ARTICLE

Genomic partitioning of inbreeding depression in humans

Loic Yengo,[1,*] Jian Yang,[1,2,3] Matthew C. Keller,[4,5] Michael E. Goddard,[6,7] Naomi R. Wray,[1,8] and Peter M. Visscher[1]

Summary

Across species, offspring of related individuals often exhibit significant reduction in fitness-related traits, known as inbreeding depression (ID), yet the genetic and molecular basis for ID remains elusive. Here, we develop a method to quantify enrichment of ID within specific genomic annotations and apply it to human data. We analyzed the phenomes and genomes of ~350,000 unrelated participants of the UK Biobank and found, on average of over 11 traits, significant enrichment of ID within genomic regions with high recombination rates (>21-fold; $p < 10^{-5}$), with conserved function across species (>19-fold; $p < 10^{-4}$), and within regulatory elements such as DNase I hypersensitive sites (~5-fold; $p = 8.9 \times 10^{-7}$). We also quantified enrichment of ID within trait-associated regions and found suggestive evidence that genomic regions contributing to additive genetic variance in the population are enriched for ID signal. We find strong correlations between functional enrichment of SNP-based heritability and that of ID ($r = 0.8$, standard error: 0.1). These findings provide empirical evidence that ID is most likely due to many partially recessive deleterious alleles in low linkage disequilibrium regions of the genome. Our study suggests that functional characterization of ID may further elucidate the genetic architectures and biological mechanisms underlying complex traits and diseases.

Introduction

Mating between genetically related individuals, i.e., inbreeding, has detrimental phenotypic consequences in resulting offspring.[1–8] This phenomenon, known as inbreeding depression (ID), has been reported for multiple human traits, such as stature, intelligence, lung function, and fertility.[5–7,9] Inbreeding results in increased homozygosity across the genome and ID can be explained by the increased homozygosity of (partially) recessive deleterious alleles, thereby exposing their detrimental effects on fitness and fitness-related traits. Detecting ID and quantifying its strength in human populations is particularly challenging because inbreeding is less common and less extreme and the effective population is larger in humans than in some other animal species.[10,11] As a consequence, sample sizes of 100,000s or more participants are required to obtain reliable estimates of ID.[12] Over recent years, the advent of large scale biobanks such as the UK Biobank (UKB) has allowed detection and quantification of ID in a wide variety of traits, including hip-to-waist ratio, heart rate, facial aging, or hemoglobin levels, which previously were not known to be associated with inbreeding.[7]

Despite a growing catalog of traits associated with inbreeding, the genetic basis of ID in humans remains elusive. Notably, genes, biological pathways, or functions involved in ID are still largely unknown. Previous studies have attempted to identify genes involved in ID by testing the association between runs of homozygosity (ROHs)[13] and traits or diseases.[14] Although the latter studies have not robustly identified such genes in humans,[15,16] the approach has been more powerful in other animal species with smaller effective population sizes, such as cattle.[17,18] In addition, studies in *Drosophila*[19–21] have tested the association between inbreeding coefficients (F) and gene expression and have thereby identified genes implicated or affected by inbreeding. Overall, contrasting findings from human and non-human studies highlight the limited statistical power to dissect ID in humans and thus calls for larger sample sizes and new analytical methods to overcome this challenge.

Here, we develop a method to detect and quantify ID at a finer scale by breaking its effects down to genomic regions with specific annotations. An enrichment of ID within a given genomic region means that homozygosity in that region (or regions sharing the same annotation) has a disproportional effect on the mean value of a trait as compared to the effect of homozygosity at any equally sized, randomly selected genomic region. We chose genomic annotations as a higher-level unit (as opposed to lower-level units, such as genes, for example) in order to measure the functional impact on traits of homozygosity at particular types of variants. Using this method, we analyzed genetic and phenotypic data from a large sample of ~350,000 unrelated participants of the UKB. Our method utilizes (but is not restricted to) functional annotations introduced in

[1]Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia; [2]School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310024, China; [3]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang 310024, China; [4]Department of Psychology and Neuroscience, University of Colorado, Boulder, CO 80309, USA; [5]Department of Integrative Physiology, University of Colorado, Boulder, CO 80309, USA; [6]Faculty of Veterinary and Agricultural Science, Na, Parkville, VIC 3010, Australia; [7]Agriculture Victoria Research, Department of Jobs, Precincts and Regions, Bundoora, VIC 3083, Australia; [8]Queensland Brain Institute, University of Queensland, Brisbane, QLD 4072, Australia
*Correspondence: l.yengodimbou@uq.edu.au
https://doi.org/10.1016/j.ajhg.2021.06.005.

Finucane et al.,[22] Gazal et al.,[23] and Hujoel et al.,[24] covering diverse properties of the human genome. We specifically quantify the degree to which variants located within these functional genomic regions contribute to ID across 11 traits. These 11 traits were selected on the basis of prior evidence of ID[11] and are height, hip-to-waist ratio, handgrip strength (average of left and right hand), lung function measured as the peak expiratory flow, visual acuity, auditory acuity, number of years of education, fluid intelligence score, cognitive function measured as the mean time to correctly identify matches, fertility measured as the number of offspring, and overall health measured as the number of diseases diagnosed in an individual. We use our method to quantify the enrichment of ID within trait-associated genomic regions identified through genome-wide association studies (GWASs) and finally compare functional enrichments of heritability with that of ID.

Material and methods

Quantification of the enrichment of ID via individual-level data

Model definition

Let y denote a quantitative trait subject to ID. We assume an infinitesimal model where all SNPs contribute to ID. We can therefore write y as

$$y = \sum_{j=1}^{M} \beta_j F_j + e, \qquad \text{(Equation 1)}$$

where F_j denotes an estimator of the inbreeding coefficient based on genotypes at SNP j ($j = 1,\ldots,M$), β_j the contribution of SNP j to ID, and e a residual term capturing all other effects, including additive genetic and environmental effects.

Estimation of ID from inbreeding measures based on genome-wide average homozygosity (as opposed to runs of homozygosity [ROHs]) relies on the assumption that β_j is random and has a constant expectation $E[\beta_j] = b/M$, where b denotes the genome-wide ID. Other inbreeding measures, such as the excess of homozygosity measure, rely on the assumption that $E[\beta_j]$ is proportional to mean heterozygosity $h_j = 2p_j(1-p_j)$, where p_j is the minor allele frequency (MAF) at SNP j in the population. We previously introduced a flexible MAF and linkage disequilibrium (LD)-stratified method to estimate ID, which we showed to yield unbiased estimation of ID even when the relationship between $E[\beta_j]$ and h_j is misspecified.[9] Here, we further extend this approach by allowing SNPs within different genomic annotations to have a specific contribution to ID. We propose the following model:

$$E[\beta_j] = \frac{b_0}{M} + \sum_{k=1}^{K}\left(z_{jk}\left[\frac{b_k}{m_k}\right] + (1-z_{jk})\left[\frac{b-b_k}{M-m_k}\right]\right), \qquad \text{(Equation 2)}$$

where b denotes the genome-wide ID parameter, $b_0 = -(K-1)b$, z_{jk}. the indicator of membership of SNP j to annotation k ($k = 1\ldots K$), b_k the contribution of annotation k to ID, and m_k. ($1 \leq m_k \leq M$)he number of SNPs in annotation k. We hereafter denote $\pi_k = m_k/M$ as the proportion of SNPs in annotation k. Under the null hypothesis that each SNP contributes equally to ID, each genomic annotation is expected to contribute proportionally to the number of SNPs it contains. Therefore, we can define the enrichment of ID in annotation k, hereter denoted δ_k, as the ratio between the contribution of annotation k to ID that is b_k over the expected contribution of annotation k to ID that is $\pi_k b$, i.e., $\delta_k = b_k/(\pi_k b)$.

Combining Equation 1 and Equation 2 leads to (Appendix A)

$$y = bF_g + \sum_{k=1}^{K}\left(\frac{b_k}{\pi_k} - b\right)\left[\left(\frac{\pi_k}{1-\pi_k}\right)\left(\overline{F}_k - F_g\right)\right] + e, \qquad \text{(Equation 3)}$$

where $\overline{F}_k$ denotes the average inbreeding coefficient across SNPs in annotation k, F_g denotes the average inbreeding coefficient across all M SNPs, and e the residual term from Equation 1.

Estimation of model parameters

Equation 3 implies that $\tau_k = b_k/\pi_k - b$ can be directly estimated by performing a multivariate regression of y onto F_g and the $(\Delta F_k = \pi_k(\overline{F}_k - F_g)/(1-\pi_k))$ terms. Therefore, enrichment of ID in annotation k can be detected with standard regression p values quantifying the statistical significance of τ_k. We use the latter approach in all analyses.

It is important to note, as with methods for partitioning heritability,[22] that certain combinations of annotations cannot be fitted jointly with Equation 3. Indeed, in some cases, overlap between annotations can induce a strong collinearity between inbreeding measures, which can make estimates unstable. For example, when the annotation is defined by chromosome number, fitting F_g as well as all chromosome-specific inbreeding measures will lead to identifiability issues. In the latter case, a simple solution is to choose one class of the annotation as a reference (e.g., chromosome 1) and therefore remove it from the model.

One of the challenges of our model is to provide estimates that can be interpreted in terms of enrichment (or depletion) of ID. However, enrichment is often intuitively conceptualized on a multiplicative scale, whereas the inference in our model is done on an additive scale (linear regression). Parameters τ_k and $\delta_k = 1 + \tau_k/b$ measure the same information, but τ_k is on an additive scale and is directly estimated, while δ_k is on a multiplicative scale (the scaled used for interpretation) and derived from estimated parameters. Previous studies aiming at partitioning additive genetic variance (e.g., Finucane et al.[22]) faced similar challenges and addressed them in the same way by using two scales for their model parameters. However, the parameterization chosen in our study leads to a much simpler relationship between τ_k and δ_k, thereby allowing for the use of the p value of δ_k, which is readily available from fitting the model to test the significance of δ_k for τ_k.

Average ID enrichment across traits

The significance of the average ID enrichment across traits (i.e., $\overline{\delta}_k$) is determined by that of the average estimate of τ_k's across traits, hereafter denoted $\overline{\tau}_k$. Given that $\overline{\tau}_k$ is a linear combination of ordinary-least-squares (OLS) estimators, its asymptotic distribution is also Gaussian and its sampling variance defined as

$$\mathrm{var}\left(\overline{\tau}_k\right) = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\rho_{t,t'}\left[\mathrm{var}(\tau_{k,t})\mathrm{var}(\tau_{k,t'})\right]^{1/2}, \qquad \text{(Equation 4)}$$

where $\mathrm{var}(\tau_{k,t})$ is the sampling variance of the OLS estimator of τ_k for trait t and $\rho_{t,t'}$ the correlation between OLS estimators of $\tau_{k,t}$ and $\tau_{k,t'}$. In practice, we propose to approximate $\rho_{t,t'}$ with the sample correlation between traits, which is a valid approximation given that OLS estimators are linear transformations of the phenotypes. Therefore, we test the significance of $\overline{\tau}_k$ by using the following test statistic, $\overline{\tau}_k^2/\mathrm{var}(\overline{\tau}_k)$, which follows a chi-square

distribution with 1 degree of freedom under the null. Finally, we consistently define the average ID enrichment across traits as $\overline{\delta}_k = 1 + \overline{\tau}_k/\overline{b}$, where $\overline{b}$ is the average ID across traits.

ID enrichment within continuous annotations

We analyzed continuous genomic annotations defined by non-negative values (e.g., recombination rate or posterior causal probability). We scaled these annotations by the largest value observed across the genome such that each SNP is assigned a value $\tilde{z}_{jk}$ between 0 and 1. We then defined $\tilde{\pi}_k$ as the mean of the $\tilde{z}_{jk}$'s across the genome (which is equivalent to the proportion of annotated SNPs when genomic annotations are binary variables) and used that value as our reference to calculate enrichments of ID for continuous annotations. Note that the use of another statistic from the distribution of $\tilde{z}_{jk}$'s (e.g., the median) would not change statistical significance but will influence the magnitude of the estimated ID enrichment.

When reporting results for continuous annotations, we used terms such as "high" or "low" (e.g., "high recombination rates" or "low nucleotide diversity") only to indicate the direction of the effect but not to imply that the annotation was discretized in any way.

Depletion of ID enrichment

In some cases, we observed that estimates of τ_k can lead to negative values of the enrichment statistic δ_k, which are difficult to interpret. We refer to this situation as "depletion of ID signal" and report instead the estimate $\breve{\delta}_k = 1 - \pi_k(\delta_k - 1)/(1 - \pi_k)$ corresponding to a transformed annotation defined by (1-z_{jk}) or (1-$\tilde{z}_{jk}$).

Inbreeding measures

To maximize power to detect ID enrichment across a wide range of causal allele frequencies, we use two inbreeding measures in our analyses. The first is the correlation between uniting gametes[1,2] (F_{UNI}) and the second is the proportion, F_{ROH}, of SNPs contained in >1.5 Mb-long runs. Note that our definition of F_{ROH} is slightly different from the widely used definition of F_{ROH}, which is the proportion of an individual genome that is covered by their ROHs. We chose this alternative definition because it offers more flexibility for partitioning analyses while still retaining the same amount of information (correlation > 0.99; Figure S1). F_{UNI} was previously shown to be more powerful to detect ID caused by well-tagged causal variants (e.g., SNPs with a minor allele frequency [MAF] > 1%),[9,25] while F_{ROH} can perform better when alleles causing ID are rarer and therefore poorly imputed via current imputation reference panels and methods.[12,26] We re-assessed the relative fraction of ID captured with either inbreeding measure and show that common and rare alleles both contribute to ID and that their relative contribution is trait specific (Figure 1).

For ROH-based analyses, the null hypothesis is defined such that each SNP has an equal probability of being covered by identical-by-descent genomic segments, which we approximate in this study by using long ROHs. We show in the supplemental methods section that violation of this assumption because of non-uniform genomic distribution of ROHs has little impact on our results.

We used PLINK 1.9 with the –ibc command to estimate F_{UNI} for each UKB participant. For continuous annotations, we used a custom C++ program (see data and code availability and web resources) that calculates weighted sums of per-SNP estimates of F_{UNI} with weights proportional to the value of the annotation. For binary annotations, we calculated F_{UNI} by only using SNPs assigned to those annotations. ROHs were called with the following PLINK command: –maf 0.05–homozyg–homozyg-density 50–homozyg-gap 1000–homozyg-kb 1500–homozyg-snp 50–homozyg-window-het 1–homozyg-window-missing 5–homozyg-window-snp 50. We then used 19,476,620 imputed SNPs with MAF > 0.1%, which were annotated to 187 functional categories (web resources) and defined F_{ROH} as the proportion of these SNPs contained in a ROH. Similarly, annotation-specific inbreeding measures were calculated as the (weighted) proportion of SNPs assigned to a given annotation that is contained in a ROH.

SNP genotyping

We used allele counts at 44,741,804 SNPs genotyped and imputed in 487,409 participants of the UK Biobank[27,28] (UKB). Although an extensive description of our dataset was given previously,[29] we briefly summarize the main steps of data preparation. We identified 456,414 UKB participants of European ancestry by using projected principal components based on sequenced participants of the 1000 Genomes Projects with known ancestry.[29] We then restricted our analyses to a subset of these participants that contains 348,501 conventionally unrelated participants, i.e., whose estimated pairwise SNP-based genomic relationships are <0.05. Genomic relationships were estimated with 1,124,803 common (MAF ≥ 1%) HapMap3[30] SNPs via GCTA (v.1.9).[31] Imputed SNPs included in our F_{UNI}-based analyses were selected on the basis of the following criteria: MAF > 1%, p value from the Hardy-Weinberg equilibrium test > 10^{-6}, and imputation quality r^2_{INFO} statistic > 0.3. After quality control, 9,326,198 imputed SNPs were included in our analyses. Genotyped SNPs used to call ROH were selected, as previously described,[11] via the following criteria: missingness rate < 1%, MAF > 5%, and Hardy-Weinberg equilibrium test p value > 0.0001. Quality control was performed with PLINK 1.9[32] (v.1.90b6.13 64 bit from November 30, 2019).

Association between levels of inbreeding and phenotypes measured in the UKB

We tested the association between F_{UNI} or F_{ROH} and traits by using linear regression adjusted for age at recruitment (UKB field 21022-0.0), sex, assessment center (UKB field 54-0.0), genotyping chip and batch, year of birth (UKB field 34-0.0), year of birth square, and the top ten genetic principal components calculated via PLINK 2.0. For each trait, we excluded phenotypic values larger than 4 standard deviations and then pre-adjusted trait values for the covariates listed above. Residuals obtained from these pre-adjustment analyses were then inverse normally transformed and used as focal traits. UKB identifiers for our 11 focal traits are height (UKB field 50-0.0), hip-to-waist ratio (ratio of UKB field 49-0.0 over UKB field 48-0.0), handgrip strength (average of UKB fields 46-0.0 and 47-0.0), lung function measured as the peak expiratory flow (UKB field 3064-0.0), visual acuity measured on log MAR scale (average between UKB field 5201-0.0 and UKB field 5208-0.0), auditory acuity measured as the speech reception threshold (average between UKB field 20019-0.0 and UKB field 20021-0.0), number of years of education (EA), fluid intelligence score (UKB field 20016-0.0), cognitive function measured as the mean time to correctly identify picture card matches (UKB field 20023-0.0), fertility measured as the number of children (for males UKB field 2405-0.0 and for females UKB field 2734-0.0), and number of diseases diagnosed estimated as the number International Classification of Diseases, Tenth Revision (ICD10) codes reported for UKB participants (release prior to December 2020). The North West Multi-Centre Research Ethics Committee (MREC) approved the study, and all participants in the UKB study analyzed here provided written informed consent.

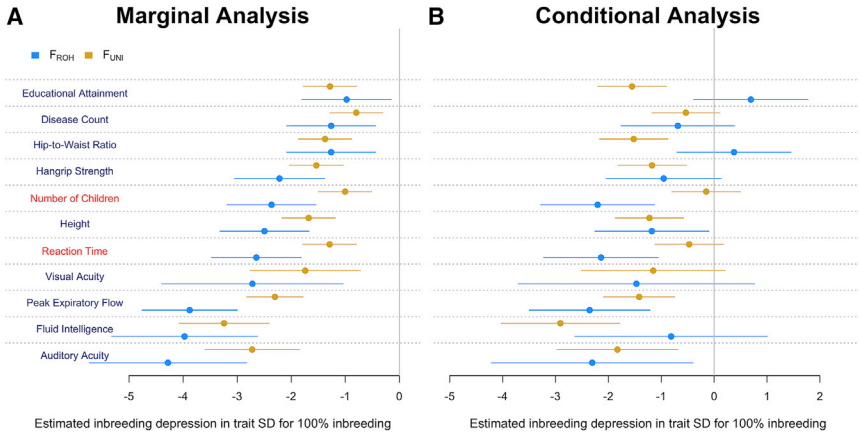

Figure 1. Partitioning of inbreeding depression (ID) between F_{ROH} and F_{UNI}
Left shows marginal analyses where ID is estimated with either F_{ROH} and F_{UNI}. Right shows estimates of ID obtained from fitting F_{ROH} and F_{UNI} jointly. Error bars represents 95% confidence intervals. Highlighted in red font: two traits (number of children and reaction time) for which ID is mostly captured by F_{ROH}. SD, standard deviation.

Genomic annotations used in individual-level analyses

We used genomic annotations previously compiled and processed by Finucane et al.,[22] Gazal et al.,[23] and Hujoel et al.[24] The 44 genomic annotations analyzed in our study were derived from 187 annotations downloaded from the LD score regression repository (baselineLF_v2.2.UKB model; web resources).

Binary annotations include putative evolutionary old enhancers and promoters (referred to in Hujoel et al.[24] as ancient sequence age: × 2 annotations); flanking bivalent transcription start sites or enhancers from the Roadmap Epigenomics Project;[33] coding; intronic; promoter; 3′ UTR and 5′ UTR genomic regions from the UCSC Genome Browser[34] (post-processed by Gusev et al.[35]); synonymous and non-synonymous; conserved genomic regions across mammals, primates, and vertebrates (phastCons 46-way[24]); CTCF (annotation from Hoffman et al.[36]); digital genomic footprint (DGF data from ENCODE[37] and post-processed by Gusev et al.[35]); DNase-I hypersensitive sites (ENCODE and Roadmap Epigenomics data post-processed by Trynka et al.[38] and also merged with fetal DHS annotation); enhancers (merged × 2 annotations: from Andersson et al.[39] and Hoffman et al.[36]); histone marks (merged H3K27ac annotations from Hnisz et al.[40] and Kundaje et al.[33] post-processed by PGC2,[41] H3K4me1,[38] H3K4me3,[38] and H3K9ac[38]); enhancers and promoters from Villar et al.;[42] promoters of loss-of-function-intolerant genes from ExAC (annotation from Hujoel et al.[24]); promoter flanking;[36] repressed genomic regions;[36] super enhancers;[40] transcription factor binding sites;[37] transcribed genomic regions;[36] transcription start sites;[36] weak enhancers;[36] and >4 rejected substitution from GERP++ score.[43]

Continuous annotations include maximum posterior probability from fine-mapping of molecular QTL[44] (expression, methylation, H3K4me1, and H3K27ac), background selection statistic (McVicker *B* statistic),[45] predicted allele age,[46] time to most recent common ancestor,[47] nucleotide diversity within 10 kb, recombination rate (within 10 kb) based on the Oxford recombination map,[48] CpG content within 50 kb, GERP++ score (number of substitution; referred to as GERP NS), and number of species sharing a given putative enhancer.[24] We also analyzed LD score with 1 Mb chi-square association test-statistic with a given trait or level of heterozygosity (i.e., MAF × (1 − MAF)) as continuous annotations.

Enrichment of heritability and ID from GWAS summary statistics

Enrichment of ID in trait-associated genomic regions

We first performed a standard GWAS of each of the 11 traits included in our analyses. We tested association between SNPs and traits by using an additive model (hereafter referred to as additive GWAS) where allele counts are directly correlated with traits. We then used resulting chi-square association test statistics as a continuous annotation and estimated its corresponding ID enrichment by using the method described above.

Estimation of heritability and ID enrichment via stratified LD score regression

We quantified the enrichment of SNP-based heritability in the 44 genomic annotations described above by using stratified LD score regression[22] (SLDSC). In brief, the SLDSC method, as implemented in this study, was based upon the following regression model:

$$E\left[\chi_j^2\right] = 1 + Na_{h^2} + N\sum_{k=1}^{K}\theta_{k,h^2}\left(\frac{\ell_{jk}}{M_k}\right), \qquad \text{(Equation 5)}$$

where a_{h^2} measures the level of confounding (e.g., due to uncorrected population stratification) in the GWAS, χ_j^2 the chi-square association statistic of SNP *j*, N the GWAS sample size, θ_{k,h^2} the effect size of annotation k on trait heritability (h^2), ℓ_{jk} (the annotation-weighted LD score), and M_k defined as

$$\ell_{jk} = \sum_{i=1}^{M} z_{ik} r_{ij}^2 \text{ and } M_k = \sum_{l=1}^{M} z_{ik}, \qquad \text{(Equation 6)}$$

where r_{ij}^2 is the squared correlation of allele counts between SNP *i* and SNP *j*.

We used publicly available annotation-weighted LD scores of 1,190,321 sequenced HapMap 3 SNPs from Finucane et al.,[22] Gazal et al.,[23] and Hujoel et al.[24] (web resources), which were calculated against a larger set of 9,997,231 sequenced SNPs with MAF > 0.5% in a European ancestry sample from the 1000 Genomes Project. LD scores of duplicated annotations were averaged. We analyzed each annotation independently and included the standard LD score ($\ell_{j0} = \sum_{i=1}^{M} r_{ij}^2$) and ten MAF-class-specific LD scores (i.e., LD score with SNPs within a given MAF class only) as covariates.[23] Parameters were estimated via weighted least-squares with weights proportional to $1/\ell_{j0}$.

Next, we re-tested the association between SNPs and traits by using a model fitting both allele counts and the indicator of heterozygosity, which estimates dominance deviation at each SNP. We refer to the latter analysis as additive-dominance GWAS. We show in Appendix B that the LD score regression methodology can be extended to estimate ID from summary statistics of an additive-dominance GWAS and further extend it below to quantify enrichment of ID in genomic annotations by using the following regression model:

$$E\left[-Z_{d,j}\right] = a_b + \sqrt{N}\sum_{k=1}^{K}\theta_{k,b}\left(\frac{\ell_{jk}}{M_k}\right), \qquad \text{(Equation 7)}$$

where $Z_{d,j}$ is the Z scores of estimates of dominance deviation at SNP j, a_b the intercept term measuring confounding effects, and $\theta_{k,b}$ the effect size of annotation k on ID. Standard errors for θ_{k,h^2}, $\theta_{k,b}$ as well as their correlation across multiple annotations were estimated with a block-jackknife (leave one block out) procedure based on 240 ~10 Mb-long genomic segments used as blocks.

Results

The relative contribution of rare variants to ID varies across human traits

Clark et al.[16] previously claimed that ID is predominantly caused by rare, recessive variants made homozygous in long ROH and not by homozygosity at variants that are in LD with common SNPs. To reach that conclusion, their study compared the magnitude of conditional estimates of ID from a bivariate linear regression model fitting F_{UNI} (referred to as F_{GRM} in their study) and F_{ROH} jointly. Clark et al. restricted their calculation of F_{UNI} to SNPs with an MAF $>$ 5% across diverse studies in which participants were genotyped with different arrays. We show that the same analysis performed in a homogeneous and large sample like the UKB leads to a different and trait-specific conclusion regarding the contribution of rare variants to ID. As described in Clark et al., we calculated both F_{UNI} and F_{ROH} by using 301,412 quality-controlled genotyped SNPs (as described in Yengo et al.[11]) with MAF $>$ 5% and also used the same definition of ROH. For this analysis, F_{ROH} was defined as in Clark et al., i.e., the cumulated length of ROH in Mb divided by 3,000.

We found that the relative fraction of ID captured by each inbreeding measure varies between traits (Figure 1). For example, the estimated effects of F_{UNI} on the number of children and on the mean time to identify matches (reaction time) both became not statistically significant ($p < 0.05$) once conditioned on F_{ROH}. This observation is consistent with Clark et al.'s claims and suggests a larger relative contribution of rare variants to ID in these traits. However, we find that ID in hip-to-waist ratio as well as on fluid intelligence is mostly captured by F_{UNI}. Finally, for height and peak expiratory flow, the conditional effects F_{ROH} and F_{UNI} are of similar magnitudes, suggesting that both low frequency variants and more common partially recessive alleles are causal of ID in these traits. On average across traits, we found that the conditional effects of F_{UNI} and F_{ROH} were not statistically different from each another. Altogether, our results imply that alleles causing ID span a larger spectrum of the frequencies (from rare to common, e.g., >5%) and that the relative contribution of rare variants is trait specific.

Enrichment of ID within functional genomic annotations

We analyzed 44 genomic annotations (including 12 continuous annotations; material and methods), which cover ~98% of the autosome (Figure S2). In total, our analysis involves 11 traits, 44 annotations, and 2 inbreeding measures (i.e., $11 \times 44 \times 2 = 968$ trait-annotation-measure triplets). Therefore, statistical significance accounting for multiple testing (Bonferroni correction) was set to $0.05/968 \approx 5.2 \times 10^{-5}$. We also quantified the average ID enrichment across traits, in which case statistical significance was set to $0.05/(44 \times 2) \approx 5.7 \times 10^{-4}$. Note that Bonferroni correction is most likely too conservative because the overlap between genomic annotations (Figure S3) induces a positive correlation across statistical tests. Therefore, we also report enrichments detected at a false discovery rate (FDR) threshold of 5%.

On average across traits, we detected significant enrichment of ID within 8 annotations (Figure 2; Table S1). Importantly, the largest enrichment was detected with both inbreeding measures within genomic regions with high recombination rates ($\overline{\delta}_{F_{ROH}}$~33.5, $p = 3.2 \times 10^{-4}$; and $\overline{\delta}_{F_{UNI}}$~21.9, $p = 1.5 \times 10^{-6}$). Other significant enrichments were detected within conserved genomic regions across mammals ($\overline{\delta}_{F_{UNI}}$~20.9, $p = 7.8 \times 10^{-5}$) and primates ($\overline{\delta}_{F_{UNI}}$~19.1, $p = 1.9 \times 10^{-4}$); DNase I hypersensitive regions (DHS: $\overline{\delta}_{F_{UNI}}$~5.7, $p = 2.3 \times 10^{-7}$); chromatin accessible regions identified through digital genomic foot printing (DGF: $\overline{\delta}_{F_{UNI}}$~5.3, $p = 2.1 \times 10^{-5}$); regions with large GERP++ score,[43] which measures the strength of purifying selection at a locus ($\overline{\delta}_{F_{UNI}}$~3.9, $p = 4.5 \times 10^{-6}$); genomic regions with low nucleotide diversity ($\overline{\delta}_{F_{ROH}}$~3.0, $p = 1.8 \times 10^{-4}$); and H3K4me1 histone marks ($\overline{\delta}_{F_{UNI}}$~1.9, $p = 4.3 \times 10^{-5}$).

Previous studies have shown that nucleotide diversity and linkage disequilibrium (LD) can influence ROH detection.[49,50] Therefore, errors in ROH calling due to these two annotations could potentially confound our results. Using data from the UKB and from an independent sample from the UK10K Project,[51] we show in the supplemental methods section (Figures S4–S6) that ROH genomic density as well as errors in ROH calling cannot explain the enrichment of ID in low nucleotide diversity regions nor that in high recombination rate regions. Importantly, ID enrichment in high recombination rate regions is also detected via F_{UNI}, which minimizes the likelihood of this finding's being caused by errors in ROH calling. However, we found that exclusion of the major histocompatibility complex (MHC) locus (hg19: chr6: 25,000,000–35,000,000) from our analyses has a strong impact on the significance of $\overline{\delta}_{F_{ROH}}$ in low nucleotide diversity regions (without MHC: $\overline{\delta}_{F_{ROH}}$~2.0, $p = 0.29$; Table S8). Given that ROH genomic density alone cannot explain that observation, we further investigated whether population stratification of rare variants may confound our results. The latter hypothesis is justified by the fact that enrichment in low nucleotide diversity is detected with F_{ROH}, but not with F_{UNI}, which suggests a signal (or a confounding effect) coming from low frequency variants. To test this hypothesis, we used birth coordinates as proxies for geographically localized rare variants and fitted those as covariates in our analyses. To account for non-linear effects of rare variants stratification, we used a *k-means* clustering approach to create 100 geographical clusters from participants' birth

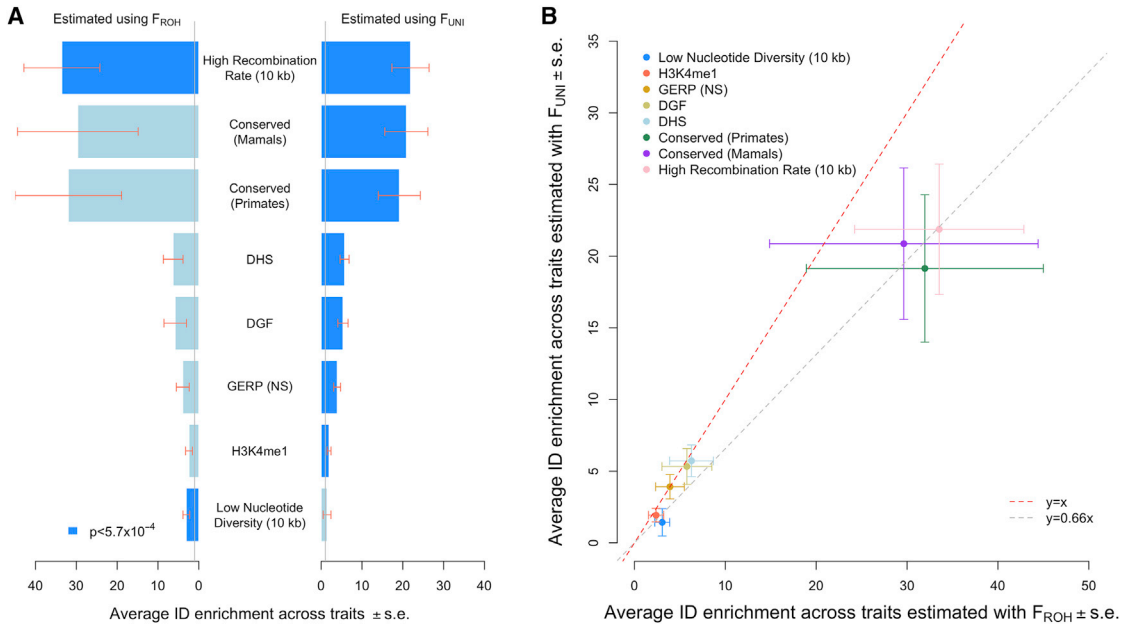

Figure 2. Significant inbreeding depression (ID) enrichment within eight genomic annotations
(A and B) Estimates of ID enrichment on average across 11 traits obtained via F_{UNI} and F_{ROH}. Statistical significance was set at $p < 0.05/(44 \times 2) \approx 5.7 \times 10^{-4}$. Recombination rate and nucleotide diversity were analyzed as continuous annotations. "High recombination rate" denotes that recombination rate is positively correlated with ID, and "low nucleotide diversity" denotes that nucleotide diversity is negatively correlated with ID. Data underlying this figure are reported in Table S1. DHS, DNA-se I hypersensitive sites; DGF, chromatin accessible regions from digital genomic footprint; GERP (NS), GERP++ score number of substitution; H3K4Me1, histone mark (binary annotation). Error bars represent standard errors (SEs).

coordinates and fitted those as categorical covariates. Although these additional analyses were based on a slightly reduced number of participants who reported their birth location in the UK, we still found a significant enrichment of ID within low nucleotide diversity regions (adjusted for 20 PCs + 100 birth coordinates clusters and including MHC: $\overline{\delta}_{F_{ROH}} \sim 3.3$, $p = 1.4 \times 10^{-4}$), consistent with our main analysis. Altogether, our results show that the MHC locus disproportionately contributes to the enrichment of ID in low nucleotide diversity regions, which is not explained by higher ROH frequency at this locus nor by uncorrected population stratification.

Recognizing that the Bonferonni correction is conservative, at FDR < 5%, we detected 16 additional annotations, which include other regulatory elements (e.g., transcription factor binding sites: $\overline{\delta}_{F_{UNI}} \sim 4.0$) and conservation annotations (e.g., promoters containing evolutionary old DNA sequences: $\overline{\delta}_{F_{UNI}} \sim 27.0$) as shown in Table S1.

Next, we assessed the independence between significantly enriched annotations by fitting them jointly. On average across traits, only ID enrichment within genomic regions with low nucleotide diversity remained significant at our initial Bonferroni threshold ($\overline{\delta}_{F_{ROH}} \sim 3.1$, $p = 3.0 \times 10^{-4}$; Table S2).

Moreover, we detected significant enrichment of ID in height within transcribed genomic regions ($\delta_{F_{ROH}} \sim 8.0$, $p = 4.9 \times 10^{-5}$) as well as in genomic regions with a low nucleotide diversity ($\delta_{F_{ROH}} \sim 7.6$, $p = 4.8 \times 10^{-9}$) and those with a shorter time to most recent common ancestor (TMRCA: $\delta_{F_{ROH}} \sim 5.8$, $p = 7.6 \times 10^{-6}$). Note that alleles with a shorter TMRCA are more likely to have arisen recently in the population. Therefore, the latter result suggests a larger contribution of recent alleles to ID, consistent with previous findings.[7,52] Importantly, these three genomic annotations were detected with F_{ROH} but did not pass statistical significance when quantified via F_{UNI} ($p > 0.27$), which suggests that rare variants within (or with large values of) these annotations could be the main drivers of our results for height. Interestingly, we also detected a significant depletion of ID signal (material and methods) for height within transcription start sites (enrichment outside TSS: $\delta_{F_{ROH}} \sim 2.0$, $p = 3.2 \times 10^{-5}$; Table S3). Furthermore, we detected suggestive ID enrichment at FDR < 5% for the majority of traits except for fertility and visual and auditory acuity (Table S3). Among these additional associations, we highlight the strong enrichment of ID in height and cognitive function within promoters of loss-of-function-intolerant genes from the Exome Aggregation Consortium[53] (>84 and $p < 0.001$ with both inbreeding measures; Table S3). We also highlight suggestive enrichment of ID in peak expiratory flow within DGF ($\delta_{UNI} \sim 7.2$, $p = 5.5 \times 10^{-5}$), where ~13% common SNPs contribute ~90% of the genome-wide ID for that trait.

ID enrichment within GWAS-associated genomic regions

Our analyses based on F_{UNI} show interesting similarities between functional enrichment of ID and that of additive

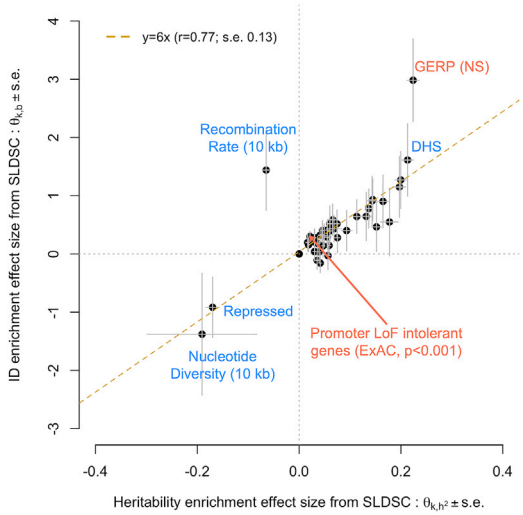

Figure 3. Correlation between functional enrichment of SNP-based heritability and inbreeding depression (ID) estimated from GWAS summary statistics

Enrichment statistics for heritability (x axis: θ_{k,h^2}) and ID (y axis: $\theta_{k,b}$) were estimated with stratified LD score regression (SLDSC) as described in the material and methods section. Enrichments statistics were averaged over 11 traits. Each dot represents a genomic annotation, and errors bars represent standard errors (SEs) calculated with a block-jackknife procedure. Annotations highlighted in coral font show a statistically significant ID enrichment at $p < 0.05/88$. DHS, DNA-se I hypersensitive sites; GERP (NS), GERP++ score number of substitution; LoF, loss of function. Data underlying this figure are reported in Table S5.

genetic variance as reported previously.[22] In both cases, we see a strong enrichment of ID and SNP-based heritability within genomic regions with conserved functions across species as well as within regulatory elements such as promoters, enhancers, and DHSs and a much smaller enrichment within transcriptionally repressed or transcribed regions.

To further explore these similarities, we sought to directly quantify ID enrichment within trait-associated loci identified through GWASs. Therefore, we performed GWASs of all 11 traits in our sample of ~350,000 unrelated UKB participants (material and methods) and used chi-square association statistics of each SNP to define a new continuous annotation. Thus, we calculated 11 trait-specific inbreeding measures derived from F_{UNI} such that each SNP is weighted proportionally to the strength of its association with the target trait. We then tested the ID enrichment of these 11 GWAS-derived annotations. Given that stronger association is expected at SNPs with large LD scores,[54] each analysis was also conditioned on another continuous annotation defined by the LD score of each SNP. On average across traits, we found marginally significant evidence that genomic regions with large chi-square association statistics, independently of LD, are enriched for ID signal ($\bar{\delta}_{F_{\mathrm{UNI}}} \sim 10.6$, $p = 0.03$; Table S4).

In addition, we sought to estimate the correlation between functional enrichment of heritability and that of ID across the 44 genomic annotations analyzed previously. To ensure a fair comparison, we used stratified LD score regression (SLDSC) as a unified framework to assess enrichment of SNP-based heritability and that of ID. We denote θ_{k,h^2} and $\theta_{k,b}$ as the SLDSC effect size of annotation k on heritability and ID, respectively. We provide a theoretical justification for using LD score regression to estimate ID by highlighting its direct connection with F_{UNI} (Appendix B). Consistently, the mean correlation between Z scores of τ_k (i.e., our individual-level data-based statistic to quantify enrichment of ID: $\tau_k = b(\delta_k - 1)$; material and methods) and $\theta_{k,b}$ is ~0.7 (range: 0.5 up to 0.8) across traits, which shows that both enrichment measures largely capture the same information (Figure S7).

On average across traits, we find a significant positive correlation between θ_{k,h^2} and $\theta_{k,b}$of ~0.77 with a block jack-knife standard error (b.s.e.) of ~0.13 (Figure 3). We also detect a significant ($p < 0.05/11$) correlation between θ_{k,h^2} and $\theta_{k,b}$ for specific traits (Figure S8; Table S6), such as handgrip strength ($r = 0.91$; b.s.e. = 0.07), cognitive function ($r = 0.82$; b.s.e. = 0.14), and peak expiratory flow ($r = 0.93$; b.s.e. = 0.05). We show in the supplemental methods section and Figure S9, through theory and simulations, that a large positive correlation between enrichments of heritability and ID in fitness could be expected if the distribution of fitness effects of mutations causing ID is moderately skewed, i.e., if the proportion of mutations with large fitness effect is small. However, given that this distribution is largely unknown, our theoretical results therefore only provide a set of sufficient, but not necessary, conditions that can explain our observations. Furthermore, we performed additional simulations to show that the correlation between enrichments of ID and heritability is not due to an artifact in our methods, which would systematically induce such a correlation (supplemental methods, Figure S10).

Despite an overall large correlation between enrichments of heritability and ID, we find that the effect of recombination rate on ID and on heritability are in opposite directions (Figure 3). To further explore what could explain this observation, we performed extensive forward-time evolutionary simulations by using SLiM3[55] (Supplemental Methods) to quantify the effect of recombination rate on both heritability and ID enrichments. In these simulations, we considered various combinations of selection (s) and dominance (h) coefficients for mutations causing ID and additive genetic variance in fitness (material and methods). Using 1,000 simulation replicates (for each scenario), we find that recombination rate affects the enrichment of heritability and ID in a non-linear fashion that is modulated by the strength of both selection and dominance (Figures S11 and S12). More specifically, our simulations show a strong enrichment of ID in high recombination rate regions (>2 cM/Mb) only when ID is caused by partially recessive ($h = 0.1$) near-neutral mutations, i.e., when s is between $\sim 1/(2N_e)$ and $\sim 2/N_e$, where N_e denotes the effective population size. Under the same scenarios, heritability was either slightly enriched or even depleted in high recombination regions.

However, for large selection coefficients ($s > 2/N_e$) and large dominance coefficients ($h > 0.3$), enrichments of heritability and ID behave consistently, and both monotonically decrease with recombination rate.

In summary, our findings imply that functional enrichments of ID and additive genetic variance in the population are most likely caused by similar mechanisms, such as variable recombination rates between genomic regions and disruption of evolutionary old regulatory elements (e.g., older than the split between marsupial and placental mammals ~160 million years ago).[56] However, although co-localized, variants causing these two phenomena may only partially overlap, while variants causing ID may tend to be more recessive than those causing additive genetic variance.

Discussion

In this study, we introduce a method to detect and quantify the enrichment of ID within a wide range of binary and continuous genomic annotations and applied it to analyze 11 traits with prior evidence of ID. We show that ID is independently enriched within regions with high recombination rates, consistent with prior evidence in potatoes,[57] and also, to a lesser extent, within regions with a low nucleotide diversity, a marker of high functional importance of DNA sequences.[58,59] Interestingly, we find that the MHC locus contributes disproportionately to the enrichment of ID in low nucleotide diversity regions, even when analyses are adjusted for population stratification (genetic principal components + geographical birth coordinates). An enrichment of ID in low nucleotide diversity regions and in high recombination rate regions may seem counterintuitive given the negative correlation between these two annotations ($r \sim -0.2$; Figure S3), which translates the fact that higher recombination rates are associated with higher nucleotide diversity. However, high recombination rates also contribute to minimize the strength of background selection[60] (i.e., purging of neutral and near-neutral alleles in LD with deleterious mutations) and selective sweeps, thereby contributing to increase diversity of slightly deleterious alleles.[61] Consistent with reduced selection efficiency, recombination rate and strength of background selection annotations show a larger negative correlation ($r \sim -0.3$; Figure S3). Another potential explanation for the enrichment of ID within genomic regions with high recombination rates is because frequency of deleterious *de novo* mutation is increased in those regions.[62,63] However, the relatively small variance in *de novo* mutations explained by recombination rate ($R^2 \sim 5\%$, in Kessler et al.[62]) implies that variation in mutation rates most likely contributes a second order effect to our observations.

Using estimated coalescence times, we confirm that recent alleles contribute more to ID and also show that homozygosity in DNA sequences that are conserved across mammal and primate species contribute ~20-fold more ID than any random equally sized portion of the genome. In humans, inbreeding between close relatives often manifests as severe Mendelian syndromes caused by homozygous loss-of-function mutations.[8] Consistent with this observation, our study shows a large ID enrichment ($\bar{\delta} > 30$; Table S1) within promoters of loss-of-function-intolerant genes. It is noteworthy that ID enrichment within coding regions was comparatively smaller ($\bar{\delta}$ between 1 and 7.1, $p > 0.15$; Table S1) and did not survive any of our corrections for multiple testing. The latter observation emphasizes that impairment of essential genes, as opposed to genes in general, is a key contributor to ID. Finally, our study shows a strong ID enrichment within multiple regulatory elements, which opens interesting avenues to quantify the effect of homozygosity on gene and protein expression in humans.

The genetic basis of ID is predicated on two hypotheses.[1] The first one, the "dominance hypothesis," stipulates that ID is caused by directional dominance effects of partially recessive deleterious alleles. On the other hand, the "overdominance hypothesis" assumes that the genetic architecture of traits subjected to ID is such that heterozygotes express better phenotypes than both homozygotes, for example due to balancing selection. Although major progress has been made in model species,[1] disentangling these two explanations remains a challenge in natural populations. One possibility to address it could be by leveraging the distinct evolutionary consequences of these two hypotheses. The "dominance hypothesis" predicts a larger contribution to ID from loci where mutation rate is higher than average, while the "overdominance hypothesis" predicts a larger ID from loci where balancing selection is stronger.[64] Our study shows patterns of ID enrichment that are consistent with both hypotheses. However, we found a much larger magnitude of ID enrichment from annotations that are correlated with mutation rates (e.g., recombination rates[62,63] or CpG content;[23] Table S1) as compared to those correlated with strength of selection (e.g., GERP++ score or background selection; Table S1). Altogether, our results imply that both hypotheses probably contribute to explain ID in human traits, but overdominance does so to a lesser extent.

We also revisited previous claims[7] regarding the contribution of common variants to ID. We provide new evidence that variants with an MAF > 1% substantially contribute to ID in height and educational attainment (Figure 1), while their contribution to ID in fertility remains limited. While some of the traits analyzed in this study have been shown to drive assortative mating in the population, we do not think that assortment based on polygenic traits (e.g., height) would implicitly create an enrichment of ID within any specific part of the genome, as reported here. In fact, the increased homozygosity at trait-associated loci that is induced by assortative mating is inversely proportional to the number of causal variants and therefore negligible for highly polygenic traits.[65,66]

Finally, our results also contribute to illuminating the direction of causality between homozygosity and traits. Indeed, differential contribution of genomic regions to ID would not be expected if traits of parents (e.g., intelligence or socio-economic status) were causally associated with the phenotypes in their children and only incidentally associated with their relatedness, i.e., with offspring homozygosity. Note that the latter configuration has been shown to induce biases in estimates of ID, as suggested previously.[67]

Our study has a number of limitations. First, we lacked statistical power to detect more trait-specific patterns of ID enrichment beyond height. Similarly, we also lacked power to estimate conditional enrichment across multiple functional annotations. Therefore, estimates of ID enrichment reported here may reflect overlap between annotations (e.g., the correlation between DHS and DGF is ~0.5; Figure S3), including ones not observed in our study. A second limitation of our study is that estimates of ID obtained via our extension of the LD score regression method have too large standard errors to be efficiently used for partitioning analyses (Table S7; Figure S13). However, we emphasize that such methods remain valid and can still be used when sample size is sufficiently large (e.g., n ~ 1,000,000, which is becoming more common in human studies).

A third limitation of our study is that our method assumes a linear relationship between ID and continuous annotations (Equation 2), while this relationship may be more complex in reality. For example, our forward-time evolutionary simulations showed a non-linear relationship between the recombination rate and enrichment of ID (and that of additive genetic variance), which is modulated by the distribution of selection coefficients of deleterious alleles, as well as their dominance coefficients (Figures S11 and S12). Note that such a non-linear relationship has also been reported in previous studies[23,68] that investigated the functional enrichment of trait heritability, although these studies did not emphasize that observation. As an alternative to assuming a linear relationship between continuous annotations and ID, we recommend discretizing the annotations into quantiles or testing the significance of the conditional effects of the annotation and its squared value. We show in Figure S14 that the empirical relationship between ID enrichment and recombination rate is largely monotonic, which suggests that the linear assumption made here would have a limited impact on our conclusions regarding this annotation.

In conclusion, we have proposed in this study a refined characterization of the functional effects of variants contributing to ID. Beyond conceptual parallels between estimation of ID and that of heritability established previously,[9] our study demonstrates that functional mechanisms involved in ID are also relevant for characterizing genetic variance and genetic risk in the population. We foresee that the application of the different methods developed here to large collections of whole-genome or whole-exome sequencing data available in the near future will lead to major discoveries regarding the genetic basis of ID.

Appendix A: Derivation of Equation 3

Equation 3 is derived from combining Equation 1 and Equation 2 recalled below:

$$y = \sum_{j=1}^{M} \beta_j F_j + e, \qquad \text{(Equation A1)}$$

$$E[\beta_j] = \frac{b_0}{M} + \sum_{k=1}^{K} \left(z_{jk} \left[\frac{b_k}{m_k} \right] + (1 - z_{jk}) \left[\frac{b - b_k}{M - m_k} \right] \right). \qquad \text{(Equation A2)}$$

Note that Equation 1 is the same as Equation A1 and that Equation 2 the same as Equation A2.

We write $\beta_j = E[\beta_j] + \varepsilon_j$ then replace $E[\beta_j]$ with its expression from Equation 2. We get

$$y = \sum_{j=1}^{M} \beta_j F_j + e = \sum_{j=1}^{M} E[\beta_j] F_j + e', \qquad \text{(Equation A3)}$$

where $e' = e + \sum_{j=1}^{M} \varepsilon_j F_j$ and such as $E[e'] = 0$.

Besides, we can rewrite Equation 2 from the main text (methods section) as

$$E[\beta_j] = \frac{b(1-K)}{M} + \sum_{k=1}^{K} \frac{b - b_k}{M - m_k} + \sum_{k=1}^{K} z_{jk} \left[\frac{b_k}{m_k} - \frac{b - b_k}{M - m_k} \right],$$

$$E[\beta_j] = \frac{1}{M} \left[b(1-K) + \sum_{k=1}^{K} \frac{M(b - b_k)}{M - m_k} \right] + \frac{1}{m_k} \sum_{k=1}^{K} z_{jk} \left[b_k - \frac{m_k}{M - m_k} (b - b_k) \right],$$

$$E[\beta_j] = \frac{1}{M} \left[b(1-K) + \sum_{k=1}^{K} \left(\frac{b - b_k}{1 - \pi_k} \right) \right] + \frac{1}{m_k} \sum_{k=1}^{K} z_{jk} \left[b_k - \frac{\pi_k}{1 - \pi_k} (b - b_k) \right].$$

$$E[\beta_j] = \frac{1}{M} \left[b(1-K) + \sum_{k=1}^{K} \left(\frac{b - b_k}{1 - \pi_k} \right) \right] + \frac{1}{m_k} \sum_{k=1}^{K} z_{jk} \left[\frac{b_k - \pi_k b}{1 - \pi_k} \right]. \qquad \text{(Equation A4)}$$

where $\pi_k = m_k / M$. If we denote $\overline{F}_k$ and F_g as

$$\overline{F}_k = \frac{1}{m_k} \sum_{k=1}^{M} z_{jk} F_j \text{ and } F_g = \frac{1}{M} \sum_{k=1}^{K} F_j$$

Then, combining Equation A3 and Equation A4 leads to

$$y = \left[b(1-K) + \sum_{k=1}^{K} \left(\frac{b - b_k}{1 - \pi_k} \right) \right] F_g + \sum_{k=1}^{K} \left[\frac{b_k - \pi_k b}{1 - \pi_k} \right] \overline{F}_k + e',$$

$$= \left[b(1-K) + \sum_{k=1}^{K} \left(\frac{b-b_k}{1-\pi_k} \right) \right] F_g + \sum_{k=1}^{K} \left[\frac{b_k - \pi_k b}{1-\pi_k} \right] \times \left(\overline{F}_k - F_g + F_g \right) + e',$$

$$= \left[b(1-K) + \sum_{k=1}^{K} \left(\frac{b-b_k}{1-\pi_k} \right) + \sum_{k=1}^{K} \left(\frac{b_k - \pi_k b}{1-\pi_k} \right) \right] F_g + \sum_{k=1}^{K} \left(\frac{b_k}{\pi_k} - b \right) \left[\left(\frac{\pi_k}{1-\pi_k} \right) \left(\overline{F}_k - F_g \right) \right] + e',$$

$$= bF_g + \sum_{k=1}^{K} \left(\frac{b_k}{\pi_k} - b \right) \left[\left(\frac{\pi_k}{1-\pi_k} \right) \left(\overline{F}_k - F_g \right) \right] + e'.$$

(Equation A5)

QED.

Appendix B: Enrichment of ID from additive-dominance GWAS summary statistics

We assume the following model

$$y = \sum_{j=1}^{M} \left[\frac{x_j - 2p_j}{\sqrt{2p_j(1-p_j)}} \right] \beta_j^{(a)} + \left[\frac{-x_j^2 + (1+2p_j)x_j - 2p_j^2}{2p_j(1-p_j)} \right] \beta_j^{(d)} + e,$$

(Equation B1)

where x_j denotes the minor allele counts ($x_j = 0$, 1, or 2) at SNP j, p_j the minor allele frequency in the population, $\beta_j^{(a)}$ and $\beta_j^{(d)}$ additive and dominance effects at SNP j, respectively, and e a residual term. This model is the same as in Zhu et al.[69] (Equation 3) and Yengo et al. (Equations S6 and S7).[9] Zhu et al. further assumed that $E[\beta_j^{(a)}] = E[\beta_j^{(d)}] = 0$ and that $\mathrm{var}[\beta_j^{(a)}] = h_a^2/M$, $\mathrm{var}[\beta_j^{(d)}] = h_d^2/M$, and $\mathrm{var}[\beta_j^{(a)}] = 1 - h_a^2 - h_d^2$. However, in order to account for ID, Equation B1 must be generalized to model the directionality of dominance effects. For that, we assume $E[\beta_j^{(d)}] = -b/M$, where b is the genome-wide ID.

Now we denote $z_j = (x_j - 2p_j)/\sqrt{2p_j(1-p_j)}$ and $F_j = [x_j^2 - (1+2p_j)x_j - 2p_j^2]/[2p_j(1-p_j)]$. We note that F_j is exactly F_{UNI} at SNP j. Therefore, Equation B1 can be rewritten as

$$y = \sum_{j=1}^{M} z_j \beta_j^{(a)} - F_j \beta_j^{(d)} + e = b \left(\frac{1}{M} \sum_{j=1}^{M} F_j \right) + g_a + g_d + e,$$

(Equation B2)

where $E[g_a] = E[g_d] = 0$.

In an additive-dominance GWAS, the association between SNP i and y is tested by fitting simultaneously both additive (a_i) and dominance (d_i) effects, i.e., by fitting the following model

$$y = a_i x_i + d_i H_i + \text{ residual},$$

(Equation B3)

where $H_i = x_i(2 - x_i)$ is the indicator of heterozygosity at SNP i. Under the model above, we define $\widehat{d}$ as the ordinary least-squares estimator of d_i and $\mathrm{SE}(\widehat{d}_i)$ as its standard error. We then denote the Z score of $\widehat{d}_i$ as $Z_{d,i} = \widehat{d}_i/\mathrm{SE}(\widehat{d}_i)$and shows that it verifies (proof is given below)

$$E\left[-Z_{d,i} \Big/ \sqrt{N} \right] = E\left[\widehat{B}_i \right] = b \ \mathrm{cov}\left(F_i, \frac{1}{M} \sum_{j=1}^{M} F_j \right) \approx \left(\frac{b}{M} \right) \ell_i,$$

(Equation B4)

where ℓ_i is the LD score of SNP i and N is the GWAS sample size. Note that Equation B4 is analogous to the main result underlying the LD score regression methodology, which we recall below in Equation B5:

$$E\left[(\chi_i^2 - 1) / N \right] \approx \left(\frac{h^2}{M} \right) \ell_i.$$

(Equation B5)

Equation B4 and Equation B5 imply that the mean of the $\widehat{B}_i$'s divided by the mean LD score over a given set of SNPs is in fact a measure of the "per-SNP" ID just like the mean χ^2 (minus 1) divided by the mean LD score is a measure of the per-SNP heritability. Therefore, $E_k(h^2)$ and $E_k(b)$ quantify the relative per-SNP heritability and ID for SNPs within annotation k as compared with the rest of the genome.

In principle, Equation B4 can be further extended to estimate enrichment of ID as done with the stratified LD score regression methodology. However, we found the standard errors of estimates of ID based on Equation B4 to be ~1.8-fold larger than that of estimates obtained from individual-level data (Table S7), which substantially reduces statistical power to detect enrichments. Note that estimates from LD score regression with an intercept constrained to 0 have slightly lower standard errors but also suffer large biases as shown in Table S7.

Proof of Equation B4

The proof of Equation B4 relies on two arguments. The first one is to note that under classical linear regression theory, the ordinary least-squares estimate $(\widehat{a}_i, \widehat{d}_i)$ of (a_i, d_i) in Equation B3 verifies that

$$E\begin{pmatrix} \widehat{a}_i \\ \widehat{d}_i \end{pmatrix} = \begin{pmatrix} \mathrm{var}(x_i) & \mathrm{cov}(x_i, H_i) \\ \mathrm{cov}(x_i, H_i) & \mathrm{var}(H_i) \end{pmatrix}^{-1} \begin{bmatrix} \mathrm{cov}(x_i, y) \\ \mathrm{cov}(H_i, y) \end{bmatrix}.$$

In particular for $\widehat{d}$, we can write that

$$E\left[\widehat{d}_i \right] = \frac{\mathrm{var}(x_i)\mathrm{cov}(H_i, y) - \mathrm{cov}(x_i, H_i)\mathrm{cov}(x_i, y)}{\mathrm{var}(x_i)\mathrm{var}(H_i) - \mathrm{cov}(x_i, H_i)^2} = \mathrm{cov}\left(\frac{\mathrm{var}(x_i)H_i - \mathrm{cov}(x_i, H_i)x_i}{\mathrm{var}(x_i)\mathrm{var}(H_i) - \mathrm{cov}(x_i, H_i)^2}, \mathrm{y} \right).$$

Similarly, we derive the standard error of $\widehat{d}_i$ (under the assumption that the phenotypic variance is equal to 1 and that each SNP explains a negligible part of the trait variance) as

$$\mathrm{SE}\left[\widehat{d}_i\right] \approx \frac{\mathrm{SD}(y)}{h_i\sqrt{N}} = \frac{1}{\left[h_i\sqrt{N}\right]},$$

where $\mathrm{SD}(y)$ denotes the standard deviation of y.

If we denote p as the minor allele frequency, then under Hardy-Weinberg equilibrium, we can show that $\mathrm{var}(x_i) = h_i = 2p_i(1-p_i)$, $\mathrm{var}(H_i) = h_i(1-h_i)$, and $\mathrm{cov}(x_i, H_i) = h_i(1-2p_i)$. Using these relationships, we can further show that $\mathrm{var}(x_i)\mathrm{var}(H_i) - \mathrm{cov}(x_i, H_i)^2 = h_i^3$ and $\mathrm{var}(x_i)H_i - \mathrm{cov}(x_i, H_i)x_i = -h_i^2 F_i$. Therefore,

$$E\left[\widehat{d}_i\right] = \mathrm{cov}\left(\frac{-F_i}{h_i}, \mathrm{y}\right)$$

or, equivalently, $E[-(\widehat{d}_i/\mathrm{SE}[\widehat{d}_i])/\sqrt{N}] = E[-Z_{d,i}/\sqrt{N}] \approx \mathrm{cov}(F_i, \mathrm{y})$. Now using Equation B2, we can subsequently write that

$$E\left[-\frac{Z_{d,i}}{\sqrt{N}}\right] = \mathrm{cov}\left[F_i, b\left(\frac{1}{M}\sum_{j=1}^{M} F_j\right) + g_a + g_d + e\right]$$

$$= \frac{b}{M}\sum_{j=1}^{M}\mathrm{cov}(F_i, F_j) + \mathrm{cov}(F_i, g_a + g_d + e).$$

Under the assumption that SNP effects are independent of genotypes and that e (environment) is independent of F_i, we have that $\mathrm{cov}(F_i, g_a + g_d + e) = 0$. Moreover, Yengo et al.[9] previously showed that $\mathrm{cov}(F_i, F_j) \approx r_{ij}^2$, where r_{ij}^2 is the square LD correlation between SNP i and j. Hence,

$$E\left[-\frac{Z_{d,i}}{\sqrt{N}}\right] \approx \left(\frac{b}{M}\right)\sum_{j=1}^{M} r_{ij}^2 = \left(\frac{b}{M}\right)\ell_i.$$

Data and code availability

This study makes use of genotype and phenotype data from the UK Biobank data under project 12505. UKB data can be accessed upon request once a research project has been submitted and approved by the UKB committee. Data sources underlying all figures are provided as supplemental tables. Example R scripts from our pipeline to simulate and estimate enrichment of ID and a C++ code describing the algorithm utilized to calculate annotation-specific inbreeding measures are provided at the following URL: https://github.com/loic-yengo/Code_for_Genomic_Partitioning_of_Inbreeding_Depression.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.06.005.

Acknowledgments

This research was supported by the Australian Research Council (DE200100425, FT180100186, and FL180100072), the Australian National Health and Medical Research Council (1173790 and 1113400) and the National Institute of Health (MH100141). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding bodies. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This research has been conducted with the UK Biobank Resource under project 12505. The authors are grateful to Deborah Charlesworth and Brian Charlesworth for helpful discussions and comments on the manuscript.

Declaration of interests

The authors declare no competing interests.

Received: December 21, 2020
Accepted: June 1, 2021
Published: July 1, 2021

Web resources

Extended set of 189 functional annotations for variants with MAF>0.1% in the UK Biobank, https://storage.googleapis.com/broad-alkesgroup-public/LDSCORE/baselineLF_v2.2.UKB.tar.gz

GCTA software, https://cnsgenomics.com/software/gcta/

General description (README file) for all sets of functional annotations, https://alkesgroup.broadinstitute.org/LDSCORE/readme_baseline_versions

PLINK software, https://www.cog-genomics.org/plink2/

Set of 97 functional annotations from stratified LD score regression (baseline LD model), https://alkesgroup.broadinstitute.org/LDSCORE/baselineLD_v2.2_bedfiles.tgz

References

1. Charlesworth, D., and Willis, J.H. (2009). The genetics of inbreeding depression. Nat. Rev. Genet. *10*, 783–796.
2. Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., Li, W., Zhan, Q., Cheng, B., Xia, J., et al. (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nat. Commun. *6*, 6258.
3. Huisman, J., Kruuk, L.E.B., Ellis, P.A., Clutton-Brock, T., and Pemberton, J.M. (2016). Inbreeding depression across the lifespan in a wild mammal population. Proc. Natl. Acad. Sci. USA *113*, 3585–3590.
4. Pemberton, J.M., Ellis, P.E., Pilkington, J.G., and Bérénos, C. (2017). Inbreeding depression by environment interactions in a free-living mammal population. Heredity *118*, 64–77.
5. McQuillan, R., Eklund, N., Pirastu, N., Kuningas, M., McEvoy, B.P., Esko, T., Corre, T., Davies, G., Kaakinen, M., Lyytikäinen, L.P., et al.; ROHgen Consortium (2012). Evidence of inbreeding depression on human height. PLoS Genet. *8*, e1002655.
6. Joshi, P.K., Esko, T., Mattsson, H., Eklund, N., Gandin, I., Nutile, T., Jackson, A.U., Schurmann, C., Smith, A.V., Zhang, W., et al. (2015). Directional dominance on stature and cognition in diverse human populations. Nature *523*, 459–462.

7. Clark, D.W., Okada, Y., Moore, K.H.S., Mason, D., Pirastu, N., Gandin, I., Mattsson, H., Barnes, C.L.K., Lin, K., Zhao, J.H., et al. (2019). Associations of autozygosity with a broad range of human phenotypes. Nat. Commun. *10*, 4957.
8. Bittles, A.H., and Neel, J.V. (1994). The costs of human inbreeding and their implications for variations at the DNA level. Nat. Genet. *8*, 117–121.
9. Yengo, L., Zhu, Z., Wray, N.R., Weir, B.S., Yang, J., Robinson, M.R., and Visscher, P.M. (2017). Detection and quantification of inbreeding depression for complex traits from SNP data. Proc. Natl. Acad. Sci. USA *114*, 8602–8607.
10. Marshall, T.C., Coltman, D.W., Pemberton, J.M., Slate, J., Spalton, J.A., Guinness, F.E., Smith, J.A., Pilkington, J.G., and Clutton-Brock, T.H. (2002). Estimating the prevalence of inbreeding from incomplete pedigrees. Proc. Biol. Sci. *269*, 1533–1539.
11. Yengo, L., Wray, N.R., and Visscher, P.M. (2019). Extreme inbreeding in a European ancestry sample from the contemporary UK population. Nat. Commun. *10*, 3719.
12. Keller, M.C., Visscher, P.M., and Goddard, M.E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics *189*, 237–249.
13. Ceballos, F.C., Joshi, P.K., Clark, D.W., Ramsay, M., and Wilson, J.F. (2018). Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet. *19*, 220–234.
14. Keller, M.C., Simonson, M.A., Ripke, S., Neale, B.M., Gejman, P.V., Howrigan, D.P., Lee, S.H., Lencz, T., Levinson, D.F., Sullivan, P.F.; and Schizophrenia Psychiatric Genome-Wide Association Study Consortium (2012). Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. PLoS Genet. *8*, e1002656.
15. Christofidou, P., Nelson, C.P., Nikpay, M., Qu, L., Li, M., Loley, C., Debiec, R., Braund, P.S., Denniff, M., Charchar, F.J., et al. (2015). Runs of Homozygosity: Association with Coronary Artery Disease and Gene Expression in Monocytes and Macrophages. Am. J. Hum. Genet. *97*, 228–237.
16. Johnson, E.C., Bjelland, D.W., Howrigan, D.P., Abdellaoui, A., Breen, G., Borglum, A., Cichon, S., Degenhardt, F., Forstner, A.J., Frank, J., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2016). No Reliable Association between Runs of Homozygosity and Schizophrenia in a Well-Powered Replication Study. PLoS Genet. *12*, e1006343.
17. Pryce, J.E., Haile-Mariam, M., Goddard, M.E., and Hayes, B.J. (2014). Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genet. Sel. Evol. *46*, 71.
18. Ferenčaković, M., Sölkner, J., Kapš, M., and Curik, I. (2017). Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population. J. Dairy Sci. *100*, 4721–4730.
19. Ayroles, J.F., Hughes, K.A., Rowe, K.C., Reedy, M.M., Rodriguez-Zas, S.L., Drnevich, J.M., Cáceres, C.E., and Paige, K.N. (2009). A genomewide assessment of inbreeding depression: gene number, function, and mode of action. Conserv. Biol. *23*, 920–930.
20. García, C., Avila, V., Quesada, H., and Caballero, A. (2012). Gene-expression changes caused by inbreeding protect against inbreeding depression in Drosophila. Genetics *192*, 161–172.
21. García, C., Ávila, V., Quesada, H., and Caballero, A. (2013). Are transcriptional responses to inbreeding a functional response to alleviate inbreeding depression? Fly (Austin) *7*, 8–12.
22. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
23. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., and Price, A.L. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat. Genet. *49*, 1421–1427.
24. Hujoel, M.L.A., Gazal, S., Hormozdiari, F., van de Geijn, B., and Price, A.L. (2019). Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species. Am. J. Hum. Genet. *104*, 611–624.
25. Yengo, L., Zhu, Z., Wray, N.R., Weir, B.S., Yang, J., Robinson, M.R., and Visscher, P.M. (2018). Reply to Kardos et al.: Estimation of inbreeding depression from SNP data. Proc. Natl. Acad. Sci. USA *115*, E2494–E2495.
26. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat. Rev. Genet. *11*, 499–511.
27. Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P., Pell, J., et al. (2012). UK Biobank: Current status and what it means for epidemiology. Health Policy Technol. *1*, 123–126.
28. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
29. Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., Visscher, P.M.; and GIANT Consortium (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum. Mol. Genet. *27*, 3641–3649. https://doi.org/10.1093/hmg/ddy271.
30. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
31. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
32. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
33. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
34. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.
35. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E.,

et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
36. Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. *41*, 827–841.
37. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
38. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.
39. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.
40. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.
41. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
42. Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.
43. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.
44. Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H.K., Ju, C.J., Loh, P.R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al. (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nat. Genet. *50*, 1041–1047.
45. McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. *5*, e1000471.
46. Rasmussen, M.D., Hubisz, M.J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. PLoS Genet. *10*, e1004342.
47. Palamara, P.F., Terhorst, J., Song, Y.S., and Price, A.L. (2018). High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. Nat. Genet. *50*, 1311–1317.
48. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science *310*, 321–324.
49. Howrigan, D.P., Simonson, M.A., and Keller, M.C. (2011). Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. BMC Genomics *12*, 460.
50. Gazal, S., Sahbatou, M., Perdry, H., Letort, S., Génin, E., and Leutenegger, A.L. (2014). Inbreeding coefficient estimation with dense SNP data: comparison of strategies and application to HapMap III. Hum. Hered. *77*, 49–62.
51. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
52. Szpiech, Z.A., Xu, J., Pemberton, T.J., Peng, W., Zöllner, S., Rosenberg, N.A., and Li, J.Z. (2013). Long runs of homozygosity are enriched for deleterious variation. Am. J. Hum. Genet. *93*, 90–102.
53. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
54. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
55. Haller, B.C., and Messer, P.W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Mol. Biol. Evol. *36*, 632–637.
56. Luo, Z.-X., Yuan, C.-X., Meng, Q.-J., and Ji, Q. (2011). A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature *476*, 442–445.
57. Zhang, C., Wang, P., Tang, D., Yang, Z., Lu, F., Qi, J., Tawari, N.R., Shang, Y., Li, C., and Huang, S. (2019). The genetic basis of inbreeding depression in potato. Nat. Genet. *51*, 374–378.
58. Booker, T.R., and Keightley, P.D. (2018). Understanding the Factors That Shape Patterns of Nucleotide Diversity in the House Mouse Genome. Mol. Biol. Evol. *35*, 2971–2988.
59. Tatarinova, T.V., Chekalin, E., Nikolsky, Y., Bruskin, S., Chebotarov, D., McNally, K.L., and Alexandrov, N. (2016). Nucleotide diversity analysis highlights functionally important genomic regions. Sci. Rep. *6*, 35730.
60. Pouyet, F., Aeschbacher, S., Thiéry, A., and Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. eLife *7*, e36317.
61. Nordborg, M., Charlesworth, B., and Charlesworth, D. (1996). The effect of recombination on background selection. Genet. Res. *67*, 159–174.
62. Kessler, M.D., Loesch, D.P., Perry, J.A., Heard-Costa, N.L., Taliun, D., Cade, B.E., Wang, H., Daya, M., Ziniti, J., Datta, S., et al. (2020). De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. Proc. Natl. Acad. Sci. USA *117*, 2560–2569.
63. Halldorsson, B.V., Palsson, G., Stefansson, O.A., Jonsson, H., Hardarson, M.T., Eggertsson, H.P., Gunnarsson, B., Oddsson, A., Halldorsson, G.H., Zink, F., et al. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science *363*, eaau1043.
64. Lynch, M., and Walsh, B. (1998). Genetics and Analysis of Quantitative Traits (Sinauer).
65. Crow, J.F., and Felsenstein, J. (1982). The effect of assortative mating on the genetic composition of a population. Soc. Biol. *29*, 22–35.

66. Crow, J.F., and Kimura, M. (2009). An Introduction to Population Genetics Theory (Blackburn Press).
67. Johnson, E.C., Evans, L.M., and Keller, M.C. (2018). Relationships between estimated autozygosity and complex traits in the UK Biobank. PLoS Genet. *14*, e1007556.
68. Gazal, S., Loh, P.R., Finucane, H.K., Ganna, A., Schoech, A., Sunyaev, S., and Price, A.L. (2018). Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. Nat. Genet. *50*, 1600–1607.
69. Zhu, Z., Bakshi, A., Vinkhuyzen, A.A., Hemani, G., Lee, S.H., Nolte, I.M., van Vliet-Ostaptchouk, J.V., Snieder, H., Esko, T., Milani, L., et al.; LifeLines Cohort Study (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. Am. J. Hum. Genet. *96*, 377–385.

The American Journal of Human Genetics, Volume *108*

Supplemental Data

Genomic partitioning of inbreeding depression in humans

Loic Yengo, Jian Yang, Matthew C. Keller, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher

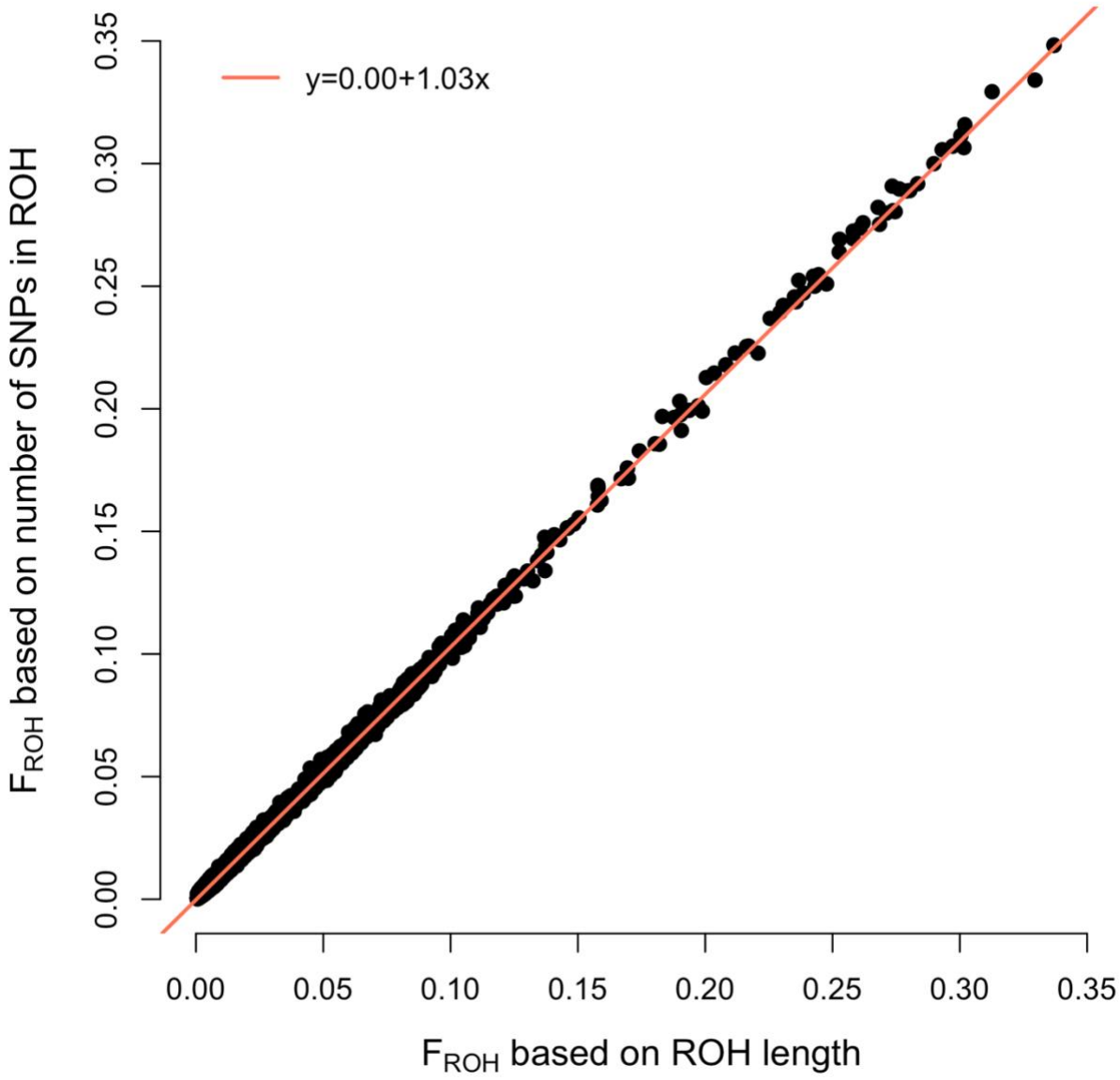

Figure S1. Consistency between two definitions of F_{ROH}. On the x-axis, F_{ROH} is the defined as the cumulated length of ROH in bp divided by 2,785,774,901 that is total autosomal length covered by genotyped SNPs. One the y-axis, F_{ROH} is defined as the proportion of 19,476,620 imputed SNPs with a minor allele frequency >0.1% with 187 functional annotations from previous studies (**URLs**). The correlation between these two definition is >0.99.

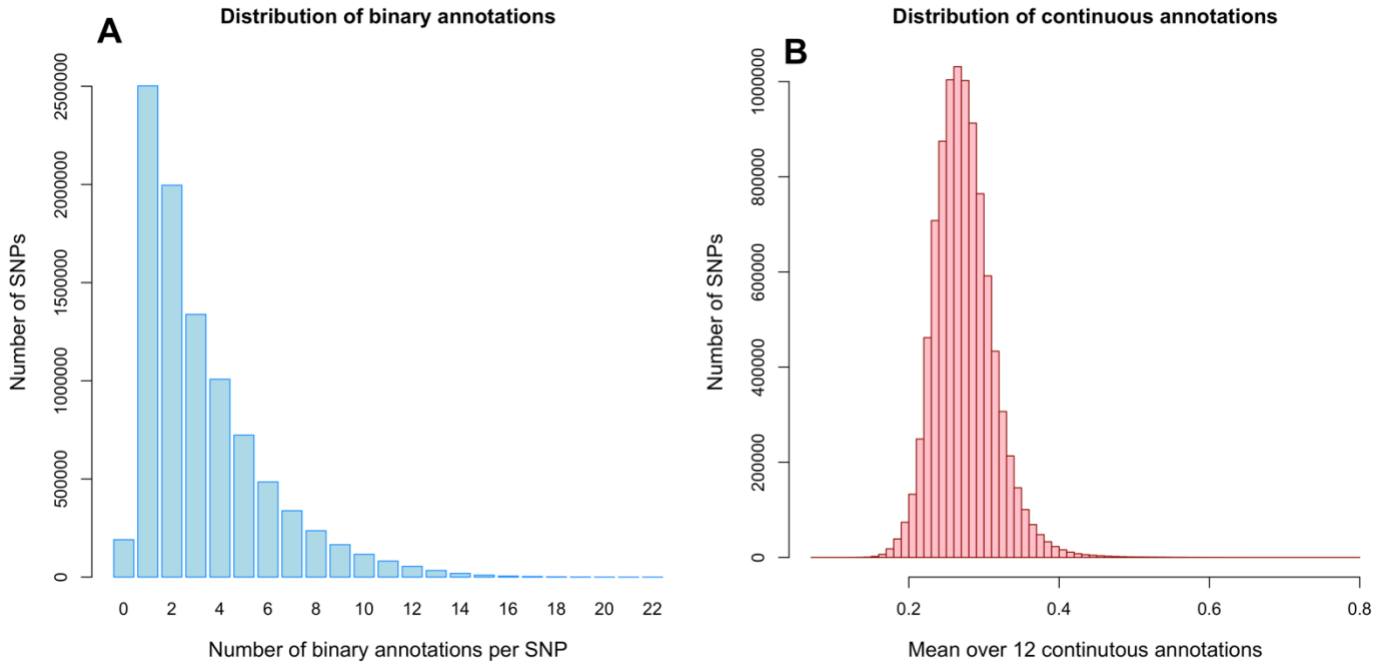

Figure S2. Distribution genomic annotations across SNPs. Left panel shows the histogram of number of binary annotations per SNPs (x-axis). 97.9% of SNPs have at least 1 binary annotation. Right panel shows the histogram across SNPs of the average of 12 normalised continuous annotation. Continuous annotations were normalised by scaling them with the largest value of the annotation across the entire genome such that normalised values range between 0 and 1. Mean and standard deviation of average continuous annotation are ~0.27 and ~0.04 respectively.

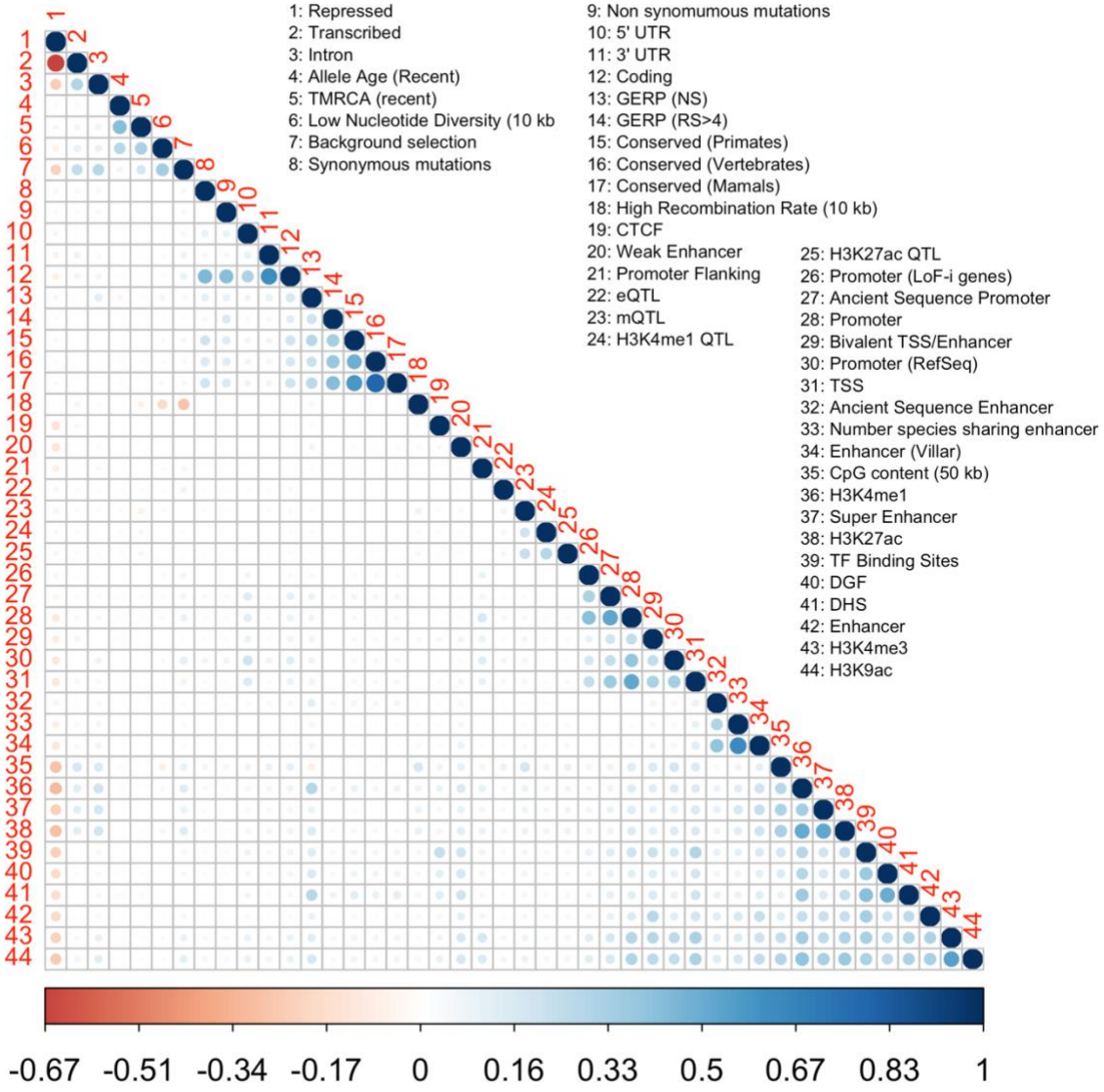

Figure S3. Correlation between genomic annotations estimated from SNPs assigned at least two (binary or continuous) annotations.

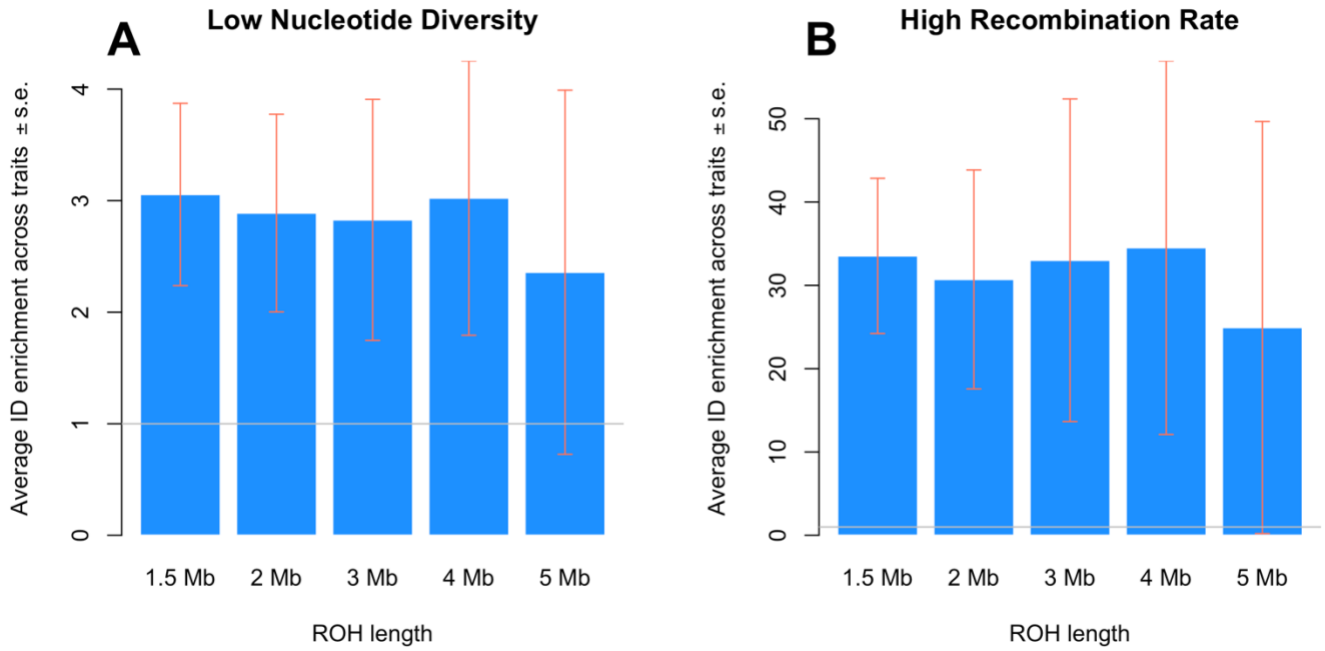

Figure S4. Average enrichment of ID across 11 traits as a function of ROH length. Panel **A** shows F_{ROH}-based estimates of ID enrichment within genomic regions with low nucleotide diversity. Nucleotide diversity was defined, for each SNP, as the mean diversity with 10 kb. Panel **B** shows F_{ROH}-based estimates of ID enrichment within genomic regions with high recombination rates. Recombination rates and nucleotide diversity was determined, for each SNP, as the mean recombination rate within 10 kb. Recombination rate and nucleotide diversity were analysed as continuous annotations. "High recombination rate" denotes that recombination rate is positively correlated with ID; and "Low nucleotide diversity" denotes that nucleotide diversity is negatively correlated with ID. Error bars are standard errors (s.e.).

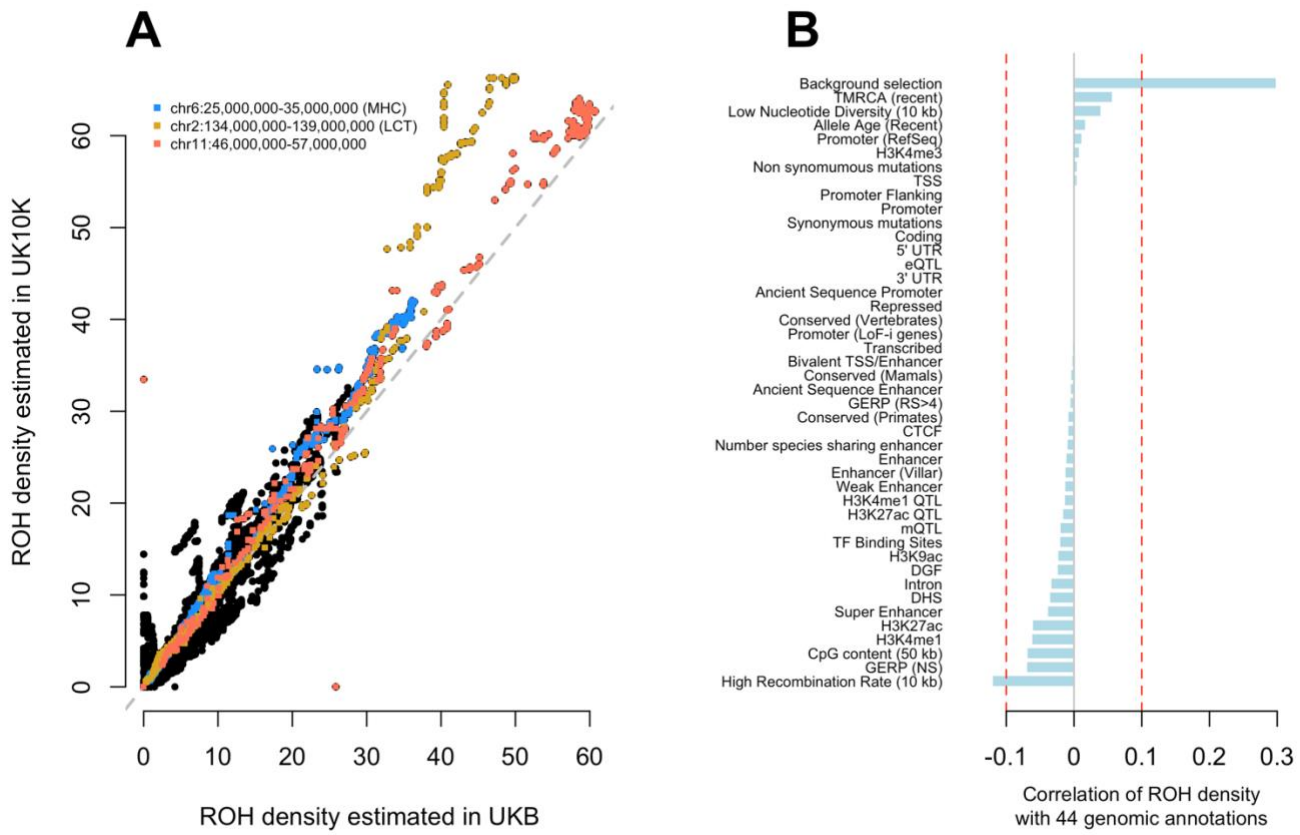

Figure S5. Genomic density of ROHs and its correlation with functional genomic annotations. Panel **A** shows consistent ROH density estimated in two independent samples from the UK: the UK Biobank (UKB; on the x-axis) and the UK10K sample (on the y-axis). Correlation of estimated ROH density from these two UK samples is >0.99 (jackknife standard error <0.001). In each sample, ROH density was estimated over 9,309,159 genomic positions by counting the number of ROHs overlaping that position. Values of ROH density shown on panel **a** are divided by the mean density in each sample. Panel **B** shows the correlation (x-axis) between ROH density (in the UKB) and 44 genomic annotations (y-axis). ROH density is most largely correlated with the McVicker *B* statistic measuring the strength of background selection. Recombination rate and nucleotide diversity were analysed as continuous annotations. "High recombination rate" denotes that recombination rate is positively correlated with ID; and "Low nucleotide diversity" denotes that nucleotide diversity is negatively correlated with ID.

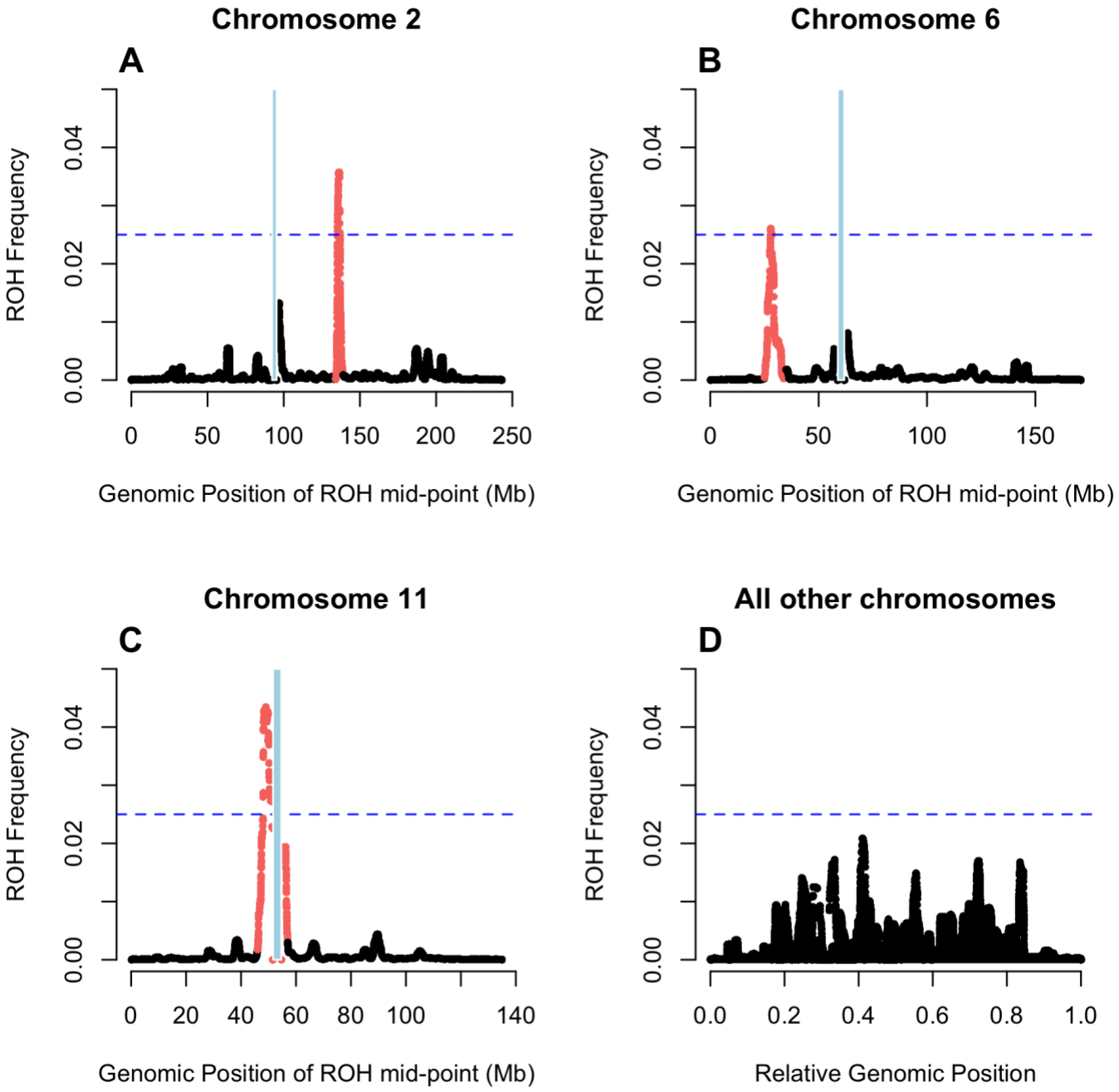

Figure S6. Genomic distribution of ROHs. ROH density was estimated in 456,414 European ancestry participants of the UK Biobank over 9,309,159 genomic positions by counting the number of ROHs overlapping that position. Values of ROH density shown on y-axes of all four panels are divided by the mean density in a sample. In panel **D**, relative genomic positions were calculated by dividing each genomic position (in base-pair unit) by the length of their corresponding chromosome. We highlight 3 genomic regions, where ROH frequency is >2.5% (i.e. >10 standard deviations above the mean ROH frequency across the genome): the Major Histocompatibility Complex (MHC) locus (hg19: chr6:25,000,000-35,000,000), the lactase locus (LCT; hg19:chr2:134,000,000-139,000,000) and the centromere region on chromosome 11 (hg19:chr11:46,000,000-57,000,000). The locations of the centromeres are depicted by a blue vertical line.

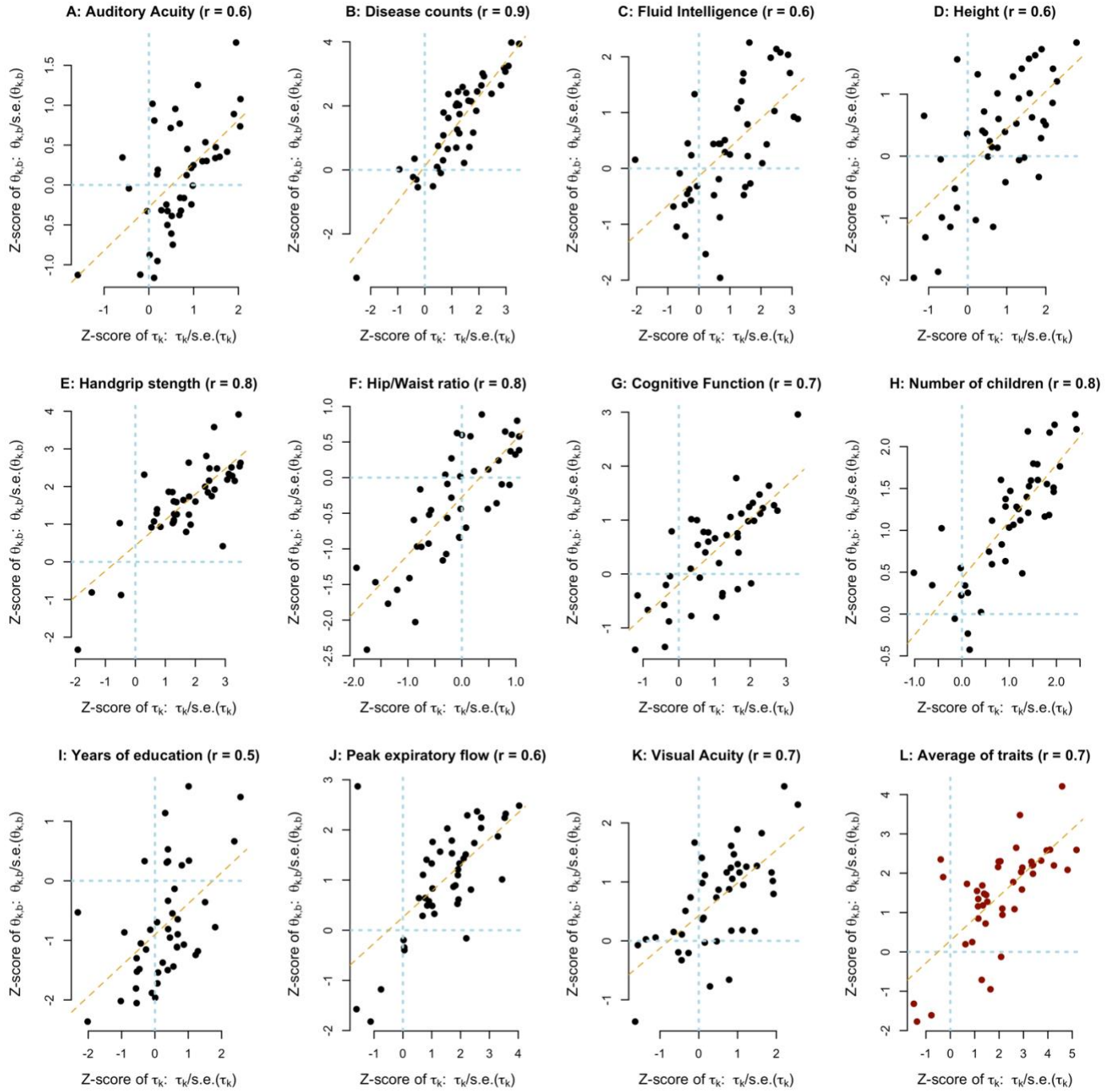

Figure S7. Correlation between individual-level data-based (τ_k) and GWAS-based ($\theta_{k,b}$) estimates of ID enrichment statistics for 11 traits and 44 functional annotations. Each panel represent a trait and the bottom right panel the average across traits. Within each panel, a dot present a genomic annotation. Correlation between enrichment measures for each trait is reported in the title of the figure (range of correlation (r): 0.5 to 0.9).

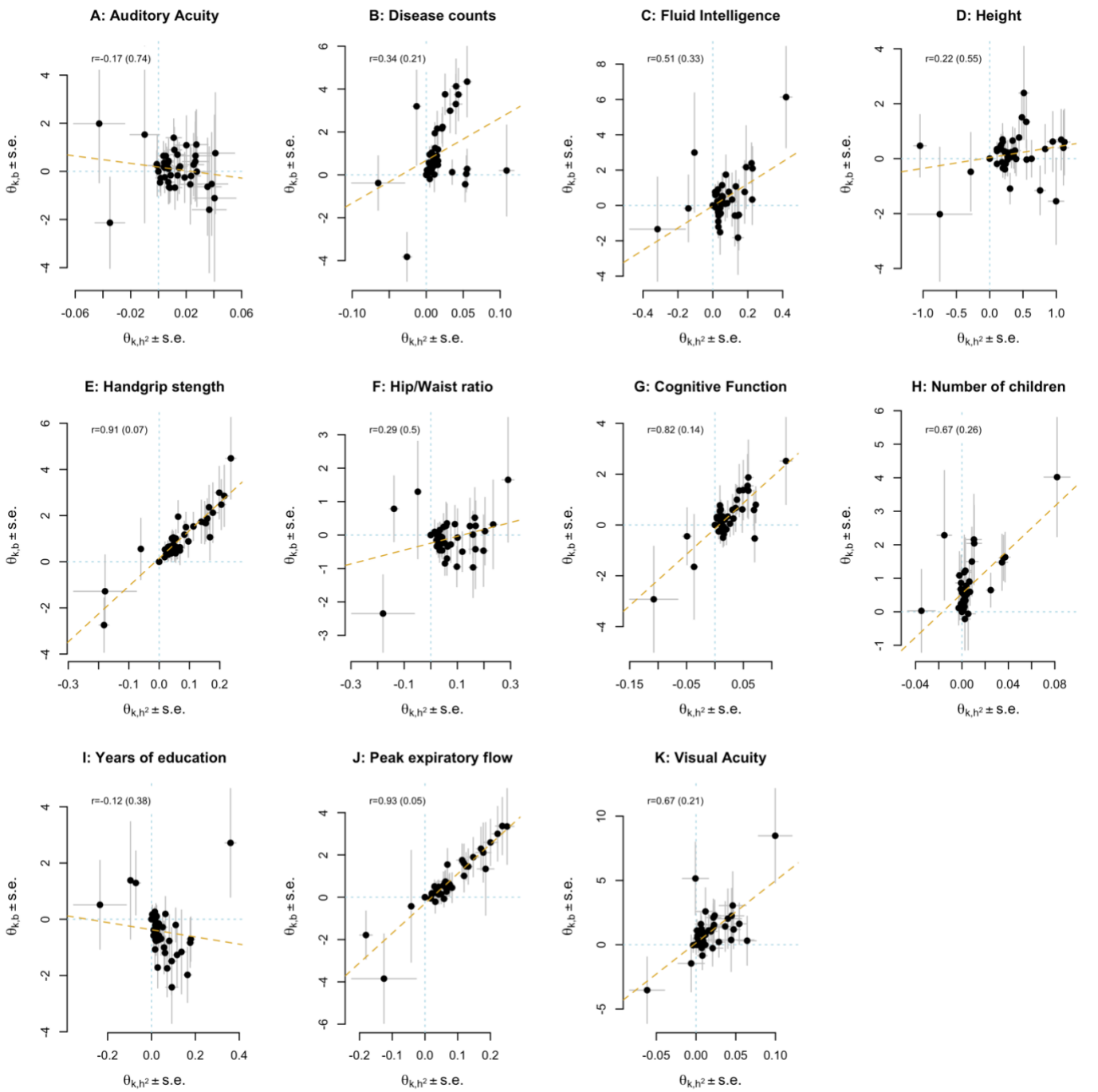

Figure S8. Correlation (r) between GWAS-based estimates of enrichment of heritability (using the θ_{k,h^2} statistic; x-axis) and ID (using the $\theta_{k,b}$ statistic; y-axis) across 44 genomic annotations and 11 traits associated with inbreeding. Enrichment statistics were estimated using stratified LD score regression (SLDSC) as described in the Methods section. For each trait, r is estimated over 44 pairs of enrichment statistics ($\theta_{k,h^2}, \theta_{k,b}$) and the corresponding standard error (shown in brackets) is obtained using block-jackknife. Error bars represent standard errors. Data for each trait is shown in a specific panel. Data underlying this figure are reported in **Table S6**.

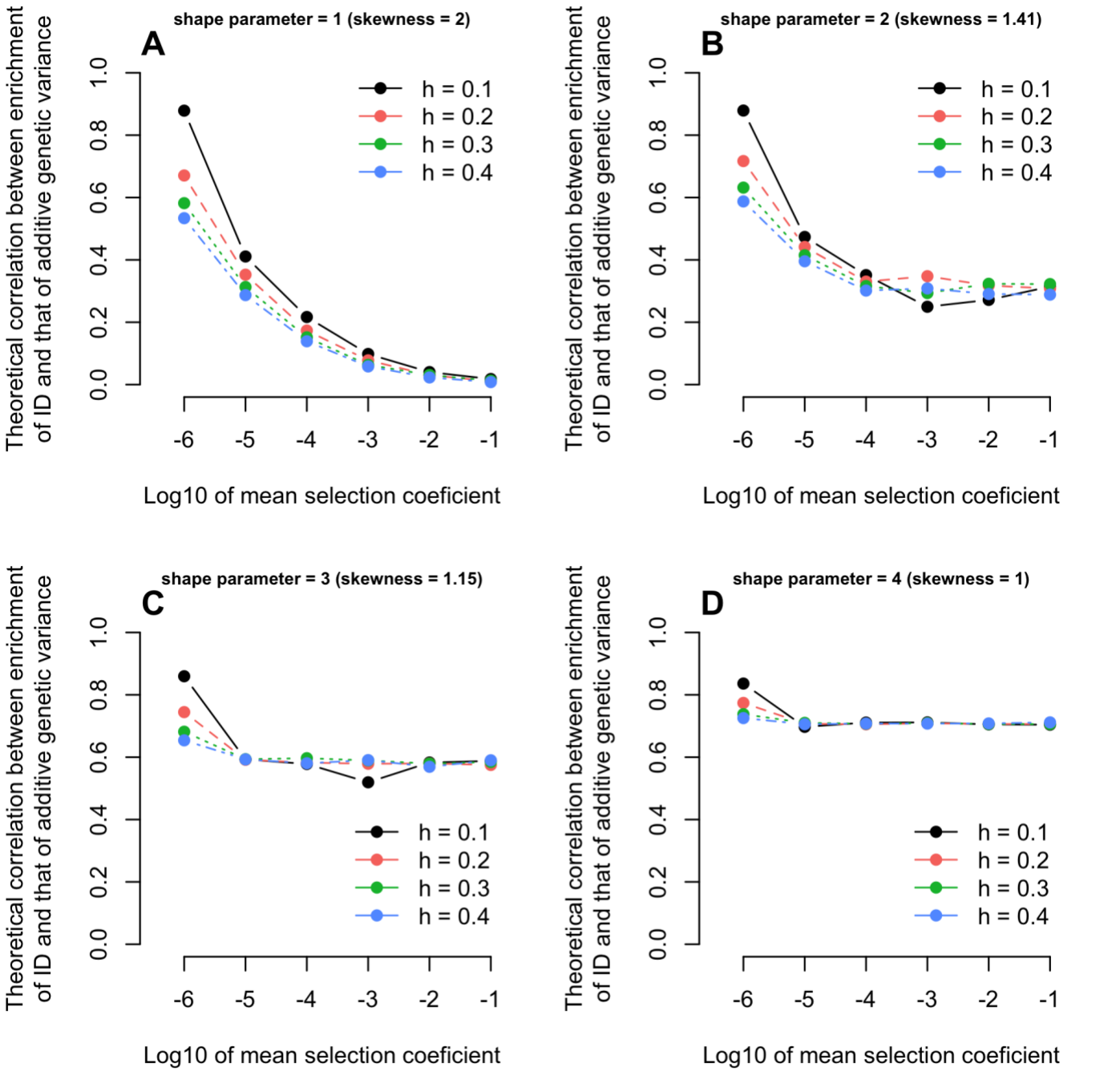

Figure S9. Expected correlation between enrichments of heritability and ID under various assumed distributions of selection and dominance coefficients of fitness mutations (**Supplemental Methods**). Selection coefficients were assumed to be Gamma-distributed with a mean varying between 10^{-6} (weak selection) and 10^{-1} (strong selection) and a shape parameter between 1 (strong skewness) and 4 (moderate skewness). Expected correlation were calculated using Monte Carlo approximation based on 1,000,000 samples of selection coefficients.

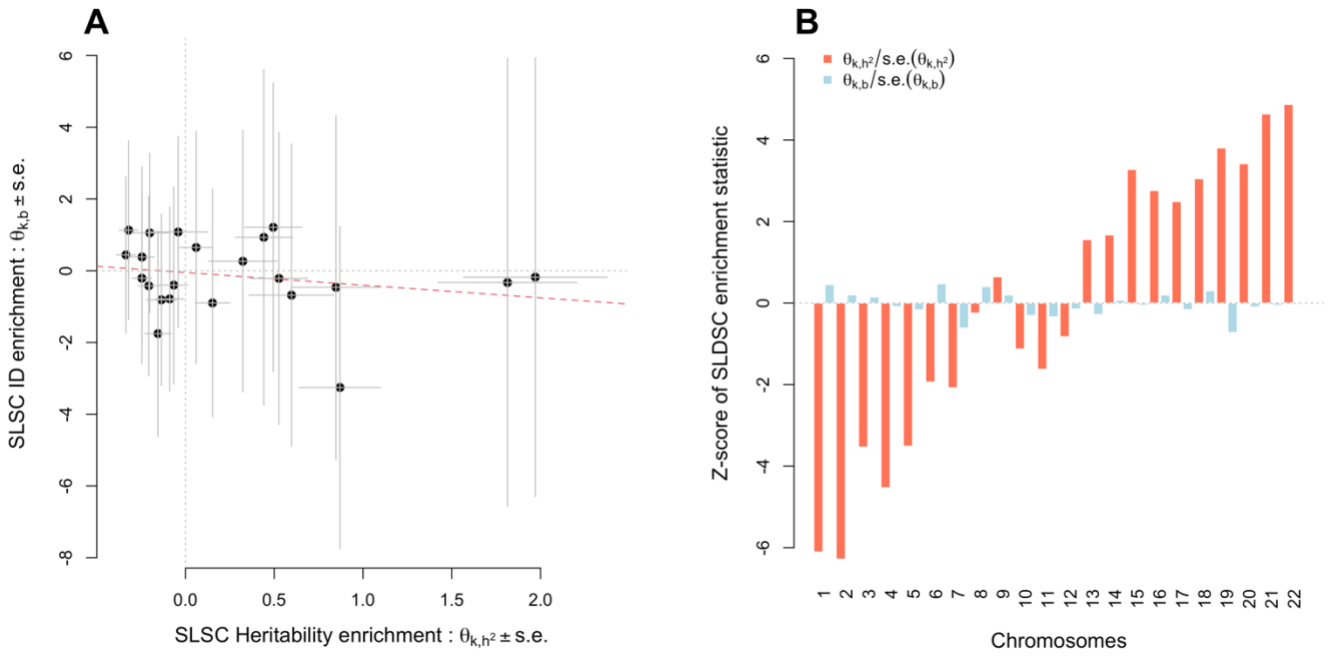

Figure S10. Relationship between GWAS-based estimates of enrichment of heritability (θ_{k,h^2} statistic; x-axis) and ID (with the $\theta_{k,b}$ statistic; y-axis) in simulated data. Enrichment statistics were estimated using stratified LD score regression (SLDC) as described in the Methods section. Data were simulated using genotypes of 348,501 UK Biobank participants and such that heritability is enriched in small chromosomes (e.g., chromosome 22) and depleted in large chromosomes (e.g., chromosome 2), while assuming a uniform contribution of all chromosomes to ID. Full description of the simulations is given in the **Supplemental Methods** section. Error bars are standard errors (s.e.).

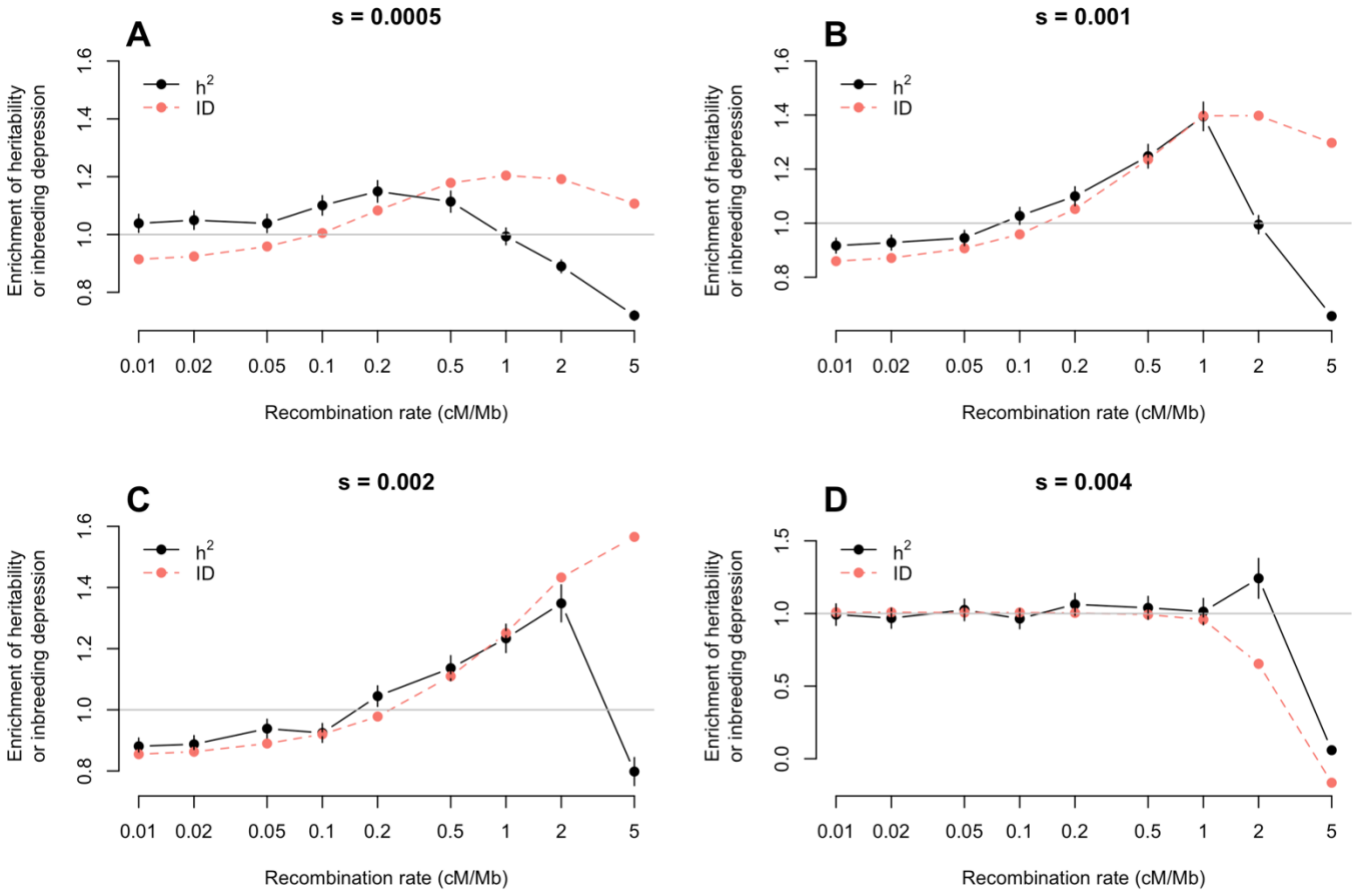

Figure S11. Enrichments of heritability (h^2) and inbreeding depression (ID) in various recombination rate regions as a function of the strength of selection (s) of fitness mutations. Data were generated using forward-time evolutionary simulation (details in **Supplemental Methods**) assuming a fixed dominance coefficient and a fixed selection coefficient for all fitness mutations. In all scenarios (i.e. the four panels), the dominance coefficient is h=0.1 (partially recessive) and the selection coefficient varies between 0.0005, 0.001, 0.002 and 0.004.

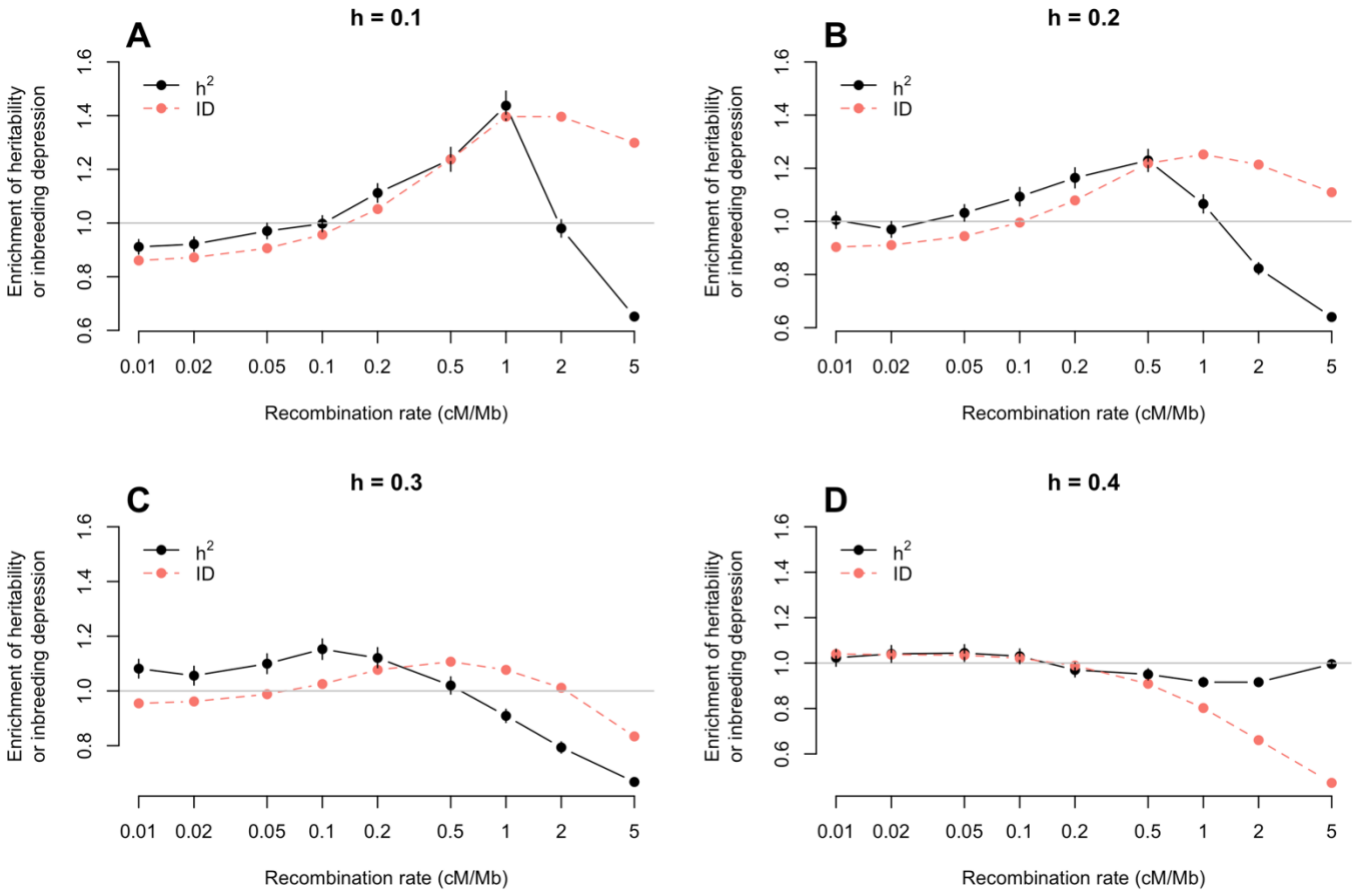

Figure S12. Enrichments of heritability (h^2) and inbreeding depression (ID) in various recombination rate regions as a function of the strength of dominance (h) of fitness mutations. Data were generated using forward-time evolutionary simulation (details in **Supplemental Methods**) assuming a fixed selection coefficient and a fixed dominance coefficient for all fitness mutations. In all scenarios (i.e. four panels), the selection coefficient is s=0.001 (nearly neutral mutation) and the dominance coefficient varies between 0.1, 0.2, 0.3 and 0.4.

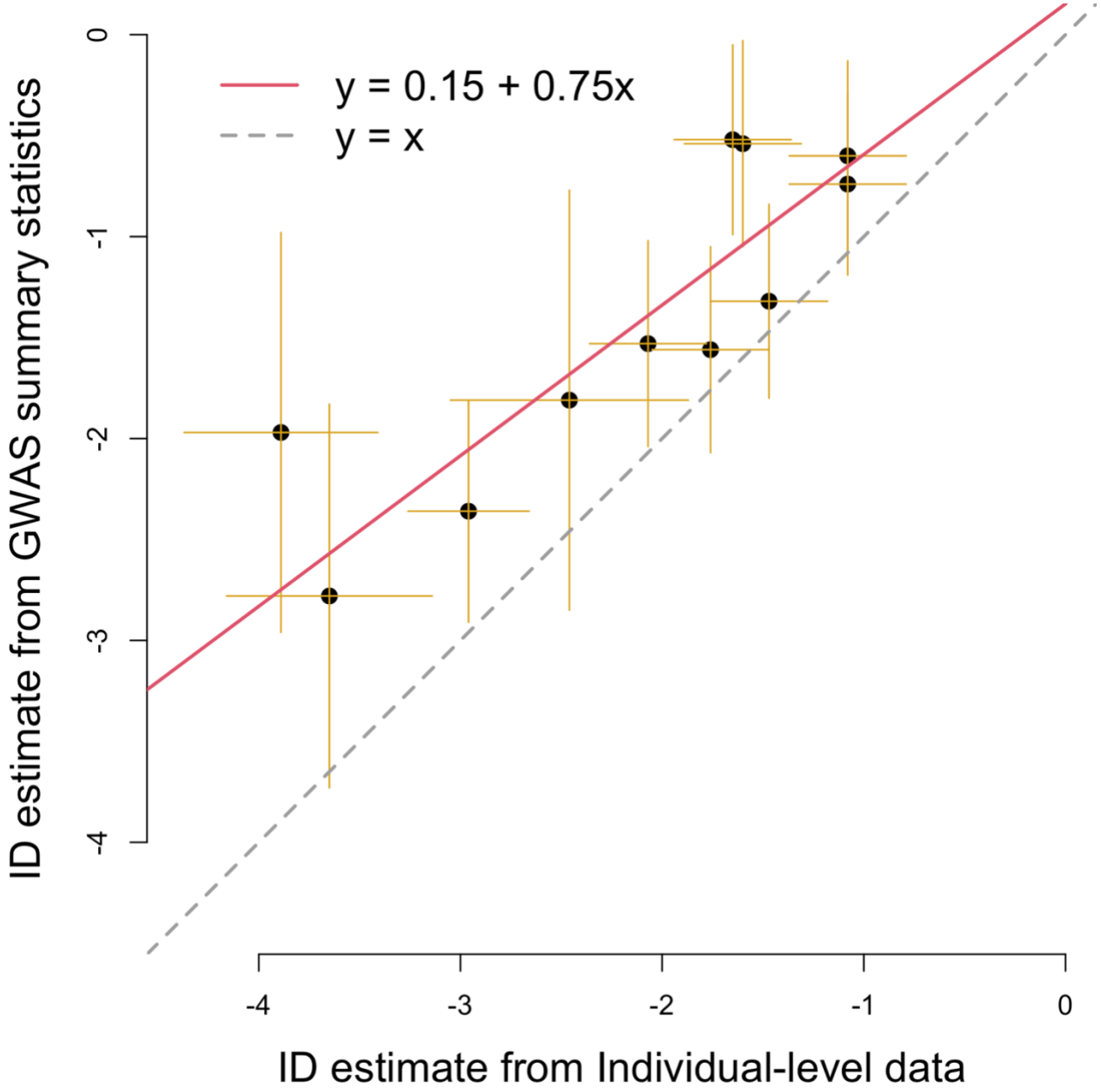

Figure S13. Comparison of estimates of genome-wide inbreeding depression (ID) from individual-level data using the F_{UNI} inbreeding measure (x-axis) and from summary statistics (y-axis) of additive-dominance genome-wide association study (GWAS) of 11 traits. Data underlying this figure are reported in **Table S7**. Estimation of ID from GWAS summary-statistics is based upon LD score regression as described in **Appendix B**. LD scores were calculated for 9,326,198 imputed SNPs (with minor allele frequency >1% and imputation accuracy >0.3, Methods) in 348,501 unrelated participants of the UK Biobank. Error bars are standard errors.

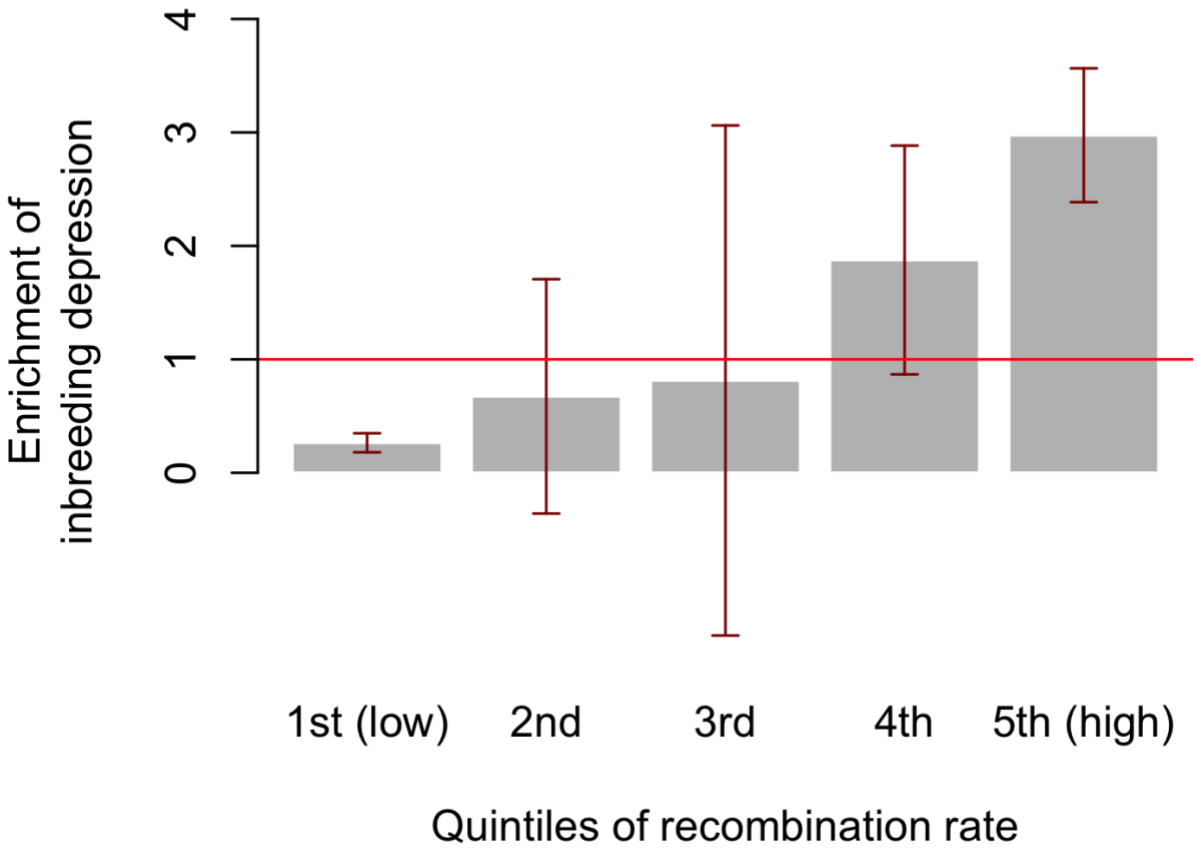

Figure S14. Enrichment of ID in quintiles of the recombination rate distribution. Error bars represent standard errors.

Supplemental Methods

1. Impact of ROH calling and ROH density on the enrichment of ID in low nucleotide diversity and high recombination rates regions

Here we evaluate how much the enrichment of ID in high recombination rate (HRR) regions and in low nucleotide diversity (LND) regions could be explained by potential errors and artefacts in ROH calling or because of the non-uniform distribution of ROHs across the genome.

First, we assessed the sensitivity of our F_{ROH}-based results to potential errors in ROH calling by re-estimating ID enrichment using increasing lengths of ROHs (from 2 Mb to 5 Mb). Although standard errors expectedly increased with ROH length threshold (as fewer ROHs are included in our analyses), we found little change in the estimates of ID enrichment in genomic regions with LND as well as those with HRR (**Figure S4**).

Next, we quantified the genomic density of ROHs in a sample of 455,414 European ancestry participants of the UKB.[11] ROH density was defined, at a given genomic position, as the number of ROHs covering that position. We estimated the density of ROHs over 9.3 million genomic positions across the autosome. We also estimated the density of ROHs in an independent sample from the UK (*N*=3,781 from the UK10K Project[68]) using the same set of SNPs genotyped and quality-controlled as in the UKB and the same parameters to call ROHs (Method section). Given the high consistency between the two estimated ROH densities (r>0.99; **Figure S5**), we therefore hereafter focus on ROH density estimated in the UKB, which has the larger sample size.

Consistent with previous studies, we found that ROHs genomic distribution is not uniform across the genome. In particular, we identified 3 genomic regions with extreme density of ROHs (**Figure S6**;>10 standard deviations above the mean density), which includes the MHC locus (chr6:25,000,000-35,000,000), the lactase locus (LCT) on chromosome 2 (chr2:134,000,000-139,000,000) and the centromere region on chromosome 11 (chr11:46,000,000-57,000,000). We show in **Figure S6**, the correlation between ROH density and all 44 annotations analysed in this study. ROH density was mostly correlated (r~0.3) with the McVicker *B* statistic measuring the strength of background selection. The second largest correlation was observed with recombination rate (r~-0.12), while nucleotide diversity only came at the 8th place (r~0.04) over 44 annotations tested.

One of the assumptions underlying our method is that SNPs have an equal probability to fall into identical-by-descent (IBD) genomic segments. However, the observed genomic distribution of ROHs seems to violate this assumption, at least to the extent that long ROHs were used as proxies for IBD segments. To test the impact of that violation, we analysed ROH density as a continuous genomic annotation and quantified its associated enrichment of ID. On average across traits, we found no significant enrichment of ID associated with ROH frequency (Enrichment=1.01,P=0.42), which overall implies little confounding due to ROH density.

In summary, we have shown in this note that ROH density is not enriched for ID signal and therefore cannot confound any of our results; and also that errors in ROH calling are unlikely to explain the enrichment of ID in HRR and LND.

2. Forward-time evolutionary simulation to quantify the effect of recombination rate on the enrichment of ID and on additive genetic variance

Description of the simulation and enrichment metrics

We performed a forward-time evolutionary simulation using SLIM v3.5 to quantify the effect of recombination rate on the genomic distribution of additive genetic variance and ID. In each simulation replicate, we simulated a population of fixed size $N_e = 1{,}000$ individuals, whose genomes are each made of 9 chromosomes, each 1Mb long. Chromosomes were numbered from 1 to 9 and differed in their recombination rates. Recombination rate values were set to be 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2 and 5 cM/Mb for chromosomes 1 to 9 respectively. Each simulated chromosome contains only deleterious mutations with a fixed selection coefficient (s) and a fixed dominance coefficient (h), such that the relative fitness (i.e. multiplicative fitness model) of an individual carrying one of those mutations is w=1, $1-hs$ and $1-s$ for ancestral allele homozygotes, heterozygotes; and derived allele homozygotes, respectively. We performed two series of simulations. In the first one we fixed the dominance coefficient $h = 0.1$ and varied s=0.0005, 0.001, 0.002 and 0.004; while in the second one, we fixed $s = 0.001$ and varied h between 0.1, 0.2, 0.3 and 0.4.

We assumed a constant mutation rate $\mu = 2 \times 10^{-7}$ per-bp per-generation. We simulated random mating for $10N_e = 10{,}000$ generations then sampled simulated genotypes of 1,000 individuals in the last generation to quantify enrichment of ID and additive genetic variance in log fitness defined below as

$$(S2.1)\ \log(w) = C + \sum_{j=1}^{M} \log(w_j) = C + \sum_{j=1}^{M} x_j \log(1-s_j)/2 + H_j[\log(1-\mathrm{h}_j s_j) - \log(1-s_j)/2]$$
$$\approx C + \sum_{j=1}^{M} x_j(-s_j/2) + H_j(s_j/2)(1-2h_\mathrm{j})$$

where, M is the number of segregating mutations in the last generation, C is an arbitrary constrant, w_j is the relative fitness of carrier of mutation j, s_j and h_j the selection and dominance coefficients of mutation j, x_j(values between 0, 1 and 2) count of mutation j in an individual and $H_j = x_j(2-x_j)$ the indicator of heterozygosity for mutation j. The approximation in Equation (S2.1) is made under the assumption that $s_j \ll 1$.

Let q_j denote the frequency of mutation j. Therefore, the average effect α_j of mutation j on log-fitness can be expressed as

$$(S2.2)\ \alpha_j = -s_j/2 + (1-2q_j)(s_j/2)(1-2h_\mathrm{j}) = -s_j[h_\mathrm{j} + q_j(1-2h_j)],$$

the total ID in log fitness as

$$(S2.3)\ b = -\sum_{j=1}^{M} s_j q_j (1-q_j)(1-2h_\mathrm{j}),$$

and the total additive genetic variance as

$$(S2.4)\ \sigma_A^2 = \mathrm{var}\left(\sum_{j=1}^{M} x_j \alpha_j\right).$$

Note that, because of linkage disequilibrium between mutations, $\mathrm{var}\left(\sum_{j=1}^{M} x_j \alpha_j\right)$ is not expected to be equal to $\sum_{j=1}^{M} 2q_j(1-q_j)\alpha_j^2$, unless recombination rate is extremely large.

For each simulation replicate, we analysed the log-fitness of each individual as the phenotype of interest and quantified enrichment of ID in each recombination rate class using the same approach defined in the main text. We also analysed recombination rate as a continuous annotation, which showed

consistent results. We defined the enrichment of additive genetic variance in for each recombination rate class k (hereafter denoted $\mathrm{Enr}\left[\sigma^2_{A,k}\right]$) as the ratio of additive genetic variance due to SNPs in that class ($\sigma^2_{A,k}$) over the total additive genetic variance (σ^2_A) multiplied by the proportion π_k of SNP in that class, i.e. $\mathrm{Enr}\left[\sigma^2_{A,k}\right] = \sigma^2_{A,k}/(\pi_k \sigma^2_A)$.

Correlation between enrichment of additive genetic variance and that of ID

In this section, we show under classical assumptions that a large correlation between enrichment of ID and that of additive genetic variance, as reported in this study, is not unexpected. Using Equation (S2.3), we define the relative contribution of mutation j to ID as

$$(S2.5)\quad \mathrm{Enr}\left[ID_j\right] = \frac{s_j q_j (1-q_j)(1-2h_{\mathrm{j}})}{\sum_{k=1}^{M} s_k q_k (1-q_k)(1-2h_{\mathrm{k}})}$$

Similarly, and assuming independence between mutations, we can define the relative contribution mutation j to σ^2_A as

$$(S2.6)\quad \mathrm{Enr}\left[\sigma^2_{A,j}\right] = \frac{q_j (1-q_j)\alpha_j^2}{\sum_{k=1}^{M} q_k (1-q_k)\alpha_k^2}$$

Under a mutation-drift-selection equilibrium the frequency (q) of the derived allele is expected to reach a value $q *= \mu/hs$ if h>0 (or $q *= \sqrt{\mu/s}$ if $h = 0$, i.e. fully recessive; Crow & Kimura 1970), where μ is the mutation rate at the locus. Replacing q with its equilibrium frequency in equations (S2.5) and (S2.6) and assuming a constant mutation rate across the genome leads to express the locus contribution to both ID and σ^2_A only as a function of (h, s). However, determining the theoretical correlation between $\mathrm{Enr}\left[\mathrm{ID_j}\right]$ and $\mathrm{Enr}\left[\sigma^2_{\mathrm{A,j}}\right]$ remains intractable because the joint distribution of (h, s) is unknown. Nevertheless, we can show numerically that a large correlation between $\mathrm{Enr}\left[\mathrm{ID_j}\right]$ and $\mathrm{Enr}\left[\sigma^2_{\mathrm{A,j}}\right]$ is expected under various assumed distributions for (h, s).

For example, we fixed h and varied its value between 0.1 and 0.4, while modelling the distribution of s using an Gamma distribution with a mean between 10^{-6} (weak selection) and 10^{-1} (strong selection) and a shape parameter between 1 (strong skewness) and 4 (moderate skewness).

We found that the correlation between enrichment of ID and that of σ^2_A decreases with the mean selection coefficient and with the dominance coefficient of the derived allele. However, we found that moderately skewed distribution of fitness effects (i.e. such that the proportion of mutations with strong fitness effect is low) can yield large correlations between enrichment of ID and that of σ^2_A as shown in **Figure S9**. Consistently, we also report large positive correlations between enrichments of ID and that of additive genetic variance in our forward-time evolutionary simulations (**Figure S11-S12**).

Overall, this analysis highlights a few sufficient (but not necessary) conditions that can lead to a positive and large correlation between enrichment of ID and that of σ^2_A. We acknowledge that this is a simplified model, which nonetheless demonstrates the plausibility of our observations.

Furthermore, we sought to test the observed correlation between enrichments of ID and heritability reported in **Figure 3** could be due to an artefact in our method such that an enrichment of heritability (which has been previously reported) would systematically induce an enrichment of ID. To test this hypothesis, we performed a series of simulations in which heritability is enriched in specific chromosomes, while the per-SNP contribution to ID is uniform across the genome. We used genotypes of all 348,501 unrelated participants of the UKB included in our study to simulate a trait (y) controlled by 11,000 causal variants, i.e. 500 on each of the 22 autosomes. Such a simulation setting generates an enrichment of heritability in smaller chromosomes (e.g., chromosomes 10 to 22) and a depletion in larger ones (e.g., chromosomes 1 to 6). The simulated trait was defined as y = bF + g + e, where g is the

additive genetic value, F the genome-wide inbreeding coefficient (F_{UNI}) and e an environmental value. We simulated a genome-wide ID b=-5 trait standard deviation for 100% inbreeding and a heritability h^2=0.5.

On average over 100 simulation replicates, we found a significant enrichment of heritability in smaller chromosomes and a significant depletion of heritability signal in larger chromosomes (both expected). However, we found no enrichment of ID in any of the 22 chromosomes (**Figure S10**). Altogether, this simulation demonstrates that the correlation between enrichment of heritability and that of ID is not likely to be an artefact of our method.