Bioinformatics, 34, 2018, i195–i201
doi: 10.1093/bioinformatics/bty254
ISMB 2018

OXFORD

A unifying framework for joint trait analysis under a non-infinitesimal model

Ruth Johnson[1,*], Huwenbo Shi[2], Bogdan Pasaniuc[2,3,4,†] and Sriram Sankararaman[1,2,3,†]

[1]Department of Computer Science, [2]Bioinformatics Interdepartmental Program, [3]Department of Human Genetics, David Geffen School of Medicine and [4]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90024, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that these authors contributed equally.

Abstract

Motivation: A large proportion of risk regions identified by genome-wide association studies (GWAS) are shared across multiple diseases and traits. Understanding whether this clustering is due to sharing of causal variants or chance colocalization can provide insights into shared etiology of complex traits and diseases.
Results: In this work, we propose a flexible, unifying framework to quantify the overlap between a pair of traits called UNITY (Unifying Non-Infinitesimal Trait analYsis). We formulate a Bayesian generative model that relates the overlap between pairs of traits to GWAS summary statistic data under a non-infinitesimal genetic architecture underlying each trait. We propose a Metropolis–Hastings sampler to compute the posterior density of the genetic overlap parameters in this model. We validate our method through comprehensive simulations and analyze summary statistics from height and body mass index GWAS to show that it produces estimates consistent with the known genetic makeup of both traits.
Availability and implementation: The UNITY software is made freely available to the research community at: https://github.com/bogdanlab/UNITY.
Contact: ruthjohnson@ucla.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have identified thousands of regions in the genome that contain variants that contribute to risk for many diseases. Many of these risk regions have been implicated in multiple phenotypes such as autism and schizophrenia (Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium *et al.*, 2017), multiple autoimmune diseases (Cotsapas *et al.*, 2011; Ramos *et al.*, 2011; Richard-Miceli and Criswell, 2012), Crohn's disease and psoriasis (Ellinghaus *et al.*, 2012), and many others. Understanding which causal variants are shared among diseases can provide novel etiological insight as well as provide evidence of potential shared causal mechanisms between complex traits. In addition, identifying which variants contribute to multiple traits can help decipher which molecular traits (e.g. gene expression) contribute to disease risk (Giambartolomei *et al.*, 2014; Hormozdiari *et al.*, 2016); genetic variants that causally alter gene expression as well as disease risk can link a particular gene to a given disease.

Genetic overlap has been analyzed both at the genome-wide level and local level, where the latter refers to analysis done within a given genomic region. *Genetic correlation*, a measure that quantifies the similarity in the genetic effects on pairs of traits, is commonly used for assessing the relationship between two traits and can be applied either genome-wide or to local data (Bulik-Sullivan *et al.*, 2015; Shi *et al.*, 2017). Many of the models for estimating genome-wide genetic correlation assume an *infinitesimal* genetic architecture where all SNPs, or single nucleotide polymorphisms, are assumed to have a very small effect on the trait. In contrast to genetic correlation, *colocalization* methods aim to estimate whether the GWAS association signals for two traits at the same region are due to the same causal variant across the traits or chance (Giambartolomei *et al.*, 2014; Hormozdiari *et al.*, 2016). The methods that relax the infinitesimal assumption either assume a single causal variant per region or limit the number of potential causal variants a priori, often due to computational considerations (Giambartolomei *et al.*, 2014; Hormozdiari *et al.*, 2016). Although, both genetic correlation

and colocalization aim to describe the genetic sharing between traits, these methods have been utilized largely independently of each other.

In this work, we present a unifying statistical model that ties together genetic correlation and colocalization. To accomplish this, we present a fully generative Bayesian statistical model that models the shared as well distinct genetic variants underlying a pair of traits. The model allows for sparse genetic architectures (where only a small fraction of variants are causally impacting the traits). The model is richly parametrized: allowing us to jointly model global parameters such as the proportion of variants that are causal for both as well for either trait, the trait heritability, the correlation of the effect sizes at the causal SNPs and local parameters such as the effect of a single SNP on each of the traits.

A challenge of a non-infinitesimal genetic architecture is that it presents a computationally challenging inference problem. Performing inference under this model often involves explicitly enumerating all causal configurations of the SNPs. This exponential search space of 2^{2M}, where M is the number of SNPs analyzed, proves intractable given the large genetic datasets now available. We propose Unifying Non-Infinitesimal Trait analYsis (UNITY) that relies on Markov Chain Monte Carlo (MCMC) to approximate the posterior probabilities of the model parameters. In this work, we focus on estimating the proportion of shared and trait-specific causal variants since parameters such as heritability and genetic correlation can be estimated using previous methods (Bulik-Sullivan *et al.*, 2015). Additionally, a key advantage of the method is that it only requires summary level association statistic data, which bypasses many of the privacy concerns associated with individual level data. With the widespread availability of GWAS summary statistics (Pasaniuc and Price, 2017), we expect that a method operating only on summary statistics would prove most useful for the research community. Through comprehensive simulations and an analysis of height and body mass index (BMI), we show that our method can accurately estimate the proportion of shared causal SNPs between two complex traits.

2 Materials and methods

2.1 Generative model

Here, we introduce a Bayesian framework for estimating the proportion of causal variants shared between a pair of complex traits. The input to our method is the vector of signed effect sizes at each SNP for each trait (we only analyze SNPs for which effect size estimates are available for both traits). We model the genetic as well as non-genetic variances in each trait, the genetic correlation among the traits, and the proportion of causal SNPs that are shared across traits as well as are unique to each. The proportion of causal SNPs shared between the traits is denoted by p_{11}, the proportion of causal SNPs specific to trait 1 and trait 2 as p_{10} and p_{01}, respectively, and the proportion of non-causal SNPs is denoted by p_{00}, where $p_{00}+p_{10}+p_{01}+p_{11}=1$. For each trait $p \in \{1,2\}$, we denote the genetic variance σ_p^2 (which is the same as its heritability as h_p^2 if the trait is standardized), the environmental noise as $\sigma_{e_p}^2 = \frac{1-h_p^2}{N_p}$, where N_p denotes the sample size for trait p, and the genetic correlation between the two traits as ρ. Altogether, our model has the following parameters: $(\sigma_1^2, \sigma_2^2, \rho, p_{00}, p_{10}, p_{01}, p_{11})$.

We assume that trait p ($p \in \{1,2\}$) measured in individual i, $y_{p,i}$ is a linear function of standardized genotypes $\boldsymbol{x}_{=}(x_{i,1},\ldots,x_{i,M})$ measured across M SNPs with SNP effect sizes $\boldsymbol{\beta}_p = (\beta_{p,1},\ldots,\beta_{p,M})$ and independent additive noise term $\epsilon_{p,i}$. Further, we assume that there are no sample overlaps across the two studies.

$$y_{1,i} = \sum_{m=1}^{M} \beta_{1,m} x_{i,m} + \epsilon_{1,i} \qquad y_{2,i} = \sum_{m=1}^{M} \beta_{2,m} x_{i,m} + \epsilon_{2,i}$$

$$i \in \{1,\ldots,N_1\} \qquad i \in \{1,\ldots,N_2\}$$

$$\epsilon_{p,i} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_{e_p}^2\right)$$

A SNP m is causal for trait p if its true effect $\beta_{p,m} \neq 0$ and it is not causal otherwise. We denote the probability of a SNP being causal for every combination of the two traits as: $\boldsymbol{p} = (p_{00}, p_{10}, p_{01}, p_{11})$.

Denoting the causal effect sizes for trait $p, p \in \{1,2\}$ across all SNPs $\boldsymbol{\gamma}_p = (\gamma_{p,1},\ldots,\gamma_{p,M})$, we assume that the causal effect sizes for each SNP are independent, allowing us to model the effect sizes at SNP m for each of the two traits $(\gamma_{1,m}, \gamma_{2,m})$ as a random vector drawn from a bi-variate normal distribution centered at zero with the following covariance matrix:

$$\begin{pmatrix}\gamma_{1,m}\\ \gamma_{2,m}\end{pmatrix} \mid (\sigma_1^2, \sigma_2^2, \rho, \boldsymbol{p}) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\frac{\sigma_1^2}{M(p_{11}+p_{10})} & \frac{\sigma_1\sigma_2\rho}{M(p_{11})}\\ \frac{\sigma_1\sigma_2\rho}{M(p_{11})} & \frac{\sigma_2^2}{M(p_{11}+p_{01})}\end{pmatrix}\right)$$

$\boldsymbol{C}_p = (C_{p,1},\ldots,C_{p,M})$ denotes the causal indicator vector for trait p, where $C_{p,m}=1$ if SNP m is causal for trait p and 0 otherwise. $(C_{1,m}, C_{2,m})$ is a random vector drawn from a discrete distribution with parameters given by $\boldsymbol{p}$:

$$P\left(\begin{pmatrix}C_{1,m}\\ C_{2,m}\end{pmatrix} = \begin{pmatrix}a\\ b\end{pmatrix} \mid \boldsymbol{p}\right) = p_{ab}, \quad a,b \in \{0,1\}$$

The true effect sizes for each trait p at SNP m, $\beta_{p,m}$, conditioned on the causal status at a SNP is the element-wise product of the causal indicator vector and the true causal effect sizes.

$$\begin{pmatrix}\beta_{1,m}\\ \beta_{2,m}\end{pmatrix} \mid \begin{pmatrix}C_{1,m}\\ C_{2,m}\end{pmatrix}, \begin{pmatrix}\gamma_{1,m}\\ \gamma_{2,m}\end{pmatrix} = \begin{pmatrix}\gamma_{1,m}C_{1,m}\\ \gamma_{2,m}C_{2,m}\end{pmatrix}$$

We can model the conditional distribution of the GWAS summary statistics given the true effect sizes, where $\widehat{\beta}_{p,m}$ is the estimated marginal effect size of the mth SNP for trait p (Shi *et al.*, 2017):

$$\begin{pmatrix}\widehat{\beta}_{1,1:M}\\ \widehat{\beta}_{2,1:M}\end{pmatrix} \mid \left(\begin{pmatrix}\beta_{1,1:M}\\ \beta_{2,1:M}\end{pmatrix}, \sigma_{e_1}^2, \sigma_{e_2}^2\right) \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{V}\beta_{1,1:M}\\ \boldsymbol{V}\beta_{2,1:M}\end{pmatrix}, \Sigma_e\right)$$

$$\Sigma_e = \begin{pmatrix}\sigma_{e_1}^2\boldsymbol{V} & 0\\ 0 & \sigma_{e_2}^2\boldsymbol{V}\end{pmatrix}$$

$\boldsymbol{V}$ is the matrix of correlations among the SNPs, i.e. the linkage disequilibrium (LD) matrix. $\boldsymbol{V}$ can be estimated from a reference panel of genotypes collected from a population that is genetically similar to the populations for which summary statistics are available. Alternately, when performing inference at the genome-wide level, we can prune the list of SNPs such that they come from independent LD blocks. LD-pruning creates an approximately independent subset of SNPs in which case $\boldsymbol{V}$ can be approximated by the identity matrix, $\boldsymbol{I}$. In this work, we restrict our attention to the case where $\boldsymbol{V} \approx \boldsymbol{I}$.

We impose a Dirichlet prior on $\boldsymbol{p}$:

$$\rho|\lambda \sim \mathrm{Dir}(\lambda)$$

Here $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. In practice, we set $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda = 0.20$:

In principle, we can also impose priors on the remaining parameters, i.e. the trait heritability (σ_1^2, σ_2^2) and their genetic correlation ρ and estimate all of these parameters jointly with $\boldsymbol{p}$ in a fully Bayesian model. These parameters can be estimated using other methods (Bulik-Sullivan *et al.*, 2015) and, in this work, we fix the values of these parameters to their estimates and focus on estimating $\boldsymbol{p}$.

Given the parameters $(\sigma_1^2, \sigma_2^2, \rho, \lambda)$, the joint distribution of the probability of causal configurations $\boldsymbol{p}$, the causal indicator vectors $\mathbf{C}_1, \mathbf{C}_2$, the causal effect sizes $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$, and the estimated effect sizes $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\beta}}_2$ is given by:

$$\begin{aligned}
&P\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \mathbf{C}_1, \mathbf{C}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{p} | \sigma_1^2, \sigma_2^2, \rho, \boldsymbol{\lambda}\right) \\
&= P(\boldsymbol{p}|\boldsymbol{\lambda}) \\
&\times \prod_{m=1}^{M} \left\{ P\left(\begin{pmatrix} C_{1,m} \\ C_{2,m} \end{pmatrix} | \boldsymbol{p} \right) P\left(\begin{pmatrix} \gamma_{1,m} \\ \gamma_{2,m} \end{pmatrix} | (\sigma_1^2, \sigma_2^2, \rho, \boldsymbol{p}) \right) \right. \\
&\left. \times P\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix} | \left(\begin{pmatrix} \gamma_{1,m} \\ \gamma_{2,m} \end{pmatrix}, \begin{pmatrix} C_{1,m} \\ C_{2,m} \end{pmatrix}, \sigma_{e_1}^2, \sigma_{e_2}^2 \right) \right) \right\}
\end{aligned}$$

Integrating over the hidden variables $\mathbf{C}_1, \mathbf{C}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$, we obtain:

$$\begin{aligned}
&P(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \boldsymbol{p} | \sigma_1^2, \sigma_2^2, \rho, \boldsymbol{\lambda})) \\
&= P(\boldsymbol{p}|\boldsymbol{\lambda}) \times \prod_{m=1}^{M} \left[\int \sum_{\substack{C_{1,m} \\ C_{2,m}}} \left\{ P\left(\begin{pmatrix} C_{1,m} \\ C_{2,m} \end{pmatrix} | \boldsymbol{p} \right) P\left(\begin{pmatrix} \gamma_{1,m} \\ \gamma_{2,m} \end{pmatrix} | (\sigma_1^2, \sigma_2^2, \rho, \boldsymbol{p}) \right) \right. \right. \\
&\left. \left. \times P\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix} | \left(\begin{pmatrix} \gamma_{1,m} \\ \gamma_{2,m} \end{pmatrix}, \begin{pmatrix} C_{1,m} \\ C_{2,m} \end{pmatrix}, \sigma_{e_1}^2, \sigma_{e_2}^2 \right) \right) \right\} d\gamma_{1,m} d\gamma_{2,m} \right] \\
&= \mathrm{Dir}(\boldsymbol{p}; \boldsymbol{\lambda}) \\
&\times \prod_{m=1}^{M} \left[p_{00} \mathcal{N}\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{e_1}^2 & 0 \\ 0 & \sigma_{e_2}^2 \end{pmatrix} \right) \right. \\
&\quad + p_{10} \mathcal{N}\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\sigma_1^2}{M(p_{11}+p_{10})} + \sigma_{e_1}^2 & 0 \\ 0 & \sigma_{e_2}^2 \end{pmatrix} \right) \\
&\quad + p_{01} \mathcal{N}\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{e_1}^2 & 0 \\ 0 & \frac{\sigma_2^2}{M(p_{11}+p_{01})} + \sigma_{e_2}^2 \end{pmatrix} \right) \\
&\left. + p_{11} \mathcal{N}\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\sigma_1^2}{M(p_{11}+p_{10})} + \sigma_{e_1}^2 & \frac{\sigma_1\sigma_2}{M(p_{11})}\rho \\ \frac{\sigma_1\sigma_2}{M(p_{11})}\rho & \frac{\sigma_2^2}{M(p_{11}+p_{01})} + \sigma_{e_2}^2 \end{pmatrix} \right) \right]
\end{aligned}$$

2.2 Parameter inference in our model

Given the generative model described in the previous section, the inference problem lies in computing the posterior distribution of $\boldsymbol{p}$ given the estimated summary statistics

$$P\left(\boldsymbol{p} | \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \sigma_1^2, \sigma_2^2, \rho, \boldsymbol{\lambda}\right) = \frac{P\left(\boldsymbol{p}, \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2 | \sigma_1^2, \sigma_2^2, \rho, \boldsymbol{\lambda}\right)}{P\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2 | \sigma_1^2, \sigma_2^2, \rho, \boldsymbol{\lambda}\right)}$$

The true joint posterior distribution is intractable. Thus, we use MCMC (Brooks *et al.*, 2011) to approximate the posterior distribution. MCMC approximates the target posterior distribution $P(\boldsymbol{p} | \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \sigma_1^2, \sigma_2^2, \rho)$ by a sequence of random samples $\left(\boldsymbol{p}^{(t)}\right)_{t=1}^{T}$ drawn from a Markov chain constructed so that the stationary distribution of the chain is the target posterior.

$$P\left(\boldsymbol{p} | \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \sigma_1^2, \sigma_2^2, \rho, \boldsymbol{\lambda}\right) \approx \frac{1}{T} \sum_{t=1}^{T} \delta_{\boldsymbol{p}^{(t)}}(\boldsymbol{p})$$

In our setting, we use a random-walk Metropolis–Hastings algorithm (Metropolis *et al.*, 1953) that generates a sample $\boldsymbol{p}^{(t+1)}$ at iteration $t+1$ given the sample $\boldsymbol{p}^{(t)}$ at the previous iteration using the following proposal distribution that generates a proposed sample $\boldsymbol{p}^*$ which is then accepted or rejected depending on the Metropolis–Hastings ratio (which depends on the ratio of the posterior probability density at the proposed parameter to the previous parameter):

$$\boldsymbol{p}^* \sim \mathrm{Dir}(\boldsymbol{d})$$

$$\boldsymbol{d} = \boldsymbol{\lambda} + B\boldsymbol{p}^{(t)}$$

Here, B is a constant that controls the variance of the proposal distribution. In practice, we found that $B = 10$ yields effective mixing.

The final step in specifying the MCMC algorithm lies in computing the ratio of the posterior probability density at the proposed parameter to the original parameter. Computation of the ratio requires the evaluation of the posterior probability only up to a normalization constant:

$$\begin{aligned}
&P\left(\boldsymbol{p} | \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \sigma_1^2, \sigma_2^2, \rho, \boldsymbol{\lambda}\right) \\
&\propto P\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \boldsymbol{p} | \sigma_1^2, \sigma_2^2, \rho, \boldsymbol{\lambda}\right) \\
&= \mathrm{Dir}(\boldsymbol{p}; \boldsymbol{\lambda}) \\
&\times \left[\prod_{m=1}^{M} \mathcal{N}\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{e_1}^2 & 0 \\ 0 & \sigma_{e_2}^2 \end{pmatrix} \right) \cdot (p_{00}) \right. \\
&\quad + \mathcal{N}\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\sigma_1^2}{M(p_{11}+p_{10})} + \sigma_{e_1}^2 & 0 \\ 0 & \sigma_{e_2}^2 \end{pmatrix} \right) \cdot (p_{10}) \\
&\quad + \mathcal{N}\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{e_1}^2 & 0 \\ 0 & \frac{\sigma_2^2}{M(p_{11}+p_{01})} + \sigma_{e_2}^2 \end{pmatrix} \right) \cdot (p_{01}) \\
&\left. + \mathcal{N}\left(\begin{pmatrix} \widehat{\beta}_{1,m} \\ \widehat{\beta}_{2,m} \end{pmatrix}; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\sigma_1^2}{M(p_{11}+p_{10})} + \sigma_{e_1}^2 & \frac{\sigma_1\sigma_2}{M(p_{11})}\rho \\ \frac{\sigma_1\sigma_2}{M(p_{11})}\rho & \frac{\sigma_2^2}{M(p_{11}+p_{01})} + \sigma_{e_2}^2 \end{pmatrix} \right) \cdot (p_{11}) \right]
\end{aligned}$$

2.3 Efficient mixing of MCMC chains

In any practical application of MCMC, the number of iterations, burn-in period, and initialization point are critical to ensuring convergence and accurate estimates. Slow mixing of the MCMC chains can occur if the starting point is at a region of low posterior density. As opposed to selecting a random starting point, we carefully select the initialization of each chain by choosing the set of parameters that yields the highest posterior density. We use the limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Byrd *et al.*, 1994) to determine the maximum a posteriori estimates for $p_{00}, p_{10}, p_{01}, p_{11}$. We repeat this 10 times, initializing the optimization algorithm with random starting values drawn from the prior. We compute the posterior density of all 10 candidate starting values and select the set that yields the highest density. This set of

Table 1. Displayed is a summary of current methods that perform joint trait analysis and the relationship to the parameters in UNITY

Method	h^2	ρ	$\boldsymbol{p}$	Misc.
UNITY	*	*	*	
Cross-trait LD Score regression (Bulik-Sullivan *et al.*, 2015)	*	*	$p_{11} \approx 1$	
PleioPred (Hu *et al.*, 2017a)	*	*	*	infers $\boldsymbol{p}$ to estimate effect sizes
COLOC (Giambartolomei *et al.*, 2014)	–	–	*	max 1 causal
eCAVIAR (Hormozdiari *et al.*, 2016)	–	–	*	max 6 causals

Boxes with an (*) denote the values that a method models. Note that this summary is not exhaustive.

parameters is then used as the starting point for our MCMC chain. In addition, to diagnose convergence, we use 100 Markov chains all initialized using the scheme described above. Our final estimate is the mean of all samples drawn from the 100 chains.

2.4 Note on runtime

We assessed the performance based on the number of seconds per iteration of the MCMC sampler. The main computation is calculating the likelihood at each iteration, which is directly dependent on the number of SNPs per trait. The complexity of the algorithm is $\mathcal{O}(m)$, where m is the number of SNPs. We empirically demonstrate that our method is linear in the number of SNPs through simulation (Supplementary Fig. S1). In addition, the runtime is invariably connected to the number of iterations required for the MCMC to converge. We find that using the maximum a posteriori probability (MAP) estimate as an initialization value leads to fast convergence, requiring only 500 iterations in practice.

3 Results

UNITY provides a novel generalized framework to jointly model GWAS summary statistics data of two complex traits, incorporating fundamental genetic parameters, such as heritability and genetic correlation, and makes minimal assumptions in inference procedures. Since UNITY assumes a non-infinitesimal model, it allows for very sparse genetic architectures, i.e. by setting $p_{00} \approx 1$. However, this non-infinitesimal model can also be generalized to the infinitesimal model by setting $p_{00} \approx 0, p_{10} \approx 0, p_{01} \approx 0, p_{11} \approx 1$.

3.1 UNITY generalizes colocalization and genetic correlation

We discuss a comparison of the parameters of UNITY with those obtained by other methods that perform cross-trait analysis and the underlying assumptions of each method. We first analyze the cross-trait LD score regression model (Bulik-Sullivan *et al.*, 2015), which estimates genome-wide genetic correlation based on the random-effect model, making the implicit assumption that every SNP has a non-zero effect. In contrast to cross-trait LD score regression, UNITY assumes a generalized non-infinitesimal model, explicitly modeling a sparse genetic architecture. We also compare UNITY with methods that do not make the infinitesimal model assumption. While models such as PleioPred explicitly model the proportion of trait-specific and shared causal variants $p_{00}, p_{10}, p_{10}, p_{11}$, the main goal of this method is to perform genetic risk prediction (Hu *et al.*, 2017a) rather than estimating these proportions.

We compare UNITY with COLOC (Giambartolomei *et al.*, 2014) and eCAVIAR (Hormozdiari *et al.*, 2016), Bayesian methods to assess the evidence of colocalization, i.e. whether GWAS signals of two traits are driven the same underlying causal variants. Both methods explicitly model $\boldsymbol{p} = (p_{00}, p_{10}, p_{10}, p_{11})$ (Giambartolomei *et al.*, 2014; Hormozdiari *et al.*, 2016). However, COLOC makes the simplifying assumption that there is at most one-causal variant at a region (Giambartolomei *et al.*, 2014), allowing it to not explicitly model LD. And although eCAVIAR allows for multiple causal variants and explicitly models LD, it restricts the maximum number of causal variants at six per region for computational efficiency (Hormozdiari *et al.*, 2016). In comparison with these methods, UNITY allows for any number of causal variants while making the assumption that there is no LD between the SNPs. We outline a summary of the relationship between UNITY and all methods described in Table 1.

To empirically demonstrate the benefit of the relaxed assumptions of UNITY as compared to current methods, we conduct a modest comparison against COLOC (Giambartolomei *et al.*, 2014). We simulated 100 regions of 500 SNPs with multiple causal variants. We perform colocalization analysis over all of the regions using COLOC. When there are causal variants independently associated with each trait and shared variants, COLOC estimates that the association within the region is driven only by two independent variants, where one is specific to trait 1 and the other is specific to trait 2. Because COLOC assumes at most one-causal variant per region, the method is unable to distinguish between a variant that independently drives only one trait versus a variant that is colocalized when both cases are present. For completeness, we also included a simulation that follows the assumption underlying COLOC of the one-causal setting. The full table listing these results in outlined in Supplementary Table S2. However, we are unable to directly compare estimates with COLOC because there is not a clear mapping between the estimates of COLOC and the estimated parameters of UNITY, thus any direct comparison would be an unfair comparison due to the mismatch in the models.

3.2 Simulations

We generated summary statistics for 500 SNPs from two synthetic GWAS. The causal effect sizes for each SNP, $\gamma_{p,m}$, were drawn jointly from a multivariate normal distribution where h_1^2, h_2^2, ρ denote the heritability of each trait and the genetic correlation. We denote the number of SNPs as M and the proportion of causal variants for each trait as p_{10}, p_{01} and the proportion of shared casuals as p_{11}:

$$\begin{pmatrix} \gamma_{1,m} \\ \gamma_{2,m} \end{pmatrix} \sim \begin{pmatrix} \frac{h_1^2}{M(p_{11}+p_{10})} & \frac{h_1 h_2 \rho}{M(p_{11})} \\ \frac{h_1 h_2 \rho}{M(p_{11})} & \frac{h_2^2}{M(p_{11}+p_{01})} \end{pmatrix}$$

To simulate causal SNPs, we drew an $M \times 4$ matrix from a multinomial distribution parametrized by $\boldsymbol{p}$ where the mth row of values denotes whether a SNP is causal for neither trait, only trait 1, only trait 2, or neither trait. Using this, we constructed two $M \times 1$ causal indicator vectors, C_1, C_2, where $C_{1,m}, C_{2,m} = 1$ if the mth SNP was causal for both traits, $C_{1,m} = 1, C_{2,m} = 0$ if the SNP was

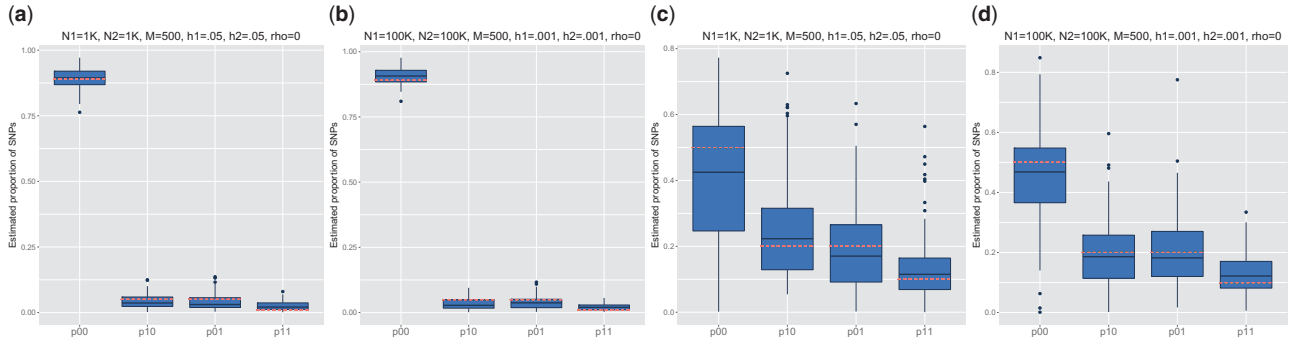

Fig. 1. We estimate the proportion of causal variants under four simulation frameworks where we vary the sample size (N_1, N_2), heritability ($h_1^2 = h_2^2$), and proportion of causal variants. First, we first simulated values where the total proportion of causal variants is low: $p_{00} = 0.89$, $p_{10} = 0.05$, $p_{01} = 0.05$, $p_{11} = 0.01$, along with a low sample size and high heritability: $h_1^2 = 0.05$, $h_2^2 = 0.05$, $\rho = 0$, $N_1 = 1000$, $N_2 = 1000$, as shown in (**a**). Second, we tested the model with the same proportion of causal variants, but with a larger sample size and smaller heritability: $h_1^2 = 0.001$, $h_2^2 = 0.001$, $\rho = 0$, $N_1 = 100\,000$, $N_2 = 100\,000$, shown in (**b**). Third, we simulated data with a higher proportion of causal variants, $p_{00} = 0.50$, $p_{10} = 0.20$, $p_{01} = 0.20$, $p_{11} = 0.10$. Using the same sets of heritabilities and sample sizes from the first two simulations, we tested the prediction accuracy of our model. (**c**) denotes the simulation with low sample size and high heritability, and (**d**) denotes the simulation with high sample size and low heritability. The dotted red lines denote the true proportion of causal SNPs in each simulation

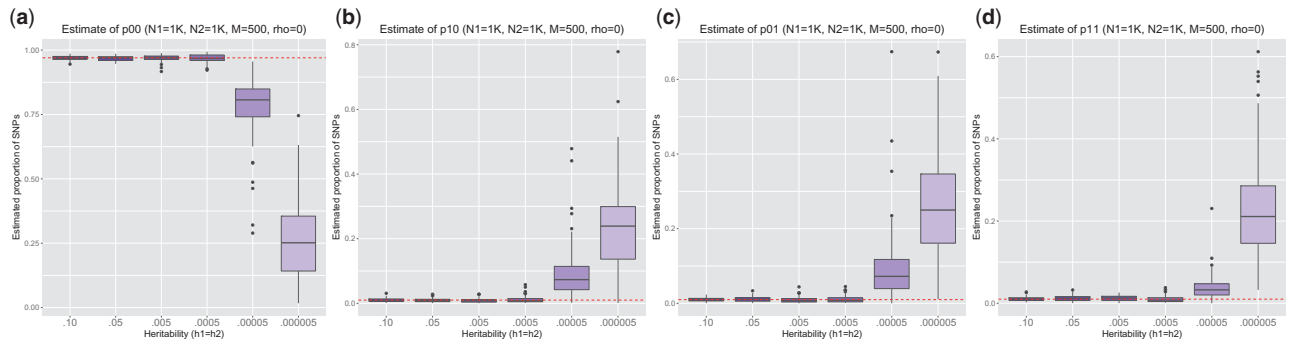

Fig. 2. We simulate the following proportion of causal variants $p_{00} = 0.97$, $p_{10} = 0.01$, $p_{01} = 0.01$, $p_{11} = 0.01$ and vary the heritability ($h_1^2 = h_2^2$) while fixing ρ, N_1, N_2, M. We vary the heritability from.01 to 5e−7 and plot the estimated proportion of non-causal variants (**a**), proportion of causal variants for trait 1 (**b**), proportion of causal variants for trait 2 (**c**) and proportion of shared causal variants (**d**). We note that as the heritability goes down, the data become less informative and the estimates reflect the prior

only causal for trait 1, $C_{1,m} = 0, C_{2,m} = 1$ if it was only causal for trait 2, and $C_{1,m}, C_{2,m} = 0$ if the SNP was non-causal. To get the true effect sizes, we multiplied element-wise $\boldsymbol{\beta}_1 = \boldsymbol{C}_1 \circ \boldsymbol{\gamma}_1$ and $\boldsymbol{\beta}_2 = \boldsymbol{C}_2 \circ \boldsymbol{\gamma}_2$ where we are essentially zeroing out any entry from the causal effect vector where a SNP is non-causal.

To compute the estimated GWAS effect sizes, $\widehat{\boldsymbol{\beta}}_p$, we assumed $\text{cov}(\epsilon_1, \epsilon_2) = 0$, so random noise terms ϵ_1, ϵ_2 were drawn from two normal distributions $\mathcal{N}\left(0, \frac{1-h_1^2}{N_1}\right)$ and $\mathcal{N}\left(0, \frac{1-h_2^2}{N_2}\right)$ respectively. We assume that the SNPs being used at the genome-wide level will be LD-pruned such that there is very little or no correlation structure. Thus, we set the LD matrix $V = I_M$, where I_M is an $M \times M$ identify matrix. We then draw the estimated effect sizes from a conditional distribution of the GWAS summary statistics, as described in Section 2.

First, we confirm that our method accurately predicts the proportion of causal variants under varying sample sizes and heritability estimates. We tested a variety of simulation frameworks where we fixed the genetic correlation and heritabilities of the two traits. We ran each simulation for 500 iterations and used the first quarter of the iterations as burn-in. We vary the proportion of causal variants contributing to only trait 1 (p_{10}), proportion of causal variants for only trait 2 (p_{01}), and the proportion of casual variants contributing to both traits (p_{11}). As shown in Figure 1, we can see that UNITY performs robustly across each scenario.

Next, to assess how UNITY performs with varying levels of heritability, we continued to fix $\rho = 0$, but varied the values of the heritability. Note that we used low heritability values due to the low number of simulated SNPs ($M = 500$). From Figure 2, we can see that the estimates reflect the prior distribution of $(p_{00}, p_{10}, p_{01}, p_{11})$ when the heritability is very low. We also show in Figure 3 that our estimates are invariant to the correlation between phenotypes.

To assess the role of sample size in our inference, we performed simulations where we varied the number of individuals from 1000 to 250 000. We find that the recommended sample size should be at least 50 000 individuals to yield precise results (Supplementary Fig. S2). Additionally, to further assess the performance of the method, we also performed simulations where $h_1^2 \neq h_2^2$ and when $p_{10} \neq p_{01}$. Through simulation, we demonstrate that our method is robust to these scenarios, with detailed results provided in Supplementary Figures S3 and S4.

Finally, through simulations, we empirically demonstrate that our method is well calibrated under the null hypothesis, defined as: (i) $p_{10} = 0$, (ii) $p_{01} = 0$ and (iii) $p_{11} = 0$. To demonstrate this, we simulated 100 000 SNPs with 100 000 individuals where $h_1^2 = 0.25, h_2^2 = 0.25, \rho = 0$. For each hypothesis, we set the parameter of interest exactly to 0 and then then simulated 2% causal variants between the remaining parameters. For example, for null hypothesis (1), the corresponding set of simulation parameters would be: $p_{10} = 0, p_{01} = 0.01, p_{11} = 0.01$. Using UNITY, we estimated the null parameter and report the posterior mean and standard deviation in Table 2. Note that UNITY estimates the null

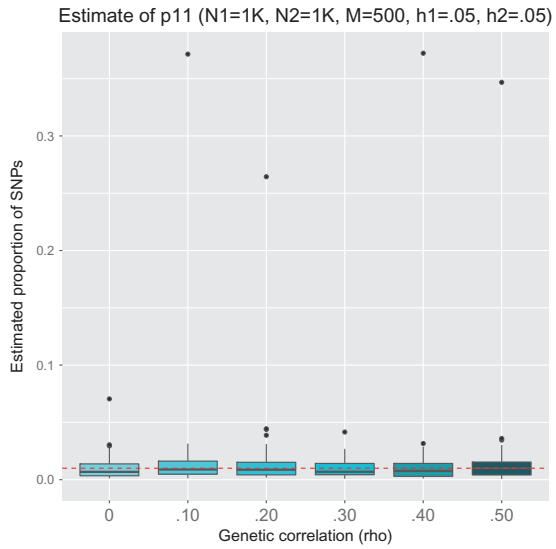

Fig. 3. We simulate the following proportion of causal variants $p_{00} = 0.97$, $p_{10} = 0.01$, $p_{01} = 0.01$, $p_{11} = 0.01$ and vary the genetic correlation from 0 to 0.50 while fixing $h_1^2, h_2^2, N_1, N_2, M$. We only show the estimate of p_{11}, since this would be the only estimate directly affected by the presence of genetic correlation

Table 2. We present the posterior means and standard deviations estimated when the proportion of causal variants is set exactly to zero for trait 1 and trait 2, and when the shared proportion is exactly zero

Hypothesis	Null parameter	Mean	SD
1	p_{10}	0.0006	0.0023
2	p_{01}	0.0004	0.0005
3	p_{11}	0.0002	0.0003

parameter very close to zero, but not exactly zero. This is because there is a non-zero prior on the set of parameters, making it not possible to be exactly zero, but can instead be asymptotically close.

3.2.1 LD-pruning to identify approximately independent SNPs

To rigorously assess the role of LD in our model, we demonstrate a sufficient LD-pruning scheme through simulations. To model a realistic LD structure, we used SNPs from 1000 Genomes (Consortium *et al.*, 2012a) to compute the LD for each of the approximately independent LD blocks identified in Berisa and Pickrell (2016). We filtered rare SNPs by minor allele frequency, MAF ≤ 0.05, and used 1 million SNPs sampled across the LD blocks. We chose only a subset of 1 million SNPs because this closely reflects the number of SNPs genotyped on SNP arrays. We simulated the GWAS effect sizes as outlined in Section 3.1, where the heritabilities for each of the each traits was set to $h_1^2 = 0.50$ and $h_2^2 = 0.50$ (which is similar to the estimated SNP heritability for height), and genetic correlation $\rho = 0$.

To assess the role of LD-pruning, we divided the genome into K kilobase non-overlapping windows and selected a SNP from each window. We varied K to assess the minimal window size necessary to create a subset of approximately independent SNPs. In addition, we used cross-trait LD Score regression to estimate the heritabilities for both traits and the genetic correlation after pruning, which were subsequently used in the inference. Through simulations, we determined that a 5 KB window provides precise estimates (Supplementary Table S1).

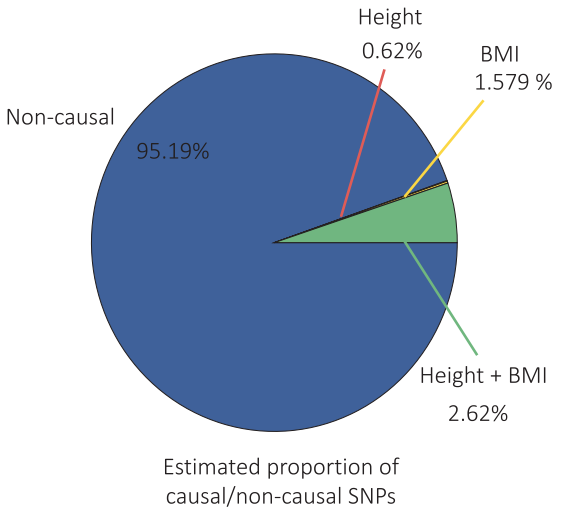

Fig. 4. We show the distribution of estimated non-causal and causal SNPs from the height and BMI analysis

3.3 Empirical analysis of BMI and height

We downloaded GWAS summary data for both height and BMI from the GIANT consortium (Allen *et al.*, 2010; Speliotes *et al.*, 2010) where each study has >170 000 individuals. First, we overlapped each GWAS by rsid to get SNPs present in both studies. Then for each trait, we filtered out SNPs with a minor allele frequency ≤ 0.05. Additionally, we performed LD-pruning by taking a SNP from every 5 KB window.

We used cross-trait LD Score to estimate the heritability and genetic correlation parameters: $h_H^2 = 0.2390$, $h_B^2 = 0.1566$, $\rho = -0.0845$. Denoting height as the first trait and BMI as the second, we estimated the proportion of causal variants for each trait as, $p_{00} = 0.9519$, $p_{10} = 0.0062$, $p_{01} = 0.01579$, $p_{11} = 0.0262$. We summarize the distribution of estimated causal SNPs in Figure 4.

Our results are consistent with the known genetic makeup of BMI and height. Since BMI is a function of an individual's height and weight, we expect all of the contributing variants for height to also contribute to BMI. UNITY predicts more BMI-only specific variants than height-only variants. We hypothesize that the BMI specific variants are those that contribute to weight, whereas the variants that contribute to height in the BMI dataset were already captured in the p_{11} estimate. In principal, we would expect p_{10} to be zero since SNPs contributing to height also contribute to BMI. We expect this could be due to the non-zero prior on p_{10}. Because of this, the estimate can never truly be zero but can be asymptotically close.

4 Discussion

In this work, we introduce a statistical framework for quantifying the relationship between two complex traits. The key advantage of our method is that it makes very few assumptions about the data and few restrictions during inference. Rather than relying on assumptions about a trait's genetic architecture, we let the data describe the underlying genetics. By using a Metropolis–Hastings sampling framework, we can calculate a variety of likelihoods without relying on any conjugate prior pairings. For example, although we

choose to model the causal effect sizes through a multivariate normal, one could choose another distribution, and the sampling procedure would still hold even if the new distribution did not have a conjugate prior. Finally, by operating exclusively on GWAS summary statistic data, we aim to encourage future large-scale meta analyses, since obtaining individual level data are not always readily available.

We conclude with several limitations and potential future directions of our framework. First, as the size of genetic datasets grow, subsampling methods such as MCMC may prove computationally intractable. Alternatives include using adaptive MCMC to accelerate mixing and convergence or variational methods that do not require subsampling. Additionally, we have yet to rigorously quantify the effects of LD in our model in practice for local inference. We leave rigorous comparison between UNITY and other relevant methods as future work.

Additionally, recent integrative methods have shown that the incorporation of a variants functional genomic context can improve both power and accuracy in identifying potential causal variants (Hu *et al.*, 2017b; Kichaev *et al.*, 2014; Li and Kellis, 2016; Pickrell, 2014). Large-scale initiatives such as the ENCODE (Consortium *et al.*, 2012b) and ROADMAP (Kundaje *et al.*, 2015) projects have provided comprehensive databases of tissue-specific functional genomic annotations. Combining this rich atlas of functional data and the genetic information from GWAS will likely uncover novel insights into disease biology. We leave the incorporation of functional elements as a potential direction for future work.

Acknowledgements

We are grateful to Kathryn Burch, Claudia Giambartolomei and Gleb Kichaev for helpful and insightful discussions.

Funding

This work was funded in part by National Institutes of Health (NIH), under awards R01HG009120, R01HG006399, U01CA194393, R00GM111744, R35GM125055, National Science Foundation Grant III-1705121, an Alfred P. Sloan Research Fellowship, and a gift from the Okawa Foundation.

Conflict of Interest: none declared.

References

Allen,H.L. *et al.* (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832.

Autism Spectrum Disorders Working Group of the Psychiatric Genomics Consortium, Anney,R.J. *et al.* (2017) Meta-analysis of gwas of over 16, 000 individuals with autism spectrum disorder highlights a novel locus at 10q24. 32 and a significant overlap with schizophrenia. *Mol. Autism*, **8**, 1–17.

Berisa,T. and Pickrell,J.K. (2016) Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, **32**, 283.

Brooks,S. *et al.* (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton, FL.

Bulik-Sullivan,B. *et al.* (2015) An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, **47**, 1236.

Byrd,R.H. *et al.* (1994). Representations of quasi-Newton matrices and their use in limited memory methods. *Math. Program.*, 63, 129–156.

1000 Genomes Project Consortium, Abecasis,G.R. *et al.* (2012) An integrated map of genetic variation from 1, 092 human genomes. *Nature*, **491**, 56.

Cotsapas,C. *et al.* (2011) Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.*, **7**, e1002254.

Ellinghaus,D. *et al.* (2012) Combined analysis of genome-wide association studies for crohn disease and psoriasis identifies seven shared susceptibility loci. *Am. J. Hum. Genet.*, **90**, 636–647.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57.

Giambartolomei,C. *et al.* (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, **10**, e1004383.

Hormozdiari,F. *et al.* (2016) Colocalization of gwas and eqtl signals detects target genes. *Am. J. Hum. Genet.*, **99**, 1245–1260.

Hu,Y. *et al.* (2017) Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genet.*, **13**, e1006836.

Hu,Y. *et al.* (2017) Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.*, **13**, e1005589.

Kundaje,A. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317.

Kichaev,G. *et al.* (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722.

Li,Y. and Kellis,M. (2016) Joint bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res.*, **44**, e144.

Metropolis,N. *et al.* (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Pasaniuc,B. and Price,A.L. (2017) Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.*, **18**, 117.

Pickrell,J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, **94**, 559–573.

Ramos,P.S. *et al.* (2011) A comprehensive analysis of shared loci between systemic lupus erythematosus (sle) and sixteen autoimmune diseases reveals limited genetic overlap. *PLoS Genet.*, **7**, e1002406.

Richard-Miceli,C. and Criswell,L.A. (2012) Emerging patterns of genetic overlap across autoimmune disorders. *Genome Med.*, **4**, 6.

Shi,H. *et al.* (2017) Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.*, **101**, 737–751.

Speliotes,E.K. *et al.* (2010) Association analyses of 249, 796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.*, **42**, 937.