Evidence of natural selection in human complex phenotypes from GWAS summary statistics

Dominic Holland*a,b, Oleksandr Freid, Rahul Desikan*c, Chun-Chieh Fan*a,e,f, Alexey A. Shadrin*d, Olav B. Smeland*a,d,g, Ole A. Andreassen*d,g, Anders M. Dale*a,b,f,h

*aCenter for Multimodal Imaging and Genetics, University of California at San Diego, La Jolla, CA 92037, USA,
bDepartment of Neurosciences, University of California, San Diego, La Jolla, CA 92093, USA,
cDepartment of Radiology, University of California, San Francisco, San Francisco, CA 94158, USA,
dNORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo 0424 Oslo, Norway,
eDepartment of Cognitive Sciences, University of California at San Diego, La Jolla, CA 92093, USA,
fDepartment of Radiology, University of California, San Diego, La Jolla, CA 92093, USA,
gDivision of Mental Health and Addiction, Oslo University Hospital, 0407 Oslo, Norway,
hDepartment of Psychiatry, University of California, San Diego, La Jolla, CA 92093, USA.

Abstract

We propose an extended Gaussian mixture model for the distribution of causal effects of common single nucleotide polymorphisms (SNPs) for human complex phenotypes, taking into account linkage disequilibrium (LD) and heterozygosity (H), while also allowing for independent components for small and large effects. Using a precise methodology showing how genome-wide association studies (GWAS) summary statistics (z-scores) arise through LD with underlying causal SNPs, we applied the model to multiple GWAS. Our findings indicated that causal effects are distributed with dependence on a SNP’s total LD and H, whereby SNPs with lower total LD are more likely to be causal, and causal SNPs with lower H tend to have larger effects, consistent with the influence of negative pressure from natural selection. The degree of dependence, however, varies markedly across phenotypes.

Keywords: GWAS, Polygenicity, Discoverability, Heritability, Causal SNPs, Effect size, Linkage disequilibrium, Minor allele frequency, Natural selection

INTRODUCTION

There is currently great interest in the distribution of causal effects among trait-associated single nucleotide polymorphisms (SNPs), and recent analyses of genome-wide association studies (GWAS) have begun uncovering deeper layers of complexity in the genetic architecture of complex traits [21, 13, 14, 59, 60]. This research is facilitated by using new analytic approaches to interrogate structural features in the genome and their relationship to phenotypic expression. Some of these analyses take into account the fact that different classes of SNPs have different characteristics and play a multitude of roles [42, 43, 12]. Along with different causal roles for SNPs, which in itself would suggest differences in distributions of effect-sizes for different sets of causal effects, the effects of minor allele frequency (MAF) of the causal SNPs and their total correlation with neighboring SNPs are providing new insights into the action of selection on the genetic architecture of complex traits [14, 56, 60].

Selection pressure can be assessed by measuring the extent of long-range linkage disequilibrium (LD), given heterozygosity and local recombination rates. In general, the signature of positive selection is higher heterozygosity of variants given the extent of LD (or higher extended LD, given the heterozygosity). Negative selection acts to keep variants (mutations) deleterious to fitness at low frequency (or ultimately remove them), a process that is facilitated by higher recombination rates [18]. There are more ways of being wrong than being right: any given mutation is more likely to be deleterious to, rather than aid, fitness. Since weakly deleterious variants will take a while to be removed, recent variants are likely to be deleterious [29]. In addition, the larger the effect of a deleterious variant the more efficient negative selection will be, which all suggests that the lower the MAF the larger the effect size— which indeed is expected under evolutionary models [36] and consistent with empirical findings [34]. On the other hand, older variants, being old, are likely to have been positively selected (perhaps pleiotropically with a fitness-related trait), and over time will acquire LD with recent variants [14]. So when negative selection is operating, we should expect to find more and more effects with larger and larger effect size at lower LD score (or total LD) and lower heterozygosity.

Several recent reports focused on individual aspects of genetic architecture and their potential roles in natural selection, or excluded important quantities like polygenicity, but it is important to treat disparate (and overlapping) architectural features simultaneously to avoid over-
interpreting the role of any individual feature and more precisely predict quantities of interest like heritability. Here we present a unifying approach, using Bayesian analysis on GWAS summary statistics and maximum likelihood estimation of model parameters, to explore the role of total LD (TLD) on the distribution of causal SNPs, and the impact of MAF on causal effect size, while simultaneously incorporating multiple effect-size distributions (TLD used here and previously [22, 13] is similar to LD score [3] but takes into account more neighboring SNPs; it is defined in the Methods section). Treating each of these factors independently is likely to provide a misleading assessment. Thus, we posit a model for the prior distribution of effect sizes, and building on previous work [21], fit it to the GWAS summary statistics for a wide range of phenotypes. From the estimated set of model parameters for a phenotype, we generate simulated genotypes and calculate the corresponding z-scores. With a wide range of model parameter values across real phenotypes, the specificity of the individual parameters for a given phenotype make them narrowly defining of the distribution of summary statistics for that phenotype; since each parameter is physically motivated, they allow for direct interpretation, and the more accurate the model fit the more accurate are estimates of physical quantities based on the parameter values. We find that the detailed distributions of GWAS summary statistics for real phenotypes are reproduced to high accuracy, while many of the parameters used to set up a simulated phenotype are accurately reproduced by the model. The distribution of causal SNPs with respect to their TLD will be shown to vary widely across different phenotypes. Additionally, accurate detailed model reproduction of empirical distributions of GWAS summary statistics, while taking into account complexity arising from selection pressure and categories of strong and weak effects, argues in favor of greater accuracy in estimating “SNP heritabilty” [46], the measure of trait heritability that can in principle be gleaned from GWAS summary statistics.

In previous work [21], which builds on earlier reports of others (e.g., [61, 10, 15]), we presented a Gaussian mixture model to describe the distribution of underlying causal SNP effects (the “β” simple linear regression coefficients that indicate the strength of association between causal variants and a phenotype). Due to extensive and complex patterns of LD among SNPs, many non-causal SNPs will exhibit a strong association with phenotypes, resulting in a far more complicated distribution for the summary z-scores. The basic model for the distribution of the causal β’s is a mixture of non-null and null normal distributions, the latter (denoted $\mathcal{N}(0,0)$) being just a delta function (or point mass at zero):

$$
β \sim π_1 \mathcal{N}(0,σ_1^2) + (1 − π_1) \mathcal{N}(0,0)
$$

where $π_1$ is the polygenicity, i.e., the proportion of SNPs that are causal (equivalently, the prior probability that any particular SNP is causal), and $σ_1^2$ is the discoverability, i.e., the variance of the non-null normal distribution, which was taken to be a constant across all causal SNPs. The distribution of z-scores arising from this shows strong heterozygosity and TLD dependence (Eqs. 21 and 29 in [22]). Correctly implementing this relatively simple model (Eq. 1) and applying it to GWAS summary statistics involved a good deal of complexity, but resulted in remarkably accurate estimation of the overall distribution of z-scores, as demonstrated by the fit between the actual distribution of z-scores for multiple phenotypes and the model predictions, visualized on quantile-quantile (QQ) plots. As noted above, the more accurate the fit the more correct are the values of the physically meaningful parameters; additionally, these values help distinguish the genetic architectures of different phenotypes, and enable prediction, for example of sample size required to detect the preponderance of causal SNPs.

Given that GWAS, due at least to lack of power, consistently failed to discover the SNPs that explain the bulk of heritability, the objective of the model was to characterize the likely large number of causal SNPs of weak signal that evaded detection – and likely would explain missing or hidden heritability. Thus, the model was not intended to accurately predict the tails of the distribution of z-scores, which for large enough GWAS would correspond to already discovered SNPs (i.e., whose summary p-values reached genome-wide significance), and involved the simplifying assumption that the bulk of causal effects roughly followed the same Gaussian. Nevertheless, the predicted distributions captured the main characteristics of the SNP associations for many phenotypes. For some phenotypes, however, the fit was relatively poor, and even for phenotypes where the overall fit was good, a breakdown of the z-score distributions for SNPs stratified by heterozygosity and TLD indicated some unevenness in how well the tails of the sub-distributions were predicted. Additionally, recent work by others focusing on selection pressure [59, 41] indicated that it is important to take SNP heterozygosity into account as a component in discoverability (initially explored by examining four fixed relationships in [48]; see also [26]), and that an additional Gaussian distribution for the $β$’s might be appropriate if large and small effects are distributed differently [60]. These approaches, however, were not combined (the former involving a single causal Gaussian incorporating heterozygosity, the latter involving two Gaussians with no heterozygosity dependence).

An extra complexity is that TLD might play an important role in the distribution of causal effects [14]. It is unclear how all these factors impact each other. A final and very important matter is that most analyses in the literature by other groups were conducted in the “infinitesimal” model framework [41, 46, 14, 12, 3], where all SNPs are causal, though there is compelling evidence that only a very small fraction of SNPs are in fact causal for any given phenotype [59, 62, 60, 13, 43, 22].

In the current work, we sought to extend our earlier work to incorporate multiple Gaussians, while taking into account TLD and selection effects reflected in heterozy-
gosity, in modeling the distribution of causal β’s. Note that this is independent of the need to correctly take heterozygosity and TLD into account when building a distribution for z-scores even when the distribution of causal effects (β’s) does not depend on these factors, as in our earlier work [22]. We then applied the model to study evolutionary aspects of complex human phenotypes, using large-GWAS summary statistics.

MODEL

The appropriate methodology calls for using an extensive reference panel that likely involves all possible causal SNPs with MAF greater than some threshold (e.g., 1%), and regard the z-scores for typed or imputed SNPs – a subset of the reference SNPs – as arising, directly or through LD, from the underlying causal SNPs in the reference panel.

Since the single causal Gaussian, Eq. 1, has provided an appropriate starting point for many phenotypes, it is reasonable to build from it. With additional terms included, if it turns out that this original term is not needed, the fitting procedure, if implemented correctly, should eliminate it. Also, anticipating extra terms in the distribution of causal β’s, we introduce a slight change in labeling the Gaussian variance (σ_β^2 → σ_c^2), and write the distributions for the causal component only – it being understood that the full distribution will include the last term on the right side of Eq. 1 for the prior probability of being null.

Given that for some phenotypes there is strong evidence that rarer SNPs have larger effects, we next include a term that reflects this: a Gaussian whose variance is proportional to \( H^S \), where \( H \) is the SNP’s heterozygosity and \( S \) is a new parameter which, if negative, will reflect the noted behavior [59]. With the addition of the new term, the total prior probability for the SNP to be causal is still given by \( \pi_1 \). Thus, extending Eq. 1, we get:

\[
\beta(H) \sim \pi_1 \{ (1 - p_c) N(0, \sigma_c^2) + p_c N(0, H^S \sigma_c^2) \} 
\]

(ignoring, as noted, the null component \( (1 - \pi_1)N(0,0) \)), where \( p_c (0 \leq p_c \leq 1) \) is the prior probability that the SNP’s causal component comes from the “c” Gaussian (with variance \( H^S \sigma_c^2 \)), and \( p_b \equiv 1 - p_c \) is the prior probability that the SNP’s causal component comes from the “b” Gaussian (with variance \( \sigma_c^2 \)). This extension introduces an extra three parameters \( p_c, \sigma_c, \text{ and } S \), assumed for the moment to be the same for all SNPs. Ignoring inflation and implementation details like choice of reference panel and parameter estimation scheme, setting \( p_c \equiv 1 \) recovers the model distribution assumed in [59]; additionally setting \( \pi_1 \equiv 1 \) recovers the primary model distribution assumed in [41]; and additionally setting \( S \equiv -1 \) recovers the model assumed in [3], while instead setting \( S \equiv -0.25 \) partially recovers the “recommended” LD-adjusted kinships (LDAK) model distribution in [47, 46].

The LDAK model of Speed et al also incorporates two predetermined fixed factors that scale the variance of the causal effect size distribution: SNP-specific local LD weightings and “information scores” (the latter a quantification of SNP quality [47]). The local LD weighting (which remains implicitly dependent on MAF) is designed to scale down the effect size in the context of the infinitesimal model, to try to compensate (in empirical z-scores) for the effect itself being replicated through LD with neighboring SNPs. It should be noted that this problem of LD is implicitly taken care of in our basic model by: (1) using a point-normal distribution for non-causal SNPs; (2) an extensive underlying reference panel; and (3) directly modeling the effects of LD on z-scores in our PDF for GWAS summary statistics – e.g., see Eq. 21 in [22]. The LDAK local LD weighting scheme is quite distinct from exploring the possibility of whether or not a SNP is causal depending on its TLD, which is the role of \( \pi_1 \times p_c \) in the extended model presented here (see below), or the degree to which its effect size is LD-dependent. Rather, within the infinitesimal framework (i.e., assuming a polygenicity of 1) and not explicitly taking the direct effects of LD on z-scores into account (as is done in Eq. 21 noted above), it unwinds the otherwise implicit problem – that SNPs with larger TLD would have over-inflated z-scores – by means of an assumed exponential decay function with respect to base-pair distance (such a function provides an approximation for the gross decay of long-range LD [37, 24]).

In the LD Score regression model, which is another instance of the infinitesimal framework, Finucane et al [12] introduce an annotation-weighted LD score \( l(j,c) \) for variant \( j \) in LD with reference SNPs with (potentially overlapping) categorical annotation \( c \), finding that different categories of SNPs are differentially enriched for heritability. Gazal et al [14] extend this to include continuous-valued annotations, and also introduce rank-based inverse normal transformation of the LD score, denoted LLD, which is calculated independently for different MAF bins and, in this way, unlike LD Score, is intended to be independent of MAF, with the effect size variance for SNP \( j \) given as a sum of terms: a baseline parameter common to all SNPs, plus one of ten parameters depending on which of ten bins the MAF of \( j \) is in, plus another parameter times the LLD of \( j \).

Schoech et al [41] implement a variation of their main model, again in the infinitesimal framework, by augmenting their variance with an additional factor \( (1 - \tau \cdot \text{LLD}_j) \), searching over five values of the new parameter \( \tau \) along with the original 20 values of the selection parameter in a fixed 2D grid.

Within these infinitesimal models [48, 47, 46, 12, 14, 41], it is not clear the degree to which explicit LD-dependence in the variance of the causal effect size merely takes into account the effect on z-scores due to LD with causal SNPs, and how much it models any true effect of LD on underlying causal effect size. Also, such models preclude examining if the TLD of a variant has any bearing on whether the variant is causal. In contrast is the “BayesS” model of Zeng et al [59], a causal mixture model (i.e., not
infinite-simal), using individual genotype data and a reference panel of ~484k non-imputed SNPs, that examines the effects of heterozygosity on effect size. The model we present here is in some respects an extension of that, but based on summary statistics, adding TLD dependence and an additional Gaussian, and we fit the model from a reference panel of 11 million SNPs.

Along with heterozygosity, SNP effect size might independently depend on TLD (for which we use the variable $L$ in equations below), and in principle this could be explored in a manner similar to how heterozygosity is incorporated in the “$c$” causal effect Gaussian variance. However, TLD and heterozygosity are often related (given TLD, the expected heterozygosity shows a distinct well-defined pattern – which may exist [14] – might be difficult to disentangle. Instead, here we explore a separate potential role for TLD.

Patterns of TLD among SNPs are effectively constant in a population over a few generations, but there is no a priori reason why the probability (in a Bayesian sense) of a SNP’s being causally associated with any particular trait in the population should be independent of the SNP’s TLD. For example, there might be lower probability of association for SNPs with larger TLD. Indeed, with the complex interaction of multiple genetic forces such as mutation, genetic drift, and selection, the net role of LD (particularly TLD), through mechanisms like background selection and genetic hitching, on causal association with a particular phenotype is not clear. As noted above, when negative selection is operating (when effects are deleterious to fitness), it is likely that there are increasing numbers of larger effects at lower TLD. If the “$c$” Gaussian is capturing larger effects from rarer SNPs, reflecting selection pressure, it seems reasonable to inquire if the prior probability for a causal SNP’s contribution from the “$c$” Gaussian is TLD-mediated. Here, instead of treating $p_c$ as a constant, we explore the possibility that it is larger for SNPs with lower TLD (in the event that this probability is in fact constant, or increasing, with respect to TLD, we would at least not expect to find it decreasing). This can be accomplished by means of a generalized sigmoidal function that will have a maximum at very low TLD, might maintain that maximum for all SNPs (equivalently, $p_c$ is a constant), or decrease in amplitude slowly or rapidly, possibly to 0, for SNPs with higher TLD. Letting the variable $L$ be the TLD of a SNP, such a function of TLD can be characterized by three parameters: its amplitude (at $L = 1$), the TLD at the midpoint of the sigmoidal transition, and the width of the sigmoidal transition (over a wide or narrow range of TLD). We use the following general form with three parameters, $y_{\text{max}}$, $x_{\text{mid}}$, and $x_{\text{width}}$:

$$y(x) = \frac{y_{\text{max}}}{1 + \exp((x - x_{\text{mid}})/x_{\text{width}})}, \quad (3)$$

defined in the range $-\infty < x < \infty$, for which $y(x)$ is monotonically decreasing and bounded $0 < y(x) < y_{\text{max}}$, with $0 \leq y_{\text{max}} \leq 1$; $-\infty < x_{\text{mid}} < \infty$ locates the mid point of the overall sigmoidal transition ($y(x_{\text{mid}}) = y_{\text{max}}/2$), and $0 < x_{\text{width}} < \infty$ controls its width ($y(x)$ smoothly changing from a Heaviside step function at $x_{\text{mid}}$ as $x_{\text{width}} \to 0$ to a constant function as $x_{\text{width}} \to \infty$). Examples are shown in Figure 1 (scaled by $\pi_1$, giving the physically interesting total prior probability of a SNP being causal with respect to the selection “$c$” Gaussian, as a function of the SNP’s TLD); parameter values are in Supplementary Table S10. Mathematically, the curve can continue into the “negative TLD” range, revealing a familiar full sigmoidal shape; since we are interested in the range $1 \leq x \leq \text{max(TLD)}$, below we report the actual mid point (denoted $m_c$) and width ($w_c$, defined below) of the transition that occurs in this range. Then

$$\beta(H, L) \sim \pi_1 \{ (1 - p_c(L))N(0, \sigma_b^2) + p_c(L)N(0, H^2 \sigma_c^2) \}, \quad (4)$$

where $p_c(L)$ is the sigmoidal function ($0 \leq p_c(L) \leq 1$ for all $L$) given by $y(x)$ in Eq. 3 for $L = x \geq 1$, which numerically can be found by fitting for its three characteristic parameters.

As a final possible extension, we add an extra term – a “$d$” Gaussian – to describe larger effects not well captured in the range, but it seems reasonable to inquire if the prior probability for a causal SNP’s contribution from the “$d$” Gaussian is TLD-mediated.
Figure 2: QQ plots of (pruned) z-scores for qualitative phenotypes (dark blue, 95% confidence interval in light blue) with model prediction (yellow). See Supplementary Material Figures S15 to S22. The value given for $p_\text{u}$ is the amplitude of the full $p_k(L)$ function, which occurs at $L = 1$; the values $(w_c, w_d)$ in parentheses following it are the total LD $(w_c)$ where the function falls to half its amplitude (the middle gray dashed lines in Figure 1 are examples), and the total LD width $(w_d)$ of the transition region (distance between flanking red dashed lines in Figure 1). Similarly for $p_{da} (m_d, w_d)$, where $h_2^b, h_2^c$, and $h_2^d$ are the heritabilities associated with the "b", "c", and "d" Gaussians, respectively. $h_2^b$ is the total SNP heritability, reexpressed as $h_2^b$ on the liability scale for binary phenotypes. Parameter values are also given in Table 1 and heritabilities are also in Table 3; numbers of causal SNPs are in Table 2. Reading the plots: on the vertical axis, choose a $p$-value threshold for typed or imputed SNPs (SNPs with z-scores; more extreme values are further from the origin), then the horizontal axis specified by Eq. 1, 4, or 5.

by the “b” and “c” Gaussians. This gives finally:

$$\beta(H, L) \sim \pi_1 \{ (1 - p_c(L) - p_d(L))N(0, \sigma_0^2) + p_c(L)N(0, H^2 \sigma_2^2) + p_d(L)N(0, \sigma_d^2) \}. \tag{5}$$

where $\sigma_0^2$ is a new parameter, $p_d(L)$ is another general sigmoid function ($0 \leq p_d(L) \leq 1$ for all $L$) where now there is the added constraint $0 \leq p_c(L) + p_d(L) \leq 1$, and the prior probability for the “b” Gaussian becomes $p_b(L) \equiv 1 - p_c(L) - p_d(L)$.

Depending on the phenotype and the GWAS sample size, it might not be feasible, or meaningful, to implement the full model. In particular, for low sample size and/or low discoverability, the “b” Gaussian is all that can be estimated, but in most cases both the “b” and “c” Gaussians can be estimated, and $\beta$ will be well characterized by Eq. 4.

As we described in our previous work [22], a z-score is given by a sum of random variables, so the a posteriori pdf (given the SNP’s heterozygosity and LD structure, and the phenotype’s model parameters) for such a composite random variable is given by the convolution of the pdfs for the component random variables. This facilitates an essentially exact calculation of the z-score’s a posteriori pdf arising from the underlying model of causal effects as specified by Eq. 1, 4, or 5.

For our reference panel we used the 1000 Genomes phase 3 data set for 503 subjects/samples of European ancestry [6, 52]. In ref. [22] we describe how we set up the general framework for analysis, including the use of the reference panel and dividing the reference SNPs into a sufficiently fine 10×10 heterozygosity×TLD grid to facilitate computations.

In our earlier work on the “b” model we gave an expression, denoted $G(k)$, for the Fourier transform of the genetic contribution to a z-score, where $k$ is the running Fourier parameter. The extra complexity in the “c” and “d” models here requires a modification only in this term, which we describe in the Methods. In addition to the parameters presented above, we also include an inflation parameter $\sigma_0$: if $z_0$ denotes uninflated GWAS z-scores and $z$ denotes actual GWAS z-scores, then $\sigma_0$ is defined by $z = \sigma_0 z_0$ [22, 9]. The optimal model parameters for a particular phenotype are estimated by minimizing the negative of the log likelihood of the data (z-scores) as a function of the parameters. This is done with the “b” model as before [22], and then proceeding iteratively with the more complex models, continually re-estimating the new values of the earlier parameters that maximize the likelihood when a new parameter is introduced, with extensive single and multiple parameter searches (to avoid being trapped in local minima) until all parameters simultaneously are at a convex minimum. This involved many explicit and repeated coarse- and fine-
grained linear (for each individual parameter) and grid (for multiple parameters simultaneously) searches as well as many Nelder-Mead multidimensional unconstrained nonlinear minimizations. There is no artificial constraining on the parameters; for example, no prior assumption is made about the relative sizes of the causal $\sigma$’s, or the parameter values of the LD-dependent prior probabilities (the full range of amplitudes, transition widths, and location of the mid-point of the transitions are searched).

The total SNP heritability is given by the sum of heritability contributions of each SNP in the reference panel from each of the relevant Gaussians. In the Bayesian approach, we do not know the value of the causal effect of any particular SNP, but we assume it comes from a distribution, $\beta(H, L)$, which characterize our ignorance of it, where $H$ is the SNP’s heterozygosity and $L$ is its total LD.

For a specific Gaussian ("b", "c", or "d") in our model, the contribution of the SNP to heritability is given by the prior probability that the SNP is causal is $\pi_p c(L)$, times the expected value of the square of the effect size, $E(\beta^2)$, times $H$. For the "c" Gaussian, for example, the prior probability that the SNP is causal is $\pi_p c(L)$, and $E(\beta^2)$ is just the variance, $H^2 \sigma_c^2$. Thus, the contribution of this SNP to the overall heritability associated with the "c" Gaussian, $h_c^2$, is $\pi_p c(L)H^2 \sigma_c^2 H$. Below we report the sums over all such contributions. The number of causal SNPs associated with the "c" Gaussian, $n_c$, is given by summing $\pi_p c(L)$ for each reference panel SNP, and similarly for the other Gaussians. All heritabilities and discoverabilities are, as before [22], corrected with respect to the inflation parameter, i.e., divided by $\hat{\sigma}_0^2$.

All code used in the analyses, including simulations, is publicly available on GitHub [20, 19].

**Data Preparation**

We analyzed summary statistics for fourteen phenotypes—genotypes (in what follows, where sample sizes varied by SNP, we quote the median value): (1) bipolar disorder ($N_{cases} = 20,352$, $N_{controls} = 31,358$) [50]; (2) schizophrenia ($N_{cases} = 35,476$, $N_{controls} = 46,839$) [40]; (3) coronary artery disease ($N_{cases} = 60,801$, $N_{controls} = 123,504$) [32]; (4) ulcerative colitis ($N_{cases} = 12,366$, $N_{controls} = 34,915$) and (5) Crohn’s disease ($N_{cases} = 12,194$, $N_{controls} = 34,915$) [8]; (6) late onset Alzheimer’s disease (LOAD; $N_{cases} = 17,008$, $N_{controls} = 37,154$) [25] (in the Supplementary Material we present results for a more recent GWAS with $N_{cases} = 71,880$ and $N_{controls} = 383,378$) [23]; (7) amyotrophic lateral sclerosis (ALS) ($N_{cases} = 12,577$, $N_{controls} = 23,475$) [53]; (8) number of years of formal education ($N = 293,723$) [33]; (9) intelligence ($N = 262,529$) [44, 39]; (10) body mass index (UKB-GIANT 2018) ($N = 690,519$) [58]; (11) height (2010) ($N = 133,735$) [57]; and height (2014) ($N = 251,747$) [55]; (12) low- ($N = 94,295$) and (13) high-density lipoprotein ($N = 94,579$) [54]; and (14) total cholesterol ($N = 94,579$) [54]. Most participants were of European ancestry. (A spreadsheet giving data sources is provided in the Supplement.) In the tables, we also report results for body mass index (GIANT 2015) ($N = 233,554$) [28], and height (UKB-GIANT 2018) [58].

In Figures 2 and S1 we report effective sample sizes, $N_{eff}$, for the case-control GWASs. This is defined as $N_{eff} = 4/(1/N_{cases} + 1/N_{controls})$, so that when $N_{cases} =$
In estimating heritabilities on the liability scale for the qualitative phenotypes [22], we assumed prevalences of: BIP 0.5% [31], SCZ 1% [47], CAD 3% [38], UC 0.1% [4], CD 0.1% [4], AD 14% (for people aged 71 and older in the USA [35, 1]), and ALS 5 × 10⁻⁵ [30].

Confidence intervals for parameters were estimated using the inverse of the observed Fisher information matrix (FIM). The full FIM was estimated for up to eight parameters used in Model C, and for the remaining parameters that extend the analysis to Model D the confidence intervals were approximated ignoring off-diagonal elements. Additionally, the $w_d$ parameter was treated as fixed quantity, the lowest value allowing for a smooth transition of the $p_d(L)$ function to 0 (see Supplementary Material Figure S7; for CD, UC, and TC, however, the function $p_d(L)$ was a constant (= $p_d(1)$). For the derived quantities $h^2$ and $n_{causal}$, which depend on multiple parameters, the covariances among the parameters, given by the off-diagonal elements of the inverse of the FIM, were incorporated. Numerical values are in Supplementary Material Tables S5-S7.

In order to carry out realistic simulations (i.e., with realistic heterozygosity and LD structures for SNPs), we used HAPGEN2 [27, 49, 51] to generate genotypes for 10⁵ samples; we calculated SNP MAF and LD structure from 1000 simulated samples.

RESULTS

Phenotypes

Summary QQ plots for pruned z-scores are shown in Figure 2 for seven binary phenotypes (for AD we separate out chromosome 19, which contains the APOE gene), and in Figure 3 for seven quantitative phenotypes (including two separate GWAS for height), with model parameter values in Table 1; breakdowns of these summary plots with respect to a 4x4 grid of heterozygosity×TLD (each grid a subset of a 10×10 grid) are in Supplementary Material Figures S15-S29. For each phenotype, model selection (B, C, or D) was performed by testing the Bayesian information criterion (BIC) – see Supplementary Table S4. For comparison, all QQ figures include the basic (B) model in green; the extended model C or D, consistently demon-

| Phenotype | $\pi_1$ | $\sigma_b^2$ | $\sigma_c^2$ | $S$ | $p_{c1}$ | $m_c$ | $w_c$ | $\sigma_d^2$ | $p_{d1}$ | $m_d$ | $w_d$ | $\sigma_0^2$ |
|-----------|---------|--------------|-------------|-----|---------|------|------|------------|---------|------|------|----------|
| SCZ 2014 | 5.38e-2 | 1.46-6       | 3.65-5      | -0.52 | 0.07 | 87 | 352 | 1.4e-4 | 5.2e-3 | 519 | 7 | 1.07 |
| BIP      | 5.91e-2 | 1.26-6       | 4.56-5      | -0.40 | 6.7e-2 | 102 | 414 | —       | —       | —       | —       | 1.01 |
| CD       | 9.55e-4 | 5.06-5       | 5.45-6      | -0.64 | 0.20 | 176 | 604 | 7.6e-2 | 1.0e-4 | —       | —       | 1.14 |
| UC       | 1.16e-3 | 3.65-6       | 4.06-4      | -0.67 | 0.16 | 173 | 627 | 8.0e-2 | 1.0e-4 | —       | —       | 1.12 |
| CAD      | 1.88e-3 | 1.16-5       | 9.25e-5     | -0.51 | 6.9e-2 | 171 | 615 | 5.3e-3 | 3.6e-4 | 102 | 7 | 0.92 |
| AD Chr19 | 4.3ae-4 | 1.06-4       | 6.1e-3      | -0.57 | 0.73 | 35 | 89 | —       | —       | —       | —       | 1.09 |
| AD NoC19 | 1.05e-3 | 1.85-6       | 2.6e-4      | -0.52 | 2.9e-2 | 264 | 6 | —       | —       | —       | —       | 1.04 |
| ALS Chr9 | 1.12e-2 | 7.36-6       | 3.9e-3      | -0.01 | 1.8e-3 | 106 | 6 | —       | —       | —       | —       | 0.99 |
| Edu      | 1.43e-2 | 1.76-6       | 7.8e-6      | -0.44 | 0.45 | 111 | 339 | 8.5e-5 | 6.3e-3 | 411 | 7 | 0.94 |
| IQ 2018  | 1.27e-2 | 7.5e-6       | 6.2e-6      | -0.51 | 0.38 | 122 | 309 | 3.6e-5 | 7.8e-2 | 561 | 6 | 1.17 |
| Height 2010 | 1.02e-3 | 4.15-6     | 2.0e-4      | -0.44 | 0.20 | 322 | 1243 | —       | —       | —       | —       | 0.90 |
| Height 2014 | 1.15e-3 | 3.76-6     | 1.6e-4      | -0.46 | 0.21 | 242 | 929 | —       | —       | —       | —       | 1.57 |
| Height 2018 | 2.50e-3 | 8.76-6     | 8.9e-5      | -0.43 | 0.37 | 210 | 739 | —       | —       | —       | —       | 2.12 |
| HDL      | 2.54e-3 | 1.16-5       | 4.5e-4      | -0.79 | 1.4e-2 | 143 | 599 | 2.2e-2 | 1.1e-3 | 66 | 7 | 0.91 |
| LDL      | 5.84e-3 | 3.36-6       | 2.4e-4      | -0.52 | 8.9e-3 | 336 | 1471 | 7.3e-3 | 2.2e-4 | 346 | 6 | 0.92 |
| BMI GIANT 2015 | 1.54e-3 | 2.2e-5     | 4.5e-4      | 0.00 | 4.6e-3 | 288 | 12 | —       | —       | —       | —       | 0.85 |
| BMI 2018 | 2.34e-3 | 1.6e-5       | 3.0e-4      | 0.11 | 3.8e-3 | 268 | 8 | —       | —       | —       | —       | 1.72 |
| TC       | 1.15e-3 | 1.7e-5       | 6.2e-4      | -0.97 | 2.1e-2 | 140 | 583 | 2.9e-4 | 3.4e-2 | —       | —       | 0.92 |

Table 1: Model parameters for phenotypes, case-control (upper section) and quantitative (lower section). $\pi_1$ is the overall proportion of the 11 million SNPs from the reference panel that are estimated to be causal. $p_c(L \geq 1)$ is the prior probability multiplying the “c” Gaussian, which has variance $H^2\sigma_c^2$, where $H$ is the reference SNP heterozygosity. Note that $p_c(L)$ is just a sigmoidal curve, and can be characterized quite generally by three parameters: the value $p_c(L) \equiv p_c(1)$ at $L = 1$; the total LD value $L = m_c$ at the mid point of the transition, i.e., $p_c(m_c) = p_c(1)/2$ (see the middle gray dashed lines in Figure 1, which shows examples of the function $p_c(L)$); and the width $w_c$ of the transition, defined as the distance (in L) between where the curve falls to 95% and 5% of $p_c(1)$ (distance between the flanking red dashed lines in Figure 1). Note that for AD Chr19, AD NoC19, and ALS Chr9, $\pi_1$ is the fraction of reference SNPs on chromosome 19, on the autosome excluding chromosome 19, and on chromosome 9 respectively. For bipolar disorder, Crohn’s disease, ulcerative colitis, and total cholesterol $p_d(L) = p_{d1}$ for all L. Examples of $H^2\sigma_d^2$ multiplying $\sigma_d^2$ are shown in Supplementary Material Figure S9. Estimated Bayesian information criterion (BIC) values for three models (B, C, and D) are shown in Supplementary Material Table S4: the 3-parameter model B with only the “b” Gaussian ($\pi_1, \sigma_b, \sigma_c$); the 8-parameter model C with both the “b” with “c” Gaussians (Eq. 4); and the 12-parameter model D with “b”, “c” and “d” Gaussians (Eq. 5). 95% confidence intervals are in Supplementary Materials Tables S5-S7.
Table 2: Numbers of causal SNPs. \( n_{\text{causal}} \) is the total number of causal SNPs (from the 11 million in the reference panel); \( n_b \), \( n_c \), and \( n_d \) are the numbers associated with the “b”, “c”, and “d” Gaussians, respectively. 95% confidence intervals are in Supplementary Materials Table S8.

| Phenotype        | \( n_b \)  | \( n_c \)  | \( n_d \)  | \( n_{\text{causal}} \) |
|------------------|------------|------------|------------|--------------------------|
| SCZ 2014         | 5.6e5      | 2.3e4      | 3.0e3      | 5.82e5                   |
| BIP              | 6.2e5      | 2.6e4      | —          | 6.51e5                   |
| CD               | 9.0e3      | 1.5e3      | 1          | 1.05e4                   |
| UC               | 1.1e4      | 1.4e3      | 2          | 1.27e4                   |
| CAD              | 2.0e4      | 1.0e3      | 5          | 2.07e4                   |
| AD Chr9          | 80         | 33         | —          | 113                      |
| AD NoC19         | 1.1e4      | 294        | —          | 1.12e4                   |
| ALS Chr9         | 5.3e3      | 7          | —          | 5.29e3                   |
| Edu              | 1.1e5      | 4.4e4      | 956        | 1.58e5                   |
| IQ 2018          | 9.6e4      | 3.4e4      | 1.1e4      | 1.40e5                   |
| Height 2010      | 9.4e3      | 1.9e3      | —          | 1.13e4                   |
| Height 2014      | 1.1e4      | 2.0e3      | —          | 1.26e4                   |
| Height 2018      | 2.0e4      | 7.6e3      | —          | 2.75e4                   |
| HDL              | 2.7e4      | 264        | 15         | 2.79e4                   |
| LDL              | 6.4e4      | 463        | 14         | 6.43e4                   |
| BMI 2015         | 1.7e4      | 68         | —          | 1.69e4                   |
| BMI 2018         | 2.6e4      | 88         | —          | 2.57e4                   |
| TC               | 1.2e4      | 180        | 434        | 1.27e4                   |

The distributions of z-scores for different phenotypes are quite varied. Nevertheless, for most phenotypes analyzed here, we find evidence for larger and smaller effects being distributed differently, with strong dependence on total LD, \( L \), and heterozygosity, \( H \).

Our model estimates polygenicity as a one-dimensional function of \( L \). We find that polygenicity is dominated by SNPs with low \( L \). However, the degree of restriction varies widely across phenotypes, depending on the shapes and sizes of \( p_c(L) \) and \( p_d(L) \) in Eq. 5, the prior probabilities that a causal SNP belongs to the “c” and “d” Gaussians. These prior probabilities are shown in Figure 1 and Supplementary Figures S5-S7. Taking into account the underlying distribution of reference SNPs with respect to heterozygosity, these distributions lead to a varied pattern across phenotypes of the expected number of causal SNPs in equally-spaced elements in a two-dimensional \( H \times TLD \) grid, as shown for height (2014) in Figure 4 C, and for all phenotypes in Supplementary Figures S11-S14 (third columns). Further, for any given phenotype, the effect sizes of causal variants come from distributions whose variances can be widely different – by up to two orders of magnitude. Thus, given the prior probabilities \( (p_b, p_c, \) and \( p_d) \) by which these distributions are modulated as a function of \( L \), we are able to estimate the expected effect size per causal SNP, \( E(\beta^2) \), in each \( H \times TLD \) grid element, as shown in Figure 4 D and Supplementary Figures S11-S14 (fourth columns). In general, depending on the

Table 3: Heritabilities: \( h^2 \) is the total additive SNP heritability, re-expressed on the liability scale as \( h^2_l \) for the qualitative traits (upper section). \( h^2_{bc}, h^2_{cd}, \) and \( h^2_d \) are the heritabilities associated with the “b”, “c”, and “d” Gaussians, respectively. The last column, labeled \( %c \), gives the percentage of SNP heritability that comes from the “c” Gaussian. 95% confidence intervals are in Supporting Materials Table S9. A comparison of the total heritabilities with estimates from our basic model alone [22], and with estimates from other models [60, 59, 41] are given in Supplementary Table S1.

| Phenotype        | \( h^2_b \)  | \( h^2_c \)  | \( h^2_d \)  | \( h^2 \)  | \( h^2_d \)  | \( %c \)  |
|------------------|------------|------------|------------|----------|------------|--------|
| SCZ 2014         | 0.16       | 0.31       | 0.09       | 0.56     | 0.31       | 55.1   |
| BIP              | 0.16       | 0.37       | —          | 0.53     | 0.26       | 69.7   |
| CD               | 0.10       | 0.40       | 0.02       | 0.52     | 0.24       | 77.5   |
| UC               | 0.09       | 0.29       | 0.02       | 0.41     | 0.18       | 72.3   |
| CAD              | 0.05       | 0.04       | 0.00       | 0.09     | 0.07       | 41.3   |
| AD Chr9          | 0.00       | 0.08       | —          | 0.08     | 0.11       | 97.4   |
| AD NoC19         | 0.04       | 0.03       | —          | 0.07     | 0.10       | 42.1   |
| ALS Chr9         | 0.01       | 0.00       | —          | 0.01     | 0.00       | 37.1   |
| Edu              | 0.04       | 0.11       | 0.02       | 0.18     | —          | 64.9   |
| IQ 2018          | 0.02       | 0.08       | 0.08       | 0.18     | —          | 44.4   |
| Height 2010      | 0.08       | 0.13       | —          | 0.22     | —          | 61.0   |
| Height 2014      | 0.09       | 0.12       | —          | 0.21     | —          | 58.4   |
| Height 2018      | 0.04       | 0.24       | —          | 0.28     | —          | 86.0   |
| HDL              | 0.06       | 0.08       | 0.05       | 0.19     | —          | 40.1   |
| LDL              | 0.05       | 0.05       | 0.02       | 0.11     | —          | 40.6   |
| BMI 2015         | 0.08       | 0.01       | —          | 0.08     | —          | 7.6    |
| BMI 2018         | 0.00       | 0.00       | —          | 0.09     | —          | 5.2    |
| TC               | 0.04       | 0.10       | 0.03       | 0.18     | —          | 59.0   |
Simulations
To test the specificity of the model for each real phenotype, we constructed simulations where, in each case, the true causal \( \beta \)'s (a single vector instantiation) for all reference panel SNPs were drawn from the overall distribution defined by the real phenotype’s parameters (thus being the "true" simulation parameters). We set up simulated phenotypes for 100,000 samples by adding noise to the genetic component of the simulated phenotype [22], and performed a GWAS to calculate z-scores. We then sought to determine whether the true parameters, and the component heritabilities, could reasonably be estimated by our model. In Supplementary Figures S3 and S4 we show the results for the simulated case-control and quantitative phenotypes, respectively. Overall heritabilities were generally faithful to the true values (the values estimated for the real phenotypes) – see Supplementary Table S3 – though for Crohn’s disease the simulated value was overestimated due to the \( h_d \) component. Note that for the case-control simulated phenotypes, the heritabilities on the observed scale, denoted \( \hat{h}^2 \) in Supplementary Figure S3, should be compared with the corresponding values in Figure 2, not with \( \hat{h}_i^2 \), which denotes heritability on the liability scale, i.e., adjusted for population prevalence; note also that for case-control phenotypes, we implicitly assume the same proportion \( P \) of cases in the real and simulated GWAS (for the basic model, assuming the liability-scale model is correct, one can easily see that discoverability \( \sigma_d^2 \propto P(1-P) \) – see Eqs. 38 and 39 in [22]; this carries over to \( \sigma_b, \sigma_c, \) and \( \sigma_d \) used here). Polygenicities and discoverabilities were also generally faithfully reproduced. However, for ALS restricted to chromosome 9, and BMI, the selection parameter was incorrectly estimated, owing to the weak signal in these GWAS (e.g., for BMI \( \pi_{1P_L}(1) \simeq 8 \times 10^{-6} \), Fig. S6 (A), and only around 5% of SNP heritability was found to be associated with the "c" Gaussian, Table 3) and very low polygenicity (small number of causal SNPs) for the "c" Gaussian. Given the wide variety and even extreme ranges in parameters and heritability components across diverse simulated phenotypes, the multiple simulated examples provide checks respect to each-other for correctly discriminating phenotypes by means of their model parameter estimates: the results for individual cases are remarkably faithful to the respective true values, demonstrating the utility of the model in distinguishing different phenotypes.

DISCUSSION
We propose an extended Gaussian mixture model for the distribution of underlying SNP-level causal genetic effects in human complex phenotypes, allowing for the phenotype-specific distribution to be modulated by heterozygosity, \( H \), and total LD, \( L \), of the causal SNPs, and also allowing for independent distributions for large and small effects. The GWAS z-scores for the typed or imputed SNPs, in addition to having a random environmental and error contribution, arise through LD with the causal SNPs. Thus, taking the detailed LD and heterozygosity structure of the population into account by using a reference panel, we are able to model the distribution of z-scores and test the applicability of our model to human complex phenotypes.

Complex phenotypes are emergent phenomena arising from random mutations and selection pressure. Underlying causal variants come from multiple functional categories [42], and heritability is known to be enriched for some functional categories [12, 43]. Thus, it is likely that different variants will experience different evolutionary pressure – negative, neutral, or positive – due at least to pleiotropic roles. Indeed, heritability has been found partially to depend on functional annotations differentially correlated with levels of LD [14], consistent with the operation of negative selection pressure.

Here, we find evidence for markedly different genetic architectures across diverse complex phenotypes, where the polygenicity (or, equivalently, the prior probability that a SNP is causal) is a function of SNP total LD (\( L \)), and discoverability is multi-component and MAF dependent.

In contrast to previous work modeling the distribution of causal effects that took total LD and multiple functional annotation categories into account while implicitly assuming a polygenicity of 1 [14], or took MAF into account while ignoring total LD dependence and different distributions for large and small effects [59], or took independent distributions for large and small effects [45], or took independent distributions for large and small effects into account (which is related to incorporating multiple functional annotation
categories) while ignoring total LD and MAF dependence, here we combine all these issues in a unified way, using an extensive underlying reference panel of ~11 million SNPs and an exact methodology using Fourier transforms to relate summary GWAS statistics to the posited underlying distribution of causal effects [22]. We show that the distributions of all sets of phenotypic z-scores, including extreme values that are well within genome-wide significance, are accurately reproduced by the model, both at overall summary level and when broken down with respect to a 10×10 H×TLD grid – even though the various phenotypic polygenicities and per-causal-SNP heritabilities range over orders of magnitude. Improvement with respect to the basic model can be seen in the summary QQ plots Figs. 2 and 3, and also the H×TLD grids of subplots Figs. S15-S29. But even when model fits are similar, the extended model fit taking selection pressure into account results in higher likelihood and lower Bayesian information criterion.

As noted in the Introduction, negative selection can be expected to result in an increasing number of effects with increasing effect size at lower total LD and lower heterozygosity. It was found in [14] – which, in addition to analyzing total LD, modeled allele age and recombination rates – that common variants associated with complex traits are weakly deleterious to fitness, in line with an earlier model result that most of the variance in fitness comes from rare variants that have a large effect on the trait in question [11]. Thus, larger per-allele effect sizes for less common variants is consistent with the action of negative selection.

Further, based on a model equivalent to Eq. 2 with \( p_c \equiv 1 \), it was argued in [59], using forward simulations and a commonly used demographic model [16], that negative values for the selection parameter, \( S \), which leads to larger effects for rarer variants, is a signature of negative selection. Similar results were found for the related infinitesimal model \( (p_c \equiv 1 \text{ and } \pi_1 \equiv 1) \) in [41].

We find negative selection parameter values for most traits, which is broadly in agreement with [59] with the exception of BMI, which we can find to be modeled with two Gaussians with no or weakly positive (\( S \geq 0 \)) heterozygosity dependence, though it should be noted that the polygenicity for the larger-effects Gaussian (the “c” Gaussian with the S parameter) is very low, amounting to an estimate of < 100 common causal SNPs of large effect. A similar situation (\( S \approx 0 \)) obtains with ALS restricted to chromosome 9; here, the sample size is relatively low, leading to a weak signal, and we estimate only 7 common causal SNPs associated with the “c” Gaussian. Our BMI results are not in agreement with the results of others. For the UKB (2015) data, Zeng at al [59] report a selection parameter \( S = -0.283 \) [-0.377, -0.224], \( h^2 = 0.276 \), and \( n_{\text{causal}} = 45k \), while Schoech et al [41] report a selection parameter \( \alpha = -0.24 \) [-0.38, -0.06], and \( h^2 = 0.31 \). For the same GIANT 2015 dataset used here [28], Zhang et al [60] report \( h^2 = 0.20 \) and \( n_{\text{causal}} = 18k \). It is not clear why our BMI results are in such disagreement with these. For height, our selection parameter \( S = -0.46 \) (2014 GWAS) is in reasonable agreement with \( S = -0.422 \) reported in [59], and with \( \alpha = -0.45 \) reported in [41].

Compared with our basic model results [22], we generally have found that using the extended model presented here heritabilities show marked increases, with large contributions form the selection (“c”) Gaussian – see Supplementary Table S1. However, due to the small effect sizes associated with the “b” Gaussian as a component of the extended model, the overall number of causal SNPs can be considerably larger. Supplementary Figures S11-S14 capture the breakdown in SNP contributions to heritability as a function of heterozygosity and total linkage disequilibrium.

Generally, we find evidence for the existence of genetic architectures where the per-causal-SNP heritability is larger for more common SNPs and/or for SNPs with lower total LD. But the trend is not uniform across phenotypes – see Supplementary Figures S11-S14, second columns. The observed-scale heritability estimates given by \( h^2 \) correspond to effects not experiencing much selection pressure. The new final values presented here result from a model that gives better fits between data and model prediction of the summary and detailed QQ plots, and thus constitute more accurate estimates of SNP heritability. For 2010 and 2014 height GWASs, we obtain very good consistency for the model parameters and therefore heritability, despite considerable difference in apparent inflation. The 2018 height GWAS [58] has a much larger sample size (almost three quarters of a million); the slightly different parameter estimates for it might arise due to population structure not fully captured by our model [45, 2]. Our \( h^2 \) estimates for height, however, remain consistently lower than other reported results (e.g., \( h^2 = 0.33 \) in [60], \( h^2 = 0.527 \) in [59], \( h^2 = 0.61 \) in [41]). For educational attainment, our heritability estimate agrees with [59] (\( h^2 = 0.182 \), despite our using a sample more than twice as large. The difference in selection parameter value, \( S = -0.335 \) in [59] versus \( S = -0.44 \) here, might partially be explained by the model differences (Zeng et al use one causal Gaussian with MAF-dependence but no LD dependence). A comparison of the total heritabilities reported here with estimates from the work of others [60, 59, 41], in addition our basic model alone [22], are given in Supplementary Table S1.

For most traits, we find strong evidence that causal SNPs with low heterozygosity have larger effect sizes (\( S < 0 \) in Table 1; the effect of this as an amplifier of \( \sigma^2 \) in Eq. 5 is illustrated in Supplementary Material Figure S9) – see also Supplementary Figures S11-S14, fourth columns. Thus, negative selection seems to play an important role in most phenotypes-genotypes. This is also indicated by the extent of the region of finite probability for variants of large effect sizes (which are enhanced by having \( S \lesssim -0.4 \) being relatively rare (low \( H \)), which will be greater for larger \( p_{c1} \), the amplitude of the prior probability for the “c” Gaussian (see Supplementary Figures S5 and S6).

The “b” Gaussian in Eq. 4 or 5 does not involve a selection parameter: effect size variance is independent of
MAF. Thus, causal SNPs associated with this Gaussian are likely undergoing neutral (or very weakly negative) selection. It should be noted that in all traits examined (S < 0), the effect size variance of the “b” Gaussian is many times smaller – sometimes by more than an order of magnitude – than that for the “c” Gaussian. Thus, it appears there are many causal variants of weak effect undergoing neutral (or very weakly negative) selection. For the nine phenotypes where the “d” Gaussian could be implemented, its variance parameter was several times larger than that of the “c” Gaussian. However, the amplitude of the prior probability for the “d” Gaussian, pD1, was generally much smaller than the amplitudes of the prior probabilities for the “b” or “c” Gaussians, which translated into a relatively small number of causal variants with very large effect associated with this Gaussian. (Due to lack of power, in three instances – CD, UC, and TC – pD(L) was treated as a constant, i.e., independent of L.) Interestingly, intelligence had the highest number of causal SNPs associated with this Gaussian, while the extent of total LD for associated SNPs was also liberal (mD = 561; see also Supplementary Figure S7). It is possible that some of these SNPs are undergoing positive selection, but we did not find direct evidence of that.

We find a diversity of genetic architectures across multiple human complex phenotypes. We find that SNP total LD plays an important role in the likelihood of the SNP being causal, and in the effect size of the SNP. In general, we find that lower total LD SNPs are more likely to be causal with larger effects. Furthermore, for most phenotypes, while taking total LD into account, we find that causal SNPs with lower MAF have larger effect sizes, a phenomenon indicative of neutral selection. Additionally, for all phenotypes, we found evidence of neutral selection operating on SNPs with relatively weak effect. Apart from a weak effect in BMI, we did not find direct evidence of positive selection. Future work will explore SNP functional annotation categories and their differential roles in human complex phenotypes.

**METHODS**

**Total Linkage Disequilibrium**

Sequentially moving through each chromosome in contiguous blocks of 5,000 SNPs in the reference panel, for each SNP in the block we calculated its Pearson r² correlation coefficients (that arise from linkage disequilibrium, LD) with all SNPs in the central block itself and with all SNPs in the pair of flanking blocks of size up to 25,000 each. For each SNP we calculated its total linkage disequilibrium (TLD), given by the sum of LD r²’s thresholded such that if r² < r² min we set that r² to zero (r² min = 0.05). The fixed window size corresponds on average to a window of ±8 centimorgans. This is deliberately larger than the 1-centimorgan window used to define LD Score [3], because the latter appears to exclude a noticeable part of the LD structure.

In applying the model to summary statistics, we calculated histograms of TLD (using 100 bins) and ignoring SNPs whose TLD was so large that their frequency was less than a hundredth of the respective histogram peak; typically this amounted to restricting to SNPs for which TLD ≤ 600. We also ignored summary statistics of SNPs for which MAF ≤ 0.01.

Since we are estimating three parameters from millions of data points, it is reasonable to coarse-grain the data. Knowing the TLD and heterozygosity (H) of each SNP, we divided the full set of GWAS SNPs into a H×TLD coarse-grained grid; we found that 10×10 is more than sufficient for converged results.

**Model PDF**

When implementing the discrete Fourier transform (DFT) to calculate model a posteriori probabilities for z-score outcomes for a single SNP, we discretize the range of possible z-scores into the ordered set of n (equal to a power of 2) values z₁, . . . , zₙ with equal spacing between neighbors given by ∆z (zₙ = −z₁ − ∆z, and zₙ/2+1 = 0). Taking z₁ = −38 allows for the minimum p-values of 5.8 × 10⁻³¹⁶ (near the numerical limit); with n = 2¹⁰, ∆z = 0.0742. Given ∆z, the Nyquist critical frequency is f_c = 1/2∆z, so we consider the Fourier transform function for the z-score pdf at n discrete values k₁, . . . , kₙ, with equal spacing between neighbors given by ∆k, where k₁ = −f_c (kₙ = −k₁ − ∆k, and kₙ/2+1 = 0; the DFT pair ∆z and ∆k are related by ∆z∆k = 1/n).

In our earlier work on the “b” model [22] we gave an expression, denoted G(k_j), for the Fourier transform of the genetic contribution to a z-score, where k_j is the discrete Fourier variable described above. In constructing the a posteriori pdf for a z-score, the extra complexity in the “c” and “d” models presented in the current work requires a modification only in this term.

The set of typed SNPs are put in a relatively fine 10×10 grid, called the “H-L” grid, based on their heterozygosity and total LD (coarser grids are refined until converged results are achieved, and 10×10 is more than adequate). Given a typed SNP in LD with many tagged SNPs in the reference panel (some of which might be causal), divide up those tagged SNPs based on their LD with the typed SNP into w max LD-r²² windows (we find that w max = 20, dividing the range 0 ≤ r² ≤ 1 into 20 bins, is more than adequate to obtain converged results); let r²ₜ denote the LD of the wᵗʰ bin, w = 1, . . . , w max. Denote the number of tagged SNPs in window w as nₜ and their mean heterozygosity as Hₜ. Then, given model “b” parameters π₁, σ₂, and σ₄, and sample size N, G(k_j) is given by

\[
G(k_j) = \prod_{w=1}^{w max} \left( \pi_1 \exp(B_{w k_j^2}) + (1 - \pi_1) \right)^{n_w}, \tag{6}
\]
\[ B_w = -2\pi^2 N H_w r_w^2 \hat{\sigma}_b^2 \]  \hspace{1cm} (7)
and \[ \hat{\sigma}_b^2 \equiv \sigma_b^2 / \sigma_0^2 \]. For model “c”, this becomes
\[ G(k_j) = \prod_{w=1}^{w_{\text{max}}} \left( \pi_1 \left[ (1 - p_c(L_w)) \exp(B_w k_j^2) + p_c(L_w) \exp(C_w k_j^2) \right] + (1 - \pi_1) \right)^{n_w} \]  \hspace{1cm} (8)
where
\[ C_w = -2\pi^2 N H_w r_w^2 \hat{\sigma}_c^2 H_w^S \]  \hspace{1cm} (9)
\( S \) is the selection parameter, \( \hat{\sigma}_c^2 \equiv \sigma_c^2 / \sigma_0^2 \) (\( \sigma_c^2 \) is defined in Eq. 2), \( L_w \) is the mean total LD of reference SNPs in the \( w \) bin, and \( p_c(L_w) \) is the sigmoidal function (see Eq. 3) giving, when multiplied by \( \pi_1 \), the prior probability of reference SNPs with this TLD being causal (with effect size drawn from the “c” Gaussian). For model “d”, \( G(k_j) \) becomes
\[ G(k_j) = \prod_{w=1}^{w_{\text{max}}} \left( \pi_1 \left[ (1 - p_c(L_w) - p_d(L_w)) \exp(B_w k_j^2) + p_c(L_w) \exp(C_w k_j^2) + p_d(L_w) \exp(D_w k_j^2) \right] + (1 - \pi_1) \right)^{n_w} \]  \hspace{1cm} (10)
where
\[ D_w = -2\pi^2 N H_w r_w^2 \hat{\sigma}_d^2 \]  \hspace{1cm} (11)
\( \hat{\sigma}_d^2 \equiv \sigma_d^2 / \sigma_0^2 \) (\( \sigma_d^2 \) is defined in Eq. 5), and \( p_d(L_w) \) is the sigmoidal function (again, see Eq. 3) giving, when multiplied by \( \pi_1 \), the prior probability of reference SNPs with this TLD being causal (with effect size drawn from the “d” Gaussian).

Let \( H \) denote the LD and heterozygosity structure of a particular SNP, a shorthand for the set of values \{\( n_w, H_w, L_w : w = 1, \ldots, w_{\text{max}} \)\} that characterize the SNP, and let \( M \) denote the set of model parameters for whichever model – “b”, “c”, or “d” – is implemented. The Fourier transform of the environmental contribution [22], denoted \( E_j \equiv E(k_j) \), is
\[ E(k_j) = \exp(-2\pi^2 \sigma_0^2 k_j^2). \]  \hspace{1cm} (12)
Let \( F_F = (G_1 E_1, \ldots, G_n E_n) \), where \( G_j \equiv G(k_j) \), denote the vector of products of Fourier transform values, and let \( F^{-1} \) denote the inverse Fourier transform operator. Then for the SNP in question, the vector of pdf values, \( \text{pdf}_z \), for the uniformly discretized possible z-score outcomes \( z_1, \ldots, z_n \) described above, i.e., \( \text{pdf}_z = (f_1, \ldots, f_n) \) where \( f_i \equiv \text{pdf}(z_i | H, M, N) \), is
\[ \text{pdf}_z = F^{-1} [F_F]. \]  \hspace{1cm} (13)
Thus, the \( i \)th element \( \text{pdf}_z_i = f_i \) is the \( a \ posteriori \) probability of obtaining a z-score value \( z_i \) for the SNP, given the SNP’s LD and heterozygosity structure, the model parameters, and the sample size.

### Funding
Research Council of Norway (262656, 248984, 248778, 223273) and KG Jebsen Stiftelsen; NIH Grant Number U24DA041123.

### References
[1] Alzheimer’s Association, 2018. 2018 Alzheimer’s disease facts and figures. Alzheimer’s & Dementia 14 (3), 367–429.
[2] Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., Boyle, E. A., Zhang, X., Racimo, F., Pritchard, J. K., et al., 2018. Reduced signal for polygenic adaptation of height in uk biobank. eLife 8, e39725.
[3] 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., of the Psychiatric Genomics Consortium, S. W. G., et al., 2015. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics 47 (3), 291–295.
[4] Burish, J., Jess, T., Martinato, M., Lakatos, P. L., ECO- Epicom, 2013. The burden of inflammatory bowel disease in europe. Journal of Crohn’s and Colitis 7 (4), 322–337.
[5] Consortium, G. P., et al., 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491 (7422), 56–65.
[6] Consortium, G. P., et al., 2015. A global reference for human genetic variation. Nature 526 (7571), 68–74.
[7] Consortium, I. H., et al., 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467 (7311), 52.
[8] de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., et al., 2017. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nature genetics 49 (2), 256.
[9] Devlin, B., Roeder, K., Dec 1999. Genomic control for association studies. Biometrics 55 (4), 997–1004.
[10] Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., Mason, B., Goddard, M., 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of dairy science 95 (7), 4114–4129.
[11] Eyre-Walker, A., 2010. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proceedings of the National Academy of Sciences 107 (suppl 1), 1752–1756.
[12] Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Fagh, R., et al., 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature genetics 47 (11), 1228.
[13] Frei, O., Holland, D., Smeland, O. B., Shadrin, A. A., Fan, C. C., Maeland, S., O’Connell, K. S., Wang, Y., Djurovic, S., Thompson, W. K., et al., 2019. Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. Nature communications 10 (1), 1–11.
[14] Gazal, S., Finucane, H. K., Furtlette, N. A., Loh, P.-R., Palamara, P. F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B. M., Gusev, A., et al., 2017. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature genetics 49 (10), 1421.
[15] George, E. I., McCulloch, R. E., 1993. Variable selection via gibbs sampling. Journal of the American Statistical Association 88 (424), 881–889.
[16] Gravel, S., Lun, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., Bustamante, C. D., Project, G., et al., 2011. Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences 108 (29), 11983–11988.
multiple disease SNPs. Bioinformatics 27 (16), 2304–2305.

[52] Sveinbjornsson, G., Albrechtsen, A., Zink, F., Gudjonsson, S. A., Oddson, A., Mässon, G., Holm, H., Kong, A., Thorsteindottir, U., Sulem, P., et al., 2016. Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nature genetics.

[53] Van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pult, S. L., Van Der Spek, R. A., Vösa, U., De Jong, S., Robinson, M. R., et al., 2016. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nature genetics.

[54] Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., et al., 2013. Discovery and refinement of loci associated with lipid levels. Nature genetics 45 (11), 1274.

[55] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., Visscher, P. M., Jul 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42 (7), 565–569.

[56] Zhang, Y., Qi, G., Park, J.-H., Chatterjee, N., 2018. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nature genetics 50 (9), 1318.

[57] Zhou, X., Carbonetto, P., Stephens, M., 2013. Polygenic modeling with bayesian sparse linear mixed models. PLoS genetics 9 (2), e1003264.

[58] Zhu, X., Stephens, M., 2017. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. The annals of applied statistics 11 (3), 1561.