Linkage Disequilibrium and Heterozygosity Modulate the Genetic Architecture of Human Complex Phenotypes: Evidence of Natural Selection from GWAS Summary Statistics

Dominic Holland[a,b,*], Oleksandr Frei[d], 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]

[a] *Center for Multimodal Imaging and Genetics, University of California at San Diego, La Jolla, CA 92037, USA,*
[b] *Department of Neurosciences, University of California, San Diego, La Jolla, CA 92093, USA,*
[c] *Department of Radiology, University of California, San Francisco, San Francisco, CA 94158, USA,*
[d] *NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo 0424 Oslo, Norway,*
[e] *Department of Cognitive Sciences, University of California at San Diego, La Jolla, CA 92093, USA,*
[f] *Department of Radiology, University of California, San Diego, La Jolla, CA 92093, USA,*
[g] *Division of Mental Health and Addiction, Oslo University Hospital, 0407 Oslo, Norway,*
[h] *Department 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

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 [7, 3, 24, 25]. This research is facilitated by using new analytic approaches to interrogate structural features in the genome and their relationship to phenotypic expression. These analyses take into account the fact that different classes of SNPs have different characteristics and play a multitude of roles [16]. 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 (total linkage disequilibrium, TLD) are providing new insights into the action of selection on the genetic architecture of complex traits [3, 22, 25].

Here we present a unifying approach, using Bayesian analysis on GWAS summary statistics, to explore the role of TLD on the distribution of causal SNPs, and the impact of MAF on causal effect size, while simultaneously incorporating multiple effect-size distributions. Treating each of these factors independently is likely to to provide a misleading assessment. Thus, we posit a model for the prior distribution of effect sizes, and building on previous work [7], fit it to the GWAS summary statistics for a wide range of phenotypes. From the estimated set of model parameters for a phenotypes, we generated simulated genotypes and calculated the corresponding z-scores. With a wide range of model parameters across real phenotypes, the specificity of the model parameters for a given phenotype are narrowly defining of the distribution of summary statistics for that phenotype. We find that the detailed distribution of GWAS summary statistics for real phenotypes are reproduced to high accuracy, while many of the parameters used to setup 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.

In earlier work [6] 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 linkage disequilibrium between SNPs, many

*Corresponding author:
email: dominic.holland@gmail.com
Phone: 858-822-1776
Fax: 858-534-1078

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 βs is a mixture of non-null and null normal distributions, the latter (denoted $\mathcal{N}(0,0)$) being just a delta function:

$$\beta \sim \pi_1 \mathcal{N}(0, \sigma_\beta^2) + (1-\pi_1)\mathcal{N}(0,0) \tag{1}$$

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_β^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 total LD dependence (Eqs. 20 and 28 in [6]). Implementing this relatively simple model (Eq. 1) and applying it to GWAS summary statistics involved a good deal of complexity, but resulted in remarkably good 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.

Given that GWAS, due 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 total LD indicated some unevenness in how well the tails of the sub-distributions were predicted. Additionally, recent work by others [24] indicated that it is important to take SNP heterozygosity into account as a component in discoverability, and that an additional Gaussian distribution for the βs might be appropriate if large and small effects are distributed differently [25]. 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 total LD might play an important role in the distribution of causal effects [3]. It is unclear how all these factors impact each other.

METHODS

In the current work, we sought to extend our earlier work to incorporate multiple Gaussians, while taking into account total LD and selection effects reflected in heterozygosity, in modeling the distribution of causal βs. Note that this is independent of the need to correctly take heterozygosity and total LD into account when building a distribution for z-scores even when the distribution of causal effects (βs) does *not* depend on these factors, as in our earlier work [6]. 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 βs, we introduce a slight change in labeling the Gaussian variance ($\sigma_\beta^2 \to \sigma_b^2$), 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

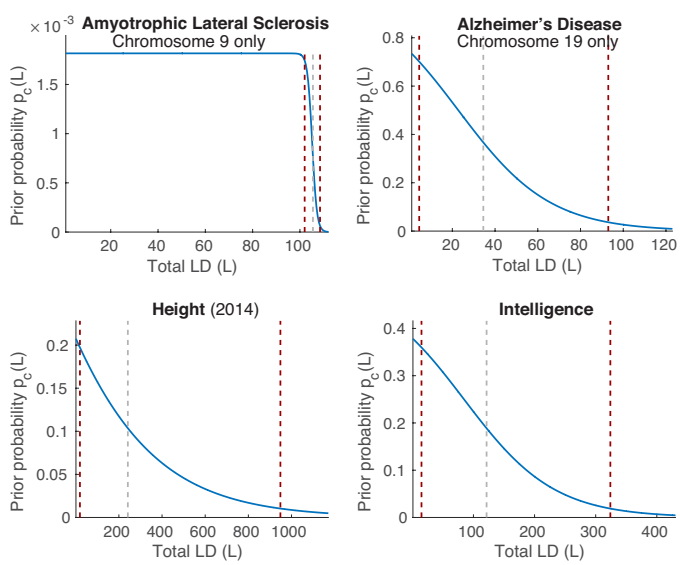

Figure 1: Examples of prior probability functions $p_c(L)$ used in Eqs. 3 and 4, where L is reference SNP total LD. These functions can be summarized by three quantities: the maximum value, p_{c1}, which occurs at $L = 1$; the total LD value, $L = m_c$, where $p_c(m_c) = p_{c1}/2$, give by the gray dashed lines in the figure; and the total LD width of the transition region, w_c, defined as the distance between where $p_c(L)$ falls to 95% and 5% of p_{c1} given by the flanking red dashed lines in the figure. Numerical values of p_{c1}, m_c, and w_c are given in Table 1 and Figures 2 and 3. $p_d(L)$ is similar. Plots of $p_c(L)$ and $p_d(L)$, where relevant, for all phenotypes are shown in Supplementary Material Figures 8-10.

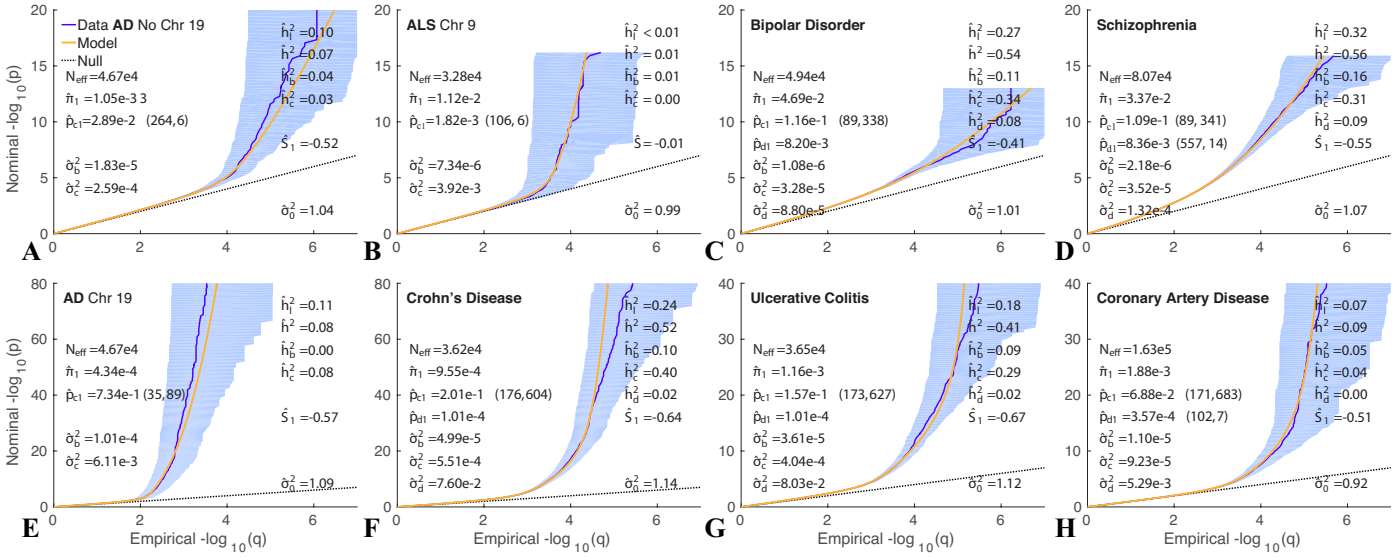

Figure 2: QQ plots of (pruned) z-scores for qualitative phenotypes (dark blue, 95% confidence interval in light blue) with model prediction (yellow). See Supplementary Material Figures 16 to 23. The value given for p_{c1} is the amplitude of the full $p_c(L)$ function, which occurs at $L = 1$; the values (m_c, w_c) in parentheses following it are the total LD (m_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_c) of the transition region (distance between flanking red dashed lines in Figure 1). Similarly for p_{d1} (m_d, w_d), where given. h^2_b, h^2_c, and h^2_d are the heritabilities associated with the "b", "c", and "d" Gaussians, respectively. h^2 is the total SNP heritability, reexpressed as h^2_l 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 gives the proportion, q, of typed SNPs exceeding that threshold (higher proportions are closer to the origin).

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 [24]. With the addition of the new term, the total prior probability for the SNP to be causal is still given by π_1. Thus, extending Eq. 1, we get:

$$\beta(H) \sim \pi_1 \left\{(1-p_c)\mathcal{N}(0,\sigma_b^2) + p_c\mathcal{N}(0, H^S\sigma_c^2)\right\} \qquad (2)$$

(ignoring, as noted, the null component $(1-\pi_1)\mathcal{N}(0,0)$), where p_c $(0 \le p_c \le 1)$ 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 σ_b^2). This extension introduces an extra three parameters p_c, σ_c, and S, assumed for the moment to be the same for all SNPs.

If the "c" Gaussian is capturing larger effects from rarer SNPs, reflecting selection pressure, it is reasonable to inquire if the prior probability for a causal SNP's contribution from the "c" Gaussian is total LD-mediated – with lower probability for a SNP that has large total LD (TLD). Thus, instead of treating p_c as a constant, we explore the possibility that it is larger for SNPs with weaker TLD. 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. Such a function of TLD can be characterized by three parameters: its amplitude, the TLD at the midpoint of the sigmoidal transition, and the width of the sigmoidal transition (over a wide or narrow range of TLD). Examples are shown in Figure 1 (mathematically, the curve can continue into the "negative TLD" range, revealing a familiar sigmoidal shape). Let the variable L be the TLD of a SNP. Then

$$\beta(H,L) \sim \pi_1\{(1-p_c(L))\mathcal{N}(0,\sigma_b^2) + p_c(L)\mathcal{N}(0, H^S\sigma_c^2)\} \qquad (3)$$

where $p_c(L)$ is the sigmoidal function ($0 \le p_c(L) \le 1$ for all L), 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 by the "b" and "c" Gaussians. This gives finally:

$$\beta(H,L) \sim \pi_1\{(1-p_c(L)-p_d(L))\mathcal{N}(0,\sigma_b^2) + p_c(L)\mathcal{N}(0, H^S\sigma_c^2) + p_d(L)\mathcal{N}(0,\sigma_d^2)\}. \qquad (4)$$

where σ_d^2 is a new parameter, $p_d(L)$ is another general sigmoid function ($0 \le p_d(L) \le 1$ for all L) where now there is the added constraint $0 \le p_c(L) + p_d(L) \le 1$, 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 β will be well characterized by Eq. 3.

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 ($N_{cases} = 20{,}352$, $N_{controls} = 31{,}358$) [18]; (2) schizophrenia ($N_{cases} = 35{,}476$, $N_{controls} = 46{,}839$) [15]; (3) coronary artery disease ($N_{cases} = 60{,}801$, $N_{controls} = 123{,}504$) [12]; (4) ulcerative colitis ($N_{cases} = 12{,}366$, $N_{controls} = 34{,}915$) and (5) Crohn's disease ($N_{cases} = 12{,}194$, $N_{controls} = 34{,}915$) [1]; (6) late onset Alzheimer's disease (LOAD; $N_{cases} = 17{,}008$, $N_{controls} = 37{,}154$) [9] (in the Supplementary Material we present results for a more recent GWAS with $N_{cases} = 71{,}880$ and $N_{controls} = 383{,}378$ [8]); (7) amyotrophic lateral sclerosis (ALS) ($N_{cases} = 12{,}577$, $N_{controls} = 23{,}475$) [19]; (8) number of years of formal education ($N = 293{,}723$) [13]; (9) intelligence ($N = 262{,}529$) [17, 14]; (10) body mass index ($N = 233{,}554$) [10]; (11) height (2010) ($N = 133{,}735$) [23]; and height (2014) ($N = 251{,}747$) [21]; (12) low- ($N = 89{,}873$) and (13) high-density lipoprotein ($N = 94{,}295$) [20]; and (14) total cholesterol ($N = 94{,}579$) [20]. Most participants were of European ancestry.

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 Supporting Material Figure 10; for BIP, 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 Supporting Material Tables 5-9.

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 4×4 grid of heterozygosity×total LD (each grid a a subset of a 10×10 grid) are in Supplementary Material Figures 16-34. For each phenotype, model selection (B, C, or D) was performed by testing the Bayesian information criterion (BIC) – see SI Table 1.

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. 4, the prior probabilities that a causal SNP belongs to the "c" and "d" Gaussians. These prior probabilities are shown in Figure 1 and SI Figures 8-10. 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 $H \times L$ grid, as shown for height (2014) in Figure 4 C, and for all phenotypes in SI Figures 12-15 (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 $H \times L$ grid element, as shown in Figure 4 D and SI Figures 12-15 (fourth columns). In general, depending on the shapes and sizes of $p_c(L)$ and $p_d(L)$, SNPs with lower L have larger $E(\beta^2)$. However, the selection parameter S in the "c" Gaussian has a large impact on $E(\beta^2)$ as a function of H (see SI Figure 5). As a result, for most phenotypes, we find that the effect sizes for low MAF causal SNPs ($H < 0.05$) are several times larger than for more common causal SNPs ($H > 0.1$). We find that heritability per causal SNP is larger for lower L, a general pattern that follows from at least one of the prior probabilities, $p_c(L)$ or $p_d(L)$, being non-constant. However, because heritability per causal SNP is proportional to H, we find that, even with negative selection parameter, S (and thus larger $E(\beta^2)$ for lower H), the heritability per causal SNP is largest for the most common causal SNPs ($H > 0.45$).

Simulations

To test the specificity of the model for each real phenotype, we constructed simulations where, in each case, the true causal βs (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 [6], 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 SI Figures 6 and 7 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 phe-

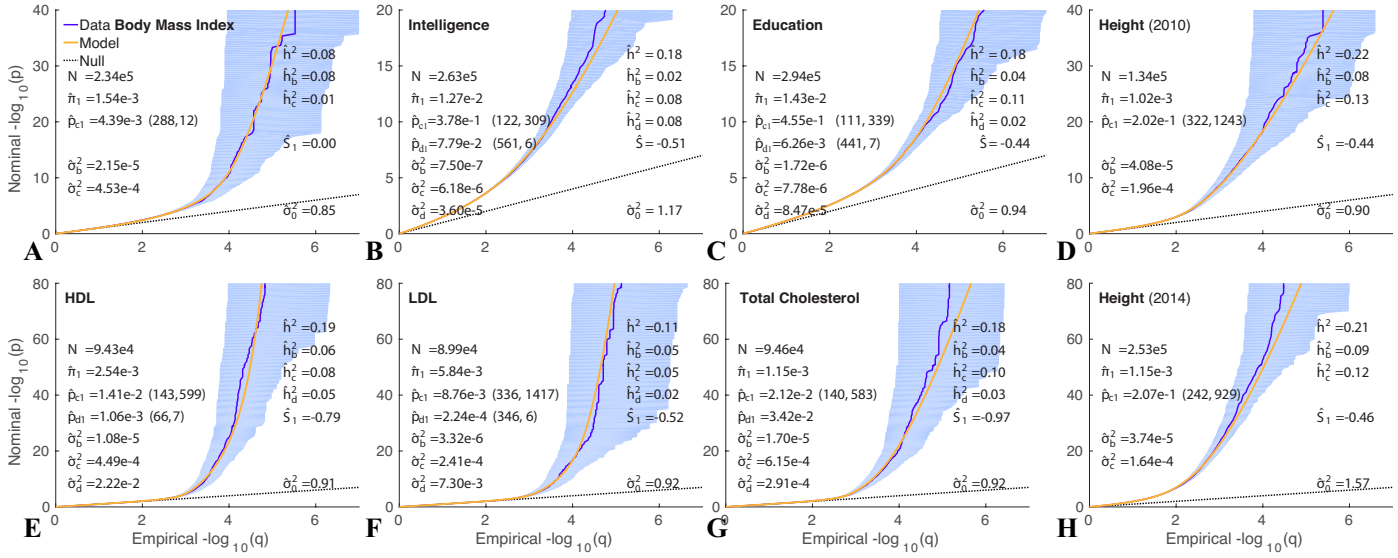

Figure 3: QQ plots of (pruned) z-scores for quantitative phenotypes (dark blue, 95% confidence interval in light blue) with model prediction (yellow). See Supplementary Material Figures 24 to 34. For HDL, $p_c(L) = p_{c1}$ for all L; for bipolar disorder and LDL, $p_d(L) = p_{d1}$ for all L. See caption to Figure 2 for further description.

Phenotype	π_1	σ_b^2	σ_c^2	S	p_{c1}	m_c	w_c	σ_d^2	p_{d1}	m_d	w_d	σ_0^2
SCZ 2014	3.37e-2	2.2e-6	3.5e-5	-0.55	0.11	89	341	1.3e-4	8.4e-3	557	14	1.07
BIP	4.69e-2	1.1e-6	3.3e-5	-0.41	0.12	89	338	8.8e-5	8.2e-3	—	—	1.01
CD	9.55e-4	5.0e-5	5.5e-4	-0.64	0.20	176	604	7.6e-2,	1.0e-4	—	—	1.14
UC	1.16e-3	3.6e-5	4.0e-4	-0.67	0.16	173	627	8.0e-2	1.0e-4	—	—	1.12
CAD	1.88e-3	1.1e-5	9.2e-5	-0.51	6.9e-2	171	683	5.3e-3	3.6e-4	102	7	0.92
AD Chr19	4.34e-4	1.0e-4	6.1e-3	-0.57	0.73	35	89	—	—	—	—	1.09
AD NoC19	1.05e-3	1.8e-5	2.6e-4	-0.52	2.9e-2	264	6	—	—	—	—	1.04
ALS Chr9	1.12e-2	7.3e-6	3.9e-3	-0.01	1.8e-3	106	6	—	—	—	—	0.99
Edu	1.43e-2	1.7e-6	7.8e-6	-0.44	0.45	111	339	8.5e-5	6.3e-3	441	7	0.94
IQ 2018	1.27e-2	7.5e-7	6.2e-6	-0.51	0.38	122	309	3.6e-5	7.8e-2	561	6	1.17
Putamen	5.33e-5	9.5e-4	—	—	—	—	—	—	—	—	—	1.00
Height 2010	1.02e-3	4.1e-5	2.0e-4	-0.44	0.20	322	1243	—	—	—	—	0.90
Height 2014	1.15e-3	3.7e-5	1.6e-4	-0.46	0.21	242	929	—	—	—	—	1.57
Height 2018	2.50e-3	8.7e-6	8.9e-5	-0.43	0.37	210	739	—	—	—	—	2.12
HDL	2.54e-3	1.1e-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	3.3e-6	2.4e-4	-0.52	8.8e-3	336	1417	7.3e-3	2.2e-4	346	6	0.92
BMI	1.54e-3	2.2e-5	4.5e-4	0.00	4.4e-3	288	12	—	—	—	—	0.85
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). π_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 $H^S \sigma_c^2$, 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 $p_{c1} \equiv p_c(1)$ at $L = 1$; the total LD value $L = m_c$ at the mid point of the transition, i.e., $p_c(m_c) = p_{c1}/2$ (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_{c1} (distance between the flanking red dashed lines in Figure 1). Note that for AD Chr19, AD NoC19, and ALS Chr9, π_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^S multiplying σ_c^2 are shown in Supplementary Material Figure 5. Estimated Bayesian information criterion (BIC) values for three models (B, C, and D) are shown in Supplementary material Table 4: the 3-parameter model B with only the "b" Gaussian $(\pi_1, \sigma_b, \sigma_0)$; the 8-parameter model C with both the "b" with "c" Gaussians (Eq. 3); and the 12-parameter model D with "b", "c" and "d" Gaussians (Eq. 4). 95% confidence intervals are in Supporting Materials Tables 5-7

Phenotype	n_b	n_c	n_d	n_{causal}
SCZ 2014	3.5e5	2.3e4	3.0e3	3.71e5
BIP	4.8e5	3.4e4	4.2e3	5.16e5
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 Chr19	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
Putamen	587	—	—	587
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	1.7e4	68	—	1.69e4
TC	1.2e4	180	434	1.27e4

Table 2: Numbers of causal SNPs. n_{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 Supporting Materials Table 8

Phenotype	h_b^2	h_c^2	h_d^2	h^2	h_l^2
SCZ 2014	0.16	0.31	0.09	0.56	0.32
BIP	0.11	0.34	0.08	0.54	0.27
CD	0.10	0.40	0.02	0.52	0.24
UC	0.09	0.29	0.02	0.41	0.18
CAD	0.05	0.04	0.00	0.09	0.07
AD Chr19	0.00	0.08	—	0.08	0.11
AD NoC19	0.04	0.03	—	0.07	0.10
ALS Chr9	0.01	0.00	—	0.01	0.00
Edu	0.04	0.11	0.02	0.18	—
IQ 2018	0.02	0.08	0.08	0.18	—
Putamen	0.12	—	—	0.12	—
Height 2010	0.08	0.13	—	0.22	—
Height 2014	0.09	0.12	—	0.21	—
Height 2018	0.04	0.24	—	0.28	—
HDL	0.06	0.08	0.05	0.19	—
LDL	0.05	0.05	0.02	0.11	—
BMI	0.08	0.01	—	0.08	—
TC	0.04	0.10	0.03	0.18	—

Table 3: Heritabilities: h^2 is the total additive SNP heritability, re-expressed on the liability scale as h_l^2 for the qualitative traits (upper section). h_b^2, h_c^2, and h_d^2 are the heritabilities associated with the "b", "c", and "d" Gaussians, respectively. 95% confidence intervals are in Suporting Materials Table 9.

notypes), though for Crohn's disease the simulated value was overestimated. Note that for the case-control simulated phenotypes, the heritabilities on the observed scale, denoted $\hat{h}^2$ in SI Figure 6, should be compared with the corresponding values in Figure 2, not with $\hat{h}_l^2$, which denotes heritability on the liability scale, i.e., adjusted for population prevalence. 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 and very low polygenicity (small number of causal SNPs) for the "c" Gaussian.

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

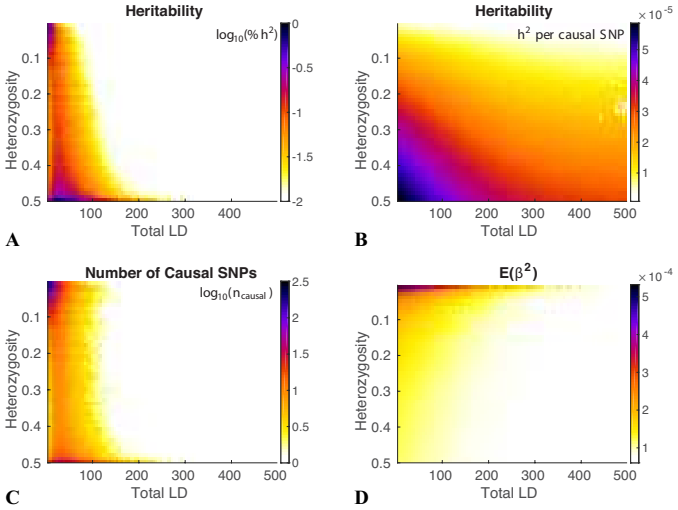

Figure 4: Model results for height (2014) using the BC model. The reference panel SNPs are binned with respect to both heterozygosity (H) and total LD (L) in a 50×50 grid for $0.02 \leq H \leq 0.5$ and $1 \leq L \leq 500$. Shown are model estimates of: (A) $\log_{10}$ of the percentage of heritability in each grid element; (B) for each element, the average heritability per causal SNP in the element; (C) $\log_{10}$ of the number of causal SNPs in each element; and (D) the expected β^2 for the element-wise causal SNPs. Note that H increases from top to bottom.

from random mutations and selection pressure. With many underlying causal variants, coming from a multitude of functional categories [16], it is likely that different variants will experience different evolutionary pressure – negative,

neutral, or positive – due at least to pleiotropic roles.

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 [3], or took MAF into account while ignoring total LD dependence and different distributions for large and small effects [24], 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 [6]. 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 $H\times L$ grid – even though the various phenotypic polygenicities and per-causal-SNP heritabilities range over orders of magnitude.

Selection pressure can be assessed by measuring the extent of long-range 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 [5]. 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 [11]. 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. 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 [3]. So when negative selection is operating, we should expect to find more and more effects with larger and larger effect size at lower total LD and lower heterozygosity.

It was found in [3] – 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 the earlier model result that most of the variance in fitness comes from rare variants that have a large effect on the trait in question [2]. 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 [24], using forward simulations and a commonly used demographic model [4], that negative values for the selection parameter, S, which leads to larger effects for rarer variants, is a signature of negative selection.

We find negative selection parameter values for most traits, which is broadly in agreement with [24] with the exception of BMI, which we find can be modeled with two Gaussians with no 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 only 68 common causal SNPs of large effect. A similar situation ($S \simeq 0$) 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.

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 SI Figures 12-15, second columns.

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 a amplifier of σ_c^2 in Eq. 4 is illustrated in Supplementary Material Figure 5) – see SI Figures 12-15, 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 "c" Gaussian (see SI Figures 8 and 9).

The "b" Gaussian in Eq. 3 or 4 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 here, whether or not there is evidence of negative selection ($S < 0$), 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 ten 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, p_{d1}, 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 four instances – BIP, CD, UC, and TC – $p_d(L)$ 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 ($m_d = 561$; see also SI Figure 10). 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 negative selection. Additionally, for all phenotypes, we found evidence of neutral selection operating on SNPs with relatively weak effect. 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.

Funding

Research Council of Norway (262656, 248984, 248778, 223273) and KG Jebsen Stiftelsen; ABCD-USA Consortium (5U24DA041123).

References

[1] 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.

[2] 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.

[3] 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., et al., 2017. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature genetics 49 (10), 1421.

[4] Gravel, S., Henn, 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.

[5] Hill, W., Robertson, A., 2007. The effect of linkage on limits to artificial selection. Genetics Research 89 (5-6), 311–336.

[6] Holland, D., Desikan, R., Frei, O., Fan, C., Shadrin, A., Smeland, O., Sundar, V., Thompson, P., Andreassen, O., Dale, A., 2018. Beyond snp heritability: Polygenicity and discoverability of phenotypes estimated with a univariate gaussian mixture model.

[7] Holland, D., Fan, C.-C., Frei, O., Shadrin, A. A., Smeland, O. B., Sundar, V. S., Andreassen, O. A., Dale, A. M., 2017. Estimating degree of polygenicity, causal effect size variance, and confounding bias in gwas summary statistics. bioRxiv. URL https://www.biorxiv.org/content/early/2017/05/24/133132

[8] Jansen, I., Savage, J., Watanabe, K., Bryois, J., Williams, D., Steinberg, S., Sealock, J., Karlsson, I., Hagg, S., Athanasiu, L., et al., 2018. Genetic meta-analysis identifies 10 novel loci and functional pathways for alzheimer's disease risk. bioRxiv, 258533.

[9] Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., et al., 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. Nature genetics 45 (12), 1452–1458.

[10] Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al., 2015. Genetic studies of body mass index yield new insights for obesity biology. Nature 518 (7538), 197.

[11] Maruyama, T., 1974. The age of a rare mutant gene in a large population. American journal of human genetics 26 (6), 669.

[12] Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., et al., 2015. A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. Nature genetics 47 (10), 1121.

[13] Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S. F. W., et al., 2016. Genome-wide association study identifies 74 loci associated with educational attainment. Nature 533 (7604), 539–542.

[14] Savage, J., Jansen, P., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C., Nagel, M., Awasthi, S., Barr, P., Coleman, J., Grasby, K., Hammerschlag, A., Kaminski, J., Karlsson, R., et al., 2018. Genome-wiide association meta-analysis (n=269,867) identifies new genetic and functional links to intelligence. Nature genetics forthcoming. URL https://ctg.cncr.nl/software/summary_statistics

[15] Schizophrenia Working Group of the Psychiatric Genomics Consortium, Jul 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511 (7510), 421–427.

[16] Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., Kelsoe, J. R., O'Donovan, M. C., Furberg, H., Schork, N. J., et al., 2013. All snps are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps. PLoS genetics 9 (4), e1003449.

[17] Sniekers, S., Stringer, S., Watanabe, K., Jansen, P. R., Coleman, J. R., Krapohl, E., Taskesen, E., Hammerschlag, A. R., Okbay, A., Zabaneh, D., et al., 2017. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. Nature genetics 49 (7), 1107.

[18] Stahl, E., Breen, G., Forstner, A., McQuillin, A., Ripke, S., Cichon, S., Scott, L., Ophoff, R., Andreassen, O. A., Kelsoe, J., Sklar, P., 2018. Genomewide association study identifies 30 loci associated with bipolar disorder. bioRxiv. URL https://www.biorxiv.org/content/early/2018/01/24/173062

[19] Van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., Van Der Spek, R. A., Võsa, 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 48 (9), 1043.

[20] 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.

[21] Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., et al., 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature genetics 46 (11), 1173–1186.

[22] Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M., et al., 2018. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature genetics 50 (5), 668.

[23] 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.

[24] Zeng, J., Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F., et al., 2018. Signatures of negative selection in the genetic architecture of human complex traits. Nature genetics 50 (5), 746.

[25] 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.