bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Re-evaluation of SNP heritability in complex human traits

Doug Speed,[1,*] Na Cai,[2,3] the UCLEB Consortium,[4] Michael R. Johnson,[5] Sergey Nejentsev,[6] David J Balding.[1,7]
*Corresponding author: doug.speed@ucl.ac.uk

[1] *UCL Genetics Institute, University College London, United Kingdom.*
[2] *Wellcome Trust Sanger Institute, Hinxton, United Kingdom.*
[3] *European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom.*
[4] *A full list of members and affiliations appears in the Supplementary Material.*
[5] *Division of Brain Science, Imperial College London, United Kingdom.*
[6] *Department of Medicine, University of Cambridge, United Kingdom.*
[7] *Centre for Systems Genomics, School of BioSciences and School of Mathematics & Statistics, University of Melbourne, Australia.*

**Abstract**

SNP heritability, the proportion of phenotypic variance explained by SNPs, has been estimated for many hundreds of traits, and these estimates are being used to explore genetic architecture and guide future research. To estimate SNP heritability requires strong assumptions about how heritability is distributed across the genome, but the assumptions in current use have not been thoroughly tested. By analyzing imputed data for 42 human traits, we empirically derive an improved model for heritability estimation. It is commonly assumed that the expected heritability of a SNP does not depend on its allele frequency; we instead identify a more realistic relationship which reflects that heritability tends to decrease with minor allele frequency. Two methods for estimating SNP heritability, GCTA and LDAK, make contrasting assumptions about how heritability varies with linkage disequilibrium; we demonstrate that the model used by LDAK better reflects the properties of real data. Additionally, we show how genotype certainty can be incorporated in the heritability model; this enables the inclusion of poorly-imputed SNPs, which can capture substantial extra heritability. Our revised method typically results in substantially higher estimates of SNP heritability: for example, across 19 traits (mainly diseases), the estimates based on common SNPs (minor allele frequency $> 0.01$) are on average 40% (SD 3) higher than those obtained using original GCTA, and 25% (SD 2) higher than those from the recently-proposed extension GCTA-LDMS. We conclude that for a wide range of traits, common SNPs tag a greater fraction of causal variation than is currently appreciated. When we also include rare SNPs (minor allele frequency $< 0.01$), we find that across 23 quantitative traits, estimates of SNP heritability increase by on average 29% (SD 12), and that rare SNPs tend to contribute about half the heritability of common SNPs.

**Introduction**

The SNP heritability ($h^2_{\text{SNP}}$) of a trait is the total proportion of phenotypic variance explained by additive contributions from SNPs.[1] Accurate estimates of $h^2_{\text{SNP}}$ are central to resolving the missing heritability debate, indicate the potential utility of SNP-based prediction and help design future genome-wide association studies (GWAS).[2,3] The first method for estimating $h^2_{\text{SNP}}$ from GWAS data was proposed in 2010,[1] and has subsequently been applied to many hundreds of traits. Extensions of this method are now being used to partition heritability across chromosomes, biological pathways and by SNP function, and to calculate the genetic correlation between pairs of traits.[4–6]

As a GWAS usually contains many more SNPs than individuals, estimation of $h^2_{\text{SNP}}$ requires steps to avoid over-fitting. Most reported estimates of $h^2_{\text{SNP}}$ are based on assigning the same Gaussian prior distribution to each SNP effect size, in a way which implies that all SNPs are expected to contribute equal heritability. By examining a large collection of real datasets, we derive approximate relationships between the expected heritability of a SNP and minor allele frequency (MAF), levels of linkage disequilibrium (LD) with other SNPs and genotype certainty. This provides us with an improved model for heritability estimation, and a better understanding of

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the genetic architecture of complex traits.

## Methods

Suppose the length-$n$ vector $\boldsymbol{Y}$ contains phenotypic values, the $n \times p$ matrix $\boldsymbol{Z}$ contains covariates, while the $n \times m$ matrix $\boldsymbol{S}$ contains allele counts: $S_{i,j} \in [0, 2]$ indicates the number of copies of the minor allele possessed by Individual $i$ at SNP $j$. The genotype matrix $\boldsymbol{X}$ is constructed from $\boldsymbol{S}$ by centering and scaling the allele counts for each SNP:[1,7]

$$X_{i,j} = (S_{i,j} - 2f_j) \times (2f_j(1-f_j))^{\alpha/2},$$

where $f_j$ is the observed MAF of SNP $j$. It is common to use $\alpha = -1$ so that SNP genotypes have the same variance, but we consider alternative values. Let $w_1, \ldots, w_m$ denote SNP weights, designed to adjust for variations in levels of tagging across the genome:[7] SNPs in regions of high linkage disequilibrium (LD) tend to have low $w_j$, and vice versa. We also calculate for each SNP an information score $r_j$, defined in Online Methods, which measures genotype certainty: high-quality SNPs have $r_j$ close to one, while low-quality SNPs have $r_j$ close to zero. Given $w_j$ and $r_j$, the "LDAK Model" for estimating $h^2_{\text{SNP}}$ is:

$$Y_i = \sum_{k=1}^{p} \theta_k Z_{i,k} + \sum_{j=1}^{m} \beta_j X_{i,j} + e_i, \quad \text{with} \quad \beta_j \sim \mathbb{N}(0, r_j w_j \sigma_g^2 / W), \quad e_i \sim \mathbb{N}(0, \sigma_e^2), \tag{1}$$

where the fixed effect $\theta_k$ denotes the coefficient for the $k$th covariate, the random coefficients $\beta_j$ and $e_i$ represent the effect size of SNP $j$ and the noise component for Individual $i$, respectively, $\sigma_g^2$ and $\sigma_e^2$ are interpreted as genetic and environmental variances, while $W$ is a normalizing constant. When the individuals are "unrelated" (only distantly related), $h^2_{\text{SNP}} = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$ is the heritability tagged by the $m$ SNPs.[1,7]

Under the LDAK Model, the expected heritability contribution of SNP $j$ is proportional to $(f_j(1-f_j))^{1+\alpha} \times w_j \times r_j$, and so the expected contribution to $h^2_{\text{SNP}}$ of a set of SNPs depends on their MAFs, weights and information scores. The "GCTA Model" is obtained by setting $w_j = 1$ and $r_j = 1$. To date, most reported estimates of $h^2_{\text{SNP}}$ have used the GCTA Model with $\alpha = -1$; this corresponds to the assumption that all SNP are expected to contribute equal heritability, in which case the expected contribution of a SNP set depends only on the number of SNPs it contains.[1,8] To appreciate the major difference between the GCTA and LDAK Models, consider a region containing two SNPs: under the GCTA Model, the expected heritability of these two SNPs is the same irrespective of the LD between them, whereas under the LDAK Model, two SNPs in perfect LD are expected to contribute only half the heritability of two SNPs showing no LD. See Figure 1A for a more detailed example.

**SNP partitioning:** We can generalize Model (1) by dividing SNPs into tranches across which the genetic variance is allowed to vary: i.e., we assign SNPs in Tranche $l$ the effect size prior distribution $\mathbb{N}(0, r_j w_j \sigma_l^2 / W_l)$, where $W_l$ is a normalizing constant. This is known as SNP partitioning.[1] Two examples are GCTA-MS[9] and GCTA-LDMS.[10] When applied to common SNPs (MAF $> 0.01$), GCTA-MS divides the genome into five tranches based on MAF, using the boundaries 0.1, 0.2, 0.3 and 0.4, while GCTA-LDMS first divides SNPs into four tranches based on local average LD Score[8] (high LD Score indicates a region of high LD), then divides each of these into five based on MAF, resulting in a total of 20 tranches. Our main use of SNP partitioning is to test different model assumptions on real data. The basic idea is to use the partitioned model to obtain estimates of the per-SNP heritability for different SNP tranches, then compare these to the per-SNP heritabilities predicted by the non-partitioned model (Figures 1B & 2A and Supplementary Figures 1 & 2). Formally, we compute a likelihood ratio test (LRT) statistic, equal to twice the difference in log likelihood between the partitioned and non-partitioned models; if the assumptions of the non-partitioned model are accurate, the statistic will follow a $\chi^2(L-1)$ distribution, where $L$ is the number of tranches. We also use SNP partitioning (by MAF, with boundaries at 0.001, 0.0025, 0.01 and 0.1) to to obtain reliable estimates of $h^2_{\text{SNP}}$ when rare variants (MAF $< 0.01$) are included.

**Datasets:** In total, we analyze data for 42 traits. Table 1 describes the 19 "GWAS traits" (17 case-control, 2 quantitative). For these, individuals were genotyped using either genome-wide Illumina or Affymetrix arrays (typically 500 K to 1.2 M SNPs). We additionally examine data from eight cohorts of the UCLEB consortium,[11] which comprise about 14 000 individuals genotyped using the Metabochip[12]

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

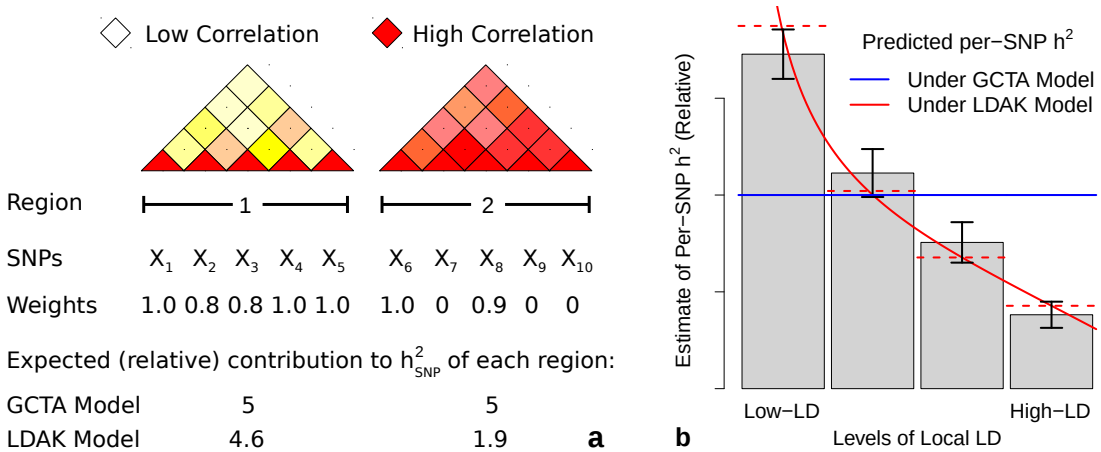


**Figure 1: (a) Comparison of the GCTA and LDAK Models.** Region 1 contains five SNPs in low LD (lighter colors indicate weaker pairwise correlations). Each SNP contributes unique genetic variation, reflected by SNP weights close to one. Region 2 contains five SNPs in high LD (strong correlations). The total genetic variation tagged by the region is effectively captured by two of the SNPs, and so the others receive zero weight. Under the GCTA Model, the regions are expected to contribute heritability proportional to their numbers of SNPs, here equal. Under the LDAK Model, they are expected to contribute proportional to their sums of SNP weights, here in the ratio 4.6:1.9. Note that the expected heritability can also depend on the allele frequencies and genotype certainty of the SNPs, but for simplicity, these factors are ignored here. **(b) Using SNP partitioning to model test.** SNP partitioning allows us to estimate the per-SNP heritabilities for different tranches, which we can then compare to those predicted by different heritability models. For example, to compare the GCTA and LDAK Models, we divide SNPs based on local levels of LD. The gray bars report (relative) estimates of per-SNP heritability, averaged across the 19 GWAS traits (vertical lines provide 95% confidence intervals). Under the GCTA Model, these are predicted to be constant (blue line). The LDAK Model assumes an inverse relationship between heritability and LD (solid red line), and thus higher-LD tranches are expected to have lower per-SNP heritability (dashed red lines). Note that in practice, we generally use two SNP tranches, four are used here for additional clarity.

(a relatively sparse array of 200 K SNPs selected based on previous GWAS) and recorded for a wide range of clinical phenotypes. From these, we consider 23 quantitative phenotypes (average sample size 8 200), which can loosely be divided into anthropomorphic (height, weight, BMI and waist circumference), physiological (lung capacity and blood pressure), cardiac (e.g., PR and QT intervals), metabolic (glucose, insulin and lipid levels) and blood chemistry (e.g., fibrinogen, interleukin 6 and haemoglobin levels). In general, our quality control is extremely strict; after imputing using IMPUTE2[13] and the 1000 Genome Phase 3 (2014) reference panel,[14] we retain only autosomal SNPs with MAF $> 0.01$ and information score $r_j > 0.99$. We only relax quality control when, using the UCLEB data, we explicitly examine the consequences of including lower-quality and rare SNPs.

Further details of our methods and datasets are provided in Online Methods. These include: (i) a description of LD Score Regression (LDSC),[8] which can loosely be viewed as a way to estimate $h^2_{\text{SNP}}$ under the GCTA Model (with $\alpha = -1$) using only summary statistics; (ii) how when estimating $h^2_{\text{SNP}}$ we give special consideration to highly-associated SNPs, which we define as those with $P < 10^{-20}$ from single-SNP analysis; (iii) how for the UCLEB data, we confirm that genotyping errors do not correlate with phenotype (which is important for the analyses where we include lower-quality SNPs).

## Results

**Relationship between heritability and MAF:** Varying the value of $\alpha$ used when scaling genotypes changes the assumed relationship between heritability and MAF; three example relationships are shown in Figure 2A. To investigate suitable $\alpha$, we analyze each of the 42 traits using seven values: $-1.25, -1, -0.75, -0.5, -0.25$, 0 and 0.25. We compare different $\alpha$ by calculating the LRT statistic corresponding to partitioning SNPs into MAF $< 0.1$ and MAF $> 0.1$; a lower statistic indicates better-fitting $\alpha$ (Supplementary Fig. 1).

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

| Collection | Trait (Disease Prevalence, %) | $n$ | $m$ | $\sum_{j=1}^{m} w_j$ | $h^2_{\text{GWAS}}$ | Estimates of $h^2_{\text{SNP}}$ (SD) Previous | LDAK |
|---|---|---|---|---|---|---|---|
| **Welcome Trust Case Control Consortium 1 (WTCCC 1)** | Bipolar Disorder (0.5) | 1 840 + 2 913 | 3 729 K | 79 K | 0.02 | 0.24 (0.04)[5] | 0.35 (0.03) |
| | Coronary Artery Disease (6) | 1 907 + 2 918 | 3 739 K | 80 K | 0.03 | 0.25 (0.06)[5] | 0.40 (0.06) |
| | Crohn's Disease (0.5) | 1 691 + 2 905 | 3 724 K | 79 K | 0.21 | 0.26 (0.01)[15] | 0.32 (0.03) |
| | Hypertension (5) | 1 918 + 2 916 | 3 740 K | 80 K | <0.01 | 0.33 (0.06)[5] | 0.46 (0.06) |
| | Rheumatoid Arthritis (0.5) | 1 846 + 2 918 | 3 736 K | 80 K | 0.19 | 0.09 (0.03)[5] | 0.21 (0.03) |
| | Type 1 Diabetes (0.5) | 1 941 + 2 907 | 3 732 K | 80 K | 0.27 | 0.13 (0.03)[5] | 0.30 (0.02) |
| | Type 2 Diabetes (8) | 1 896 + 2 917 | 3 736 K | 80 K | 0.08 | 0.42 (0.07)[5] | 0.54 (0.07) |
| **Welcome Trust Case Control Consortium 2 (WTCCC 2)** | Barrett's Oesophagus (1.6) | 1 861 + 5 138 | 4 831 K | 116 K | <0.01 | 0.25 (0.05)[16] | 0.32 (0.04) |
| | Ischaemic Stroke (2) | 3 769 + 5 139 | 4 797 K | 115 K | <0.01 | 0.25 (0.03)[17] | 0.34 (0.03) |
| | Parkinson's Disease (0.2) | 1 687 + 5 136 | 4 820 K | 116 K | 0.03 | 0.27 (0.05)[18] | 0.20 (0.03) |
| | Psoriasis (0.5) | 2 267 + 5 143 | 4 815 K | 116 K | 0.21 | 0.35 (0.06)[19] | 0.34 (0.02) |
| | Schizophrenia (1) | 2 068 + 2 615 | 3 481 K | 111 K | 0.07 | 0.23 (0.01)[20] | 0.30 (0.04) |
| | Ulcerative Colitis (0.2) | 2 614 + 5 327 | 4 062 K | 115 K | 0.12 | 0.19 (0.01)[15] | 0.28 (0.02) |
| **WTCCC 2 +** | Celiac Disease (1) | 2 492 + 7 376 | 3 682 K | 88 K | 0.29 | 0.33 (0.04)[21] | 0.37 (0.02) |
| | Multiple Sclerosis (0.1) | 8 553 + 5 667 | 4 702 K | 113 K | 0.17 | 0.17 (0.01)[5] | 0.24 (0.01) |
| | Partial Epilepsy (0.3) | 1 217 + 5 152 | 3 399 K | 108 K | <0.01 | 0.33 (0.05)[3] | 0.27 (0.04) |
| **RPTB** | Pulmonary Tuberculosis (4) | 5 142 + 5 283 | 3 987 K | 102 K | <0.01 | None Found | 0.26 (0.03) |
| **Blue Mountain** | Intraocular Pressure | 2 235 | 4 149 K | 125 K | 0.02 | None Found | 0.38 (0.17) |
| **CHOP** | Wide-Range Achievement Test | 3 747 | 3 593 K | 88 K | <0.01 | 0.43 (0.10)[22] | 0.21 (0.09) |
| **UCLEB** | 23 Quantitative Traits | 6 458 to 11 005 | 353 K | 39 K | - - - | Supplementary Table 1 | |

**Table 1: Properties of datasets and estimates of $h^2_{\text{SNP}}$.** $n$ = sample size (cases + controls), $m$ = number of SNPs, $\sum_{j=1}^{m} w_j$ = sum of SNP weights which can be interpreted as an effective number of independent SNPs. All values are post quality control; values for $m$ and $\sum w_j$ are rounded to the nearest K (thousand). For UCLEB, $m$ and $\sum w_j$ refer to our main analysis, which considers only high-quality, common SNPs. The final column provides our best estimates of $h^2_{\text{SNP}}$ from common SNPs, computed using LDAK with $\alpha = -0.25$ (see main text for explanation of the scaling parameter $\alpha$). For comparison, we include previously published estimates of $h^2_{\text{SNP}}$ (note that the analyses for rheumatoid arthritis, type 1 diabetes and multiple sclerosis excluded major histocompatibility SNPs, which we estimate contribute 0.07, 0.16 and 0.05, respectively), as well as $h^2_{\text{GWAS}}$, the proportion of phenotypic variance explained by SNPs reported as GWAS significant ($P < 5 \times 10^{-8}$). For disease traits, estimates of $h^2_{\text{SNP}}$ and $h^2_{\text{GWAS}}$ have been converted to the liability scale assuming the stated prevalence.

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

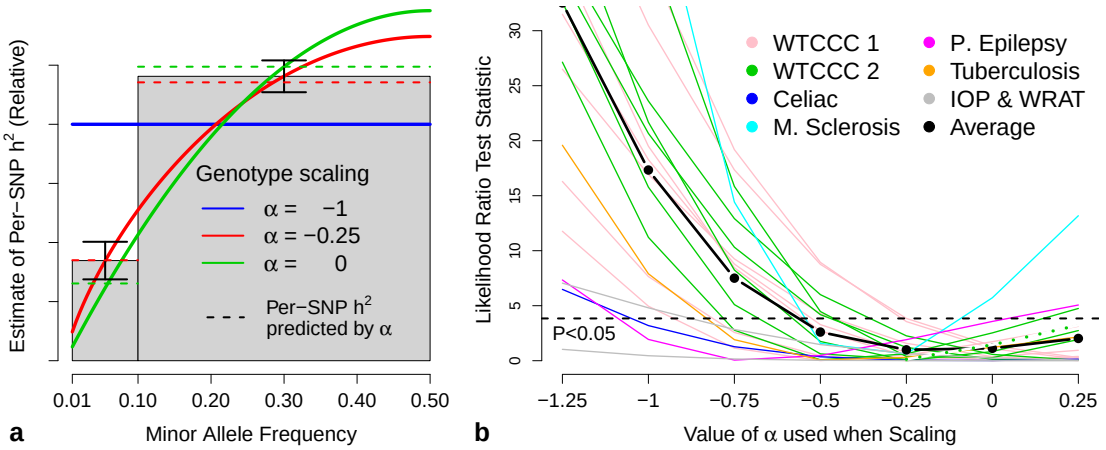


**Figure 2: (a) Specifying the relationship between heritability and MAF.** The parameter $\alpha$ specifies the assumed relationship between heritability and MAF: in human genetics, $\alpha = -1$ is typically used (solid blue line), while in animal and plant genetics, $\alpha = 0$ is more common (green); we instead recommend $\alpha = -0.25$ (red). To test different $\alpha$, we divide SNPs into MAF $< 0.01$ and MAF $> 0.01$. The gray bars report (relative) estimates of the per-SNP heritability for each tranche, averaged across the 19 GWAS traits (vertical lines provide 95% confidence intervals); the dashed lines indicate the per-SNP heritability predicted by each $\alpha$. **(b) Testing different $\alpha$ for the GWAS traits.** For seven values of $\alpha$, points report likelihood ratio test (LRT) statistics corresponding to partitioning SNPs into MAF $< 0.1$ and MAF $> 0.1$; lower statistics indicate better-fitting $\alpha$. Values above the horizontal line are significant ($P < 0.05$). Line colors indicate the seven trait categories, while the black line reports averages.

Full results are provided in Supplementary Figure 3. First, to remove any confounding due to LD, we perform this analysis using only a pruned subset of SNPs (with $w_j = 1$); next, we repeat without LD pruning (the results for the GWAS traits are shown in Figure 2B); finally, for the UCLEB traits, we repeat including lower-quality and rare SNPs. While optimal $\alpha$ varies by trait, the most widely-used value, $\alpha = -1$, tends to perform poorly, whereas $-0.5 \leq \alpha \leq 0$ tend to perform well. On the basis that it performs consistently well across different traits and SNP filterings, we recommend that $\alpha = -0.25$ becomes the default scaling. This value implies a tendency for heritability to decline with MAF, as seen in Figure 2A, which reports, averaged across the 19 GWAS traits, the (weight-adjusted) per-SNP heritability for low- and high-MAF SNPs (see Supplementary Figure 4 for further details).

While $\alpha = -0.25$ fits the data best overall, for individual traits, best-fitting $\alpha$ is likely to differ, and therefore we investigate the sensitivity of $h^2_{\text{SNP}}$ estimates to the value of $\alpha$. Full results are in provided Supplementary Figures 5, 6 & 7, while Figure 6A provides a summary for the UCLEB traits. When analyzing only common SNPs, we find that changes in $\alpha$ have little impact on $h^2_{\text{SNP}}$. For example, across the 23 UCLEB traits, estimates from high-quality common SNPs using $\alpha = -0.25$ are on average only 5% (SD 4) lower than those using $\alpha = -1$, and 4% (SD 4) higher than those using $\alpha = 0$. However, this is no longer the case when rare SNPs are included in the analysis: for example, when the MAF threshold is reduced to 0.0005, estimates using $\alpha = -0.25$ are on average 18% (SD 4) lower than those using $\alpha = -1$ and 30% (SD 6) higher than those from $\alpha = 0$. Therefore, when including rare SNPs, we guard against misspecification of $\alpha$ by partitioning based on MAF (with boundaries at 0.001, 0.0025, 0.01 and 0.1); we find that this provides stable estimates of $h^2_{\text{SNP}}$ and also allows estimation of the relative contributions of rare and common variants (Figure 6A and Supplementary Figure 8).

**Relationship between heritability and LD:** For each of the 19 GWAS traits, Figure 3A reports relative estimates of $h^2_{\text{SNP}}$ from GCTA, GCTA-MS, GCTA-LDMS and LDAK (all using $\alpha = -0.25$). We find that estimates of $h^2_{\text{SNP}}$ from LDAK are on average 45% (SD 3) higher than estimates from GCTA or GCTA-MS, and 14% (SD 2) higher than those from GCTA-LDMS. Results are very similar if instead we use $\alpha = -1$ (Supplementary Fig. 9). For the 23 UCLEB traits, estimates from LDAK are on average 88% (SD 7) higher than those from GCTA (Supplementary Fig. 10; convergence issues prevent us from comparing with GCTA-MS and GCTA-LDMS). Figure 3A also includes estimates from LDSC, which we run as described in the original publication;[8] in Supplementary Figure 11 we consider alternative versions of LDSC (e.g., varying how LD Scores are computed, forcing the intercept term to be zero and excluding

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

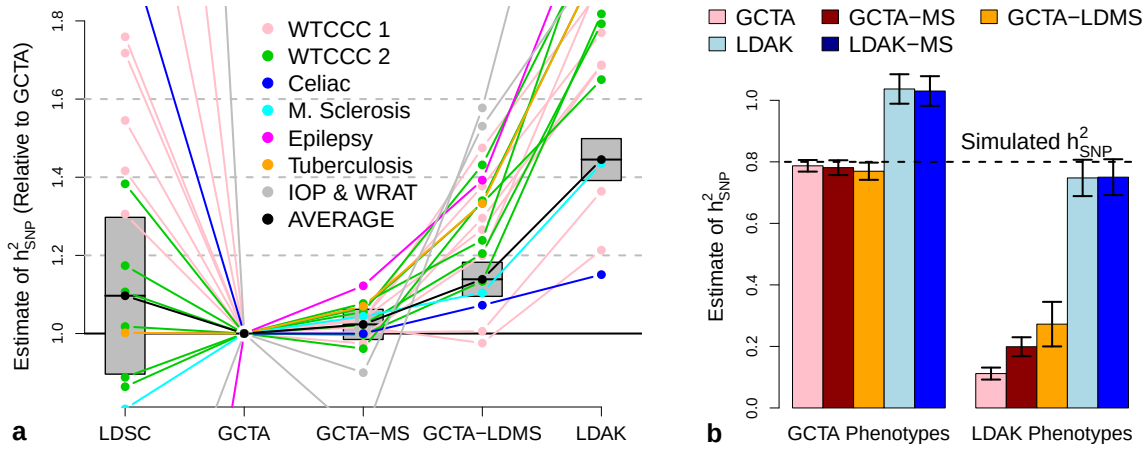


**Figure 3: (a) Relative estimates of $h^2_{\text{SNP}}$ for the GWAS traits.** $h^2_{\text{SNP}}$ estimates from LDSC, GCTA-MS (SNPs partitioned by MAF), GCTA-LDMS (SNPs partitioned by LD and MAF) and LDAK are reported relative to those from GCTA. Line colors indicate the seven trait categories; the black line reports the (inverse variance weighted) averages, with gray boxes providing 95% confidence intervals for these averages. **(b) Simulation studies can be misleading.** Phenotypes are simulated with 1000 causal SNPs and $h^2_{\text{SNP}} = 0.8$ (black horizontal line), then analyzed using GCTA, GCTA-MS, GCTA-LDMS, LDAK and LDAK-MS (LDAK with SNPs partitioned by MAF). Bars report average $h^2_{\text{SNP}}$ across 200 simulated phenotypes (vertical lines provide 95% confidence intervals). **Left:** copying the study of Yang *et al.*,[1] causal SNP effect sizes are sampled from $\mathbb{N}(0, 1)$, similar to the GCTA Model. **Right:** causal SNP effect sizes are sampled from $\mathbb{N}(0, w_j)$, similar to the LDAK Model.

highly-associated SNPs). While changing settings can have a large impact, in all cases the average estimate of $h^2_{\text{SNP}}$ from LDSC remains substantially below that from LDAK.

A recent article asserting that GCTA estimates $h^2_{\text{SNP}}$ more accurately than LDAK, based this claim on a simulation study in which 1 000 causal SNPs were chosen at random and assigned effect sizes from the same Gaussian distribution, irrespective of LD.[4] This resembles the GCTA Model but not the LDAK Model, and as expected GCTA performs better. Figure 3B shows that if instead effect size variances are scaled by SNP weights, and thus depend on local LD, the LDAK analysis is superior to GCTA. See Supplementary Figure 12 for a full reanalysis of the reported simulation study.

**Which LD model is best?** To empirically test which of the GCTA and LDAK models better reflects the properties of real datasets, we partition each dataset into low- and high-LD SNP tranches based on the average LD Score[8] of non-overlapping 100 kb segments (i.e., low-LD SNPs are those in segments with low average LD Score, and vice versa). For Partition I, we divide SNPs so that the two tranches are predicted to contribute equally under the GCTA Model, while for Partition II, we ensure that they are predicted to contribute equally under the LDAK Model (the predicted contribution of a tranche is determined by how much its SNPs contribute towards $\sum_{j=1}^{m}(f_j(1-f_j))^{1-\alpha}w_j r_j$). We analyze each trait using GCTA and LDAK (both with $\alpha = -0.25$), then compute the LRT statistic corresponding to partitioning SNPs by LD: a lower statistic indicates the corresponding model better reflects the data (Supplementary Fig. 2). Full numerical results are provided in Supplementary Table 2. We first consider the 19 GWAS Traits. For Partition I, when assuming the GCTA Model, the mean (median) statistic is 10.6 (10.5); when assuming the LDAK Model, the mean (median) is 1.5 (0.7), and the statistic is lower than under the GCTA Model for 16 of the 19 traits. For Partition II, the mean (median) statistic is 11.6 (7.7) under the GCTA Model and 2.0 (1.1) under the LDAK Model, and assuming the LDAK model leads to a lower statistic for 18 of the 19 traits. Results are similar for the UCLEB data: for 31 of the 46 comparisons the LRT statistic is lower when assuming the LDAK Model (mean 3.7, median 1.1) than when assuming the GCTA Model (mean 7.2, median 3.4). Thus we conclude that, for a wide-range of traits, the LDAK Model fits substantially better than the GCTA Model. Supplementary Table 3 shows that the conclusion remains the same if we instead perform the analyses using $\alpha = -1$ (rather than $\alpha = -0.25$).

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

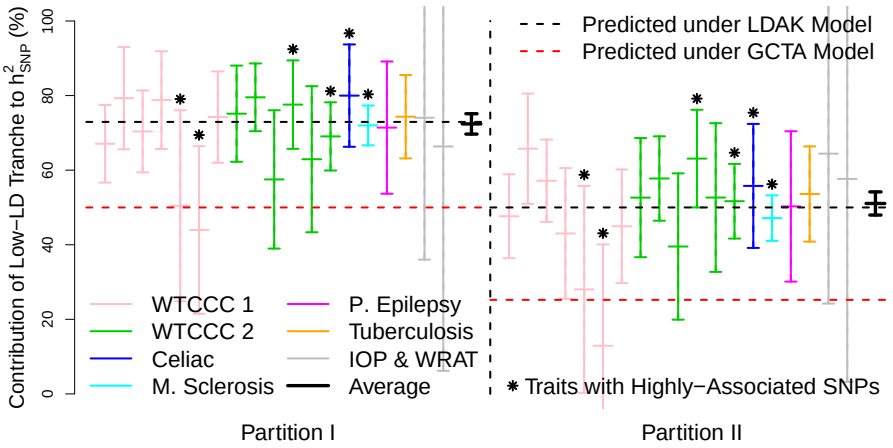


**Figure 4: Comparing the GCTA and LDAK Models for the GWAS traits:** We consider two partitions of SNPs into low- and high-LD tranches. The horizontal red lines indicate, for Partition I (left) and Partition II (right), the predicted contribution of the low-LD tranche to $h^2_{\text{SNP}}$ under the GCTA Model; the horizontal black lines indicate its predicted contribution under the LDAK Model. For each trait (colored by category), vertical lines provide point estimates and 95% confidence intervals for the contribution of the low-LD tranche to $h^2_{\text{SNP}}$, estimated assuming the GCTA Model; the black lines provides the (inverse variance weighted) averages. $*$ marks traits with one or more highly-associated SNPs. For this plot, the variance explained by these SNPs has been excluded; see Supplementary Figure 13 for results with these included.

To display these results visually, we compare the estimated contributions to $h^2_{\text{SNP}}$ of each tranche with those predicted under the GCTA and LDAK Models. Figure 4 reports, for each of the GWAS traits, the contribution of the low-LD tranche, estimated using the GCTA Model (with $\alpha = -0.25$). For Partition I, the estimate is consistent with the GCTA Model (95% confidence interval overlaps 0.50) for only 6 of the 19 traits, whereas it is consistent with the LDAK Model (confidence interval overlaps 0.73) for 17 traits. Similarly, for Partition II, the estimate is consistent with the GCTA Model for only 5 traits, but consistent with the LDAK Model for 16. Supplementary Figure 13 provides the corresponding plots when we instead estimate heritabilities under the LDAK Model or using $\alpha = -1$, and when we consider the UCLEB traits.

**Relationship between heritability and genotype certainty:** We now consider including lower-quality SNPs, focusing on the 23 UCLEB traits, as for these we were able to test for correlation between genotyping errors and phenotype (Supplementary Fig. 14). We partition the 8.8 M common SNPs into high-quality ($r_j > 0.6$; 5.0 M SNPs) and low-quality ($r_j \leq 0.6$; 3.9 M SNPs). Under the LDAK Model (with $\alpha = -0.25$), the LRT statistic corresponding to this partitioning has mean (median) 2.2 (1.3). If instead we set $r_j = 1$ for all SNPs, and thus ignore genotype certainty,[9,10,23] the LRT statistic has mean (median) 3.4 (1.3) and is increased for 17 out of 23 traits (Supplementary Fig. 15). This indicates that when lower-quality SNPs are included, weighting the effect-size prior variance by $r_j$ improves model specification.

**Estimates of $h^2_{\text{SNP}}$ for the GWAS traits:** Table 1 presents our final estimates of $h^2_{\text{SNP}}$ for the 19 GWAS traits, obtained using the LDAK Model (with $\alpha = -0.25$). For comparison, we report previous estimates of $h^2_{\text{SNP}}$, as well as the proportion of phenotypic variance explained by SNPs reported as genome-wide significant (see Supplementary Table 4). For the disease traits, estimates are on the liability scale, obtained by scaling according to the observed case-control ratio and (assumed) trait prevalence.[24,25] We are unable to find previous estimates of $h^2_{\text{SNP}}$ for tuberculosis or intraocular pressure, indicating that for these two traits, we are the first to establish that common SNPs contribute sizable heritability. Extended results are provided in Supplementary Table 5. These show that our final estimates of $h^2_{\text{SNP}}$ are on average 40% (SD 3) and 25% (SD 3) higher than, respectively, those obtained using the original versions (i.e., with $\alpha = -1$) of GCTA[26] and GCTA-LDMS.[10]

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

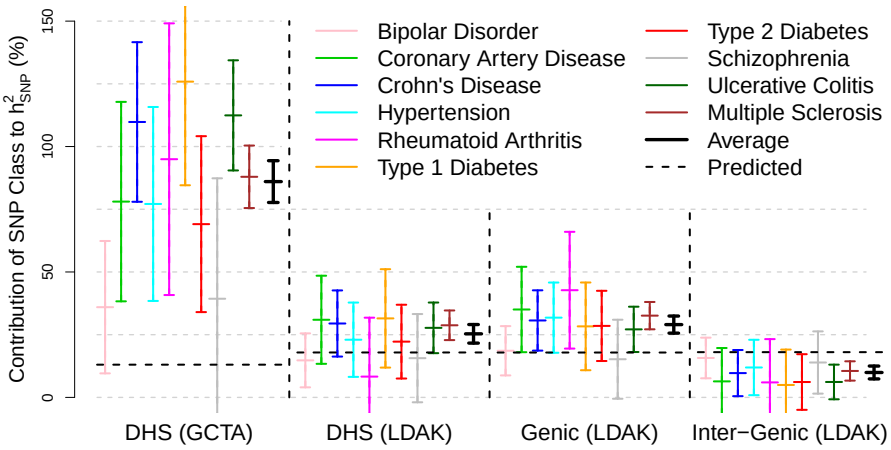


**Figure 5: Enrichment of SNP Classes.** Block 1 reports the contributions to $h^2_{\text{SNP}}$ of DNaseI hypersensitivity sites (DHS), estimated under the GCTA Model. The vertical lines provide point estimates and 95% confidence intervals for each trait, and for the (inverse variance weighted) average; for 3 of the traits, the point estimate is above 100%, as was also the case for Gusev *et al.*[5] Block 2 repeats this analysis, but instead assuming the LDAK Model. Blocks 3 & 4 examine the contribution of "genic SNPs" (those inside or within 2 kb of exons) and "inter-genic SNPs" (further than 125 kb), again assuming the LDAK Model. To calculate enrichment, estimated contributions are compared to those predicted under the GCTA or LDAK Model, as appropriate (horizontal lines).

**Role of DNaseI hypersensitivity sites (DHS):** Gusev *et al.*[5] used SNP partitioning to assess the contributions of SNP classes defined by functional annotations. Across 11 diseases they concluded that the majority of $h^2_{\text{SNP}}$ was explained by DHS, despite these containing less than 20% of all SNPs. For Figure 5, we perform a similar analysis using the 10 traits we have in common with their study (for 9 of these, we are using the same data). When we copy Gusev *et al.* and assume the GCTA Model with $\alpha = -1$, we estimate that DHS contribute 86% (SD 4) of $h^2_{\text{SNP}}$, close to the value they reported (79%). When instead we assume the LDAK Model (with $\alpha = -0.25$), the estimated contribution of DHS reduces to 25% (SD 2). Under the LDAK Model, DHS are predicted to contribute 18% of $h^2_{\text{SNP}}$ so 25% represents 1.4-fold enrichment. To add context, we also consider "genic" SNPs, which we define as SNPs inside or within 2 kb of an exon (using RefSeq annotations[27]), and "inter-genic," SNPs further than 125 kb from an exon; these definitions ensure that these two SNP classes are also predicted to contribute 18% of $h^2_{\text{SNP}}$ under the LDAK Model. We estimate that genic SNPs contribute 29% (SD 2), while inter-genic SNPs contribute 10% (SD 2), representing 1.6- and 0.6-fold enrichment, respectively. When we extend this analysis to all 42 traits, DHS on average contribute 24% (SD 2) of $h^2_{\text{SNP}}$, and in contrast to Gusev *et al.*, enrichment remains constant when we reduce SNP density (Supplementary Fig. 16 & 17 and Supplementary Table 6).

**Relaxing quality control:** For the UCLEB data, we consider nine alternative SNP filterings. Supplementary Figure 18 reports estimates of $h^2_{\text{SNP}}$ for each trait / filtering, while Figure 6A provides a summary. First we vary the information score threshold: $r_j > 0.99$, $> 0.95$, $> 0.9$, $> 0.6$, $> 0.3$ and $> 0$ (each time continuing to require MAF $> 0.01$). Simulations suggest that by including all 8.8 M common SNPs ($r_j > 0$), instead of using just the 353 K high-quality ones ($r_j > 0.99$), we can expect estimates of $h^2_{\text{SNP}}$ to increase by 50-60% (Supplementary Fig. 19). This is similar to what we observe in practice, as across the 23 traits, estimates of $h^2_{\text{SNP}}$ (using $\alpha = -0.25$) are on average 45% (SD 8) higher. The simulations further predict that, even though the Metabochip provides relatively low coverage of the genome (after quality control, it contains only 60 K SNPs, predominately within genes), we can expect estimates of $h^2_{\text{SNP}}$ to be approximately 80% as high as those obtained starting from genome-wide genotyping arrays. While we are unable to test this claim directly, it is consistent with our results for height, body mass index and QT Interval, the three traits for which reasonably precise estimates of $h^2_{\text{SNP}}$ are available[4] (Figure 6B). For the final three SNP filterings, we vary the MAF threshold: MAF $> 0.0025$, MAF $> 0.001$ and MAF $> 0.0005$ (all with $r_j > 0$). Across the 23 traits, we find that rare SNPs contribute substantially to $h^2_{\text{SNP}}$: for example, when we use the 17.3 M SNPs with MAF $> 0.0005$, estimates of $h^2_{\text{SNP}}$ (using $\alpha = -0.25$ and MAF partitioning) are on average 29% (SD 12) higher than those based on the 8.8 M common SNPs, with rare variants contributing on average 33% (SD 5) of $h^2_{\text{SNP}}$ (Figure 6A).

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

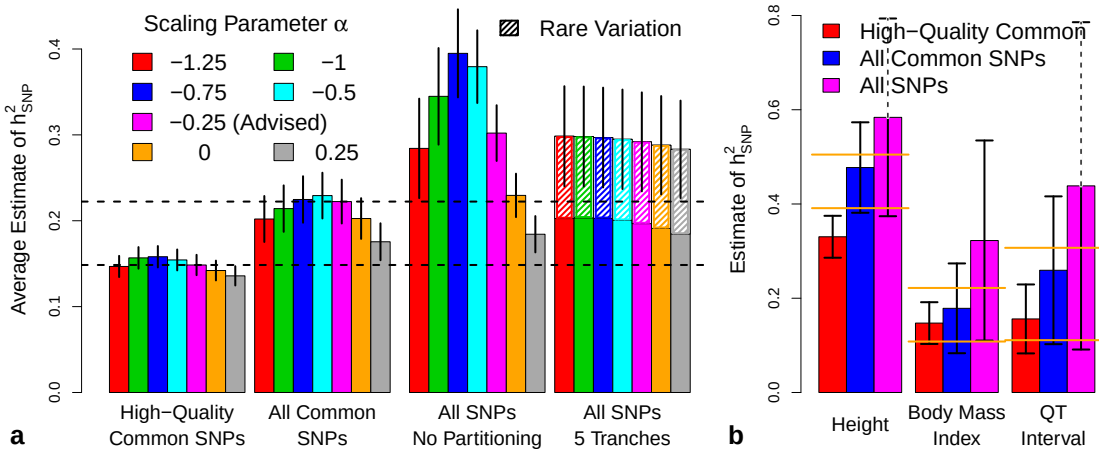


**Figure 6: Varying quality control for the UCLEB traits.** We consider three SNP filterings: 353 K high-quality common SNPs (information score $>0.99$, MAF $>0.01$), 8.8 M common SNPs (MAF $>0.01$), and all 17.3 M SNPs (MAF $>0.0005$). **(a)** Blocks indicate SNP filtering; bars report (inverse variance weighted) average estimates of $h^2_{\text{SNP}}$ (vertical lines provide 95% confidence intervals). Bar color indicates the value of $\alpha$ used. For Blocks 1, 2 & 3, $h^2_{\text{SNP}}$ is estimated using the non-partitioned model. For Block 4, SNPs are partitioned by MAF; we find this is necessary when rare SNPs are included, and also allows estimation of the contribution of MAF $<0.01$ SNPs (hatched areas). **(b)** Bars report our final estimates of $h^2_{\text{SNP}}$ for height, body mass index and QT interval, the three traits for which common SNP heritability has been previously estimated with reasonable precision[4] (orange lines mark the 95% confidence intervals from these previous studies). Bar colors now indicate SNP filtering; all estimates are based on $\alpha = -0.25$, using either a non-partitioned model (red and blue bars) or with SNPs partitioned by MAF (purple bars).

## Discussion

With estimates of $h^2_{\text{SNP}}$ so widely reported, it is easy to forget that calculating the variance explained by large numbers of SNPs is a challenging problem. To avoid over-fitting, it is necessary to make strong prior assumptions about SNP effect sizes, but different assumptions can lead to substantially different estimates of $h^2_{\text{SNP}}$. Previous attempts to assess the validity of different assumptions have used simulation studies,[9,10] but this approach will tend to favor the assumptions used to generate the phenotypes. Instead, we have used SNP partitioning to empirically test different models for the distribution of heritability across the genome. For example, the GCTA Model assumes that heritability does not depend on local levels of LD; by partitioning the genome into low- and high-LD regions, we have demonstrated the inadequacy of this assumption over a large collection of real datasets. By contrast, applying the same test to the LDAK Model indicates that the relationship it specifies between heritability and LD provides a better description of reality.

This finding has important consequences for complex trait genetics. Firstly, it implies that for many traits, common SNPs explain considerably more phenotypic variance than previously reported, which represents a significant advance in the search for missing heritability.[2] It also impacts a large number of closely-related methods. For example, LDSC,[8] like GCTA, assumes that heritability contributions are independent of LD and therefore also tends to under-estimate $h^2_{\text{SNP}}$. Similarly, we have shown that estimates of the relative importance of SNP classes via SNP partitioning can be misleading when the GCTA Model is assumed.[5] Further afield, most software for mixed model association analyses (e.g., FAST-LMM, GEMMA and MLM-LOCO) use an extension of the GCTA Model,[28–30] and likewise most bivariate analyses, including those performed by LDSC.[6,31–33] It remains to be seen how much these methods would be affected if they employed more realistic heritability models.

We also used SNP partitioning to investigate the relationship between heritability and MAF. By dividing the genome into low- and high-MAF SNPs, we showed that prior specification is improved by using $\alpha = -0.25$ when scaling genotypes, which implies that heritability varies with $(\text{MAF}(1-\text{MAF}))^{0.75}$. This is important because many penalized and Bayesian regression methods, for example, the Lasso, ridge regression and Bayes A,[34–36] assign the same penalty function or prior distribution to all SNPs. Whenever such a

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

method is applied to standardized genotypes, this implicitly assumes that heritability is independent of MAF. Thus our result suggests that the performance of these methods (i.e., their detection power and/or prediction accuracy) could be improved simply by changing how genotypes are scaled. We have investigated a one-parameter family of relationships (defined by varying the scaling parameter $\alpha$); with additional data, there is the potential to improve model specification further by considering other relationships. In particular, with a better understanding of the relationship between heritability and MAF for low frequencies, it may no longer be necessary to partition by MAF when rare variants are included.

Although we used SNP partitioning to test different model assumptions, we could have reached the same conclusions by comparing model likelihoods (Supplementary Table 7). An advantage of SNP partitioning is that it allowed us to systematically examine different features of the heritability model. We investigated the role of MAF, LD and genotype certainty, but there are other factors on which heritability could depend, in particular the known functions of different genomic locations.[37] For example, our comparison of genic and inter-genic SNPs indicates that the effect-size prior distribution could be improved by taking into account proximity to coding regions. By way of demonstration, Supplementary Figure 20 shows that model fit is improved by assuming $\beta_j \sim \mathbb{N}(0, \exp(\frac{-(D+50)}{500}) r_j w_j \sigma_g^2 / W)$, where $D$ is the distance (in kb) between SNP $j$ and the nearest exon (this prior distribution ensures that genic SNPs have expected heritability about twice as high as inter-genic SNPs). We believe that modifications of this type will have a relatively small impact; we note that across the 19 GWAS traits, scaling the effect-size prior variance by $\exp(\frac{-(D+50)}{500})$ increases model log likelihood by on average only 1.6, much less than the average increase obtained by using $\alpha = -0.25$ instead $\alpha = -1$ (9.1), or by choosing the LD-model specified by LDAK instead of GCTA (17.1), and does not significantly change estimates of $h^2_{\text{SNP}}$. However, with sufficient data, it may be possible to obtain more substantial improvement by tailoring model assumptions to individual traits.

When estimating $h^2_{\text{SNP}}$, care should be taken to avoid possible sources of confounding. Previously, we advocated a test for inflation of $h^2_{\text{SNP}}$ due to population structure and familial relatedness.[3] The conclusions of a recent paper claiming that $h^2_{\text{SNP}}$ estimates are unreliable,[38] would have changed substantially had this test been applied (Supplementary Fig. 21). We also recommend testing for inflation due to genotyping errors, particularly before including lower-quality and/or rare SNPs. For the 23 UCLEB traits, we showed that including poorly-imputed SNPs resulted in significantly higher estimates of $h^2_{\text{SNP}}$, and made it possible to capture the majority of genome-wide heritability despite the very sparse genotyping provided by the Metabochip. We found that including rare SNPs also led to significantly higher $h^2_{\text{SNP}}$. Although sample size prevented us from obtaining precise estimates of $h^2_{\text{SNP}}$ for individual traits, our analyses indicated that for larger datasets, including rare SNPs will be both practical and fruitful in the search for the remaining missing heritability.[2]

## Acknowledgments

Access to Wellcome Trust Case Control Consortium data was authorized as work related to the project "Genome wide association study of susceptibility and clinical phenotypes in epilepsy. "We thank Anne Molloy, James Mills and Lawrence Brody for permission to use genotype data from the Trinity College Dublin Student Study.[39] This work is funded by the UK Medical Research Council under grant MR/L012561/1, by the British Heart Foundation under grant RG/10/12/28456, and supported by researchers at the National Institute for Health Research (NIHR) University College London Hospitals Biomedical Research Centre. SN is a Wellcome Trust Senior Research Fellow in Basic Biomedical Science and is also supported by the NIHR Cambridge Biomedical Research Centre. Analyses were performed with the use of the UCL Computer Science Cluster and the help of the CS Technical Support Group, as well as the use of the UCL Legion High Performance Computing Facility (Legion@UCL) and associated support services.

## Author Contributions

DS and NC performed the analyses. DS and DJB wrote the manuscript with assistance from NC, MRJ, SN and members of the UCLEB Consortium.

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Online Methods

Supplementary Note 1 summarizes the different analyses we performed, and the conclusions we drew from each, while Supplementary Protocol 1 provides step-by-step instructions for estimating $h^2_{\text{SNP}}$ starting from raw genotype data. As a reminder, the length-$n$ vector $\boldsymbol{Y}$ contains phenotypic values, the $n \times p$ matrix $\boldsymbol{Z}$ contains covariates, while the $n \times m$ matrix $\boldsymbol{S}$ contains allele counts.

**Information score $r_j$:** Let the vector $\boldsymbol{S_j} = (S_{1,j}, \ldots, S_{n,j})^T \in [0,2]^n$, denote the (expected) allele counts for SNP $j$ (i.e., $S_j$ is Column $j$ of $\boldsymbol{S}$). Our information score $r_j$ estimates the squared correlation between $\boldsymbol{S_j}$ and $\boldsymbol{G_j} = (G_{1,j}, \ldots, G_{n,j})^T \in \{0,1,2\}^n$, the true genotypes for SNP $j$. When using imputed data, $\boldsymbol{G_j}$ is typically not known; instead for each individual we have a triplet of state probabilities $(p_{i,j,0}, p_{i,j,1}, p_{i,j,2})$, where $p_{i,j,g} = \mathbb{P}(G_{i,j} = g)$ and $p_{i,j,0} + p_{i,j,1} + p_{i,j,2} = 1$. Therefore, we define $r_j$ by taking expectations over the $3^n$ possible realizations of $G_j$.

$$r_j = \frac{\mathbb{E}\big[\sum_{i=1}^n (S_{i,j} - \bar{S}_j)(G_{i,j} - \bar{G}_j)\big]^2}{\big(\sum_{i=1}^n (S_{i,j} - \bar{S}_j)^2\big)\,\mathbb{E}\big[\sum_{i=1}^n (G_{i,j} - \bar{G}_j)^2\big]}, \quad \text{where} \quad \bar{S}_j = \frac{1}{n}\sum_{i=1}^n S_{i,j} \quad \text{and} \quad \bar{G}_j = \frac{1}{n}\sum_{i=1}^n G_{i,j}.$$

$S_j$ is known, so computing $\sum_{i=1}^n (S - \bar{S}_j)^2$ is straightforward. The two expectations can also be calculated explicitly:

$$\mathbb{E}\big[\textstyle\sum_{i=1}^n (S_{i,j} - \bar{S}_j)(G_{i,j} - \bar{G}_j)\big] = \sum_i (S_j - \bar{S}_j)\mathbb{E}[G_j - \mu] = \sum_i (S_j - \bar{S}_j)(p_{i,1} + 2p_{i,2} - \mu)$$
$$\mathbb{E}\big[\textstyle\sum_i (G_j - \bar{G}_j)^2\big] = \sum_i \mathbb{E}\big[(G_j - \mu)^2\big] = \sum_i \big[p_{i,j,0}(-\mu)^2 + p_{i,j,1}(1-\mu)^2 + p_{i,j,2}(2-\mu)^2\big],$$

where $\mu = \mathbb{E}[\bar{G}_j] = \frac{1}{n}\sum_i (p_{i,j,1} + 2p_{i,j,2})$. For our analyses, we use expected allele counts (dosages), so $S_{i,j} = p_{i,j,1} + 2p_{i,j,2}$. In this case, $\mathbb{E}\big[\sum_{i=1}^n (S_{i,j} - \bar{S}_j)(G_{i,j} - \bar{G}_j)\big] = \sum_{i=1}^n (Si,j - \bar{S}_j)^2$ and so the score reduces to $r_j = \sum_{i=1}^n (S_{i,j} - \bar{S}_j)^2 / \sum_{i=1}^n (G_{i,j} - \bar{G}_j)^2$. For a directly genotyped SNP, each triplet of state probabilities will be (1,0,0), (0,1,0) or (0,0,1), which will result in $S_{i,j} = G_{i,j}$ for all $i$ and $r_j = 1$; so for these, in place of $r_j$, we use the metric `r2_type0` reported by IMPUTE2.[13] Additional details on our information score are provided in Supplementary Figure 22.

**Estimating $h^2_{\text{SNP}}$:** We first construct the $n \times m$ genotype matrix $\boldsymbol{X}$, by centering and scaling the allele counts for each SNP according to $X_{i,j} = (S_{i,j} - 2f_j) \times (2f_j(1-f_j))^{\alpha/2}$, where $f_j = \sum_i S_{i,j}/2n$. If $w_j$ and $r_j$ denote the LD weighting[7] and information score for SNP $j$, then the LDAK Model for estimating SNP heritability $h^2_{\text{SNP}} = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$ is:

$$Y_i = \sum_{k=1}^p \theta_k Z_{i,k} + \sum_{j=1}^m \beta_j X_{i,j} + e_i, \quad \text{with} \quad \beta_j \sim \mathbb{N}(0, r_j w_j \sigma_g^2/W), \quad e_i \sim \mathbb{N}(0, \sigma_e^2) \quad \text{and} \quad W = \sum_{j=1}^m r_j w_j (2f_j(1-f_j))^{1-\alpha}. \quad (2)$$

$\theta_k$ denotes the fixed-effect coefficient for $k$ covariate, $\beta_j$ and $e_i$ are random-effects indicating the effect size of SNP $j$ and the noise component for Individual $i$, while $\sigma_g^2$ and $\sigma_e^2$ are interpreted as genetic and environmental variances, respectively. Note that the introduction of $r_j$ is an addition to the model we proposed in 2012.[7] Model (2) is equivalent to assuming:[40,41]

$$\boldsymbol{Y} \sim \mathbb{N}(\boldsymbol{Z\theta}, \boldsymbol{K}\sigma_g^2 + \boldsymbol{I}\sigma_e^2), \quad \text{with} \quad \boldsymbol{K} = \frac{\boldsymbol{X\Omega X}^T}{W}, \quad (3)$$

where $\boldsymbol{I}$ is an $n \times n$ identity matrix and $\boldsymbol{\Omega}$ denotes a diagonal matrix with diagonal entries $(r_1 w_1, \ldots, r_m w_m)$. The kinship matrix $\boldsymbol{K}$, also referred to as a genetic relationship matrix (GRM)[1] or genomic similarity matrix (GSM),[42] consists of average allelic correlations across the SNPs (adjusted for LD and genotype certainty). Model (3) is typically solved using REstricted Maximum Likelihood (REML), which returns estimates of $\theta_1, \ldots, \theta_p$, $\sigma_g^2$ and $\sigma_e^2$.[43]

The heritability of SNP $j$ can be estimated by $h_j^2 = \beta_j^2\,\text{Var}(X_j)/\text{Var}(Y)$, which under Model (2), and assuming Hardy-Weinberg Equilibrium,[44,45] has expectation

$$\mathbb{E}[h_j^2] = \frac{\mathbb{E}[\beta_j^2] \times \text{Var}(X_j)}{\text{Var}(Y)} = \frac{r_j w_j \sigma_g^2/W \times (2f_j(1-f_j))^{1+\alpha}}{\text{Var}(Y)}. \quad (4)$$

If $P_1$ and $P_2$ index two sets of SNPs of size $|P_1|$ and $|P_2|$, then under the LDAK Model, they are expected to contribute heritability in the ratio $W_1 : W_2$, where $W_l = \sum_{j \in P_l} r_j w_j (2f_j(1-f_j))^{1-\alpha}$. The GCTA Model corresponds to setting $w_j = r_j = 1$, in which case

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$W_l = \sum_{j \in P_l} (f_j(1-f_j))^{1-\alpha}$. Most applications of GCTA have further assumed $\alpha = -1$, so that $W_l = |P_l|$, which corresponds to the assumption that SNP sets are expected to contribute heritability proportional to the number of SNPs they contain.

Model (2) assumes that all effect-sizes can be described by a single prior distribution. This assumption is relaxed by SNP partitioning. Suppose that the SNPs are divided into tranches $P_1, \ldots, P_L$ of sizes $|P_1|, \ldots, |P_L|$; typically these will partition the genome, so that each SNP appears in exactly one tranche and $\sum_l |P_l| = m$, but this is not required. This correspond to generalizing Model (2), so that SNPs in Tranche l have effect-size prior distribution $\beta_j \sim \mathbb{N}(0, r_j w_j \sigma_l^2 / W_l)$. Letting $\Sigma = \sigma_1^2 + \ldots + \sigma_L^2$, then $h^2_{\text{SNP}} = \Sigma/(\Sigma + \sigma_e^2)$, while $\sigma_l^2/\Sigma$ represents the contribution to $h^2_{\text{SNP}}$ of SNPs in Tranche $l$. This model can equivalently be expressed as $\boldsymbol{Y} \sim \mathbb{N}(\boldsymbol{Z\theta}, \boldsymbol{K}_1\sigma_1^2 + \ldots + \boldsymbol{K}_L\sigma_L^2 + \boldsymbol{I}\sigma_e^2)$, where $\boldsymbol{K}_l$ represents allele correlations across the SNPs in Tranche $l$.

For analyses under the LDAK Model, we used LDAK v.5; for analyses under the GCTA Model, we used GCTA v.1.26. For about a third of GCTA-LDMS analyses, the GCTA REML solver failed with the error "information matrix is not invertible," in which case we rerun using LDAK (while the GCTA and LDAK solvers are both based on Average Information REML,[26,46] subtle differences mean that when using a large number of tranches, one might complete while the other fails). For the few occasions when both solvers failed, we instead used "GCTA-LD" (i.e., SNPs divided only by LD, rather than by LD and MAF), which we found gave very similar results to GCTA-LDMS for traits where both completed (Supplementary Fig. 9). For diseases, we converted estimates of $h^2_{\text{SNP}}$ to the liability scale based on the observed case-control ratio and assumed prevalence.[24,25] In general, we copied the prevalences used by previous studies, however for tuberculosis, where no previous estimate of $h^2_{\text{SNP}}$ is available, we derived an estimate of prevalence from World Health Organization data[47] (Supplementary Note 2).

**LD Score Regression (LDSC):** Originally designed as a way to quantify confounding in a GWAS, LDSC[8] also provides a method for estimating $h^2_{\text{SNP}}$ using only summary statistics from single-SNP analysis (rather than raw genotype and phenotype data). LDSC is based on the principal that in a single-SNP analysis, the $\chi^2(1)$ test statistic for SNP $j$ has expected value $\mathbb{E}[\chi_j^2] = h_j^2 + \sum_{k \neq j} r_{j,k}^2 h_k^2 + a_j$, where $r_{j,k}^2$ denotes the squared correlation between SNPs $j$ and $k$, while $a_j$ represents bias due to confounding factors (e.g., population structure and familial relatedness).[8] Under a polygenic model where every SNP is expected to contribute equally (i.e., $\mathbb{E}[h_j^2] = h^2_{\text{SNP}}/m$), and the (widely-used) assumption that the bias is constant across SNPs ($a_j = a$), we have $\mathbb{E}[\chi_j^2] = 1 + n l_j h^2_{\text{SNP}}/m + a$, where $l_j = 1 + \sum_{k \neq j} r_{j,k}^2$ is referred to as the LD Score of SNP $j$ (as it is not feasible to compute pairwise correlations across all SNPs, in practice these are approximated using a sliding window of, say, 1 centiMorgan). Therefore, LDSC estimates $h^2_{\text{SNP}}$ and $a$ by regressing test statistics on LD Scores. In the absence of confounding ($a = 0$), LDSC can be viewed as estimating $h^2_{\text{SNP}}$ under the GCTA Model with $\alpha = -1$ (as this satisfies the assumption that every SNP is expected to contribute equal heritability). As the authors of LDSC point out,[8] it is straightforward to accommodate alternative relationships between $\mathbb{E}[h_j^2]$ and MAF (i.e., $\alpha \neq -1$) by changing how genotypes are scaled when computing LD Scores, and potentially genotype certainty could be accommodated. However, the assumption that heritability is independent of LD appears intrinsic to LDSC, and so we can not envisage how the method could be modified to accommodate a more realistic LD Model. For LDSC analyses, we used LDSC v.1.0.0 both for calculating LD Scores and estimating $h^2_{\text{SNP}}$.

**Accommodating very large effect loci:** Equation (2) assumes that all SNP effect sizes can be modeled by a single Gaussian distribution. Estimates are generally robust to violations of this assumption,[7] but problems can occur when individual SNPs have very large effect sizes, because a single Gaussian distribution cannot accommodate both these SNPs and the very many with small effect sizes. This is a common concern when analyzing autoimmune traits for which the major histocompatibility complex (MHC) can contribute substantial heritability. In response to this problem, some authors exclude MHC SNPs from analyses.[5,26,48,49] Another approach is to model effect sizes as a mixture of Gaussians,[36,50] but this is not computationally feasible for millions of SNPs and many thousands of individuals. Therefore, our proposed strategy is to first identify SNPs with $P < 10^{-20}$ from single-SNP analysis, to prune these using a correlation squared threshold of 0.95, then to include those which remain as fixed-effect covariates. Thus in place of Equation (3), we assume $\boldsymbol{Y} \sim \mathbb{N}(\boldsymbol{Z\theta} + \boldsymbol{T\phi}, \boldsymbol{K}\sigma_g^2 + \boldsymbol{I}\sigma_e^2)$, where columns of the matrix $\boldsymbol{T}$ contain allele counts of the highly-associated SNPs (i.e., $\boldsymbol{T}$ is a submatrix of $\boldsymbol{S}$), and the vector $\boldsymbol{\phi}$ represents their effect sizes. In contrast to standard (non-SNP) covariates, the variance explained by $\boldsymbol{T}$ counts towards SNP heritability: $h^2_{\text{SNP}} = (\sigma_g^2 + \sigma_T^2)/(\sigma_g^2 + \sigma_T^2 + \sigma_e^2)$, where $\sigma_T^2 = (\boldsymbol{T\phi})^T(\boldsymbol{T\phi})$. Supplementary Figures 23 & 24

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

provides further details. In particular, we appreciate that our definition of highly-associated is somewhat arbitrary, so we confirm that estimates of $h^2_{\text{SNP}}$ are almost unchanged if instead we use $P < 5 \times 10^{-8}$.

**Datasets and phenotypes:** When searching for GWAS datasets, we preferred those with sample size at least 4 000 to ensure reasonable precision of $h^2_{\text{SNP}}$.[51] In total, our datasets were constructed from 40 independent cohorts, all of which have been previously described (see Supplementary Tables 8 & 9 for references and details of how cohorts were merged to form datasets). For the UCLEB data, there were in total 28 quantitative traits with measurements recorded for at 7 000 individuals. For each of these, we quantile normalized, then applied a test for inflation due to genotyping errors (Supplementary Fig. 14). Specifically, our test, inspired by Bhatia *et al.*[52] and valid for quantitative phenotypes where individuals are recruited from multiple cohorts, first estimates $h^2_{\text{SNP}}$ using only pairs of individuals in different cohorts, then using only pairs of individuals in the same cohort; a significant difference between the two estimates indicates possible inflation due to genotyping errors. We excluded five traits that showed evidence of inflation ($P < 0.05/28$), leaving us with 23: height, weight, body mass index, waist circumference, forced vital capacity, one second forced vital capacity, systolic blood pressure (adjusted), diastolic blood pressure (adjusted), PR Interval, QT Interval, Corrected QT Interval, QRS Voltage Product, Sokolow Lyon, glucose, insulin, total cholesterol (adjusted), LDL cholesterol (adjusted), triglyceride (adjusted), viscosity, fibrinogen, Interleukin 6, C-Reactive Protein and haemoglobin. Approximately 40% of individuals were receiving medication to reduce blood pressure, 25% to reduce lipid levels, so where indicated, phenotypes had been adjusted for this: for individuals on medication, their raw measurements had been increased either by adding on (blood pressure) or scaling by (lipid levels) a constant.[53,54] We note that some pairs of traits are highly correlated. However, as the overall correlation is not that extreme (we estimate the effective number of independent traits to be about 15), and most of our UCLEB analyses serve to support conclusions drawn from the GWAS traits, we decide to retain all 23 traits (rather than, say, consider only a subset). See Supplementary Note 3 for further details on phenotyping.

**Quality control:** We processed each of the 40 cohorts in identical fashion; see Supplementary Note 4 for full details. In summary, after excluding apparent population outliers, samples with extreme missingness or heterozygosity, and SNPs with MAF $< 0.01$, call-rate $< 0.95$ or $P < 10^{-6}$ from a test for Hardy-Weinberg Equilibrium, we imputed using IMPUTE2[13] and the 1000 Genome Phase 3 (2014) Reference Panel.[14] When merging cohorts to construct the GWAS datasets, we retained only autosomal SNPs which in all cohorts have MAF $> 0.01$ and $r_j > 0.99$ (using IMPUTE2 `r2_type2` in place of $r_j$ for directly genotyped SNPs). For the 8 UCLEB cohorts, we applied these filters only after merging. We only relax quality control for the analyses of the UCLEB data where we explicitly examine the consequences of including lower-quality and rare SNPs. When possible, the matrix $\boldsymbol{S}$ contains expected allele counts (dosages); i.e., $S_{i,j} = p_{i,j,1} + 2 \times p_{i,j,2}$, where $p_{i,j,1}$ and $p_{i,j,2}$ denote the probabilities of allele counts 1 and 2, respectively. If hard genotypes are required, for example when using LDSC to compute LD Scores,[8] we round $S_{i,j}$ to the nearest integer. As this was only necessary when considering high-quality SNPs ($r_j > 0.99$), we expect this rounding to have negligible impact on results. For each trait, Table 1 reports $m$, the total number of SNPs after imputation, and $\sum_{j=1}^{m} w_j$, the sum of SNP weights, which can be interpreted as an effective number of independent SNPs. For the GWAS datasets, $\sum w_j$ ranges from 79 K to 125 K. By contrast, when restricted to only high-quality SNPs, the UCLEB data has $\sum w_j = 39$ K, reflecting that the Metabochip directly captures a much smaller amount of genetic variation than standard genome-wide SNP arrays.

When analyzing quantitative traits, genotyping errors will tend only to be a concern when there are systematic differences between phenotypes across cohorts, and this is something we are able to explicitly test (Supplementary Fig. 14). However, for disease traits, when cases and controls have been genotyped separately (as is the design of most of our GWAS datasets), any errors will almost certainly correlate with phenotype and therefore cause inflation of $h^2_{\text{SNP}}$.[7,25] To test the effectiveness of our quality control for the GWAS traits, we construct a pseudo case-control study using two control cohorts; we confirm that the resulting estimate of $h^2_{\text{SNP}}$ is not significantly greater than zero, suggesting that the quality control steps we use for the GWAS datasets are sufficiently strict (Supplementary Note 5).

Accurate estimation of $h^2_{\text{SNP}}$ requires samples of unrelated individuals with similar ancestry. Prior to imputation, we removed ethnic outliers identified through principal component analyses (Supplementary Fig. 25). Post imputation, we computed (unweighted) allelic correlations using a pruned set of SNPs, then filtered individuals so that no pair remained with correlation greater than $c$, where $-c$ is the smallest observed pairwise correlation ($c$ ranges from 0.029 to 0.038, depending on dataset). For all analyses, we include

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a minimum of 30 covariates: the top 20 eigen-vectors from the allelic correlation matrix just described, and projections onto the top 10 principal components computed from 1000 Genomes samples.[14] For the 19 GWAS traits, we also include sex as a covariate, while for intraocular pressure and wide range achievement test scores, we additionally include age. Supplementary Figure 26 reports the proportion of phenotypic variance explained by each covariate. To check our filtering and covariate choices, we estimate the inflation of $h^2_{\text{SNP}}$ due to population structure and residual relatedness[3] (Supplementary Fig. 21). For the GWAS traits, we estimate that on average $h^2_{\text{SNP}}$ estimates are inflated by at most 3.1%, with the highest observed for ischaemic stroke (7.1%). For the 23 UCLEB traits, the average inflation is 0.3% (highest 2.3%).

**Single-SNP analysis:** Supplementary Figure 27 provides Manhattan Plots from logistic (case-control traits) and linear regression (quantitative traits), performed using PLINK v.1.9. These analyses provide the summary statistics required by LDSC. For the GWAS traits, we identified highly-associated SNPs ($P < 10^{-20}$) within the MHC for 6 of the GWAS traits (rheumatoid arthritis, type 1 diabetes, psoriasis, ulcerative colitis, celiac disease and multiple sclerosis), while rs2476601, a SNP within *PTPN22*, is highly associated with both rheumatoid arthritis and type 1 diabetes.[55,56] For the UCLEB traits, we find highly associated SNPs within *SCN10A* (PR Interval), *APOE* (total cholesterol, LDL cholesterol and C-reactive protein) and *ZPR1* (triglyceride levels). For heritability analysis, these SNPs were pruned, then included as additional fixed-effect covariates as described above.

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

3. Speed, D. *et al.* Describing the genetic architecture of epilepsy through heritability analysis. *Brain* **137**, 26802689 (2014).

4. Yang, J. *et al.* Genomic partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).

5. Gusev, A. *et al.* Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).

6. Lee, S., Yang, J., Goddard, M., Visscher, P. & Wray, N. Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).

7. Speed, D., Hemani, G., Johnson, M. & Balding, D. Improved heritability estimation from genome-wide SNP data. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).

8. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2014).

9. Lee, S. *et al.* Estimation of SNP-heritability from dense genotype data. *Am. J. Hum. Genet.* **93**, 1151–1155 (2013).

10. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).

11. Shah, T. *et al.* Population genomics of cardiometabolic traits: Design of the University College London-London School of Hygiene and Tropical Medicine-Edinburgh-Bristol (UCLEB) Consortium. *PLoS One* **8**, e71345 (2013).

12. Voight, B. *et al.* The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8**, e1002793 (2012).

13. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1**, 457–470 (2011).

14. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

15. Chen, G. *et al.* Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum. Mol. Genet.* **23**, 4710–4720 (2014).

16. Ek, W. *et al.* Germline genetic contributions to risk for esophageal adenocarcinoma, Barretts Esophagus, and gastroesophageal reflux. *J. Nal. Cancer Inst.* **105**, 1711–1718 (2013).

17. Bevan, S. *et al.* Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. *Stroke* **43**, 3161–3167 (2012).

18. Keller, M. *et al.* Using genome-wide complex trait analysis to quantify 'missing heritability' in parkinson's disease. *Hum. Mol. Genet.* **21**, 4996–5009 (2012).

19. Yin, X. *et al.* Common variants explain a large fraction of the variability in the liability to psoriasis in a han chinese population. *BMC Genomics* **15** (2014).

20. Lee, S. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).

21. Stahl, E. *et al.* Bayesian inference of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).

22. Robinson, E. *et al.* The genetic architecture of pediatric cognitive abilities in the Philadelphia Neurodevelopmental Cohort. *Mol. Psychiatry* **20**, 454–458 (2015).

23. Mancuso, N. *et al.* The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* **48**, 3035 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

24. Dempster, E. & Lerner, I. Heritability of threshold characters. *Genetics* **35**, 212–236 (1950).

25. Lee, S., Wray, N., Goddard, M. & Visscher, P. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).

26. Lee, J. Y. S., Goddard, M. & Visscher, P. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

27. Bethesda (MD): National Library of Medicine (US), N. C. f. B. I. The ncbi handbook [internet] (2002).

28. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).

29. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).

30. Yang, J., Zaitlen, N., Goddard, M., Visscher, P. & Price, A. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).

31. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).

32. Deary, I. *et al.* Genetic contributions to stability and change in intelligence from childhood to old age. *Nature* **18**, 212–215 (2012).

33. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

34. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2001).

35. Habier, D., Fernando, R., Kizilkaya, K. & Garrick, D. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **186**, 186–197 (2011).

36. Moser, G. *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* **11**, e1004969 (2015).

37. The ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature* **489**, 57–74 (2012).

38. Kumar, S., Feldman, M., Rehkopf, D. & Tuljapurkar, S. Limitations of GCTA as a solution to the missing heritability problem. *PNAS* **113**, E61E70 (2015).

39. Molloy, A. *et al.* A common polymorphism in HIBCH influences methylmalonic acid concentrations in blood independently of cobalamin. *Am. J. Hum. Genet.* **5**, 869–882 (2016).

40. Hayes, B., Visscher, P. & Goddard, M. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* **91**, 47–60 (2009).

41. Habier, D., Fernando, R. & Dekkers, J. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397 (2007).

42. Speed, D. & Balding, D. Relatedness in the post-genomic era: is it still useful? *Nat. Rev. Genet.* **16**, 33–44 (2014).

43. Corbeil, R. & Searle, S. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* **18**, 31–38 (1976).

44. Hardy, G. Mendelian proportions in a mixed population. *Science* **28**, 49–50 (1908).

45. Weinberg, W. Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins fur Vaterländische Naturkunde in Württemberg* **64**, 368–382 (1908).

46. Lee, S. & van der Werf, J. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet. Sel. Evol.* **38**, 25–43 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted September 9, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

47. World Health Organization. Global tuberculosis report (2014).

48. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLoS Genet.* **9**, e1003993 (2013).

49. Speed, D. & Balding, D. MultiBLUP: improved SNP-based prediction for complex traits. *Gen. Res.* **24**, 1550–1557 (2014).

50. Zhou, X., Carbonetto, P. & Stephens, M. Polygeneic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).

51. Visscher, P. *et al.* Statistical power to detect genetic (co)variance of complex traits using snp data in unrelated samples. *PLoS Genet.* **10**, e1004269 (2014).

52. Bhatia, G. *et al.* Haplotypes of common SNPs can explain missing heritability of complex diseases (2016). Preprint available on BioR$\chi$iv.

53. Tobin, M., Sheehan, N., Scurrah, K. & Burton, P. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Stat. Med.* **24**, 2911–2935 (2005).

54. Asselbergs, F. *et al.* Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am. J. Hum. Genet.* **91**, 8230838 (2012).

55. Todd, J. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **39**, 857–864 (2007).

56. Plenge, R. *et al.* TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study. *N. Engl. J. Med.* **20**, 1199–1209 (2007).