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

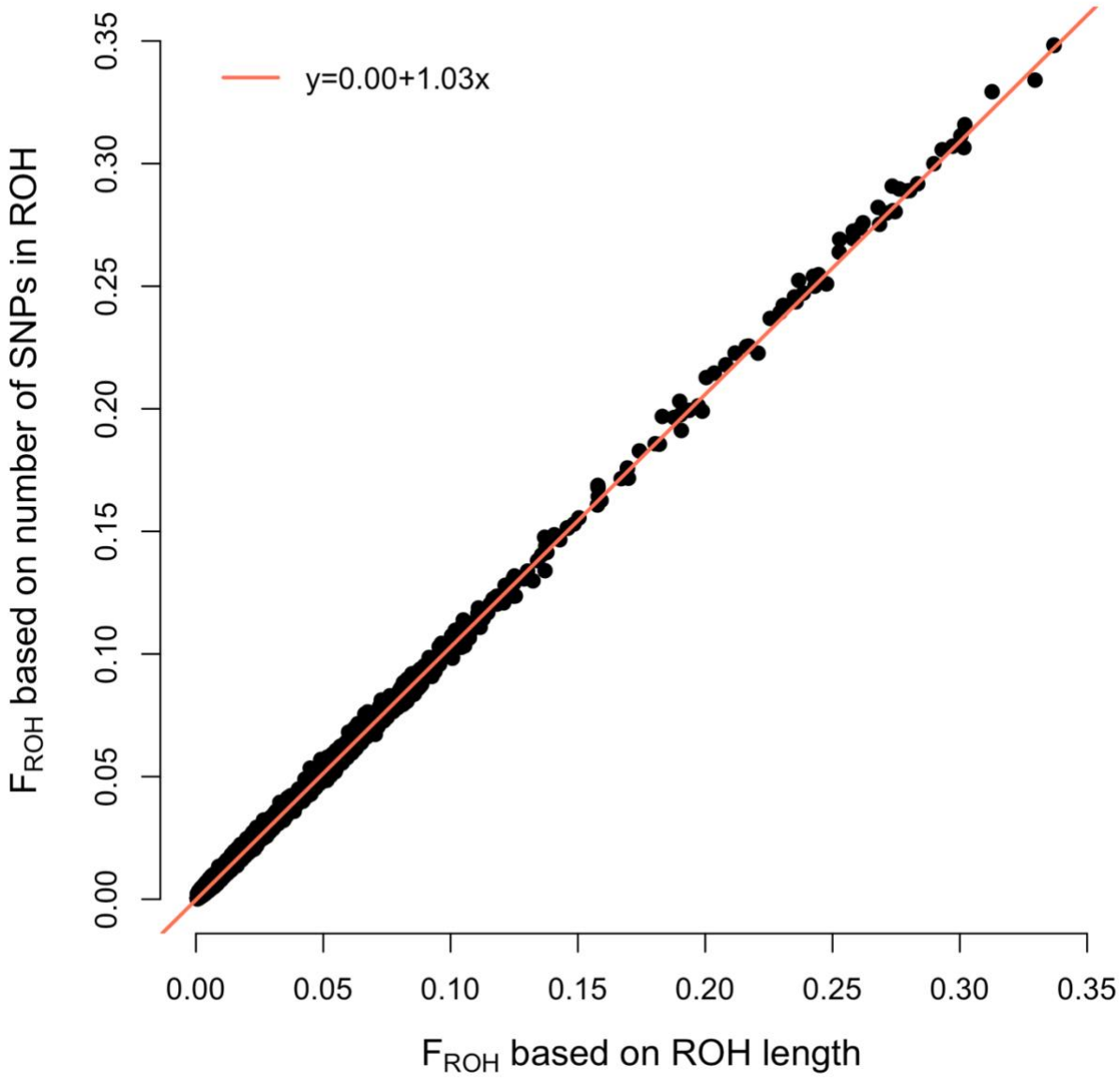

Figure S1. Consistency between two definitions of F_{ROH}. On the x-axis, F_{ROH} is the defined as the cumulated length of ROH in bp divided by 2,785,774,901 that is total autosomal length covered by genotyped SNPs. One the y-axis, F_{ROH} is defined as the proportion of 19,476,620 imputed SNPs with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

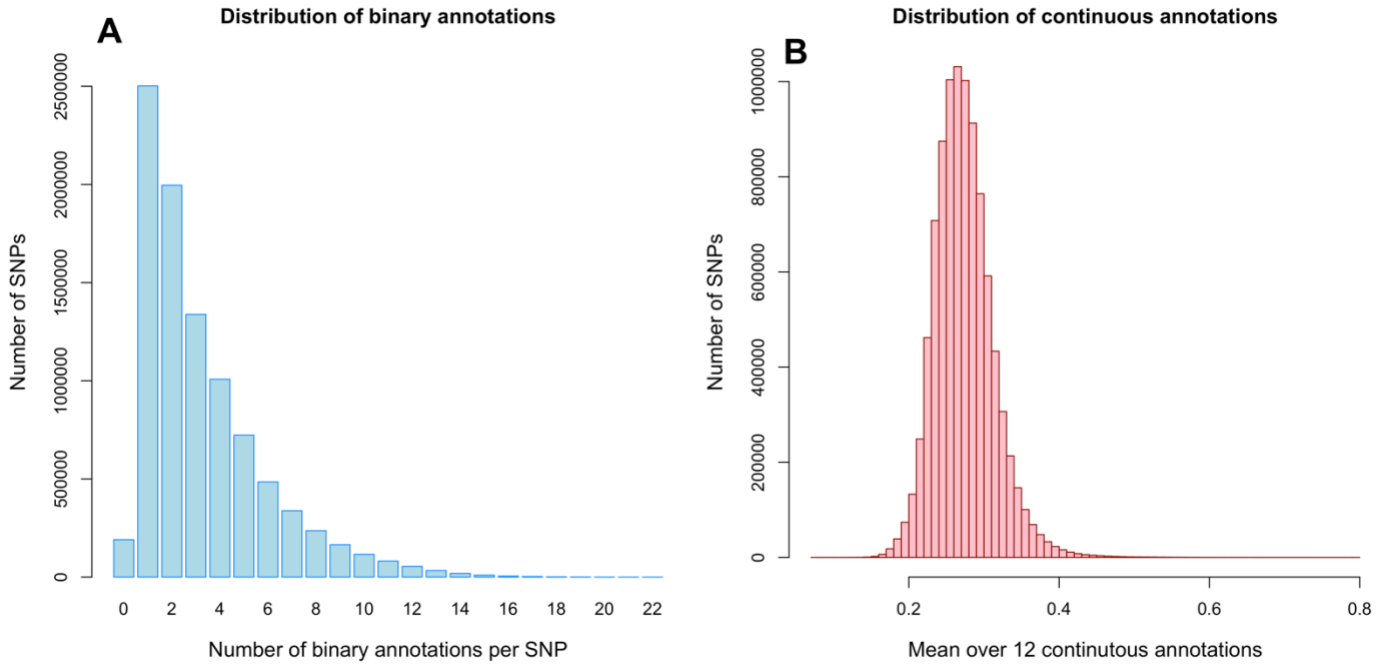

Figure S2. Distribution genomic annotations across SNPs. Left panel shows the histogram of number of binary annotations per SNPs (x-axis). 97.9% of SNPs have at least 1 binary annotation. Right panel shows the histogram across SNPs of the average of 12 normalised continuous annotation. Continuous annotations 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.

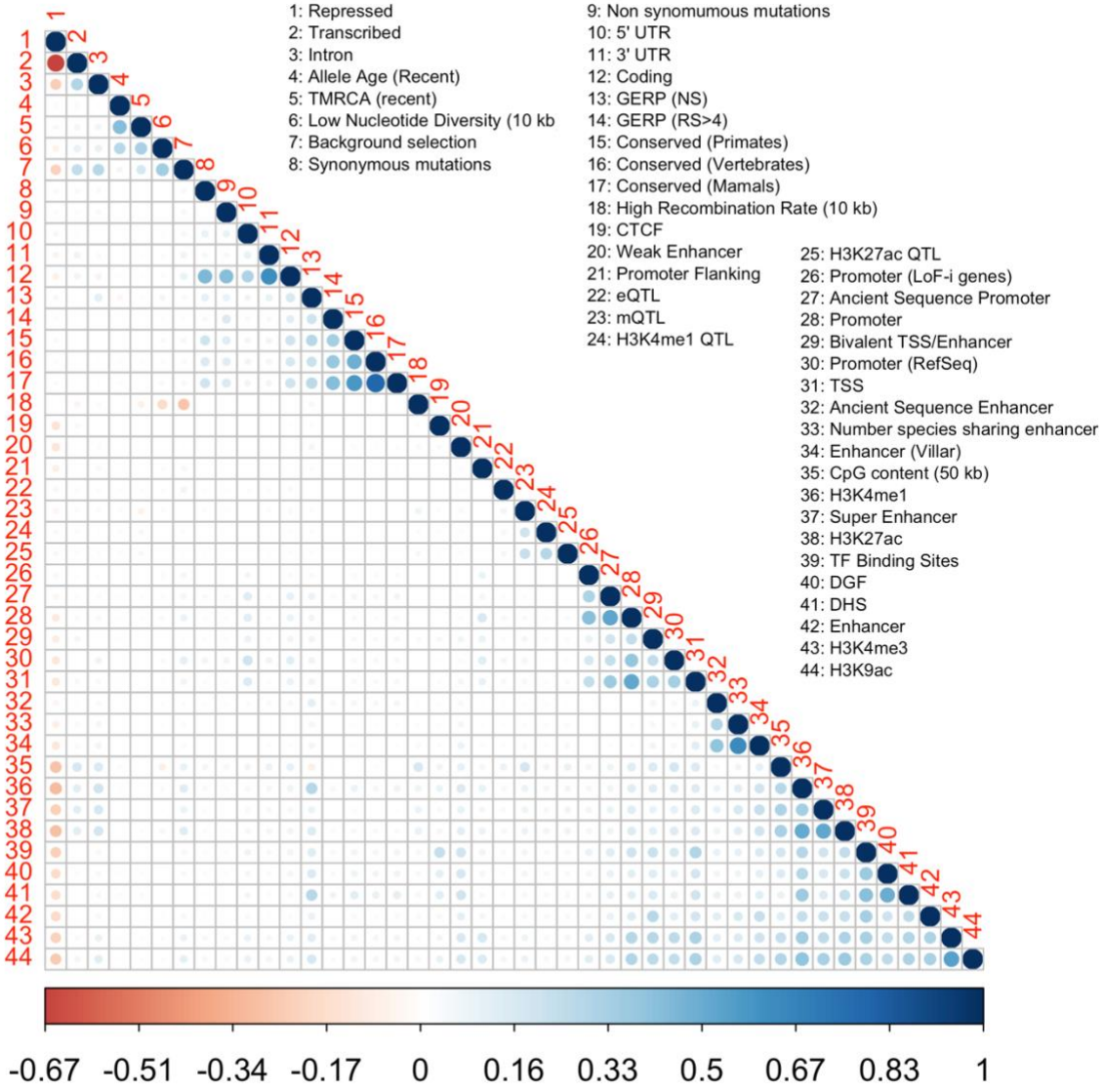

Figure S3. Correlation between genomic annotations estimated from SNPs assigned at least two (binary or continuous) annotations.

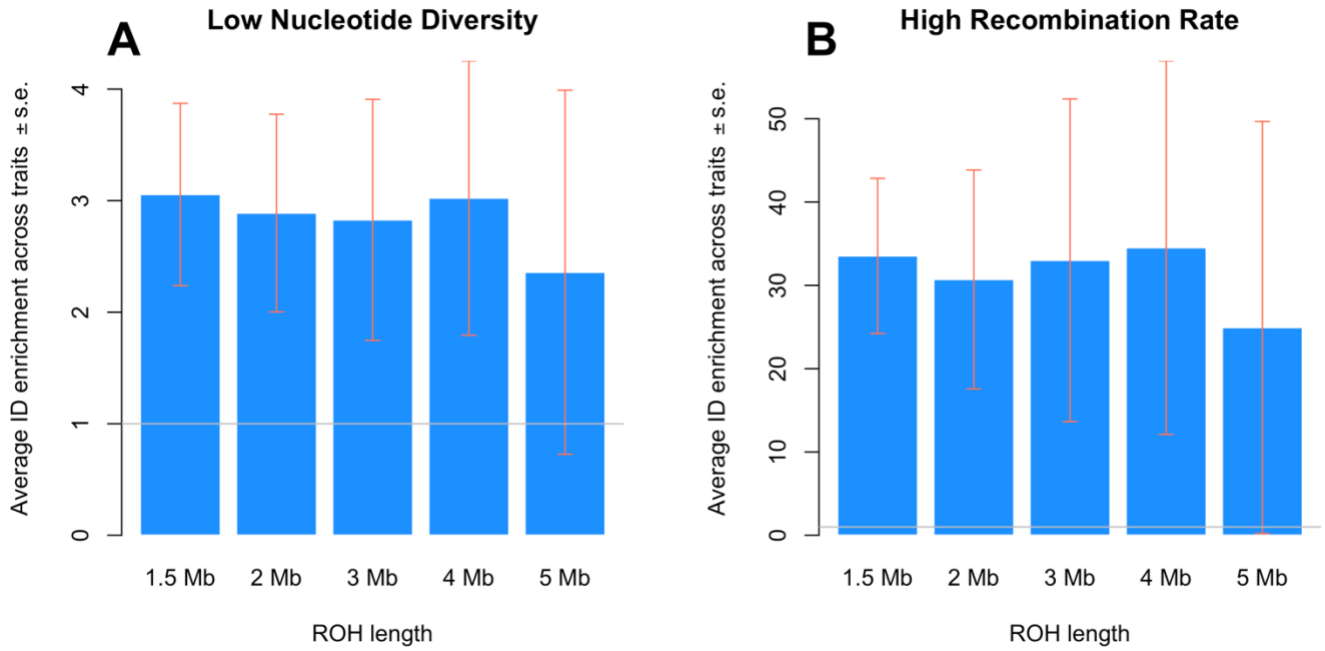

Figure S4. Average enrichment of ID across 11 traits as a function of ROH length. Panel **A** shows F_{ROH}-based estimates of ID enrichment within genomic regions with low nucleotide diversity. Nucleotide diversity was defined, for each SNP, as the mean diversity with 10 kb. Panel **B** shows F_{ROH}-based 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.).

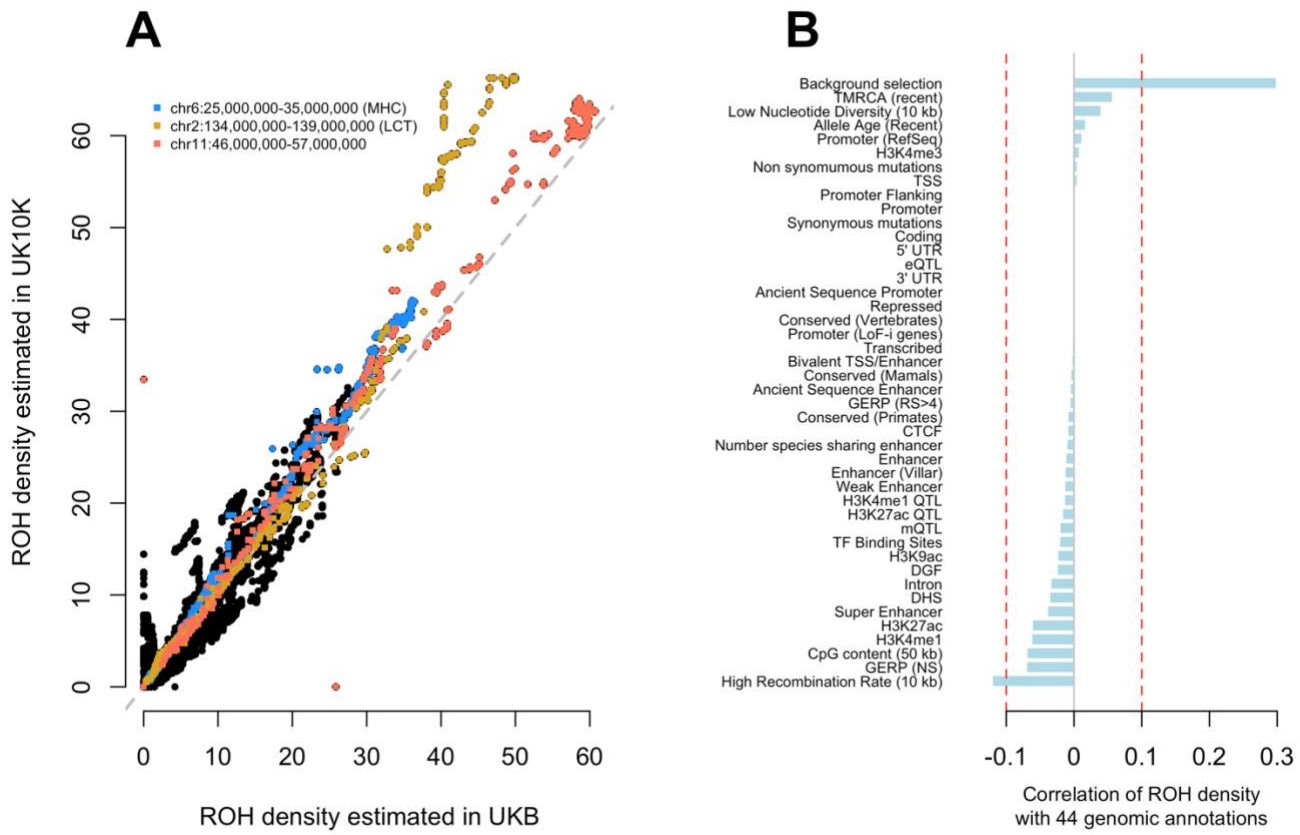

Figure S5. Genomic density of ROHs and its correlation with functional genomic annotations. Panel **A** shows consistent ROH density estimated in two independent samples from the UK: the UK Biobank (UKB; on the x-axis) and the UK10K sample (on the y-axis). Correlation of estimated ROH density from these 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.

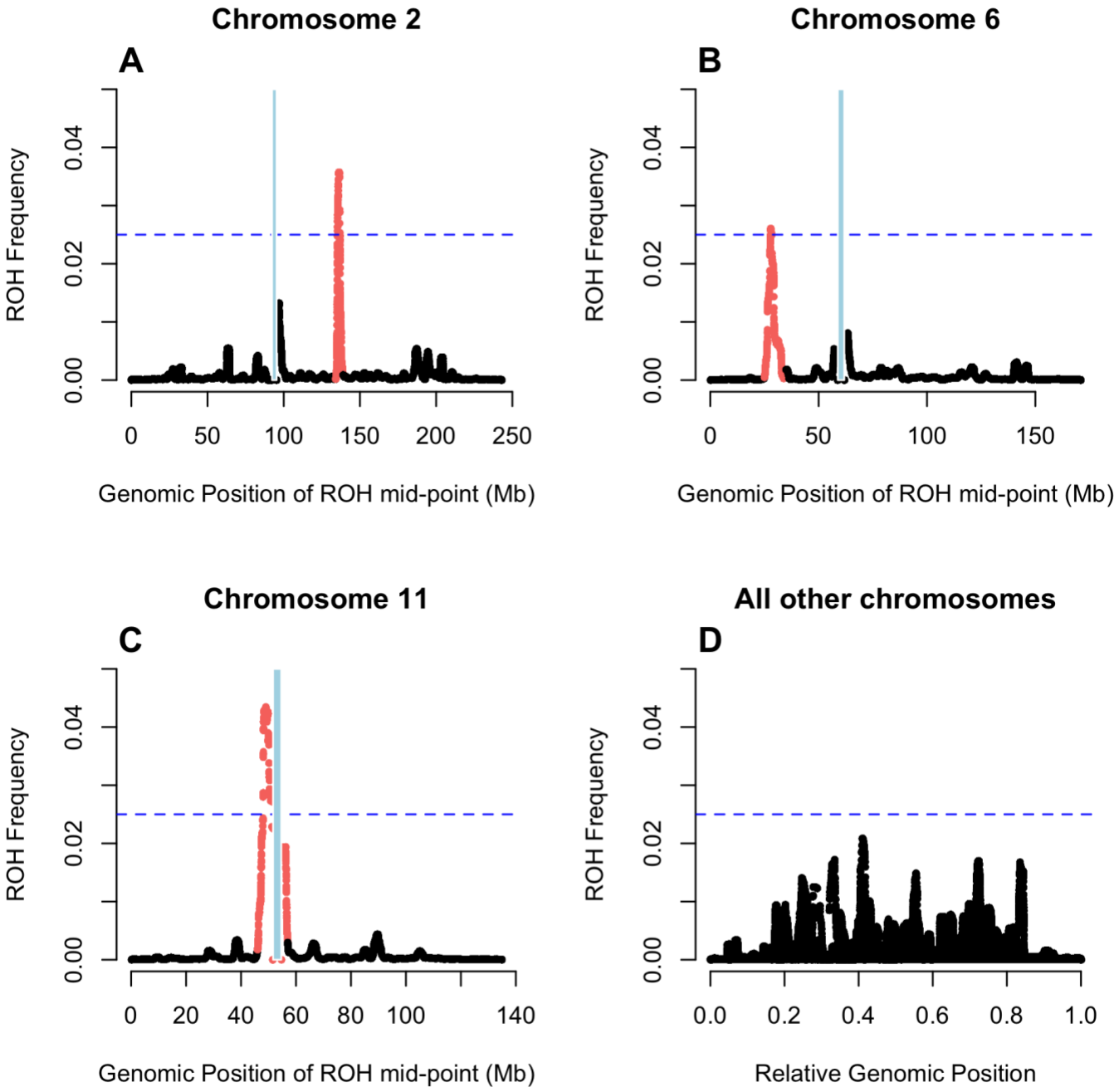

Figure S6. Genomic distribution of ROHs. ROH density was estimated in 456,414 European ancestry participants of the UK Biobank over 9,309,159 genomic positions by counting the number of ROHs overlapping that position. Values of ROH density shown on y-axes of all four panels are divided by the mean density 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.

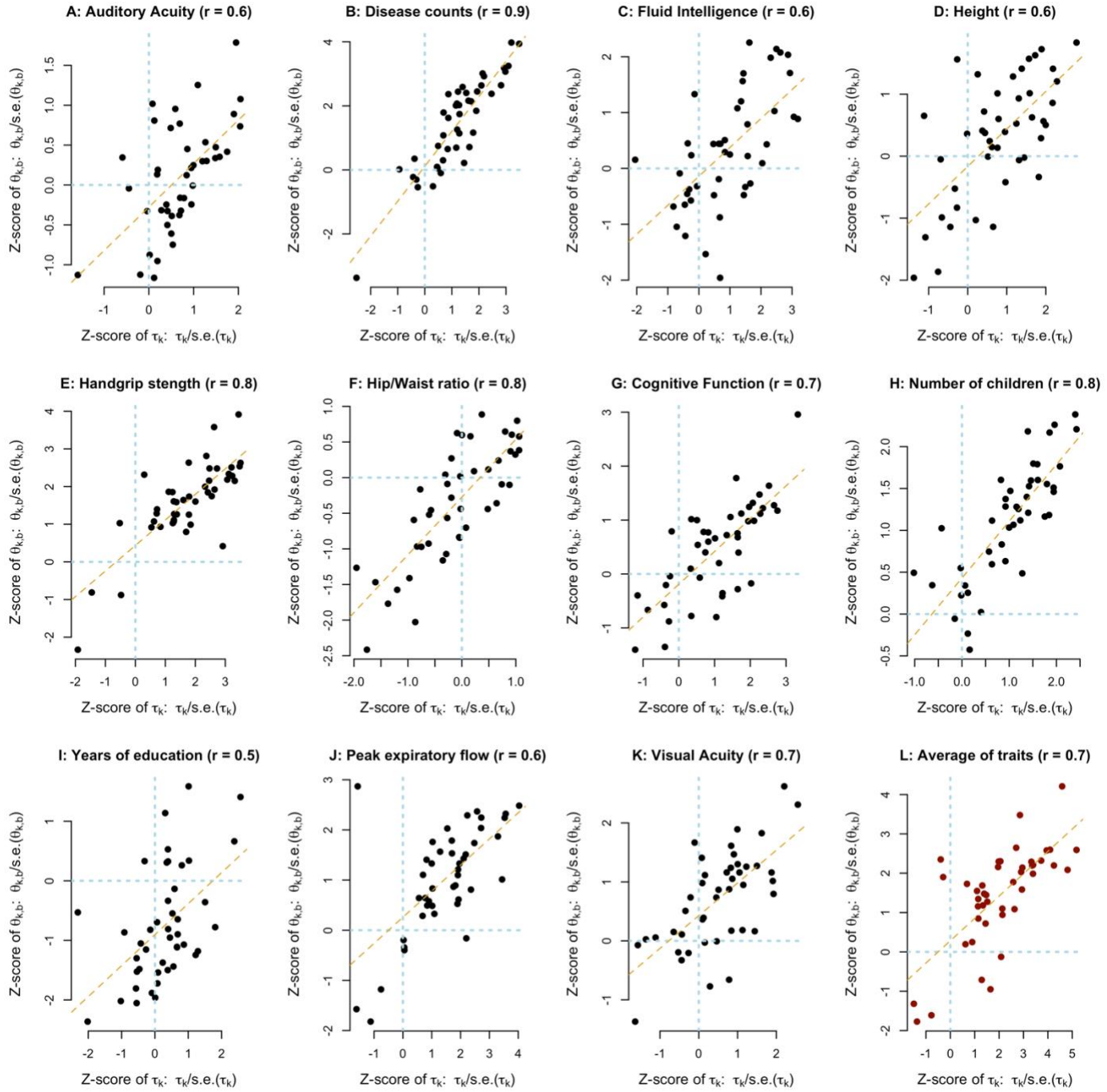

Figure S7. Correlation between individual-level data-based (τ_k) and GWAS-based ($\theta_{k,b}$) estimates of ID enrichment statistics for 11 traits and 44 functional annotations. Each panel represent a trait and the bottom right panel the average across traits. Within each panel, a dot present a genomic 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).

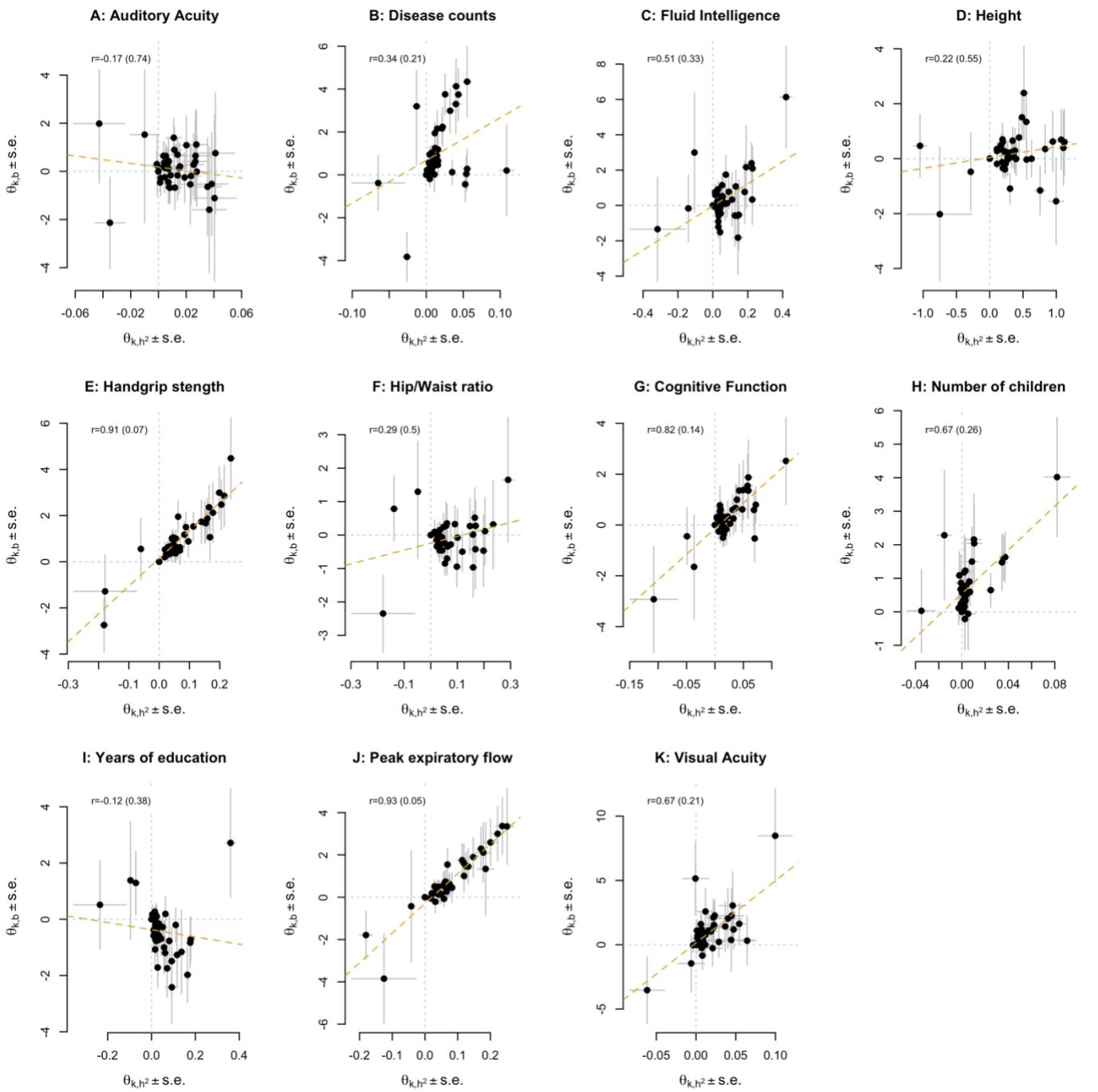

Figure S8. Correlation (r) between GWAS-based estimates of enrichment of heritability (using the θ_{k,h^2} statistic; x-axis) and ID (using the $\theta_{k,b}$ statistic; y-axis) across 44 genomic annotations and 11 traits associated with inbreeding. Enrichment statistics were estimated using stratified 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**.

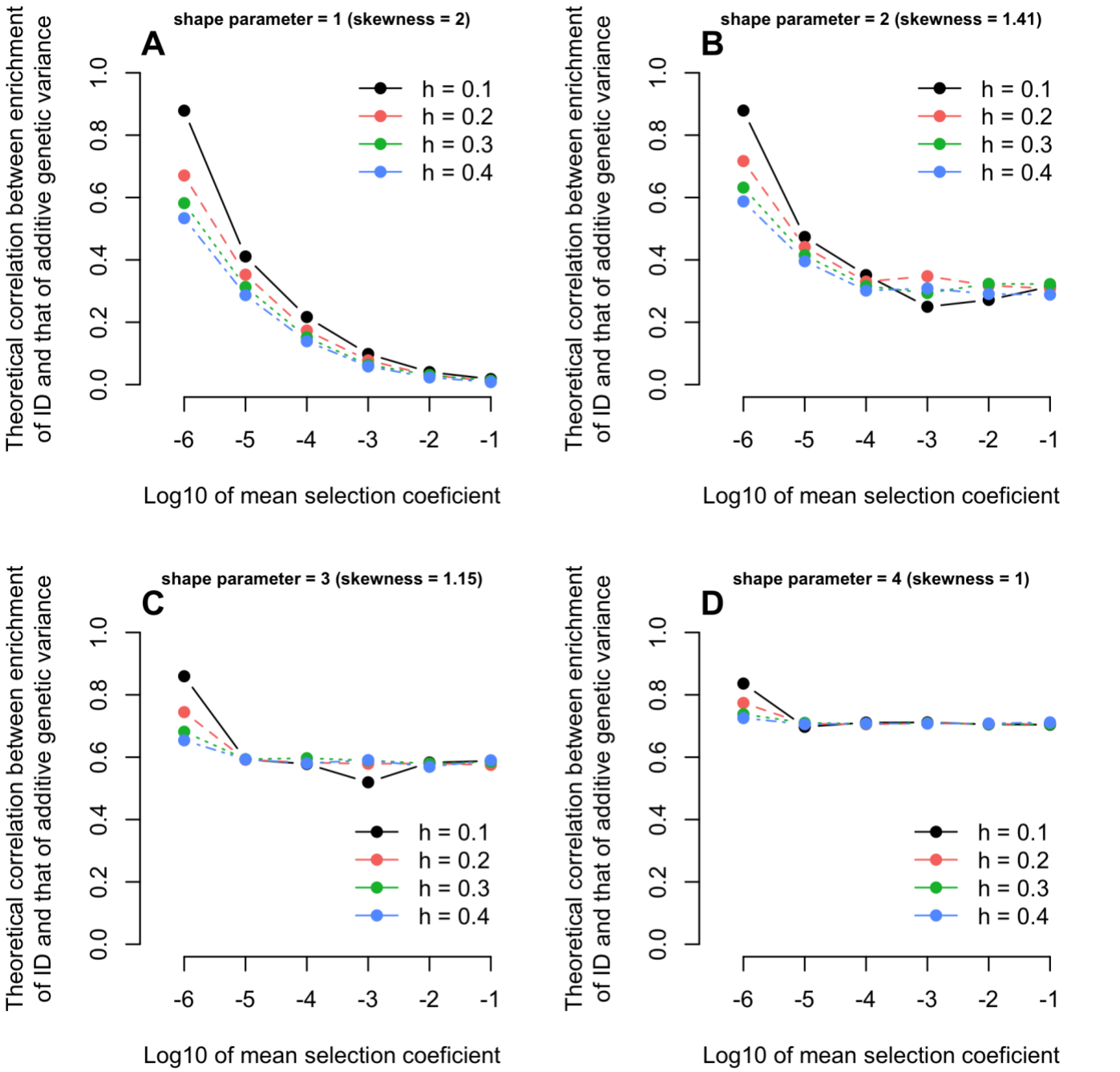

Figure S9. Expected correlation between enrichments of heritability and ID under various assumed distributions of selection and dominance coefficients of fitness mutations (**Supplemental Methods**). Selection coefficients were assumed to be Gamma-distributed with a mean varying between 10^{-6} (weak 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.

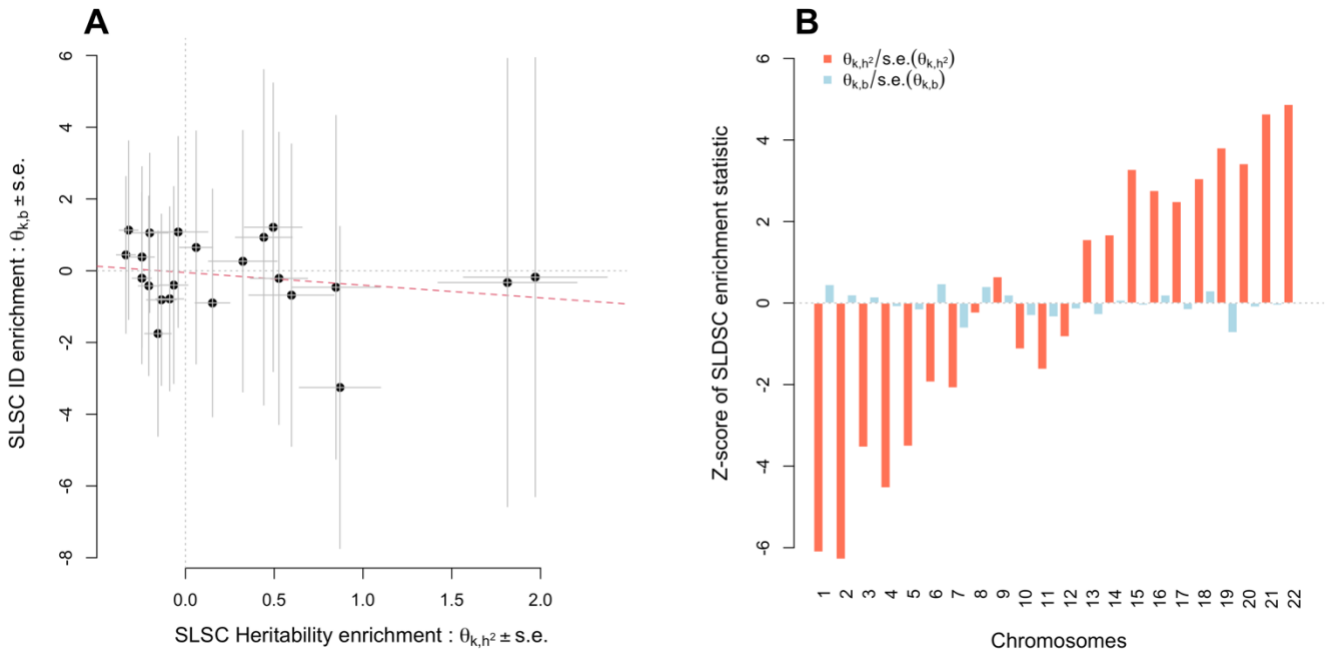

Figure S10. Relationship between GWAS-based estimates of enrichment of heritability (θ_{k,h^2} statistic; x-axis) and ID (with the $\theta_{k,b}$ statistic; y-axis) in simulated data. Enrichment statistics were estimated using stratified LD score regression (SLDC) as described in the Methods section. 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.).

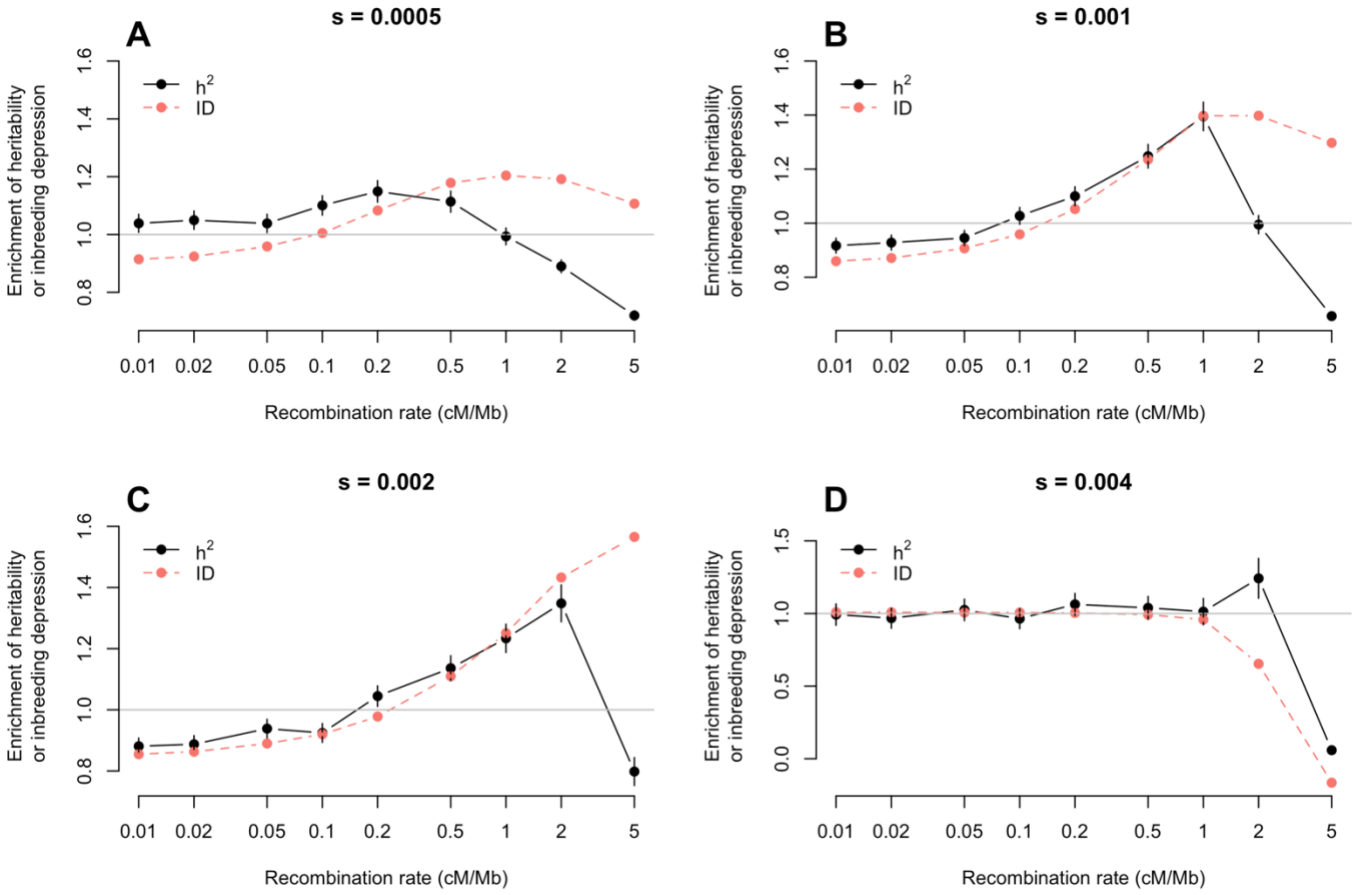

Figure S11. Enrichments of heritability (h^2) and inbreeding depression (ID) in various recombination rate regions as a function of the strength of selection (s) of fitness mutations. Data were generated using forward-time evolutionary simulation (details in **Supplemental Methods**) assuming a fixed 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.

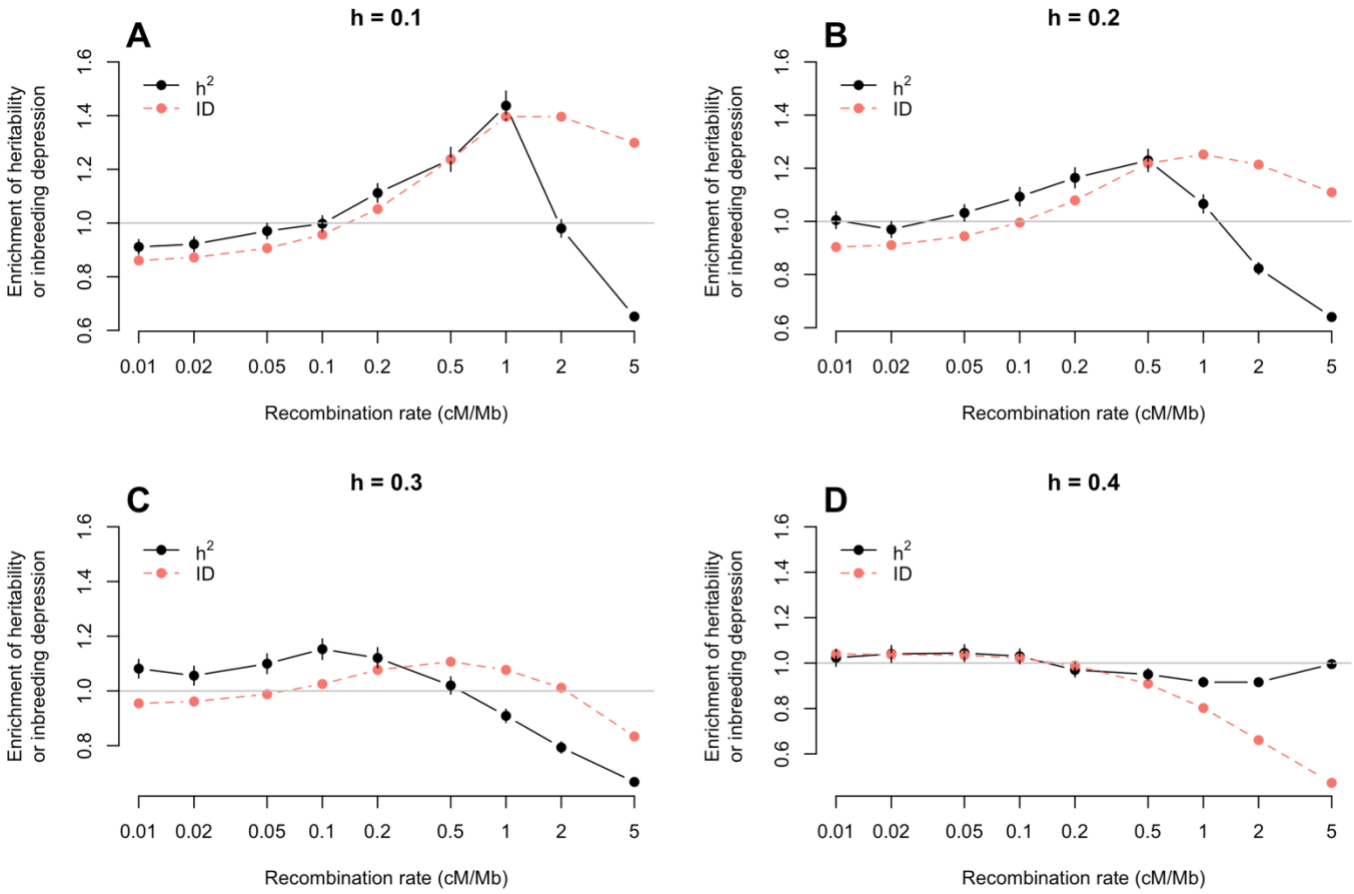

Figure S12. Enrichments of heritability (h^2) and inbreeding depression (ID) in various recombination rate regions as a function of the strength of dominance (h) of fitness mutations. Data were generated using forward-time evolutionary simulation (details in **Supplemental Methods**) assuming a fixed 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.

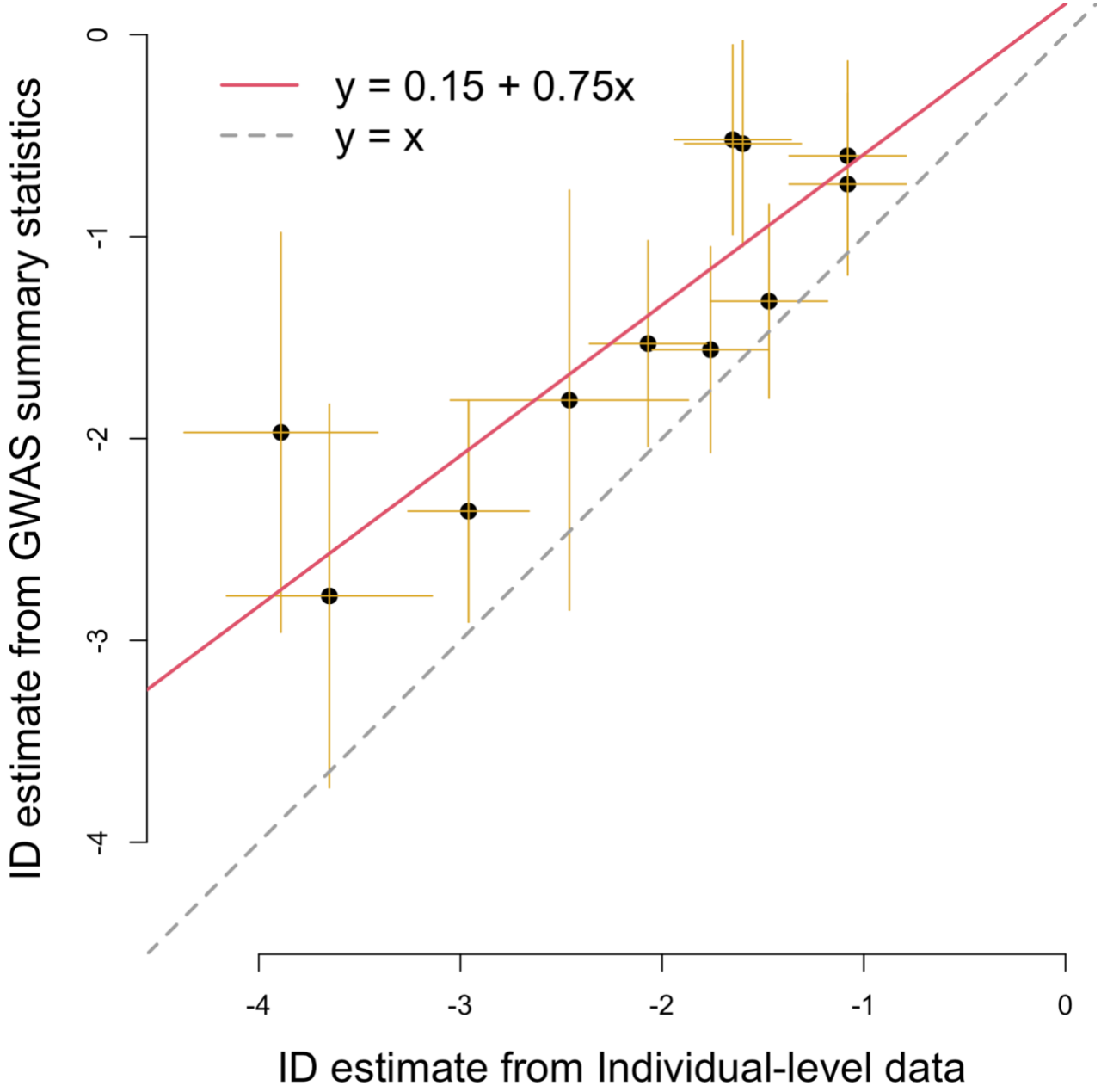

Figure S13. Comparison of estimates of genome-wide inbreeding depression (ID) from individual-level data using the F_{UNI} inbreeding measure (x-axis) and from summary statistics (y-axis) of additive-dominance genome-wide association study (GWAS) of 11 traits. Data underlying this figure are reported 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.

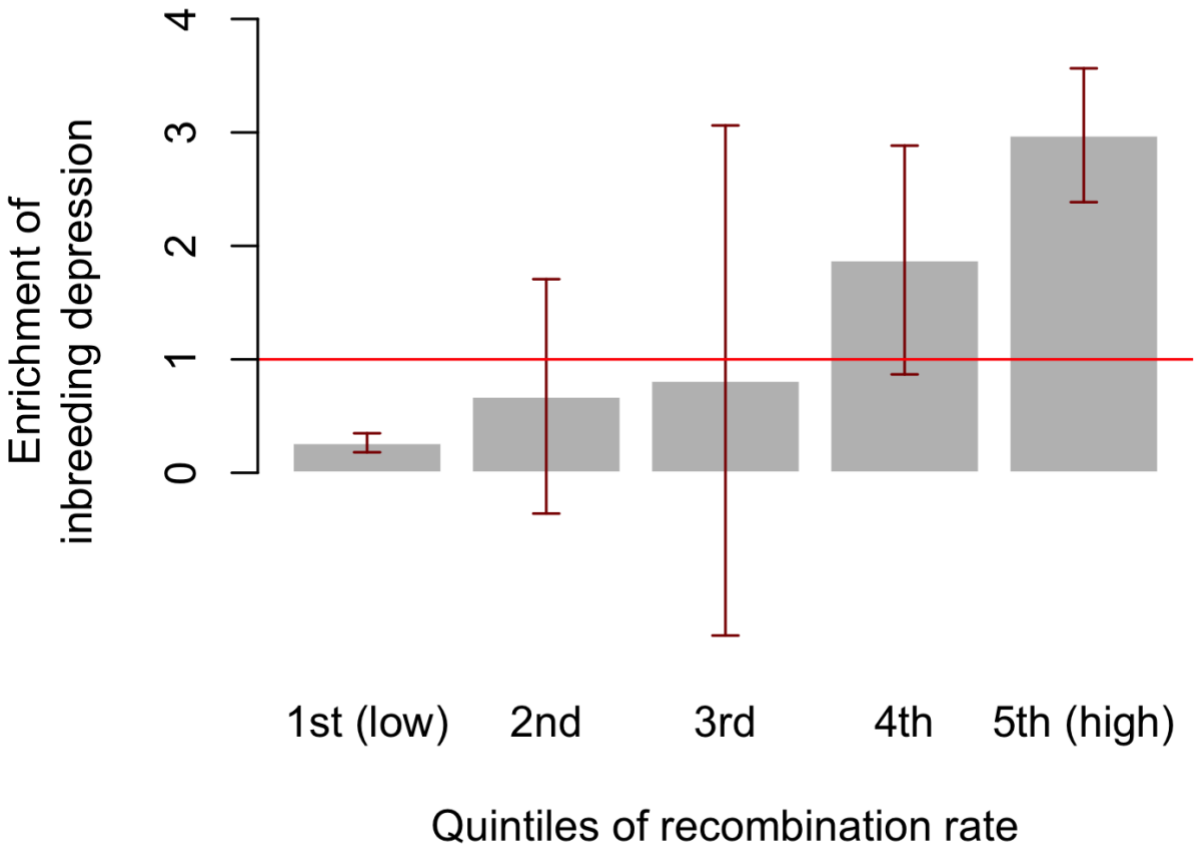

Figure S14. Enrichment of ID in quintiles of the recombination rate distribution. Error bars represent standard errors.

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.