Bioinformatics, 36(18), 2020, 4749–4756
doi: 10.1093/bioinformatics/btaa568
Advance Access Publication Date: 15 June 2020
Original Paper

Genetics and population analysis

Phenotype-specific differences in polygenicity and effect size distribution across functional annotation categories revealed by AI-MiXeR

Alexey A. Shadrin [1,2,*], Oleksandr Frei[1,2,3], Olav B. Smeland[1,2], Francesco Bettella [1,2], Kevin S. O'Connell[1,2], Osman Gani[1,2], Shahram Bahrami[1,2], Tea K. E. Uggen[1,2], Srdjan Djurovic[4,5], Dominic Holland[6,7], Ole A. Andreassen[1,2,7] and Anders M. Dale[6,7,8,9,*]

[1]NORMENT, Institute of Clinical Medicine, University of Oslo, Oslo 0424, Norway, [2]Division of Mental Health and Addiction, Oslo University Hospital, Oslo 0424, Norway, [3]Center for Bioinformatics, Department of Informatics, University of Oslo, Oslo 0373, Norway, [4]Department of Medical Genetics, Oslo University Hospital, Oslo 0424, Norway, [5]NORMENT, Department of Clinical Science, University of Bergen, Bergen 5020, Norway, [6]Center for Multimodal Imaging and Genetics, University of California, San Diego, La Jolla, CA, 92037, USA, [7]Department of Neurosciences, University of California, San Diego, La Jolla, CA 92093, USA, [8]Department of Radiology, University of California, San Diego, La Jolla, CA 92093, USA and [9]Department of Psychiatry, University of California, San Diego, La Jolla, CA 92093, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on February 11, 2020; revised on June 3, 2020; editorial decision on June 6, 2020; accepted on June 9, 2020

Abstract

Motivation: Determining the relative contributions of functional genetic categories is fundamental to understanding the genetic etiology of complex human traits and diseases. Here, we present Annotation Informed-MiXeR, a likelihood-based method for estimating the number of variants influencing a phenotype and their effect sizes across different functional annotation categories of the genome using summary statistics from genome-wide association studies.

Results: Extensive simulations demonstrate that the model is valid for a broad range of genetic architectures. The model suggests that complex human phenotypes substantially differ in the number of causal variants, their localization in the genome and their effect sizes. Specifically, the exons of protein-coding genes harbor more than 90% of variants influencing type 2 diabetes and inflammatory bowel disease, making them good candidates for whole-exome studies. In contrast, <10% of the causal variants for schizophrenia, bipolar disorder and attention-deficit/hyperactivity disorder are located in protein-coding exons, indicating a more substantial role of regulatory mechanisms in the pathogenesis of these disorders.

Availability and implementation: The software is available at: https://github.com/precimed/mixer.

Contact: a.a.shadrin@medisin.uio.no or andersmdale@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The rapid technological advances of the last years have provided an enormous amount of genetic data, promoting the development of statistical methods aimed at unraveling the genetic architecture of complex traits (Evans *et al.*, 2018). A key effort has been to estimate single nucleotide polymorphism (SNP)-based heritability, either using individual-level genotype data (Yang *et al.*, 2010), or summary-level statistics from genome-wide association studies (GWAS) (Bulik-Sullivan *et al.*, 2015). However, heritability estimates provide a limited picture of the genetic architecture underlying complex phenotypes. For example, they are agnostic about the number of genetic variants influencing a phenotype and their effect sizes (Timpson *et al.*, 2018): both of these quantities can vary and still result in the same heritability, which is proportional to their product (Frei *et al.*, 2019; Holland *et al.*, 2020b). Importantly, the proportion of variants influencing a phenotype (polygenicity) and the

variance their effect sizes (discoverability) substantially affects the power of GWAS and may inform the design of future genetic studies to maximize discovery (Schork *et al.*, 2016; Smeland *et al.*, 2019).

Recently, we developed a model which allows the breakdown of SNP-heritability into the number of variants influencing a given phenotype (non-null variants) and the distribution of their effect sizes using summary statistics from GWAS and detailed population-specific linkage disequilibrium (LD) structure (Frei *et al.*, 2019; Holland *et al.*, 2020b). The model assumes that the non-null variants are distributed uniformly throughout the genome and that their effect sizes are drawn from a Gaussian distribution. However, prior genetic studies suggest that non-null variants are differentially enriched across functional genomic categories and complex phenotypes (Schaub *et al.*, 2012; Schork *et al.*, 2013). Here, we present a model, Annotation Informed (AI)-MiXeR, which extends our previous work by allowing different (non-overlapping) predefined functional annotation categories of the genome to have various densities of non-null variants with different effect size distributions.

Several conceptually related methods that aim to characterize the genetic architecture of phenotypes using GWAS summary statistics have recently been developed. The partitioned LD score regression (LDSC) analysis (Finucane *et al.*, 2015) estimates the proportion of SNP-based heritability explained by variants within predefined functional categories but does not estimate the abundance of non-null variants or assess their effect sizes. The RSS-E method (Zhu and Stephens, 2018) only estimates the abundance of non-null variants in different annotation categories, while the distribution of effect sizes of non-null variants is assumed to be the same for all annotation categories. The GENESIS model (Zhang *et al.*, 2018) allows several groups of trait-susceptibility variants with different densities and effect size distributions but assumes the non-null variants to be uniformly distributed among the groups and does not support prior group definition (e.g. in terms of functional annotation categories). In contrast to these methods, AI-MiXeR allows simultaneous modeling of abundance and effect size magnitudes of non-null variants in arbitrary predefined functional annotation categories.

Here, we extensively tested AI-MiXeR on synthetic GWAS data generated under various setups to establish scenarios where it reconstructs the underlying parameters correctly. We then applied AI-MiXeR to GWAS summary statistics for 11 complex phenotypes representing a range of diverse human traits and diseases. Our analysis suggests that both densities and effect sizes of non-null variants vary considerably across different genomic annotation categories and reveals diverse patterns of genetic architecture in different phenotypes.

2 Materials and methods

2.1 Ai-MiXeR model overview

We consider an additive model of genetic effects ignoring gene–environment interactions, epistasis and dominance effects. Variant effect sizes are modeled with point-normal mixture priors, where both proportion of non-null variants and distribution of their effect sizes can vary between different predefined functional genomic categories. Each functional category in the model is characterized by the proportion of non-null variants (polygenicity, π) and the variance of their effect sizes (discoverability, σ^2). The pure (i.e. not induced by LD) effect of the kth variant (β_k) is modeled as a mixture of null and non-null components: $\beta_k = \begin{cases} 0, & 1-\pi_C \\ N(0,\sigma_C^2), & \pi_C \end{cases}$, where π_C and σ_C^2, respectively are proportion and variance of non-null variants effect sizes in the functional category C, and $N(0,\sigma_C^2)$ denotes the normal distribution with mean 0 and variance σ_C^2. The signed association test statistics (z-score) of the jth variant is then given by: $z_j = \sum_{k=1}^{M} \sqrt{NH_k} r_{jk} \beta_k + \epsilon$, where N is the GWAS sample size, H_k is the heterozygosity of variant k, M is the number of variants in LD with variant k, r_{jk} is the Pearson's correlation coefficient between the genotypes of the jth and kth variants and ϵ is a $N(0,\sigma_0^2)$ distributed residual factor. Functional category-specific polygenicities and discoverabilities of a GWAS trait are estimated by maximizing the likelihood of the GWAS summary statistics (z-scores). To reduce computational burden, we randomly select a subset of 10^6 variants of all GWAS variants to use for maximization of the likelihood function. For specific details of the model and its implementation, please refer to the following sections.

2.2 Ai-MiXeR model details

Consider a quantitative phenotype standardized to mean 0 and variance 1. Let y be a random variable representing a phenotype measurement for an individual in the population ($\mathrm{E}(y) = 0$, $\mathrm{var}(y) = 1$). Let $G = \{g_j\}_{j=1,\dots M}$ be a fixed set of M random variables representing genotypes of bi-allelic variants. These are assumed to be centered ($\mathrm{E}(g_j) = 0$) but not scaled ($\mathrm{var}(g_j) = 2f_j(1-f_j) = H_j$, where f_j is the minor allele frequency of variant j and h_j is its heterozygosity). We assume an additive genetic model for the phenotype generation:

$$y = \sum_{j=1}^{M} g_j \beta_j(G) + \epsilon, \tag{1}$$

where ϵ is a normally distributed error term with mean 0 and variance σ_e^2. $\beta_j(G)$ is understood here as the (unknown) effect of variant j as would be obtained from a multiple linear regression of the phenotype y on all genotypes G in a hypothetical infinite sample. This definition of the effect size $\beta_j(G)$ implies that $\beta_j(G)$ will reflect only the true causal effect of the jth variant (thus $\beta_j(G) = 0$ if the jth variant is not causal) whenever G includes all causal variants for the trait. On the other hand, if any causal variants are missing in the set G, $\beta_j(G)$ will also include the effects of those missing causal variants that happen to be tagged by the jth variant. Any variant j with $\beta_j(G) \neq 0$ will be called a non-null variant. Further, we will henceforth consider G to be fixed and omit it from the notation.

Consider now a GWAS on a quantitative phenotype. Let N be the number of individuals in the GWAS and assume that N is sufficiently large so that the allelic composition (i.e. genotype frequencies) of the variants observed in the GWAS is approximately equivalent to the allelic composition of the same variants in the population. Then $\hat{y} = (\hat{y}_1 \dots \hat{y}_N)$ is a vector of phenotypes, $\hat{G} = [\hat{g}_{ij}]_{i=1\dots N, j=1\dots M}$ is the $N \times M$ matrix of genotypes observed in the GWAS and $\hat{\epsilon} = (\hat{\epsilon}_1 \dots \hat{\epsilon}_N)$ is a vector of residuals ($\hat{\epsilon}_i$ represents the residual term for the ith individual). Using (1) and this notation we can write:

$$\hat{y} = \hat{G}\beta + \hat{\epsilon}, \tag{2}$$

which is a sample-equivalent of Equation (1). Denote also with $\hat{g}_j = (\hat{g}_{1j} \dots \hat{g}_{Nj})$ the vector of genotypes of variant j observed in the GWAS (jth column of $\hat{G}$ matrix). The marginal effect of variant j estimated in GWAS ($\hat{\beta}_j$) is obtained from the simple linear regression of the phenotype on the genotype of variant j: $\hat{y}_i = \alpha_j' + \beta_j' \hat{g}_{ij} + \hat{\varepsilon}_{ij}$, $i = 1, \dots, N$, where α_j' and β_j' are the (unknown) intercept and slope of the simple linear regression, respectively, and $\hat{\varepsilon}_{ij}$, $i = 1 \dots N$, are its residuals. The value of the slope $\hat{\beta}_j'$ minimizing the sum of squared residuals is:

$$\begin{aligned} \hat{\beta}_j' &:= \frac{\mathrm{cov}(\hat{y}, \hat{g}_j)}{\mathrm{var}(\hat{g}_j)} = \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \hat{g}_{ij}}{\frac{1}{N}\sum_{i=1}^{N} \hat{g}_{ij}^2} = [\text{substitute } \hat{y}_i \text{ using (2)}] \\ &= \frac{\frac{1}{N}\sum_i \left[\left(\sum_{k=1}^{M} \hat{g}_{ik}\beta_k + \hat{\epsilon}_i \right) \hat{g}_{ij} \right]}{\hat{H}_j} = \frac{\frac{1}{N}\sum_k \left(\beta_k \sum_i \hat{g}_{ij}\hat{g}_{ik} \right)}{\hat{H}_j} + \frac{\sum_i \hat{\epsilon}_i \hat{g}_{ij}}{N\hat{H}_j} \\ &= \frac{\sum_k \sqrt{\hat{H}_k} \hat{r}_{jk} \beta_k}{\sqrt{\hat{H}_j}} + \frac{\sum_i \hat{\epsilon}_i \hat{g}_{ij}}{N\hat{H}_j}, \end{aligned} \tag{3}$$

where $\hat{r}_{jk} := \frac{\sum_i \hat{g}_{ij}\hat{g}_{ik}}{\sqrt{\sum_i \hat{g}_{ij}^2 \sum_i \hat{g}_{ik}^2}} = \frac{\sum_i \hat{g}_{ij}\hat{g}_{ik}}{N\sqrt{\hat{H}_j \hat{H}_k}}$ is the sample correlation coefficient between genotype vectors $\hat{g}_j$ and $\hat{g}_k$, $\hat{H}_j = \mathrm{var}(\hat{g}_j) :=$

$\frac{1}{N}\sum_{i=1}^{N}\hat{g}_{ij}^2\left[\cong H_j := \mathrm{var}(g_j)\right]$ is the sample heterozygosity of variant j and β_k is the hypothetical effect size of variant k from a multiple linear regression in an infinite population (as discussed above).

Assuming the considered phenotype is complex, i.e. it is influenced by many variants each explaining only a tiny fraction of phenotypic variance, then the variance of the simple linear regression error is approximately equal to the sample variance of the phenotype, $\mathrm{var}(\hat{\varepsilon}_j) \cong \mathrm{var}(\hat{y}) := 1$, where $\hat{\varepsilon}_j = (\hat{\varepsilon}_{1j} \ldots \hat{\varepsilon}_{Nj})$. Using this approximation, we can write an expression for the standard error of $\hat{\beta}_j$:

$$\mathrm{SE}\left(\hat{\beta}_j^{'}\right) := \sqrt{\frac{\frac{1}{N-2}\sum_{i=1}^{N}\hat{\varepsilon}_{ij}^2}{\sum_i \hat{g}_{ij}^2}} \cong \sqrt{\frac{\mathrm{var}(\hat{y})}{N\hat{H}_j}} = \frac{1}{\sqrt{N\hat{H}_j}}. \tag{4}$$

Combining Equations (3) and (4), we can write an expression for the z-score of variant j observed in GWAS:

$$\begin{aligned} z_j|\hat{r}_{jk}, \hat{H}_k, \hat{g}_{ij}, N &:= \frac{\hat{\beta}_j^{'}}{SE\left(\hat{\beta}_j^{'}\right)} = \sum_{k=1}^{M}\sqrt{N\hat{H}_k}\hat{r}_{jk}\beta_k + \frac{1}{\sqrt{N\hat{H}_j}}\sum_{i=1}^{N}\hat{\epsilon}_i\hat{g}_{ij} \\ &= \sum_{k=1}^{M}\sqrt{N\hat{H}_k}\hat{r}_{jk}\beta_k + \hat{\epsilon}^{'}. \end{aligned} \tag{5}$$

In Equation (5), $\hat{r}_{jk}$, $\hat{H}_k$, $\hat{g}_{ij}$ and N are known constant factors and $\hat{\epsilon}^{'} = \frac{1}{\sqrt{N\hat{H}_j}}\sum_{i=1}^{N}\hat{\epsilon}_i\hat{g}_{ij}$ is an unknown (because the $\hat{\epsilon}_i$ are) residual. Remembering that, by definition, $\hat{\epsilon}_i$, $i = 1, \ldots, N$ are realizations of the ϵ random variable [see Equations (1) and (2)] and assuming that these realizations are independent from each other, $\hat{\epsilon}^{'}$ can be modeled as a normally distributed random variable with mean 0 and variance (equally for all variants):

$$\mathrm{var}(\hat{\epsilon}^{'}) \cong \mathrm{var}\left(\frac{1}{\sqrt{N\hat{H}_j}}\sum_{i=1}^{N}\epsilon\hat{g}_{ij}\right) = \frac{\mathrm{var}(\epsilon)}{N\hat{H}_j}\sum_{i=1}^{N}\hat{g}_{ij}^2 = \mathrm{var}(\epsilon) = \sigma_e^2.$$

By construction [Equation (1)], when there is no genetic effect on the phenotype, $\sigma_e^2 = 1$. However, the assumption of independence of all $\hat{\epsilon}_i$ is often violated in GWAS due to the presence of various confounding factors such as population stratification and cryptic relatedness. Moreover, both $\hat{H}_k$ and $\hat{r}_{jk}$ are usually estimated from external genotyping panels where variant frequencies and correlations may differ from those in the GWAS sample. In addition, due to technical limitations, $\hat{r}_{jk}$ estimates are commonly truncated (e.g. disregarding all correlations below a certain threshold). For the model to be able to mitigate these discrepancies we introduce a σ_0^2 parameter and model $\hat{\epsilon}^{'}$ in (5) as a random variable distributed as $N(0, \sigma_0^2)$. It was shown that, in the framework of the infinitesimal model, σ_0^2 has the same mathematical meaning as the intercept term in the LDSC model (Frei *et al.*, 2019). The last unknown factor in (5), β_k, is modeled as a random variable with point-normal mixture distribution, where the variance is allowed to differ between different variant annotation categories:

$$\beta_k = \begin{cases} 0, & 1-\pi_C \\ N(0, \sigma_C^2), & \pi_C \end{cases}, \tag{6}$$

where variant $k \in C$, $C \subseteq G$ is a subset of variants in G constituting some annotation category, π_C is the proportion of variants with non-zero effect (non-null variants) in the annotation category C and σ_C^2 is the variance of the effect sizes among all non-null variants in C. The set of annotation categories $\{C_j\}_{j=1\ldots T}$ defined on G must form a partition of G (i.e. each variant from the G must belong to one and only one annotation category C_j).

Modeling β_k as Equation (6), $\hat{\epsilon}^{'}$ as $N(0, \sigma_0^2)$ and taking r_{jk}, h_k and N as known constant factors, Equation (5) allows to derive the probability density function (pdf) of z_j (pdf_z) as the convolution of β_k ($k = 1 \ldots M$) and $\hat{\epsilon}^{'}$ random variables.

2.3 Estimation of pdf of *z*-scores

We derive the pdf of a random variable z representing a variant's association z-score from the convolution of β_k ($k = 1 \ldots M$) and $\hat{\epsilon}^{'}$ random variables. To simplify notation, we omit the indices reflecting the annotation category, replace $\hat{\epsilon}^{'}$ with ϵ and denote:

$$\xi_k = \sqrt{Nh_k}r_{jk}\beta_k = \begin{cases} 0, & 1-\pi \\ N\left(0, \sigma_{e,k}^2\right), & \pi \end{cases}, \tag{7}$$

where $\sigma_{e,k}^2 = N_i r_{ik}^2 H_k \sigma^2$, $\epsilon \sim N(0, \sigma_0^2)$. We can then rewrite Equation (5) as:

$$z = \epsilon + \sum_{k=1}^{M}\xi_k.$$

The pdf of z at z_0 (in our case z_0 is the z-score from the GWAS) can be written as the inverse Fourier transform of its characteristic function $\phi_z(t)$:

$$\mathrm{pdf}_z(z_0) = \frac{1}{2\mathrm{pi}}\int_{-\infty}^{\infty} e^{-itz_0}\phi_z(t)\mathrm{d}t,$$

where pi is Archimedes' constant (i.e. pi $\approx$ 3.14) and i is the unit imaginary number.

Assuming that the non-null effects (β_j) are independent from each other and from the error term ϵ:

$$\phi_z(t) = \phi_\epsilon(t)\prod_k \phi_{\xi_k}(t).$$

Using the definition of characteristic function and expression (7), we can write the characteristic function of ξ_k as:

$$\phi_{\xi_k}(t) = \int_{-\infty}^{\infty} e^{itx} f_{\xi_k}(x)dx = (1-\pi) + \pi e^{-\frac{t^2\sigma_{e,k}^2}{2}},$$

and similarly for ϵ:

$$\phi_\epsilon(t) = e^{-\frac{t^2\sigma_0^2}{2}}.$$

Combining the last two expressions, the characteristic function of z can be written as:

$$\phi_z(t) = e^{-\frac{t^2\sigma_0^2}{2}}\prod_k\left[(1-\pi) + \pi e^{-\frac{t^2\sigma_{e,k}^2}{2}}\right],$$

from which we can obtain the point estimate of pdf_z at z_0:

$$\begin{aligned} pdf_z(z_0) &= \frac{1}{2\mathrm{pi}}\int_{-\infty}^{\infty} e^{-itz_0}\phi_z(t)\mathrm{d}t \\ &= \frac{1}{2\mathrm{pi}}\int_{-\infty}^{\infty}\cos(tz_0)\phi_z(t)\mathrm{d}t - \frac{i}{2\mathrm{pi}}\int_{-\infty}^{\infty}\sin(tz_0)\phi_z(t)\mathrm{d}t \\ &= \frac{1}{\mathrm{pi}}\int_{0}^{\infty}\cos(tz_0)\phi_z(t)\mathrm{d}t. \end{aligned}$$

The result is a definite integral (i.e. a number), which can be computed numerically.

2.4 Optimization setup

The polygenicity (π) and discoverability (σ^2) parameters are estimated by maximizing the likelihood of the z-scores observed in the GWAS summary-level data (z_0): $pdf_z(z_0)\underset{\pi,\sigma^2,\sigma_0^2}{\longrightarrow}\max$, where the

probability density function of the z-scores (pdf_z) is modeled as described in the section above. Specific estimation details are given below.

The following optimization setup was used:

- Nelder–Mead method (maxiter = 1200, fatol = 1e-7, xatol = 1e-4, adaptive = True) was applied starting from the best point obtained after a single iteration of differential evolution [maxiter = 1, popsize = 50, init = latinhypercube, bounds: π = (5E-5,5E-1), σ^2 = (5E-6,5E-2), σ_0^2 = (0.9,2.5)], as implemented in SciPy (Virtanen *et al.* 2020).
- Variants from the extended major histocompatibility complex region (genome build 19 locations chr6:25119106–33854733) were excluded from the optimization due to the high complexity of the LD structure in this region.
- The z-scores of 10^6 randomly selected variants were used for the optimization of the cost function. This procedure was replicated 50 times to limit selection bias.
- The cost function was defined as -log(likelihood)/10^6, where 10^6 reflects the number of variants (z-scores) used at each replica of the optimization procedure.

2.5 Implementation

Data management and configuration procedures were implemented in Python. For optimization SciPy implementations of both Nelder–Mead and differential evolution methods were used. Evaluation of the cost function was implemented in C using GNU Scientific Library (http://www.gnu.org/software/gsl/) for numeric integration and OpenMP (https://www.openmp.org/) for parallelization. AI-MiXeR's source code is freely available at https://github.com/precimed/mixer.

2.5.1 Computational time

A single optimization run using the setup described in the 'Optimization setup' section above took between 3 and 6 h on a computing node with dual Intel Sandy Bridge E5-2670 (16 physical computing cores) running at 2.6 GHz, and 64 Gb RAM.

2.6 Simulations with synthetic data

We analyzed the performance of the model on GWAS summary statistics generated from synthetic genotypes and phenotypes with various genetic architectures under model assumptions.

2.6.1 Synthetic genotypes

10^5 synthetic genotypes were generated with Hapgen2 (Su *et al.*, 2011) using 503 European samples from 1000 Genomes Phase 3 data (1000 Genomes Project Consortium *et al.*, 2015) as described in the study by Frei *et al.* (2019). A set of 11 015 833 biallelic variants was considered. The LD structure was estimated from a subset of 10^4 genotypes using PLINK 1.9 (Chang *et al.*, 2015) ignoring all correlations between variant genotypes at r^2<0.01 and trans-chromosome correlations.

2.6.2 Functional annotation categories

Two non-overlapping functional annotation categories were considered: exonic and non-exonic. The exonic annotation category includes all variants within exons (including 5′ and 3′ untranslated regions) of protein-coding genes, while the non-exonic category contains all remaining variants. This choice was motivated by previous research showing that protein-coding exons (including 5′ and 3′ untranslated regions) are most strongly enriched for association with many complex human phenotypes (Schork *et al.*, 2013). Additionally, the exonic category, as defined above, largely overlaps with the genomic regions investigated in whole-exome genotyping and whole-exome sequencing studies. Its modeling can therefore serve as a projection for future discoveries in whole-exome studies. All variants were functionally annotated using UCSC's Table Browser (hg19/GRCh37) (Karolchik *et al.*, 2004). With this definition, the non-exonic category contains approximately 70 times more variants than the exonic category.

2.6.3 Synthetic phenotypes

Synthetic phenotypes were generated using SIMU (Frei, 2016). A given number of non-null variants was selected at random for each functional annotation category. Effect sizes for the selected non-null variants were sampled from the standard normal distribution and then rescaled to obtain the required level of heritability, given different predefined ratios (see below) between the average effect sizes of the two dichotomous functional annotation categories. For each synthetic genotype, a quantitative synthetic phenotype was then generated as the sum of allelic effects over all non-null variants complemented by a certain proportion of a random Gaussian noise (representing effects of the environment) required to keep the predefined level of heritability. Finally, association tests were performed using PLINK 1.9 to obtain GWAS summary statistics.

2.6.4 Simulation setup

All possible combinations of the following parameter values were used for generating synthetic phenotypes: π_{exonic}=10^{-1}, 10^{-2}, 10^{-3}; $\pi_{non-exonic}$=10^{-2}, 10^{-3}, 10^{-4}; $\sigma^2_{exonic}/\sigma^2_{non-exonic}$=0.1, 1.0, 10.0; h^2_{total}=0.1, 0.4, 0.7, resulting in 81 different parameter setups covering a broad range of genetic architectures. Ten different phenotypes with independently generated locations of non-null variants and effect sizes thereof were generated for each combination of parameters, resulting in 810 synthetic phenotypes (and corresponding GWAS summary statistics).

2.7 GWAS summary statistics

We applied the model to GWAS summary statistics on 11 phenotypes (Table 1). Like in simulations with synthetic data, we considered here two functional annotation categories for the variants: exonic and non-exonic. For ease of comparison with partitioned LDSC method (Finucane *et al.*, 2015), the LD structure was estimated with PLINK 1.9 using genotype data from LDSC's template containing 9 997 231 biallelic variants for 489 unrelated European individuals [originally derived from 1000 Genomes Phase 3 data (1000 Genomes Project Consortium *et al.*, 2015)]. Trans-chromosome correlations as well as correlations between variant genotypes at r^2 <0.05 were disregarded. For each phenotype, 50 independent optimization runs were performed to maximize the likelihood of the GWAS z-scores observed in different subsets of 10^6 randomly selected variants.

3 Results

3.1 Simulations with synthetic data

The simulations with synthetic data demonstrate that the true parameters are estimated accurately when the proportions of heritability carried by both functional categories are comparable and each category individually carries >2% of the total heritability; if one of the functional categories carries a negligible fraction (<2%) of the total heritability the model often fails to reconstruct its parameters accurately (Supplementary Fig. S1). Selected simulation cases representing scenarios closely resembling complex human phenotypes analyzed in this study are presented in Figure 1. These simulations show that in the range of parameters observed (according to the model) in the 11 phenotypes analyzed in this study, the model is able to provide instructive unbiased estimates of π and σ^2 parameters for both exonic and non-exonic functional annotation categories. A complete comparison of true simulation parameters and corresponding model estimates for all 810 simulated phenotypes is shown in Supplementary Figures S2–S4, and the corresponding numerical results are given in Supplementary Table S3. Of note is that, in general, heritability estimates are more robust than estimates of π and σ^2.

Table 1. Details of GWAS on 11 phenotypes analyzed in the study

Phenotype	Publication	Sample size (total or cases/controls)
Schizophrenia (SCZ), 49 European sub-studies	Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014)	33 640/43 456
Bipolar disorder (BD)	Stahl *et al.* (2019)	20 352/31 358
Attention-deficit/hyperactivity disorder (ADHD)	Demontis *et al.* (2019)	19 099/34 194
General cognitive ability (COG)	Savage *et al.* (2018)	269 867
Educational attainment (EA)	Lee *et al.* (2018)	766 345
Type 2 diabetes (T2D)	Mahajan *et al.* (2018)	74 124/824 006
Inflammatory bowel disease (IBD)	de Lange *et al.* (2017)	25 042/34 915
Low-density lipoproteins (LDL)	Willer *et al.* (2013)	188 577
Body mass index (BMI)	Yengo *et al.* (2018)	795 640
Height	Yengo *et al.* (2018)	709 706
Waist-hip ratio (WHR)	Shungin *et al.* (2015)	224 459

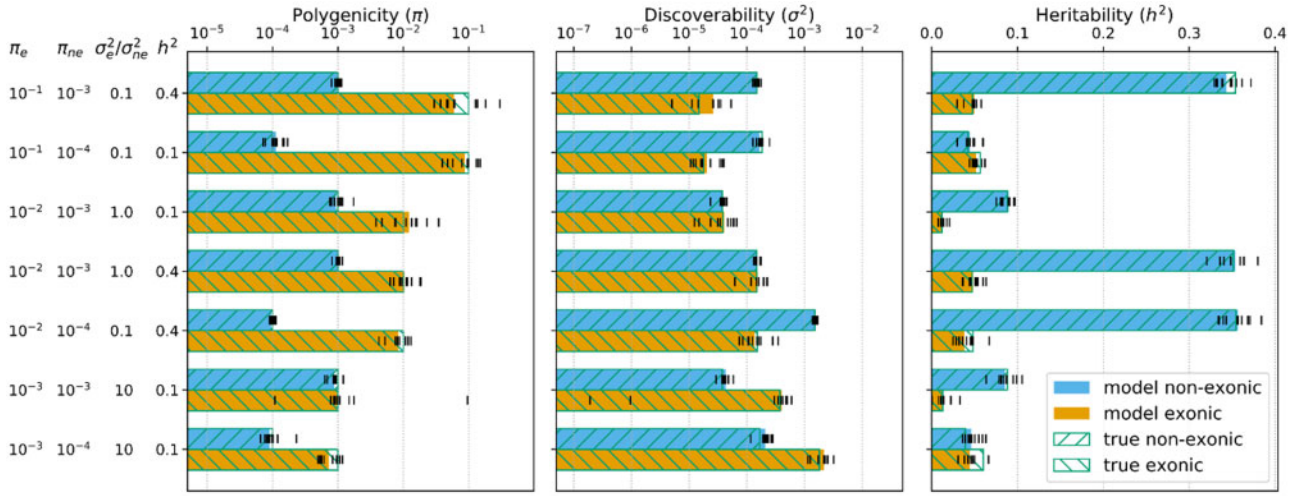

Fig. 1. Performance of the model on a selected set of scenarios with synthetic GWAS data. True simulation parameters (π_{exonic}, $\pi_{\text{non-exonic}}$, $\sigma^2_{\text{exonic}}/\sigma^2_{\text{non-exonic}}$ and h^2_{total}) are shown on the left. The blue bars represent the non-exonic category, the orange bars represent the exonic category. The bar lengths represent the median values obtained from 10 optimization runs with independently generated GWAS (locations and effect sizes of non-null variants). The parameter values from individual optimization runs are shown with vertical black dashes. Empty bars with green borders and hatches show the true values of the corresponding parameters used for GWAS simultion

3.2 GWAS summary statistics

The model was used to estimate exonic and non-exonic polygenicities, discoverabilities and heritabilities of 11 complex phenotypes presented in Table 1 (Fig. 2, Supplementary Table S1). We also obtained estimates of SNP-heritability per functional category for all 11 phenotypes using partitioned LDSC. These estimates are compared with AI-MiXeR's in Figure 2 (right) and Supplementary Table S2. The polygenicity parameters (π_{exonic} and $\pi_{\text{non-exonic}}$) can be converted into the number of non-null variants by multiplying them by the total number of variants within the corresponding annotation category. The numbers ensuing for the analyzed phenotypes are presented in Supplementary Table S1.

The majority of 11 analyzed phenotypes fall into the portion of parameter space where, according to our simulations, the model is expected to produce robust parameter estimates (Supplementary Fig. S1, Fig. 1). However, two phenotypes (ADHD and WHR) fall in a portion of parameter space where the model is prone to return inconsistent results in the exonic category (due to this category's limited size, its very low polygenicity in ADHD and its very low discoverability in WHR). This is reflected in larger error bars for the exonic category in these phenotypes (Fig. 2, Supplementary Table S1). However, the observed consistency of parameter estimates across all 50 independent optimization runs for all analyzed phenotypes suggests that some robust conclusions can be drawn about actual features of the underlying genetic architecture.

The model suggests, that despite having similar heritability, phenotypes may differ substantially in polygenicity and discoverability of non-null variants. For example both AI-MiXeR and partitioned LDSC provide comparable estimates of total and partitioned heritability for LDL and T2D (AI-MiXeR LDL: h^2_{total}= 0.14, h^2_{exonic}= 0.09; AI-MiXeR T2D: h^2_{total}= 0.13, h^2_{exonic}= 0.06) (Fig. 2, Supplementary Table S2). However, our model suggests that the genetic architectures underlying these phenotypes differ drastically, with T2D being approximately 5 times more polygenic than LDL and having 91% (versus 15% in LDL) of non-null variants within exons. The polygenicity deficit is compensated in LDL with a discoverability 3.5 times larger than in T2D (50 times larger in the exonic category). According to the model, EA has the largest number of non-null variants (48 000, with only 0.6% of exonic variants) among all analyzed phenotypes, while IBD has the smallest (3000, 91% exonic) (Supplementary Table S1). However, the effects of the non-null variants are on average five times stronger in IBD than in EA and result in a larger total heritability for the former (0.25 in IBD versus 0.1 in EA). SCZ and BD show similar polygenicity

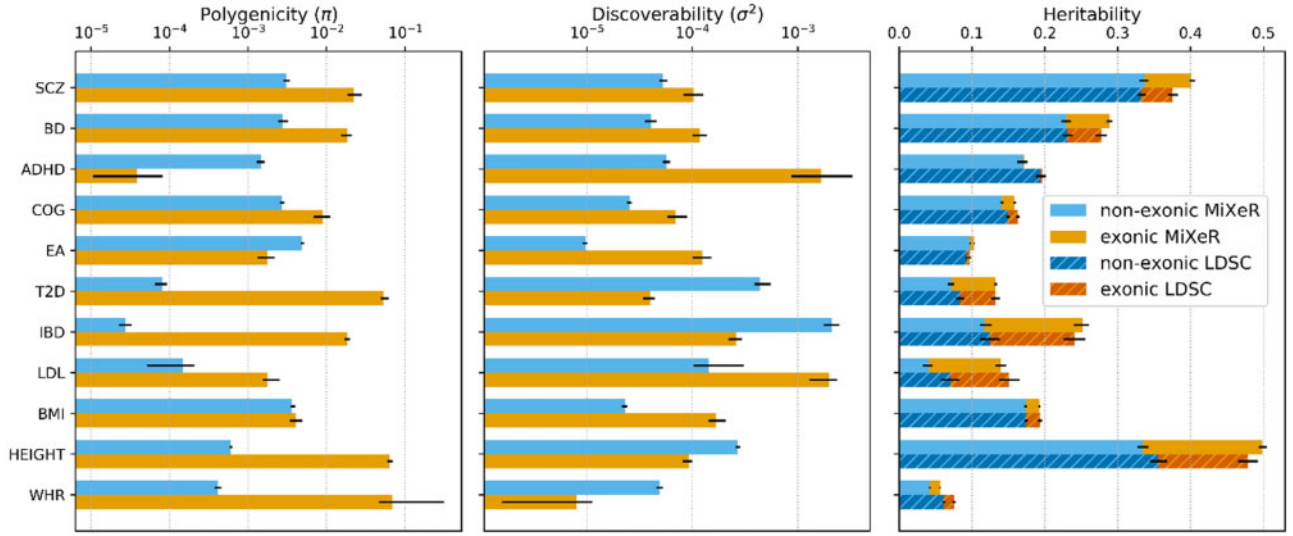

Fig. 2. Estimated polygenicity (proportion of non-null variants), discoverability (variance of non-null effect sizes) and heritability of exonic and non-exonic functional categories in 11 traits. Traits are shown in the left column: schizophrenia (SCZ), bipolar disorder (BD), attention-deficit/hyperactivity disorder (ADHD), general cognitive ability (COG), educational attainment (EA), type 2 diabetes (T2D), inflammatory bowel disease (IBD), low-density lipoproteins (LDL), body mass index (BMI), height and waist-hip ratio (WHR). For AI-MiXeR results, a bar's length shows the mean value of the parameter obtained from 50 independent optimization runs with 106 randomly selected variants used to maximize the likelihood of the observed GWAS *z*-scores, while the black bars show the range (min, max) of such estimates. The heritability estimates obtained with partitioned LDSC are shown in darker colors with hatching (right) and black bars representing the standard errors of the estimates

(34 000 and 30 000 non-null variants, of which 9% and 10%, respectively, are exonic) and discoverability (with exonic effects being approximately twice stronger). In contrast, most non-null variants for height and T2D are exonic (62% out of 15 500 and 90% out of 9000, respectively) having on average 10- and 3-times weaker effects, respectively, than non-exonic variants.

4 Discussion

We present the AI-MiXeR model, which can be used to decouple and partition a phenotype's heritability into functional category-specific polygenicity (proportion of non-null variants in a given category) and discoverability (variance of non-null effect sizes) components and thus better characterize the phenotype's genetic architecture. This may inform the design of future genetic studies, as efforts to improve discoverability in different genomic categories are likely to have different impact across complex phenotypes depending on their unique genetic architecture.

It is widely assumed that protein-coding exons contain a higher proportion of causal variants (higher polygenicity) and have on average stronger effects (higher discoverability) on complex phenotypes compared to non-exonic regions (Minelli *et al.*, 2013; Schork *et al.*, 2013; Sveinbjornsson *et al.*, 2016). In our study, the AI-MiXeR model suggests that less than half (5 of 11) of the analyzed phenotypes (SCZ, BD, COG, LDL and BMI) support this assumption. Four other phenotypes (T2D, IBD, height and WHR) show higher density of non-null variants in exonic regions but stronger average effects in the non-exonic portion of the genome. In the two remaining traits (ADHD and EA), the pattern appears to be reversed, with a higher density of weaker effect variants in non-exonic regions. Since non-exonic regions cover a substantially larger fraction of the genome compared to exonic regions (containing roughly 70 times more variants), the former account for a greater portion of SNP-heritability than the latter for most phenotypes. Only IBD and T2D present substantially higher fractions of non-null variants in exonic variants.

From our simulation studies on synthetic GWAS, we can infer that the balance of h^2 partition between the functional annotation categories has a strong effect on the model's performance. Extremely small values of polygenicity (π) or discoverability (σ^2) in a functional annotation category (relative to the complementary category) result in a heavily unbalanced heritability partition between the categories and can thus lead to substantial errors in the estimates of π and σ^2 for the category with smaller absolute heritability (Supplementary Fig. S1 top and bottom). Despite this, heritability estimates were generally robust (Supplementary Figs S1–S4, Supplementary Table S3).

Decoupling the heritability of different functional categories into polygenicity and discoverability may facilitate trait-specific experimental designs prioritizing certain genomic regions for detailed investigation. For instance, by looking only at the heritability pertaining exons in T2D ($h^2_{\text{exonic}} = 0.06$) and LDL ($h^2_{\text{exonic}} = 0.09$), one could expect the yield of an exome-wide scan for both phenotypes to be comparable. However, AI-MiXeR predicts that the average effect size (square root of discoverability) of exonic non-null variants is approximately seven times larger in LDL than in T2D. An exome study of the former therefore could be expected to result in a higher yield of statistically significant findings, given a moderately sized sample. This speculation may be indirectly supported by comparing existing exome-wide studies of T2D and LDL. One of the largest exome sequencing studies on T2D published so far (20 791 cases, 24 440 controls) identified 15 variants in 7 distinct genomic loci reaching exome-wide significance level (Flannick *et al.*, 2019). In contrast, an exome-wide association study of serum lipids in a comparable sample ($N = 39\,087$) reported 66 exome-wide significant LDL susceptibility variants within 14 loci (Dewey *et al.*, 2016). AI-MiXeR's predictions also suggest, however, that a significant increase in the sample size in T2D whole-exome studies will yield more phenotype-associated variants than an equivalent sample size increase in LDL whole-exome studies, since T2D has substantially larger polygenicity.

AI-MiXeR relies on design and implementation quality of the specific GWAS. In general, model predictions for a given phenotype may differ depending on a GWAS's sample size, as well as on the coverage of the tested variants. The sample sizes of the GWAS tested here vary by more than one order of magnitude, from approximately 5×10^4 for BD and ADHD to more than 7×10^5 for EA and height. In all simulations, we kept the sample size constant ($N = 10^5$) and varied only the heritability ($h^2 = 0.1, 0.4, 0.7$). Since these quantities contribute to the GWAS *z*-scores distribution only through their product [follows from formula (5)], our simulation scenario with

$N = 10^5$ and h^2=0.7 is equivalent to a scenario with, for example, $N = 7 \times 10^5$ and $h^2 = 0.1$ (given that polygenicities are equal in both scenarios). Other aspects of potential GWAS-related issues (e.g. coverage of tested variants) were not tested.

The model underlying AI-MiXeR is sensitive to the LD structure estimates. Ideally, the LD structure should be estimated on the same sample used for association testing. However, this is mostly impractical if not impossible. Here, in the analysis of GWAS summary statistics for 11 phenotypes, we estimated the LD structure using the 1000 Genomes Phase 3 genotype panel. Inconsistencies between the LD structure of the samples used for association testing and that of the 1000 Genomes Phase 3 panel could skew the model's results. Additionally, roughening the LD structure (e.g. by ignoring all correlations with r^2 below a certain threshold) also could result in biased parameter estimates. In our simulations, r^2 was estimated from 10 000 synthetic genotypes (randomly sampled from the complete set of 100 000 synthetic genotypes used for association testing) ignoring all correlations with $r^2 < 0.01$. A subset of European ancestry samples from the 1000 Genomes Phase 3 panel ($N = 489$) was used to estimate LD r^2 values for the GWAS data because of the wide availability of these data, ease of comparison with LDSC and the fact that genotypes in a majority of analyzed GWAS were imputed using this panel as a reference. The limited size of the 1000 Genomes panel, however, results in relatively low confidence r^2 values, especially for weak correlations involving low-frequency variants. To mitigate this issue, we increased the r^2 cutoff, disregarding all correlations with $r^2 < 0.05$. Nevertheless, consistency between partitioned heritability estimates produced by AI-MiXeR and LDSC (Fig. 2, Supplementary Table S2) suggests the absence of the model-specific systematic biases.

AI-MiXeR makes further simplifying assumptions, including uniform distribution of non-null variants within functional annotation categories and the effect size's independence of allele frequency and LD. It has recently been shown that these simplified assumptions, which have been used implicitly or explicitly in many earlier methods, can lead to substantial biases in heritability estimates (Speed *et al.*, 2017). We previously demonstrated (Frei *et al.*, 2019) that these factors also bias the model's estimates of π and σ^2 when no distinction is made between annotation categories. We did not investigate how disregarding them affect AI-MiXeR's category-specific estimates. These assumptions are likely violated to different degrees in different phenotypes and make the model more suitable for some phenotypes than for others. In this report, we provide examples of successful applications of AI-MiXeR but advise prospective users to carefully assess the model's suitability for a given phenotype. Recently, we proposed a model (Holland *et al.*, 2020a) for the distribution of non-null variants and their effect sizes which takes allele frequency and LD into account. Combining the latter with the AI model presented here involves considerable complexity, however, it is a logical next step. Additionally, the model assumes additivity of genetic effects. Allowing dominance genetic effects and epistasis may further improve model fitness for some phenotypes. Introducing this flexibility requires significant technical complexity and will be considered in our future work.

It is also important to note that the numbers of non-null variants presented (Supplementary Table S1) are estimated for the hypothesized distributions of variant effects, capturing the broad outlines of polygenicity. The observed relative proportions of non-null variants (and their effect sizes) between functional annotation categories within one trait or across different traits might be more reliable indicators of actual genetic architecture features or differences.

The model allows for simulating any number of annotation categories simultaneously (in the marginal scenario, each variant can be treated as a separate annotation category). However, with the current implementation, the computational cost of the optimization increases rapidly as the number of annotation categories grows. For this reason, we restricted the main analysis of this study to exonic and non-exonic categories, which we reckoned could be informative in the context of whole-exome studies. An exploratory analysis of several other functional annotation categories (including intronic, promoter and enhancer regions) for SCZ and T2D can be found in Supplementary Material.

The AI-MiXeR method presented here considers predefined annotation categories allowing both different proportions of non-null variants and different effect size distributions in various functional annotation categories, which is not possible with other methods available to date (Finucane *et al.*, 2015; Zhang *et al.*, 2018; Zhu and Stephens, 2018). The ability to model predefined annotation categories separately allows hypothesis-driven studies of complex phenotypes, which in turn can provide a better understanding of the genetic architecture of those complex phenotypes. Our analysis suggests that both the polygenicity and the discoverability in different functional categories vary considerably across human traits and disorders. Knowing such patterns may facilitate trait-specific experimental designs prioritizing specific genomic regions for detailed investigation.

Acknowledgements

The simulations were performed on the Abel Cluster, owned by the University of Oslo and Uninett/Sigma2, and operated by the Department for Research Computing at USIT, the University of Oslo IT-department (http://www.hpc.uio.no/).

Funding

This work was supported by the Research Council of Norway [#223273, #225989, #248778]; South-East Norway Health Authority [#2016-064, #2017-004]; KG Jebsen Stiftelsen [#SKGJ-Med-008] and the National Institutes of Health [U24DA041123 to A.M.D.].

Conflict of Interest: A.M.D. is a Founder of and holds equity in CorTechs Labs, Inc., and serves on its Scientific Advisory Board. He is a member of the Scientific Advisory Board of Human Longevity, Inc. and receives funding through research agreements with General Electric Healthcare and Medtronic, Inc. The terms of these arrangements have been reviewed and approved by UCSD in accordance with its conflict of interest policies. O.A.A. is a consultant for HealthLytix. The remaining authors have no competing interest.

References

1000 Genomes Project Consortium *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Bulik-Sullivan,B.K. *et al.*; Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

Chang,C.C. *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.

de Lange,K.M. *et al.* (2017) Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.*, **49**, 256–261.

Demontis,D. *et al.*; ADHD Working Group of the Psychiatric Genomics Consortium (PGC). (2019) Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.*, **51**, 63–75.

Dewey,F.E. *et al.* (2016) Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR Study. *Science*, **354**, aaf6814.

Evans,L.M. *et al.*; Haplotype Reference Consortium. (2018) Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.*, **50**, 737–745.

Finucane,H.K. *et al.*; ReproGen Consortium. (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, **47**, 1228–1235.

Flannick,J. *et al.*; Broad Genomics Platform. (2019) Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature*, **570**, 71–76.

Frei,O. (2016) Precimed/SIMU GitHub page: https://github.com/precimed/simu.

Frei,O. *et al.* (2019) Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nat. Commun.*, **10**, 2417.

Holland,D. *et al.* (2020a) Linkage disequilibrium and heterozygosity modulate the genetic architecture of human complex phenotypes: evidence of natural selection from GWAS summary statistics. *bioRxiv* 2020:705285.

Holland,D. *et al.* (2020b) Beyond SNP heritability: polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLoS Genet.*, **16**, e1008612.

Karolchik,D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–496.

Lee,J.J. *et al.*; 23andMe Research Team. (2018) Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.*, **50**, 1112–1121.

Mahajan,A. *et al.* (2018) Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.*, **50**, 1505–1513.

Minelli,C. *et al.* (2013) Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. *Genet. Epidemiol.*, **37**, 205–213.

Savage,J.E. *et al.* (2018) Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.*, **50**, 912–919.

Schaub,M.A. *et al.* (2012) Linking disease associations with regulatory information in the human genome. *Genome Res.*, **22**, 1748–1759.

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

Schork,A.J. *et al.*; The Tobacco and Genetics Consortium. (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet.*, **9**, e1003449.

Schork,A.J. *et al.* (2016) New statistical approaches exploit the polygenic architecture of schizophrenia–implications for the underlying neurobiology. *Curr. Opin. Neurobiol.*, **36**, 89–98.

Shungin,D. *et al.*; The ADIPOGen Consortium. (2015) New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, **518**, 187–196.

Smeland,O.B. *et al.* (2019) The emerging pattern of shared polygenic architecture of psychiatric disorders, conceptual and methodological challenges. *Psychiatr. Genet.*, **29**, 152–159.

Speed,D. *et al.*; the UCLEB Consortium. (2017) Reevaluation of SNP heritability in complex human traits. *Nat. Genet.*, **49**, 986–992.

Stahl,E.A. *et al.*; eQTLGen Consortium. (2019) Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.*, **51**, 793–803.

Su,Z. *et al.* (2011) HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27**, 2304–2305.

Sveinbjornsson,G. *et al.* (2016) Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.*, **48**, 314–317.

Timpson,N.J. *et al.* (2018) Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.*, **19**, 110–124.

Virtanen,P. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, **17**, 261–272.

Willer,C.J. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.

Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.

Yengo,L. *et al.*; the GIANT Consortium. (2018) Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum. Mol. Genet.*, **27**, 3641–3649.

Zhang,Y. *et al.* (2018) Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.*, **50**, 1318–1326.

Zhu,X. and Stephens,M. (2018) Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.*, **9**, 4361.