
Supplementary information

**Estimation and mapping of the missing
heritability of human phenotypes**

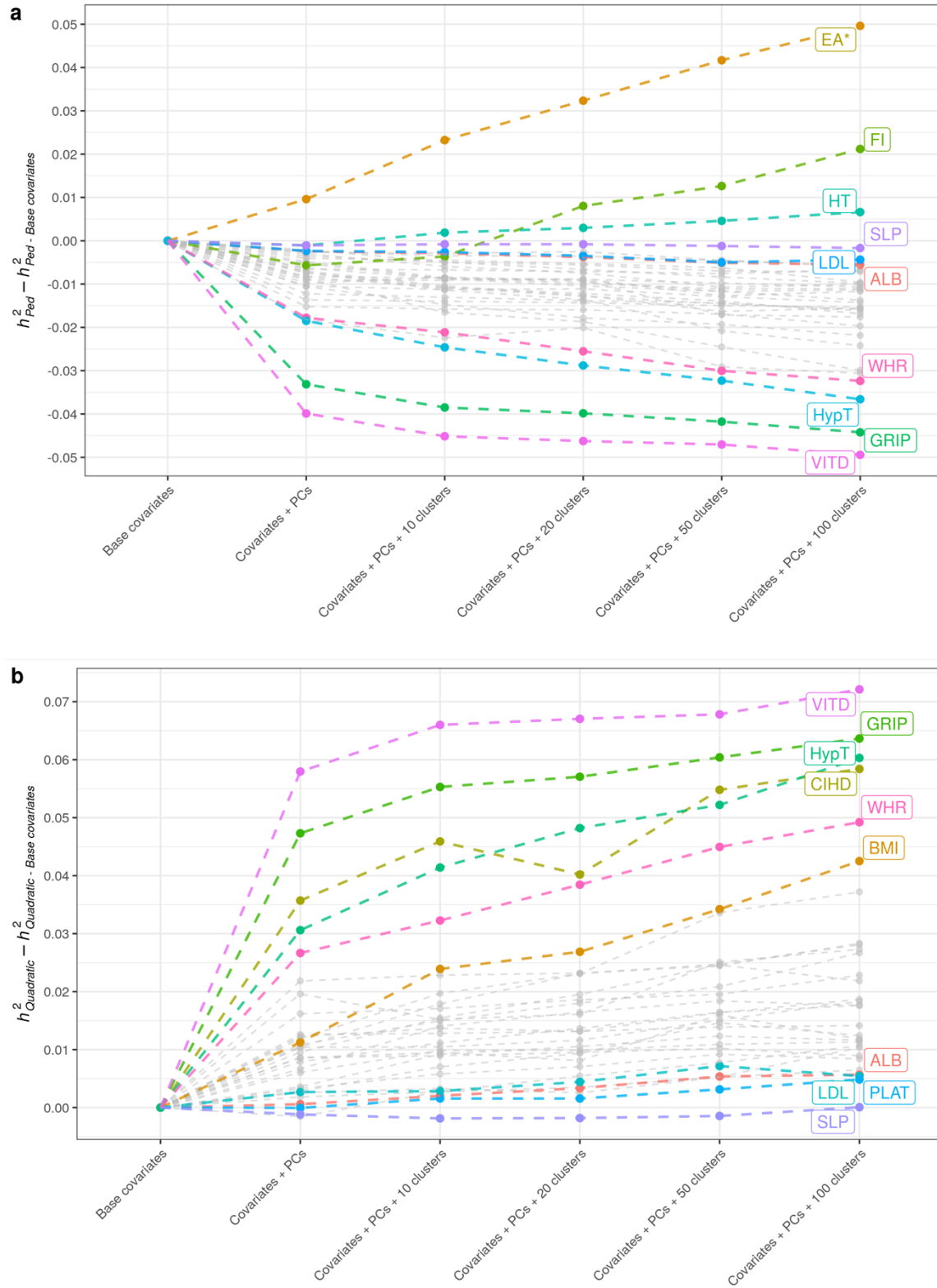
In the format provided by the
authors and unedited

SUPPLEMENTARY INFORMATION

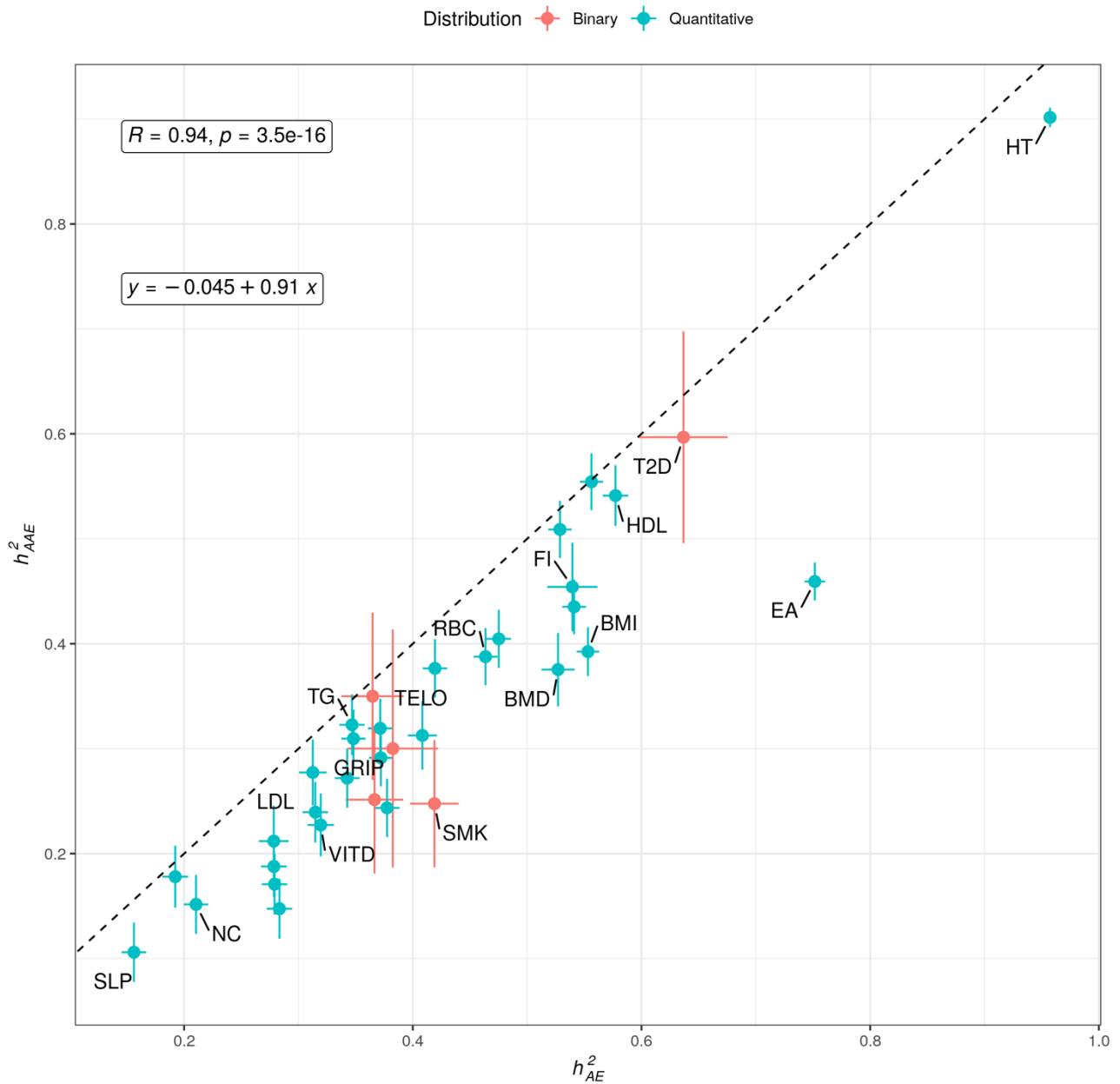
Table of Contents

SUPPLEMENTARY FIGURES	2
<i>Supplementary Figure 1: Effects of different covariates adjustments on REML-based pedigree estimates for 34 traits.</i>	2
<i>Supplementary Figure 2: Pedigree-based estimates of narrow sense heritability for 34 traits under different model of resemblance between relatives.</i>	3
<i>Supplementary Figure 3: Enrichment of trait heritability in coding and non-coding variants within genomic regions covered by whole-exome sequencing (WES) technologies.</i>	4
<i>Supplementary Figure 4: Heritability enrichment within genomic regions conserved across species.</i>	5
<i>Supplementary Figure 5: RVA annotations enrichments.</i>	6
<i>Supplementary Figure 6: Defining pathogenicity.</i>	7
<i>Supplementary Figure 7: Distance to nearest common variant association for trait-associated rare variants.</i>	8
<i>Supplementary Figure 8: Relationship between GWAS signal density and colocalization with structural variants.</i>	9
<i>Supplementary Figure 9: Distance of trait-associated variants to gene boundaries.</i>	10
<i>Supplementary Figure 10: Estimates from linkage disequilibrium score regression (LDSC) as a function of the minimum allele frequencies of SNPs included in the analysis.</i>	11
<i>Supplementary Figure 11: Analyses of ultra-rare variants (MAC between 1 and 69).</i>	12
<i>Supplementary Figure 12: Relationship between allele frequency and missing heritability.</i>	13
<i>Supplementary Figure 13: Quantification of SNP-based heritability captured outside of hg38 genome build.</i>	14
SUPPLEMENTARY NOTES	15
Supplementary Note 1: Estimates of genetic correlation from WGS data	15
Supplementary Note 2: Enrichment in functional annotations	16
2.1. Heritability enrichment in genomic loci regions covered by whole-exome sequencing	16
2.2. Heritability enrichment in genomic loci conserved across species	16
Supplementary Note 3: Replication in the Alliance for Genomic Discovery dataset	18
3.1. Overview of the data and quality control.....	18
3.2. AGD replication analysis	18
Supplementary Note 4: Gain of WGS over imputation for GWAS discovery	19
4.1. Background.....	19
4.2. Results	19
4.3. Discussion	19
Supplementary Note 5: Relationship between still missing heritability and negative selection	21
5.1. Definition of the α AUC statistic.....	21
5.2. Relationship between the α AUC statistic and other models of negative selection	21
5.3. Derivation of diagonal GRM for ultra-rare variants.....	22
Supplementary Note 6: SNP-based heritability captured outside of the hg38 genome build	24
6.1. Overview and main results	24
6.2. Methods	24
SUPPLEMENTARY REFERENCES	25

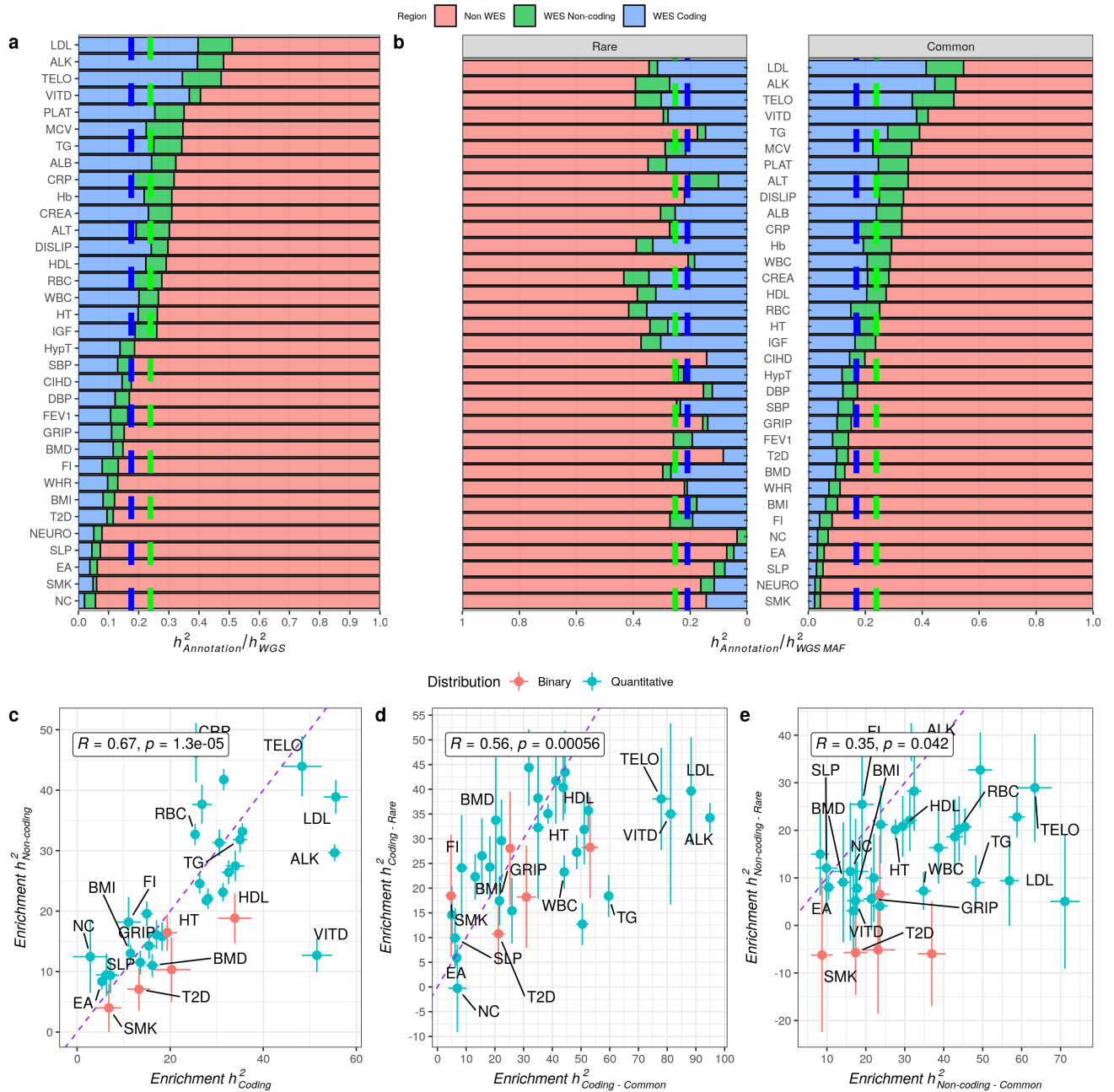
SUPPLEMENTARY FIGURES



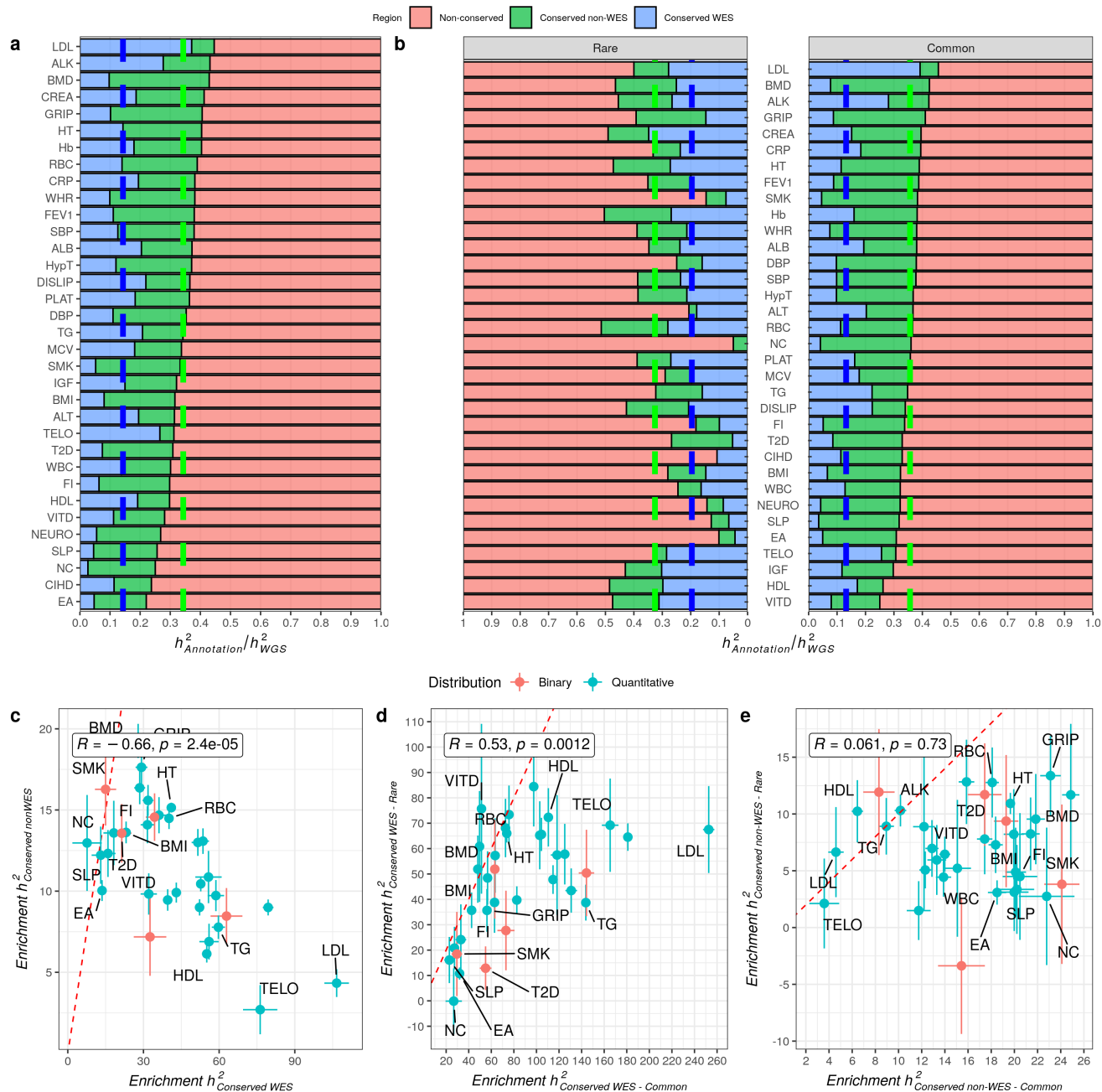
Supplementary Figure 1: Effects of different covariates adjustments on REML-based pedigree estimates for 34 traits. (a) Change for \hat{h}^2_{PED} estimates – Note that estimates for height (HT), fluid intelligence score (FI) and educational attainment (EA) were obtained using a different model (as compared to other traits) accounting for assortative mating (AM). (b) Change for the quadratic term, capturing non-additive genetic and shared environmental effects. Traits subject to AM are not shown in this plot.



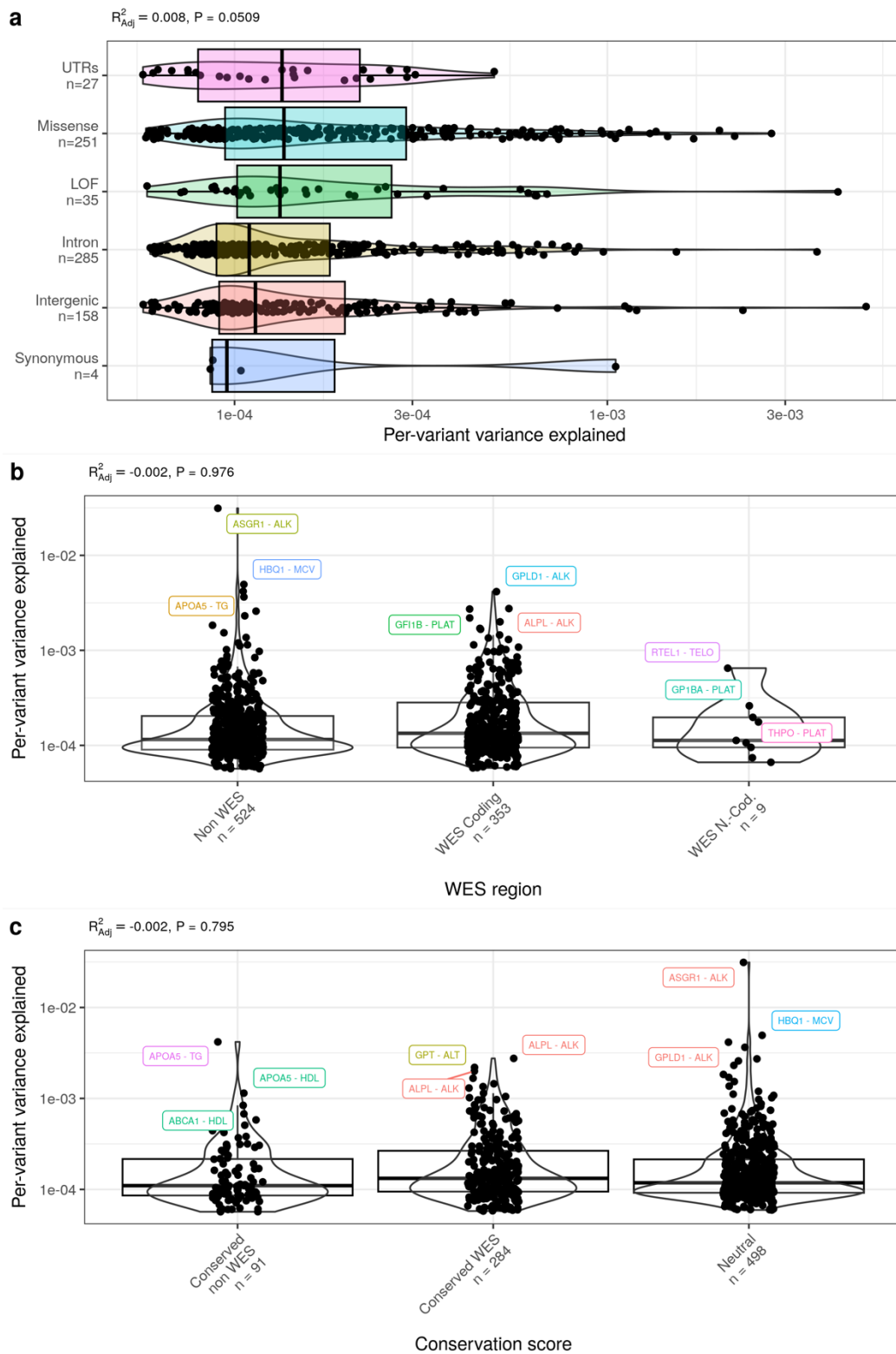
Supplementary Figure 2: Pedigree-based estimates of narrow sense heritability for 34 traits under different model of resemblance between relatives. The x-axis represents estimates under a model assuming all resemblance between relatives is due to additive genetic effects (AE model). For all traits but height (HT), educational attainment (EA) and fluid intelligence score (FI), the y-axis represents estimates under a model accounting for non-additive genetic effects (AAE model). This model assumes that phenotypic resemblance between relatives varies as a quadratic function of their estimated genetic relatedness. For HT, EA and FI, estimates were obtained using another model accounting for assortative mating described in the METHODS section. Error bars represent standard errors. The correlation between heritability estimates shown at the top-left corner of the Figure was calculated using a Pearson's correlation coefficient (R) over n=34 traits. The p-value measuring the significance of that correlation is denoted as p in the bottom-right corner of the panel and is based on a two-side Pearson's test.



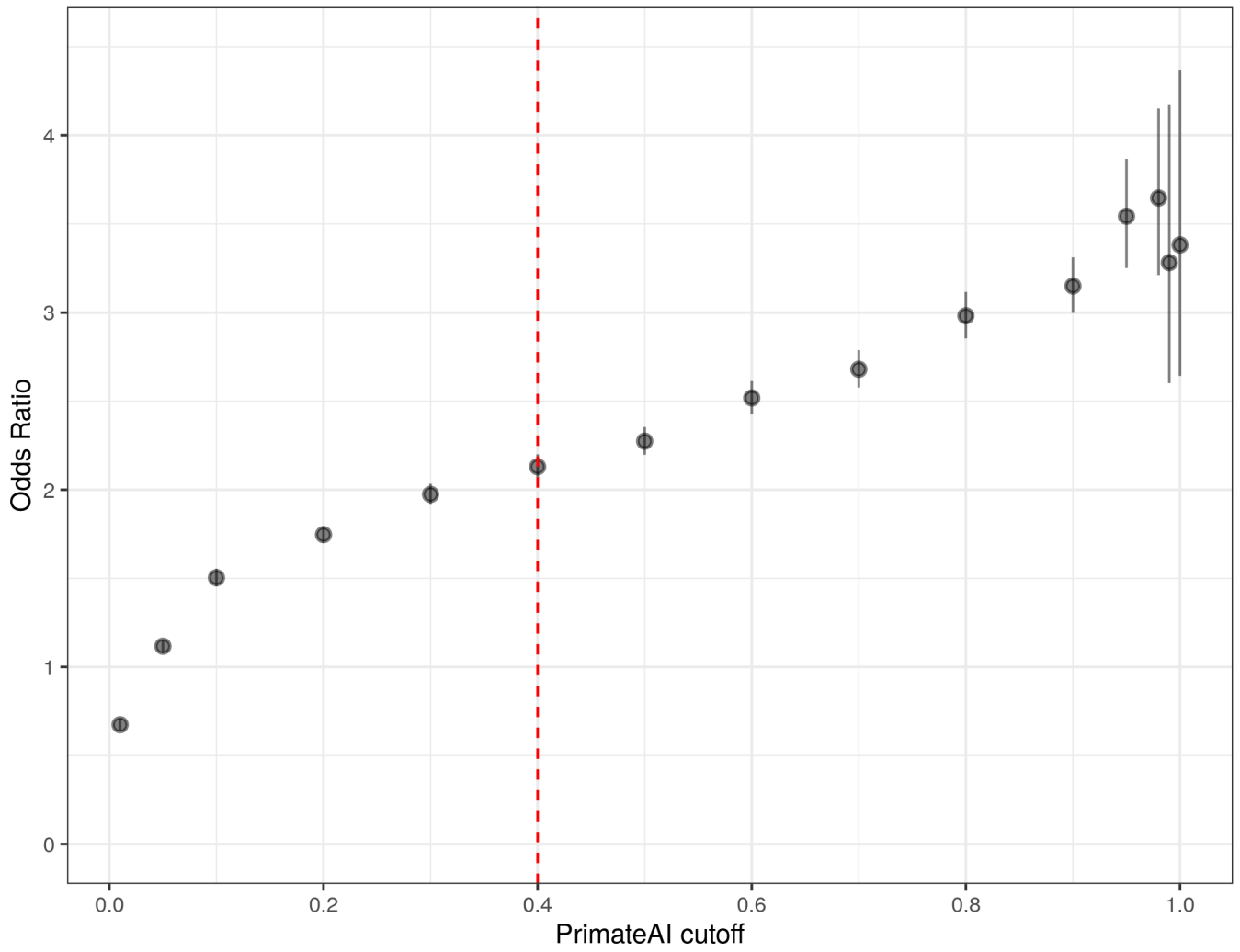
Supplementary Figure 3: Enrichment of trait heritability in coding and non-coding variants within genomic regions covered by whole-exome sequencing (WES) technologies. (a) Proportion of h^2_{WGS} explained by coding and non-coding variants in WES regions. Dotted lines represent the mean of each annotation. (b) Proportion of h^2_{WGS} explained by common and rare coding and non-coding WES variants. (c) Genome-wide enrichments in heritability for coding and non-coding WES variants. (d-e) Enrichments decomposed for common and rare variants. Error bars represent standard errors. Correlations between heritability enrichments shown in panels (c), (d) and (e) were calculated using a Pearson's correlation coefficient (R) over $n=34$ traits. P -values measuring the significance of these correlation were denoted as p in the top-left corner of the corresponding panel and were based on a two-side Pearson's test.



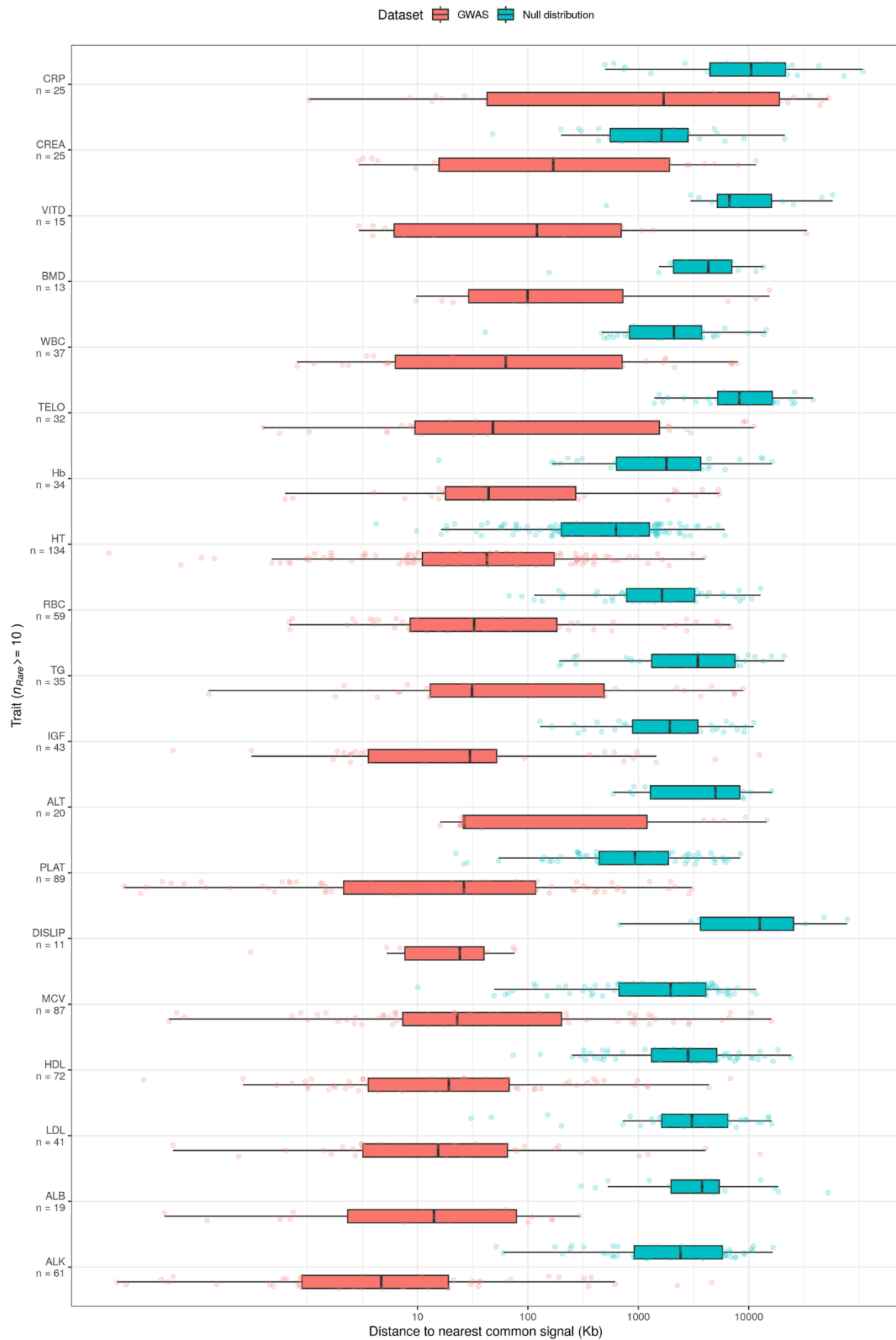
Supplementary Figure 4: Heritability enrichment within genomic regions conserved across species. Conserved variants were stratified as those located within genomic regions covered by whole-exome sequencing (WES) technologies versus those outside of WES loci. (a) Proportion of h^2_{WGS} explained by conserved variants in and out of WES regions. Dotted lines represent the mean of each annotation. (b) Proportion of h^2_{WGS} explained by common and rare conserved variants in/out WES. (c) Genome-wide enrichments in heritability for conserved variants in/out of WES. (d-e) Enrichments decomposed for common and rare variants. Error bars represent standard errors. Correlations between heritability enrichments shown in panels (c), (d) and (e) were calculated using a Pearson's correlation coefficient (R) over $n=34$ traits. P-values measuring the significance of these correlation were denoted as p in the top-left corner of the corresponding panel and were based on a two-side Pearson's test.



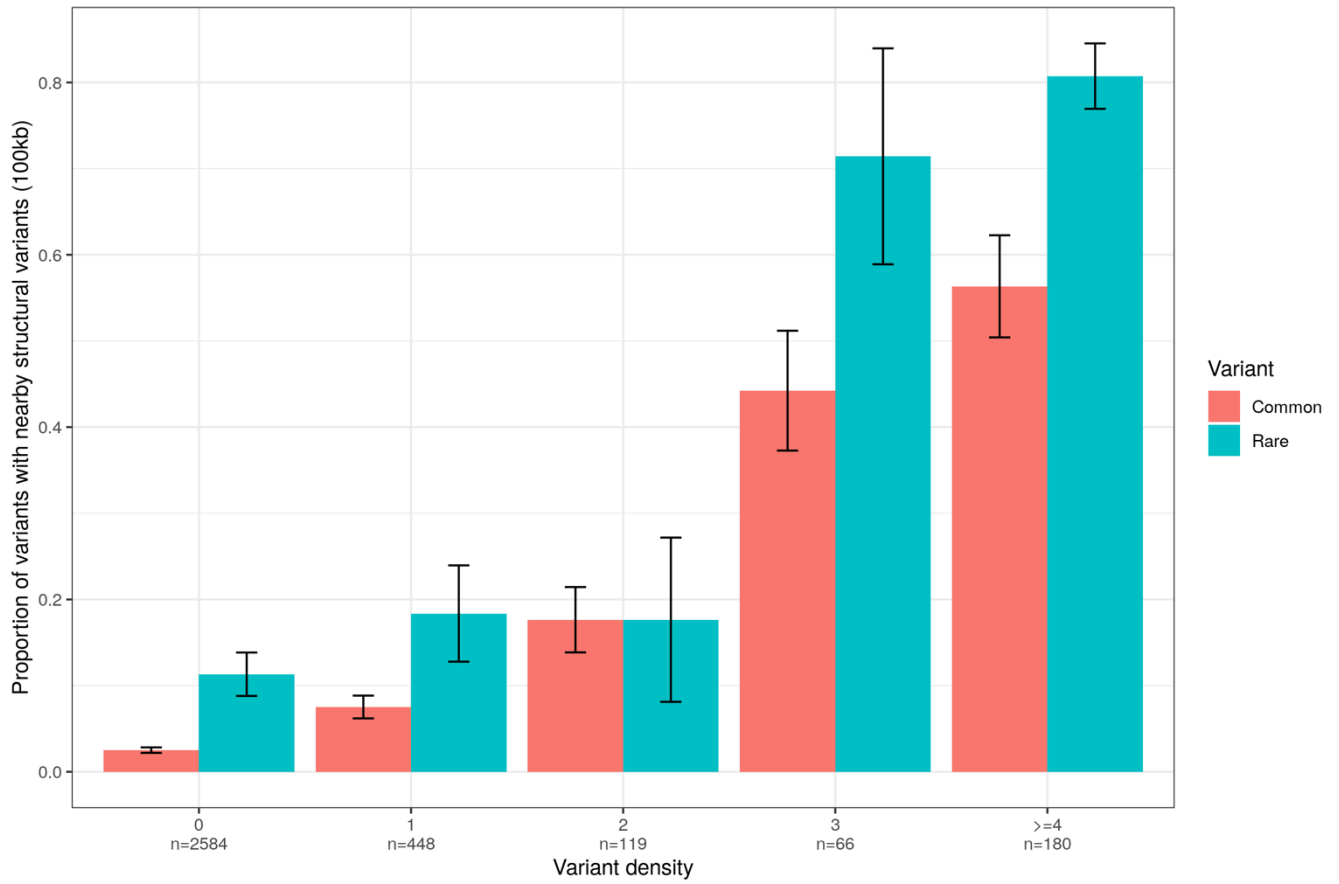
Supplementary Figure 5: RVA annotations enrichments. (a) Variance explained by each RVA across all traits, partitioned by their functional roles. (b) Similar variance showed but partitioned by WES and coding coverage. (c) Using Zoonomia conservation annotations. Boxplot shown here represent the first quartile, the median and the third quartile of the corresponding distribution. P-values shown at the top-left corner of each panel is based on a two-sided F-test.



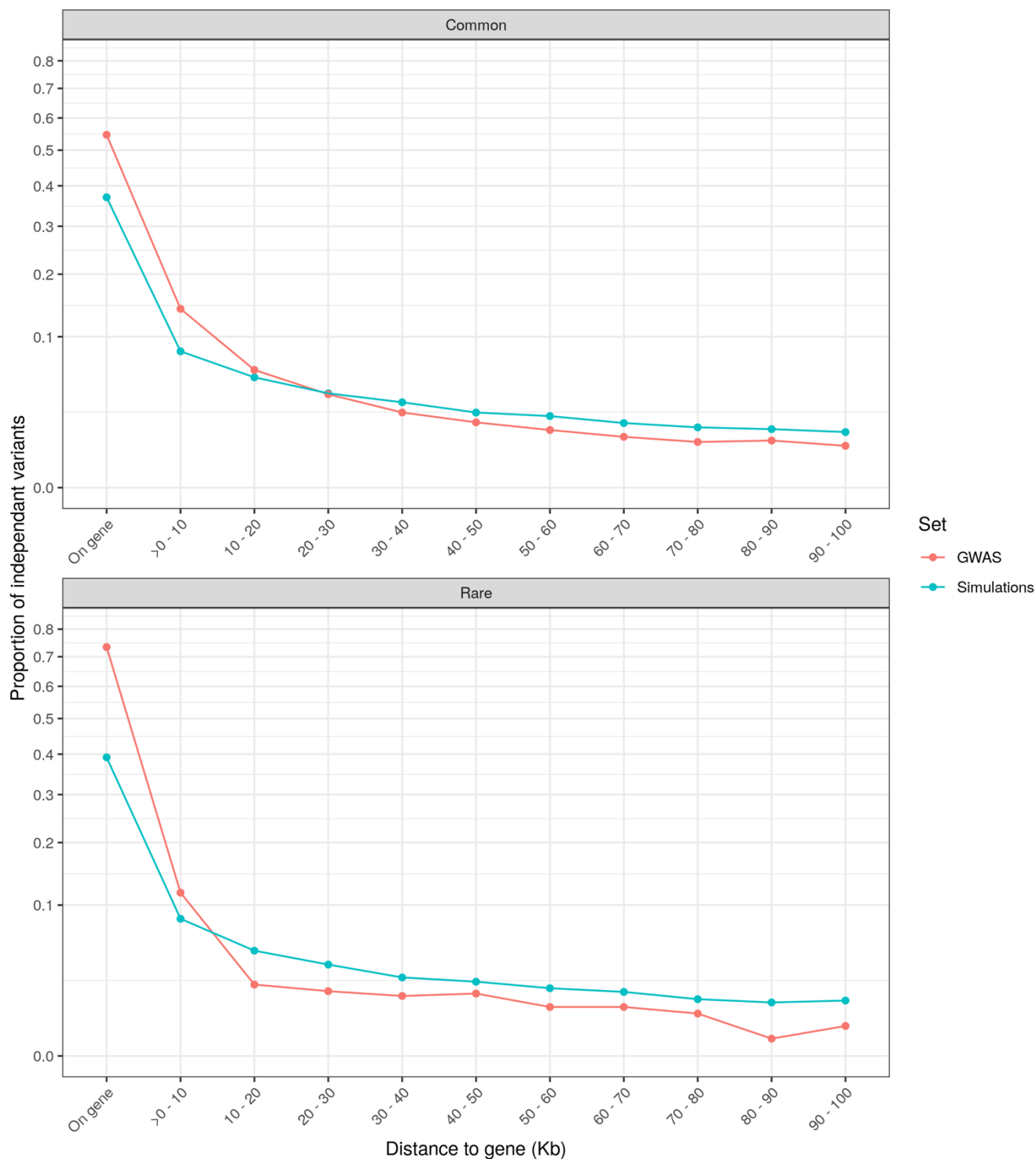
Supplementary Figure 6: Defining pathogenicity. Odds-ratio of a Fisher exact test to evaluate the enrichment between common and rare variants association at different pathogenicity cutoffs based on their PrimateAI 3D percentile score. Subsequent analyses used a pathogenicity threshold of 0.4. Error bars represent standard errors.



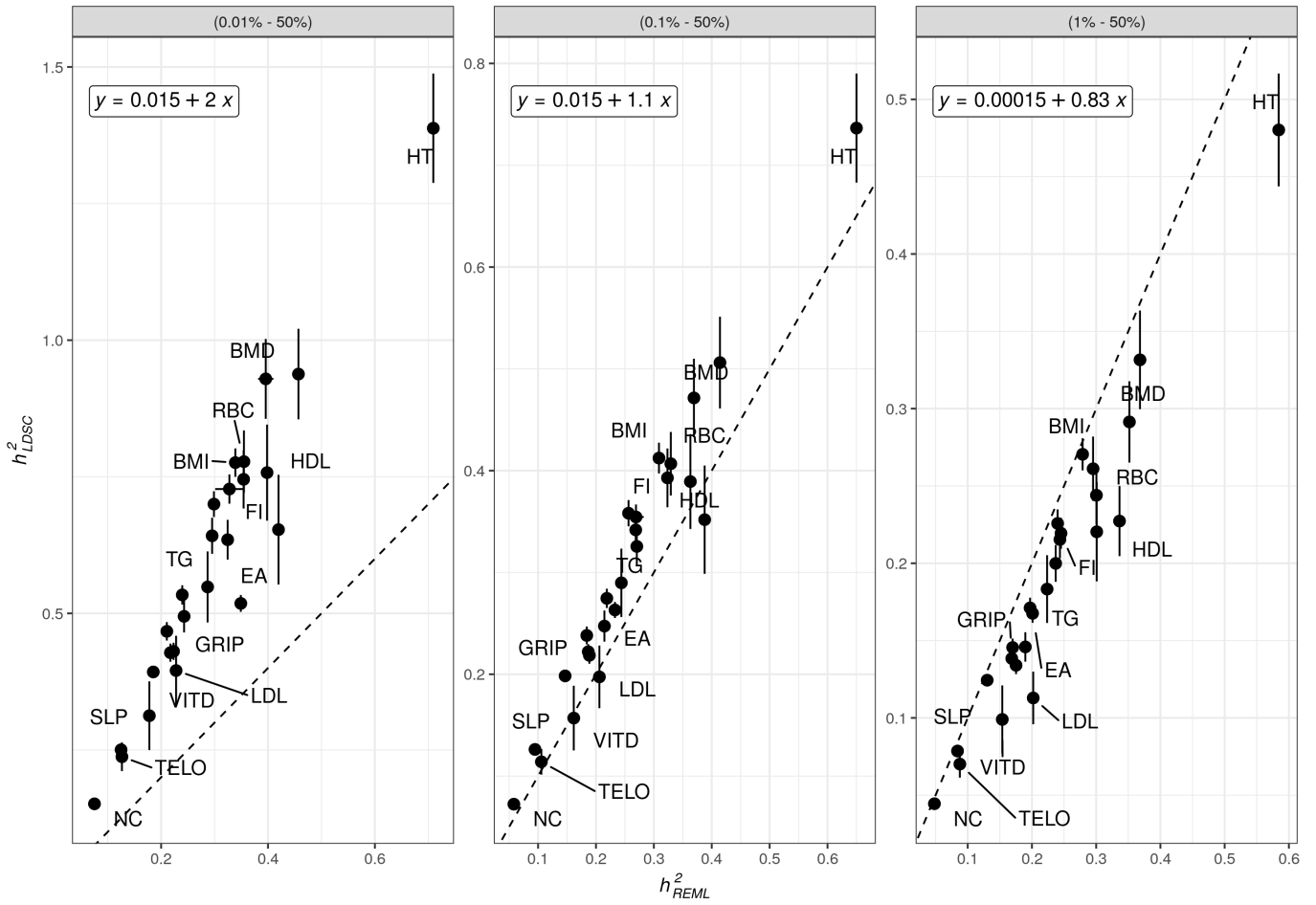
Supplementary Figure 7: Distance to nearest common variant association for trait-associated rare variants. These analyses focus on 19 traits with at least 10 rare-variant associations (RVAs). Each dot represents a RVA. Red dots and boxes represent observed, while blue dots and boxes represent trait-specific null distributions. Null distributions were obtained for each trait by randomly sampling the same number of variants as the number of RVA detected for that trait, matched on allele frequencies and linkage disequilibrium score. Boxplot shown here represent the first quartile, the median and the third quartile of the corresponding distribution.



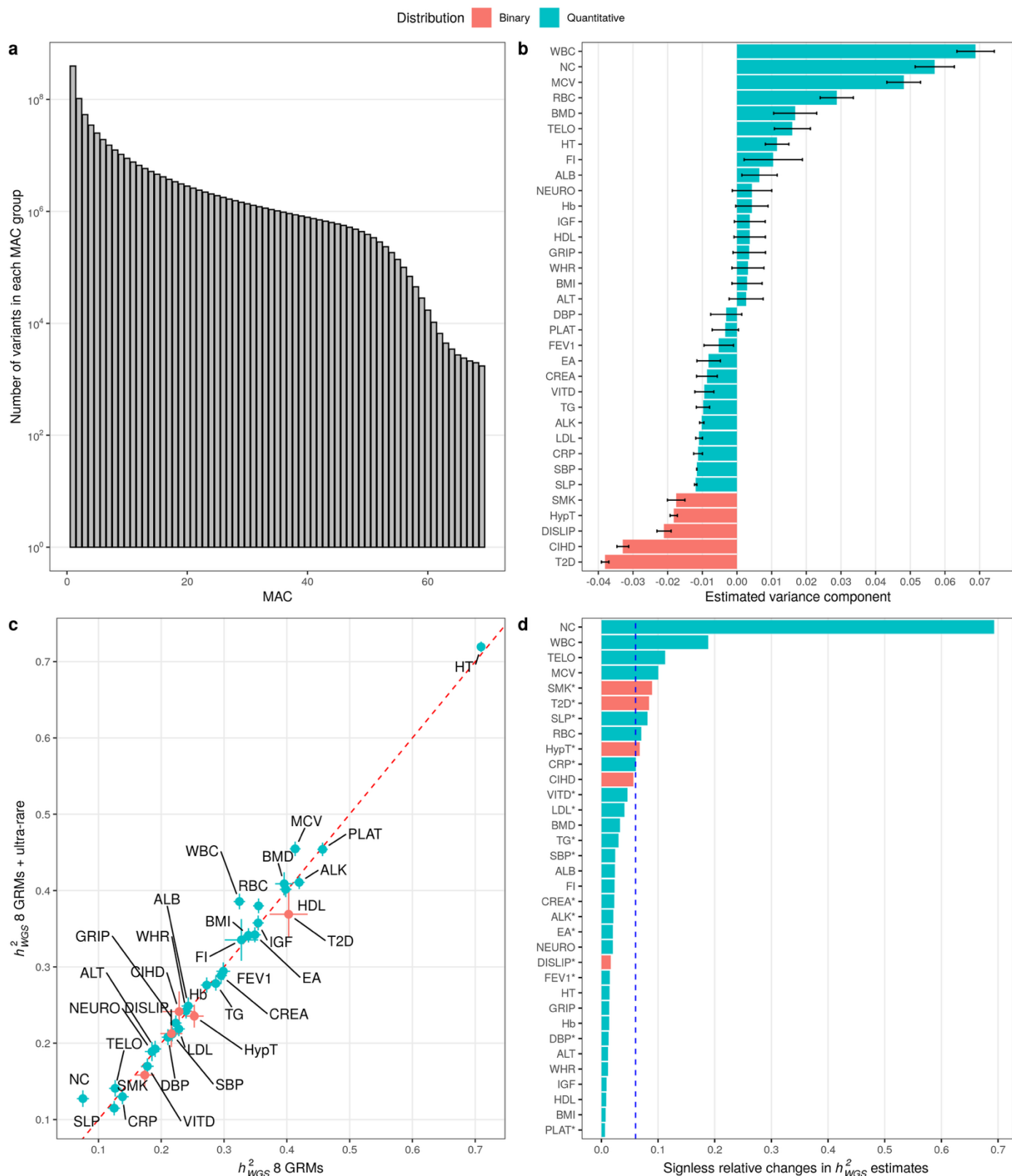
Supplementary Figure 8: Relationship between GWAS signal density and colocalization with structural variants. Trait-associated variants for 20 phenotypes were binned into 5 groups of signal density (x-axis) representing the number of other associations detected within 100 kb of their own position (CVA-CVA and RVA-RVA). Within each density group, we also calculated the proportion of trait-associated variants located within 100 kb of a structural variant associated with the same trait, which appears to be monotonically related with signal density. Finally, we fitted two logistic regression models (for common and rare variants separately) regressing a binary indicator of the presence of a nearby (within 100 kb) trait-associated SV onto a binary indicator of CVA-CVA or RVA-RVA density equal or larger than 2. The odds ratios and p-values for the RVA-RVA and CVA-CVA logistic regression models were 1.8 ($P = 1.2 \times 10^{-35}$) and 1.4 ($P = 2.3 \times 10^{-100}$), respectively. Error bars represent standard errors.



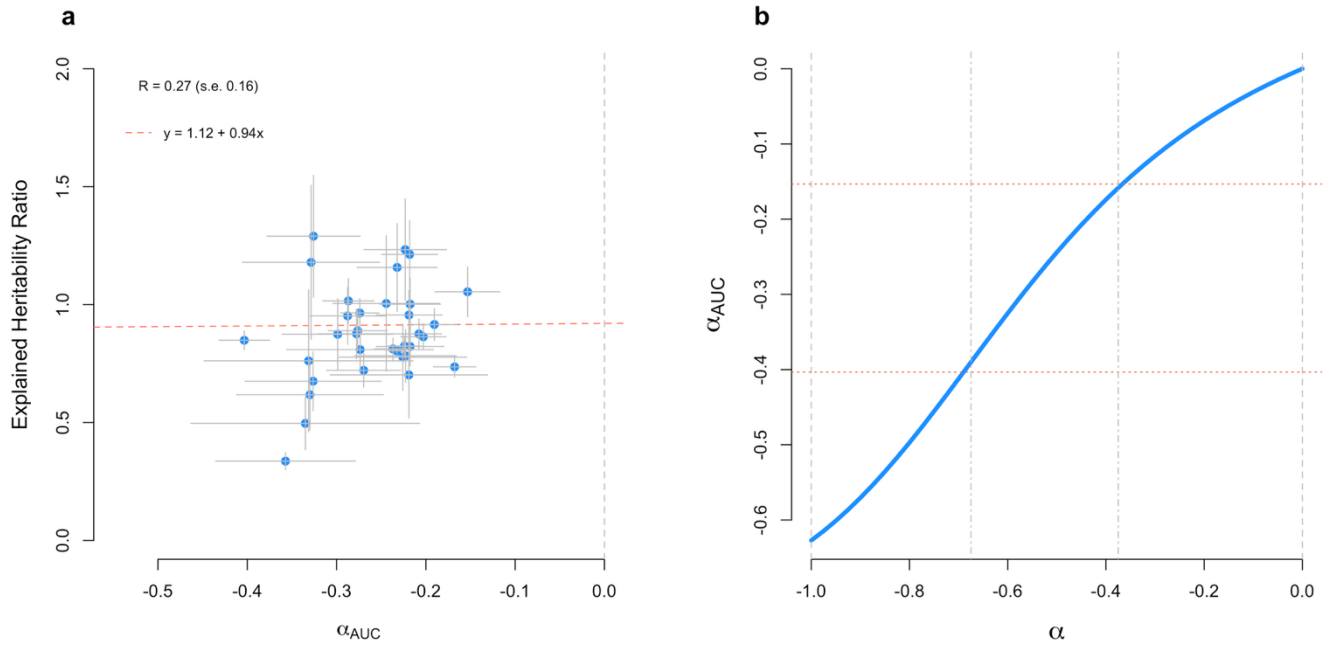
Supplementary Figure 9: Distance of trait-associated variants to gene boundaries. This figure represents the proportion of trait-associated detected on average across traits as a function their distance to genes. Red curves represent observed proportion while blue curves that from a null distribution based on randomly sampled SNPs matched on number, allele frequency and linkage disequilibrium score. The proportion was calculated across all unique variants, independently of the trait, and split between common variant and rare variant associations. Gene boundaries were defined from Encode (gene_start to gene_end).



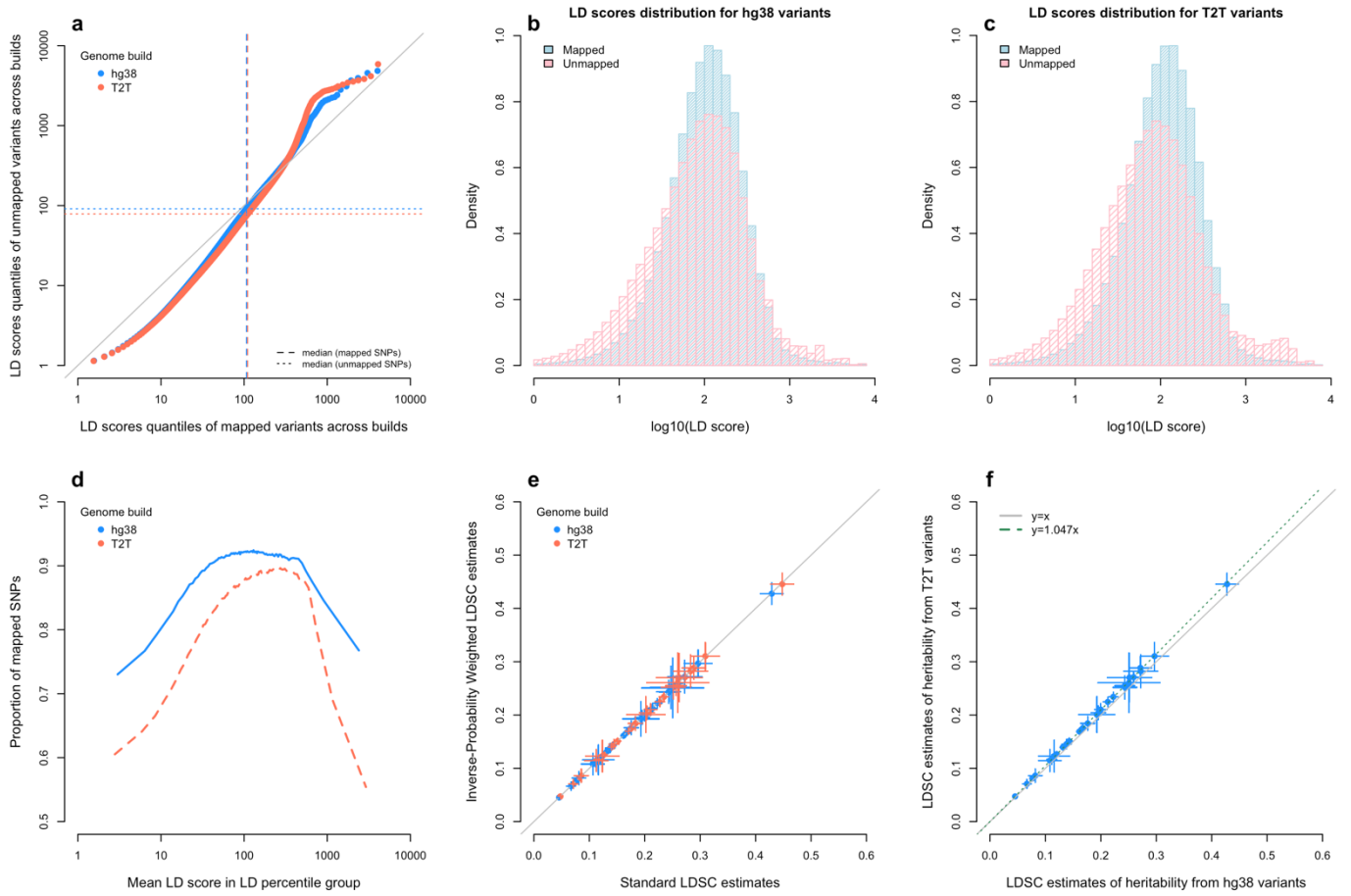
Supplementary Figure 10: Estimates from linkage disequilibrium score regression (LDSC) as a function of the minimum allele frequencies of SNPs included in the analysis. Each panel compares heritability estimates from GREML (x-axis) with those from LDSC (y-axis). The difference between each panel lies in the range of allele frequencies used to obtain those estimates, as indicated at the top of each panel. LDSC estimates were obtained using an estimate of effective sample size calculated as the observed sample size (N) multiplied by an inflation factor (λ). The inflation factor λ was calculated as mean test statistic at genome-wide significant SNPs over the expected test statistic in standard linear regression model, that $1 + N\bar{q}^2 / (1 - \bar{q}^2)$, where \bar{q}^2 denotes the average proportion of trait-variance explained by a detected association (after correction for winner's curse). LDSC was applied to summary statistics obtained from GWAS analyses fitting all covariates. Error bars represent standard errors.



Supplementary Figure 11: Analyses of ultra-rare variants (MAC between 1 and 69). (a) Number of variants for each MAC group from 1 to 69. Variants with MAC > 69 were considered in the main REML- WGS analysis. (b) Estimates of the ultra-rare variant heritability, fitted in a model with common and rare variants GRMs. (c) Total h^2_{WGS} estimates with and without adding the ultra-rare variants GRM. (d) Relative change in h^2_{WGS} estimates induced by including ultra-rare variants. The vertical dotted line represents the mean change across phenotypes (6.0%). Error bars represent standard errors.



Supplementary Figure 12: Relationship between allele frequency and missing heritability. (a) α_{AUC} statistic (**Supplementary Note 5**) calculated for each of the 34 selected traits (each dot is a trait) on the x-axis against the explained heritability ratio (a) on the y-axis. The α_{AUC} measures the relationship between allele frequency and heritability and is shown in **Supplementary Note 5** to be highly consistent with other measures (e.g., the α statistic; Zeng et al. 2018 and Schoech et al. 2019) of the strength of natural selection also based on the relationship between allele substitution effects and heterozygosity. Error bars represent standard errors (s.e.). The correlation between the α_{AUC} statistic and the explained heritability ratio was calculated using a Pearson's correlation coefficient (R) over $n=34$ traits. (b) Relationship between the α_{AUC} statistic (x-axis) and the α statistic (y-axis).



Supplementary Figure 13: Quantification of SNP-based heritability captured outside of hg38 genome build. (a) Quantile-quantile plot comparing LD score distribution for hg38 variants and T2T variants between mapped and unmapped SNPs across genome build. (b) and (c) are histograms of the \log_{10} of LD scores for hg38 and T2T variants stratified by whether those variants are uniquely mapped or not across the two genome builds. (d) Relationship between mean LD score and mapping probability. SNPs were groups into percentile groups. The x-axis shows the mean LD score (on the log-scale) within percentile groups, and the y-axis represents the proportion of mapped variants within the corresponding percentile group. (e) Comparison of LD score regression (LDSC) estimates for common SNPs before (x-axis) and after (y-axis) correction for ascertainment bias using inverse probability weighting (details are given in **Supplementary Note 6**). (f) Comparison of LDSC estimates of common variants heritability from hg38 variants (x-axis) and T2T variants (y-axis). The grey lines in panels (a), (e) and (f) represents the $y=x$ line. The dotted line in panel (f) represents the $y=1.047x$ regression line fitted to model the relationship between T2T-based estimates and hg38-based estimates. Error bars in panels (e) and (f) are jackknife standard errors.

SUPPLEMENTARY NOTES

Supplementary Note 1: Estimates of genetic correlation from WGS data

The extent to which the genetic correlation (r_g) between traits varies across the allele frequency spectrum is largely unknown. We sought to address this question by quantifying potential differences between r_g estimates (\hat{r}_g) obtained from common versus rare variants.

We first estimated genetic correlations from common SNPs using LD score regression (LDSC)⁷¹ for all trait-pairs among our selected set of 34 traits. We then focused on 86 pairs of traits for which the magnitude of the LDSC-based estimated genetic correlation was larger than 0.2. For each of the 86 pairs of traits, we selected samples with both phenotypes available and created a set of common covariates by applying a SVD from each trait-specific SVD matrices (from all covariates available). Due to computational constraints (bivariate analyses with MPH requiring >2 Tb of RAM if all samples were used simultaneously), trait-pairs with more than 200,000 samples (81/86 pairs) were randomly split into two equally sized subsets. We estimated genetic variances and covariances in each subset fitting 8 MAF/LD GRMs (above) and covariates, meta-analysed those estimates across subsets (using a fixed-effect framework), then used the result to calculate genetic correlations for the entire sample. We applied this approach separately for common and rare variants. Standard errors were calculated using the delta-method based on the meta-analysed variance-covariance matrix of estimated genetic (co)variances.

Overall, we found for each trait-pair, that r_g estimated from rare variants was highly concordant with estimates based on common variants (**Extended Data Figure 3c-d**), consistent with results from Wiener et al. (2023) based on whole-exome sequences. However, unlike the latter study, we did not observe a larger magnitude of \hat{r}_g for rare variants (**Extended Data Figure 3d**), which likely reflects the fact that our estimates are dominated by non-coding variants. The most significant difference (Two-sided Wald-test: $P=3.8 \times 10^{-3}$) was observed between haemoglobin levels and red blood cell count, for which the common-variant \hat{r}_g (0.73, s.e. 0.004) was 1.8-fold larger than the rare-variant \hat{r}_g (0.40, s.e. 0.11). However, none of the differences observed between rare-variant and common-variant estimates were significant after multiple testing correction (Wald test $P<0.05/86$). Estimates for all 86 pairs of traits are reported in **Supplementary Table 7**.

Supplementary Note 2: Enrichment in functional annotations

In this note we describe patterns of heritability enrichment in two genomic annotations (genomic loci covered by whole-exome sequencing (WES) technologies, and genomic loci that are conserved across species) and contrast those patterns between common variants and rare variants.

2.1. Heritability enrichment in genomic loci regions covered by whole-exome sequencing

We partitioned \hat{h}_{WGS}^2 to assess the contribution of coding and non-coding variants within WES loci, relative to the rest of the genome. Coding and non-coding WES variants respectively represent 0.71% and 0.29% of the total number of WGS variants included in our analysis (**Supplementary Table 2**).

We found a disproportionate contribution of WES variants to trait heritability with an average contribution to \hat{h}_{WGS}^2 of 17.5% and 6.4% from coding and non-coding WES variants, respectively. Importantly, these estimates are conditional on each other (and conditional on other variants) such that the significant enrichment in non-coding WES variants is unlikely to be explained by LD between coding and non-coding WES SNPs. Moreover, we found that patterns of heritability enrichment in coding and non-coding WES variants were also largely consistent across traits (Pearson's correlation of enrichment statistics across traits: $R=0.67$, two-sided test p-value: $P = 1.3 \times 10^{-5}$; **Supplementary Figure 3c**; **Supplementary Table 10**). Nonetheless, we still observed examples, like that of vitamin D levels, where the heritability enrichment in coding WES variants (51-fold, s.e. 3.2) was four times stronger than in non-coding WES variants (13-fold, s.e. 2.8). Other notable trait-specific observations include LDL, which showed the largest heritability enrichment in coding WES variants (~56-fold, s.e. 2.6 for both traits), while educational attainment showed the lowest significant enrichment (5.3-fold, s.e. 0.8) and number of children had no significant enrichment at all (2.8-fold, s.e. 3.7). Conversely, C-reactive protein had the largest heritability enrichment in non-coding WES variants (46-fold, s.e. 4.9).

Across traits, coding (resp. non-coding) WES variants accounted for 21.0% (resp. 4.1%) of rare-variant heritability and 16.9% (resp. 7.0%) of common-variant heritability (**Supplementary Figure 3b**). However, relative to the proportion of SNPs in those annotations that represents a larger heritability enrichment in coding and non-coding WES variants for common SNPs (coding WES: 36-fold; non-coding WES: 30-fold) as compared to rare SNPs (coding WES: 26-fold; non-coding WES: 13-fold; **Supplementary Table 10**). Moreover, we found that traits whose common-variant heritability is enriched in coding WES variants also tend to display a significant enrichment of rare-variant heritability in the same annotation (Pearson's correlation: $R = 0.56$; two-sided test p-value: $P = 5.6 \times 10^{-4}$; **Supplementary Figure 3d**). Enrichment of rare-variant and common-variant heritability in non-coding WES variants were less correlated (Pearson's correlation: $R = 0.35$; two-sided test p-value: $P = 0.042$; **Supplementary Figure 3e**) across traits and, thus, displayed more discrepancies.

2.2. Heritability enrichment in genomic loci conserved across species

We partitioned \hat{h}_{WGS}^2 to assess the relative contribution of variants within genomic regions that are conserved across species. Variants were labelled as conserved if their Zoonomia phylogenetic score³³ exceeded 2.27. We partitioned the genome into 3 groups of loci: conserved regions overlapping with WES loci (0.35% of SNPs), which includes mostly coding variants, conserved region not overlapping with WES loci (1.72% of SNPs), and the rest of the genome (97.9% of SNPs).

On average across traits, we found that conserved variants contribute 34.3% of \hat{h}_{WGS}^2 (14.3% from conserved variants in WES loci, and 20.0% from conserved variants outside of WES loci). Interestingly, we found, across traits, a significantly negative correlation between heritability enrichment in coding conserved variants as compared to non-coding conserved variants (Pearson's correlation: $R = -0.66$; two-sided test p-value: $P = 2.4 \times 10^{-5}$; **Supplementary Figure 4c**). For example, traits like LDL and telomere length displayed a >60-fold heritability enrichment in conserved variants within WES loci,

while the enrichment in conserved variants outside of WES loci was lower than 5-fold (**Supplementary Figure 4c**). Another interesting observation was the difference in heritability enrichment between telomere length and number of children (two traits with <50% of their pedigree-based heritability accounted for by WGS variants with MAF>0.01%). For telomere length, heritability enrichment was only significant in conserved variants within WES loci (76-fold, two-sided Wald test $P = 1.7 \times 10^{-28}$) but not in conserved variants outside of WES loci (2.7-fold, $P = 0.26$). We observed the opposite for number of children, which only display a significant enrichment of its heritability in conserved variants outside of WES loci (13-fold; two-sided Wald test $P = 5.1 \times 10^{-5}$) but not in conserved variants within WES loci (7.6-fold; two-sided Wald test $P = 0.25$).

Moreover, we found a significant and positive correlation of patterns of heritability enrichment in conserved WES loci between common and rare variants (Pearson's correlation: $R = 0.53$; two-sided Pearson test p-value: $P = 1.2 \times 10^{-3}$; **Supplementary Figure 4d**). However, patterns of heritability enrichment were not correlated outside of WES loci ($R = 0.06$; two-sided test p-value: $P = 0.73$; **Supplementary Figure 4e**).

3.1. Overview of the data and quality control

The Alliance for Genomic Discovery (AGD) dataset is comprised of whole-genome sequencing and phenotypic data derived from Vanderbilt University Medical Center's BioVU biobank. Electronic health record data was collected during clinical encounters. Whole genome sequencing data was generated using Illumina sequencing technology and variant calling used the DRAGEN pipeline (version 3.7.8). European and African ancestry samples in AGD were identified by merging 250,105 AGD samples with samples from the 1000 Genomes (1KG) reference panel, with the first 50 principal components (PCs) calculated on 47,683 LD pruned autosomal HapMap3 SNPs using PLINK2⁵³. A total of 191,454 AGD samples were identified within 3 standard deviations of the 1KG reference European ancestry population mean in the first 10 PCs and, therefore, we classified as European ancestry. A total of 28,232 AGD samples were similarly classified as AGD African ancestry. KING⁷² was applied using the '—kinship' and '—degree 2' options to identify 159,073 and 21,236 unrelated European and African ancestry samples respectively.

Phenotypes (ALK, HDL, LDL) quality control included harmonizing units and correcting for the measurement time of day. Likely pregnancies and individuals under 15 years of age were excluded. Individuals with raw measurements more than 6 standard deviations from the mean were also excluded. After these exclusions, phenotypes were mean-centred and scaled to variance 1 within each sex group. Scaled phenotypic measurements were then regressed on baseline covariates including age, sex, age², age×sex and age²×sex interactions and 20 genotypic PCs. Residuals from the latter regression analyses were treated as the covariate-corrected phenotypes and used in downstream analyses.

Trait-associated variants identified in the UKB were matched to those available in AGD. Variant quality control included the following criteria: p-value > 0.001 for the Hardy-Weinberg Equilibrium test, genotype missingness rate <0.10, and minor allele counts >10. Quality control was performed in each ancestry group, separately.

3.2. AGD replication analysis

We quantified the proportion of trait-variance explained by RVAs identified in UKB using two approaches. First, we fitted all RVAs (passing quality control in AGD) into a multivariate regression model and assessed how much including these variants improved model fit (measured using the adjusted R² statistic) over and above a model only fitting age, sex and 20 genotypic PCs (hereafter referred to as baseline model). For the second approach we assessed the incremental adjusted R² statistic induced by adding to the baseline model a polygenic score combining these RVAs with effect sizes estimated in the UKB. We also performed a direct replication analysis by testing the association of each RVA identified in the UKB in our AGD samples. We used fastGWA⁷³ for these replication analyses.

⁶³A complete overview of the sample-sizes for each ancestry group, estimates of variance explained and summary statistics for the replication analyses are available in **Supplementary Table 14** and **Supplementary Table 16**.

4.1. Background

Wainschtein and colleagues¹² previously showed that a larger proportion of pedigree-based heritability is recoverable using WGS data as opposed to imputed SNPs. We sought to approach this question from a different angle by quantifying how many associations detected with WGS could also be detected using imputed SNPs. Therefore, we performed similar GWAS analyses as described in the main text using imputed SNPs with MAF>0.01%. For these analyses, we used genotypes imputed from two reference panels: HRC+UK10K⁷⁴ (30.8M SNPs tested) and TOPMed⁷⁵ (31.9M SNPs tested).

4.2. Results

Compared to the 12,129 (11,243 CVAs and 886 RVAs) independent associations detected with WGS, we detected 11,597 (10,903 CVAs and 694 RVAs) and 11,770 (10,942 CVAs and 828 RVAs) associations using HRC+UK10K and TOPMed panel, respectively (**Extended Data Figure 5a; Supplementary Table 17, Supplementary Table 18**). We assessed the colocalization between WGS-based and imputation-based associations by quantifying the density of imputation-based associations within a specified window around each WGS-based association. Specifically, we focused on WGS-based associations with a density of 0, meaning that those loci could not be detected using imputed SNPs. For ~95% of WGS-based associations (11,474 / 12,129), we found at least one imputation-based association detected within a 100 kb window (on both sides), with a stronger colocalization for TOPMed-imputed variants as compared to HRC+UK10K-imputed variants (**Extended Data Figure 5b**). Nevertheless, we also identified 247 loci (37 RVA loci and 210 CVA loci) where no imputation-based associations were detected within 1 Mb on each side of WGS-based associations (**Supplementary Table 19**). Interestingly, while this observation is not surprising for the 37 RVAs because of the known (relatively) low imputation quality for rare variants,^{76,77} we found that 210 of these 247 missed loci were, in fact, centred around a CVA. We showcase two of these missed loci in **Extended Data Figure 5c** and **Extended Data Figure 5d**, which are centred around an *EBF3* intronic splicing indel (SpliceAI score 0.11 and MAF = 45%) associated with WHR, and a rare pathogenic SNP (PrimateAI 3D percentile score of 0.78 and MAF = 0.8%) downstream *TINF2* associated with telomere length and reported in a recent study by Burren and colleagues²⁵.

Finally, we applied SuSiE^{65,66} to compare the resolution of fine-mapping of GWAS loci detected with WGS versus imputed variants (METHODS). Overall, we found higher posterior inclusion probabilities with WGS as compared to imputed data (**Extended Data Figure 6a-b**). Consistently, we also observed smaller credible sets, on average, for WGS-based analyses (**Extended Data Figure 6c-d**). We found a minimal gain of WGS over imputation at loci where the lead SNP was a common variant, while fine-mapping of RVAs using WGS yielded much smaller credible sets. Specifically, we identified 13% more single-variant credible sets for RVAs identified using WGS as compared to RVAs identified and fine-mapped using imputed SNPs. In contrast, WGS and imputation yielded similar numbers of single-variant credible sets. We provide the credible sets for all independent associations detected with WGS and imputed variants (**Supplementary Data**).

Altogether, our results are consistent with previous studies suggesting that imputation captures 70-80% of WGS-based heritability¹², highlight that using WGS over imputation improves fine-mapping resolution for rare-variant associations and that existing imputation panels may still be missing common haplotypes in European ancestry populations.

4.3. Discussion

The cost-effectiveness of WGS for GWAS discovery is still debated relative to alternatives combining WES with SNP-array genotyping followed by imputation⁷⁸. Our study shows that while imputation captures the bulk of genetic signals, it misses a small fraction of common-variant associations and, expectedly, a larger fraction of rare-variant associations. Although many rare-variant associations missed by imputation can be detected by WES, we nevertheless highlighted multiple variants explaining

notable amounts of phenotypic variance, which were detected outside of genomic regions covered by WES technologies.

Moreover, while choosing the right genomic technology is a critical part of designing future genetic studies, determining an optimal sample size is equally important. By providing precise estimates of rare-variant heritability (and therefore, of average per-SNP variance explained), our study informs the expected accuracy of WGS-based polygenic scores and sample size calculation to detect rare-variant associations in future WGS-based GWAS. Finally, by quantifying the degree of colocalization between common- and rare-variant associations, our study also provides a key indication of how often rare-variant associations can be expected near common-variant associations, which could also inform optimal statistical methods to detect these rare-variant associations.

In this note, we sought to test if the gap between \hat{h}_{WGS}^2 over \hat{h}_{PED}^2 could be explained by negative selection, which is expected to reduce I by inflating the relative contribution to heritability from ultra-rare variants (MAF<0.01%) not been included in our primary analyses. As in previous studies, we quantified negative selection by modelling the relationship between allele frequency and heritability using a single parameter α_{AUC} , which values below 0 indicate stronger strengths of selection (or rapid population expansion). Importantly, the impact of population expansion on heritability is not expected to differ between traits and, therefore, would not contribute to I variation across traits. Details about the definition and properties of α_{AUC} , as well as its relationship with other models of negative selection are discussed in the remainder of the note. Across phenotypes, we found estimates of α_{AUC} significantly lower than 0 (two-sided Wald test p-value: $P<0.05/34$) for 29 out of 34 traits (**Supplementary Table 21**), consistent with widespread evidence of negative selection on human phenotypes. However, we did not find a significant correlation between estimates of α_{AUC} and I (Pearson's correlation: $R=0.007$, two-sided Pearson's test p-value $P=0.97$; **Supplementary Figure 12a**), suggesting that the variation across traits is unlikely to be explained by negative selection causing an enrichment of ultra-rare-variant heritability.

5.1. Definition of the α_{AUC} statistic

We define $K \geq 2$ contiguous MAF intervals $[x_k; x_{k+1})$ such that $0 \leq x_1 < x_2 < \dots < x_{K+1} \leq 0.5$. We denote h_k^2 as the proportion of trait variance explained by SNPs in the $[x_k; x_{k+1})$ MAF interval. Using the trapezoidal rule, we can express the area under the curve (hereafter denoted AUC) representing the cumulative proportion of heritability between $[x_1; x_{K+1})$ as a function of x_k as

$$(5.1) \quad \text{AUC} = x_{K+1} - \frac{\sum_{k=1}^K h_k^2 [(x_k + x_{k+1})/2]}{\sum_{k=1}^K h_k^2}$$

Under neutral evolution, allele substitution effects of causal variants are expected to be independent of their frequencies. In that case, h_k^2 would be expected to be proportional to $m_k \bar{H}_k$, where m_k and \bar{H}_k are the number and average heterozygosity of SNPs in the $[x_k; x_{k+1})$ MAF interval. We used empirical data from 40,575,204 quality-controlled sequenced variants to determine m_k and \bar{H}_k for each MAF class, then utilised those values to calculate an expected AUC under neutral evolution of $\text{AUC}_0 = 0.241$. Note that AUC_0 is approximately equal to 0.25, which is the value expected when assuming a constant effective population size (N_e) and a probability density function for allele frequencies (p) proportional to $1/[p(1-p)]$. A difference between AUC and AUC_0 , indicates that lower frequency variants disproportionately contribute to heritability, which is expected under negative selection but also under rapid population expansion⁷⁹. For each trait, the sampling variance of AUC was approximated using the delta-method based on the sample covariance matrix of estimates of h_k^2 provided as an output of the GREML method. Finally, we define α_{AUC} as $\alpha_{\text{AUC}} = \log(\text{AUC}_0) - \log(\text{AUC})$, which is a monotonic transformation of AUC and thus retains the same interpretation with values of α_{AUC} below 0 indicating a disproportionate contribution to heritability from lower frequency variants.

5.2. Relationship between the α_{AUC} statistic and other models of negative selection

Previous studies^{80,81} have proposed a genetic architecture model assuming the variance of causal effect at SNP j to be proportional to $[2p_j(1-p_j)]^\alpha$, with p_j being the minor allele frequency of that SNP, and α a scalar whose magnitude is correlated with the strength of negative selection. We hereafter refer to this model as the α -model. Under the (additional) assumption that causal variants are unlinked and randomly sampled across the genome, the α -model implies that the fraction $h_k^2(\alpha)$ of heritability in the $[x_k; x_{k+1})$ MAF interval is expected to be proportional to

$$(5.2) \quad h_k^2(\alpha) \propto \sum_{p_j \in [x_k; x_{k+1})} [2p_j(1 - p_j)]^\alpha$$

In practice, we propose to quantify $h_k^2(\alpha)$ using the empirical distribution of allele frequencies, which has already been shaped by natural selection. However, one may also consider a null distribution of allele frequencies (e.g., assuming a probability density function of p_j proportional to $1/[p_j(1 - p_j)]$) to make these calculations. We used Eqn. (5.2) to derive the expected AUC statistic under the α -model as

$$(5.3) \quad \text{AUC}(\alpha) = x_{K+1} - \frac{\sum_{k=1}^K h_k^2(\alpha) [(x_k + x_{k+1})/2]}{\sum_{k=1}^K h_k^2(\alpha)}$$

Supplementary Figure 12b is based on Eqn. 5.2 and shows that α and $\alpha_{\text{AUC}} = \log(\text{AUC}_0) - \log[\text{AUC}(\alpha)]$ are almost perfectly correlated for values of α between -0.675 and -0.375 ($\alpha_{\text{AUC}} \approx 0.14 + 0.78 S$; $R^2 = 0.997$), although a nonlinear (yet monotonic) relationship is expected over a wider range of values. Importantly, values of α between -0.675 and -0.375 correspond to α_{AUC} between -0.40 and -0.15, which is the range of estimates observed across the 34 traits analysed in our study. These results are also consistent with findings from Zeng et al.⁸⁰ who showed that estimates of α across 27 complex traits measured in the UK Biobank were also strongly correlated with AUC ($R=0.896$). Therefore, given that estimates of α are more robust to the effect of population expansion, this also suggests that α_{AUC} should be more influenced by negative selection than population expansion.

5.3. Derivation of diagonal GRM for ultra-rare variants

In this section, we present the detailed derivation of Eqn. (2) from Eqn. (1). From Eqn. (1), we can express diagonal elements of the GRM as

$$(5.4) \quad D_{ii} = \frac{1}{M} \sum_{j=1}^M \frac{(x_{ij} - 2p_j)^2}{2p_j(1 - p_j)}$$

where x_{ij} is the minor allele count at SNP j for individual i , M the number of variants used to quantify relatedness and p_j the minor allele frequency (MAF) at SNP j . This equation applies for all SNPs, including ultra-rare variants (MAF<0.01%).

We now stratify each ultra-rare variants into K groups of variants with the same minor allele count (MAC). Our set of ultra-rare variants spans $K = 69$ MAC groups. Note that $K=69$ corresponds to our MAF threshold for defining ultra-rare variants because $69/(2N) < 0.0001$ and $(69+1)/(2N) > 0.0001$, where $N = 347,630$ is the sample size in our primary analyses.

Moreover, we can notice that $p_j = k/(2N)$ for all variants in the k -th MAC group (hereafter denote G_k). Therefore, Eqn. (5.4) can then be rewritten as

$$(5.5) \quad D_{ii} = \frac{1}{M} \sum_{j=1}^M \frac{(x_{ij} - 2p_j)^2}{2p_j(1 - p_j)} = \frac{1}{M} \sum_{k=1}^K \sum_{j \in G_k} \frac{(x_{ij} - k/N)^2}{(k/N) \left[1 - \frac{k}{2N}\right]} = \frac{1}{M} \sum_{k=1}^K \sum_{j \in G_k} \frac{x_{ij}^2 - \left(\frac{2k}{N}\right) x_{ij} + \left(\frac{k}{N}\right)^2}{(k/N) \left[1 - \frac{k}{2N}\right]}$$

Given that homozygotes for ultra-rare alleles are rarely observed, we can assume that x_{ij} mainly takes values between 0 and 1 and, therefore, that $x_{ij}^2 \approx x_{ij}$. This leads to

$$(5.6) \quad D_{ii} \approx \frac{1}{M} \sum_{k=1}^K \sum_{j \in G_k} \frac{\left(1 - \frac{2k}{N}\right) x_{ij} + \left(\frac{k}{N}\right)^2}{(k/N) \left[1 - \frac{k}{2N}\right]} = \frac{1}{M} \sum_{k=1}^K \sum_{j \in G_k} \frac{N(N - 2k)x_{ij} + k^2}{k(N - k/2)}$$

Finally, if we denote M_k as the number of ultra-rare variants in the k -th MAC group and $S_{ik} = \sum_{j \in G_k} x_{ij}$ then

$$(5.7) \quad D_{ii} \approx \frac{1}{M} \sum_{k=1}^K \frac{N(N - 2k)S_{ik} + k^2 M_k}{k(N - k/2)}$$

which proves Eqn. (2).

6.1. Overview and main results

This note compares SNP-based heritability estimates from common variants (i.e., $MAF > 1\%$) mapped to the GRCH38 (hg38 in short) genome build versus those mapped to the Telomere-to-Telomere CHM13 (T2T in short) genome build. T2T contains an extra ~ 200 Mb (that is, $\sim 6\%$) of the DNA sequence as compared to hg38, which may potentially explain a fraction of the still missing heritability. We focused on common variants because large samples with WGS data mapped to T2T are not currently available.

We used LD score regression (LDSC) to obtain SNP-based heritability estimates for T2T and hg38 variants separately. We focused on 29 quantitative traits for which we showed in **Supplementary Figure 10** that LDSC estimates are well-calibrated relative to GREML. Our analyses were based on a subset of 9,094,785 variants uniquely mapped between hg38 and T2T (details in Section 4.2 below). For each of these mapped variants, we calculated two LD score statistics: (i) $\ell^{(hg38)}$ measuring tagging of other common variants available in hg38, and (ii) $\ell^{(T2T)}$ measuring tagging of common variants available in T2T (including those that can and cannot be mapped to hg38). To ensure a fair comparison of LDSC estimates from these two LD score statistics, $\ell^{(hg38)}$ and $\ell^{(T2T)}$ were calculated from the same set of $n = 490$ European ancestry individuals in the 1000 Genomes Project, whose genotypes were available under the two genome builds. LD scores for hg38 variants were calculated from a panel of $M_{hg38} = 11,413,052$ autosomal common SNPs, while LD scores for T2T variants were obtained from on a panel of $M_{T2T} = 12,511,561$ autosomal common SNPs.

LDSC estimates heritability as the slope from a (weighted) linear regression across variants. Therefore, it is possible to obtain unbiased estimates of that slope from a subset of SNPs such as those uniquely mapped between hg38 and T2T. However, unbiasedness would only be guaranteed if the distribution of LD scores is the same between mapped and unmapped variants. Overall, we found the LD scores distributions to vary significantly between mapped and unmapped variants (Kolmogorov-Smirnov test $p\text{-value} < 10^{-10}$; **Supplementary Figure 13a-d**). Nevertheless, these differences did not seem to substantially affect LDSC analyses. In fact, estimates obtained before and after correction for ascertainment biases using inverse-probability weighting were virtually identical (Section 5.2 below; **Supplementary Figure 13e**). Note that mean differences in LD scores (resp. log LD scores) between mapped and unmapped explain $< 0.3\%$ of LD scores (resp. $< 1\%$ of log LD scores) variance across SNPs, which might explain why ascertainment bias seemed minimal.

Finally, we found, on average across traits, that T2T variants capture 4.7% (range across traits: 3.5% to 7.4%) more heritability than hg38 variants. This gain is ~ 2 -fold lower than the 9.6% increase in total number of common variants detected using T2T (12,511,561 vs. 11,413,052), suggesting that loci missing from hg38 are relatively depleted of heritability signal (**Supplementary Figure 13f**). Importantly, this conclusion still holds even when comparing the 4.7% increase in heritability estimates to the more localized 7.8% average increase in variant density around mapped SNP within the 10 Mb window used to calculate LD scores.

6.2. Methods

6.2.1. Quality control and LD score calculations

We downloaded genotypes of 1000 Genomes participants under the hg38 genome build from the PLINK website (URL) and under the T2T CHM13 genome build from the GitHub repository of the Telomere-to-telomere consortium CHM13 project (URL). We focused on $n = 490$ samples whose ancestry group ('super population') was labelled as European. For each dataset, we selected biallelic autosomal variants with a minor allele frequency larger than 1%, a genotype call rate $> 95\%$ and a $p\text{-value} > 10^{-8}$ for the Hardy-Weinberg test. These quality control steps were performed using PLINK2 and led to include $M_{hg38} = 11,413,052$ autosomal common hg38 SNPs and $M_{T2T} = 12,511,561$ autosomal

common T2T SNPs in our analyses. LD scores were calculated within 10 Mb of each SNP using a custom C++ code (URL). LD scores were adjusted for sampling biases by subtracting from each squared Pearson correlation (\hat{r}^2) its expected bias, $(1 - \hat{r}^2)/n$. We then restricted our analyses to SNPs with a corrected LD score statistics larger than 1. Less than 0.01% of SNPs were excluded.

6.2.2. Mapping SNPs across builds

We used *liftOver* to map quality-controlled SNPs between hg38 and T2T. We then retained a subset of 9,094,785 SNPs uniquely mapped between the two builds. Unique mapping was defined using the following criteria: (i) mapping from hg38 to T2T and back from T2T to hg38 yields the same result, (ii) SNPs are mapped to the same chromosome across the two builds, and (iii) the squared correlation of genotypes between builds is >0.95 .

6.2.3. Inverse-probability weighted LDSC analyses

We grouped SNPs into LD score percentile groups and quantified the proportion of uniquely mapped variants (see criteria listed above) within the corresponding percentile group. **Supplementary Figure 13d** shows a non-linear relationship between the mean LD score in each percentile group (x-axis) and the proportion of mapped variants in the group (y-axis). We used the proportion of mapped variants as an estimator of the mapping probability for SNPs within the corresponding LD score percentile group. Inference in LDSC already uses weights to account for non-independence between SNPs and heteroskedasticity. Therefore, we used the same framework and added another multiplicative weight equal to the inverse of the estimated mapping probability. Inverse-probability-weighted LDSC was implemented in a custom R function (URL). Standard errors were calculated using block jackknife based on approximately 300 10-Mb wide blocks across autosomes.

6.2.4. URLs

- 1000 Genomes hg38 genotypes:
<https://www.cog-genomics.org/plink/2.0/resources>
- 1000 Genomes T2T genotypes:
[https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/assemblies/variants/1000 Genomes Project/chm13v2.0/](https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/assemblies/variants/1000%20Genomes%20Project/chm13v2.0/)
- Custom C++ code for calculating LD scores
<https://doi.org/10.5281/zenodo.16550864>

SUPPLEMENTARY REFERENCES

71. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* 47, 1236–1241 (2015).
72. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873 (2010).
73. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet* 51, 1749–1755 (2019).
74. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209 (2018).
75. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290–299 (2021).
76. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat Genet* 53, 1260–1269 (2021).
77. Rubinacci, S., Hofmeister, R. J., Sousa da Mota, B. & Delaneau, O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nat Genet* 55, 1088–1090 (2023).
78. Gaynor, S. M. *et al.* Yield of genetic association signals from genomes, exomes and imputation in the UK Biobank. *Nat Genet* 56, 2345–2351 (2024).
79. Uricchio, L. H., Zaitlen, N. A., Ye, C. J., Witte, J. S. & Hernandez, R. D. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Res* 26, 863–873 (2016).
80. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat Genet* 50, 746–753 (2018).
81. Schoech, A. P. *et al.* Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat Commun* 10, 790 (2019).