ≤ 0.1 (θ) in each MAF group, using a regression model $y = \alpha + x\beta + e$ with $y = \hat{h}^2_{1KGP}$ (or its s.e.) and $x = \theta - \text{mean}(\theta)$. Shown in panel (d) is the adjusted estimate is calculated as $\hat{h}^2_{1KGP} - x\hat{\beta}$ with $\hat{\beta}$ estimated from the regression above.



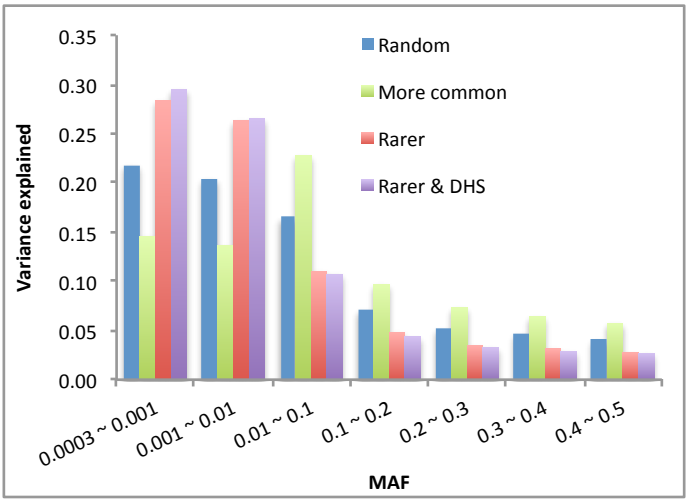
Supplementary Figure 15 Power comparison between WGS and imputation. Power is measured by the noncentrality parameter (NCP) of a χ_1^2 test-statistic. Given a fixed budget, NCP(imputation) / NCP(WGS) = $r^2 n_1 / n_2$ where n_1 is the sample sizes of a study using SNP genotyping followed imputation, n_2 is the sample size of a WGS study, r^2 is the proportion of variation at a sequence variant tagged by imputed variant(s). Assuming genotyping cost per individual using SNP array is 20 times cheaper than that using WGS, NCP(imputation) / NCP(WGS) = $20r^2$. Error bar is the s.e.m.. Multi-variant tagging was quantified as in **Fig. 3**, and single-variant tagging is calibrated as in **Fig. 5**, by imputing variants on Illumina CoreExome array from the UK10K-seq data to 1KGP reference panels.



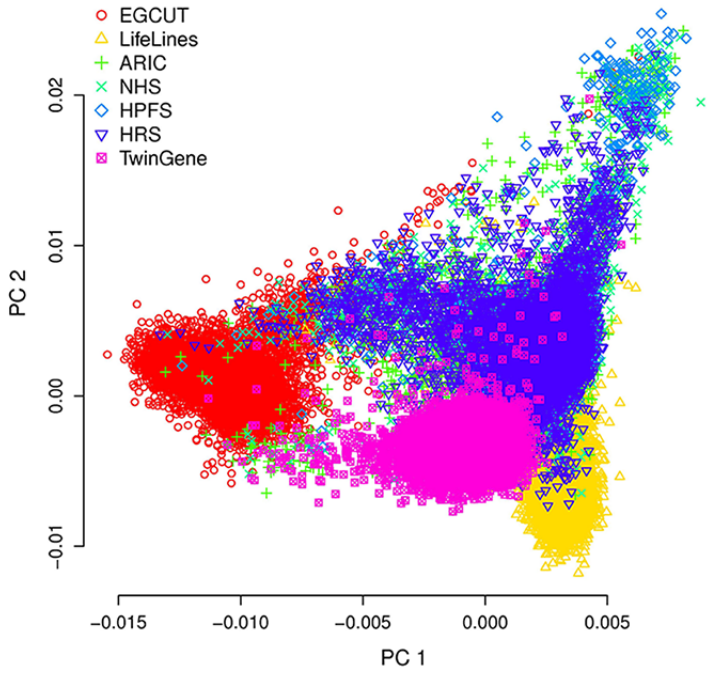
Supplementary Figure 16 Single-variant tagging of sequence variants by 1KGP-imputed variants based Illumina OmniExpress array vs. that based on Illumina CoreExome array. Shown is the squared correlation (r^2_{max}) between a sequence variant from the UK10K-WGS data and the best tagging variant from 1KGP imputation within ±1Mb distance. The 1KGP imputation analyses are based on variants on Illumina OmniExpress and Illumina CoreExome arrays extracted from the UK10K-WGS data (see Online Methods for details about the imputation analyses based on the UK10K-WGS data). There are 10.2% (8.3%) rare variants and 1.1% (0.3%) common variants that are almost not tagged by imputation, i.e. $r^2_{max} < 0.05$, based on Illumina CoreExome (OmniExpress) array.



Supplementary Figure 17 MAF distribution of variants in the UK10K-WGS data. After quality controls (Online Methods), there are 17.6M variants in total (9.3M variants with MAF < 0.01).



Supplementary Figure 18 Proportions of variance explained by the simulated causal variants in 7 MAF groups under different simulation scenarios. I) Random: 1,000 causal variants randomly sampled from all the sequence variants; II) More common: 1,000 random and additional 500 common causal variants; III) Rarer: 1,000 random and additional 500 rare causal variants; IV) Rarer & DHS: 1,000 random and additional 500 rare causal variants all sampled from the variants at DHSs. Shown are averages from 200 simulation replicates. The total heritability is 0.8.



Supplementary Figure 19 Principal component analysis (PCA) of ancestry in the combined GWAS data set. The PCA is based on 1KGP-imputed variants that are on HM3, with MAF ≥ 0.01 and imputation $R^2 \geq 0.3$.

Supplementary Tables

Supplementary Table 1 Estimates of heritability using WGS variants under different simulation scenarios based on the UK10K-WGS data.

		h^2 = 0.2		h^2 = 0.5		h^2 = 0.8	
		Est.	**s.e.m.**	**Est.**	**s.e.m.**	**Est.**	**s.e.m.**
Random	GREML-SC	0.19	0.010	0.49	0.013	0.79	0.013
	GREML-MS	0.20	0.014	0.50	0.015	0.79	0.016
	LDAK	0.26	0.024	0.61	0.027	0.98	0.026
	LDAK-MS	0.32	0.028	0.72	0.031	1.15	0.029
	LDres	0.20	0.011	0.52	0.013	0.77	0.010
	LDres-MS	0.24	0.015	0.54	0.017	0.85	0.017
	GREML-LDMS	0.20	0.016	0.52	0.017	0.79	0.017
More common	GREML-SC	0.21	0.011	0.55	0.012	0.89	0.013
	GREML-MS	0.21	0.013	0.50	0.014	0.79	0.014
	LDAK	0.22	0.024	0.54	0.025	0.85	0.028
	LDAK-MS	0.30	0.029	0.70	0.029	1.10	0.031
	LDres	0.22	0.012	0.57	0.013	0.82	0.010
	LDres-MS	0.23	0.015	0.54	0.015	0.86	0.016
	GREML-LDMS	0.19	0.017	0.51	0.017	0.79	0.016
Rarer	GREML-SC	0.18	0.011	0.46	0.011	0.71	0.013
	GREML-MS	0.21	0.013	0.50	0.014	0.79	0.016
	LDAK	0.32	0.028	0.72	0.026	1.11	0.027
	LDAK-MS	0.34	0.033	0.79	0.030	1.21	0.030
	LDres	0.20	0.011	0.49	0.012	0.72	0.011
	LDres-MS	0.23	0.016	0.55	0.016	0.86	0.017
	GREML-LDMS	0.19	0.016	0.49	0.017	0.80	0.017
Rarer & DHSs	GREML-SC	0.18	0.011	0.40	0.011	0.63	0.013
	GREML-MS	0.18	0.014	0.43	0.014	0.69	0.016
	LDAK	0.33	0.026	0.75	0.026	1.13	0.028
	LDAK-MS	0.36	0.031	0.81	0.029	1.24	0.032
	LDres	0.19	0.012	0.44	0.012	0.69	0.012
	LDres-MS	0.20	0.016	0.48	0.017	0.78	0.017
	GREML-LDMS	0.20	0.017	0.49	0.017	0.79	0.017

Random: 1,000 causal variants sampled from the WGS variants at random; **More common:** 1,000 causal variants sampled at random + 500 causal variants with MAF > 0.01; **Rarer:** 1,000 causal variants sampled at random + 500 causal variants with MAF < 0.01; **Rarer & DHSs:** 1,000 causal variants sampled at random from the variants at DHSs + 500 causal variants with MAF < 0.01 sampled from the variants at DHSs; **Est.**: mean estimate averaged from 200 simulations; **s.e.m.**: standard error of the mean.

Supplementary Table 2 Number of SNPs used in the imputation analysis based on the UK10K-WGS data

SNP array	# Variants (common)	# Variants (rare)	# Variants (total)
Illumina CoreExome	261,136	51,128	312,264
Affymetrix 6.0	572,172	0	572,172
Affymetrix Axiom	562,477	10,764	573,241
Illumina OmniExpress	607,344	0	607,344
Illumina Omni2.5	1,367,530	173,187	1,540,717

These are the number of SNPs that are in common between the SNPs arrays and the UK10K data after QC (i.e. UK10K-WGS). The list of SNPs for Affymetrix 6.0, Illumina OmniExpress, and Illumina Omni2.5 were from the ARIC, TwinGene, and HRS data, respectively, after QC. The SNPs for Illumina CoreExome and Affymetrix Axiom were from the strand alignment data files produced and hosted by Will Rayner (http://www.well.ox.ac.uk/~wrayner/strand/).

Nature Genetics: doi:10.1038/ng.3390

Supplementary Table 3 Estimates of variance explained by 1KGP-imputed variants from GREML-LDMS analysis for height and BMI in the combined GWAS data set.

		1st LD quartile		2nd LD quartile		3rd LD quartile		4th LD quartile		Row sum	s.e.
		Est	s.e.	Est	s.e.	Est	s.e.	Est	s.e.		
Height	**2.5×10^{-5} < MAF ≤ 0.001**	-0.002	0.014	-0.003	0.011	0.004	0.010	-0.003	0.006	**-0.005**	**0.016**
	0.001 < MAF ≤ 0.01	0.048	0.013	-0.001	0.011	0.019	0.010	0.018	0.007	**0.084**	**0.019**
	0.01 < MAF ≤ 0.1	0.031	0.008	0.043	0.007	0.036	0.006	0.021	0.004	**0.130**	**0.011**
	0.1 < MAF ≤ 0.2	0.029	0.005	0.019	0.004	0.018	0.004	0.014	0.003	**0.081**	**0.008**
	0.2 < MAF ≤ 0.3	0.024	0.005	0.028	0.004	0.028	0.004	0.012	0.003	**0.092**	**0.008**
	0.3 < MAF ≤ 0.4	0.024	0.005	0.029	0.004	0.010	0.004	0.022	0.003	**0.085**	**0.008**
	0.4 < MAF ≤ 0.5	0.034	0.004	0.019	0.004	0.022	0.004	0.013	0.003	**0.088**	**0.007**
	Column sum	**0.188**	**0.019**	**0.134**	**0.016**	**0.137**	**0.015**	**0.097**	**0.009**		
	Total sum	**0.555**	**0.023**								
	Log likelihood	**-20189.2**									
	Sample size	**43,599**									
BMI	**2.5×10^{-5} < MAF ≤ 0.001**	0.003	0.015	0.015	0.013	0.009	0.011	0.010	0.008	**0.038**	**0.018**
	0.001 < MAF ≤ 0.01	0.001	0.014	-0.009	0.012	-0.012	0.010	0.006	0.007	**-0.014**	**0.019**
	0.01 < MAF ≤ 0.1	0.027	0.008	0.008	0.006	0.007	0.005	0.009	0.004	**0.051**	**0.011**
	0.1 < MAF ≤ 0.2	0.009	0.005	0.020	0.004	0.020	0.004	0.010	0.003	**0.059**	**0.008**
	0.2 < MAF ≤ 0.3	0.012	0.005	0.014	0.004	0.013	0.004	0.003	0.003	**0.042**	**0.007**
	0.3 < MAF ≤ 0.4	0.013	0.004	0.010	0.004	0.010	0.003	0.011	0.003	**0.044**	**0.007**
	0.4 < MAF ≤ 0.5	0.023	0.004	0.015	0.004	0.014	0.003	0.002	0.002	**0.054**	**0.006**
	Column sum	**0.088**	**0.020**	**0.073**	**0.018**	**0.061**	**0.015**	**0.052**	**0.010**		
	Total sum	**0.274**	**0.025**								
	Log likelihood	**-21234.5**									
	Sample size	**43,366**									

Note: s.e. for "column sum" or "row sum" refers to standard error of the sum estimate rather than sum of the standard errors.

Supplementary Table 4 Descriptive summary of the cohorts and quality control criteria of the genotype data

Cohort	Sample size	Genotyping array	Inclusion criteria				SNPs that met QC
			MAF	SNP call rate	Sample call rate	*p* for HWE	
ARIC + NHS + HPFS	14,347	Affymetrix 6.0	≥ 1%	≥ 98%	≥ 98%	$> 10^{-3}$	565,040
TwinGene	10,729	Illumina OmniExpress	≥ 1%	≥ 97%	≥ 97%	$> 10^{-7}$	627,305
HRS	8,652	Illumina Omni2.5	≥ 0.1%	≥ 98%	≥ 98%	$> 10^{-6}$	2,152,114
EGCUT	7,967	Illumina OmniExpress	≥ 1%	≥ 95%	≥ 95%	$> 10^{-6}$	633,182
Lifelines	13,386	Illumina Cyto SNP12 v2	≥ 1%	≥ 95%	≥ 95%	$> 10^{-6}$	245,943

Supplementary Note

1. The circumstances under which the estimate of variance explained by all the variants is a (un)biased estimate of heritability

Under an additive genetic model, the quantitative phenotype of an individual *i* can be expressed as

$$y_i = g_i + e_i \quad [1]$$

where g_i is the genetic value, $g_i \sim N(0,\sigma_g^2)$ with σ_g^2 being the genetic variance, and e_i is the residual with $e_i \sim N(0,\sigma_e^2)$. The narrow-sense heritability is defined as $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$. If causal variants are known, g_i can be expressed as

$$g_i = \mathbf{z}_i'\mathbf{q} \quad [2]$$

where $z_i = \{z_{ik}\}$ is an $m_q \times 1$ vector of the standardized genotype variables of m_q causal variants, $z_{ik} = (x_{ik} - 2p_k) / \sqrt{2p_k(1-p_k)}$, x_{ik} is the genotype variable of the *k*-th causal variant (coded as 0, 1 or 2), p_k is the frequency of the coded allele, and $\mathbf{q}$ is an $m_q \times 1$ vector of the effect sizes of the causal variants with $\mathbf{q} \sim N(0,\mathbf{I}\sigma_g^2 / m_q)$. Since z_{ik} is standardized, $E(z_{ik}) = 0$ and $\text{var}(z_{ik}) = 1$. The phenotypic covariance between two individuals based on this model is

$$\text{cov}(y_i, y_j) = \text{cov}(g_i, g_j) = \text{cov}(\mathbf{z}_i'\mathbf{q}, \mathbf{z}_j'\mathbf{q}) = \mathbf{z}_i' \,\text{var}(\mathbf{q})\mathbf{z}_j = \sigma_g^2 \mathbf{z}_i'\mathbf{z}_j / m_q \quad [3]$$

If we define $G_{ij} = \mathbf{z}_i'\mathbf{z}_j / m_q$ as the genetic relationship between two individuals at the causal variants[3], then $\text{cov}(y_i, y_j) = \sigma_g^2 G_{ij}$.

In practice, however, causal variants are largely unknown. We therefore estimate the variance explained using all the variants from whole genome sequencing based on the model below

$$y_i = g_{v(i)} + e_i \text{ with } g_{v(i)} = \mathbf{w}_i'\mathbf{u} \quad [4]$$

where $g_{v(i)}$ is the genetic value captured by all the variants with $g_{v(i)} \sim N(0,\sigma_v^2)$, $\mathbf{w}_i = \{w_{ik}\}$ are the standardized genotype variables of *m* variants, and **u** is an $m \times 1$ vector of the effect sizes of the variants with $\mathbf{u} \sim N(0,\mathbf{I}\sigma_v^2 / m)$. The phenotypic covariance between two individuals based on this model is

$$\text{cov}(y_i, y_j) = \text{cov}(g_{v(i)}, g_{v(j)}) = \text{cov}(\mathbf{w}_i'\mathbf{u}, \mathbf{w}_j'\mathbf{u}) = \mathbf{w}' \text{var}(u)\mathbf{w}_j = \sigma_v^2 \mathbf{w}_i'\mathbf{w}_j / m \quad [5]$$

If we define $A_{ij} = \mathbf{w}_i'\mathbf{w}_j / m$ as the genetic relationship between two individuals at all the variants, then $\text{cov}(y_i, y_j) = \sigma_v^2 A_{ij}$. The proportion of variance explained by all the variants is defined as $h_v^2 = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2)$.

We have known previously[3] that $h_v^2 = h^2$ if A_{ij} is an unbiased predictor of G_{ij} in a sense that $E(G_{ij} \mid A_{ij}) = A_{ij}$, i.e. $h_v^2 = \frac{\text{cov}(G_{ij}, A_{ij})}{\text{var}(A_{ij})} h^2$. We show below the circumstance under which the estimate of h_v^2 is an unbiased or biased estimate of h^2. Assuming causal variants are a subset of all the variants, we have

$$\begin{aligned} &\text{cov}(G_{ij}, A_{ij}) = \text{cov}(\mathbf{z}_i'\mathbf{z}_j / m_q, \mathbf{w}_i'\mathbf{w}_j / m) = E(\mathbf{z}_i'\mathbf{z}_j \mathbf{w}_i'\mathbf{w}_j) / (m_q m) \\ &= E(\sum_k^{m_q} z_{ik} z_{jk} \sum_l^m w_{il} w_{jl}) / (m_q m) = E(\sum_k^{m_q} \sum_l^m z_{ik} z_{jk} w_{il} w_{jl}) / (m_q m) \\ &= \sum_k^{m_q} \sum_l^m E(z_{ik} w_{il}) E(z_{jk} w_{jl}) / (m_q m) = \sum_k^{m_q} \sum_l^m r_{kl}^2 / (m_q m) = \overline{r}_{MQ}^2 \end{aligned} \quad [6]$$

with $\overline{r}_{QM}^2$ being the mean linkage disequilibrium (LD) r^2 between the causal variants and all the variants (including the causal variants), and

$$\begin{aligned} \text{var}(A_{ij}) &= \text{var}(\mathbf{w}_i'\mathbf{w}_j / m) = E(\mathbf{w}_i'\mathbf{w}_j \mathbf{w}_i'\mathbf{w}_j) / m^2 \\ &= \sum_k^m \sum_l^m E^2(w_{ik} w_{il}) / m^2 = \sum_k^m \sum_l^m r_{kl}^2 / m^2 = \overline{r}_{MM}^2 \end{aligned} \quad [7]$$

with $\overline{r}_{QM}^2$ being is the mean LD r^2 between all the variants (including the causal variants), suggesting that $h_v^2 = \frac{\overline{r}_{QM}^2}{\overline{r}_{MM}^2} h^2$. We therefore can conclude that

1) If causal variants are a random subset of all the variants used in the analysis, $\overline{r}_{MQ}^2 = \overline{r}_{MM}^2$ so that $h_v^2 = h^2$.
2) Since LD r^2 is a function of allele frequencies between two variants[4], if there are disproportionally more common or rare causal variants, then $\overline{r}_{MQ}^2$ will be different from $\overline{r}_{MM}^2$, so that $h_v^2 \neq h^2$.
3) If causal variants are enriched in genomic regions with lower or higher LD than average, then $\overline{r}_{MQ}^2 \neq \overline{r}_{MM}^2$ so that $h_v^2 \neq h^2$.

It is notable that if there is a difference between h_v^2 and h^2 due to the difference between $\overline{r}_{MQ}^2$ and $\overline{r}_{MM}^2$, such difference (bias) is caused by the difference in models rather than the difference in methods for parameter estimation.

2. The standard error of the estimate of cumulative contribution of variants with MAF ≤ θ to the genetic variance

Let **L** be the variance and covariance matrix of the estimates of genetic components from GREML-LDMS analysis (**L** matrix is available in GCTA-GREML output). The diagonal elements of **L** are $\text{var}(\hat{\sigma}_{v(i)}^2)$ (see Online Methods for definition of $\sigma_{v(i)}^2$), and the off-diagonal elements of LD are $\text{cov}(\hat{\sigma}_{v(i)}^2, \hat{\sigma}_{v(j)}^2)$. The cumulative contribution of variants with MAF ≤ θ to the genetic variance is calculated as $\hat{\sigma}_v^2(\text{MAF} \leq \theta) / \hat{\sigma}_v^2(\text{MAF} \leq 0.5)$, where is the sum of estimates of $\hat{\sigma}_{v(i)}^2$ for variants with

MAF < 0.1 and $\hat{\sigma}_v^2(\text{MAF} \le 0.5)$ is the sum of all the 28 components. Let $x = \hat{\sigma}_v^2(\text{MAF} \le \theta)$ and $y = \hat{\sigma}_v^2(\text{MAF} \le 0.5)$, the sampling variance of $\hat{\sigma}_v^2(\text{MAF} \le \theta) / \hat{\sigma}_v^2(\text{MAF} \le 0.5)$ can be calculated from the Delta method[5] as $\text{var}[\frac{x}{y}] \approx \frac{E(x)^2}{E(y)^2}\left\{\frac{\text{var}(x)}{E(x)^2} + \frac{\text{var}(y)}{E(y)^2} - \frac{2\,\text{cov}(x,y)}{E(x)E(y)}\right\}$. In practice, $E(x)$ and $E(y)$ are unknown and is therefore replaced by their estimates (x and y). The var(x), var(y), and cov(x, y) are calculated from the sum of the relevant sub-matrix of **L**.

3. Variance explained by population stratification is very small for both height and BMI in our data

In all the analyses described above, we fitted in the model as fixed covariates the first 10 principal components (PCs) calculated from common (MAF > 0.01) imputed variants on HapMap3. We also performed analyses fitting the first 20 PCs or the PCs calculated from all the imputed variants. The results were all very similar (**Supplementary Fig. 9**). This is because the first 20 PCs computed from either all the imputed variants or the common variants on HapMap3 only explained up to 0.8% of variance for height and 0.08% for BMI.

4. GREML-LDMS is robust to the model assumption about the relationship between MAF and effect size

We found strong evidence that variants with MAF < 0.1 explained a larger proportion of variance for height than what we would expect under a neutral model (**Fig. 4a**). We then asked whether or not the estimate could be biased due to our model assumption. If we define $w = (x - 2p)/\sqrt{2p(1-p)}$ with x being the genotype variable of a variant (coded as 0, 1 or 2) and p being the allele frequency, i.e. w is the standardized form x, and define b and u as the regression coefficients of phenotype on x and w, respectively, then the genetic variance attributable to this variant can be written as var(xb) = $2p(1 - p)b^2$

or var(wu) = u^2, meaning that $u^2 = 2p(1-p)b^2$. By default, the GREML approach[3] calculates the genetic relationship using the equation $A_{ij} = \frac{1}{m_t}\sum_k^{m_t}\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)}$ (see Equation 2 in Online Methods for notations). This model assumes a normal distribution of u with constant variance regardless of MAF, which implicitly assumes that the per-allele effect size (b) of a rare variant is on average larger than that of a common variant (Assumption A). We therefore re-analysed the data assuming a normal distribution of b (Assumption B), which means that variance explained for a rare variant is on average smaller than that for a common variant. Assumption B is equivalent to calculating the genetic relationship using the following equation[6,7], $A_{ij} = \sum_k^{m_t}[(x_{ik}-2p_k)(x_{jk}-2p_k)] / \sum_k^{m_t}[2p_k(1-p_k)]$. We found that the GREML-LDMS estimates based on these two assumptions were remarkably consistent (**Supplementary Fig. 12**), demonstrating the robustness of the GREML-LDMS method to different assumptions about the distribution of effect sizes. In addition, we observed from the genome-wide association analysis a correlation between effect size and MAF, and showed by simulations (mimicking the observed genetic architecture for height) that the GREML-LDMS was also robust to such a correlation (**Supplementary Fig. 13**).

5. An approximate method to estimate h^2_{WGS} from h^2_{1KGP}

We quantified by simulations based on 5 different types of SNP arrays under four different scenarios that on average ~97% and ~68% of variation at common and rare variants, respectively, can be captured by KGP imputation (**Supplementary Fig. 4**). This implies that $h^2 \approx h^2_{\mathrm{1KGP(rare)}}/0.68 + h^2_{\mathrm{1KGP(common)}}/0.97$, where $h^2_{\mathrm{1KGP(rare)}}$ and $h^2_{\mathrm{1KGP(common)}}$ denote h^2_{1KGP} for rare and common variants, respectively. Note that this is still likely to be an underestimate of heritability because complicated structure variation and extremely rare variation are likely to be less well captured

by 1KGP imputation than those used in our simulation. Using this approximate method, we quantified the lower limit of heritability of 0.61 (s.e. = 0.045) for height and 0.29 (s.e. = 0.047) for BMI.

6. Other possible sources of missing heritability

It has been suggested that epistasis and GxE can explain a substantial amount of variance for allele-specific expression traits[8]. However, the extent to which epistasis and GxE contribute to complex trait variation in general remains unclear. The missing heritability problem refers to the gap between additive genetic variance estimated from family studies and that from population-based studies (e.g. GWAS), so that non-additive genetic variation is irrelevant unless h^2 is overestimated in family-based studies due to the confounding with non-additive genetic variation. However, such confounding is negligible if the total amount of variance attributed to non-additive genetic variation is small. In fact, a recent study[9] shows that on average across a number of quantitative traits dominance variation explains < 4% of phenotypic variance. Quantitative genetics theory also predicts that epistasis and other higher order interaction terms are unlikely to explain a substantial amount of variance for complex traits[10]. Therefore, non-additive genetic variation is unlikely to contribute substantially to the missing heritability. GxE can be relevant only if there is a strong component of GxE variance and h^2 is estimated from family-based samples in a single homogenous environment where GWAS is performed in samples from multiple heterogeneous environments.

7. Enrichment of variance explained in lower-LD regions is driven by MAF

Our results also seem to suggest that variants in genomic regions with lower LD tend to explain a larger proportion of variance than those in regions with higher LD, 18.8% vs. 9.7% ($P_{\text{difference}} = 3.0\times10^{-5}$) for height and 8.8% vs. 5.2% ($P_{\text{difference}} = 0.12$) for BMI (**Supplementary Table 3**). This, however, is

likely due to the enrichment of lower-MAF variants in regions with lower LD (**Supplementary Fig. 14**).

8. The Morrison et al. WGS study

Morrison et al.[11] applied the GREML method in a WGS data set of 962 individuals and concluded that the majority of the heritability of high-density lipoprotein cholesterol level can be attributable to common variation. The s.e. of $\hat{h}^2_{\mathrm{WGS}}$ for common variants from Morrison et al. is 0.14, much smaller than that expected from theory[12] given the sample size. The reported s.e. is consistent with the variance of the off-diagonal elements of the genetic relationship matrix (GRM) being 1.1×10^{-4} for common variants, ~7 times larger than what we observed in the UK10K data (1.6×10^{-5}). This suggests that the variance of GRM in Morrison et al. was highly inflated, possibly due to the adjustment that they made to the GRM (the adjustment is under strong assumptions[3] and actually not necessary for WGS data) and/or due to cryptic relatedness. If all the 962 individuals are unrelated and there is no adjustment to the GRM, the s.e. of $\hat{h}^2_{\mathrm{WGS}}$ for common variants should be ~0.37 rather than 0.14. Such inflation in the variance of estimated genetic relatedness could potentially bias their results from simulations and their conclusion about variance explained by common variation from the analysis of real data.

Nature Genetics: doi:10.1038/ng.3390

9. Acknowledgements

ARIC: The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National HumanGenome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

EGCUT: EGCUT study was supported through the Estonian Genome Center of University of Tartu by the Targeted Financing from the Estonian Ministry of Science and Education [SF0180142s08]; the Development Fund of the University of Tartu (grant SP1GVARENG); the European Regional Development Fund to the Centre of Excellence in Genomics (EXCEGEN; grant 3.2.0304.11-0312); and through FP7 grant 313010.

HRS: HRS is supported by the National Institute on Aging (NIA U01AG009740). The genotyping was funded separately by the National Institute on Aging (RC2 AG036495, RC4 AG039029). Genotyping was conducted by the NIH Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotyping quality control and final preparation of the data were performed by the Genetics Coordinating Center at the University of Washington.

Lifelines: The LifeLines Cohort Study, and generation and management of GWAS genotype data for the LifeLines Cohort Study is supported by the Netherlands Organization of Scientific Research NWO (grant 175.010.2007.006), the Economic Structure Enhancing Fund (FES) of the Dutch government, the Ministry of Economic Affairs, the Ministry of Education, Culture and Science, the Ministry for

Health, Welfare and Sports, the Northern Netherlands Collaboration of Provinces (SNN), the Province of Groningen, University Medical Center Groningen, the University of Groningen, Dutch Kidney Foundation and Dutch Diabetes Research Foundation. The authors are grateful to the study participants, the staff from the LifeLines Cohort Study and the contributing research centers delivering data to LifeLines and the participating general practitioners and pharmacists.

NHS & HPFS: Funding support for the GWAS of Gene and Environment Initiatives in Type 2 Diabetes was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004399). The human subjects participating in the GWAS derive from The Nurses' Health Study (NHS) and Health Professionals' Follow-up Study (HPFS) and these studies are supported by National Institutes of Health grants CA87969, CA55075, and DK58845. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies, GENEVA Coordinating Center (U01 HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH GEI (U01HG004424).

TwinGene: This work was supported by grants from the Ministry for Higher Education, the Swedish Research Council (M-2005-1112 and 2009-2298), GenomEUtwin (EU/QLRT-2001-01254; QLG2-CT-2002-01254), NIH grant DK U01-066134, The Swedish Foundation for Strategic Research (SSF; ICA08-0047).

UK10K: The UK10K project was funded by the Wellcome Trust award WT091310. Twins UK (TUK): TUK was funded by the Wellcome Trust and ENGAGE project grant agreement HEALTH-F4-2007–201413. The study also receives support from the Department of Health via the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St. Thomas' NHS Foundation Trust in partnership with King's College London. Dr Spector is an NIHR senior Investigator and ERC Senior Researcher. Funding for the

project was also provided by the British Heart Foundation grant PG/12/38/29615 (Dr Jamshidi). A full list of the investigators who contributed to the UK10K sequencing is available from www.UK10K.org.

10. LifeLines Cohort Study group author

Behrooz Z Alizadeh (1), Paul IW de Bakker (2,3), H Marike Boezen (1), Lude Franke (4), Pim van der Harst (5), Gerjan Navis (6), Marianne Rots (7), Harold Snieder (1), Ronald P Stolk (1,8), Morris Swertz (4), Bruce HR Wolffenbuttel (9), Cisca Wijmenga (4)

(1) Department of Epidemiology, University of Groningen, University Medical Center Groningen, The Netherlands

(2) Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, The Netherlands

(3) Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands

(4) Department of Genetics, University of Groningen, University Medical Center Groningen, The Netherlands

(5) Department of Cardiology, University of Groningen, University Medical Center Groningen, The Netherlands

(6) Department of Internal Medicine, Division of Nephrology, University of Groningen, University Medical Center Groningen, The Netherlands

(7) Department of Medical Biology, University of Groningen, University Medical Center Groningen, The Netherlands

(8) LifeLines Cohort Study, University of Groningen, University Medical Center Groningen, The Netherlands

(9) Department of Endocrinology, University of Groningen, University Medical Center Groningen, The Netherlands

11. References

1. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-90 (2015).
2. Galinsky, K.J. *et al.* Fast principal components analysis reveals independent evolution of ADH1B gene in Europe and East Asia. *bioRxiv*, 018143 (2015).
3. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565-9 (2010).
4. Wray, N.R. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. *Twin Res Hum Genet* **8**, 87-94 (2005).
5. Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*, (Sunderland, MA: Sinauer Associates, 1998).
6. Lee, S.H. *et al.* Estimation of SNP heritability from dense genotype data. *Am. J. Hum. Genet.* **93**, 1151-5 (2013).
7. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011-21 (2012).
8. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet* **47**, 88-91 (2015).
9. Zhu, Z. *et al.* Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet* **96**, 377-85 (2015).
10. Maki-Tanila, A. & Hill, W.G. Influence of gene interaction on complex trait variation with multilocus models. *Genetics* **198**, 355-67 (2014).
11. Morrison, A.C. *et al.* Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat. Genet.* **45**, 899-901 (2013).
12. Visscher, P.M. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet* **10**, e1004269 (2014).