Zhang et al. Genome Biology (2021) 22:262
https://doi.org/10.1186/s13059-021-02478-w

Genome Biology

METHOD Open Access

SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits

Yiliang Zhang[1†], Qiongshi Lu[2,3,4†], Yixuan Ye[5], Kunling Huang[3], Wei Liu[5], Yuchang Wu[2], Xiaoyuan Zhong[2], Boyang Li[1], Zhaolong Yu[5], Brittany G. Travers[6,7], Donna M. Werling[7,8], James J. Li[7,9] and Hongyu Zhao[1,5,10*]

* Correspondence: hongyu.zhao@yale.edu
†Yiliang Zhang and Qiongshi Lu contributed equally to this work.
[1]Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06520, USA
[5]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06510, USA
Full list of author information is available at the end of the article

Abstract

Local genetic correlation quantifies the genetic similarity of complex traits in specific genomic regions. However, accurate estimation of local genetic correlation remains challenging, due to linkage disequilibrium in local genomic regions and sample overlap across studies. We introduce SUPERGNOVA, a statistical framework to estimate local genetic correlations using summary statistics from genome-wide association studies. We demonstrate that SUPERGNOVA outperforms existing methods through simulations and analyses of 30 complex traits. In particular, we show that the positive yet paradoxical genetic correlation between autism spectrum disorder and cognitive performance could be explained by two etiologically distinct genetic signatures with bidirectional local genetic correlations.

Keywords: GWAS, Local genetic covariance, Eigen decomposition, Autism spectrum disorder, Chromatin modifiers

Background

Genome-wide association study (GWAS) has achieved remarkable success in the past 15 years and has identified numerous single-nucleotide polymorphisms (SNPs) associated with complex human traits and diseases [1]. Increasingly accessible summary statistics from GWAS, in conjunction with advances in analytical methods that use marginal association statistics as input, have circumvented logistical challenges in data sharing and greatly accelerated research in complex trait genetics [2].

With these advancements, multi-trait modeling has undergone rapid developments, leading to the emergence of numerous methods that study the shared genetic basis across multiple phenotypes [3–8]. Among these methods, genetic correlation analysis is a statistically powerful and biologically interpretable approach to quantifying the overall genetic similarity of two traits [9–15]. It has gained popularity in the field, provided new insights into the shared genetics of many phenotypes [10, 16], and has a variety of downstream applications [9]. Properly modeling genetic correlation could

enhance statistical power in genetic association studies [3, 4], improve risk prediction accuracy [17–19], and facilitate causal inference and mediation analysis [5, 7, 20–22]. A number of methods have been developed for genetic correlation estimation. Built upon the GREML approach [14, 23], cross-trait linkage disequilibrium (LD) score regression (LDSC) was the first method that uses GWAS summary statistics alone as input [10, 24]. Methods have also been developed to estimate annotation-stratified [12] and trans-ethnic [13] genetic correlation. Bioinformatics servers have been built to improve the computation and visualization of genetic correlations [25].

Local genetic correlation analysis is another important approach to tackling the underlying etiological mechanisms shared by multiple complex traits [11, 26]. Instead of estimating the average correlation of genetic effects across the genome, local genetic correlation quantifies the genetic similarity of two traits in specific genomic regions. This approach could reveal local, heterogeneous architecture of etiological sharing and is critical for understanding the heterogeneity in pleiotropic genetic effects. Existing methods have struggled to provide statistically principled and robust results due to technical challenges including extensive LD in local chromosomal regions and pervasive sample overlap across GWASs.

Here, we introduce a novel statistical framework named SUPERGNOVA for local genetic correlation estimation. Based on the GNOVA approach which was designed for partitioning genetic correlation by functional annotation [12], SUPERGNOVA is a principled framework for diverse types of genetic correlation analyses. Through extensive simulations, we demonstrate that SUPERGNOVA provides statistically rigorous and computationally efficient inference for both global and local genetic correlations and substantially outperforms existing methods when applied to local genomic regions. Additionally, our approach uses GWAS summary statistics alone as input and is robust to overlapping GWAS samples even when the shared sample size is unknown. We applied SUPERGNOVA to 30 complex traits and report 150 pairs of phenotypes with significant local genetic correlations. In particular, we investigated an empirical paradox—the robust, positive genetic correlation between autism spectrum disorder (ASD) and cognitive ability, which contradicts the comorbidity between ASD and intellectual disability [27]. We demonstrate that multiple distinct etiologic pathways contribute to the shared genetics between ASD and cognitive ability which could only be revealed by genetic correlation analysis at a local scale.

Results

Overview of SUPERGNOVA analytical framework

Genetic covariance (correlation) is defined as the covariance (correlation) of genetic effects on two traits. It is commonly used as an informative metric to quantify the shared genetic basis between traits. Given the marginal association statistics from two GWASs (i.e., z scores z_1 and z_2), genetic covariance ρ between two traits can be estimated by minimizing the "distance" between the empirical covariance of z scores, i.e., $\widehat{Cov}(z_1, z_2) = \frac{1}{2}(z_1 z_2^T + z_2 z_1^T)$, and the theoretical covariance

$$Cov(z_1, z_2) = \frac{\sqrt{n_1 n_2}\rho}{m} V^2 + \frac{n_s \rho_t}{\sqrt{n_1 n_2}} V \quad (1)$$

where m is the number of SNPs, n_1 and n_2 are the sample sizes of two GWASs, n_s is the number of individuals included in both studies, V is the LD matrix, and $\rho_t = \rho + \rho_e$ is the sum of genetic covariance (i.e., ρ) and the covariance of non-genetic effects (i.e., ρ_e) on the two traits among shared individuals. Derivation of the theoretical covariance and other statistical details are reported in the Additional file 1: Supplementary Note. In the "Methods" section, we show that with different definitions of "distance", existing methods such as LDSC [10] and GNOVA [12] are special cases of this unified framework.

Local genetic covariance (correlation) can be defined in a similar way by focusing only on SNPs in a pre-specified genomic region (the "Methods" section). Despite the conceptual similarity between global and local genetic correlation, local z scores from each GWAS can be highly correlated due to the extensive LD in local regions. Hence, most methods developed for global genetic correlation cannot be directly applied to estimate local correlations. In addition, ubiquitous sample overlap across GWASs introduces additional correlations among association statistics from different studies, which further complicates the estimation of genetic correlation. SUPERGNOVA resolves these statistical challenges by decorrelating local z scores with eigenvectors of the local LD matrix (Fig. 1). In practice, LD can be estimated from an external reference panel (e.g., 1000 Genomes Project [28]) and the independent LD blocks are determined by the LD patterns. For example, in our applications, we used LDetect [29] to partition the genome. Due to the noise in LD estimation, we only use the first K_i eigenvectors to transform and decorrelate association statistics in any given region i where K_i can be determined adaptively in SUPERGNOVA. After decorrelation, local genetic covariance ρ_i is estimated through a weighted least squares regression in each region. In contrast, LDSC directly applies weighted least squares on the correlated products of z scores. In the Additional file 1: Supplementary Note, we show that when the per-SNP heritability is small, SUPERGNOVA is equivalent to GNOVA which is a method that has been proven to achieve theoretical optimality compared to LDSC [12]. Another technical

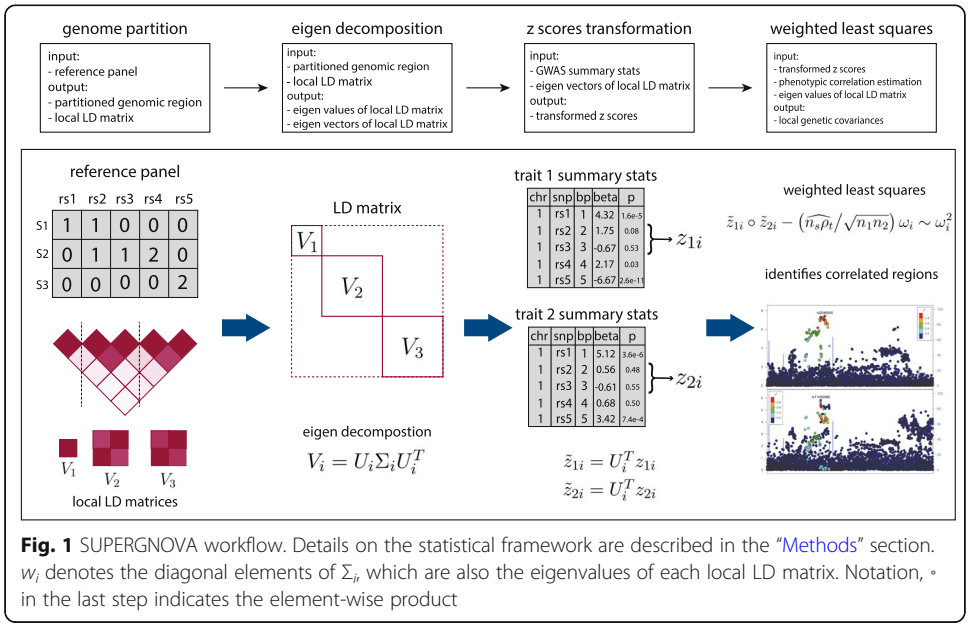

Fig. 1 SUPERGNOVA workflow. Details on the statistical framework are described in the "Methods" section. w_i denotes the diagonal elements of Σ_i, which are also the eigenvalues of each local LD matrix. Notation, $\circ$ in the last step indicates the element-wise product

challenge is that numerically unstable estimates of local heritability will lead to extreme variability in the estimates of local genetic correlation (Additional file 1: Supplementary Note; Additional file 2: Supplementary Figure 1). Therefore, we base our inference on local genetic covariance which is statistically equivalent. We discuss more statistical details in the "Methods" section.

Simulations

We performed simulations to assess the performance of SUPERGNOVA for both global and local genetic correlation analyses. We compared SUPERGNOVA with multiple state-of-the-art methods in six different simulation settings and repeated each setting 100 times. We used real genotype data from the Wellcome Trust Case Control Consortium (WTCCC) to simulate quantitative traits. After quality control, 15,918 samples and 287,539 SNPs remained in the dataset. We equally divided 15,918 samples into two subsets which we denote as set 1 and set 2. To assess the robustness of our approach to sample overlap between GWASs, we generated another dataset by combining 3979 samples from set 1 and 3980 samples from set 2. We refer to it as set 3. This results in a 50% sample overlap between set 1 and set 3. Detailed simulation settings and quality control procedures are described in the "Methods" section.

We compared the performance of LDSC, GNOVA, and SUPERGNOVA on global genetic covariance estimation. We set the heritability to be fixed at 0.5 and the genetic covariance to range from 0 to 0.25. The covariance of non-genetic effects was 0.2 for the overlapped samples. The effect sizes of SNPs were generated from a multivariate normal distribution. Both SUPERGNOVA and GNOVA showed superior statistical power compared to LDSC in all settings (Fig. 2A–C). No method showed inflated type I error rates when the true covariance was 0. All three approaches provided unbiased estimates for global genetic covariance but LDSC estimates had substantially larger variance compared to GNOVA and SUPERGNOVA (Additional file 2: Supplementary Figures 2-4).

Next, we compared ρ-HESS and SUPERGNOVA on their performance of estimating local genetic covariance. We used 395 SNPs from a genomic region of about 3.3 Mb on chromosome 2 as the local region of interest. The remaining SNPs on chromosome 2 (23,839 SNPs) were used as the "background SNPs" in the analysis. We set the covariance in the small local region to be from 0 to 0.005. Outside of this region on chromosome 2, covariance was fixed as 0. The total heritability was set to be 0.5 and was equally distributed among all SNPs on chromosome 2 (24,234 SNPs). Both SUPERGNOVA and ρ-HESS assume the SNPs in different regions to be independent. However, in practice, there can be weak LD between nearby regions which could bias the estimates towards the average genetic covariance of adjacent regions. When there is no overlapping sample between two studies, SUPERGNOVA estimates showed lower bias, well-controlled type I error, and good statistical power. On the other hand, ρ-HESS consistently underestimated local genetic covariance and had lower statistical power (Fig. 2D; Additional file 2: Supplementary Figure 5).

We repeated these simulations in set 1 and set 3 with a 50% sample overlap. SUPERGNOVA estimates of local genetic covariance remained less biased with well-controlled type I error (Additional file 2: Supplementary Figure 6). Compared

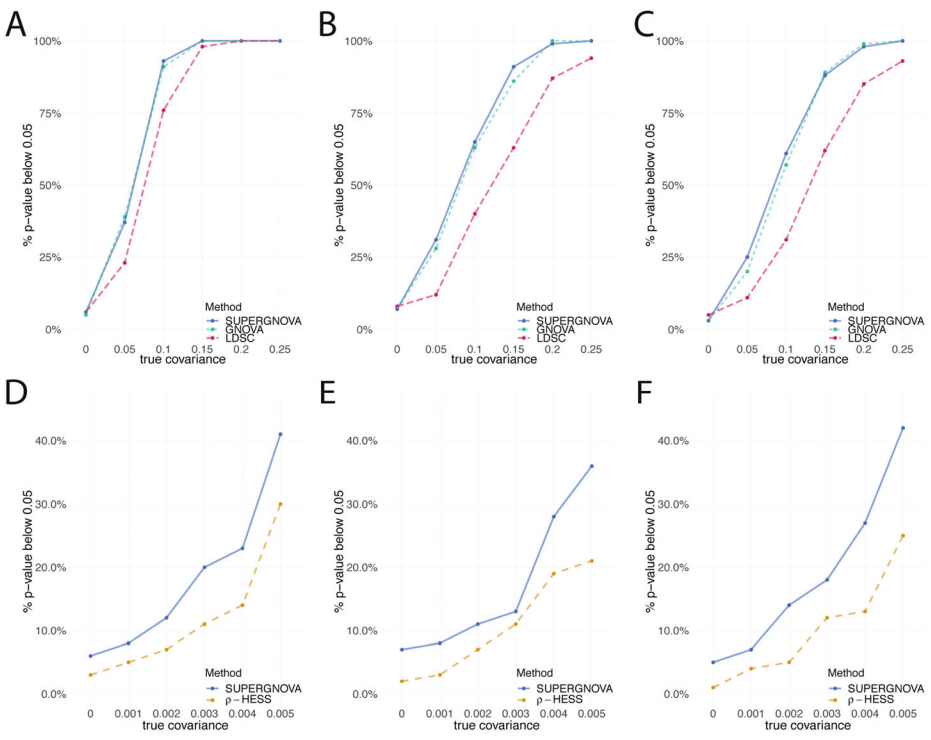

Fig. 2 Simulation results. **A**–**C** compare the type-I error and statistical power of SUPERGNOVA, GNOVA, and LDSC in global genetic covariance estimation. We use the proportion of *p* values that are less than 0.05 to estimate type I error or statistical power when true parameters are zero or nonzero, respectively. **A** Two GWASs were simulated on two non-overlapping datasets (set 1 and set 2). **B** GWASs were simulated on two datasets with a 50% sample overlap (set 1 and set 3). **C** Two GWASs were both simulated on the same dataset (set 1) with a 100% sample overlap. **D**–**F** compare SUPERGNOVA and *ρ*-HESS on local genetic covariance estimation using GWASs **D** without sample overlap (set 1 and set 2), **E** with a partial sample overlap (set 1 and set 3), and **F** with a complete sample overlap (set 1 only)

to *ρ*-HESS, SUPERGNOVA showed superior statistical power (Fig. 2E). *ρ*-HESS underestimated local genetic covariance even though we provided the correct sample size for shared samples (Additional file 2: Supplementary Figure 6). We also performed simulations under a complete sample overlap by simulating two traits on set 1. SUPERGNOVA still achieved less biased estimation and valid inference (Fig. 2F; Additional file 2: Supplementary Figure 7). *ρ*-HESS lacked statistical power in all settings. Additionally, when provided with inaccurate values of the overlapping sample, *ρ*-HESS showed even lower statistical power (Additional file 2: Supplementary Figure 8-9). We note that SUPERGNOVA does not need the shared sample size or phenotypic correlation as input.

Finally, we repeated the simulations on densely imputed genotype data from the UK Biobank (UKBB) and further evaluated the robustness of SUPERGNOVA under a set of mis-specified models, including models with sparse genetic architecture and effects dependent on minor allele frequencies (MAF) and LD. We also assessed how the size of local genomic regions affects the performance of SUPERGNOVA and the type-I inflation of LDSC on local genetic covariance estimation. These additional simulations showed highly consistent results. We describe the settings and results of these simulations in the Additional file 1: Supplementary Note and Additional file 2: Supplementary Figures 10-15.

Global and local genetic correlations among 30 complex traits

We applied SUPERGNOVA to estimate local and global genetic correlations among 30 phenotypes (Additional file 3: Supplementary Table 1). We partitioned the genome into 2353 approximately independent regions (about 1.6 centimorgan on average) using LDetect [29], with LD estimated from the 1000 Genomes Project phase III samples of European ancestry [28]. One hundred twenty-seven pairs of traits were globally correlated ($p < 0.05/435 = 1.1e{-}4$; Additional file 2: Supplementary Figure 16) and 150 pairs of traits were locally correlated in 109 different regions under Bonferroni correction ($p < 0.05/1{,}006{,}072 = 5.0e{-}8$; Fig. 3A; Additional file 3: Supplementary Tables 2-3). All significant regions had at least one SNP with $p < 1e{-}4$ in both GWASs.

The sums of local covariance across 2353 regions were highly concordant with the estimated global genetic covariance (Fig. 3B; $R^2 = 0.99$), but local genetic covariance revealed diverse architecture of genetic sharing locally. We estimated the proportion of correlated regions for each pair of traits using ashr [30] (the "Methods" section; Fig. 3A; Additional file 3: Supplementary Table 4). The proportion of correlated regions predicted global genetic correlation in general, with some notable outlier trait pairs (Fig. 3C). Two subtypes of inflammatory bowel disease (IBD), Crohn's disease and ulcerative colitis (UC), had strong pairwise global correlations but relatively sparse local

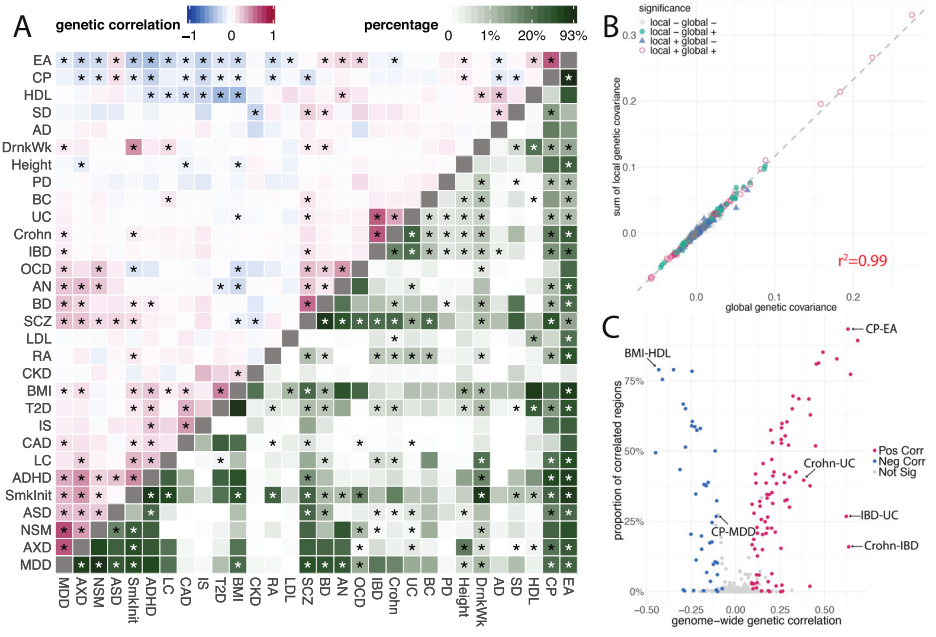

Fig. 3 Global and local genetic correlations among 30 complex traits. **A** Estimates of global genetic correlations (upper triangle) and estimated proportions of correlated regions among 435 trait pairs (lower triangle). Asterisks in the upper triangle highlight significant genetic correlations after Bonferroni correction for 435 pairs. Asterisks in the lower triangle indicate at least one significantly correlated region between the traits after Bonferroni correction for all 1,006,072 regions in 435 trait pairs. We grouped traits with hierarchical clustering applied to global genetic correlations. We summarized detailed information about each trait, including abbreviations, in Additional file 3: Supplementary Table 1. **B** Global genetic covariance estimates were highly concordant with the sums of local genetic covariance. Each point represents a trait pair. Color and shape of each data point denote the significance status in global and local correlation analyses. **C** Volcano plot comparing the global genetic correlation and proportion of correlated local regions. Each point represents a trait pair. Color of each data point represents the significance and direction of global correlation. Trait pairs discussed in the main text are labeled in the plot

genetic correlations (Additional file 2: Supplementary Figure 17). In fact, all 8 identified regions were positively correlated among Crohn's disease, UC, and IBD and harbored genome-wide significant loci reported in the GWAS on IBD [31], suggesting that a limited fraction of the genome contribute to different subtypes of IBD with strong and concordant effects. In contrast, SNPs in 93% of regions had correlated effects between cognitive performance (CP) and educational attainment (EA; global genetic correlation = 0.63; p = 6.1e−115), the highest among all trait pairs. Seventy-nine percent of regions showed correlated effects between body mass index (BMI) and high-density lipoprotein (HDL) cholesterol (global genetic correlation = − 0.43; p = 3.6e−41), the highest among negatively correlated traits. These results suggest extensive and "omnigenic" genetic sharing between these traits, which is also reflected in the substantial shift in the distribution of local genetic covariances (Additional file 2: Supplementary Figure 17). Bidirectional correlations were also observed in several trait pairs, including ASD and CP [32] (Additional file 2: Supplementary Figure 17; global correlation = 0.15; 15% of regions were correlated). Across all trait pairs, we observed a modest association between the sample size of traits and the proportion of correlated regions (Additional file 2: Supplementary Figure 18).

We identified significant local genetic covariance for 86 trait pairs that were not significantly correlated in the global analysis (Additional file 2: Supplementary Figure 19; Additional file 3: Supplementary Table 2-3), including HDL cholesterol and low-density lipoprotein (LDL) cholesterol [11], CP and major depressive disorder (MDD), obsessive-compulsive disorder (OCD) and anxiety disorder (AXD), and ASD and bipolar disorder (BD).

Our analyses also implicated several genomic regions showing correlated genetic effects on more than two traits. The *BDNF* locus on chromosome 11 (hg19 coordinate: 27,019,873–28,741,185) is known to control the development of neurons and synapses and is vital to learning, memory, and vulnerability to stress [33–36]. We identified significant genetic covariance at this locus between 6 trait pairs among schizophrenia (SCZ), EA, smoking initiation (SmkInit), drinks per week (DrnkWk), and attention deficit/hyperactivity disorder (ADHD) (Additional file 2: Supplementary Figure 20-21). Another locus on chromosome 11 (111,985,737–113,103,996) was identified among 7 neuropsychiatric traits: anorexia nervosa (AN), BD, MDD, CP, SCZ, SmkInit, and neuroticism (NSM) (Additional file 2: Supplementary Figure 22-23). *NCAM1* at this locus is involved in development and maintenance of the nervous system and is associated with SCZ and comorbid alcohol and drug dependence [37–39]. These hub regions with pervasive correlations among psychiatric disorders hint at key regulators in the nervous system and provide guidance to functional genomic studies that interrogate the mechanisms of pleiotropic effects [40].

Local genetic covariance that did not achieve statistical significance may still be worth follow-up investigations. Despite evidence on phenotypic correlations, previous studies have suggested that Alzheimer's disease (AD) is not genetically correlated with neuropsychiatric traits except education and cognition [16]. We identified suggestive local correlations of AD with 7 neuropsychiatric traits: NSM (p = 1.2e−6), OCD (p = 3.0e−6), CP (p = 3.4e−6), DrnkWk (p = 2.0e−5), MDD (p = 2.7e−5), and AN (p = 4.2e−4), at the *SPI1* locus (chr11: 46,876,411–48,200,127). We replicated the local correlations with DrnkWk (p = 7.5e−2) and NSM (p = 1.4e−2) using an independent GWAS of AD

family history (the "Methods" section; Additional file 3: Supplementary Tables 5-6). The estimates for local genetic covariance were highly consistent between two analyses (R^2 = 0.84; Additional file 2: Supplementary Figure 24). The *SPI1* locus has been consistently identified in AD GWASs [41, 42]. A recent genome-wide survival study of AD onset convincingly demonstrated that transcription factor (TF) PU.1 encoded by *SPI1* is a key regulator for the development and function of myeloid cells and lower *SPI1* expression delays the onset of AD by regulating gene expression in myeloid cells [43]. However, genetic covariance of AD with DrnkWk and NSM was not statistically significant in TF binding sites of PU.1 in macrophages and monocytes (the "Methods" section; Additional file 3: Supplementary Table 7). Transcriptome-wide association study (TWAS) identified a number of PU.1-regulated genes associated with these phenotypes in macrophages and monocytes (the "Methods" section; Additional file 3: Supplementary Tables 8-9), but all the genes shared by multiple traits are located at the *SPI1* locus (Additional file 2: Supplementary Figure 24). These results suggest that although the *SPI1* locus may have correlated roles in multiple psychiatric and neurodegenerative diseases, PU.1 may modulate the risk of these diseases through regulating the transcription of distinct susceptible genes in myeloid cells.

Dissecting the shared genetic basis of ASD and cognitive ability

We further demonstrate the power of SUPERGNOVA through an in-depth case study of the shared genetics between ASD and cognitive ability (Additional file 2: Supplementary Figure 25). Paradoxically, previous studies based on multiple different approaches have found a positive genetic correlation between ASD and CP [16, 44, 45]. We also identified significant positive global genetic correlations between ASD and measures of cognitive ability (Fig. 3), e.g., CP (standardized score on neuropsychological tests; correlation = 0.15, p = 3.2e−8) and EA (years of schooling; correlation = 0.18, p = 3.8e−14). Cognitive phenotypes in these GWASs have been previously described in detail [46]. However, such a positive correlation contradicts the known comorbidity of intellectual disability and ASD with regard to de novo variants of high penetrance [27, 47]. In addition, other neurodevelopmental disorders such as ADHD showed negative genetic correlations with cognitive measures (correlation = − 0.29 and p = 2.9e−29 with CP; correlation = − 0.41 and p = 2.0e−59 with EA), but the genetic correlation between ASD and ADHD was positive (correlation = 0.28, p = 2.3e−9).

A total of 64 genomic regions with significant local genetic covariance were identified among ADHD, ASD, and CP at a false discovery rate (FDR) cutoff of 0.1 (Additional file 2: Supplementary Figure 26; Additional file 3: Supplementary Table 10). The local covariances of CP with ASD and ADHD were bidirectional. No region with a negative covariance between ASD and ADHD was identified. The paradox that ASD and ADHD show opposite correlations with CP was not observed in any local region (Fig. 4A; Additional file 3: Supplementary Table 11). Eighteen regions showed significant positive correlations between ASD and CP, among which 3 regions were also significant and positive between ADHD and ASD and 2 regions were significant and positive between ADHD and CP. Similarly, we identified 32 regions with significant negative correlations between ADHD and CP. Among these regions, 3 were positive between ADHD and ASD and 3 were negative between ASD and CP. Three regions reached statistical

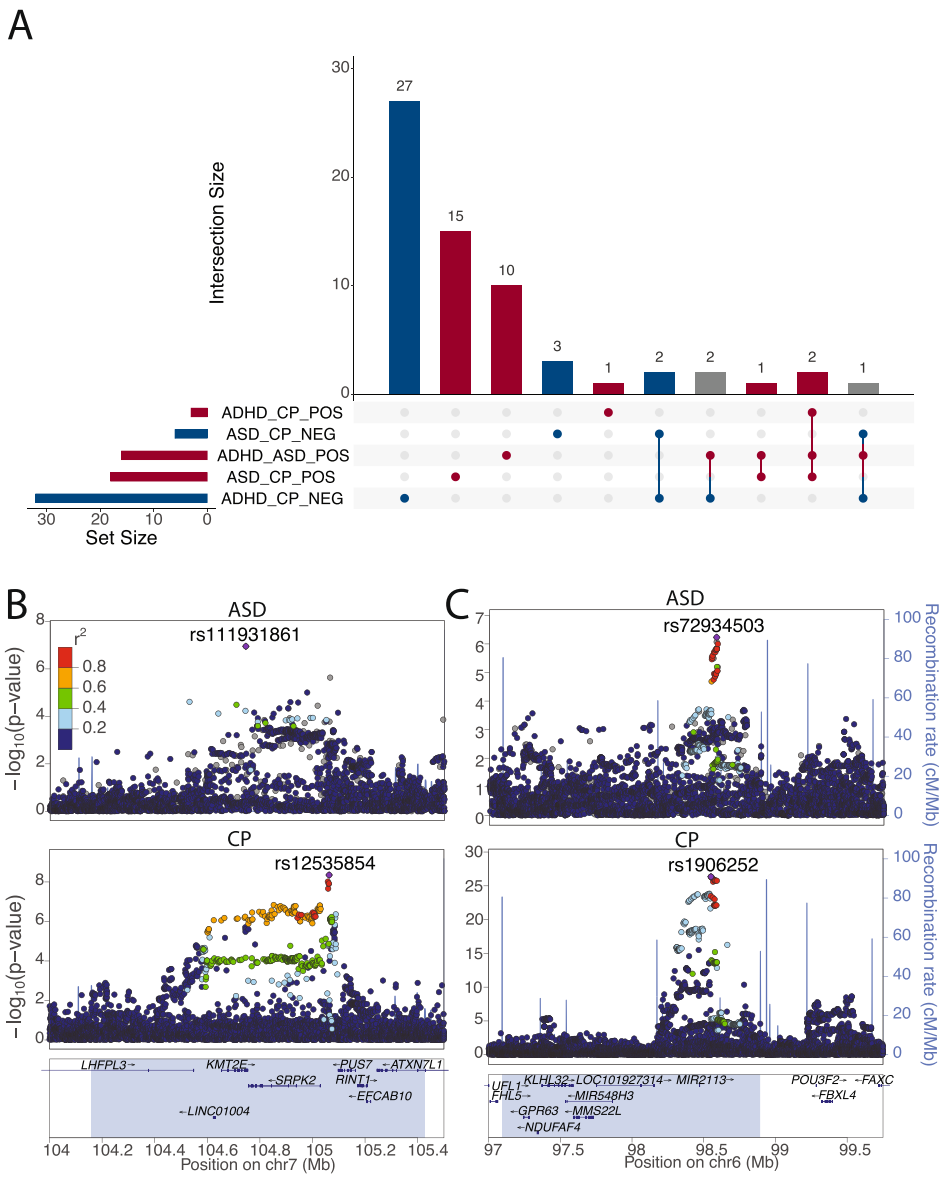

Fig. 4 Bidirectional local genetic covariance between ASD and CP. **A** Regions with significant local genetic covariance among ADHD, ASD, and CP (FDR < 0.1). This plot uses bars to break down the Venn diagram of overlapped regions in different categories. The five categories shown in the lower panel are correlated regions of ADHD and CP (positive and negative), ASD and CP (positive and negative), and ASD and ADHD (positive only). We use different colors (red, blue, and gray) to annotate region categories of positive, negative, and mixed covariance directions. **B** LocusZoom plots for ASD and CP GWAS associations at the *KMT2E* locus. ASD and CP are negatively correlated in the highlighted region. **C** LocusZoom plots for ASD and CP at the *POU3F2* locus. ASD and CP are positively correlated in the highlighted region. *POU3F2* is 700 kb downstream of the GWAS association peak

significance in all three trait pairs. ASD and ADHD were positively correlated in all three regions. ASD and ADHD were both positively correlated with CP in the regions on chromosomes 4 (150,634,191–153,226,998) and 14 (36,683,516-38,481,516) (Additional file 2: Supplementary Figures 27-28) and were both negatively correlated with CP in the region on chromosome 7 (104,158,491–105,425,027) (Fig. 4B; Additional file 2: Supplementary Figure 29).

The locus on chromosome 7 (104,158,491–105,425,027) showed significant and negative correlations between CP and both neurodevelopmental disorders (Additional file 3: Supplementary Table 11). We also identified a significant negative correlation of CP and SCZ in this region (p = 1.8e−8; Additional file 2: Supplementary Figure 29; Additional file 3: Supplementary Table 12). Among genes at this locus, *PUS7* is associated with intellectual disability and neurological defects [48]. De novo mutations in *KMT2E* cause a spectrum of neurodevelopmental disorders including ASD [49]. An intronic SNP in *KMT2E*, rs111931861, with a MAF of 0.034, reached genome-wide significance in a recent ASD GWAS [44] (Fig. 4B). *KMT2E* was also implicated by a recent exome sequencing study [50]. It is the only gene that reached genome-wide significance in both GWAS and exome sequencing studies of ASD. TWAS did not identify any genes associated with ADHD, ASD, SCZ, or CP in this region (the "Methods" section; Additional file 3: Supplementary Tables 13). These results, coupled with the findings about de novo and ultra-rare variants in *KMT2E*, suggest that common variants in this region may be tagging protein-altering variants instead of regulatory variants for transcriptional activities. A missense SNP in *KMT2E*, rs117986340, was nominally associated with ASD (p = 5.7e−2) and ADHD (p = 4.4e−2) in GWAS (the "Methods" section; Additional file 3: Supplementary Table 14) but this hypothesis needs to be investigated in the future using sequencing data.

POU3F2 (also known as *BRN2*) is a key TF in the central nervous system and a master regulator of gene expression changes in BD and SCZ [51, 52]. It is the first genome-wide significant locus identified for EA [53]. It has also been identified in a recent TWAS for ASD [54]. In our analysis, the *POU3F2* locus on chromosome 6 (97,093,295-98,893,182) showed significant positive correlations between ASD and CP (p = 1.8e−5; Fig. 4C) and among many neuropsychiatric phenotypes including AN, BD, DrnkWk, EA, and SmkInit (Additional file 2: Supplementary Figures 30-31; Additional file 3: Supplementary Table 12). In addition, genes in other regions showing nominal negative correlations between ASD and CP were significantly enriched for *POU3F2* protein-protein interactors (PPIs) (odds ratio = 24.8; p = 2.8e−3; the "Methods" section). This is consistent with our recent finding that genes regulated by TF *POU3F2* showed a 2.7-fold enrichment for loss-of-function de novo mutations in ASD probands which are known to cause comorbid intellectual disability [54]. These results hint at a pervasive, regulatory role of *POU3F2* in cognitive ability and many neuropsychiatric disorders [55, 56].

Regions showing opposite correlation directions between ASD and CP were enriched for distinct mechanistic pathways (the "Methods" section; Fig. 5; Additional file 3: Supplementary Tables 15-18). Genomic regions with negative correlations between ASD and CP were significantly enriched for chromatin modifier genes (enrichment = 3.2; p = 3.8e−4; Additional file 3: Supplementary Tables 15-16). De novo protein-truncating mutations in these genes are known to cause ASD, intellectual disability, and a variety of congenital anomalies [27, 57, 58]. Regions positively correlated between ASD and CP were significantly enriched for postsynaptic density (PSD) proteins (enrichment = 1.8; p = 3.5e−4; Additional file 3: Supplementary Tables 17-18). *FMRP* targets also showed a significant enrichment in positively correlated regions (enrichment = 1.9; p value = 2.7e−3; Additional file 3: Supplementary Table 17). The enrichment of *FMRP* targets in negatively correlated regions was comparable but did not reach statistical significance

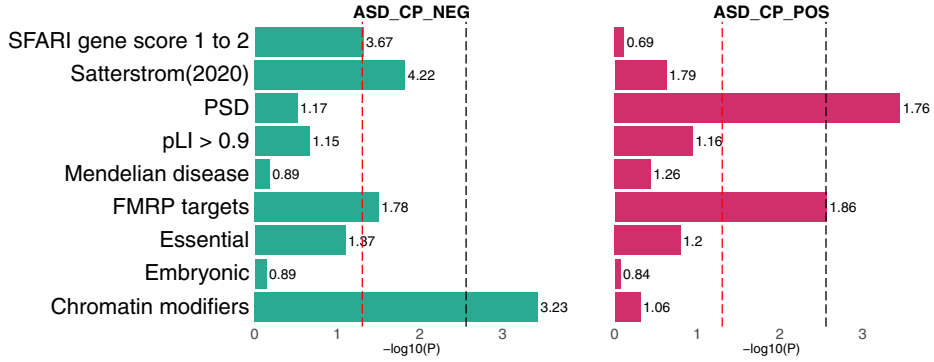

Fig. 5 Enrichment for gene sets in correlated regions between ASD and CP. Regions with opposite correlations between ASD and CP were enriched for different mechanistic pathways. Fold enrichment values are labeled next to each bar. The red dashed lines mark the *p* value cutoff of 0.05, and the black dashed lines denote the *p* value thresholds after Bonferroni correction ($p = 2.8e{-}3$)

after multiple testing correction (enrichment = 1.8; p value = 0.032; Additional file 3: Supplementary Tables 15). PSD genes are known to be enriched for associations identified in ASD TWAS [54]. *FMRP* targets are enriched for both ASD heritability quantified using common variants [59] and de novo mutations of ASD [60, 61]. *FMRP* target genes showed a 12.4-fold enrichment ($p = 3.5e{-}15$) in the 102 risk genes identified in the latest exome sequencing study of ASD [50]. Notably, findings from exome-sequencing studies (e.g., the 102 ASD genes [50]) and gene sets known to be enriched for ultra-rare or de novo protein-truncating variants in ASD probands (e.g., chromatin modifiers [27]) showed substantially stronger enrichment in the regions with negative ASD-CP correlations than the regions with positive correlations ($p = 0.034$, the "Methods" section; Additional file 3: Supplementary Tables 15-18).

We then assessed the enrichment of associations for other complex traits in genetically correlated regions between ASD and CP (the "Methods" section). Regions with positive correlations between ASD and CP were significantly enriched for associations for 10 traits documented in GWAS Catalog ($p < 0.05/664 = 7.5e{-}5$), including extremely high intelligence (odds ratio = 9.7; $p = 3.5e{-}9$), household income (odds ratio = 52.6; adjusted $p = 5.7e{-}9$), and loneliness (odds ratio = 5.5; $p = 4.1e{-}5$) (Additional file 2: Supplementary Figure 32; Additional file 3: Supplementary Table 19). Negatively correlated regions were enriched for associations with a variety of neurodevelopmental and psychiatric disorders including SCZ (odds ratio = 6.1; $p = 2.2e{-}31$), BD (odds ratio = 13.3; $p = 2.7e{-}24$), and NSM (odds ratio = 10.0; $p = 6.0e{-}12$) (Additional file 2: Supplementary Figure 32; Additional file 3: Supplementary Table 20). We also estimated stratified genetic covariance of 28 other traits with ASD and CP in these identified regions (Additional file 2: Supplementary Figure 33). EA, MDD, and rheumatoid arthritis (RA) showed significant stratified covariance with ASD or CP ($p < 0.05/112 = 4.5e{-}4$) in regions positively correlated between ASD and CP (Additional file 3: Supplementary Table 21). On the other hand, ADHD, EA, and AXD showed significant stratified covariance with ASD or CP in regions showing significant negative correlations between ASD and CP (Additional file 3: Supplementary Table 22). Overall, traits showed the same directions of covariances with ASD and CP in regions with positive ASD-CP covariances, while they showed opposite directions of genetic covariances with ASD and

CP in regions with negative ASD-CP covariances (Additional file 2: Supplementary Figure 33). In other words, no paradoxical covariances were present when we zoomed in by ASD-CP correlated regions.

Genes in positively correlated regions of ASD and CP were expressed in a substantially higher proportion of cells in fetal brains compared to background genes (p = 0.012; log-rank test) (the "Methods" section; Additional file 2: Supplementary Figure 34; Additional file 3: Supplementary Table 23) while the elevation of gene expression rate in negatively correlated regions was not significant (p = 0.15, log-rank test). We did not identify a significant difference in the expression rate between genes in the ASD-CP positively correlated regions and genes in the ASD-CP negatively correlated regions (p = 0.71, log-rank test). The average expression of both gene sets was significantly higher than background genes across prenatal and postnatal stages (p = 9.7e−525 and 2.5e−99 for genes in positively and negatively correlated regions, respectively) (the "Methods" section; Additional file 2: Supplementary Figure 35). We also identified significantly higher expression of genes in positively correlated regions than in negatively correlated regions across developmental stages (p = 1.92e−61; Additional file 2: Supplementary Figure 35). We did not identify differential expression between prenatal and postnatal brains for either gene set (p = 0.83 and 0.81; the "Methods" section; Additional file 3: Supplementary Table 24).

These results hinted that different pathways and biological processes were underlying the positive and negative genetic correlation of ASD and cognitive ability. We further investigated if these two sets of genetic components were associated with different clinical symptoms and subtypes of ASD. We constructed two polygenic risk scores (PRSs) of ASD based on independent SNPs from genomic regions with positive and negative local correlations between ASD and CP, respectively, for 5469 ASD probands and 2132 healthy siblings in the Simons Foundation Powering Autism Research for Knowledge (SPARK) cohort (the "Methods" section). We refer to these scores as PRS+ and PRS−. Both PRS+ and PRS− are normally distributed in SPARK (Additional file 2: Supplementary Figure 36). PRS+ could significantly distinguish ASD probands and healthy siblings (odds ratio = 1.08; p = 0.026) while the association between PRS- and ASD status was not significant (odds ratio = 1.02; p = 0.71; the "Methods" section). One thousand eight hundred three probands had both genotype data and intelligence quotient (IQ) information. Probands with high PRS+ had higher IQ compared to probands with high PRS-, with the average IQ changing sharply in the right tail of the PRS distribution, from 93.8 and 94.7 (p = 0.64; two-sample t-test) in the 75% percentile to 101.7 and 84.0 (p = 0.046; two-sample t test) in the 99% percentile (Fig. 6A; Additional file 2: Supplementary Figure 37). The proband subgroups above the 99% percentiles of PRS+ and PRS- did not have overlapping samples. 10.5% of probands above the 99% percentile of PRS+ and 31.6% of probands above the 99% percentile of PRS− had an IQ below 70 (Figs. 6A, B). Four probands above the 99% percentile of PRS- had relatively high PRS+ (greater than the 90% percentile of PRS+). All of them had IQ > 70 (Fig. 6B). No proband in the 99% percentile group of PRS+ had high PRS− (greater than the 90% percentile of PRS−) (Fig. 6C).

Four thousand two hundred sixty-seven probands in SPARK had genotype data and social communication questionnaire (SCQ) scores (the "Methods" section). We used SCQ score as a proxy for ASD symptom severity. Probands with PRS− above the 99%

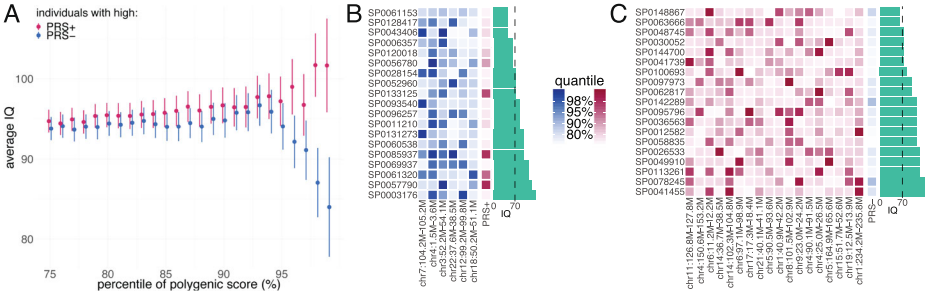

Fig. 6 Phenotypic heterogeneity of ASD probands with high PRS+ and PRS−. **A** Average IQ is computed for different groups defined by PRS. Each interval indicates standard error of the estimated mean. **B** PRS percentiles and IQ of probands above the 99% percentile of PRS−. PRS− was calculated using six negatively correlated regions between ASD and CP. The blue heatmap indicates the percentile of ASD PRS in each contributed region for each proband. The percentiles of PRS+ values are shown in the red boxes. IQ is shown as green bars. **C** PRS percentiles and IQ of probands above the 99% percentile of PRS+. PRS+ was calculated using 18 positively correlated regions between ASD and CP and the per-locus percentiles are shown in red. The percentiles of PRS+ are shown in blue and the green bars denote IQ

percentile showed significantly elevated SCQ scores compared to other probands (p = 0.03; two-sample t test) (Additional file 3: Supplementary Table 25). Average SCQ score rose from 22.4 in the 75% percentile to 24.4 in the 99% percentile (Additional file 2: Supplementary Figure 38). We did not identify a significant elevation in probands with PRS+ above 99% percentile (p = 0.56). The repetitive behaviors scale-revised (RBS-R) questionnaire was used to quantify repetitive behaviors, including self-injuries, restricted behavior, compulsive behavior, stereotyped behavior, ritualistic behavior, and sameness behavior [62]. We observed a significant increase of RBS-R scores in probands with PRS+ above the 99% percentile (p = 0.016; Additional file 2: Supplementary Figure 38) but not in probands with PRS− above the 99% percentile (p = 0.4; Additional file 3: Supplementary Table 25). We also investigated motor ability quantified by the developmental coordination disorder questionnaire [63, 64] (DCDQ) in SPARK. We observed a downward trend of DCDQ score (i.e., worse motor ability) as PRS increases (Additional file 2: Supplementary Figure 38), but the changes were not statistically significant (Additional file 3: Supplementary Table 25). Follow-up analyses examining RBS-R and DCDQ subscales found that the pattern of results was not driven by any one of the subdomains. Finally, we assessed the enrichment of ASD subtypes in PRS+ and PRS− 99% percentile groups. No subtype reached statistical significance (Additional file 3: Supplementary Table 26), with Asperger's disorder showing the strongest yet modest enrichment (enrichment = 1.58; p = 0.082) in probands with PRS+ above 99% percentile (Additional file 2: Supplementary Figure 39). We note that these identified associations only achieved suggestive statistical evidence after accounting for multiple testing and need to be validated in the future using larger samples.

Discussion

Owing to increasingly accessible GWAS summary statistics and advances in statistical methods to directly model summary-level data, genetic correlation estimation, especially at the genome-wide scale, has become a routine procedure in post-GWAS analyses. These correlation estimates effectively summarize the complex etiologic sharing of multiple traits into concise, robust, and interpretable values, which provided novel

insights into the shared genetic architecture of a spectrum of phenotypes. However, genome-wide genetic correlations only reflect the average concordance of genetic effects across the genome and often fail to reveal the local, heterogeneous pleiotropic effects, especially when the underlying genetic basis involves multiple etiologic pathways. To this end, methods that partition genetic covariance by functional annotation or local genetic region have achieved some success [11, 12]. These methods generally use more sophisticated statistical models and are more adaptive to diverse types of shared genetic architecture. On the downside, it is statistically more challenging to estimate all the parameters in these models using GWAS summary statistics alone. The problem is further exacerbated by technical issues such as strong LD among SNPs in local regions and sample overlap across different GWASs. Due to these challenges, stratified genetic correlation analysis has not been as popular as its genome-wide counterpart.

In this paper, we have introduced SUPERGNOVA, a unified framework for both genome-wide and stratified genetic correlation analysis. Improved upon our previous work [12], SUPERGNOVA directly addresses the technical challenges in local genetic correlation inference while retaining the statistical optimality in analyses at the genome-wide scale. Through extensive simulations, we demonstrated that SUPERGNOVA provides statistically robust and efficient estimates and substantially outperforms other methods in estimation accuracy and statistical power. Notably, SUPERGNOVA uses GWAS summary statistics as the input and is robust to arbitrary sample overlap between GWAS datasets.

Applied to 30 complex traits, SUPERGNOVA identified 150 trait pairs with significant local genetic covariance, including 86 pairs without a significant global correlation. We identified various patterns in the shared genetic architecture between traits, with some traits (e.g., EA and CP) showing ubiquitous genetic covariance in a large fraction of the genome and other traits (e.g., Crohn's disease and UC) showing relatively sparse genetic sharing with strong pleiotropic effects. Our analyses also implicated hub regions in the genome that are significantly correlated across numerous neuropsychiatric phenotypes. These results can guide future modeling efforts on these traits as well as functional genomic studies that interrogate key regions with pervasive regulatory roles across many phenotypes.

ASD and cognitive ability showed significant, bidirectional local genetic correlations in our analysis. We performed in-depth analyses to further dissect the shared genetics of ASD and cognition. For many years, GWASs of ASD have failed to identify robust associations that can be consistently replicated, most likely due to "omnigenicity" [65], weak effect of common SNPs, and insufficient sample size. However, exome sequencing studies for ASD have been fruitful in the past decade. Numerous consortium-scale whole-exome and whole-genome sequencing studies have been conducted to assess the roles of de novo mutations and very rare transmitted variants in ASD. These studies have convincingly identified more than 100 risk genes harboring pathogenic rare or de novo variants and implicated a number of etiologic pathways for ASD [27, 50, 66, 67]. Additionally, overwhelming evidence suggests that rare and de novo pathogenic variants in pathways such as chromatin modifiers and *FMRP* target genes contribute to the comorbidity of ASD and intellectual disability [27], which shaped our understanding of ASD genetics until very recently.

In contrast, successful GWASs for ASD have just begun to emerge [44]. It was notable that risk genes implicated by common SNPs do not have an apparent overlap with ASD genes identified in rare variant studies. Large well-powered GWASs, coupled with methodological advances in multi-trait analysis, have led to exciting findings about the shared genetic basis of ASD and other genetically correlated traits. However, a finding that surprised many in the field was the highly significant genetic correlation between ASD and higher IQ [16]. This genetic correlation was first identified using relatively underpowered ASD GWAS [10], but have since been replicated in well-powered large studies [16, 44]. A recent study further demonstrated that PRS of EA is over-transmitted from healthy parents to ASD probands, including probands who have pathogenic de novo mutations in known ASD genes, but not to the unaffected siblings [45].

These findings seemed paradoxical—whole exome sequencing (WES) reveals that shared genetic components contribute to ASD and intellectual disability while GWAS suggests that shared genetics contribute to ASD and higher cognitive ability. It raised two important questions. Why are ASD genes affected by common SNPs different from genes harboring rare protein-altering variants? Why do common and rare variants suggest opposite genetic relationships between ASD and cognition?

We aimed to address these questions head-on using local genetic correlations. We identified significant positive correlations of ASD and CP in 18 genomic regions but also 6 regions showing significant negative correlations. Locally, we did not observe the paradoxical correlation pattern seen in the global analysis, i.e., two positively correlated neurodevelopmental disorders ASD and ADHD showing opposite correlations with cognitive measures. Regions that were significantly correlated in all three trait pairs (e.g., the *KMT2E* locus) all showed consistent local correlations between both ASD and ADHD with CP. Of note, the set of regions negatively correlated between ASD and CP had a 3.2-fold enrichment for chromatin modifier genes. Thus, a genetic signature with consistent results between common and rare variants was hidden in plain sight. These genes, affected by both rare protein-altering variants and common (possibly regulatory) SNPs, may contribute to ASD with comorbid intellectual impairment in part through dysregulating chromatin modification in the developing brain. The positive global correlation between ASD and cognition was explained by a second genetic signature driven by a different set of regions that showed positive local correlations and were significantly enriched for PSD genes. When calculating the total genetic covariance between ASD and CP in the genome, negatively correlated regions were overwhelmed by the positive covariance in regions involved in the second signature, thus showing a positive global covariance. PRS based on these two signatures (PRS+ and PRS−) showed distinct associations with ASD phenotypes in the SPARK cohort. Compared to PRS−, PRS+ could better distinguish ASD cases from healthy controls. Both PRS+ and PRS− were associated with IQ in ASD probands but with opposite directions. In addition, PRS− significantly predicted overall ASD symptom severity while PRS+ significantly predicted repetitive behaviors. We also observed an enrichment of Asperger's disorder in probands with high PRS+ (and a slight depletion in probands with high PRS−), but these results only showed moderate statistical evidence and remain to be validated using larger samples in the future.

Our method still has some limitations. Although SUPERGNOVA can effectively estimate local genetic covariance, local genetic correlation estimates are numerically unstable due to the non-negligible noise in the estimates of local heritability. Second, due to the distal regulatory nature of common genetic variations, causal genes may not always be included in the pre-defined genetic region harboring GWAS associations. We suggest researchers also investigate regions adjacent to the identified region when interpreting local correlation results from SUPERGNOVA. Also note that although local genetic covariance highlights important genomic regions and hints at the involvement of physically proximal genes, it cannot prioritize genes directly. Functional evidence needs to be considered when linking the identified regions with genes. Third, the analyses we conducted in this paper were based on hypothesis-free scans in the genome but it is not the only possible study design. Filtering candidate regions based on strength of GWAS associations may reduce multiple testing burden and consequently improve statistical power in SUPERGNOVA. Our implemented software allows users to re-define their local region of interest if needed. Fourth, the performance of SUPERGNOVA can be affected under some settings of model misspecification. Whether the current model assumptions can be relaxed remains to be studied in the future. Fifth, the association results of PRS+/PRS− with clinical symptoms and ASD subtypes were not statistically significant after Bonferroni correction. Replications in probands with larger sample size will be implemented in the future. Finally, some other future directions include extending our method to estimate transethnic local genetic correlation [13]. The local correlation estimates provided by SUPERGNOVA may also improve other types of multi-trait analysis such as multi-trait association mapping [3] and genomic structural equation modeling (GenomicSEM) [4]. We believe SUPERGNOVA may play a critical role in accelerating the development of novel statistical genetics tools in the future.

Conclusions

Local genetic correlation analysis could reveal heterogeneous architecture of etiological sharing between complex traits and is critical for understanding the genetic basis of phenotypic correlations among traits. SUPERGNOVA provides a biologically motivated and statistically principled analytical strategy to tackle etiologic sharing of complex traits. A combination of global and local genetic correlation could provide new insights into the shared genetic basis of many phenotypes. As a case study to illustrate the power of SUPERGNOVA, we performed in-depth analyses to dissect the shared genetics of ASD and cognitive abilities. Given the biological difference between two sets of genomic regions with opposite correlations between ASD and CP, we concluded that the "paradoxical" genetic correlation could be explained by genetic heterogeneity. We believe SUPERGNOVA will have wide applications in complex trait genetics research.

Methods

Statistical model

We start with the statistical framework for global genetic covariance. Assume there are two studies with sample sizes n_1 and n_2, respectively. Standardized trait values ϕ_1 and ϕ_2 follow the linear models below:

$$\phi_1 = X\beta + \varepsilon$$

$$\phi_2 = Y\gamma + \delta,$$

where X and Y are $n_1 \times m$ and $n_2 \times m$ standardized genotype matrices; m is the number of shared SNPs between the two studies; ϵ and δ are the noise terms; and β and γ denote the genetic effects for ϕ_1 and ϕ_2. We adopt a model with random effects and random design matrices [10, 12, 24] to define genetic covariance ρ. The combined random vector of β and γ follows a multivariate normal distribution given by:

$$\begin{bmatrix}\beta\\ \gamma\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix}, \begin{bmatrix}\frac{h_1^2}{m} I_m & \frac{\rho}{m} I_m\\ \frac{\rho}{m} I_m & \frac{h_2^2}{m} I_m\end{bmatrix}\right),$$

where h_1^2 and h_2^2 are the heritability of the two traits, respectively; I_m is the identity matrix of size m. In practice, two different GWASs may share a subset of samples. Without loss of generality, we assume the first n_s samples in each study are shared ($n_s \le n_1$ and $n_s \le n_2$). The non-genetic effects of the shared samples for the two traits are correlated:

$$Cov[\varepsilon_{i_1}, \delta_{i_2}] = \begin{cases} \rho_e, 1 \le i_1 = i_2 \le n_s \\ 0, \quad otherwise \end{cases}$$

Since trait values ϕ_1 and ϕ_2 are standardized, we have $Var(\varepsilon_{i_1}) = 1 - h_1^2$ and $Var(\delta_{i_2}) = 1 - h_2^2$ for $1 \le i_1 \le n_1$ and $1 \le i_2 \le n_2$.

In GWAS summary data, we can approximate z scores of SNP j for trait 1 and trait 2 by $z_{1j} \approx X_{\cdot j}^T \phi_1 / \sqrt{n_1}$ and $z_{2j} \approx Y_{\cdot j}^T \phi_2 / \sqrt{n_2}$. We use z_1 and z_2 to denote the vectors for all SNPs' z scores and use V to denote the LD matrix. Under a random design model, $Cov(X_{i_1 \cdot}) = Cov(Y_{i_2 \cdot}) = V$ and the variance-covariance matrix of $(z_1^T, z_2^T)^T$ is given by

$$Var\left(\begin{bmatrix}z_1\\ z_2\end{bmatrix}\right) = \begin{bmatrix}\frac{n_1 h_1^2}{m} V^2 + V & \frac{\sqrt{n_1 n_2}\rho}{m} V^2 + \frac{n_s \rho_t}{\sqrt{n_1 n_2}} V\\ \frac{\sqrt{n_1 n_2}\rho}{m} V^2 + \frac{n_s \rho_t}{\sqrt{n_1 n_2}} V & \frac{n_2 h_2^2}{m} V^2 + V\end{bmatrix}, \quad (2)$$

where ρ_t is defined as the sum of genetic covariance and non-genetic effects covariance, i.e., $\rho_t = \rho + \rho_e$. We provide detailed derivations of (2) in the Additional file 1: Supplementary Note.

Results similar to (2) can be derived when one or both studies are case-control studies, where genetic covariance is on the observed scale. The observed scale genetic covariance is $\rho_{obs} = \rho\phi(\tau_1)\phi(\tau_2)\sqrt{P_1(1-P_1)P_2(1-P_2)}/[K_1(1-K_1)K_2(1-K_2)]$ when both studies are case-control studies, where ρ is the liability scale genetic covariance, ϕ is the standard normal density, τ_1 and τ_2, P_1 and P_2, and K_1 and K_2 are the liability threshold, sample prevalence, and population prevalence of study 1 and study 2, respectively. Details are provided in the Supplemental Note. Since the only distinction between observed and liability scale of genetic covariance is a positive constant, the observed and liability scale of genetic covariance is equivalent in terms of statistical significance and covariance direction between two traits.

Most existing genetic covariance methods are based on the idea of minimizing the "distance" between the empirical covariance matrix $\widehat{Cov}(z_1, z_2) = \frac{1}{2}(z_1 z_2^T + z_2 z_1^T)$ and the theoretical covariance in (1). For example, LDSC [10] regresses the diagonal elements of empirical z score covariance matrix on that of the theoretical covariance matrix. GNOVA [12] applies the method of moments estimator that compares the trace of the empirical and theoretical covariance matrices. Our new approach, SUPER-GNOVA, is also based on this unified framework.

The statistical framework we introduced above can be easily generalized to local genetic covariance. We assume ϕ_1 and ϕ_2 follow additive linear models:

$$\phi_1 = \sum_{i=1}^{I} X_i \beta_i + \varepsilon$$

$$\phi_2 = \sum_{i=1}^{I} Y_i \gamma_i + \delta,$$

where X_i and Y_i are the genotypes and β_i and γ_i are the effect sizes of SNPs in region i. In practice, I genomic regions can be mutually independent LD blocks defined by LDetect [29]. Following the same derivations as shown above, the variance-covariance matrix of local z scores z_{1i} and z_{2i} is

$$Var\left(\begin{bmatrix} z_{1i} \\ z_{2i} \end{bmatrix}\right) = \begin{bmatrix} \frac{n_1 h_{1i}^2}{m_i} V_i^2 + V_i & \frac{\sqrt{n_1 n_2}\rho_i}{m_i} V_i^2 + \frac{n_s \rho_t}{\sqrt{n_1 n_2}} V_i \\ \frac{\sqrt{n_1 n_2}\rho_i}{m_i} V_i^2 + \frac{n_s \rho_t}{\sqrt{n_1 n_2}} V_i & \frac{n_2 h_{2i}^2}{m_i} V_i^2 + V_i \end{bmatrix}, \quad (3)$$

where V_i, $h_{\cdot i}^2$, ρ_i and m_i are LD matrix, heritability, genetic covariance and number of SNPs for region i, respectively. ρ_t here is defined as the sum of local genetic covariance and non-genetic effects covariance, i.e., $\rho_t = \sum_{i=1}^{I} \rho_i + \rho_e$. Similarly, the local genetic covariance is on observed scale when one or both studies are case-control studies. Details about the construction of statistical model for local genetic covariance are provided in Additional file 1: Supplementary Note.

Local genetic covariance estimation

Following (3), the covariance of z_{1i} and z_{2i} (i.e., z scores of trait 1 and trait 2 in region i) is

$$Cov(z_{1i}, z_{2i}) = \frac{\sqrt{n_1 n_2}\rho_i}{m_i} V_i^2 + \frac{n_s \rho_t}{\sqrt{n_1 n_2}} V_i$$

Assume eigen decomposition of V_i is $V_i = U_i \Sigma_i U_i^T$, then we have

$$Cov\left(U_i^T z_{1i}, U_i^T z_{2i}\right) = \frac{\sqrt{n_1 n_2}\rho_i}{m_i} \Sigma_i^2 + \frac{n_s \rho_t}{\sqrt{n_1 n_2}} \Sigma_i$$

where $\Sigma_i = diag(w_{i1}, w_{i2}, \ldots, w_{im_i})$ ($w_{i1} \geq w_{i2} \geq \ldots \geq w_{im_i} \geq 0$ are the eigenvalues of Σ_i) and U_i is the corresponding orthogonal matrix of eigenvectors. Denote $\tilde{z}_{1i} = U_i^T z_{1i}$ and $\tilde{z}_{2i} = U_i^T z_{2i}$. For $j = 1, 2, \ldots, m_i$, the expected value and variance of $\tilde{z}_{1ij}\tilde{z}_{2ij}$ for the jth eigenvalue w_{ij} are

$$E\left[\tilde{z}_{1ij}\tilde{z}_{2ij}\right] = \frac{\sqrt{n_1 n_2}\rho_i}{m_i} w_{ij}^2 + \frac{n_s \rho_t}{\sqrt{n_1 n_2}} w_{ij}$$

and

$$Var\left[\tilde{z}_{1ij}\tilde{z}_{2ij}\right] = \left(\frac{\sqrt{n_1 n_2}\rho_i}{m_i} w_{ij}^2 + \frac{n_s \rho_t}{\sqrt{n_1 n_2}} w_{ij}\right)^2 + \left(\frac{n_1 h_{1i}^2}{m_i} w_{ij}^2 + w_{ij}\right)\left(\frac{n_2 h_{2i}^2}{m_i} w_{ij}^2 + w_{ij}\right) \tag{4}$$

where h_{1i}^2 and h_{2i}^2 can be estimated by the method of moments [68]. Derivations of (3) and (4) are in the Additional file 1: Supplementary Note. Due to the noise in LD estimation, especially for the smaller eigenvalues, we only use the first K_i eigenvalues to estimate ρ_i. The procedure to adaptively determine K_i is described in the following section. In practice, the LD matrices are estimated from an external reference panel (e.g., the 1000 Genomes Project [28]) and the intercept of cross-trait LDSC [10] provides an estimate of $n_s\rho_t/\sqrt{n_1 n_2}$, denoted as $\widehat{n_s\rho_t}/\sqrt{n_1 n_2}$. For each genomic region, we can estimate local genetic covariance and test the significance of $\hat{\rho}_i$ using the weighted regression of $\tilde{z}_{1ij}\tilde{z}_{2ij} - (\widehat{n_s\rho_t}/\sqrt{n_1 n_2})w_{ij}$, denoted by η_{ij}, on the square of eigenvalue weighted by the reciprocal of the variance in (4) which is approximated by $[(n_1 h_{1i}^2/m_i)w_{ij}^2 + w_{ij}][(n_2 h_{2i}^2/m_i)w_{ij}^2 + w_{ij}]$. Since $\tilde{z}_{1i1}\tilde{z}_{2i1}, \ldots\ldots, \tilde{z}_{1iK_i}\tilde{z}_{2iK_i}$ are independent for any region i, the theoretical variance of $\hat{\rho}_i \mid (\widehat{n_s\rho_t}/\sqrt{n_1 n_2})$ is analytically given by

$$Var\left[\hat{\rho}_i \middle| \frac{\widehat{n_s\rho_t}}{\sqrt{n_1 n_2}}\right] = \left(\frac{m_i^2}{n_1 n_2}\right) \Big/ \sum_{j=1}^{K_i} \frac{w_{ij}^4}{q_{ij}^2}. \tag{5}$$

Here, we denote $q_{ij}^2 = [(n_1 h_{1i}^2/m_i)w_{ij}^2 + w_{ij}][(n_2 h_{2i}^2/m_i)w_{ij}^2 + w_{ij}]$. In weighted regression, $[(n_1 h_{1i}^2/m_i)w_{ij}^2 + w_{ij}][(n_2 h_{2i}^2/m_i)w_{ij}^2 + w_{ij}]$ is treated as the reciprocal of the weight's square. So, the empirical variance of $\hat{\rho}_i \mid (\widehat{n_s\rho_t}/\sqrt{n_1 n_2})$ is analytically given by

$$\widehat{Var}\left[\hat{\rho}_i \middle| \frac{\widehat{n_s\rho_t}}{\sqrt{n_1 n_2}}\right] = \left[\left(\frac{m_i^2}{n_1 n_2}\right) \Big/ \sum_{j=1}^{K_i} \frac{w_{ij}^4}{q_{ij}^2}\right] \cdot \left[\sum_{j=1}^{K_i} \frac{\eta_{ij}^2}{q_{ij}^2} - \frac{\left(\sum_{j=1}^{K_i} \eta_{ij} w_{ij}^2/q_{ij}^2\right)^2}{\sum_{j=1}^{K_i} w_{ij}^4/q_{ij}^2}\right] / (K_i - 1). \tag{6}$$

Derivations of (5) and (6) are in the Additional file 1: Supplementary Note. To compensate for the variance introduced by LDSC in the estimation of $n_s\rho_t/\sqrt{n_1 n_2}$, we approximate $Var[\mathbb{E}[\hat{\rho}_i|(\widehat{n_s\rho_t}/\sqrt{n_1 n_2})]]$ by

$$Var\left[\mathbb{E}\left[\hat{\rho}_i \middle| \frac{\widehat{n_s\rho_t}}{\sqrt{n_1 n_2}}\right]\right] \approx \frac{m_i^2}{n_1 n_2} \cdot \left(\frac{\sum_{j=1}^{K_i} w_{ij}^3/q_{ij}^2}{\sum_{j=1}^{K_i} w_{ij}^4/q_{ij}^2}\right)^2 \cdot Var\left[\frac{\widehat{n_s\rho_t}}{\sqrt{n_1 n_2}}\right]. \tag{7}$$

The derivation of (7) is in the Additional file 1: Supplementary Note. The estimation of the last term in (7) $Var[\widehat{n_s\rho_t}/\sqrt{n_1 n_2}]$ is from LDSC. We use (6) to approximate $\mathbb{E}[Var[\hat{\rho}_i|(\widehat{n_s\rho_t}/\sqrt{n_1 n_2})]]$. By the law of total variance, we combine the results in (6) and (7) to obtain $Var[\hat{\rho}_i]$:

$$Var\left[\hat{\rho}_i\right] = Var\left[\mathbb{E}\left[\hat{\rho}_i|\frac{\widehat{n_s\rho_t}}{\sqrt{n_1n_2}}\right]\right] + \mathbb{E}\left[Var\left[\hat{\rho}_i|\frac{\widehat{n_s\rho_t}}{\sqrt{n_1n_2}}\right]\right].$$

Our statistical framework leverages the complete LD matrix to reduce information loss in estimation and at the same time denoises the empirical LD by properly selecting the optimal number of eigenvalues (determined by $K_i{}'s$) to include in the local analysis.

Then, local genetic correlation is estimated by $\hat{\rho}_i/\sqrt{\hat{h}_{1i}^2\hat{h}_{2i}^2}$. We approximate the standard error and the confidence interval of local genetic correlation estimation by the Delta method. However, local genetic correlation estimates might be numerically unstable due to the noise in the estimates of local heritability, which is in the denominator of the estimator of local genetic correlation. So, the estimates of genetic covariance of SUPERGNOVA are more reliable. When estimating global genetic covariance, the whole genome can be treated as a single region with non-zero submatrices only on the diagonal of its LD matrix.

An adaptive procedure to determine K_i

Sample LD information is rarely available for published GWASs. Therefore, we use an external reference panel to estimate LD. In practice, the number of SNPs is far greater than the number of individuals in the reference panel. For example, in this paper, we used 503 samples of European ancestry from the 1000 Genomes Project phase III as the reference panel. The average number of SNPs in a local region is about 2000 for common GWAS summary data. To achieve robust inference, we apply factor selection and only use the first K_i eigenvectors and eigenvalues for region i. There are several existing methods to perform factor selection [69–74]. The optimal K_i should lead to powerful inference and properly controls the type I error. Here, we propose an adaptive procedure to determine the value of optimal K_i. Under the optimal K_i, theoretical variance in (5) and empirical variance in (6) should be close. We know from (5) that theoretical variance decreases with the increase of K_i. However, the value of empirical variance rapidly increases when the cutoff for the eigenvalues approaches towards zero (Additional file 2: Supplementary Figure 40). We adaptively determine the optimal K_i as follows. First, we set an upper bound for K_i. In our paper, the upper bound is 503 which is the number of samples in the reference panel. Then, for region i, we compute the value of theoretical variance and empirical variance for K_i taking values from 10 to the upper bound. We denote the maximum of theoretical variance and empirical variance for each K_i as

$$v(K_i) = max\left\{Var^{K_i}\left[\hat{\rho}_i|\frac{\widehat{n_s\rho_t}}{\sqrt{n_1n_2}}\right], \widehat{Var}^{K_i}\left[\hat{\rho}_i|\frac{\widehat{n_s\rho_t}}{\sqrt{n_1n_2}}\right]\right\}.$$

The optimal K_i is determined by $\underset{10 \le K_i \le \min(m_i,503)}{\arg\min} v(K_i)$. After K_i is decided, we use weighted least square to obtain the estimate of genetic covariance.

Simulation settings

We used genotype data from WTCCC to conduct simulations. Samples were randomly divided into two equal subgroups with 7959 individuals. We denote them as set 1 and

set 2. We randomly sampled 3979 individuals from set 1 and 3980 individuals from set 2 to create set 3 which has a 50% sample overlap with set 1. Samples with European ancestry from the 1000 Genomes Project phase III [28] were used as the LD reference in our simulations. We kept common SNPs with MAFs greater than 5% and removed all SNPs with ambiguous alleles. After quality control, 287,539 SNPs remained in both WTCCC and 1000 Genomes Project data.

We used LDetect [29] to partition the genome into 2197 LD blocks (~ 1.6 centimorgan in width on average). To estimate local genetic covariance, we selected the largest region partitioned by LDetect on chromosome 2 (176,998,822–180,334,969) to be the local region of interest. There are 395 SNPs in this region in the genotype data from WTCCC. The effects of SNPs on two simulated traits were only correlated in the local region. The remaining 23,839 SNPs on chromosome 2 were used as background SNPs whose genetic covariance was set as 0. The effect sizes of SNPs were generated by a multivariate normal distribution and we applied Genome-wide Complex Trait Analysis (GCTA) [75] to simulate ϕ_1 and ϕ_2. We used PLINK [76] to run GWAS and obtain summary statistics of the two simulated traits. We repeated each simulation setting 100 times. Detailed simulation settings are summarized below.

For simulations of global genetic covariance analysis, we set the heritability of two traits to be 0.5 and set the genetic covariance to be 0, 0.05, 0.1, 0.15, 0.2, and 0.25, respectively. We conducted three simulations, corresponding to different levels of sample overlap.

1. Independent studies: we used set 1 and set 2 to simulate ϕ_1 and ϕ_2, respectively.
2. Partial sample overlap: ϕ_1 and ϕ_2 were simulated on set 1 and set 3. The covariance of non-genetic effects on shared samples was set to be 0.2.
3. Complete sample overlap: ϕ_1 and ϕ_2 were both simulated on set 1. The covariance of non-genetic effects on shared samples was set to be 0.2.

For simulations of local genetic covariance analysis, we set the heritability of two traits to be 0.5. The total heritability is evenly distributed to all SNPs. The covariance of the local genetic effects was set to be 0, 0.001, 0.002, 0.003, 0.004, and 0.005, respectively. Similar to global analyses, we conducted three sets of simulations.

1. Independent studies: ϕ_1 and ϕ_2 were simulated on set 1 and set 2, respectively.
2. Partial sample overlap: we simulated ϕ_1 and ϕ_2 on set 1 and set 3. The covariance of non-genetic effects was set to be 0.2.
3. Complete sample overlap: we simulated both ϕ_1 and ϕ_2 on set 1. The covariance of non-genetic effects was set to be 0.2.

To compare the performance between SUPERGNOVA and ρ-HESS [11], we input true sample overlap n_s to ρ-HESS. We followed the instruction provided by ρ-HESS software to estimate phenotypic correlation, which is another required input of ρ-HESS. To evaluate the robustness of ρ-HESS against mis-specified overlapping sample size, we provided the method with an overlapping sample size of 1000 as input in partial sample overlap and complete sample overlap scenarios to estimate local genetic covariance. The phenotypic correlation is also estimated according to the instruction of ρ-HESS.

SUPERGNOVA and ρ-HESS are compared by proportion of replicates out of 100 that have a p value less than 0.05. When the true genetic covariance is zero (i.e., null hypothesis is true), this proportion is an estimation of type I error. When the true genetic covariance is nonzero (i.e., alternative hypothesis is true), this proportion is an estimation of statistical power at significance level of 0.05.

GWAS data

GWAS summary statistics of 29 complex traits included in our analyses are publicly available. We obtained the summary statistics of a recent lung cancer GWAS directly from the authors [77]. Details of the 30 GWASs are summarized in Additional file 3: Supplementary Table 1. We used munge_sumstats.py script in LDSC to reformat these data and removed strand-ambiguous SNPs from each dataset. For each trait pair, we took the intersection of SNPs in two GWAS and the 1000 Genomes Project. We matched the effect alleles after removing SNPs with MAF lower than 5%. We only included the SNPs in autosomes and excluded the MHC region in all analyses.

We accessed samples from the SPARK study through the Simons Foundation Autism Research Initiative (SFARI). Samples in the SPARK study were genotyped by the Illumina Infinium Global Screening Array. Details on these samples have been previously reported and are available on the SFARI website [78]. Following data processing procedure in Huang et al. [54], we performed pre-imputation quality control (QC) using PLINK. The genotype data were phased and imputed to the HRC reference panel version r1.1 2016 using the Michigan Imputation server [79].

Estimation of the proportion of correlated regions

We estimated the proportion of correlated regions with an R package called *ashr* [30] after the estimation of local genetic covariance among the 30 phenotypes. The inputs were estimates of local genetic covariance and its standard error. The unimodal prior distribution was set to be "halfnormal" for all the results of pairs of traits. The method applied a Bayesian framework to compute FDR for each genomic region. To estimate the numbers of correlated regions for each pair of traits, we computed the sum of (1 − FDR) given by ashr for each region.

Follow-up analyses in the *SPI1* locus for AD and other neuropsychiatric traits

To replicate local genetic covariance identified at the *SPI1* locus, we defined a new genomic region centered at *SPI1* with a 1-Mb span. We estimated the local genetic covariance between AD (IGAP2019 [41]) and the other 29 traits for this region. For replication, we implemented a GWAS for AD family history in the UKBB and estimated the local genetic covariance of this GWAS with other traits. Details on the AD-proxy GWAS have been previously reported [41, 80].

We obtained PU.1 binding sites as ChIP-seq peaks from the ReMap datasets [81] (GEO: GSE31621; *SPI1*, blood monocyte and macrophage datasets [82]). Following Huang et al. [43], we expanded each ChIP-seq peak by 150 kb up- and downstream to define the transcription factor binding site annotation. We applied GNOVA [12] to estimate the genetic covariance between AD and 29 other traits in the PU.1 binding sites. We trained elastic net gene expression imputation models [83, 84] using expression

profiles adjusted by peer factors [85] and PCs and matched genotypes from 758 monocyte and 599 macrophage samples in the Cardiogenics Consortium [86] imputed by Michigan Imputation Server [79]. We downloaded Cardiogenics resources from European Genome-phenome Archive (EGA) platform. To investigate the regulatory relationship between PU.1 and the identified genes in myeloid cells, we used GREAT [87] to map PU.1 each binding peak in macrophages and monocytes to the nearest gene.

Cross-tissue transcriptome-wide association analysis

To identify genes associated with ADHD, ASD, CP and SCZ in brain tissues in the *KMT2E* region (chr7: 104158491–105425027), we implemented cross-tissue transcriptome-wide association analysis using UTMOST [88]. We used gene expression imputation models trained by genotype and normalized gene expression data from the GTEx project [89–92] (version V8). We considered 13 brain tissues. For individual expression data, we regressed out the effects of confounding covariates including first five genotype PCs, PEER factors optimized by sample sizes as in the GTEx V8 paper [92], sequencing platforms, library construction protocol and donor sex. Cis-genotype data was extracted for SNPs located within 1 MB distance from the transcription starting sites of all protein coding genes. Then, we trained expression imputation models based on cis-genotypes for each gene in each tissue using 10-fold elastic net with alpha being 0.5. Models with credible imputation performances (FDR < 0.05) were used in later analysis.

Functional annotation for variants in GWAS data

We used bedtools [93] to extract sequence from the *KMT2E* region. We then performed gene annotations on each of the variants using ANNOVAR [94]. For exonic and splicing variants, missense variants were represented by nonsynonymous single nucleotide variants (SNVs) and loss-of-function variants were annotated as frameshift, stopgain, or stoploss mutations by ANNOVAR. We took overlapped SNPs and matched the effect alleles between ANNOVAR annotations and GWAS summary data of ADHD, ASD, CP and SCZ respectively.

Gene set enrichment analysis

We used R package *TxDb.Hsapiens.UCSC.hg19.knownGene* to identify genes in the correlated regions between ASD and CP with nominal significant covariances ($p < 0.01$). We only included protein-coding genes in our analysis, resulting in 317 positively correlated genes and 179 negatively correlated genes. We applied *Enrichr* [95, 96] to implement enrichment analysis on GWAS catalog 2019 [97] (Additional file 3: Supplementary Tables 19-20), and TF PPI [95]. We identified *FMRP* target genes, genes encoding PSD proteins, gene preferentially expressed in human embryonic brains, essential genes, chromatin modifier genes, genes with probability of loss-of-function intolerance (pLI) > 0.9, and SFARI evidence score based on previous literature [54]. We obtained a list of 102 genes identified by the refined transmitted and de novo association (TADA) model [98] (FDR < 0.1) in the recent exome sequencing study on ASD [50]. We performed hypergeometric test to assess the enrichment of ASD-CP positively and negatively correlated genes in these gene sets. We performed permutation test to

assess the differential enrichment across gene sets. We randomly assigned these genes into positive or negative sets and computed the maximal chi-square statistics in the 102 ASD genes from Satterstrom et al. [50] and chromatin modifier genes [27]. We repeated the permutation 1000 times and quantified the empirical p value as the proportion of permutations with shuffled test statistics greater than the test statistic observed in real data.

Analysis of spatio-temporal RNA-seq data in brain tissues

We used single-cell RNA-seq data generated by the PsychENCODE Consortium [99] in fetal brains to test the elevation of gene expression of ASD-CP correlated genes in brain development. There were 762 cells collected from neocortical regions of eight fetal brains from 5 to 20 PCW. We kept only protein-coding genes which included 18,134 genes in this analysis. Three hundred ten of positively correlated and 175 of negatively correlated genes were overlapped with these genes. Following Satterstrom et al. [50], for each time point, a gene was considered expressed if at least one transcript mapped to this gene in 25% or more of cells for at least one PCW period before. By definition, gene expression rate increased with fetal development. We performed log-rank test to test the difference of gene expression rate in developmental brain between positively or negatively correlated genes and other genes.

We downloaded developmental bulk RNA-seq data from BrainSpan. Gene-level RPKMs were used across 524 samples from 42 individuals in 26 brain regions [99]. We kept protein-coding genes in our analysis. Following Satterstrom et al. [50], we removed samples with RNA integrity number (RIN) ≤ 7 and only used neocortical regions—dorsolateral prefrontal cortex (DFC), ventrolateral prefrontal cortex (VFC), medial prefrontal cortex (MFC), orbitofrontal cortex (OFC), primary motor cortex (M1C), primary somatosensory cortex (S1C), primary association cortex (A1C), inferior parietal cortex (IPC), superior temporal cortex (STC), inferior temporal cortex (ITC), and primary visual cortex (V1C). Genes were defined as expressed if their RPKMs were at least 0.5 in 80% samples from at least one neocortical region at one major temporal epoch. Consequently, 14,803 genes were defined as expressed in 325 samples from 8 post-conceptual weeks (PCW) to 40 years of age. We then log-transformed RKPM ($\log_2$[RKPM+ 1]). We followed the definition of developmental stages in Li et al. [99]. We performed t test to determine the differential expression among ASD-CP positively correlated genes, negatively correlated genes, and background genes.

To study the relative prenatal and postnatal bias, we performed linear regression for the transformed RKPM of each gene on a binary "prenatal" stage variable. Sex was included as an adjustment variable. Genes were defined as prenatally (or postnatally) biased if $\log_2$ fold change> 0.1 (or < − 0.1) and q value< 0.05 resulting in 5562 prenatally biased genes and 5361 postnatally biased genes. We followed the definition of ASD-CP positively and negatively correlated genes from gene set enrichment analysis. Chi-squared test was performed to test if the distributions of prenatally and postnatally biased genes in ASD-CP positively and negatively correlated regions were significantly different from background genes.

PRS analysis in SPARK

We used the 18 positively correlated regions and 6 negatively correlated regions (FDR < 0.1) between ASD and CP to construct PRS+ and PRS− of ASD. We clumped the SNPs by PLINK [76]. We set the significance threshold for index SNPs as 1, LD threshold for clumping as 0.1, and physical distance threshold for clumping as 250 kb. We generated scores for 5469 ASD probands and 2132 healthy siblings in the SPARK cohort. We assessed associations between two PRSs and ASD using logistic regression. We then investigated the association between the two PRSs and ASD phenotypes in probands, including IQ, SCQ score, RBS-R score, DCDQ score, and subtypes of ASD. For each phenotype, we used the maximum sample with both genotype and phenotype data. Sample sizes for these phenotypes in SPARK are summarized in Additional file 3: Supplementary Table 25. We performed two-sample t test for quantitative phenotypes between probands with extreme PRS (top 1%) and other probands. We performed hypergeometric test to test enrichment of subtypes in the extreme PRS group.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02478-w.

Additional file 1. Supplementary note, containing further derivations of the statistical models, information about additional simulations, and discussion of interpretability of local genetic covariance.

Additional file 2. Supplementary figures 1-40.

Additional file 3. Supplementary tables 1-26.

Additional file 4. Review history.

Acknowledgements

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. We conducted the research using the UKBB resource under approved data requests (refs: 29900 and 42148). This study makes use of summary statistics from many GWAS consortia. We thank the investigators in these GWAS consortia for generously sharing their data. We thank the CARDIOGENICS project for providing expression data of macrophages and monocytes and expression data. CARDIOGENICS resources was funded by the European Union FP6 program (LSHM-CT-2006-037593). We thank the PsychENCODE Consortium for providing the single-cell RNA-seq data in fetal brain. We are grateful to all the families participating in the Simons Foundation Powering Autism Research for Knowledge (SPARK) study.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 4.

Authors' contributions

Y.Z. and Q.L. developed the statistical framework. Y.Z. performed statistical analysis. Y.Y. assisted in analyzing GWAS data. K.H. curated gene list of ASD pathways. Y.W and K.H. processed SPARK data. W.L. and Z.Y. assisted in training expression imputation model. Y.W and X.Z implemented the GWAS on AD family history. B.L. performed ANNOVAR analysis. B.T., D.W., and J.L. advised on the biology of ASD and ADHD. Q.L. and H.Z. advised on statistical and genetics issues. Y.Z. implemented the software. Y.Z. and Q.L. wrote the manuscript. All authors contributed to manuscript editing and approved the manuscript.

Funding

This study was supported in part by NIH grants 3P30AG021342-16S2, 1R01GM122078, and R01GM134005 and NSF grants DMS 1713120 and DMS 1902903. Q.L is supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427. We also acknowledge research support from the University of Wisconsin-Madison Office of the Chancellor and the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation and the Waisman Center pilot grant program at the University of Wisconsin-Madison.

Availability of data and materials

Details of the GWAS summary data of the 30 complex traits are summarized in Additional file 3: Supplementary Table 1 [31, 41, 44, 46, 77, 100–118]. Genotype data used in the simulation were downloaded from the Wellcome Trust Case-Control Consortium [119]. GWAS for AD family history were conducted based on UKBB samples [120] under approved data requests. Data on ASD probands and siblings were accessed from the SPARK study through the Simons

Foundation Autism Research Initiative (SFARI) [121]. PU.1 binding site as ChIP-seq peaks were downloaded from the ReMap datasets [122]. Cardiogenics resources on expression profiles of monocyte and macrophage samples were downloaded from European Genome-phenome Archive (EGA) platform [123]. Single-cell RNA-seq data used in our study were generated by the PsychENCODE Consortium [124]. Developmental bulk RNA-seq data were downloaded from BrainSpan [125].
SUPERGNOVA software is publicly available at https://github.com/qlu-lab/SUPERGNOVA. The code used for this paper has also been deposited at Zenodo with DOI 10.5281/zenodo.5205277 [126]. We used LDetect (https://bitbucket.org/nygcresearch/ldetect/src/master/) to implement genome partition. Enrichment analysis on GWAS catalog 2019 and TF PPI was conducted on the online software of Enrichr (https://amp.pharm.mssm.edu/Enrichr/). Software for LDSC is available at https://github.com/bulik/ldsc; Software for GNOVA is available at https://github.com/xtonyjiang/GNOVA; Software for ρ-HESS is available at https://huwenboshi.github.io/hess/.

Declarations

Ethics approval and consent to participate
Ethical approval was not needed for this study.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06520, USA. [2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53706, USA. [3]Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA. [4]Center for Demography of Health and Aging, University of Wisconsin-Madison, Madison, WI 53706, USA. [5]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06510, USA. [6]Occupational Therapy Program in the Department of Kinesiology, University of Wisconsin-Madison, Madison, WI 53706, USA. [7]Waisman Center, University of Wisconsin-Madison, Madison, WI 53705, USA. [8]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA. [9]Department of Psychology, University of Wisconsin-Madison, Madison, WI 53706, USA. [10]Department of Genetics, Yale School of Medicine, New Haven, CT 06510, USA.

Received: 18 October 2020 Accepted: 23 August 2021
Published online: 07 September 2021

References
1. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101(1):5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.
2. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017;18(2):117–27.
3. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018;50(2):229.
4. Grotzinger AD, Rhemtulla M, de Vlaming R, Ritchie SJ, Mallard TT, Hill WD, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nat Hum Behav. 2019;3(5):513–25. https://doi.org/10.1038/s41562-019-0566-x.
5. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, Robinson MR, McGrath JJ, Visscher PM, Wray NR, Yang J. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun. 2018;9(1):1–2.
6. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016;48(7):709.
7. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet. 2018;50(5):693.
8. Cortes A, Albers PK, Dendrou CA, Fugger L, McVean G. Identifying cross-disease components of genetic risk across hospital data in the UK Biobank. Nat Genet. 2020;52(1):126–34.
9. van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of polygenic disease traits: from theory to practice. Nat Rev Genet. 2019;20(10):567–81.
10. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47(11):1236.
11. Shi HWB, Mancuso N, Spendlove S, Pasaniuc B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. Am J Hum Genet. 2017;101(5):737–51. https://doi.org/10.1016/j.ajhg.2017.09.022.
12. Lu Q, Li B, Ou D, Erlendsdottir M, Powles RL, Jiang T, et al. A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. Am J Hum Genet. 2017;101(6):939–64. https://doi.org/10.1016/j.ajhg.2017.11.001.
13. Brown BC, Ye CJ, Price AL, Zaitlen N, Network AGE. Transethnic genetic-correlation estimates from summary statistics. Am J Hum Genet. 2016;99(1):76–88. https://doi.org/10.1016/j.ajhg.2016.05.001.
14. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28(19):2540–2.
15. Guo ZJ, Wang WJ, Cai TT, Li HZ. Optimal estimation of genetic relatedness in high-dimensional linear models. J Am Stat Assoc. 2019;114(525):358–69. https://doi.org/10.1080/01621459.2017.1407774.

16. Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, Duncan L, et al. Analysis of shared heritability in common disorders of the brain. Science. 2018;360(6395):1313.
17. Maier RM, Zhu ZH, Lee SH, Trzaskowski M, Ruderfer DM, Stahl EA, et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nat Commun. 2018;9(1):989. https://doi.org/10.1038/s41467-017-02769-6.
18. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS genetics. 2017;13(6):e1006836.
19. Zhao B, Zhu H. On genetic correlation estimation with summary statistics from genome-wide association studies. arXiv preprint arXiv:190301301; 2019.
20. Nieuwboer HA, Pool R, Dolan CV, Boomsma DI, Nivard MG. GWIS: Genome-wide inferred statistics for functions of multiple phenotypes. Am J Hum Genet. 2016;99(4):917–27. https://doi.org/10.1016/j.ajhg.2016.07.020.
21. Deng YQ, Pan W. Conditional analysis of multiple quantitative traits based on marginal GWAS summary statistics. Genet Epidemiol. 2017;41(5):427–36. https://doi.org/10.1002/gepi.22046.
22. O'Connor LJ, Price AL. Distinguishing genetic correlation from causation across 52 diseases and complex traits (vol 50, pg 1728, 2018). Nat Genet. 2018;50(12):1753.
23. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565–9. https://doi.org/10.1038/ng.608.
24. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47:291.
25. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, 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. 2017;33(2):272–9.
26. Guo H, Li JJ, Lu Q, Hou L. Detecting local genetic correlations with scan statistics. Nat Commun. 2021;12(1):1–3.
27. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216–U136. https://doi.org/10.1038/nature13908.
28. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
29. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics. 2016;32(2):283–5.
30. Stephens M. False discovery rates: a new deal. Biostatistics. 2017;18(2):275–94. https://doi.org/10.1093/biostatistics/kxw041.
31. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet. 2017;49(2):256–61.
32. Dardani C, Riglin L, Leppert B, Sanderson E, Rai D, Howe L, Davey Smith G, Tilling K, Thapar A, Davies N, Anderson E. Is genetic liability to ADHD and ASD causally linked to educational attainment?. Int J Epidemiol. 2021.
33. Kowianski P, Lietzau G, Czuba E, Waskow M, Steliga A, Morys J. BDNF: A key factor with multipotent impact on brain signaling and synaptic plasticity. Cell Mol Neurobiol. 2018;38(3):579–93. https://doi.org/10.1007/s10571-017-0510-4.
34. Notaras M, van den Buuse M. Neurobiology of BDNF in fear memory, sensitivity to stress, and stress-related disorders. Mol Psychiatry. 2020;25(10):2251–74. https://doi.org/10.1038/s41380-019-0639-2.
35. Yamada K, Nabeshima T. Brain-derived neurotrophic Factor/TrkB signaling in memory processes. J Pharmacol Sci. 2003;91(4):267–70. https://doi.org/10.1254/jphs.91.267.
36. Martinowich K, Manji H, Lu B. New insights into BDNF function in depression and anxiety. Nat Neurosci. 2007;10(9):1089–93. https://doi.org/10.1038/nn1971.
37. Sullivan PF, Keefe RS, Lange LA, Lange EM, Stroup TS, Lieberman J, et al. NCAM1 and neurocognition in schizophrenia. Biol Psychiatry. 2007;61(7):902–10. https://doi.org/10.1016/j.biopsych.2006.07.036.
38. Yang BZ, Kranzler HR, Zhao H, Gruen JR, Luo X, Gelernter J. Haplotypic variants in DRD2, ANKK1, TTC12, and NCAM1 are associated with comorbid alcohol and drug dependence. Alcohol Clin Exp Res. 2008;32(12):2117–27. https://doi.org/10.1111/j.1530-0277.2008.00800.x.
39. Bidwell LC, McGeary JE, Gray JC, Palmer RHC, Knopik VS, MacKillop J. NCAM1-TTC12-ANKK1-DRD2 variants and smoking motives as intermediate phenotypes for nicotine dependence. Psychopharmacology. 2015;232(7):1177–86.
40. Cross-Disorder Group of the Psychiatric Genomics Consortium, Electronic address pmhe, Cross-Disorder Group of the Psychiatric Genomics C. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. Cell. 2019;179(7):1469–82 e11. https://doi.org/10.1016/j.cell.2019.11.020.
41. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. Nat Genet. 2019;51(3):414–30. https://doi.org/10.1038/s41588-019-0358-2.
42. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Genetics. 2013;45(12):1452–U206. https://doi.org/10.1038/ng.2802.
43. Huang KL, Marcora E, Pimenova AA, Di Narzo AF, Kapoor M, Jin SC, et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. Nat Neurosci. 2017;20(8):1052.
44. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51(3):431–44. https://doi.org/10.1038/s41588-019-0344-8.
45. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat Genet. 2017;49(7):978.
46. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018;50(8):1112–21. https://doi.org/10.1038/s41588-018-0147-3.
47. Moreno-De-Luca A, Myers SM, Challman TD, Moreno-De-Luca D, Evans DW, Ledbetter DH. Developmental brain dysfunction: revival and expansion of old concepts based on new genetic evidence. Lancet Neurol. 2013;12(4):406–14. https://doi.org/10.1016/S1474-4422(13)70011-5.

48. De Brouwer APM, Abou Jamra R, Kortel N, Soyris C, Polla DL, Safra M, et al. Variants in PUS7 cause intellectual disability with speech delay, microcephaly, short stature, and aggressive behavior. Am J Hum Genet. 2018;103(6):1045–52. https://doi.org/10.1016/j.ajhg.2018.10.026.
49. O'Donnell-Luria AH, Pais LS, Faundes V, Wood JC, Sveden A, Luria V, et al. Heterozygous variants in KMT2E cause a spectrum of neurodevelopmental disorders and epilepsy. Am J Hum Genet. 2019;104(6):1210–22. https://doi.org/10.1016/j.ajhg.2019.03.021.
50. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An J-Y, et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell. 2020;180(3):568–84. e23.
51. Muhleisen TW, Leber M, Schulze TG, Strohmaier J, Degenhardt F, Treutlein J, et al. Genome-wide association study reveals two new risk loci for bipolar disorder. Nat Commun. 2014;5:3339.
52. Hou L, Bergen SE, Akula N, Song J, Hultman CM, Landen M, et al. Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. Hum Mol Genet. 2016;25(15):3383–94. https://doi.org/10.1093/hmg/ddw181.
53. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science. 2013;340(6139):1467–71. https://doi.org/10.1126/science.1235488.
54. Huang K, Wu Y, Shin J, Zheng Y, Siahpirani AF, Lin Y, Ni Z, Chen J, You J, Keles S, Wang D. Transcriptome-wide transmission disequilibrium analysis identifies novel risk genes for autism spectrum disorder. PLoS genetics. 2021;17(2):e1009309.
55. Pearl JR, Colantuoni C, Bergey DE, Funk CC, Shannon P, Basu B, et al. Genome-scale transcriptional regulatory network models of psychiatric and neurodegenerative disorders. Cell Syst. 2019;8(2):122.
56. Chen C, Meng Q, Xia Y, Ding C, Wang L, Dai R, Cheng L, Gunaratne P, Gibbs RA, Min S, Coarfa C. The transcription factor POU3F2 regulates a gene coexpression network in brain tissue from patients with psychiatric disorders. Sci Transl Med. 2018;10(472).
57. Jin SC, Homsy J, Zaidi S, Lu QS, Morton S, DePalma SR, et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat Genet. 2017;49(11):1593.
58. Homsy J, Zaidi S, Shen YF, Ware JS, Samocha KE, Karczewski KJ, et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. Science. 2015;350(6265):1262–6. https://doi.org/10.1126/science.aac9396.
59. Jansen A, Dieleman GC, Smit AB, Verhage M, Verhulst FC, Polderman TJC, et al. Gene-set analysis shows association between FMRP targets and autism spectrum disorder. Eur J Hum Genet. 2017;25(7):863–8.
60. Ronemus M, Iossifov I, Levy D, Wigler M. The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet. 2014;15(2):133–41. https://doi.org/10.1038/nrg3585.
61. Steinberg J, Webber C. The roles of FMRP-regulated genes in autism spectrum disorder: single- and multiple-hit genetic etiologies. Am J Hum Genet. 2013;93(5):825–39. https://doi.org/10.1016/j.ajhg.2013.09.013.
62. Bishop SL, Hus V, Duncan A, Huerta M, Gotham K, Pickles A, et al. Subcategories of restricted and repetitive behaviors in children with autism spectrum disorders. J Autism Dev Disord. 2013;43(6):1287–97. https://doi.org/10.1007/s10803-012-1671-0.
63. Buja A, Volfovsky N, Krieger AM, Lord C, Lash AE, Wigler M, et al. Damaging de novo mutations diminish motor skills in children on the autism spectrum. Proc Natl Acad Sci U S A. 2018;115(8):E1859–E66.
64. Bishop SL, Farmer C, Bal V, Robinson EB, Willsey AJ, Werling DM, et al. Identification of developmental and behavioral markers associated with genetic abnormalities in autism spectrum disorder. Am J Psychiatry. 2017;174(6):576–85. https://doi.org/10.1176/appi.ajp.2017.16101115.
65. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169(7):1177–86.
66. Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet. 2018;50(5):727–36. https://doi.org/10.1038/s41588-018-0107-y.
67. An JY, Lin K, Zhu L, Werling DM, Dong S, Brand H, Wang HZ, Zhao X, Schwartz GB, Collins RL, Currall BB. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science. 2018;362(6420).
68. Zhou X. A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann Appl Stat. 2017;11(4):2027–51. https://doi.org/10.1214/17-AOAS1052.
69. Faber K, Kowalski BR. Critical evaluation of two F-tests for selecting the number of factors in abstract factor analysis. Anal Chim Acta. 1997;337(1):57–71.
70. Xie YL, Kalivas JH. Evaluation of principal component selection methods to form a global prediction model by principal component regression. Anal Chim Acta. 1997;348(1-3):19–27. https://doi.org/10.1016/S0003-2670(97)00035-4.
71. Sutter JM, Kalivas JH, Lang PM. Which principal components to utilize for principal component regression. J Chemometr. 1992;6(4):217–25.
72. Sun JG. A correlation principal component regression-analysis of NIR data. J Chemometr. 1995;9(1):21–9. https://doi.org/10.1002/cem.1180090104.
73. Depczynski U, Frost VJ, Molt K. Genetic algorithms applied to the selection of factors in principal component regression. Anal Chim Acta. 2000;420(2):217–27. https://doi.org/10.1016/S0003-2670(00)00893-X.
74. Malinowski ER. Determination of the number of factors and the experimental error in a data matrix. Anal Chem. 1977;49(4):612–7.
75. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
76. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75. https://doi.org/10.1086/519795.
77. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017;49(7):1126–32. https://doi.org/10.1038/ng.3892.

78. Feliciano P, Zhou X, Astrovskaya I, Turner TN, Wang T, Brueggeman L, et al. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. NPJ Genom Med. 2019;4(1):19. https://doi.org/10.1038/s41525-019-0093-8.
79. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284–7.
80. Zhao Z, Yi Y, Wu Y, Zhong X, Lin Y, Hohman TJ, Fletcher J, Lu Q. Fine-tuning polygenic risk scores with GWAS summary statistics. Genome Biol. 2019:810713.
81. Griffon A, Barbier Q, Dalino J, van Helden J, Spicuglia S, Ballester B. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Res. 2015;43(4):e27. https://doi.org/10.1093/nar/gku1280.
82. Pham TH, Benner C, Lichtinger M, Schwarzfischer L, Hu YH, Andreesen R, et al. Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. Blood. 2012;119(24):E161–E71.
83. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091.
84. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, Torstenson ES, Shah KP, Garcia T, Edwards TL, Stahl EA. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9(1):1–20.
85. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6(5):e1000770. https://doi.org/10.1371/journal.pcbi.1000770.
86. Garnier S, Truong V, Brocheton J, Zeller T, Rovital M, Wild PS, Ziegler A, Cardiogenics Consortium, Munzel T, Tiret L, Blankenberg S. Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. PLoS genetics. 2013;9(1):e1003240.
87. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28(5):495–501. https://doi.org/10.1038/nbt.1630.
88. Hu YM, Li M, Lu QS, Weng HY, Wang JW, Zekavat SM, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet. 2019;51(3):568.
89. Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204–13.
90. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–60. https://doi.org/10.1126/science.1262110.
91. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreserv Biobank. 2015;13(5):311–9. https://doi.org/10.1089/bio.2015.0032.
92. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318–30.
93. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
94. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
95. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan QN, Wang ZC, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–W7. https://doi.org/10.1093/nar/gkw377.
96. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics. 2013;14(1):1–4.
97. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005–D12. https://doi.org/10.1093/nar/gky1120.
98. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. Plos Genet. 2013;9(8):e1003671. https://doi.org/10.1371/journal.pgen.1003671.
99. Li M, Santpere G, Kawasawa YI, Evgrafov OV, Gulden FO, Pochareddy S, Sunkin SM, Li Z, Shin Y, Zhu Y, Sousa AM. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science. 2018;362(6420).
100. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet. 2019;51(1):63–75. https://doi.org/10.1038/s41588-018-0269-7.
101. Watson HJ, Yilmaz Z, Thornton LM, Hubel C, Coleman JRI, Gaspar HA, et al. Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. Nat Genet. 2019;51(8):1207–14. https://doi.org/10.1038/s41588-019-0439-2.
102. Meier SM, Trontti K, Purves KL, Als TD, Grove J, Laine M, et al. Genetic variants associated with anxiety and stress-related disorders: a genome-wide association study and mouse-model study. JAMA Psychiatry. 2019;76(9):924–32. https://doi.org/10.1001/jamapsychiatry.2019.1119.
103. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat Genet. 2019;51(5):793–803. https://doi.org/10.1038/s41588-019-0397-8.
104. Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019;22(3):343–52.
105. Arnold PD, Askland KD, Barlassina C, Bellodi L, Bienvenu OJ, Black D, et al. Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. Mol Psychiatr. 2018;23(5):1181–8.

106. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 2019;18(12):1091–102.
107. Pardinas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018;50(3):381–9. https://doi.org/10.1038/s41588-018-0059-2.
108. Okbay A, Baselmans BML, De Neve JE, Turley P, Nivard MG, Fontana MA, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses (vol 48, pg 624, 2016). Nat Genet. 2016;48(12):1591.
109. Liu MZ, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet. 2019;51(2):237.
110. Jones SE, Tyrrell J, Wood AR, Beaumont RN, Ruth KS, Tuke MA, et al. Genome-wide association analyses in 128,266 individuals identifies new morningness and sleep duration loci. Plos Genet. 2016;12(8):e1006125.
111. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506(7488):376–81.
112. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92–4. https://doi.org/10.1038/nature24284.
113. Scott RA, Scott LJ, Magi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An expanded genome-wide association study of type 2 diabetes in Europeans. Diabetes. 2017;66(11):2888–902.
114. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. Hum Mol Genet. 2018;27(20):3641–9. https://doi.org/10.1093/hmg/ddy271.
115. Pattaro C, Teumer A, Gorski M, Chu AY, Li M, Mijatovic V, Garnaas M, Tin A, Sorice R, Li Y, Taliun D. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. Nat Commun. 2016;7(1):1–9.
116. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707–13. https://doi.org/10.1038/nature09270.
117. Malik R, Traylor M, Pulit SL, Bevan S, Hopewell JC, Holliday EG, et al. Low-frequency and common genetic variation in ischemic stroke: the METASTROKE collaboration. Neurology. 2016;86(13):1217–26. https://doi.org/10.1212/WNL.0000000000002528.
118. Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011;43(4):333–8. https://doi.org/10.1038/ng.784.
119. Wellcome Trust Case Control Consortium. 2009. https://www.wtccc.org.uk.
120. UK Biobank. 2021. https://www.ukbiobank.ac.uk.
121. Simons Foundation Autism Research Initiative. Simons Foundation Powering Autism Research. 2019. https://www.sfari.org/resource/spark/.
122. ReMap. 2018. http://pedagogix-tagc.univ-mrs.fr/remap/.
123. European Genome-phenome Archive. The Cardiogenics study. EGAS00001000411. Transcriptome Analysis. 2013. https://ega-archive.org/studies/EGAS00001000411.
124. The PsychENCODE Consortium. Human Brain Development. 2018. http://development.psychencode.org/.
125. BrainSpan. Developmental transcriptome. 2013. http://brainspan.org/static/home.
126. Zhang Y. SUPERGNOVA. Github. 2021. https://github.com/qlu-lab/SUPERGNOVA.

Publisher's Note