Bioinformatics, 38, 2022, i212–i219
https://doi.org/10.1093/bioinformatics/btac228
ISCB/ISMB 2022

PolarMorphism enables discovery of shared genetic variants across multiple traits from GWAS summary statistics

Joanna von Berg [1,2], **Michelle ten Dam**[1], **Sander W. van der Laan** [3] and **Jeroen de Ridder** [1,2,*]

[1]Center for Molecular Medicine, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands, [2]Oncode Institute, 3521 AL Utrecht, The Netherlands and [3]Central Diagnostics Laboratory, Division Laboratories, Pharmacy, and Biomedical Genetics, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands

*To whom correspondence should be addressed.

Abstract

Motivation: Pleiotropic SNPs are associated with multiple traits. Such SNPs can help pinpoint biological processes with an effect on multiple traits or point to a shared etiology between traits. We present PolarMorphism, a new method for the identification of pleiotropic SNPs from genome-wide association studies (GWAS) summary statistics. PolarMorphism can be readily applied to more than two traits or whole trait domains. PolarMorphism makes use of the fact that trait-specific SNP effect sizes can be seen as Cartesian coordinates and can thus be converted to polar coordinates *r* (distance from the origin) and theta (angle with the Cartesian *x*-axis, in the case of two traits). *r* describes the overall effect of a SNP, while theta describes the extent to which a SNP is shared. *r* and theta are used to determine the significance of SNP sharedness, resulting in a *P*-value per SNP that can be used for further analysis.

Results: We apply PolarMorphism to a large collection of publicly available GWAS summary statistics enabling the construction of a pleiotropy network that shows the extent to which traits share SNPs. We show how PolarMorphism can be used to gain insight into relationships between traits and trait domains and contrast it with genetic correlation. Furthermore, pathway analysis of the newly discovered pleiotropic SNPs demonstrates that analysis of more than two traits simultaneously yields more biologically relevant results than the combined results of pairwise analysis of the same traits. Finally, we show that PolarMorphism is more efficient and more powerful than previously published methods.

Availability and implementation: code: https://github.com/UMCUGenetics/PolarMorphism, results: 10.5281/zenodo.5844193.

Contact: j.deridder-4@umcutrecht.nl

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

1 Introduction

Genetic variation in the genome partly explains phenotypic differences between individuals. Genome-wide association studies (GWAS) aim to identify the specific genetic variants [usually single nucleotide polymorphisms (SNPs)] that are associated with phenotypic variation. Over the past decades, GWAS have led to the discovery of thousands of SNP–trait associations (Buniello *et al.*, 2019; Visscher *et al.*, 2017).

From these discoveries we know that some SNPs can influence multiple traits; i.e. they are pleiotropic (Paaby and Rockman, 2013). Pleiotropy is widespread in the human genome. An association analysis between millions of SNPs and hundreds of traits found that almost ten percent of SNPs were associated with more than one trait (Watanabe *et al.*, 2019). Moreover, pleiotropic SNPs have been identified for many trait combinations. In many cases, the traits are known to be biologically related; pleiotropic SNPs have been identified for several psychiatric phenotypes (Otowa *et al.*, 2016) and different types of cancers (Graff *et al.*, 2021). However, pleiotropic SNPs have also been described for seemingly unrelated diseases; for instance for prostate cancer and type 2 diabetes (Ray and Chatterjee, 2020), schizophrenia and Human Immunodeficiency Virus (HIV) infection (Wang *et al.*, 2017) and Alzheimer's disease and lung cancer (Feng *et al.*, 2017). This could mean that those SNPs are involved in a biological process with a more general function. It could also mean that the studied traits are more biologically related than was previously known and might have a common etiology. Identifying more pleiotropic SNPs can thus transform our current classification of diseases (Sivakumaran *et al.*, 2011).

Pleiotropy analysis can also be useful to identify pleiotropic SNPs in druggable genetic targets, which can help predict adverse treatment effects (Sivakumaran *et al.*, 2011) as well as identify diseases that could be treated with existing drugs (O'Mara *et al.*, 2019). Moreover, pleiotropy can be leveraged for more accurate risk prediction (Maier *et al.*, 2015). Finally, methods like Mendelian Randomization (MR) rely on the assumption that there is no direct effect of the SNPs used on both exposure and outcome (Hemani *et al.*, 2018). Since pleiotropy methods can be used to indicate whether some SNPs are pleiotropic, they can be used to filter these SNPs.

It should be noted that analysis of similarity between traits can also be done using genetic correlation, but this answers a different question. Genetic correlation gives the overall—genome-wide—correlation of effect sizes. Pleiotropic SNPs have a shared effect regardless of the genetic correlation and may tag a specific biological pathway or process rather than describing a general relationship between two traits. If traits are correlated and often co-occur in individuals, then any SNP that affects trait X will also be associated with trait Y, even if it does not directly affect trait Y. These SNPs are not actually pleiotropic because they are only directly associated with one trait. For this reason, to identify pleiotropic SNPs it is not sufficient to take the intersection of SNPs that are associated with both traits. Even if the traits are uncorrelated, intersecting SNP-sets is not an optimal approach; both GWASs need to be sufficiently powered to discover the pleiotropic SNP. Moreover, SNPs that are found with this approach lack an important feature: we know that they are shared but we do not know how shared they are and if this might be statistically significant.

Recently, a few methods that aim to identify pleiotropic SNPs have been described. HOPS (Jordan *et al.*, 2019) and PLEIO (Lee *et al.*, 2021) both identify a SNP as shared if it is associated with at least one of the traits of interest. Problematically, SNPs with an effect on only one trait will thus also be identified and cannot readily be differentiated from truly pleiotropic SNPs. Two other methods, PLACO (Ray and Chatterjee, 2020) and PRIMO (Gleason *et al.*, 2020), identify a SNP as shared if it is associated with all traits of interest. PLACO can only be used for identification of SNPs that are shared by two traits. Moreover, we will show that PLACO has a high computational burden. PRIMO, on the other hand, only identifies a subset of the pleiotropic SNPs that PLACO finds.

Here, we present PolarMorphism, a new approach to identify pleiotropic SNPs that is more efficient, identifies the same number of pleiotropic SNPs as PLACO, but can be applied to more than two traits. This enables the identification of SNPs that have an effect on numerous traits, and possibly play a role in more general biological processes. PolarMorphism is based on a transformation of the trait-specific effect sizes x and y to polar coordinates r (*radius*, the distance from the origin) and $\boldsymbol{\theta}$ (*theta*, the angle with the x-axis). As a result, r is a measure of overall effect and $\boldsymbol{\theta}$ a measure of sharedness, which can be used for downstream significance analysis and SNP ranking.

PolarMorphism enables construction of a trait network showing which traits share SNPs. From SNP-specific networks we observe that most SNPs are associated with traits within one trait domain. We find one SNP—rs495828 in the ABO locus—that is associated with traits across seven trait domains. We show that analysis of more than two traits is more powerful than the intersection of pairwise results of those same traits. We provide PolarMorphism as an R package on Github under the MIT license: https://github.com/UMCUGenetics/PolarMorphism.

2 Materials and methods

2.1 Overview of PolarMorphism

We aim to identify pleiotropic SNPs from GWAS summary statistics using an approach that can be routinely applied to combinations of two or more traits. After obtaining summary statistics with effect size beta and standard error SE, we calculate z-scores (beta/SE) per SNP. PolarMorphism can be applied on any number of traits, but here we explain the application to two traits. Analyzing more than two traits requires a slightly different approach (see the Section 2 for a full description) but leverages the same principles.

Our aim is to identify horizontally pleiotropic SNPs. Therefore we first perform a decorrelating transform to attenuate vertical pleiotropy resulting from genetic correlation. Given summary statistics for trait x and y, we calculate a covariance matrix, and use this to apply decorrelation or whitening (see methods for details) yielding decorrelated summary statistic vectors $\overrightarrow{x}$ and $\overrightarrow{y}$. Next the trait-specific vectors $\overrightarrow{x}$ and $\overrightarrow{y}$ are used to calculate polar coordinates r_i (the distance from the origin) and θ_i (the angle with the x-axis, ranging from 0 to 2π). For SNPs that are specific to trait X, θ_i is close to 0 or π. For SNPs that are specific to trait Y, θ_i is close to $\frac{1}{2}\pi$ or $1\frac{1}{2}\pi$. For SNPs that are shared, θ_i is approximately $\frac{1}{4}\pi$ or $1\frac{1}{4}\pi$ for concordant direction of effect and $\frac{3}{4}\pi$ or $1\frac{3}{4}\pi$ for opposite direction of effect. Each quadrant of the x, y plot only differs in direction of effect in the original GWAS. To simplify further analysis we use the fourfold transform of θ (θ_{trans}), which folds the quadrants on top of each other (equivalent to using the absolute values of the z-scores) and then stretches the angles so they still describe a full circle (Fig. 1).

To assess significance of sharedness, we separately test the distance r_i and angle θ_i. Under a null hypothesis of no overall effect, r_i is the square root of a sum of squared normally distributed variables with mean 0. We thus use a central χ distribution to calculate P-values for r_i (equivalent to using a χ^2 distribution to test r_i^2). The alternative hypothesis of this test is that SNP i affects at least one of the traits, which is insufficient to determine pleiotropy. Under a null hypothesis of trait-specific effect, $\theta_{\mathrm{trans},i}$ is equal to 0. To calculate P-values for $\theta_{\mathrm{trans},i}$ we use a von Mises distribution with concentration parameter κ_i. We show that κ_i depends on r_i (see Supplementary Methods). Estimates of κ_i from simulations under the null hypothesis are included in the R package. These are used to establish one P-value per SNP. The alternative hypothesis of the second test is that SNP i has a pleiotropic rather than a trait-specific effect.

2.2 PolarMorphism for two traits

PolarMorphism works on uncorrelated, standardized data. z_x and z_y are vectors of length m containing the z-scores of SNPs 1 to m for trait x and trait y, respectively. We calculate polar coordinates r and θ per SNP i: r is the distance from the origin, and θ is the angle of the vector from the origin to the point ($z_{x,i}$, $z_{y,i}$).

$$r_i = \sqrt{{z_{x,i}}^2 + {z_{y,i}}^2} \text{ and } \theta_i = \tan^{-1}\left(\frac{z_{y,i}}{z_{x,i}}\right).$$

We first test whether r comes from a central chi distribution with degrees of freedom equal to the number of traits p. The chi distribution describes the distribution of the square root of the sum of squared normally distributed variables. The distribution of P-values from this test is used to calculate q-values, which are FDR-corrected P-values (Storey, 2003). For all SNPs that have an effect, we want to know whether that effect is shared. We perform a four-fold transform of θ that 'folds' all quadrants of the Cartesian plane on top of each other and stretches it to make sure the angles can take any value on the circle (Landler *et al.*, 2018): $\theta_{\mathrm{trans}} = 4\theta$ modulo 2π. The von Mises distribution describes angular data. It takes into account that $\theta = 0$ is equal to $\theta = 2\pi$. It has two parameters: θ_{mu} is the mean value, and kappa (κ) is a concentration parameter that is similar to the inverse of the variance. θ_{mu} is zero under the null hypothesis of trait-specific effect. See the Supplementary Methods for a description of how we obtained estimates for κ. Using the distribution of the observed r P-values for the distances of all SNPs, and the fact

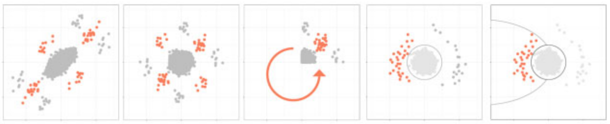

Fig. 1. Overview of the method for two traits. Orange indicates true pleiotropic SNPs, gray indicates SNPs that are either trait-specific or do not have any effect. Z-scores for each trait are plotted on each axis and the data are decorrelated. Cartesian coordinates are transformed to polar coordinates. The absolute values of the z-scores are calculated, and the angle is multiplied by four. After subsetting on SNPs with a significant distance, we calculate P-values for the angle

that P-values follow a uniform distribution under the null hypothesis, the false discovery rate (FDR) for each SNP can be calculated. This q-value gives the FDR if this SNP and all SNPs with a lower P-value would be called significant. We keep the SNPs that show a significant overall effect (r q-value < 0.05) and use the distribution of observed θ P-values for these SNPs to calculate θ q-values. We filter on θ q-value < 0.05 to obtain SNPs that are significantly shared (FDR < 0.05).

2.3 PolarMorphism for more than two traits

The distance of a SNP i in more than two dimensions is a straightforward extension of the distance in two dimensions:

$$r_i = \sqrt{\sum_{j=1}^{p} z_{i,j}^2},$$

where $z_{i,\,j}$ is the z-score of SNP i for trait j. Describing the orientation of a SNP for p traits involves calculating the corresponding p-dimensional hyperspherical coordinates. This gives an additional angle for each added trait. Fortunately, this problem can be simplified. We define $\overrightarrow{X_i}$ as the vector from the origin of the p-dimensional sphere to an observed SNP, and $\overrightarrow{\mu}$ as the vector from the origin to the expected position of the SNP under the null hypothesis of trait-specific effect, along one of the axes. The goal is to determine the angular difference between $\overrightarrow{X_i}$ and $\overrightarrow{\mu}$. We choose $\overrightarrow{\mu}$ such that it lies along the axis that is closest to $\overrightarrow{X_i}$. In other words, we construct $\overrightarrow{\mu}$ as a vector with zeros for each coordinate except for the coordinate with the highest absolute value for the SNP under consideration. We set the length of $\overrightarrow{\mu}$ equal to the length of $\overrightarrow{X_i}$ (the distance r), so the only non-zero value of $\overrightarrow{\mu}$ is set to r. The two vectors of interest always lie in a 2D plane, regardless of the number of traits p. The dot product of the vectors is a scalar and is equal to:

$$\overrightarrow{\mu} \cdot \overrightarrow{X_i} \;=\; r_\mu r_x \cos(\theta)$$

therefore

$$\theta = \cos^{-1}(\overrightarrow{\mu} \cdot \overrightarrow{X_i} \,/\, r^2)$$

which can be rewritten as

$$\theta = \cos^{-1}\left(\left(\sum\nolimits_{j=1}^{p} \mu_j x_j\right) / \left(\sum\nolimits_{j=1}^{p} x_j^2\right)\right).$$

This angle should be normalized so the maximum value is always π, regardless of p. The angle is maximal if all coordinates of a SNP have the same value (which we will call x). Recall that $\overrightarrow{\mu}$ has zeros for all coordinates but one. If θ is maximal, we can rewrite the expression for θ as:

$$\theta(p) = \cos^{-1}\left(\frac{(\sum_{j=1}^{p} \mu_j x)}{} \left(\sum_{j=1}^{p} x^2\right)\right)$$
$$= \cos^{-1}(((p-1)(0 \cdot x) + r \cdot x)/px^2) = \cos^{-1}\left(\frac{\sqrt{p}}{p}\right)$$

The final correction factor with which the angles should be multiplied can then be obtained by dividing 2π by the result of this formula.

To test the significance of r, we use the same procedure as for two traits. In this case the degrees of freedom is equal to the number of traits p. To assign significance levels to the angle θ, we use the von Mises–Fisher distribution, which is an extension of the von Mises distribution. The probability density function of the von Mises Fisher distribution is given by:

$$f \;=\; C\exp(\kappa \overrightarrow{\mu} \cdot \overrightarrow{X}),$$

where C is a normalization constant, κ is the concentration parameter, $\overrightarrow{\mu}$ is the unit vector of the expected direction and $\overrightarrow{X}$ is the observed unit vector (i.e. the vector of the SNP divided by its length to get unit length). The inner product $\overrightarrow{\mu} \cdot \overrightarrow{X}$ can be rewritten as $\cos(\theta)$, where θ is the angle between the expected and observed vectors:

$$f \;=\; C\exp(\kappa\cos(\theta)).$$

Functions to obtain the probability density function and the normalization constant C are implemented in the vMF package in R (Wood, 1994). To obtain a cumulative density function the probability density function needs to be integrated. The definite integral for $\exp(\kappa\cos(\theta))$ can not be defined using elementary functions. However, the exponent has the following series representation:

$$f \;=\; C\exp(\kappa\cos(\theta)) \;= C\sum\nolimits_{j\,=\,0}^{\infty} \frac{(k\cos(\theta))^j}{j!}\;.$$

The integral is then equal to:

$$F \;=\; C\int \sum\nolimits_{j\,=\,0}^{\infty} \frac{(k\cos(\theta))^j}{j!} \;= C\;\sum\nolimits_{j\,=\,0}^{\infty} \int \frac{(k\cos(\theta))^j}{j!}$$

The term (as a function of the iterator j) does have an indefinite integral:

$$\int \frac{(k\cos(\theta))^j}{j!}$$
$$= -\frac{\cot(\theta)\ \mathrm{abs}(\sin(\theta))(k\cos(\theta))^j\ \mathrm{hypergeo}\left(\frac{1}{2},\ \frac{j+1}{2},\ \frac{j+3}{2},\ \cos^2(\theta)\right)}{\mathrm{gamma}(j+2)},$$

where cot is the cotangent function, *hypergeo* is the hypergeometric function and gamma is the gamma function. We implemented the summation so that it stops when the last added term is smaller than a user-defined value (called 'tol' in our R package). We use the hypergeo package for the hypergeometric function (Hankin, 2015). The values for κ as a function of p that we derived for $p=2$ still apply here, because θ still describes a 2D angle.

2.4 Simulated data generation

To estimate the false positive rate (FPR) of PolarMorphism we used the R package simplePHENOTYPES (Fernandes and Lipka, 2020) to simulate GWAS data for two traits with horizontally pleiotropic SNPs and SNPs that are specific to each of the traits (49 317 SNPs for each of the three categories, approximately 10% of the total number of SNPs), a genetic correlation of 0.8 and heritability of 0.6 for each trait. This was repeated 100 times. As input to the package we used genetic data from the HD genotype chip from phase 3 of the 1000 genomes dataset (1000 Genomes Project Consortium *et al.*, 2015). We included only individuals with non-Finnish European ancestry to keep the linkage disequilibrium (LD) as homogeneous as possible while maintaining a decent sample size ($N=549$ individuals). We used bcftools (Li *et al.*, 2009) to include these samples and variants with allele frequency higher than 0.05 or lower than 0.95. We further filtered the variants to include only high-confidence SNPs, using the list of SNPs with pre-computed LD-scores from the LD-score method (Bulik-Sullivan *et al.*, 2015). The output of simplePHENOTYPES can readily be used as input for BOLT-LMM (Loh *et al.*, 2015), with which we performed a GWAS of each instance of simulated data. The resulting summary statistics were used as input for PolarMorphism. To determine FPR for the angle θ, we considered the fraction of ground-truth trait specific SNPs in our simulated data with $p_\theta < 0.05$, as these SNPs would (falsely) be considered pleiotropic in our method.

To estimate the FPR of the distance r, we permuted the phenotypes as pairs. This ensures that the correlation between the traits remains but no association between genotype and phenotype should exist beyond what is expected under the null hypothesis of no effect. Each of the 100 instances of simulated data was permuted once. We again performed GWAS in BOLT-LMM and ran PolarMorphism. To determine FPR for the significance threshold on r we determine the fraction of all SNPs with $p_r < 0.05$, as these SNPs would (falsely) be considered SNPs with a—pleiotropic or trait-specific—effect.

The mean estimated FPR$_\theta$ on the non-permuted data is 0.060 (SD = 0.001). On the permuted data, the mean estimated FPR$_r$ is 0.050 (SD = 0.0003) and the mean estimated FPR$_\theta$ is 0.060 (SD = 0.001). Boxplots of the distribution of both FPRs can be found in Supplementary Figure S1.

2.5 Preprocessing the summary statistics

We used publicly available summary statistics for the 41 traits shown in Table 1, encompassing a range of mostly cardiovascular phenotypes with relatively large sample sizes enabling biological interpretation of pleiotropic SNPs within a specific disease context. Data were obtained from the sources provided in Supplementary Table S2, which also contains references to the respective papers they were described in. We aligned reference and alternative allele across all traits, and filtered using the list of high-confidence SNPs provided with the LDSC software (Bulik-Sullivan *et al.*, 2015). We divide effect sizes by their standard error to obtain z-scores. We calculate the covariance matrix on the subset of SNPs that do not have a large overall effect. To this end, the covariance is calculated only on SNPs that have a mahalanobis distance smaller than five. We use the ZCA-cor whitening method in the 'whitening' package in R (Kessy *et al.*, 2018), to decorrelate the data while ensuring that the x and y components of the transformed z-scores maximally correlate with the x and y components of the original z-scores.

2.6 Inferring relationships between traits from pleiotropic SNPs

For all trait pairs, we ran PolarMorphism and clumped the significant SNPs with Plink, using the q-values instead of P-values (–clump-kb 5000000, –clump-p1 0.05, –clump-p2 0.05 and –clump-r2 0.2) (Purcell *et al.*, 2007). We make an adjacency matrix from the number of shared loci per trait combination and use this to construct a graph using the igraph package in R (Csárdi *et al.*, 2016). We did the same per SNP to obtain SNP-specific networks. To create domain networks from the trait networks we draw an edge between domain A and B if an edge exists between any trait of domain A and any trait of domain B.

2.7 Gene set enrichment analysis in DEPICT

We changed the following settings from the default: association_p-value_cutoff: 0.05 to accommodate for the fact that we use q-values instead of P-values. We performed gene set enrichment using the default gene sets provided by the DEPICT authors, but only considered gene sets from gene ontology (Harris *et al.*, 2004), REACTOME (Fabregat *et al.*, 2018), KEGG (Kanehisa and Goto, 2000) and the PPI networks as defined by the DEPICT authors using the InWeb database (Lage *et al.*, 2007) for further analysis.

2.8 Inferring relationships between traits from genetic correlation

To infer relationships between traits from genetic correlation, we ran LDSC (Bulik-Sullivan *et al.*, 2015) using the GenomicSEM (Grotzinger *et al.*, 2019) package in R. We calculated P-values from the correlation coefficients and their standard errors using the pnorm function in R, and used a Bonferroni corrected P-value threshold of 6.4×10^{-5} to correct for 780 trait combinations tested. For this purpose, we made an adjacency matrix from the genetic correlation for each trait combination and used this to make a graph using the igraph package in R (Csárdi *et al.*, 2016).

2.9 Comparison with other methods

Intersection refers to the straight-forward approach of finding shared SNPs: take the intersection of the SNPs that were significant for trait X and those that were significant for trait Y. We used the R package for HOPS (HOrizontal Pleiotropy Score) (Jordan *et al.*, 2019). We used our pre-processed z-scores (whitened). We ran HOPS both with and without polygenicity correction and used only the Pm P-values. We used the command line tool written in Python for PLEIO (Pleiotropic Locus Exploration and Interpretation using Optimal test) (Lee *et al.*, 2021). We used z-scores (not whitened and not corrected for LD-score) and supplied the sample sizes of the original GWAS. We used the R package for PRIMO (Package in R for Integrative Multi-Omics association analysis) (Gleason *et al.*, 2020). We used PRIMO based on P-values. For the alt_props parameter (the expected proportion of SNPs that follow the alternative hypothesis per trait) we supplied the proportion of SNPs that were significant for trait 1 (q-value < 0.05) over all SNPs, idem for trait 2 (q-value < 0.05). We supplied c(2,2) for the dfs parameter. We used the R package for PLEIO (pleiotropic analysis under composite null hypothesis) (Lee *et al.*, 2021). We used whitened z-scores (not corrected for LD-score). We used the VarZ function to calculate the covariance matrix and supplied that, with the z-scores, to the placo function.

Table 1. Trait domains and trait abbreviation as used in the figures

Domain name	Trait abbreviation	Trait name
Anthropomorphic	BMI	Body mass index
	Height	Height
Cancers	PrCa	Prostate cancer
	BC	Breast cancer
Cardiac traits	AF	Atrial fibrillation
	HF	Heart failure
	NICM	Non-ischemic cardiomyopathy
Cardiovascular	CAC	Coronary artery calcification
	CAD	Coronary artery disease
	cIMT	Carotid intima-media thickness
	Plaque	Presence of carotid plaque
Immune	IBD	Irritable bowel disease
	Asthma	Asthma
Lipids	HDL	High-density lipoprotein
	LDL	Low-density lipoprotein
	TC	Triglycerides
	TG	total cholesterol
Neurodegenerative disease	AD	Alzheimer's disease
	ALS	Amyotrophic lateral sclerosis
	PD	Parkinson's disease
Pressures	DBP	Diastolic blood pressure
	SBP	Systolic blood pressure
	PP	Pulse pressure
Psychiatric/ psychological	ASD	Autism spectrum disorder
	BIP	Bipolar disorder
	DS	Depressive symptoms
	EA	Educational attainment
	IQ	Intelligence quotient
	MDD	Major depressive disorder
	Neuroticism	Neuroticism
	SWB	Subjective well being
	Insomnia	Insomnia
Smoking	EvrSmk	Ever smoker
	FrmrSmk	Former smoker
	logOnset	Log of age at onset of smoking
	CpD	Cigarettes per day
Stroke	AS	Any stroke (hemorrhagic or ischemic)
	IS	Ischemic stroke
	CES	Cardio-embolic stroke
	LAS	Large artery stroke
	SVS	Small vessel stroke

To assess how many loci were found by each method, we LD-pruned the significantly shared SNPs. For each method and for each locus, we checked if any of the SNPs in that locus were also found by another method. If that was the case, we gave that locus the same identifier in each method. Afterwards, we determined the loci that were found by all methods and those that were found by only one or a subset of the methods. We ran Intersection, HOPS (with polyenicity correction), PRIMO, PLACO and PolarMorphism on the same data while supplying a dataframe with an increasing number of rows. For the Intersection method, we added q-value calculation from the original GWAS P-values and a filtering step on both q-values to make it a fair comparison with the other methods. All five methods are written in R, therefore we timed them in R using the tictoc package (Izrailev, 2014). Running the software in the terminal could have a different runtime, but this does allow us to compare the runtimes among the methods.

3 Results

3.1 Defining pleiotropy

Pleiotropy can be identified in different ways (Paaby and Rockman, 2013; Tyler *et al.*, 2009; Fig. 2). Horizontal pleiotropic SNPs directly affect multiple traits. Vertical (or mediated) pleiotropic SNPs directly affect one of the traits, but dependence between the traits leads to an association with both traits. The difference between horizontal and vertical pleiotropy is particularly important in the context of Mendelian randomization (MR). With MR, the causal effect of an exposure (e.g. smoking) on an outcome (e.g. lung cancer) can be determined. Genetic variants that are associated with the exposure are used as so-called ‘instrumental variables’. One important assumption is that these variants only have an effect on the outcome through the exposure. In other words, that they are vertically pleiotropic and not horizontally. Horizontally pleiotropic SNPs—which have a direct effect on both smoking and lung cancer—violate this assumption and should therefore not be used as instrumental variables in MR (Burgess *et al.*, 2019). The final pleiotropy type is spurious pleiotropy, which can arise from bias in measuring association (van Rheenen *et al.*, 2019). For example, one marker SNP can be associated with two or more traits due to that marker being in linkage disequilibrium (LD) with another SNP that directly affects one of the traits and yet another SNP that directly affects another trait. The marker SNP seems to be pleiotropic, while in reality neither the marker SNP nor the nearby linked SNPs are pleiotropic. Determining whether the same SNP is likely causal for both traits is only possible with colocalization approaches (Wallace, 2020). Another source of spurious pleiotropy is misclassification of traits. If certain symptoms are shared by two diagnoses, individuals with these overlapping symptoms can be given either diagnosis. As a result, the genetic associations for these diagnoses will be highly correlated. Finally, shared controls and ascertainment bias (participant recruitment in a specific disease field) can also cause spurious pleiotropy (Solovieff *et al.*, 2013).

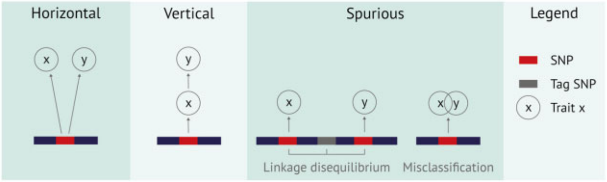

Fig. 2. Visualization of horizontal, vertical and spurious pleiotropy, respectively. A horizontally pleiotropic SNP has an effect on all traits under consideration. A vertically pleiotropic SNP has an effect on only one of the traits, but because the traits are correlated it is also associated with the other trait. A SNP can seem pleiotropic because it is in linkage disequilibrium with two SNPs that each individually have an effect on a trait. Misclassification of individuals can also give rise to a seemingly pleiotropic effect

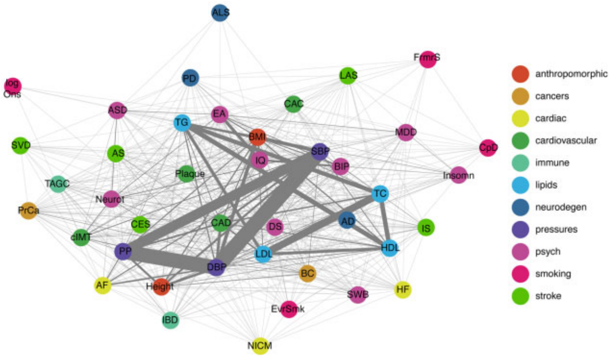

Fig. 3. Trait network based on pleiotropic SNPs. Pairwise PolarMorphism results for 41 traits. Pleiotropic SNPs were defined as having an r q-value > 0.05 and a theta q-value > 0.05. Clumping was performed based on theta q-values and linkage disequilibrium. See methods for details. The thickness of the lines (network edges) indicates how many loci are shared between two traits (network nodes). Colored by disease domain

3.2 Inferring relationships between traits from pleiotropic SNPs

We applied PolarMorphism to all pairwise combinations of 41 traits from different trait domains (Table 1). The resulting pleiotropy network is shown in Figure 3. Herein, traits are nodes and the edge weights indicate the number of pleiotropic SNPs discovered by PolarMorphism. The contribution of each SNP to the edge weights is weighted by the inverse of the total number of traits it is associated with, in order to account for the effect that SNPs affecting many traits probably tag a biological process with a general function. Sharing such a SNP is less meaningful than sharing a SNP with an effect on only some traits.

The resulting pleiotropy network is densely connected (512 out of 820 possible edges), supporting earlier descriptions of widely occurring pleiotropy among traits (Solovieff *et al.*, 2013; Watanabe *et al.*, 2019). The lipid domain (HDL, LDL, TG and TC) and blood pressure domain (DPB, SBP and PP) each form a fully connected subgraph. SBP has the highest number of edges (degree), sharing SNPs with 37 of the 41 traits. ALS, which shares SNPs with five traits, has the lowest degree. Global analysis of the pleiotropy network thus readily reveals general characteristics of traits and trait domains.

Analyzing the pleiotropy network in more detail, we find that most SNPs are associated with traits within one or across two trait domains (51% and 43%, respectively). We observe one SNP that is associated with traits across seven trait domains: rs495828, a SNP in the ABO gene, which is ubiquitously expressed across many tissues and cell types (Carithers *et al.*, 2015). For each trait domain, we determine how many SNPs only have associations within that domain (we call these single domain SNPs), and calculate the percentage of the total number of SNPs that were identified for that domain. We find that the psychiatric traits have the highest percentage of single domain SNPs; one third of all SNPs that are shared with a psychiatric trait are only associated with psychiatric traits. The smoking traits have the lowest percentage of single domain SNPs, suggesting that most smoking-associated variants tag general biological processes rather than smoking-specific processes.

3.3 A Comparison with genetic correlation

Genetic correlation (r_g) is the correlation of SNP effect sizes on two traits (van Rheenen *et al.*, 2019). Non-biological factors like sample overlap between the two GWAS can inflate the r_g estimate. LDSC (Bulik-Sullivan *et al.*, 2015) or HDL (Ning *et al.*, 2020) can be used to obtain an r_g estimate that is not biased by sample overlap. Genetic correlation leads to overall correlation of effect sizes, also in those SNPs with no effect on any of the traits. SNPs that do have an effect can influence r_g estimates; if they are very pleiotropic they can inflate r_g and if they are very trait-specific they can deflate r_g. Therefore it is generally recommended to only use the subset of SNPs with no effect on any of the traits for r_g estimation. Also note that pleiotropic effects between traits can be present without genetic correlation, as pleiotropy is a SNP-specific metric and genetic correlation is a genome-wide metric (Bulik-Sullivan *et al.*, 2015).

To assess whether genetic correlation provides the same insight into trait relationships as pleiotropy, we built a network based on genetic correlation. The resulting network is sparse (138 out of 780 possible edges) and only partially overlaps with the pleiotropy network. Figure 4 shows separate subnetworks for edges that exist in both the genetic correlation network and the pleiotropy network or in only one of the two. In total, 416 trait pairs share at least one pleiotropic SNP, but are not genetically correlated (Fig. 4A). This situation can arise if there are only a few SNPs that are shared but the rest of the genetic architecture of the traits is independent. It is also possible that some shared SNPs have the same direction of effect in both traits while other shared SNPs have an opposite direction of effect, thereby averaging out r_g Seven trait pairs are genetically correlated, but do not share any SNPs that are horizontally pleiotropic (Fig. 4B). Each SNP that is associated with one of the traits is more likely to also be associated with the other, because of the overall r_g (Burgess *et al.*, 2019). After decorrelation, vertically pleiotropic SNPs will not be identified by PolarMorphism. 96 trait pairs are genetically correlated and share horizontally pleiotropic SNPs (Fig. 4C). These are traits that share a number of vertically pleiotropic SNPs, leading to a higher r_g, as well as some horizontally pleiotropic SNPs. Our results seem to indicate that two traits are more likely to share at least one pleiotropic SNP than they are to be genetically correlated.

3.4 The stroke domain

The stroke domain consists of any stroke (AS); its subtype ischemic stroke (IS); and its subtypes cardioembolic stroke (CES), large artery stroke (LAS) and small vessel stroke (SVS). The three IS subtypes are generally believed to have different etiologies (Ay *et al.*, 2007; Malik and Dichgans, 2018; Pulit *et al.*, 2018), and previous efforts have resulted in tens of subtype-specific associations (Dichgans *et al.*, 2019; Malik *et al.*, 2018; Traylor *et al.*, 2014, 2017). In line with this, our analysis does not reveal any shared SNPs. It should be noted that shared SNPs have been described before for LAS and SVS and for LAS and CES (Malik *et al.*, 2018). However, SNPs at these loci were low-confidence and therefore not included in our analysis (see methods for details).

Given the lack of shared SNPs among the IS subtypes, we investigated which other traits share SNPs with each of the IS subtypes. To that end we looked at the subnetwork composed of the IS subtypes and their direct neighbors (Fig. 5). Our analysis reveals that six traits (CAD, DBP, Plaque, PP, SBP, TC) share SNPs with all IS subtypes. This indicates that all ischemic stroke subtypes are associated with biological pathways with a possible effect on blood pressure and lipids. CES shares most pleiotropic SNPs with atrial fibrillation (AF), which is believed to be its main cause (Pulit *et al.*, 2018). LAS, which is thought to arise from atherosclerotic plaques in the carotid arteries that rupture or block blood flow (Dichgans *et al.*, 2019), shares most SNPs with cIMT—a proxy for the extent of carotid atherosclerosis. SVS, which is thought to have a cardiovascular origin like the other IS subtypes (Lee, 2020), shares most SNPs with CAD. Notably, it also shares many SNPs with Alzheimer's and Parkinson's disease. This might indicate that many of the SNPs that are associated with risk of small vessel stroke also influence risk of neurodegenerative disease. Note that the edges LAS-HDL, SVS-AD, SVS-PD and SVS-Plaque were only found in the pleiotropy network and not in the genetic correlation network. This indicates that pleiotropic SNPs harbor information that is complementary to genome-wide correlation measures. Furthermore, zooming in on one trait domain shows how PolarMorphism can be employed to gain more detailed insight in trait relationships than the general patterns that can be gathered from the complete network.

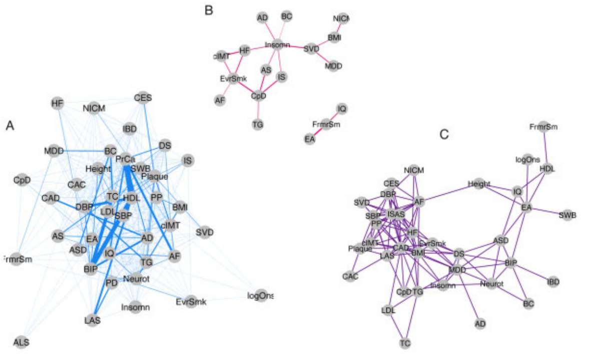

Fig. 4. (**A**) Edges denote trait pairs that share pleiotropic SNPs but are not genetically correlated. (**B**) Edges denote trait pairs that are genetically correlated but do not share pleiotropic SNPs. (**C**) Edges denote trait pairs that are genetically correlated and share pleiotropic SNPs

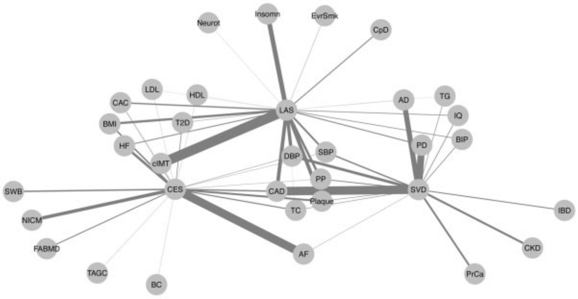

Fig. 5. Trait network of the IS subtypes and their direct neighbors, based on the weighted full network as described earlier. Only edges between any of the IS subtypes and any other trait are drawn; in other words, edges between two nodes shown here that do not include an IS subtype, are not drawn

3.5 Joint analysis of more than two traits identifies more pleiotropic SNPs than pairwise analyses of the same traits

PolarMorphism can be used to find SNPs that are shared by any number of traits. A SNP with a small effect on each trait might not be identified in univariate or even pairwise analysis, but could be if more traits are included. We therefore investigated whether analysis of three or more traits is indeed more powerful than the combined results from pairwise analyses of those same traits. Pairwise analyses of the lipid domain (HDL, LDL, TC, TG) identifies 186 shared loci. Analysis of all four traits together identifies 1029 shared loci. 180 loci are found by both approaches.

To explore whether the increased number of loci is biologically relevant, we perform gene set enrichment analysis in DEPICT (Pers *et al.*, 2015) on the significant loci from the pairwise analyses and the significant loci from the joint analysis. In order to get the relevant genes for each locus, we perform clumping using DEPICT's default settings. Hence the number of DEPICT loci differs from the loci that we identified (108 pairwise loci, 496 joint loci, see Supplementary Tables S4 and S6). The pairwise results are enriched for 12 gene sets (Supplementary Table S5) whereas the joint results are enriched for 85 gene sets (Supplementary Table S7). Moreover, the loci revealed by the joint analysis result in enrichments that are more significant: 85 of the 95 gene sets that are significant in either analysis are more significant in the joint analysis, and 2 of the 2 gene sets that are significant in both analyses are more significant in the joint analysis. Moreover, considering the 10 genes with the highest z-score for membership of these gene sets, we find that the genes implied by the joint analysis have a higher likelihood of gene set membership (see the DEPICT paper for a detailed explanation; Pers *et al.*, 2015), thus resulting in more coherent gene sets. For instance, the joint analysis identifies the LDLR (LDL receptor) gene, which has a high membership likelihood for the REACTOME 'metabolism of lipids and lipoproteins' gene set. The pairwise analysis does not identify LDLR, making this gene set less enriched. These results show that joint pleiotropy analysis of multiple traits yields more biologically relevant insights compared to pairwise analysis of those same traits.

3.6 Runtime increases marginally with the number of traits analyzed

To assess how the runtime scales with the number of traits analyzed, we picked all traits that were affected by the most pleiotropic SNP, rs495828: AS, BC, CAD, CES, DBP, HDL, HF, IS, LDL, T2D, TAGC and TC. In this order, we picked the first p traits and timed PolarMorphism (see Fig. 6). Runtime increases slightly with larger p, but the effect is small. There is a large difference between $p = 2$ and $p > 2$ because we use different approaches if $p > 2$ (see Section 2).

3.7 Comparison with other methods

To compare PolarMorphism to existing methods, we ran: PolarMorphism, intersection, PLACO and PRIMO on a selection of traits (IS and myocardial infarction). We compared the individual SNPs and loci that were identified as pleiotropic by each method. Four loci are found by all methods. Intersection does not identify more than those four loci. PLACO and PolarMorphism both find 21 loci (19 of which are identical), PRIMO finds 13 loci that were also identified by PLACO and PolarMorphism. PLACO and PolarMorphism use a fundamentally different approach to identify pleiotropy: whereas PLACO tests if the effect for both traits is not equal to zero, PolarMorphism first tests whether the overall effect (distance) is different than expected and then tests the sharedness of a SNP.

We timed each method from cleaned input data (already in memory, timing done in R) to results. The number of pleiotropic loci that were found by each method and the speed of generating results (in number of input SNPs per second) are provided in Table 2. These data show that PLACO does not identify more loci than PolarMorphism and is slower.

4 Discussion

We have developed a new method that identifies pleiotropic SNPs with an effect on multiple traits. PolarMorphism can be used on combinations of two or more traits. It uses GWAS summary statistics and corrects for correlation in effect sizes arising from genetic correlation or potential sample overlap. The potential applications of PolarMorphism include a) identifying SNPs that are shared between traits within a trait domain to learn more about the domain-wide biological processes, b) identifying SNPs that are shared among a diverse set of traits to find general biological processes and c) using the identified SNPs to inform new trait ontologies. As an example, we apply PolarMorphism to a set of traits from different domains.

The network analyses indicate that there are no trait domains that only share SNPs within the domain. We observe that most SNPs are associated with traits within one or across two trait domains. We zoomed in on the stroke domain, which has very little domain-specific SNPs. This may mean that the stroke traits are associated with general SNPs or that the stroke traits do not share many biological pathways. Each ischemic stroke subtype shares more SNPs with non-stroke traits than with the other ischemic stroke subtypes. Note that these networks are heavily influenced by the choice of included traits. Conclusions drawn about the networks in this study are therefore not necessarily general, as each trait could share SNPs with a number of traits that were not included. Future applications of PolarMorphism to a diverse set of traits will result in a more complete and precise overview of pleiotropy across the genome and across phenotypes.

We compared PolarMorphism with similar methods. PolarMorphism identifies more pleiotropic SNPs than the standard intersection method and than PRIMO. PLACO identifies the same number of pleiotropic loci as PolarMorphism. However, PolarMorphism finished analysis of 1 million SNPs in less than 20 s (compared to >25 min for PLACO), making analysis of many trait combinations feasible. Furthermore, PLACO can only be used to analyze two traits together while PolarMorphism can analyze a theoretically unlimited number of traits. A five-fold increase in the number of identified pleiotropic loci for the lipid domain indicates that analyzing more than two traits is much more powerful than combined results from the respective pairwise analyses.

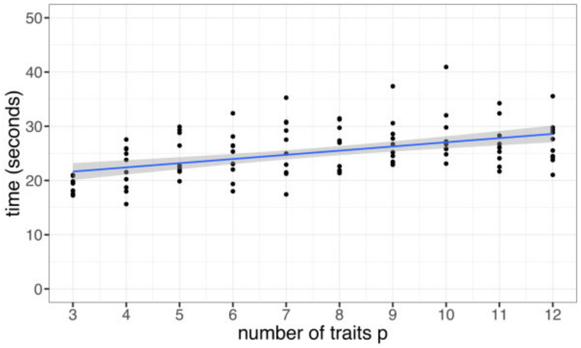

Fig. 6. Runtime scales with the number of traits p. The number of traits p ranges from 3 to 12. The slope of the regression line is 0.75 (SE = 0.13)

Table 2. Comparison of methods

	Decorrelation?	Pleiotropic loci found	Speed (1k SNPs/s)
PolarMorphism	Yes	21	63
PLACO	Yes	21	0.61
Primo	No	13	86
HOPS	Yes	—	—
PLEIO	No	—	—

Note: HOPS and PLEIO were not run because they use a pleiotropy definition that includes single-trait SNPs. Furthermore, PLACO can only be applied to two traits simultaneously.

Acknowledgements

The authors thank René Eijkemans for helpful discussions.

Funding

This work was supported by the National Institutes of Health [R01NS100178 to J.v.B.]; a Vidi Fellowship [639.072.715 to J.d.R.] from the Dutch Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO); the Netherlands CardioVascular Research Initiative of the Netherlands Heart Foundation [CVON 2011/B019 and CVON 2017-20: Generating the best evidence-based pharmaceutical targets for atherosclerosis [GENIUS I&II] to S.W.v.d.L.]. They are thankful for the support of the ERA-CVD program 'druggable-MI-targets' [01KL1802], the EU H2020 TO_AITION [848146] and the Leducq Fondation 'PlaqOmics'.

Conflict of Interest: none declared.

References

Ay,H. *et al.* (2007) A computerized algorithm for etiologic classification of ischemic stroke: the causative classification of stroke system. *Stroke*, **38**, 2979–2984.

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.

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

Burgess,S. *et al.* (2019) Guidelines for performing mendelian randomization investigations. *Wellcome Open Res.*, **4**, 186.

Carithers,L.J. *et al.*; GTEx Consortium. (2015) A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv. Biobank*, **13**, 311–319.

Csárdi,G. *et al.* (2016) *Statistical Network Analysis with Igraph*. Springer, Berlin.

Dichgans,M. *et al.* (2019) Stroke genetics: discovery, biology, and clinical applications,. *Lancet Neurol.*, **18**, 587–599.

Fabregat,A. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.

Feng,Y.-C.A. *et al.*; Transdisciplinary Research in Cancer of the Lung (TRICL). (2017) Investigating the genetic relationship between alzheimer's disease and cancer using GWAS summary statistics. *Hum. Genet.*, **136**, 1341–1351.

Fernandes,S.B. and Lipka,A.E. (2020) simplePHENOTYPES: SIMulation of pleiotropic, linked and epistatic phenotypes. *BMC Bioinformatics*, **21**, 491.

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

Gleason,K.J. *et al.* (2020) Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. *Genome Biol.*, **21**, 236.

Graff,R.E. *et al.* (2021) Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat. Commun.*, **12**, 970.

Grotzinger,A.D. *et al.* (2019) Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.*, **3**, 513–525.

Hankin,R.K.S. (2015) Numerical evaluation of the Gauss hypergeometric function with the hypergeo package. *R J.*, **72**, 81–88. https://journal.r-project.org/archive/2015/RJ-2015-022/index.html.

Harris,M.A. *et al.*; Gene Ontology Consortium. (2004) The gene ontology (GO) database and informatics resource,. *Nucleic Acids Res.*, **32**, D258–61.

Hemani,G. *et al.* (2018) *TwoSampleMR: Two Sample MR Functions and Interface to MR Base Database, R package version 030*. https://cran.r-project.org/web/packages/tictoc/index.html.

Izrailev,S. (2014) *tictoc: Functions for timing R scripts as well as implementations of Stack and List structures (R package version 1.0)*. https://cran.r-project.org/web/packages/tictoc/index.html.

Jordan,D.M. *et al.* (2019) HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. *Genome Biol.*, **20**, 222.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kessy,A. *et al.* (2018) Optimal whitening and decorrelation. *Am. Stat.*, **72**, 309–314.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Landler,L. *et al.* (2018) Circular data in biology: advice for effectively implementing statistical procedures,. *Behav. Ecol. Sociobiol.*, **72**, 128.

Lee,S.-H. (2020) Cerebral small vessel disease. In: Lee, S.-H. (ed) *Stroke Revisited Pathophysiology of Stroke*. Vol. 79, Springer Science+Business Media, Singapore. pp. 61–79.

Lee,C.H. *et al.* (2021) PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics. *Am. J. Hum. Genet.*, **108**, 36–48.

Li,H. *et al.*; 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Loh,P.-R. *et al.* (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, **47**, 284–290.

Maier,R. *et al.* (2015) Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.*, **96**, 283–294.

Malik,R. *et al.*; MEGASTROKE Consortium. (2018) Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.*, **50**, 524–537.

Malik,R. and Dichgans,M. (2018) Challenges and opportunities in stroke genetics. *Cardiovasc. Res.*, **114**, 1226–1240.

Ning,Z. *et al.* (2020) High-definition likelihood inference of genetic correlations across human complex traits. *Nat. Genet.*, **52**, 1–6.

O'Mara,T.A. *et al.* (2019) Editorial: establishing genetic pleiotropy to identify common pharmacological agents for common diseases. *Front. Pharmacol.*, **10**, 1038.

Otowa,T. *et al.* (2016) Meta-analysis of genome-wide association studies of anxiety disorders. *Mol. Psychiatry*, **21**, 1485.

Paaby,A.B. and Rockman,M.V. (2013) The many faces of pleiotropy. *Trends Genet.*, **29**, 66–73.

Pers,T.H. *et al.* (2015) Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.*, **6**, 5890.

Pulit,S.L. *et al.* (2018) Atrial fibrillation genetic risk differentiates cardioembolic stroke from other stroke subtypes. *Neurol. Genet.*, **4**, e293.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, *559–575*.

Ray,D. and Chatterjee,N. (2020) A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between type 2 diabetes and prostate cancer. *PLoS Genet.*, **16**, e1009218.

Sivakumaran,S. *et al.* (2011) Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.*, **89**, 607–618.

Solovieff,N. *et al.* (2013) Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.*, **14**, 483–495.

Storey,J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.*, **31**, 2013–2035.

Traylor,M. *et al.*; METASTROKE, International Stroke Genetics Consortium, Wellcome Trust Case Consortium 2 (WTCCC2). (2014) A novel MMP12 locus is associated with large artery atherosclerotic stroke using a genome-wide age-at-onset informed approach. *PLoS Genet.*, **10**, e1004469.

Traylor,M. *et al.*; METASTROKE, UK Young Lacunar DNA Study, NINDS Stroke Genetics Network, Neurology Working Group of the CHARGE Consortium. (2017) Genetic variation at 16q24.2 is associated with small vessel stroke. *Ann. Neurol.*, **81**, 383–394.

Tyler,A.L. *et al.* (Feb. 2009) Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays*, **31**, 220–227.

van Rheenen,W. *et al.* (2019) Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.*, **20**, 567–581.

Visscher,P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.

Wallace,C. (2020) Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.*, **16**, e1008720.

Wang,Q. *et al.* (2017) Genetic factor common to schizophrenia and HIV infection is associated with risky sexual behavior: antagonistic vs. synergistic pleiotropic SNPs enriched for distinctly different biological functions. *Hum. Genet.*, **136**, 75–83.

Watanabe,K. *et al.* (2019) A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.*, **52**, 353.

Wood,A.T.A. (1994) An R package for fast sampling from von Mises fisher distribution. https://nbviewer.jupyter.org/github/ahoundetoungan/vMF/blob/master/doc/vMF.pdf (20 October 2021, date last accessed).