bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom.*
[3] *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, 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.*

SNP heritability, the proportion of phenotypic variance explained by SNPs, has been reported for many hundreds of traits. Its estimation requires strong prior assumptions about the distribution of heritability across the genome, but the assumptions in current use have not been thoroughly tested. By analyzing imputed data for a large number of human traits, we empirically derive a model that more accurately describes how heritability varies with minor allele frequency, linkage disequilibrium and genotype certainty. Across 19 traits, our improved model leads to estimates of common SNP heritability on average 43% (SD 3) higher than those obtained from the widely-used software GCTA, and 25% (SD 2) higher than those from the recently-proposed extension GCTA-LDMS. Previously, DNaseI hypersensitivity sites were reported to explain 79% of SNP heritability; using our improved heritability model their estimated contribution is only 24%.

The SNP heritability (h^2_{SNP}) of a trait is the fraction of phenotypic variance explained by additive contributions from SNPs.[1] Accurate estimates of h^2_{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] Whereas techniques for estimating (total) heritability have existed for decades,[4,5] the first method for estimating h^2_{SNP} was proposed only in 2010,[1] but has since 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.[6–8]

As the number of SNPs in a GWAS is usually much larger than the number of individuals, estimation of h^2_{SNP} requires steps to avoid over-fitting. Most reported estimates of h^2_{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.[1,9] 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 the genetic architecture of complex traits.

Results

When estimating h^2_{SNP}, the "LDAK Model" assumes

$$\mathbb{E}[h_j^2] \propto [f_j(1-f_j)]^{1+\alpha} \times w_j \times r_j, \tag{1}$$

where $\mathbb{E}[h_j^2]$ is the expected heritability contribution of SNP j and f_j is its (observed) MAF. The parameter α determines the assumed relationship between heritability and MAF. In human genetics it is commonly assumed that heritability does not depend on MAF, which is achieved by setting $\alpha = -1$, however, we consider alternative relationships. The SNP weights $w_1, \ldots, w_m$ are computed based on local levels of LD;[9] w_j tends to be higher for SNPs in regions of low LD, and thus the LDAK Model assumes that these SNPs contribute

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

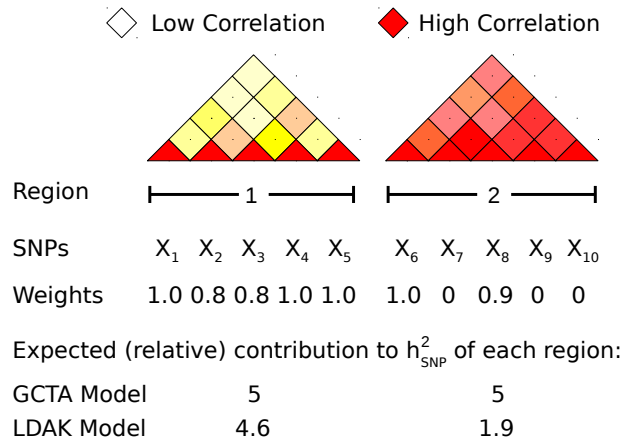

Figure 1: Comparison of the GCTA and LDAK Models. Region 1 contains five SNPs in low LD (lighter colors indicate weaker pairwise correlations). Each SNP contributes unique genetic variation, reflected by SNP weights close to one. Region 2 contains five SNPs in high LD (strong correlations). The total 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.

more than those in high-LD regions. Finally, $r_j \in [0, 1]$ is an information score measuring genotype certainty; the LDAK Model expects that higher-quality SNPs contribute more than lower-quality ones. r_j is defined in Online Methods, where we also explain how (1) arises by assuming a genome-wide random regression in which SNP effect sizes are assigned Gaussian distributions.

The "GCTA Model" is obtained from (1) by setting $w_j = 1$ and $r_j = 1$, and thus assumes that expected heritability does not vary with either LD or genotype certainty. To date, most reported estimates of h^2_{SNP} have used the GCTA Model with $\alpha = -1$, which corresponds to the assumption that $\mathbb{E}[h^2_j]$ is constant, and so the expected contribution of a SNP set depends only on the number of SNPs it contains.[1] 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 1 for a more detailed example.

An alternative method for estimating h^2_{SNP} is LDSC (LD Score Regression).[10] The LDSC Model expects that each SNP contributes equal heritability,[10,11] and therefore closely resembles the GCTA Model with $\alpha = -1$. When applied to the same dataset, estimates from LDSC will typically have standard error 25-100% higher than those from GCTA;[11] this is partly because the LDSC Model includes an extra parameter, designed to capture confounding biases, and partly because LDSC estimates are moment-based, whereas GCTA (like LDAK) uses restricted maximum likelihood (REML).[12,13] However, as LDSC requires only summary statistics (i.e., p-values from single-SNP analysis), it can be used on much larger datasets than GCTA and LDAK, which need raw genotype data, and can be applied to results from large-scale meta-analyses.[10]

SNP partitioning: (1) can be generalized by dividing SNPs into tranches across which the constant of proportionality is allowed to vary (so $\mathbb{E}[h^2_j] = c_k \times [f_j(1 - f_j)]^{1+\alpha} \times w_j \times r_j$ for SNPs in Tranche k). This is known as SNP partitioning.[6] Two examples are GCTA-MS[14] and GCTA-LDMS:[15] 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,[10] then divides each of these into five based on MAF, resulting in a total of 20 tranches. In general, we prefer to avoid SNP partitioning when estimating h^2_{SNP}, because it introduces (often arbitrary) discontinuities in the model assumptions and can cause

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Collection	Trait (Disease Prevalence, %)	n	m	$\sum_{j=1}^{m} w_j$	h^2_{GWAS}	Estimates of h^2_{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)[7]	0.35 (0.03)
	Coronary Artery Disease (6)	1 907 + 2 918	3 738 K	80 K	0.03	0.25 (0.06)[7]	0.40 (0.06)
	Crohn's Disease (0.5)	1 691 + 2 905	3 723 K	79 K	0.21	0.26 (0.01)[16]	0.32 (0.03)
	Hypertension (5)	1 918 + 2 916	3 739 K	80 K	<0.01	0.33 (0.06)[7]	0.46 (0.06)
	Rheumatoid Arthritis (0.5)	1 846 + 2 918	3 735 K	80 K	0.19	0.09 (0.03)[7]	0.21 (0.03)
	Type 1 Diabetes (0.5)	1 941 + 2 907	3 731 K	80 K	0.27	0.13 (0.03)[7]	0.31 (0.02)
	Type 2 Diabetes (8)	1 896 + 2 917	3 735 K	80 K	0.08	0.42 (0.07)[7]	0.54 (0.07)
Welcome Trust Case Control Consortium 2 (WTCCC 2)	Barrett's Oesophagus (1.6)	1 861 + 5 138	4 830 K	116 K	<0.01	0.25 (0.05)[17]	0.32 (0.04)
	Ischaemic Stroke (2)	3 769 + 5 139	4 797 K	115 K	<0.01	0.25 (0.03)[18]	0.34 (0.03)
	Parkinson's Disease (0.2)	1 687 + 5 136	4 819 K	116 K	0.03	0.27 (0.05)[19]	0.20 (0.03)
	Psoriasis (0.5)	2 267 + 5 143	4 814 K	116 K	0.21	0.35 (0.06)[20]	0.34 (0.02)
	Schizophrenia (1)	2 068 + 2 615	3 481 K	111 K	0.07	0.23 (0.01)[21]	0.30 (0.04)
	Ulcerative Colitis (0.2)	2 614 + 5 327	4 061 K	115 K	0.12	0.19 (0.01)[16]	0.28 (0.02)
WTCCC 2 +	Celiac Disease (1)	2 492 + 7 376	3 681 K	88 K	0.29	0.33 (0.04)[22]	0.35 (0.02)
	Multiple Sclerosis (0.1)	8 553 + 5 667	4 702 K	113 K	0.17	0.17 (0.01)[7]	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)[23]	0.21 (0.09)

Table 1: Properties of datasets and estimates of h^2_{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_{SNP} from common SNPs, computed using LDAK with $\alpha = -0.25$ (see main text for explanation of α). For comparison, we include previously published estimates of h^2_{SNP} (note that the previous analyses for rheumatoid arthritis, type 1 diabetes and multiple sclerosis excluded major histocompatibility SNPs, which we estimate contribute 0.07, 0.20 and 0.05, respectively), as well as h^2_{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_{SNP} and h^2_{GWAS} have been converted to the liability scale assuming the stated prevalence.

convergence problems. However, we show below that partitioning based on MAF enables reliable estimation of h^2_{SNP} when rare SNPs (MAF < 0.01) are included. Additionally, SNP partitioning provides a way to visually assess the fit of different heritability models; it allows us to estimate average h^2_j for different SNP tranches, which can then be compared to the values predicted under different assumptions.

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,[24] which comprise about 14 000 individuals genotyped using the Metabochip;[25] (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[26] and the 1000 Genome Phase 3 (2014) reference panel,[27] 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.

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

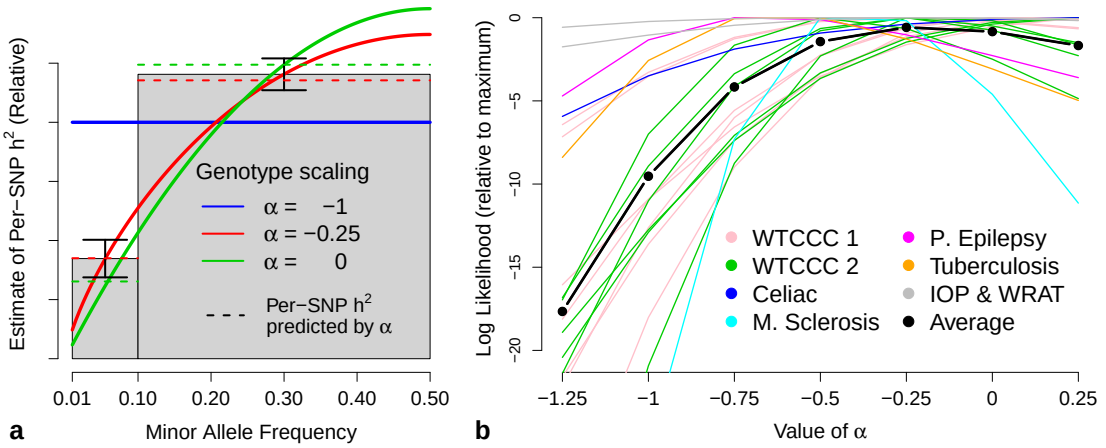

Figure 2: (a) Relationship between heritability and MAF. The parameter α specifies the assumed relationship between heritability and MAF: in human genetics, $\alpha = -1$ is typically used (solid blue line), while in animal and plant genetics, $\alpha = 0$ is more common (green); we instead found $\alpha = -0.25$ (red) provides a better fit to real data. The gray bars report (relative) estimates of the per-SNP heritability for MAF<0.1 and MAF>0.1 SNPs, averaged across the 19 GWAS traits (vertical lines provide 95% confidence intervals); the dashed lines indicate the per-SNP heritability predicted by each α. **(b) Determining best-fitting α for the GWAS traits.** We compare α based on likelihood; higher likelihood indicates better-fitting α. Lines report log likelihoods from LDAK for seven values of α, relative to the highest observed. Line colors indicate the seven trait categories, while the black line reports averages.

Further details of our methods and datasets are provided in Online Methods. In particular, we explain how when estimating h^2_{SNP} we give special consideration to highly-associated SNPs, which we define as those with $P < 10^{-20}$ from single-SNP analysis, and 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).

Relationship between heritability and MAF: Varying the value of α in (1) changes the assumed relationship between heritability and MAF; three example relationships are shown in Figure 2a. To determine suitable α, we analyze each of the 42 traits using seven values: $-1.25, -1, -0.75, -0.5, -0.25$, 0 and 0.25, seeing which lead to best model fit (highest likelihood). Full results are provided in Supplementary Figure 1 and Supplementary Table 2. First, to remove any confounding due to LD, we use 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. We find that model fit is typically highest for $-0.5 \leq \alpha \leq 0$, whereas the most widely-used value, $\alpha = -1$, reuslts in sub-optimal fit. On the basis that it performs consistently well across different traits and SNP filterings, we recommend that $\alpha = -0.25$ becomes the default. This value implies that expected heritability declines with MAF; this is 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 2 for further details).

While $\alpha = -0.25$ provides the best fit overall, for individual traits, optimal α may differ, and therefore we investigate sensitivity of h^2_{SNP} estimates to the value of α. Full results are provided in Supplementary Figures 3, 4 & 5, while Figure 6a provides a summary for the UCLEB traits. When analyzing only common SNPs, we find that changes in α have little impact on h^2_{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 α 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_{SNP} and also allows estimation of the relative contributions of rare and common variants (Figure 6a and Supplementary Figure 6).

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

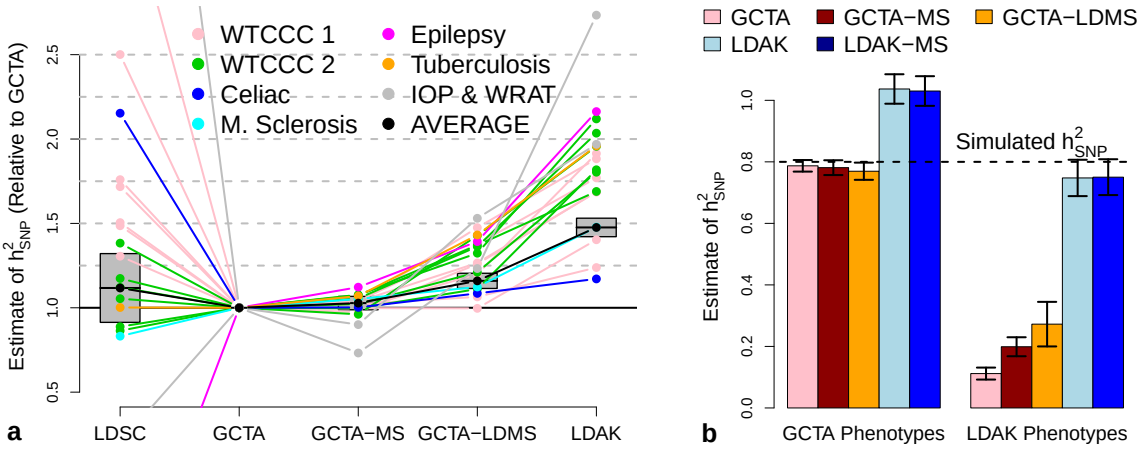

Figure 3: (a) Relative estimates of h^2_{SNP} for the GWAS traits. h^2_{SNP} estimates from LDSC, GCTA-MS (SNPs partitioned by MAF), GCTA-LDMS (SNPs partitioned by LD and MAF) and LDAK are reported relative to those from GCTA. For versions of GCTA and LDAK, we use $\alpha = -0.25$ (see main text for explanation of α). 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. Numerical values are provided in Supplementary Table 3. **(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_{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.

Relationship between heritability and LD: The LDAK Model assumes that heritability varies according to local levels of LD, whereas the GCTA Model assumes that heritability is independent of LD. First we demonstrate that choice of model matters when estimating h^2_{SNP}. For the GWAS traits, Figure 3a reports relative estimates of h^2_{SNP} from GCTA, GCTA-MS, GCTA-LDMS and LDAK (all using $\alpha = -0.25$); see Supplementary Figure 7 for an extended version. We find that estimates based on the LDAK Model are on average 48% (SD 3) higher than estimates based on the GCTA Model. For the UCLEB traits, estimates from LDAK are on average 88% (SD 7) higher than those from GCTA (Supplementary Fig. 8). Figure 3a also includes results from LDSC, run as described in the original publication[10] (see Supplementary Table 3 for numerical values). Estimates from LDSC are not significantly different to those from GCTA, which is to be expected considering that GCTA and LDSC assume the same relationship between heritability and LD. In Supplementary Figure 9 we consider alternative versions of LDSC (e.g., varying how LD Scores are computed, forcing the intercept term to be zero and excluding highly-associated SNPs). While changing settings can have a large impact, in all cases the average estimate of h^2_{SNP} from LDSC remains substantially below that from LDAK.

A recent article which asserted that GCTA estimates h^2_{SNP} more accurately than LDAK, based this claim on a simulation study in which causal SNPs were assigned effect sizes from the same Gaussian distribution, irrespective of LD.[6] This resembles the GCTA Model but not the LDAK Model, and so it is no surprise that GCTA performed better. Figure 3b shows that if instead effect size variances had been scaled by SNP weights, and so vary with LD similar to the LDAK Model, then the study would have found LDAK to be superior to GCTA. Thus using simulations to compare different heritability models is problematic, because the conclusions will depend on the assumptions used when generating phenotypes. See Supplementary Figure 10 for a full reanalysis of the reported simulation study and Supplementary Figure 11 for further simulations.

Rather than using simulations, we compare LDAK and GCTA empirically. Supplementary Table 4 shows that when $\alpha = -0.25$, assuming the LDAK Model leads to higher likelihood than assuming the GCTA Model for all 19 GWAS traits and for 17 of the 23 UCLEB traits (if we instead use $\alpha = -1$, likelihood is higher under the LDAK Model for 31 of the 42 traits). To visually demonstrate the superior fit of the LDAK Model, we partition SNPs into low- and high-LD (for this, we rank SNPs according to the average LD Score[10] of non-overlapping 100 kb segments, the metric used by GCTA-LDMS[15]). First, we partition so that the two tranches contain an

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

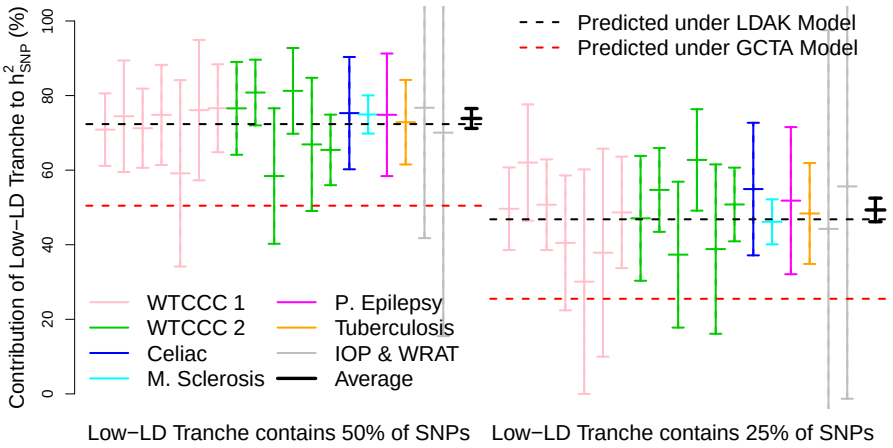

Figure 4: Comparing the GCTA and LDAK Models for the GWAS traits: We partition SNPs into low- or high-LD, with the low-LD tranche containing either 50% (left) or 25% (right) of SNPs. For each partition, the horizontal red and black lines indicate the predicted contribution of the low-LD tranche to h^2_{SNP} under the GCTA and LDAK Models, respectively. Vertical lines provide point estimates and 95% confidence intervals for the contribution of the low-LD tranche to h^2_{SNP}, estimated assuming the GCTA Model. Line colors indicate the seven trait categories, while the black lines provide the (inverse variance weighted) averages.

equal number of SNPs. The left half of Figure 4b reports, for each of the GWAS traits, the contribution of the low-LD tranche, estimated using the GCTA Model (with $\alpha = -0.25$). Under the GCTA Model, the low-LD tranche is expected to contribute 50% of h^2_{SNP}; under the LDAK Model, it is expected to contribute 72% of h^2_{SNP}. We see that the estimated contribution of the low-LD tranche is consistent with the GCTA Model (95% confidence interval includes 50%) for only 5 of the 19 traits, whereas it is consistent with the LDAK Model (confidence interval includes 72%) for 18. Next we partition so that the low-LD tranche contains a quarter of the SNPs; now the low-LD tranche is predicted to contribute 26% of h^2_{SNP} under the GCTA Model, but 47% of h^2_{SNP} under the LDAK Model. The right half of Figure 4b shows that its estimated contribution is consistent with the GCTA Model for only 7 of the 19 traits, but again consistent with the LDAK Model for 18. Additional results are provided in Supplementary Figure 12; these show that regardless of whether we estimate heritabilities using LDAK (rather than GCTA), whether we use $\alpha = -1$ (instead of $\alpha = -0.25$) or whether we analyze the UCLEB traits, it remains the case that the LDAK Model better predicts the heritability contribution of each tranche than the GCTA Model.

Relationship between heritability and genotype certainty: The LDAK Model assumes that SNP heritability contributions vary with genotype certainty (measured by the information score r_j). So far, our analyses have used only very high-quality SNPs ($r_j > 0.99$), so this assumption has been redundant. Now we also include lower-quality common SNPs; we focus on the UCLEB traits, as for these we were able to test for correlation between genotyping errors and phenotype (Supplementary Fig. 13). Supplementary Table 5 compares model fit with and without allowance for genotype certainty; it shows that including r_j in the heritability model tends to provide a modest improvement in model fit, resulting in a higher likelihood for 18 out of 23 traits.

Estimates of h^2_{SNP} for the GWAS traits: Table 1 presents our final estimates of h^2_{SNP} for the 19 GWAS traits, obtained using the LDAK Model (with $\alpha = -0.25$). For comparison, we include previously-reported estimates of h^2_{SNP}, as well as the proportion of phenotypic variance explained by SNPs reported as genome-wide significant (see Supplementary Table 6). For the disease traits, estimates are on the liability scale, obtained by scaling according to the observed case-control ratio and (assumed) trait prevalence.[28,29] We are unable to find previous estimates of h^2_{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 7. These show that our final estimates of h^2_{SNP} are on average 43% (SD 3) and 25% (SD 2) higher than, respectively, those obtained using the original versions (i.e., with $\alpha = -1$) of GCTA[30] and GCTA-LDMS.[15]

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

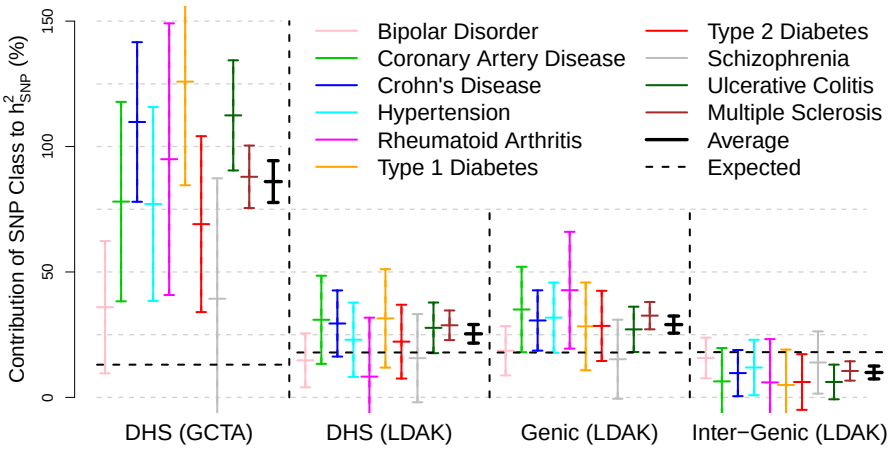

Figure 5: Enrichment of SNP Classes. Block 1 reports the contributions to h^2_{SNP} of DNaseI hypersensitivity sites (DHS), estimated under the GCTA Model with $\alpha = -1$ (see main text for explanation of α). The vertical lines provide point estimates and 95% confidence intervals for each trait, and 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.*[7] Block 2 repeats this analysis, but now assuming the LDAK Model with $\alpha = -0.25$. Blocks 3 & 4 estimate the contribution of "genic SNPs" (those inside or within 2 kb of an exon) and "inter-genic SNPs" (further than 125 kb from an exon), again assuming the LDAK Model with $\alpha = -0.25$. To assess enrichment, estimated contributions are compared to those expected under the GCTA or LDAK Model, as appropriate (horizontal lines).

Role of DNaseI hypersensitivity sites (DHS): Gusev *et al.*[7] 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_{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 on average DHS contribute 86% (SD 4) of h^2_{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_{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[31]), 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_{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-fold 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_{SNP}, and in contrast to Gusev *et al.*, enrichment remains constant when we reduce SNP density (Supplementary Fig. 14 & 15 and Supplementary Table 8).

Finucane *et al.*[32] performed a similar analysis, but considered 52 SNP classes and estimated enrichment using LDSC; across nine traits, they identified five classes with >4-fold enrichment, the highest of which, "conserved SNPs," had 13-fold enrichment. When we use LDAK to estimate enrichment for our 19 GWAS traits, the results are more modest; the highest enrichment is 2.5-fold, with only 1.3-fold enrichment for conserved SNPs (Supplementary Fig. 16).

Relaxing quality control: For the UCLEB data, we consider nine alternative SNP filterings. Supplementary Figure 17 reports estimates of h^2_{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_{SNP} to increase by 50-60% (Supplementary Fig. 18). This is similar to what we observe in practice, as across the 23 traits, estimates of h^2_{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_{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

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

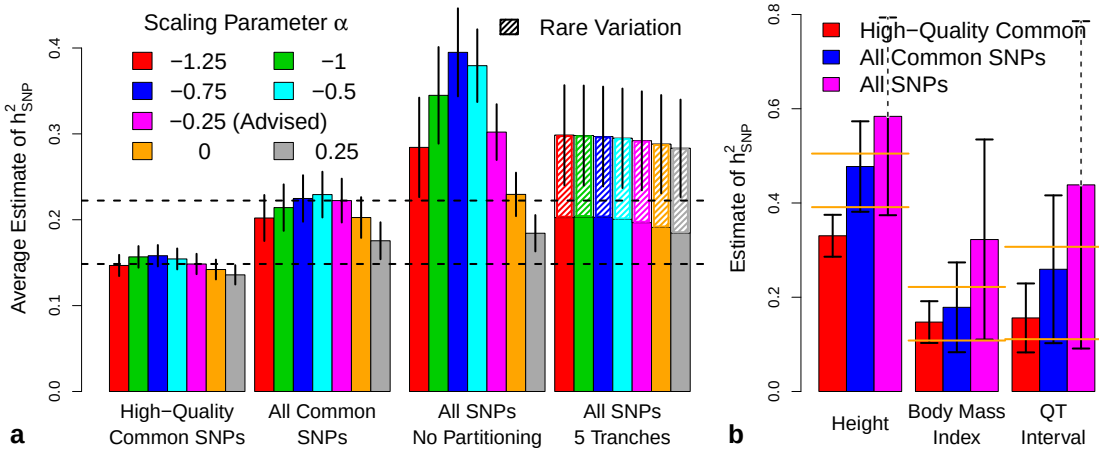

Figure 6: Varying quality control for the UCLEB traits. We consider three SNP filterings: 353 K high-quality common SNPs (information score > 0.99, MAF > 0.01), 8.8 M common SNPs (MAF > 0.01) and all 17.3 M SNPs (MAF > 0.0005). **(a)** Blocks indicate SNP filtering; bars report (inverse variance weighted) average estimates of h^2_{SNP} using LDAK (vertical lines provide 95% confidence intervals). Bar color indicates the value of α used. For Blocks 1, 2 & 3, h^2_{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_{SNP} for height, body mass index and QT interval, the three traits for which common SNP heritability has been previously estimated with reasonable precision[6] (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).

estimates of common SNP h^2_{SNP} are available[6] (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_{SNP}: for example, when we use the 17.3 M SNPs with MAF > 0.0005, estimates of h^2_{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 (median increase 22%), with rare SNPs contributing on average 33% (SD 5) of h^2_{SNP} (Figure 6a).

Discussion

With estimates of h^2_{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_{SNP}. Previous attempts to assess the validity of assumptions have used simulation studies,[14,15] but this approach will tend to favor assumptions similar to those used to generate the phenotypes. Instead, we have compared different heritability models empirically, by examining how well they fit real datasets.

We begun by investigating the relationship between heritability and MAF. Across 42 traits, we found that best fit was achieved by setting $\alpha = -0.25$ in (1), which implies that average heritability varies with $[\text{MAF}(1-\text{MAF})]^{0.75}$. As explained in Online Methods, the value of α corresponds to the scaling of genotypes. Therefore, our result indicates that the performance (i.e., detection power and/or prediction accuracy) of many penalized and Bayesian regression methods, for example, the Lasso, ridge regression and Bayes A,[33–35] could be improved simply by changing how genotypes are scaled. Although we recommend $\alpha = -0.25$ as the default value, with sufficient data available, it should be possible to estimate α on a trait-by-trait basis, or to investigate more complex relationships between heritability and MAF. 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 SNPs are included.

We also examined the relationship between heritability and LD. To date, most estimates of h^2_{SNP} have been based on the GCTA

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Model; this model can be motivated by a belief that each SNP is expected to have the same effect on the phenotype, from which it follows that the expected heritability of a region should depend on the number of SNPs it contains. By contrast, the LDAK Model views highly-correlated SNPs as tagging the same underlying variant, and therefore believes that the expected heritability of a region should vary according to the total amount of distinct genetic variation it contains. Across our traits, we found that the relationship between heritability and LD specified by the LDAK Model consistently 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 on a large number of closely-related methods. For example, LDSC,[10] like GCTA, assumes that heritability contributions are independent of LD and therefore it also tends to under-estimate h^2_{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.[7,32] Further afield, most software for mixed model association analyses (e.g., FAST-LMM, GEMMA and MLM-LOCO) use an extension of the GCTA Model,[36–38] and likewise most bivariate analyses, including those performed by LDSC.[8,39,40] It remains to be seen how much these methods would be affected if they employed more realistic heritability models.

Attempts have been made to improve the accuracy of heritability models via SNP partitioning.[14,15,41] We find that partitioning by MAF can be advantageous, as it guards against misspecification of the relationship between heritability and MAF when rare variants are included. Figure 3a and Supplementary Figure 7 indicate that the realism of the GCTA Model can be improved by partitioning based on LD; for example, across the GWAS traits, estimates from GCTA-LDMS are on average 16% (SD 2) higher than those from GCTA, and now only 23% (SD 2) lower than those from LDAK. The improvement arises because model misspecification is reduced by allowing SNPs in lower-LD tranches to have higher average heritability. However, Supplementary Table 9 illustrates why we consider such an approach sub-optimal; in particular, SNP partitioning can be computationally expensive, and even with LD-partitioning, model fit tends to be worse than that from LDAK.

While we have investigated the role of MAF, LD and genotype certainty, there remain other factors on which heritability could depend, in particular the available functional annotations of genomes.[42] 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 Table 10 shows that model fit is improved by assuming $\mathbb{E}[h^2_j] \propto [f_j(1-f_j)]^{1+\alpha} \times w_j \times r_j \times \exp(\frac{-(D_j+50)}{500})$, where D_j is the distance (in kb) between SNP j and the nearest exon (under this model, genic SNPs are expected to have about twice the heritability of inter-genic SNPs). In general, we believe that modifications of this type will have a relatively small impact; we note that across the 19 GWAS traits, scaling by $\exp(\frac{-(D_j+50)}{500})$ increases model log likelihood by on average only 1.5, much less than the average increase obtained by using $\alpha = -0.25$ instead of $\alpha = -1$ (8.9), or by choosing the LD-model specified by LDAK instead of GCTA (17.7), and does not significantly change estimates of h^2_{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_{SNP}, care should be taken to avoid possible sources of confounding. Previously, we advocated a test for inflation of h^2_{SNP} due to population structure and familial relatedness.[3] The conclusions of a recent paper claiming that h^2_{SNP} estimates are unreliable,[43] would have changed substantially had this test been applied (Supplementary Fig. 19). 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_{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_{SNP}. Although sample size prevented us from obtaining precise estimates of h^2_{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]

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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," while access to Children's Hospital of Philadelphia (CHOP) data was granted under Project 49228-1, "Assumptions underlying estimates of SNP Heritability." We thank Anne Molloy, James Mills and Lawrence Brody for permission to use genotype data from the Trinity College Dublin Student Study.[44] and Sarah Langley for help accessing the CHOP data. 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. NC is an ESPOD Fellow from the European Molecular Biology Laboratory, European Bioinformatics Institute, and Wellcome Trust Sanger Institute. 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.

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_{SNP} starting from raw genotype data. In general, we assume there are n individuals, recorded for p covariates and genotyped (either directly or via imputation) for m SNPs: 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 (expected) 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 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.[26] Additional details on our information score are provided in Supplementary Figure 20.

Estimating h^2_{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 weight[9] and information score for SNP j,

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

then the LDAK Model for estimating SNP heritability $h^2_{\text{SNP}} = \sigma_g^2/(\sigma_g^2+\sigma_e^2)$ is:

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

θ_k denotes the fixed-effect coefficient for the kth covariate, β_j and e_i are random-effects indicating the effect size of SNP j and the noise component for Individual i, while σ_g^2 and σ_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.[9] Model (2) is equivalent to assuming:[45,46]

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

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),[47] 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 $\theta_1, \ldots, \theta_p$, σ_g^2 and σ_e^2.[12]

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,[48,49] 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 $W_l = \sum_{j \in P_l} [2f_j(1-f_j)]^{1-\alpha}$. 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 σ_l^2/Σ represents the contribution to h^2_{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,[30,50] 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. 7). For diseases, we converted estimates of h^2_{SNP} to the liability scale based on the observed case-control ratio and assumed prevalence.[28,29] In general, we copied the prevalences used by previous studies; however for tuberculosis, where no previous estimate of h^2_{SNP} is available, we derived an estimate of prevalence from World Health Organization data[51] (Supplementary Note 2).

LDSC: Originally designed as a way to quantify confounding in a GWAS, LDSC[10] also provides a method for estimating h^2_{SNP}, which requires 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 $\mathbb{E}[\chi_j^2] = 1 + nh_j^2 + n\sum_{k \neq j} r_{j,k}^2 h_k^2 + na_j$, 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).[10] 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 $\mathbb{E}[\chi_j^2] = 1 + nl_j h^2_{\text{SNP}}/m +$

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

na, where $l_j = 1 + \sum_{k \neq j} r^2_{j,k}$ 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_{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_{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,[10] it is straightforward to accommodate alternative relationships between $\mathbb{E}[h^2_j]$ 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 similarity with the GCTA Model appears intrinsic to LDSC; while the assumption that heritability is independent of LD can be relaxed via SNP partitioning,[41] we can not envisage how the method could be modified to accommodate the LDAK SNP weights. For LDSC analyses, we used LDSC v.1.0.0 both for calculating LD Scores and estimating h^2_{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,[9] 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.[7,30,52,53] Another approach is to model effect sizes as a mixture of Gaussians,[35,54] 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 0.5, then to include those which remain as fixed-effect covariates. Thus in place of Equation (3), we assume $\boldsymbol{Y} \sim \mathbb{N}(\boldsymbol{Z\theta} + \boldsymbol{T\phi}, \boldsymbol{K}\sigma^2_g + \boldsymbol{I}\sigma^2_e)$, 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: $h^2_{\text{SNP}} = (\sigma^2_g + \sigma^2_T)/(\sigma^2_g + \sigma^2_T + \sigma^2_e)$, where $\sigma^2_T = (\boldsymbol{T\phi})^T(\boldsymbol{T\phi})$. Supplementary Figures 21 & 22 provides further details. In particular, we appreciate that our definition of highly-associated is somewhat arbitrary, so we confirm that estimates of h^2_{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_{SNP}.[55] In total, our datasets were constructed from 40 independent cohorts, all of which have been previously described (see Supplementary Tables 11 & 12 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. 13). Specifically, our test, inspired by Bhatia *et al.*[56] and valid for quantitative phenotypes where individuals are recruited from multiple cohorts, first estimates h^2_{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.[57,58] 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[26] and the 1000 Genome Phase 3 (2014) Reference Panel.[27] 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

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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,[10] 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; the aim of these weights is to remove duplication of signal due to LD and their sum can loosely 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. 13). 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_{SNP}.[9,29] 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_{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_{SNP} requires samples of unrelated individuals with similar ancestry. Prior to imputation, we removed ethnic outliers identified through principal component analyses (Supplementary Fig. 23). 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 our datasets, this filtering excluded relatively few individuals (on average 3.8%, with maximum 11.6%). For all analyses, we include a minimum of 30 covariates: the top 20 eigenvectors from the allelic correlation matrix just described, and projections onto the top 10 principal components computed from 1000 Genomes samples.[27] 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 24 reports the proportion of phenotypic variance explained by each covariate. To check our filtering and covariate choices, we estimate the inflation of h^2_{SNP} due to population structure and residual relatedness[3] (Supplementary Fig. 19). For the GWAS traits, we estimate that on average h^2_{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 25 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.[59,60] 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.

Computational requirements: The most time-consuming aspect of analysis was genotype imputation; for a typically-sized cohort ($\sim$3 000 individuals) this took approximately one CPU-year (i.e., a few days on a 100-node cluster). Next is computation of SNP weights, which for the GWAS traits ($\sim$4 M SNPs) took approximately one CPU-month (again, this can be near-perfectly parallelized). Finally, solving the mixed-model via REML would take between a few minutes for the smaller traits ($\sim$5 000 individuals) and a few hours for the largest ($\sim$14 000 individuals). Memory-wise, the most onerous task is solving the mixed-model, for which memory demands scale with n^2; however, even for the largest dataset, this was less than 5 Gb (when using multiple kinship matrices, LDAK allows for these to be read on-the-fly, so that the memory demands are no higher than when using only one).

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

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. Henderson, C., Kempthorne, O., Searle, S. & von Krosigk, C. The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**, 192–218 (1959).

5. Falconer, D. & Mackay, T. *Introduction to Quantitative Genetics (4th Edition)* (Longman, 1996).

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

7. 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).

8. 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).

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

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

11. Bulik-Sullivan, B. Relationship between LD Score and Haseman-Elston Regression (2015). Preprint available on BioRχiv.

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

13. Golan, D., Lander, E. & Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. *PNAS* **111**, E5272E5281 (2014).

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

15. 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).

16. 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).

17. 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).

18. 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).

19. 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).

20. 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).

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

24. 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).

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

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

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

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

29. 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).

30. 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).

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

32. Finucane, H. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

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

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

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

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

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

38. 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).

39. 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).

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

41. Gazal, S. *et al.* Linkage disequilibrium dependent architecture of human complex traits reveals action of negative selection (2016). Preprint available on BioRχiv.

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

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

44. 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).

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

bioRxiv preprint doi: https://doi.org/10.1101/074310; this version posted January 15, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

50. 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).

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

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

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

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

55. 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).

56. Bhatia, G. *et al.* Haplotypes of common SNPs can explain missing heritability of complex diseases (2016). Preprint available on BioRχiv.

57. 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).

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

59. 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).

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