ARTICLES
https://doi.org/10.1038/s41588-022-01017-y

nature genetics

An integrated framework for local genetic correlation analysis

Josefin Werme[1]✉, Sophie van der Sluis[2], Danielle Posthuma[1,2] and Christiaan A. de Leeuw[1]✉

Genetic correlation (r_g) analysis is used to identify phenotypes that may have a shared genetic basis. Traditionally, r_g is studied globally, considering only the average of the shared signal across the genome, although this approach may fail when the r_g is confined to particular genomic regions or in opposing directions at different loci. Current tools for local r_g analysis are restricted to analysis of two phenotypes. Here we introduce LAVA, an integrated framework for local r_g analysis that, in addition to testing the standard bivariate local r_gs between two phenotypes, can evaluate local heritabilities and analyze conditional genetic relations between several phenotypes using partial correlation and multiple regression. Applied to 25 behavioral and health phenotypes, we show considerable heterogeneity in the bivariate local r_gs across the genome, which is often masked by the global r_g patterns, and demonstrate how our conditional approaches can elucidate more complex, multivariate genetic relations.

Over a decade of genome-wide association studies (GWASs) have demonstrated that statistical pleiotropy is ubiquitous, with genetic variants, genes or regions often showing association with more than one phenotype[1–3]. Pleiotropy is valuable to study, as it could elucidate shared biological pathways[4–6], inform the functional significance of GWAS results[7–9] and improve our understanding of the etiology of complex traits and diseases[1,10,11].

On the variant level, pleiotropy is traditionally studied using colocalization[7,8,11,12], but a variety of different cross-trait genetic association methods now exist[5,9,13–15]. As extensive pleiotropy should induce a correlation between the genetic signals for different phenotypes, genetic correlation (r_g) analysis has been frequently employed to identify pairs of phenotypes that may be subject to widespread pleiotropy across the genome[13,16–18].

Although pleiotropy is typically discussed on a local level, such as single nucleotide polymorphisms (SNPs) or genes, r_g is traditionally studied on a global, genome-wide scale. But because a global r_g represents an average of the shared association across the genome, local r_gs in opposing directions could result in a nonsignificant global r_g, and local r_gs in the absence of any global relation may go undetected[17,19]. In addition, global r_gs offer limited insight into the underlying shared biological mechanisms, as the exact source of the genetic relation remains unidentified. To overcome this, some have partitioned the r_g by annotation (for example, GNOVA[20]) or restricted testing to variants that are assumed to be associated (MiXER[21]). And although global r_g methods cannot be easily translated to a local scale (due to the often high levels of local linkage disequilibrium (LD)), methods aimed at estimating local r_g have also started to emerge (rho-HESS[19], SUPERGNOVA[22] and LOGOdetect[23]).

To our knowledge, however, no existing tool offers the opportunity to model the local genetic relations using more than two phenotypes simultaneously, but doing so would allow a host of novel hypotheses to be tested, which could elucidate patterns of local genetic mediation and confounding. For example, multivariate local r_g analysis could answer questions such as "Can the local r_g between two phenotypes be accounted for by their genetic relation with another phenotype?" or "Within this user-defined region, to what extent is the genetic signal of a target phenotype explained by that of several predictor phenotypes?"

Although some localized methods can analyze several phenotypes at once (for example, meta-CCA[24] and Bolormaa et al.[25]), these simply aim to detect variants or genes that simultaneously influence multiple phenotypes, much like a form of meta-analysis, rather than evaluate their conditional genetic relations (Supplementary Table 1). These methods are also restricted to individuals SNPs or genes, whereas local r_g allows flexibility in region definition, which may be tailored to the goals of individual studies (including LD blocks, genes, gene complexes or clumped SNPs).

To address this gap, we have developed LAVA (Local Analysis of [co]Variant Association), a flexible and user-friendly tool that uses local r_g to detect regions of shared genetic association between phenotypes and test conditional local genetic relations. LAVA can analyze binary as well as continuous phenotypes with any degree of sample overlap and also evaluate the local univariate association signal (that is, the local h^2). Multivariate genetic association analysis can be performed via either partial correlation or multiple regression, allowing for complex, conditional genetic relationships to be examined in greater detail than currently possible.

Here we demonstrate the features of LAVA through application to real data, validate its properties and robustness using simulations, and benchmark its performance against existing methods. We first describe the method and analysis options and then apply LAVA to 25 behavioral and health-related phenotypes. We then examine the heterogeneity in the local r_g patterns compared to the global relations, identify pleiotropy hotspots containing significant r_gs between several phenotypes and finally zoom in on a subset of loci to illustrate how our multivariate approaches can elucidate the conditional genetic relations between phenotypes.

Results

Input processing and estimation of local genetic signal. For any genomic region of interest, consider a mean centered continuous phenotype vector $\mathbf{Y}_p$ (for phenotype p) with sample size N and a

[1]Department of Complex Trait Genetics, Centre for Neurogenomics and Cognitive Research, VU University, Amsterdam, the Netherlands. [2]Section Complex Trait Genetics, Department of Child and Adolescent Psychology and Psychiatry, Amsterdam Neuroscience, VU University Medical Centre, Amsterdam, the Netherlands. ✉e-mail: j.werme@vu.nl; c.a.de.leeuw@vu.nl

standardized genotype matrix X containing K_{snp} SNPs. We can model the relation between Y_p and X using a multiple linear regression model of the form $\mathbf{Y}_p = X\alpha_p + \epsilon_p$, where α_p represents the vector of joint SNP effects (accounting for the LD) and ϵ_p the vector of residuals, which are normally distributed with variance η_p^2.

Given that the standard least-squares estimate of α_p is of the form $\hat{\alpha}_p = (X^TX)^{-1}X^T\mathbf{Y}_p$, denoting the local SNP LD matrix as $S = \mathrm{cor}(X)$ and the vector of estimated marginal SNP effects as $\hat{\boldsymbol{\beta}}_p$ (that is, not accounting for LD), we can express $\hat{\alpha}_p$ as $\hat{\alpha}_p = S^{-1}\hat{\beta}_p$. After obtaining $\hat{\beta}_p$ from GWAS summary statistics for $\mathbf{Y}_p$ and using a reference genotype dataset (for example, 1000 Genomes[26]) from a population with a matching ancestry/LD structure to compute S, we can then estimate the joint SNP effects α_p without any individual-level data (this process is robust to sampling noise in S; Supplementary Fig. 1). To ensure that the direction of effect is consistent across phenotypes (which is crucial for preventing false positives; Supplementary Fig. 2), LAVA aligns the summary statistics to the reference before computing $\hat{\alpha}_p$.

Once we obtain $\hat{\alpha}_p$, we can estimate the local residual phenotypic variance η_p^2 and hence the phenotypic variance explained by the SNPs in the locus (that is, the local h^2), which we evaluate using an F test (Methods). We recommend using this test to filter out nonassociated loci prior to any r_g analysis, because r_gs will not be interpretable or reliable for loci devoid of genetic signal.

For binary phenotypes, we use a similar strategy, reconstructing a multiple logistic regression model for the locus based on the marginal SNP effects, and then use a χ^2 test to evaluate the local h^2 (Methods).

Estimation of bivariate local genetic correlations. For any region and set of P phenotypes, we define the matrix of local genetic components $G = X\alpha$, with α the K_{snp} by P matrix of joint effects on each phenotype (as outlined above). We denote the realized covariance matrix of G as $\mathbf{\Omega}$, such that each diagonal element ω_p^2 represents the local genetic variance of $\mathbf{G}_p$ for phenotype p, and each off-diagonal element ω_{pq} the local genetic covariance of $\mathbf{G}_p$ and $\mathbf{G}_q$ (for phenotypes p and q). Thus, for phenotypes p and q:

$$\mathbf{\Omega} = \begin{pmatrix} \omega_p^2 & \omega_{qp} \\ \omega_{pq} & \omega_q^2 \end{pmatrix}.$$

Given $\mathbf{\Omega}$, the local r_g can be computed as $\rho_{pq} = \frac{\omega_{pq}}{\sqrt{\omega_p^2\omega_q^2}}$, and its squared value, ρ_{pq}^2, can be interpreted as the proportion of variance in the local genetic component $\mathbf{G}_p$ that is explained by $\mathbf{G}_q$ (and vice versa). Because G is unobserved, $\mathbf{\Omega}$ cannot be computed directly but is estimated using the Method of Moments[27] (Methods). Once estimated, we evaluate significance of ρ_{pq} using simulation-based P values (Methods).

As shown in Supplementary Figs. 3–5, this approach produces unbiased parameter estimates with well-controlled type 1 error rates for binary as well as continuous phenotypes and a range of locus sizes. Note that because potential sample overlap between datasets can bias the estimated correlation, known or estimated sample overlap (for example, obtained via bivariate LD-score regression (LDSC)[13]) should be provided to LAVA to correct this (Methods and Supplementary Fig. 6).

To benchmark the performance of LAVA's bivariate test, we also applied rho-HESS to these simulations and compared their performance. We observed both inflated and deflated type 1 error rates for rho-HESS depending on simulation condition (further analysis discovered that this relates to the principal component (PC) pruning procedure; Supplementary Figs. 7 and 8), although this was not the case for LAVA, which showed well-controlled type 1 error rates regardless of condition. Because SUPERGNOVA employs a random effects model with a different null hypothesis (Supplementary Note), we did not apply it to these simulations but instead perform an application comparison using real data below.

We also compared the performance of LAVA's bivariate test to colocalization (using COLOC[12]) under infinitesimal and sparse causal scenarios. As LAVA makes no assumptions about the distribution of SNP effects, it performed well regardless of scenario, whereas COLOC was optimized for the sparse scenario (Supplementary Fig. 9 and Supplementary Table 2). For an overview of the conceptual differences between LAVA and colocalization, see the Supplementary Note.

Estimation of multivariate local genetic relations. We implemented multivariate local genetic association analysis in LAVA in two forms, partial correlation and multiple regression.

The partial correlation, denoted $\rho_{pq|Z}$, represents the local r_g between two phenotypes p and q while conditioning on some other phenotype(s) Z. This $\rho_{pq|Z}$ can be computed directly from $\mathbf{\Omega}$ (Methods) and indicates how much of the initial correlation ρ_{pq} remains once their r_gs with Z are accounted for.

In contrast, the multiple regression models the genetic signal for an outcome phenotype $\mathbf{Y}$ using that of two or more predictor phenotypes X. We formulate this as $\mathbf{G}_Y = \mathbf{G}_X\boldsymbol{\gamma} + \epsilon$ for standardized genetic components $\mathbf{G}_X$ and $\mathbf{G}_Y$, such that $\boldsymbol{\gamma}$ reflects the vector of standardized regression coefficients indicating how much the genetic component for each predictor in X contributes to $\mathbf{Y}$ (jointly), and ϵ the residuals with residual variance τ^2 (all computed directly from $\mathbf{\Omega}$; Methods). From this τ^2, we can also compute the model $R^2 = 1 - \tau^2$, reflecting the proportion of local heritability in $\mathbf{Y}$ that is explained by all predictors simultaneously.

For an overview of the relationship between partial correlation and multiple regression, as well as examples of different multivariate causal scenarios and how the conditional approaches would behave under those circumstances, see Supplementary Note.

As shown in Supplementary Figs. 10–14, our multivariate approaches provide unbiased estimates and well-controlled type 1 error rates, although for some instances of the partial correlation with binary phenotypes, this requires filtering variants with a minor allele frequency (MAF) below 0.5%.

Extensive overlap of local genetic association signals. We applied LAVA to 25 phenotypes (Table 1), testing the pairwise local r_gs within 2,495 genomic loci (genome-wide). Sample overlap was estimated using the intercepts from bivariate LDSC[13], and because the GWAS summary statistics were from European samples, we used the European panel of phase 3 of 1000 Genomes[26] (MAF > 0.5%) as a LD reference. The genomic loci were created by partitioning the genome into blocks of approximately equal size (~1 Mb) while minimizing the LD between them (Supplementary Methods). (Note that genomic coordinates are in reference to the human genome build 37.)

As detection of stable and interpretable local r_gs requires sufficient local genetic signal, we used the univariate test as a filtering step. For each phenotype pair, bivariate analysis was performed only for loci in which both phenotypes exhibited univariate signal at $P < 0.05/2{,}495$, resulting in 39,400 bivariate tests conducted in total.

With a Bonferroni-corrected P value threshold of 1.27×10^{-6} (0.05/39,400), we detected 1,119 significant bivariate local r_gs across 557 loci, of which 153 loci were associated with more than one phenotype pair. Notably, only 30% of loci with significant heritability for both phenotypes had a bivariate P value of less than 0.05, suggesting that strong local heritability occurs frequently in the absence of any local r_g (although potential power differences between the univariate and bivariate tests might also be a factor here). For 354 of the 1,119 significant correlations, 95% confidence intervals (CIs) for the explained variance included 1, consistent with the scenario that the local genetic signal of those phenotypes is completely shared (Fig. 1).

Table 1 | Overview of the 25 phenotypes included in this study

Phenotype	Short ID	*N*	Number of cases	Number of controls	Global h^2 (s.e.)	Atlas ID	Original study (year)
Coronary artery disease (including angina)	CAD	340,799	76,014	264,785	4.38% (0.24)	3,925	Nelson et al. (2017)[45]
Hypertension	Hypertension	244,890	71,332	173,558	9.93% (0.47)	3,691	Watanabe et al. (2019)[2]
Heart rate	HR	361,411	-	-	13.44% (0.82)	3,188	Watanabe et al. (2019)[2]
High-density lipoprotein cholesterol	HDL	188,577	-	-	21.49% (2.86)	71	Willer et al. (2013)[46]
Low-density lipoprotein cholesterol	LDL	188,577	-	-	20.37% (3.26)	70	Willer et al. (2013)[46]
Triglyceride cholesterol	TG	188,577	-	-	21.65% (3.53)	72	Willer et al. (2013)[46]
Total cholesterol	TC	188,577	-	-	21.24% (2.74)	73	Willer et al. (2013)[46]
Body mass index	BMI	379,831	-	-	22.03% (0.74)	3,445	Watanabe et al. (2019)[2]
Waist/hip ratio	WHR	697,734	-	-	13.28% (0.42)	4,077	Pulit et al. (2019)[47]
Type 1 diabetes	T1D	25,063	9,358	15,705	32.51% (17.15)	n/a	Forgetta et al. (2020)[48]
Type 2 diabetes	T2D	659,256	62,832	596,424	5.17% (0.31)	4,045	Xue et al. (2018)[49]
Asthma	Asthma	385,822	44,301	341,521	5.18% (0.47)	3,552	Watanabe et al. (2019)[2]
Hypothyroidism	Hypothyroid	244,890	13,043	231,847	4.31% (0.55)	3,685	Watanabe et al. (2019)[2]
Rheumatoid arthritis	RA	58,284	14,361	43,923	21.15% (7.38)	1,203	Okada et al. (2014)[50]
Lupus	Lupus	10,995	4,036	6,959	84% (21.1)	4,018	Julià et al. (2018)[51]
Crohn's disease	Crohn's	20,883	5,956	14,927	47.82% (5.51)	68	Liu et al. (2015)[52]
Ulcerative colitis	UC	27,432	6,968	20,464	25.26% (3.28)	69	Liu et al. (2015)[52]
Interleukin-6 receptor alpha	IL-6RA	21,758	-	-	21.72% (20.61)	n/a	Folkersen et al (2020)[53]
Tumor necrosis factor receptor 1	TNF-R1	21,741	-	-	7.29% (2.19)	n/a	Folkersen et al (2020)[53]
Depression	Depression	500,199	170,756	329,443	6.05% (0.23)	4,293	Howard et al. (2019)[31]
Neuroticism	Neuroticism	380,506	-	-	10.61% (0.35)	3,990	Nagel et al. (2018)[54]
Insomnia	Insomnia	386,078	-	-	6.06% (0.22)	3,232	Watanabe et al. (2019)[2]
Educational attainment	Education	766,345	-	-	10.71% (0.27)	4,066	Lee et al. (2018)[55]
Alcohol intake frequency	Alcohol	386,082	-	-	7.25% (0.34)	3,261	Watanabe et al. (2019)[2]
Current tobacco smoking	Smoking	386,150	-	-	4.22% (0.23)	3,235	Watanabe et al. (2019)[2]

Summary statistics for all but three phenotypes (with 'n/a' Atlas ID) were downloaded from the GWAS Atlas[2]; the relevant page for individual summary statistics can be accessed directly at https://atlas.ctglab.nl/traitDB/[Atlas-ID]. SNP heritabilities (h^2) were computed using LDSC[56].

Given the number of immune phenotypes analyzed, we retained the major histocompatibility complex (MHC; chr6:26–34Mb, 21 loci), as this region is highly relevant to the etiology of these phenotypes. Of the 1,119 significant local r_gs, 241 were found within the MHC (particularly for immune phenotypes), consistent with the notion that there is extensive pleiotropy within this region[2,28].

Here, we also benchmarked the performance of LAVA's bivariate test against SUPERGNOVA to evaluate their concordance and relative power. We found a greater power to detect significant associations with LAVA (Supplementary Table 3) and, with SUPERGNOVA estimates often being substantially out of bounds, also saw superior stability in LAVA estimated local r_gs (Supplementary Fig. 15).

Local r_g captures heterogeneous genetic relationships. For all phenotype pairs, we examined the strength and direction of the local r_gs by taking the average of the observed correlation coefficients across tested loci. As shown in Fig. 2, consistently positive r_gs with multiple significant loci were observed for many phenotypes for which we also saw significant global r_gs obtained via bivariate LDSC[13] (for example, neuroticism and depression, LDL and CAD, BMI and T2D, and RA and hypothyroidism). The average local r_gs from LAVA showed a correlation of 0.75 with LDSC estimates, indicating an overall good concordance between the two methods. We see no evidence of upward trends in the average LAVA estimates despite filtering on univariate h^2, indicating that this does not produce bias in the estimated correlations (Supplementary Fig. 16).

Despite the concordance between LAVA and LDSC, we saw several instances of phenotype pairs with a global r_g close to 0 that nonetheless exhibited significant local genetic correlations (for example, T1D and lupus, asthma and hypothyroidism, LDL and HDL, lupus and WHR, and asthma and RA), supporting the notion that global r_gs fail to capture heterogeneous genetic relations. As expected, these phenotypes tended to exhibit local r_gs in opposite directions and/or within a limited number of loci (Supplementary Fig. 17 contains an overview of the total number of positive versus negative local genetic correlations detected per phenotype pair).

Bivariate local r_gs implicate pleiotropy hotspots. We identified 153 regions containing significant r_gs between multiple phenotype pairs. The loci with the greatest number of significant r_gs tended to be located within the MHC (chr6:26–34Mb), a region where extensive pleiotropy has been noted previously[2,28]. Within the top MHC hotspots, immune-related phenotypes were among the most intercorrelated (Fig. 3a), consistent with the known role of the MHC in immune function[29,30].

a

Phen. 1	Phen. 2	N. sig.	$CI^{97.5} = 1$
BMI	WHR	321	37%
IL-6RA	TNF-R1	62	34%
LDL	TC	37	30%
Education	WHR	37	30%
BMI	Education	35	40%
Depression	Neuroticism	26	58%
Alcohol	BMI	19	21%
Hypothyroid	T1D	19	11%
TC	TG	18	28%
Lupus	T1D	18	0%
HR	WHR	17	18%
RA	T1D	17	0%
HDL	TC	15	47%
T2D	WHR	14	57%
BMI	HR	14	21%
Education	Neuroticism	13	69%
TG	WHR	13	23%
Asthma	RA	13	8%
Asthma	T1D	13	0%
Hypothyroid	RA	12	17%
Alcohol	WHR	12	17%
BMI	T2D	11	73%

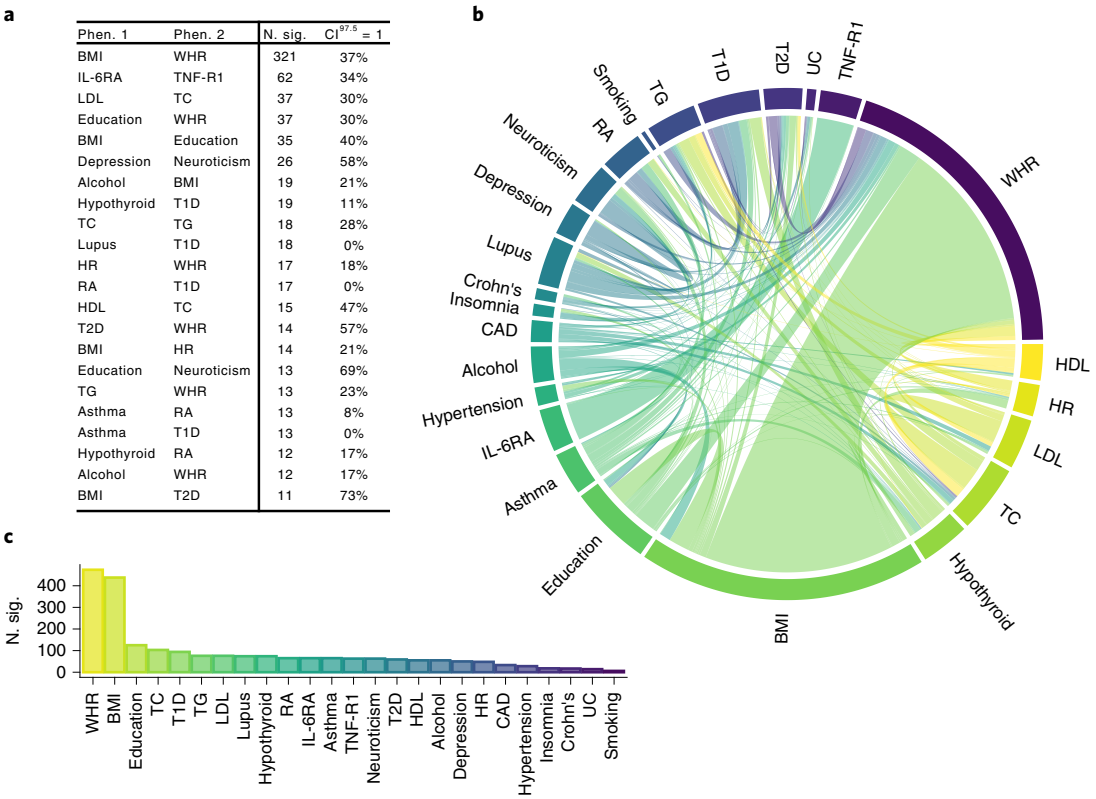

Fig. 1 | Overview of the number of significant bivariate local r_gs between all 25 phenotype pairs. a, Table showing the exact number of significant bivariate local r_gs detected between top phenotype (Phen.) pairs (N. sig.), together with proportion of significant loci for which the 95% CI included 1 ($CI^{97.5} = 1$). **b**, Chord diagram illustrating the number of significant bivariate local genetic correlations between all phenotype pairs. **c**, Bar plot showing the total number of significant bivariate correlations detected per phenotype. Colors in **b** and **c** represent individual phenotypes.

The locus with the greatest number of significant r_gs was locus 961 (chr6:31,427,210–32,208,901), which contains a total of 62 protein coding genes (including *C4A*, *C4B* and *NOTCH4*) and harbored 34 significant correlations between 15 different phenotypes (Fig. 3b and Supplementary Table 4). The second largest hotspot, locus 964 (chr6:32,539,568–32,586,784), contained 26 significant genetic correlations between 13 phenotypes and only a single protein coding gene, *HLA-DRB1* (Fig. 3c and Supplementary Table 5). Both of these loci were located within a region identified as the top pleiotropic locus in a recent large-scale investigation of pleiotropy by Watanabe et al.[2].

Outside the MHC, the two largest hotspots were locus 2,351 (chr19:45,040,933–45,893,307) and locus 1,351 (chr8:125,453,323–126,766,827), which contained 20 and 15 significant r_gs (respectively) between primarily cardiometabolic phenotypes (HDL, LDL, CAD, TG, TC, BMI and WHR; Supplementary Tables 6 and 11). These also overlapped with some of the top loci for cardiovascular and metabolic phenotypes detected by Watanabe et al.[2]. The top hotspot with a high proportion of behavioral phenotypes was locus 1,719 (chr11:112,755,447–113,889,019; Supplementary Table 27; for depression, neuroticism, educational attainment, alcohol intake, WHR and BMI), which contained *NCAM1* and *DRD2*, which have been frequently implicated in behavioral and psychiatric phenotypes[2,31–36]. This locus also overlapped with a region flagged by SUPERGNOVA where significant r_gs were identified for autism, bipolar disorder, depression, cognitive performance, schizophrenia and smoking initiation[22], as well as one of the top pleiotropic regions for psychiatric and cognitive phenotypes identified by Watanabe et al.[2].

For a complete overview of all hotspots, including the relevant statistics, genes and network plots, see Supplementary Tables 4–156.

BMI accounts for T2D–WHR relation in the *FTO* gene locus. To illustrate how the multivariate models can help disentangle patterns of genetic mediation and confounding, we zoomed in on a subset of phenotypes within a set of r_g hotspots.

First, we used locus 2,135 (chr16:53,393,883–54,866,095), which encompasses the *FTO* gene region and includes genes like *FTO*, *IRX32* and *RPGRIP1L*, which have been widely implicated in metabolic phenotypes, body composition and T2D[37–40]. Accordingly, we observed strong and highly significant local r_gs between T2D and BMI ($\rho = 0.86$, $P = 3.2 \times 10^{-52}$), T2D and WHR ($\rho = 0.80$, $P = 3.7 \times 10^{-33}$) and BMI and WHR ($\rho = 0.85$, $P = 6.2 \times 10^{-113}$) here.

Although BMI and WHR have both been implicated in T2D etiology[41,42], BMI measures the total body mass, whereas WHR measures its distribution[41]. Given the stronger relationship between BMI and T2D here, we used the partial correlation to determine whether any component of the r_g between WHR and T2D remained once accounting for BMI (T2D ~ WHR | BMI). This showed that conditioning on BMI led to a substantial reduction in the local r_g between T2D and WHR, with the association no longer nominally significant ($\rho = 0.19$, $P = 0.20$; Table 2). In the inverse model (T2D ~ BMI | WHR), however, a sizeable portion of the original association between BMI and T2D was still retained ($\rho = 0.59$, $P = 8.20 \times 10^{-6}$; Table 2).

Although the partial correlation is conceptually more suitable for this scenario (Supplementary Note), we also applied the multiple regression to demonstrate the concordance between these approaches, modeling the genetic signal for T2D as the outcome while using BMI and WHR as predictors (T2D ~ BMI + WHR). As seen in Table 2, the change in coefficients mimics that of the partial correlations, with the joint R^2 and WHR P value indicating that WHR

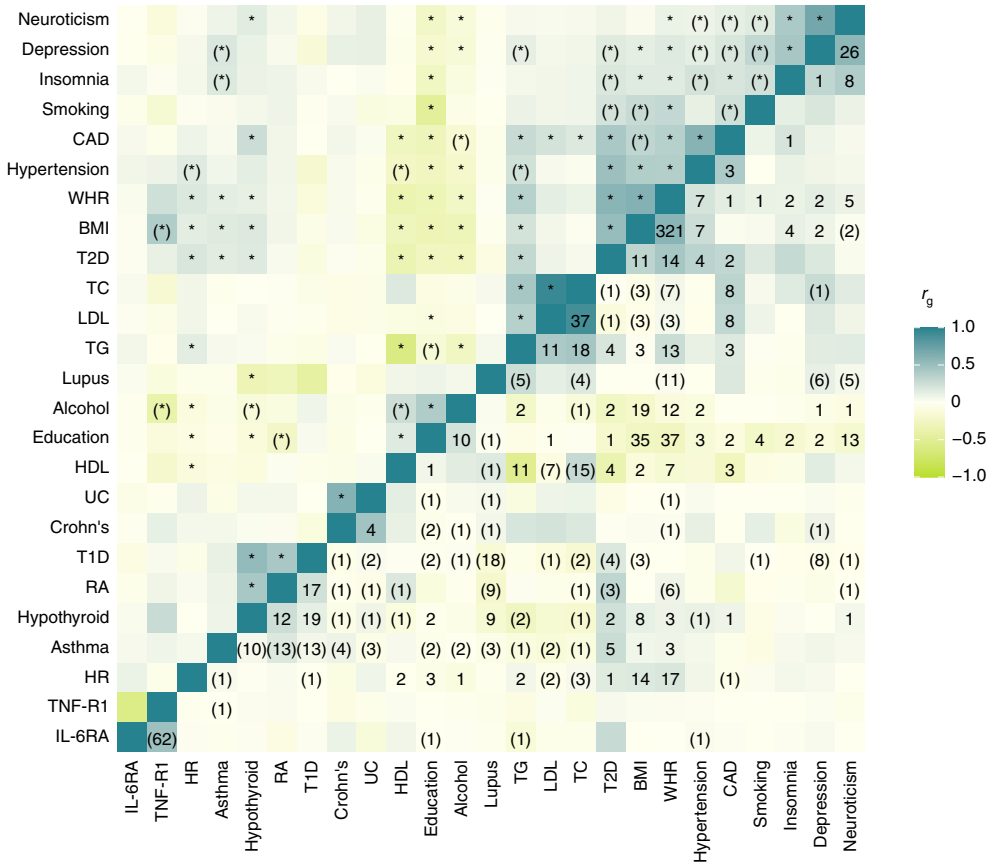

Fig. 2 | Comparison between the global genetic correlations estimated using LDSC and the mean local genetic correlations from LAVA across tested loci. Global genetic correlation estimates are shown above the diagonal, with asterisks indicating a significant global genetic correlation estimated by LDSC ($P < 0.05/300 = 1.7 \times 10^{-4}$) Mean local genetic correlations from LAVA are shown below the diagonal, where the numbers indicate how many significant local r_gs were detected ($P < 0.05/39{,}400 = 1.27 \times 10^{-6}$). Parentheses around an asterisk indicate that a significant global r_g was detected by LDSC, without any evidence for a significant local r_g, whereas parentheses around a number indicate that at least one local r_g was detected by LAVA, despite no significant global r_g. Although LDSC excludes a small region within the MHC (chr6:30–31Mb), this region concerns only three of the loci analyzed with LAVA, within which we detected 19 significant r_gs between 11 phenotypes, of which all phenotype pairs except hypothyroidism–neuroticism and neuroticism–T1D exhibited significant r_gs elsewhere.

does not notably improve the explained variance above the bivariate model with only BMI. Results from both approaches indicate that the local r_g between WHR and T2D in the *FTO* gene locus is driven by BMI and thus that the local genetic signal for body mass, rather than its distribution, confers a greater genetic risk of T2D here.

LDL accounts for CAD–HDL relation in the *APOE* gene locus. In locus 2,351 (chr19:45,040,933–45,893,307), we observed significant r_gs between CAD and LDL ($\rho = 0.90$, $P = 3.9 \times 10^{-31}$), CAD and HDL ($\rho = -0.63$, $P = 4.9 \times 10^{-10}$) and LDL and HDL ($\rho = -0.76$, $P = 3.1 \times 10^{-52}$). This locus contains *APOE*, a gene that has previously been linked with CAD, possibly via its influence on LDL metabolism[43,44]. Given the stronger relationship between CAD and LDL, with 95% CIs for the R^2 also containing 1, we hypothesized that LDL could account for the local r_g between CAD and HDL.

As seen in Table 3, the partial correlations indicate that the local genetic signal for CAD is driven by LDL, rather than HDL, which is, again, closely mimicked by the multiple regression model. In all cases, conditional parameters for LDL are largely unchanged compared to its bivariate coefficient, whereas those of HDL are close to 0 and not even nominally significant when conditioning on LDL. This suggests that the genetic signal for LDL is more closely linked to CAD etiology within the *APOE* gene region and that the association with HDL is likely driven by LDL.

Decomposing the local genetic signal for hypertension. The multiple regression model can be used to decompose the genetic signal for an outcome phenotype while examining genetic confounding between the predictors. We illustrate this with locus 1,209 (chr7:130,418,705–131,856,481) and locus 2,135 (chr16:53,393,883–54,866,095), where we model the genetic signal of hypertension using T2D, WHR and BMI as predictors.

In locus 1,209, the strong bivariate correlation between hypertension and WHR (with 95% CIs including 1) suggests that WHR might account for the relationship with the other predictors (Table 4). Fitting conditional models with either T2D or BMI as a second predictor supports this; in both cases, the joint R^2 is only marginally higher and the P values for the added predictors not even nominally significant, indicating that the fit of the two-predictor models do not significantly improve upon that with only WHR.

In locus 2135, however, WHR shows the weakest association, with BMI and T2D showing a similar, somewhat stronger association. Fitting all two-predictor models shows no improvement in R^2 over either T2D or BMI by itself, suggesting that all three predictors are tapping into the same genetic relation with hypertension in this locus (Table 4). Although unlike locus 1,209, a substantial proportion of the hypertension signal appears independent of either predictor (maximum $R^2 = 0.48$). These distinct patterns of multivariate

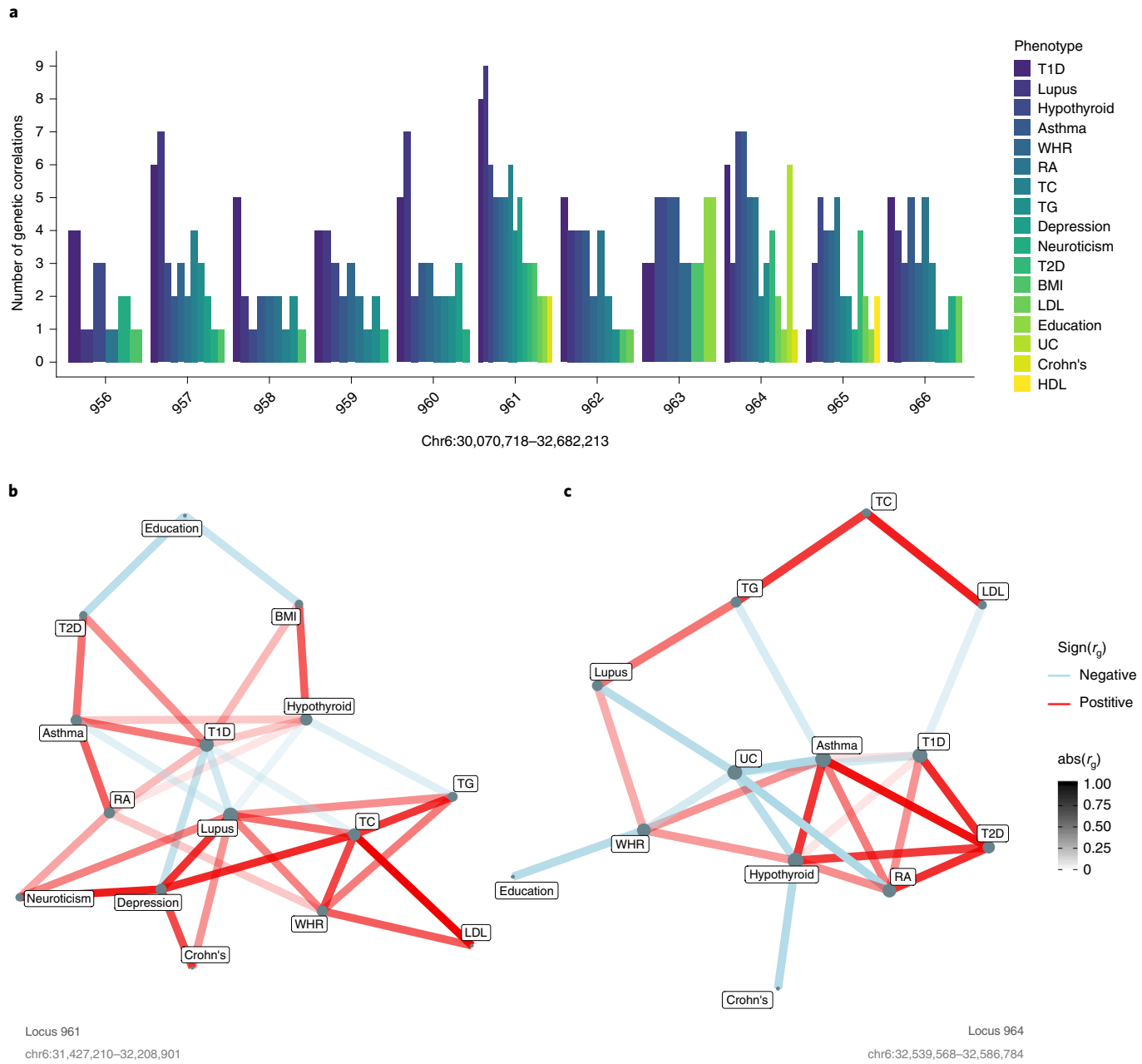

Fig. 3 | Local genetic correlation hotspots within the MHC. a, Bar plot showing how many significant bivariate r_gs were detected for each phenotype on the *y* axis, within the ten adjacent loci of the biggest hotspot on the *x* axis (chr6:30,070,718--32,682,213). The phenotypes have been ordered according to total number of significant r_gs across these loci, indicating that T1D, lupus, hypothyroidism and asthma were the most interconnected phenotypes within this region. **b**,**c**, Network plots showing the bivariate local r_gs within the two biggest MHC hotspots (locus 961, chr6:31,427,210–32,208,901 (**b**); and locus 964, chr6:32,539,568–32,586,784 (**c**)). Color indicates direction of effect (that is, sign(r_g); red, positive; blue, negative) and opacity the strength (that is, abs(r_g), whereas node size reflects the number of significant local r_gs for each phenotype.

genetic relations further emphasize the heterogeneity in the genetic relations across the genome.

Discussion

Genetic correlation analysis is used to identify pairs of phenotypes that have a shared genetic basis. The traditional, global approach considers only the average genetic correlation (r_g) across the genome and may fail to detect scenarios where the signal is confined to specific regions or in opposing directions at different loci[13,17,19].

Here, we present LAVA, a statistical framework aimed at testing the local r_g within user-defined regions. LAVA handles continuous and binary phenotypes with any degree of sample overlap (known or estimated), and, in addition to testing the standard bivariate local r_gs between two phenotypes, LAVA can evaluate the univariate local h^2 and model conditional genetic relations using partial correlation and multiple regression. Via extensive simulations, we show that LAVA produces unbiased estimates with well-controlled type 1 error rates, with superior performance compared to existing approaches.

Applied to 25 behavioral and health phenotypes across 2,495 regions, we identified 1,119 significant bivariate local r_gs in 557 regions. Although the direction of effect for individual phenotype

Table 2 | Conditional genetic relations for T2D within the *FTO* gene locus (chr16:53,393,883–54,866,095)

Model	Phenotype	Parameter	R^2	P
Bivariate	WHR	0.79	0.63	2.50×10^{-38}
Bivariate	BMI	0.86	0.75	8.80×10^{-54}
Partial correlation	WHR | BMI	0.19	-	2.00×10^{-1}
Partial correlation	BMI | WHR	0.59	-	8.20×10^{-6}
Multiple regression	WHR	0.19	0.76	2.10×10^{-1}
	BMI	0.70		1.90×10^{-4}

Empirical *P* values (unadjusted) were obtained via a permutation procedure with partial integration, evaluating the two-sided hypothesis of no association using the estimated parameters as test statistics (Methods). Note that values from the bivariate tests presented here might differ slightly from the original pairwise analyses, as these analyses were performed on the SNPs shared across all three datasets.

Table 3 | Conditional genetic relations for CAD within the *APOE* gene locus (chr19:45,040,933–45,893,307)

Model	Phenotype	Parameter	R^2	P
Bivariate	HDL	−0.65	0.42	9.45×10^{-11}
Bivariate	LDL	0.90	0.81[a]	3.72×10^{-31}
Partial correlation	HDL | LDL	0.01	-	9.62×10^{-1}
Partial correlation	LDL | HDL	0.82[a]	-	6.22×10^{-9}
Multiple regression	HDL	0.01	0.81[a]	9.62×10^{-1}
	LDL	0.91		3.09×10^{-5}

Empirical *P* values (unadjusted) were obtained via a permutation procedure with partial integration, evaluating the two-sided hypothesis of no association using the estimated parameters as test statistics (Methods). Note that values from the bivariate tests presented here might differ slightly from the original pairwise analyses, as these analyses were performed on the SNPs shared across all three datasets. [a]The 95% CIs included 1.

pairs was often consistent across loci (particularly for phenotypes showing a strong global r_g), there was substantial variability in associations across the genome, indicating that the genome-wide r_g is far from constant. We also identified significant r_gs in opposing directions for several phenotypes and observed significant local r_gs in absence of any global relationship, implying a more complex etiological relationship than that revealed by global r_g.

Furthermore, we identified several loci harboring significant r_gs between multiple phenotype pairs, implicating these regions as potential pleiotropy hotspots. As expected, many of these were located in the MHC, likely owing to the number of immune- and health-related phenotypes analyzed[29,30] and the MHC having been flagged as a pleiotropy hotspot previously[2,28]. Top hotspots identified for cardiometabolic and behavioral phenotypes also replicated previous studies investigating pleiotropy.

To demonstrate how more complex association patterns can be disentangled using LAVA's multivariate approaches, we selected a subset of phenotypes within some of the hotspots and evaluated their conditional genetic relationships. Using partial correlation, we found that the r_g between WHR and T2D in the *FTO* locus is likely driven by BMI, suggesting that genetic effects on total body mass, rather than its distribution, confer an increased genetic risk of T2D in this locus. In the *APOE* gene region, we showed that LDL could account for the r_g between HDL and CAD, implying a more direct etiological relationship between LDL and CAD in this locus. Finally, we employed multiple regression to decompose the genetic signal for hypertension in two loci using WHR, BMI and T2D as predictors, demonstrating heterogeneity in the multivariate genetic relations in these loci, and how this approach can help disentangle patterns of genetic mediation and confounding as well.

Our method is not without limitations. Analytical approaches like LAVA can only pinpoint locations where pleiotropy is likely but may be confounded by excessive LD, and ultimately, experimental validation is required to establish the nature of any observed genetic relationship. In addition, as LAVA cannot condition on the association signal from nearby loci, it is possible that local r_gs may be confounded by association signals from adjacent regions. LAVA is also limited by the number of overlapping SNPs within different summary statistics datasets, which could potentially lead to a lack of power to detect true correlations in scenarios where there are few overlapping SNPs (although we endeavor to address these limitations where possible in the future).

Here, we have presented a tool dedicated to local genetic correlation analysis, which can evaluate standard bivariate genetic relations and univariate local heritabilities and analyze the conditional genetic relations between multiple phenotypes. This framework can help elucidate disease etiology and genetic relationships as a whole, which can be aided by integrating summary statistics of molecular phenotypes or endophenotypes (such as gene expression, metabolites or regional brain volumes) to facilitate the functional interpretation of GWAS results. In this setting, the conditional models could prove particularly useful, as they might enable identification of key

Table 4 | Local multivariate genetic association signals for hypertension using T2D, WHR and BMI as predictors

Predictor	Locus 1,209, chr7:130,418,705–131,856,481			Locus 2,135, chr16:53,393,883–54,866,095		
	Parameter	R^2	P	Parameter	R^2	P
T2D	0.60	0.36	1.4×10^{-7}	0.67	0.45	9.7×10^{-9}
WHR	0.84	0.70[a]	3.2×10^{-11}	0.61	0.38	4.8×10^{-9}
BMI	0.57	0.33	3.5×10^{-6}	0.68	0.46	2.4×10^{-12}
T2D	0.52	0.59	1.9×10^{-4}	0.34	0.48	3.2×10^{-1}
BMI	0.48		7.3×10^{-4}	0.39		2.5×10^{-1}
T2D	−0.07	0.70[a]	8.3×10^{-1}	0.49	0.47	6.5×10^{-2}
WHR	0.89		8.0×10^{-2}	0.22		3.7×10^{-1}
BMI	0.26	0.76[a]	1.0×10^{-1}	0.57	0.46	2.7×10^{-2}
WHR	0.73		1.1×10^{-4}	0.12		6.2×10^{-1}

The standard bivariate relations are shown on top followed by the two-predictor models. Empirical *P* values (unadjusted) were obtained via a permutation procedure with partial integration, evaluating the two-sided hypothesis of no association using the estimated parameters as test statistics (Methods). Note that values from the bivariate tests presented here might differ slightly from the original pairwise analyses, as these analyses were performed on the SNPs shared across all four datasets. [a]The 95% CIs included 1.

tissues or regions, offering unique insight into the biological mechanisms associated with complex traits. The conditional models could also help determine whether molecular phenotypes or endophenotypes account for the genetic correlation between any higher-order phenotypes, thereby illuminating the underlying biological paths that drive their relationship.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01017-y.

Received: 28 January 2021; Accepted: 20 January 2022;
Published online: 14 March 2022

References

1. Chesmore, K., Bartlett, J. & Williams, S. M. The ubiquity of pleiotropy in human disease. *Hum. Genet.* **137**, 39–44 (2018).
2. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
3. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
4. Shikov, A. E., Skitchenko, R. K., Predeus, A. V. & Barbitoff, Y. A. Phenome-wide functional dissection of pleiotropic effects highlights key molecular pathways for human complex traits. *Sci. Rep.* **10**, 1037 (2020).
5. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* **7**, 170125 (2017).
6. Zhao, W. et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat. Genet.* **49**, 1450–1457 (2017).
7. Giambartolomei, C. et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).
8. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
9. Porcu, E. et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 3300 (2019).
10. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
11. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
12. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
13. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
14. Zeng, P., Hao, X. & Zhou, X. Pleiotropic mapping and annotation selection in genome-wide association studies with penalized Gaussian mixture models. *Bioinformatics* **34**, 2797–2807 (2018).
15. Grotzinger, A. D. et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).
16. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
17. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).
18. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
19. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* **101**, 737–751 (2017).
20. Lu, Q. et al. A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *Am. J. Hum. Genet.* **101**, 939–964 (2017).
21. Frei, O. et al. Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nat. Commun.* **10**, 2417 (2019).
22. Zhang, Y. et al. SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome Biol.* **22**, 262 (2021).
23. Guo, H., Li, J. J., Lu, Q. & Hou, L. Detecting local genetic correlations with scan statistics. *Nat. Commun.* **12**, 2033 (2021).
24. Cichonska, A. et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* **32**, 1981–1989 (2016).
25. Bolormaa, S. et al. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS Genet.* **10**, e1004198 (2014).
26. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
27. Newey, W. K. & West, K. D. Hypothesis testing with efficient method of moments estimation. *Int. Econ. Rev.* **28**, 777 (1987).
28. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
29. Flajnik, M. F. & Kasahara, M. Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* **15**, 351–362 (2001).
30. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76 (2017).
31. Howard, D. M. et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).
32. Zhou, H. et al. Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nat. Neurosci.* **23**, 809–818 (2020).
33. Pasman, J. A. et al. GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia. *Nat. Neurosci.* **21**, 1161–1170 (2018).
34. Nagel, M. et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **50**, 920–927 (2018).
35. Lane, J. M. et al. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat. Genet.* **49**, 274–281 (2017).
36. Mota, N. R. et al. NCAM1-TTC12-ANKK1-DRD2 gene cluster and the clinical and genetic heterogeneity of adults with ADHD. *Am. J. Med. Genet. B* **168**, 433–444 (2015).
37. Chauhdary, Z., Rehman, K. & Akash, M. S. H. The composite alliance of FTO locus with obesity-related genetic variants. *Clin. Exp. Pharmacol. Physiol.* **48**, 954–965 (2021).
38. Smemo, S. et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).
39. Doaei, S. et al. Changes in FTO and IRX3 gene expression in obese and overweight male adolescents undergoing an intensive lifestyle intervention and the role of FTO genotype in this interaction. *J. Transl. Med.* **17**, 1–8 (2019).
40. Bego, T. et al. Association of FTO gene variant (rs8050136) with type 2 diabetes and markers of obesity, glycaemic control and inflammation. *J. Med. Biochem.* **38**, 153–163 (2019).
41. Wang, Y., Rimm, E. B., Stampfer, M. J., Willett, W. C. & Hu, F. B. Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men. *Am. J. Clin. Nutr.* **81**, 555–563 (2005).
42. Lee, D. H. et al. Comparison of the association of predicted fat mass, body mass index, and other obesity indicators with type 2 diabetes risk: two large prospective studies in US men and women. *Eur. J. Epidemiol.* **33**, 1113–1123 (2018).
43. Afroze, D., Yousuf, A., Tramboo, N. A., Shah, Z. A. & Ahmad, A. ApoE gene polymorphism and its relationship with coronary artery disease in ethnic Kashmiri population. *Clin. Exp. Med.* **16**, 551–556 (2016).
44. Karjalainen, J. P. et al. New evidence from plasma ceramides links apoE polymorphism to greater risk of coronary artery disease in Finnish adults. *J. Lipid Res.* **60**, 1622–1629 (2019).
45. Nelson, C. P. et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).
46. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1285 (2013).
47. Pulit, S. L. et al. Meta-analysis of genome-wide association studies for body fat distribution in 694,649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).
48. Forgetta, V. et al. Rare genetic variants of large effect influence risk of type 1 diabetes. *Diabetes* **69**, 784–795 (2020).
49. Xue, A. et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes.*Nat. Commun.* **9**, 2941 (2018).
50. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

51. Julià, A. et al. Genome-wide association study meta-analysis identifies five new loci for systemic lupus erythematosus. *Arthritis Res. Ther.* **20**, 100 (2018).
52. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
53. Folkersen, L. et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).
54. Nagel, M., Watanabe, K., Stringer, S., Posthuma, D. & Van Der Sluis, S. Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat. Commun.* **9**, 905 (2018).
55. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
56. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

Methods

Ethics statement. This study relied on simulated data and secondary analysis of publicly available summary statistics, for which ethical approval was obtained by the primary researchers.

Model overview and input processing of continuous phenotypes. For any locus and phenotype p, consider a linear regression model of the standardized phenotype vector $\mathbf{Y}_p$ on a genotype matrix X (containing K_{snp} SNPs, also standardized), $\mathbf{Y}_p = X\alpha_p + \epsilon_p$, where α_p represents the vector of standardized joint SNP effects and ϵ_p the vector of normally distributed residuals with variance η_p^2. We denote the SNP LD matrix as $S = \text{cor}(X)$ and the vector of estimated marginal SNP effects $\hat{\boldsymbol{\beta}}_p$ (standardized) and obtain the estimated joint effects from these GWAS summary statistics as $\hat{\alpha}_p = S^{-1}\hat{\beta}_p$ using the genotype reference to compute S (SNP LD in the reference data is assumed to match the GWAS sample). We then estimate the residual variance as $\hat{\eta}_p^2 = \frac{N-1}{N-K-1}\left(1 - R_p^2\right)$, N being the original GWAS sample size and K the number of SNP PCs (see below), with explained variance $R_p^2 = \hat{\alpha}_p^T S \hat{\alpha}_p$. The estimated joint effects $\hat{\alpha}_p$ are distributed as $\text{MVN}\left(\alpha_p,\ \hat{\sigma}_p^2 S^{-1}\right)$, with $\hat{\sigma}_p^2 = \frac{\hat{\eta}_p^2}{N-1}$ being the sampling variance.

As we cannot know whether provided coefficients are standardized, for each SNP s, we create Z scores using the P value and sign of the provided effect size as $Z_{ps} = -\Phi(P_{ps}/2) \times \text{sign}(\hat{\beta}_{ps})$, with Φ the cumulative normal distribution function and P_{ps} the SNP P value. We then convert Z_{ps} to the corresponding correlation $r_{ps} = \frac{Z_{ps}}{\sqrt{Z_{ps}^2 + N_s - 2}}$, which equals the standardized beta coefficient (when per SNP sample size N_s is not provided, we use the overall N as a proxy). If Z scores or T statistics are provided, then we use these directly instead, in which case P values and coefficients are not necessary.

To ensure that SNPs have consistent direction of effect across phenotypes, LAVA also aligns the effect alleles to the reference before processing, flipping the sign of any marginal SNP effects for which the reported effect allele does not correspond to the reference (note that all strand-ambiguous variants are excluded, as these cannot be aligned).

Due to the substantial LD between SNPs, it is unlikely that the LD matrix S will be of full rank, in which case it is not invertible, requiring us to work in a lower dimensional space. To do so, we compute the singular value decomposition $\frac{X_{\text{ref}}}{\sqrt{N_{\text{ref}}-1}} = U\Lambda Q^T$, such that $S = Q\Lambda\Lambda Q^T$ and hence $S^{-1} = Q\left(\Lambda\Lambda\right)^{-1} Q^T$ (N_{ref} denotes the sample size of the reference data with genotype matrix X_{ref}). For each component j, the corresponding squared singular value λ_j^2 is proportional to the amount of variance of the total accounted for by that component. We order the components by decreasing singular value and select the smallest subset of the first K components such that these account for at least 99% of the total variance (pruning away the rest).

Defining Q_* as the K_{snp} by K pruned eigenvector matrix and Λ_* as the corresponding K by K diagonal singular value matrix, we approximate the inverse of S as $S^{-1} = Q_*\left(\Lambda_*\Lambda_*\right)^{-1} Q_*^T$. We then define the scaled PC matrix $W = XR$ with projection matrix $R = Q_*\,\Lambda_*^{-1}$, and then the corresponding vector of joint effects $\delta_p = R^+\alpha_p$, with $R^+ = \Lambda_* Q_*^T$ such that $W\delta_p$ closely approximates $\mathbf{G}_p = X\alpha_p$, using this sparser δ_p in place of α_p or $\mathbf{G}_p$ for parameter estimation instead.

To test the proportion of phenotypic variance that can be attributed to the local genetic signal, we construct the test statistic $T_{p_{\text{cont}}} = \frac{1}{\hat{\sigma}_p^2}\sum\frac{\hat{\delta}_p^2}{K}$, which is evaluated using an F-distribution with K and $N - K - 1$ degrees of freedom (this is a generalization of the PC regression gene analysis model in MAGMA[57] and is mathematically equivalent to that presented by Yang et al.[58]). Note that this test requires the sample size of the summary statistics to be greater than the number of PCs within a locus, although this is unlikely to not be true for typical GWAS sample sizes and the kind of loci analyzed here (on average 184 PCs per locus, with a sum of 458,468 PCs across all loci).

Processing of binary phenotypes. To obtain the joint SNP effects from GWAS summary statistics of binary phenotypes, for the scaled PCs W defined above, and with i denoting the individual, we reconstruct the multiple logistic regression model $\text{E}\left(Y_{pi}\right) = \mu_{pi} = \frac{1}{1+\exp\left(-\mathbf{W}_{i*}^T\delta_{p*}\right)}$, where $\mathbf{W}_{i*}$ is the ith row of the matrix $W_* = (1\mathbf{W})$ and $\delta_{p*} = \begin{pmatrix}\delta_{0p}\\ \delta_p\end{pmatrix}$, with δ_{0p} the model intercept. To do so, we use the iteratively reweighted least-squares (IRLS) approach[59], which iteratively updates estimates of the model coefficients according to the equation:

$$\hat{\delta}_p^{(k+1)} = \hat{\delta}_p^{(k)} + V_p^{(k)}\left(W_*^T\mathbf{Y}_p - W_*^T\mu_p^{(k)}\right).$$

Here $V_p = \left(W_*^T\text{diag}\left(\mathbf{c}_p\right)W_*\right)^{-1}$, where $\mathbf{c}_p$ is a vector with $c_{pi} = \mu_{pi}(1-\mu_{pi})$ and k is the index of the current iteration.

For the sufficient statistic $W_*^T\mathbf{Y}_p$ needed for this process, we note that $1^T\mathbf{Y}_p = N_1 = P_{\text{case}} \times N_{\text{ref}}$, with P_{case} designated as the proportion of individuals in the original sample that are cases ($Y_{pi} = 1$). In addition, we have that $W^T\mathbf{Y}_p = R^T X^T\mathbf{Y}_p$ for the standardized SNP genotype matrix X and R the projection matrix for W. This sufficient statistic can therefore be computed from the individual $\mathbf{X}_s^T\mathbf{Y}_p$ for each SNP s. To obtain these, we define the marginal logistic regression model $\text{E}\left(Y_{pi}\right) = u_{psi} = \frac{1}{1+\exp\left(-\mathbf{X}_{si}^T\beta_{ps*}\right)}$, with $\mathbf{X}_{si}$ the ith row of the matrix $X_{s*} = (1\mathbf{X}_s)$ and $\beta_{ps*} = \begin{pmatrix}\beta_{0ps}\\ \beta_{1ps}\end{pmatrix}$, and observe that at convergence of the IRLS algorithm $X_{s*}^T\mathbf{Y}_p = X_{s*}^T\mathbf{u}_{psi}$, the left side of which can be obtained by filling in the marginal SNP effect estimates $\hat{\beta}_{ps*}$. Because the intercept is unlikely to have been reported in the summary statistics and the slope may not be on the correct scale, we use a search algorithm to re-estimate these SNP effects from the GWAS Z statistics and case-control counts reported for s.

The $\hat{\delta}_p$ can then be estimated using the IRLS algorithm as outlined above, which has sampling covariance matrix $V_p\frac{N_{\text{eff}}}{N}$. Because the components in W are independent and have the same variance, V_p should be close to a diagonal matrix, and the standard errors for each $\hat{\delta}_{ps}$ should be similar. We verify this by computing the ratio between the maximum and median standard error, and if this exceeds 1.75, then the PC with the highest standard error is discarded and the process repeated until the threshold is no longer exceeded. Subsequently, we define $\hat{\sigma}_p^2$ as the mean of the diagonal elements of this sampling covariance matrix and assume $\hat{\boldsymbol{\delta}}_p \approx \text{MVN}\left(\boldsymbol{\delta}_p, \hat{\sigma}_p^2 I_K\right)$ (our simulations show that this approach has no appreciable effect on type 1 error rates; Supplementary Fig. 4). (Note that because logistic regression provides unbiased estimates of the regression parameters regardless of ascertainment (Agresti[59]), sample ascertainment will not influence the resulting genetic correlations).

Here, we also define a test for the univariate signal similar to the F test for continuous phenotypes, using the test statistic $T_{p_{\text{bin}}} = \frac{\hat{\delta}_p^T\delta_p}{\hat{\sigma}_p^2}$. Given the distribution for $\hat{\delta}_p$, this test statistic has a χ_K^2 distribution under the null of no genetic association.

Computing local heritability. Because the SNP-based heritability represents the proportion of phenotypic variance that can be explained by all SNPs in the genome, we can analogously define the local heritability as the proportion of phenotypic variance that can be explained by the SNPs in that locus. For continuous phenotypes, the total local heritability is therefore equivalent to the adjusted R^2 of the SNP PCs. To obtain this, we compute the unadjusted R^2 for phenotype p as $R^2 = \frac{\text{var}\left(\hat{\mathbf{G}}_p\right)}{\text{var}\left(\mathbf{Y}_p\right)}$, where $\text{var}\left(\hat{\mathbf{G}}_p\right) = \text{var}\left(W\hat{\delta}_p\right) = \frac{\hat{\delta}_p^T W^T W\hat{\delta}_p}{N-1} = \hat{\delta}_p^T\hat{\delta}_p$ and $\text{var}\left(\mathbf{Y}_p\right) = 1$, because continuous phenotypes are analyzed on a standardized scale. From this, the heritability (that is, the adjusted R^2) is then obtained as: $h^2 = R_{\text{adj}}^2 = 1 - (1-R^2)\frac{N-1}{N-K-1}$.

For binary phenotypes, we obtain the observed scale local heritability $h_o^2 = R_{\text{adj}}^2$ in the same way, except that $\text{var}\left(\mathbf{Y}_p\right) = P_{\text{case}}\left(1 - P_{\text{case}}\right)\frac{N}{N-1}$, with P_{case} the case proportion in the sample. Optionally, the liability scale local heritability h_l^2 can then be computed as $h_l^2 = h_o^2\left(\frac{\pi(1-\pi)}{z^2}\right)$, where π is the population prevalence of the phenotype and z the height of the standard normal distribution function at the quantile corresponding to π[60].

Estimating bivariate local genetic correlations. We define $\boldsymbol{\Omega}_{\text{G}}$ as the P x P realized covariance matrix of the genetic components $G = X\alpha$ of any P phenotypes ($\boldsymbol{\Omega}_{\text{G}} = \text{cov}(G) = \frac{G^TG}{N-1} = \frac{\alpha^TX^TX\alpha}{N-1} = \alpha^TS\alpha$), which is the main variable of interest for the estimation of the local genetic correlation. Because we are working with the sparser joint effects of the PCs δ (rather than the αs), which have the same covariance as G by a scaling factor K ($\boldsymbol{\Omega}_{\text{G}} = \alpha^TS\alpha = \alpha^TR^{+T}R^+\alpha = \delta^T\delta$, and hence $\boldsymbol{\Omega}_\delta = \text{cov}\left(\delta\right) = \frac{\delta^T\delta}{K} = \frac{\boldsymbol{\Omega}_G}{K}$), we use the $\boldsymbol{\Omega}_\delta = \text{cov}\left(\delta\right)$ instead. Because all the output is standardized, this makes no practical difference, as $\boldsymbol{\Omega}_{\text{G}}$ and $\boldsymbol{\Omega}_\delta$ have identical correlational structure. We will use $\boldsymbol{\Omega}$ to refer to $\boldsymbol{\Omega}_\delta$ henceforth.

This $\boldsymbol{\Omega}$ can be subdivided as:

$$\boldsymbol{\Omega} = \begin{pmatrix}\omega_1^2 & \cdots & \omega_{P1}\\ \vdots & \ddots & \vdots\\ \omega_{1P} & \cdots & \omega_P^2\end{pmatrix},$$

with each diagonal element ω_p^2 reflecting the (scaled) variance of the genetic component of phenotype p and each off-diagonal element ω_{pq} the (scaled) covariance of the genetic components for phenotypes p and q. We compute the corresponding bivariate local genetic correlations from the elements of this $\boldsymbol{\Omega}$ as $\rho_{pq} = \frac{\omega_{pq}}{\sqrt{\omega_p^2\omega_q^2}}$, with ρ_{pq}^2 representing the proportion of explained variance (that is, the local R^2).

For estimation of Ω, we note that the P×1 matrix of estimated joint effects of the PCs $\hat{\delta}_j$ are distributed as $\hat{\delta}_j \approx \text{MVN}\left(\boldsymbol{\delta}_j, \hat{\Sigma}\right)$, where $\hat{\Sigma}$ represents the sampling covariance matrix. We then use the Method of Moments[27] to estimate Ω as follows. With $\hat{\delta}_j \approx \text{MVN}\left(\boldsymbol{\delta}_j, \hat{\Sigma}\right)$, the expected value of $\hat{\delta}^T\hat{\delta}$ has the form $\text{E}\left[\hat{\delta}^T\hat{\delta}\right] = \delta^T\delta + K\hat{\Sigma} = K\left(\boldsymbol{\Omega} + \hat{\Sigma}\right)$, and hence $\boldsymbol{\Omega} = \frac{\text{E}\left[\hat{\delta}^T\hat{\delta}\right]}{K} - \hat{\Sigma}$. Plugging in the sample moments for $\text{E}\left[\hat{\delta}^T\hat{\delta}\right]$, we therefore obtain the estimator $\hat{\boldsymbol{\Omega}} = \frac{\hat{\delta}^T\hat{\delta}}{K} - \hat{\Sigma}$.

If there is no sample overlap, $\hat{\Sigma}$ is defined as $\text{diag}\left(\hat{\sigma}^2\right)$, where $\hat{\sigma}^2$ is a length P vector with the sampling variances of each phenotype. In the presence of possible sample overlap, estimates of the sampling correlation across phenotypes must be provided by the user. These can be obtained using cross-trait LDSC[13], creating a $P \times P$ covariance matrix with the intercepts for the genetic covariance for the off-diagonal elements (for the diagonal, use the intercept from a cross-trait analysis of a phenotype with itself or its univariate LDSC intercept). LAVA then converts this to a correlation matrix, C, and computes the sampling correlation matrix as $\hat{\Sigma} = \text{diag}\left(\hat{\sigma}\right) \times C \times \text{diag}\left(\hat{\sigma}\right)$.

Local multiple regression and partial correlations. Local conditional genetic associations between more than two phenotypes can be obtained using either multiple regression or partial correlation.

For the multiple regression, consider an outcome phenotype $\mathbf{Y}$ and set of predictor phenotypes X, with corresponding genetic components $\mathbf{G}_Y$ and G_X. Here, we can decompose $\mathbf{G}_Y$ as $\mathbf{G}_Y = G_X\gamma^{(r)} + \epsilon$ into a genetic component G_X and a residual component ϵ with $\text{cov}\left(G_X, \epsilon\right) = 0$, such that $\gamma^{(r)}$ reflects the vector of unstandardized regression coefficients. The variance of ϵ is denoted as $\tau^2_{(r)}$. Subdividing $\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_X & \boldsymbol{\Omega}_{XY} \\ \boldsymbol{\Omega}_{XY}^T & \omega_Y^2 \end{pmatrix}$, we can then compute $\gamma^{(r)} = \boldsymbol{\Omega}_X^{-1}\boldsymbol{\Omega}_{XY}$ and $\tau^2_{(r)} = \omega_Y^2 - \boldsymbol{\Omega}_{XY}^T\boldsymbol{\Omega}_X^{-1}\boldsymbol{\Omega}_{XY}$. Denoting the vector of standard deviations in $\boldsymbol{\Omega}_X$ as ω_X, we can then use these to obtain the standardized regression coefficients $\gamma = \frac{\text{diag}(\omega_X)}{\omega_Y}\gamma^{(r)}$ and standardized residual variance $\tau^2 = \frac{\tau^2_{(r)}}{\omega_Y^2}$. The corresponding explained variance for the full model is computed as $R^2 = 1 - \tau^2$.

The partial correlations between the genetic components of two phenotypes $\mathbf{X}$ and $\mathbf{Y}$, conditional on a set of other phenotypes Z (denoted $\rho_{XY|Z}$), can be expressed using the linear equations $\mathbf{G}_X = G_Z\beta_X + \epsilon_X$ and $\mathbf{G}_Y = G_Z\beta_Y + \epsilon_Y$, with $\rho_{XY|Z} = \text{cor}\left(\epsilon_X, \epsilon_Y\right)$. As with the parameters from the multiple regression, this can also be computed from the $\boldsymbol{\Omega}$ directly. Given the partial covariance $\omega_{XY|Z} = \omega_{XY} - \boldsymbol{\Omega}_{XZ}\boldsymbol{\Omega}_Z^{-1}\boldsymbol{\Omega}_{ZY}$ (with subscripts denoting the subset of relevant variances and covariances for $\mathbf{X}$ and $\mathbf{Y}$ and Z), and the partial variance $\omega^2_{X|Z} = \omega^2_X - \boldsymbol{\Omega}_{XZ}\boldsymbol{\Omega}_Z^{-1}\boldsymbol{\Omega}_{ZX}$ (and likewise for $\omega^2_{Y|Z}$), we can simply compute the partial correlation as $\rho_{XY|Z} = \frac{\omega_{XY|Z}}{\sqrt{\omega^2_{X|Z}\omega^2_{Y|Z}}}$.

Simulation *P* values and CIs. Because the sampling distributions for the local genetic correlation, partial correlation and multiple regression coefficients have no tractable closed form, we employ a simulation procedure with partial integration to obtain empirical P values for these parameters. Below, we denote the particular statistic being tested as T, with observed value T_{obs}.

First, we define a pure simulation approach, observing that the sufficient statistic $\hat{\delta}^T\hat{\delta}$ has a noncentral Wishart distribution with K degrees of freedom, scale matrix Σ and noncentrality matrix $K\boldsymbol{\Omega}$. For a statistic T, we can specify the $\boldsymbol{\Omega}_0$ corresponding to the null hypothesis being tested and use that to define the noncentrality matrix. We can then generate a random sample of null $\left(\hat{\delta}^T\hat{\delta}\right)^*$ matrices and for each of those compute the corresponding $\hat{\boldsymbol{\Omega}}^*$ and from there the statistic T^*. The sample of null T^* values can then be compared to the observed statistic T_{obs} to obtain an empirical P value, defined as the proportion of simulations for which T^* is more extreme than T_{obs}.

As empirical P values require a substantial number of simulations for sufficient accuracy at low P values, we augment the simulation procedure with a partial integration step as follows. For a single phenotype p, the distribution of $\hat{\delta}_p$ given $\hat{\delta}_{-p}$ is multivariate normal with parameters of known form, and consequently, many of the statistics of interest will have a normal distribution given $\hat{\delta}_{-p}$ (and $\boldsymbol{\Omega}_0$). We can therefore generate draws for $\left(\hat{\delta}^T\hat{\delta}\right)^*$ from the noncentral Wishart distribution, for each draw compute the parameters of the conditional distribution of the statistic T and then obtain the corresponding conditional P value for T_{obs} for that draw. We then compute the final P value as the mean of the conditional P values across draws.

Although the resulting P value is still empirical and subject to simulation uncertainty, with this procedure we can obtain sufficiently reliable P values even at very low value ranges without needing prohibitively many simulations. By default, LAVA performs 10,000 simulations to estimate the P value. This is increased this to 100,000 or 1,000,000 simulations if the P value estimate falls below 1×10^{-4} and 1×10^{-6}, respectively.

For a pair of phenotypes p and q, to test the null hypothesis of no local correlation, $H_0 : \omega_{pq} = \rho_{pq} = 0$, we use the local covariance estimate $\hat{\omega}_{pq}$ as the statistic T to test. For the integration step, to ensure symmetry, we use the conditional distribution of $\hat{\delta}_p$ given $\hat{\delta}_q$ for half of the simulations and the distribution of $\hat{\delta}_q$ given $\hat{\delta}_p$ for the other half. Similarly, to test the null hypothesis of no local partial correlation given a set of phenotypes Z, $H_0 : \omega_{pq|Z} = \rho_{pq|Z} = 0$, we use the local partial covariance estimate $\hat{\omega}_{pq|Z}$ as the statistic T. We use the conditional distribution of $\hat{\delta}_p$ given $\hat{\delta}_q$ and $\hat{\delta}_Z$ for half the simulations and the conditional distribution of $\hat{\delta}_q$ given $\hat{\delta}_p$ and $\hat{\delta}_Z$ for the other half.

For the regression model, with outcome phenotype $\mathbf{Y}$ and set of predictor phenotypes X, to test the null hypothesis of no conditional effect for predictor phenotype j, $H_0 : \gamma_j^{(r)} = \gamma_j^{(ss)} = \gamma_j = 0$, we use the semistandardized coefficient estimate $\hat{\gamma}_j^{(ss)}$ (standardized for X, but not $\mathbf{Y}$) as the statistic T. For the integration step, we use the conditional distribution of $\hat{\delta}_Y$ given $\hat{\delta}_X$.

Optionally, LAVA can also be requested to generate 95% CIs for the local correlation, partial correlation and standardized regression coefficients, as well as for the multiple R^2 parameter of the multiple regression model. These are computed by generating 10,000 draws from the noncentral Wishart distribution (with noncentrality matrix $K\hat{\boldsymbol{\Omega}}_{\text{obs}}$) and for statistic of interest T computing the simulated statistics T^* for all draws. The 2.5th and 97.5th percentiles of these T^* values are then used as estimates of the boundaries of the CI for T_{obs}.

Run times and memory usage. We ran additional analyses to evaluate run times and memory requirements of different analysis functions in LAVA. For these analyses we systematically varied the number of phenotypes from 1 to 25, and the locus size from 0.5 to 10 times the size of loci used for the main analyses here. See Supplementary Note and Supplementary Tables 157 and 158 for the results from these analyses, along with a more detailed overview.

Global r_g and estimation of sample overlap using LDSC. Bivariate LDSC[13] was used to evaluate the global r_gs between all phenotype pairs, as well as to obtain estimates of their level of sample overlap (as required for our LAVA analyses). To account for the sample overlap, we created a (symmetric) matrix based on the intercepts from the bivariate LDSC analyses (the diagonals populated by the intercepts from the bivariate analysis of each phenotype with itself). This was then converted to a correlation matrix and provided to LAVA. (An overview of how LAVA uses this information can be found in 'Estimating bivariate local genetic correlations'.) Summary statistics for each phenotype were munged using HapMap SNPs.

Statistics and reproducibility. Statistical analyses of the data were conducted using LAVA (as described above). The latest version can be downloaded from https://github.com/josefin-werme/LAVA, with the exact package version (v0.0.6) and scripts used to generate the main results of this paper deposited in https://github.com/josefin-werme/lava-scripts2021 (ref. [61]).

As this study relied on publicly available summary statistics (Table 1), sample sizes for the real data analyses were determined according to the maximum available. When per SNP imputation quality metrics were present, we filtered out any SNPs with an imputation quality ('info') score <0.9, and for the CAD summary statistics specifically, we excluded all exome chip markers due to the large sample size differences compared to the other SNPs. We also filtered out SNPs with a MAF <0.5% in the reference data, as simulations showed some risk of type 1 error rate inflation for partial correlation analysis with binary phenotypes (Supplementary Fig. 14). Note that LAVA also filters out any SNPs with strand-ambiguous alleles during processing, as these cannot be aligned.

For the simulations, the sample size was limited to 20,000 individuals to speed up processing times, achieving appropriate power by increasing the effect sizes instead (Supplementary Methods and Supplementary Note). To ensure that the set of simulation and analysis SNPs was identical, we set the alleles of strand-ambiguous SNPs in the 1000 Genomes .bim file to nonambiguous alleles to avoid these SNPs from being filtered out during analysis.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All analyses in this study relied on publicly available summary statistics that, in all but three cases, were downloaded from the GWAS Atlas[2] (https://atlas.ctglab.nl). Original sources and Atlas IDs (when applicable) are referenced in Table 1. As a reference for the estimation of LD, we used the European subset of the 1000 Genomes[26] data as downloaded from https://ctg.cncr.nl/software/magma. The locus file used for all the LAVA analyses can be accessed at https://github.com/josefin-werme/lava-scripts2021 (ref. [61]).

Code availability

LAVA is implemented as an R package, which is publicly available at the LAVA website (https://ctg.cncr.nl/software/lava) and LAVA GitHub repository (https://github.com/josefin-werme/LAVA). Analysis scripts and the exact package version (v0.0.6) used for the generation of the main results can be downloaded from https://github.com/josefin-werme/lava-scripts2021 (ref. [61]). The method used for genome partitioning is available at https://github.com/cadeleeuw/lava-partitioning (ref. [62]).

References

57. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
58. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
59. Agresti, A. *An Introduction to Categorical Data Analysis*. *Statistics in Medicine* (Wiley, 2007).
60. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
61. Werme, J. & de Leeuw, C. A. LAVA analysis scripts and package (v0.0.6). *Zenodo* https://doi.org/10.5281/ZENODO.5795207 (2021).
62. de Leeuw, C. A. LAVA partitioning algorithm (v1.0.0). Zenodo https://doi.org/10.5281/ZENODO.5583779 (2021).

Acknowledgements

This work was funded by COSYN (Comorbidity and Synapse Biology in Clinically Overlapping Psychiatric Disorders: Horizon 2020 Program of the European Union under RIA grant agreement 667301, D.P.) and the Netherlands Organization for Scientific Research (VICI 435-14-005, D.P.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The analyses were carried out on the Genetic Cluster Computer, which is financed by the Netherlands Organization for Scientific Research (480-05-003, D.P.), VU University (Amsterdam, the Netherlands) and the Dutch Brain Foundation, hosted by the Dutch National Computing and Networking Services SurfSARA.

Author contributions

J.W., S.v.d.S., D.P. and C.A.d.L. conceived of the study. J.W. and C.A.d.L. developed the statistical framework and implemented the software. J.W. performed the analyses, simulations and wrote the manuscript with contributions from C.A.d.L. J.W., S.v.d.S., D.P. and C.A.d.L. participated in the interpretation of the results and revision of the manuscript. All authors provided meaningful contributions at each stage of the project.

Competing interests

C.A.d.L. is funded by Hoffman-La Roche. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01017-y.

Correspondence and requests for materials should be addressed to Josefin Werme or Christiaan A. de Leeuw.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Josefin Werme, Christiaan de Leeuw

Last updated by author(s): Oct 26, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	The LAVA software used in this study is implemented as an R package which is publicly available at https://github.com/josefin-werme/LAVA (the IRLS algorithm was implemented directly into LAVA and did not rely on any pre-existing package). The exact package version (v0.0.6) and scripts used for the generation of the main results of this paper can be downloaded from https://github.com/josefin-werme/lava-scripts2021. The method used for genome partitioning is available at https://github.com/cadeleeuw/lava-partitioning (v1.0.0). All analyses were conducting in R v3.5.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All analyses in this study relied on publicly available summary statistics which, in all but three cases, were downloaded from the GWAS Atlas (https://atlas.ctglab.nl). Original sources and Atlas-IDs (when applicable) are referenced in Table 1. As a reference for the estimation of LD, we used the European subset of the 1,000 Genomes data as downloaded from https://ctg.cncr.nl/software/magma. The locus file used for all the LAVA analyses can be accessed at https://github.com/josefin-werme/lava-scripts2021.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	As this study relied on publicly available summary statistics, sample sizes for the real data analyses were determined according to the maximum that is publicly available. For the simulations, the sample size was limited to 20,000 individuals to optimise processing time while achieving appropriate power by increasing the effect sizes instead (see Supplementary Methods & Supplementary Note 5 for more detail)
Data exclusions	As this study relied on publicly available summary statistics, sample sizes for the real data analyses were determined according to the maximum that was available. When per SNP imputation quality metrics were present, we filtered out any SNPs with an INFO score < .9, and for the CAD summary statistics specifically, we also excluded all exome chip markers due to large sample size differences compared to the other SNPs. Finally, we filtered out all SNPs with a MAF of less than 0.5% in the reference data to prevent error rate inflation, as well as SNPs with strand ambiguous alleles since those cannot be aligned (the latter is done automatically by LAVA).
Replication	Not applicable as the study was primarily methodological in nature and the real data application served to demonstrate the method
Randomization	Not applicable as no data was gathered in this study
Blinding	Not applicable as no data was gathered in this study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging