Title: Gene dosage architecture across complex traits

Authors:

Sayeh Kazem*, Kuldeep Kumar*, Martineau Jean-Louis, Guillaume Huguet, Thomas Renne, Zohra Saci, Worrawat Engchuan, Omar Shanta, Bhooma Thiruvahindrapuram, Jeffrey R. MacDonald, Celia M. T. Greenwood, Stephen W. Scherer, Laura Almasy, Jonathan Sebat, David C Glahn, Guillaume Dumas**, Sébastien Jacquemont**.

Affiliations:

1 Centre de recherche Azrieli, CHU Sainte-Justine and University of Montréal, Canada

2 Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, PA, USA

3 Lifespan Brain Institute, Children's Hospital of Philadelphia and Penn Medicine, PA, USA

4 Department of Genetics, University of Pennsylvania, PA, USA

5 Harvard Medical School, Department of Psychiatry, 25 Shattuck St, Boston, MA, USA

6 Boston Children's Hospital, Tommy Fuss Center for Neuropsychiatric Disease Research, 300 Longwood Avenue, Boston, MA, USA

7 Department of Psychiatry, University of California San Diego, La Jolla, CA, USA

8 The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada

9 Mila – Quebec AI Institute, University of Montréal, Canada

*** authors contributed equally**

**** Shared senior authorship**

Long Abstract

Copy number variants (CNVs) have large effects on complex traits, but they are rare and remain challenging to study. As a result, our understanding of biological functions linking gene dosage to complex traits remains limited, and whether these functions sensitive to gene dosage are similar to those underlying the effects of rare single nucleotide variants (SNVs) and common variants remains unknown.

Methods We developed FunBurd, a functional burden analysis, to test the association of CNVs aggregated within functional gene sets. We applied this approach in 500,000 individuals from the UK Biobank to associate 43 complex traits with CNVs disrupting 172 gene sets across tissues and cell types. We compared CNV findings with those from common variants and LoF (Loss of Function) SNVs in the same cohort using the same functional gene sets.

Results. All 43 traits showed FDR significant associations with CNVs. Brain tissue and neuronal cell-types showed the highest levels of pleiotropy. Most of the functional gene set associations could, in part, be explained by genetic constraint, except for brain related processes. Shared genetic contributions between pairs of traits were concordant across types of variants, but on average 2-fold higher, for rare CNVs and SNVs compared to common variants.

Functional enrichment across traits found limited overlap between CNVs and common variants. Moreover, the effects of deletions and duplications were negatively correlated for most traits.

In conclusion, we present new methods to separate the contributions of genetic constraint and gene function to the associations of CNVs with complex traits. Overall, the functional convergence between different types of variants -even between deletions and duplications- remains limited.

Short Abstract

Copy number variants (CNVs) have large effects on complex traits, but they are rare and remain challenging to study. As a result, our understanding of biological functions linking gene dosage to complex traits remains limited.

FunBurd (functional burden analysis) was applied in 500,000 individuals from the UKBiobank to associate 43 complex traits with CNVs disrupting 172 tissues/cell type gene sets.

All 43 traits showed associations with CNVs. Brain tissue and neuronal cell types showed the highest levels of pleiotropy. Most of the functional gene set associations could, in part, be explained by genetic constraint, except for brain-related processes. Shared genetic contributions between pairs of traits were 2-fold higher for rare compared to common variants. The functional convergence between different types of variants -even between deletions and duplications- was limited. Our findings provide insights to separate the contributions of genetic constraint and gene function to the associations of CNVs with complex traits.

Introduction

Deciphering functions at the cell and tissue level mediating genetic associations with complex traits and conditions is a central question in human genetics[1–4]. It can offer insights into the molecular understanding of causal processes of diseases[1]. Genome-wide association studies (GWAS) have identified thousands of common variants associated with complex traits[1,2]. While functional genomic studies have shed light on the tissues and cell types disrupted by these variants and how they impact gene expression levels, it remains challenging to link common non-coding variants to genes and corresponding functions[1,4–7].

In contrast, copy number variants (CNVs) – deletions or duplications of DNA segments – invariably lead to a large decrease or increase in gene expression when a gene is fully encompassed in a deletion or a duplication, respectively [8–11]. Therefore, they provide direct insight into the effects of large changes in gene expression on complex traits. However, studies have been limited to a few of the most recurrent CNVs [10,12,13] due to statistical power. As a result, our knowledge of the CNV architecture of complex traits and biological functions that are sensitive to gene dosage remains limited.

CNVs are commonly screened in individuals referred to clinics for complex neurodevelopmental disorders and congenital malformations[14–16]. Since most clinical studies have focused on pediatric populations, there is limited understanding of the medical issues that may arise later in life. Coding CNVs are associated with broad pleiotropic effects[9–11,17,18], yet it remains unclear whether these pleiotropic effects are driven by the multigenic nature of CNVs or by particular genes (and corresponding functions) encompassed in the CNVs. Due to statistical power, previous studies have repeatedly analyzed a small set of the most frequently recurrent CNVs[16,18,19], which collectively affect only approximately 2% of the coding genome[20]. As a result, our understanding of gene functions sensitive to gene dosage is highly biased.

Previous studies have shown that measures of evolutionary/genetic constraint (such as LOEUF score[21]) explain a significant proportion of the effects of coding CNVs on cognitive, and behavioral traits[22–25]. It is however unclear if the biological function of genes can inform the effects of CNVs beyond constraint metrics. In particular, it is unknown if the effects of CNVs on traits that are not under genetic constraint can be explained by gene function.

Our overarching aim is to systematically characterize the CNV architecture of complex traits (brain and non-brain) across different levels of biological functions covering tissue and cell

types, and compare them with the common variant architecture. To study all rare coding CNVs (beyond those sufficiently frequent to establish locus-level association), we implemented the functional burden association test (FunBurd)[25], which aggregated CNVs that disrupt genes with similar functions, defined through shared expression patterns across tissues and cell types (**Figure 1**).

We quantified the effect size of biological functions disrupted by CNVs for each trait by identifying genes fully present in CNV regions. We then compared these effect sizes to those of common variants and rare SNVs. Our outcome measures included 43 complex traits from 500,000 UK Biobank participants. These traits were categorized into two broad groups: brain-related and non-brain-related. Additionally, they were divided into five specific categories: cognitive metrics, mental health, blood assays, physical measures, and reproductive and activity factors. We, therefore, compared these effect sizes to those comprehensively characterized at the common variant and SNV level[17,19,26–29]. This is to our knowledge, the first large-scale comparison of gene dosage, rare SNVs, and common variants on multiple categories of traits.

Results

Pleiotropy is higher for genes assigned to brain tissue

To identify gene-dosage-sensitive functions relevant to tissues across the human body, we performed functional burden analyses to test the association between 43 complex traits (blood assays, reproductive and activity factors, physical, cognitive, and mental health measures) and 60 gene sets assigned to 60 whole-body tissues which on average showed an overlap of 7.9% (**Supplement Figure 7**). We chose these 43 phenotypes because they were relevant to a broad spectrum of medical conditions, and their associations with common variants and rare SNVs have been well studied [19,26,28]. Out of the 2580 associations (43 traits by 60 gene sets for deletions and duplication separately, Supplement Table), deletions showed a higher proportion of FDR significant associations than duplications (n-del=568 [22%], n-dup=352 [13.6%]; proportion test p-value= 1.9e-15). Overall, brain tissue gene sets showed a higher proportion of significant associations with all complex traits (both brain-related and non-brain-related) compared to non-brain gene sets (p-value=6.5e-66), suggesting a higher level of pleiotropy for CNVs affecting brain tissue compared to non-brain tissue gene sets

(**Figure 2A and 2C, Supplement Figure 1-2**). Sensitivity analyses showed that the proportion of significant associations across brain and non-brain traits and tissues was similar using i) independent bulk RNA-seq data from GTEx[30] and Human Protein Atlas[31]; ii) different measures of the same trait (for example, online versus in-person assessments in UK Biobank), iii) demographic parameters: age (<60 versus > 60 years); sex (male and female only); iv) All ancestries; and v) unrelated traits (**Supplement Material, Supplement Figure 21**).

Gene dosage effects across whole-body and whole-brain cell types

We then asked if differences in pleiotropy observed for tissues were also observed at the cell type level. Aggregating CNVs across 81 gene sets assigned to 81 whole-body single-cell clusters (from HPA), with an average overlap of 4.6% (**Supplement Figure 8**), we performed functional burden analyses across the 43 traits. In general, there was a similar proportion of significant associations with cell type compared to tissue gene sets (p-value=4.9e-13), and deletions showed a higher proportion of significant association than duplications (n-del=494 [14.2%], n-dup=421 [12%]; proportion test p-value= 4.8e-03, **Figure 2B-C, Supplement Figure 3-4**).

Brain and non-brain cell type gene sets showed higher associations with brain (p-value=1.6e-03) and non-brain (p-value=1.3e-03) traits, respectively. Similar to brain tissue, CNVs affecting brain cell types exhibited a higher proportion of significant associations compared to those affecting non-brain cell types (p-value=1.3e-03, **Figure 2C**). This remained after removing correlated traits, suggesting higher pleiotropy for gene sets assigned to brain cells (**Supplement Figure 21**).

Building on these observations, we focused specifically on human brain cell types, using single-cell RNA-seq data from 31 distinct categories of adult brain cell types with an average overlap of 7.5% (**Supplement Figure 9**). Deletions showed a higher proportion of significant association than duplications (n-del=196 [14.7%], n-dup=162 [12.1%], p=2.7e-02; **Supplement Figure 5-6**). Inhibitory neurons, ependymal, oligodendrocytes, microglia, and the choroid plexus showed the broadest level of pleiotropy, while fibroblasts, vascular cells, and splatter neurons were not associated with any trait.

For neuronal gene sets, duplications had a higher proportion of associations with brain-related traits (p-value=1.2e-09). Whereas deletions were mostly associated with Non brain traits (p-value=8.1e-08) **(Supplement Figure 5-6)**. More specifically, within brain-related traits—including cognitive metrics and mental health—duplications exhibited a greater proportion of associations with both inhibitory and excitatory neurons compared to deletions. Interestingly, for cognitive metrics, deletions showed a higher number of associations than duplications **(Figure 3)**. These findings suggest distinct effects of deletions and duplications on brain cell types.

We found high concordance between deletion/duplication profiles for microglia, astrocytes, oligodendrocytes, and oligodendrocyte precursor cells generated from whole-body and brain cell type resources, with correlation magnitudes ranging from 0.23 to 0.88 **(Supplement Figure 22)**.

CNV burden correlations are concordant with previously reported genetic correlations across pairs of traits

Studies often correlate common-variant effect sizes across traits, providing evidence of shared biological mechanisms[26,32]. This approach has recently been extended to Loss of Function(LoF) SNVs[26] but has not yet been investigated for CNVs. We estimated the CNV-burden effect size correlation between pairs of traits, separately for deletions and duplications, and tested their concordance between each other, as well as, against previously published single nucleotide polymorphism (SNP)-based and SNV-based genetic correlations[26] (**Figure 4A, Supplement Figure 12-17**). The deletion burden effect size correlations were highly concordant with those of the duplications (Total Least Squares (TLS) regression slope: 1.09, Concordant Correlation Coefficient (CCC)=0.63, 95%CI=[0.59,0.65]) (**Figure 4B, Supplement Figure 10, Supplement Tables**). The CNV burden effect size correlations were concordant with LoF SNV-based genetic burden correlations (CCC of 0.52 for deletion and 0.69 for duplication, **Figure 4D, Supplement Tables**), and on average, 2.12-fold higher than SNP-based genetic correlation across traits (CCC=0.52, 95%CI=[0.47,0.57], **Figure 4C, Supplement Tables**). We investigated whether the directionality (sign concordance) of the effect size and genetic correlations between pairs of traits were consistent across types of

variants: deletions, duplications, SNP, and SNV-LoFs. The sign concordance between types of variants ranged from 64.6% to 77.4% across all traits (**Supplement Tables**). A sensitivity analysis comparing the CNV burden effect size correlations with common variant genetic covariance showed similar concordant results (**Supplement Material**).

Functional overlap between CNVs and common variants

We sought to quantify the potential functional convergence between CNVs and common variants associated with 43 complex traits. To do so, we used the same functional gene sets described above to compute the CNV functional burden association test and functional enrichment of GWAS summary statistics across the same 43 traits in the UKBB (**Methods**).
On average, 7.5% and 4.8% of functional gene sets were overlapping between common variants and CNVs (duplication and deletions, respectively) across all traits (**Figure 5A-B, Supplement Material**). Only 6 brain traits (but none of the non-brain traits) showed significant functional overlap (FDR permutation p-values<0.05, **Methods, Figure 5C, Supplement Figure 18**) between duplications and common variants (**Figure 5A)**. Notably, we did not observe any significant overlap/convergence between deletions and common variants (**Figure 5B**), despite the fact that deletions show a higher number of significant associations with traits compared to duplications (p-value=1.5e-05). To quantify divergence, we need more data and power.

Deletion and duplication show mirror effect sizes across most traits

Genetic association studies are not typically stratified across classes of variants. A recent study showed that different classes of variants show discordant phenotypic effects[26]. Specifically, LoF and missense variants of the same genes showed, on average, a burden genetic correlation of 0.6[26]. CNVs represent a unique opportunity to compare two classes of variants with opposing effects on gene expression[8], which is not possible for SNPs and gene-disrupting SNVs. We first compared the functional pleiotropy between deletions and duplications and observed a moderate correlation (r=0.4, p-value=4.8e-8, **Supplement Figure**

20). This suggested that only some of the functional gene sets showed higher pleiotropy when disrupted by deletions or duplication.

Using the same method described above, we then investigated the functional convergence between deletions and duplications per trait. Significant functional overlap between deletion and duplication was observed in only 2 out of 43 traits (**Figure 5D, Supplement Figure 18**). We sought to investigate this surprisingly small level of convergence and previously reported non-monotonicity[33] of the direction of the average burden effect (i.e. variants that decrease expression have the same genome-wide average direction of effect as variants that increase expression, contradicting the intuition that reducing expression should have the opposite effect on a trait as increasing expression). Towards this, first, we quantified average burden effect sizes, per trait, across 172 functional gene sets, and observed non-monotonicity for all brain traits, and monotonocity for only 4 out of 18 non-brain traits (**Supplement Figure 19**). Second, we examined the effect sizes on each trait of deletions and duplications across all functional gene sets. We observed a significant negative correlation for 5 brain traits and 3 non-brain traits (FDR p-Jaccard<0.05, **Figure 5E, Methods**). This was not related to the category of gene set used to compute effect sizes (tissue versus cell-type gene sets) or the heritability of the traits (**Supplement Material**). This suggests that the association between genes, corresponding functions, and a trait can greatly vary depending on the classes of variants and partitioning the coding genome into functional gene sets recapitulates the expected opposing effect of deletions and duplications.

Dissecting genetic constraint, pleiotropy, and function

Previous studies have shown that common and rare variant heritability enrichments are strongest in constrained genes across complex traits [21,26,28,34,35]. We, therefore, asked if the associations with functional gene sets were driven by function or mainly constraint. We observed that functional gene sets with higher proportions of genes under constraint (LOEUF top-decile, **Methods**) showed a higher level of associations and pleiotropy (**Figure 6A**); for deletion (r=0.39, p-Jaccard<1e-3, **Methods**), and duplication (r=0.44, p-Jaccard<1e-3). This was also the case for functional enrichment analysis computed for the same gene sets and traits using common variant GWAS summary statistics (r=0.52, p-Jaccard<1e-3). This remains true regardless of the constraint metric (**Figure 6B**). Using a functional burden normalized for

constraint, we demonstrate that the vast majority (~90%) of functional gene sets exhibit trait associations within a 90-percent expected range (**Figure 6C, Supplement Figures 23-24**). However, several gene sets exhibited significantly higher pleiotropy than expected based on the proportion of genes under constraint (FDR p-value < 0.05). Notably, brain-related gene sets consistently showed higher constraint-normalized pleiotropy compared to non-brain tissues, respectively (**Figure 6D**).

Discussion

This study represents the first comprehensive investigation into gene functions underlying the association between CNVs and complex traits across tissues and cell types. Genome-wide CNV functional burden tests revealed a higher number of trait associations (i.e., pleiotropy) with brain tissue and neuronal cell types, compared to non-brain tissue and non-neuronal gene sets. We show for the first time that shared genetic contributions between pairs of traits were concordant across classes of variants, but on average 2-fold higher, for deletions, duplications, and LoF SNVs compared to common variants. Common variant GWAS and CNV burden association studies showed limited functional overlap. The negative correlation between deletion and duplication effect size profiles across a majority of traits showed that the association between functional gene sets and traits was dependent on the class of disrupting variants. Normative functional burden modeling demonstrated that over 90% of functional gene set associations could be explained by genetic constraint, with the notable exception of higher pleiotropy for brain tissue and neuronal cell type.

CNV GWAS at the variant level remains challenging due to the sparsity of data. By implementing a functional burden test, we addressed this challenge to some extent. This approach showed that CNVs aggregated across many different biological functions were associated with a broad spectrum of complex traits. While levels of pleiotropy were higher for the brain compared to non-brain tissues and cell types, much of this pleiotropy could be explained by the proportion of genes under constraint in a given gene set (i.e., the fraction of constraint). In fact, disentangling constraint from function has been challenging, and many studies have highlighted the predominant role of constraint in shaping the genetic architecture of complex traits[21,26,27,34]. Using a novel approach to normative constraint modeling showed

that brain tissue and neuronal gene sets showed numbers of associations (pleiotropy) beyond what would be expected based on the fraction of constraint of any given gene set. The fact that much of the coding CNV signal (as well as the LoF SNVs) comes from pleiotropic genes under constraint may also explain why genetic correlations between traits were 2-fold higher for CNVs and LoF SNVs compared to common variants[26].

Our findings indicate limited functional overlap between common variants and rare CNVs, suggesting that biological functions show differential sensitivity to gene dosage versus the common variants. This suggests that different classes of genetic variation might exert their effects through distinct mechanisms[1,6,26].

In general, studies have failed to observe the monotonous effects of deletions and duplications on complex traits [33]. We also show non-monotonous average burden effects across functional gene sets for 90% of traits. However, investigating the association profiles, we observed that most functional gene sets were either sensitive to deletions or duplications, and this deletion/duplication sensitivity varied from one trait to another. This resulted in a negative correlation between the effect sizes of deletions and duplications across the majority of traits. This highlights that the association between a trait and genes (or biological functions) is greatly dependent on the class of variant. Previous studies showed discordant genetic correlations (0.64) between missense and LoF variants across traits, also implying that different classes of variants in the same genes often have divergent phenotypic effects[26]. As opposed to common variants or rare SNVs, deletions, and duplications are the only known variants with clearly opposing molecular effects (on transcription levels[8]). Overall, it appears that biological functions that are sensitive to haploinsufficiency are less triplosensitive and vice versa.

There are some limitations, which need to be considered. While functional burden tests address the challenges of rare variants association, overlapping gene sets introduce new challenges, including dependent results (similar to LD structure). To address this limitation, we propose the p-value Jaccard method, which effectively accounts for the overlapping gene sets. The UK Biobank, as a relatively healthy population cohort, poses limitations for disease analysis due to the under-representation of individuals with severe conditions. This bias may also result in a lower prevalence of deleterious genetic variants, potentially reducing the observed heritability of traits or conditions associated with high genetic burden in this

population. This limitation was most pronounced in our normative constraint modeling, whereas the proportion of constrained genes increased, the sample size for those gene sets decreased, leading to fewer individuals with extreme constraint genes. Larger population cohorts or clinically ascertained cohorts enriched for constrained genes will be crucial for more effectively unraveling the relationship between gene constraint and function.

Our study shows the association between gene biological functions and phenotypes may be strongly modulated by classes of variants, and aggregating heterogeneous variants within genes may provide an incomplete picture of genetic association. Systematic comparison across classes of variants, reveals the concordant structure; however, extending the burden association analysis to different classes of SNVs (LoF and missense effects) will provide a more complete understanding of how rare coding variants contribute to the genetic architecture of complex traits and conditions.

Methods

Participants

The UK Biobank recruited 502,534 individuals (54% female) ages 37 to 73 years, living in the United Kingdom, between 2006 and 2010[36]. Phenotypic measures were collected at the UK Biobank assessment centers (using touchscreen devices), or in an online follow-up. UK Biobank procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. Data were released under application number 40980. The ethics board of CHU Sainte Justine Research Center, Montreal, Canada approved this study.

Genotyping, and CNV calling

UK Biobank DNA from blood samples was genotyped using the UK BiLEVE Axiom (n=49,950, 820k probes) and UK Biobank Axiom arrays (n=438,427, 807k probes)[37]. Quality control and processing followed our previously published pipeline[23,25,38]. 733,256 shared, QC-passed, hg19-mapped biallelic probes were used. Samples with high missingness (mind>0.05,

|waviness factor|<0.05, log R ratio SD<0.35, B allele frequency SD<0.08) were excluded (n=28,522), leaving 459,855 individuals. CNVs were called using PennCNV[39] and QuantiSNP[40], combined with CNVision[41], and concatenated with CNVs INHERITANCE ANALYSIS (CIAv.2.0), following our published pipeline (https://martineaujeanlouis.github.io/MIND-GENESPARALLELCNV/#/)[23,25,38,42]. CNVs were called using both algorithms with the following parameters: number probe coverage per CNV ≥3, CNV size ≥1Kb, and confidence scores ≥ 15. CNVs detected by both algorithms were combined according to their types using CNVision to minimize the number of potential false discoveries. Following the data merging steps, CNVs were concatenated using the CNVs INHERITANCE ANALYSIS (CIAv.2.0) algorithm using the following criteria: a) CNV gapping ≤150 kb; b) CNV size ≥ 1000 bps; and c) number of probes ≥ 3. CNVs were filtered according to previously published studies[38]. We kept CNVs passing inclusion criteria such as confidence score ≥ 30 (with at least one of both detection algorithms), size ≥ 50 kb, and unambiguous type (deletions or duplications). All recurrent CNVs were verified visually.

CNVs were annotated using Gencode version 35 lifted to hg19 coordinates (https://www.gencodegenes.org/human/release_35lift37.html) We used bedtools intersect (https://bedtools.readthedocs.io/en/latest/) to identify breakpoints overlapping coding DNA elements such as UTRs, start and stop codons, exons, and introns[23,25,38].

Gene-sets

Gene sets were defined based on expression patterns across tissues and cell types. We used three resources:

a) Whole-body tissue expression data from FANTOM (Functional Annotation of the Mammalian Genome)[43] – FANTOM provides whole-body normalized gene expression data (normalized Transcripts Per Million) across 46 tissues. For sensitivity analysis, we used two independent whole-body tissue expression data: GTEx [30] and the Human Protein Atlas [44] (https://www.proteinatlas.org/about/download; version 23).

b) Whole-body single-cell data from the Human Protein Atlas (HPA)[31], which provides expression profiles across 81 cell types from 31 human tissues, based on single-cell RNA-seq (scRNAseq) data. In our analysis, we used the average expression data for 81 cell-type super-clusters across protein-coding genes (https://www.proteinatlas.org/about/download; version 23).

c) Whole brain single-cell data from the Human Brain Cell Atlas v1.0 (snRNAseq)[45] , which consists of 3.369 million nuclei successfully sequenced using snRNAseq across 106 anatomical locations in adult brains. We used cell types from 31 superclusters and 461 clusters aggregated data from (https://github.com/linnarsson-lab/adult-human-brain).

Gene-set creation

We use the Top Decile Expression Proportion (TDEP) approach [27,29] to create tissue and cell-type genesets. To do so, we divided each gene's expression in a given tissue or cell type by its total expression across all tissues or cell types, producing values between 0 and 1. This metric identifies the proportion of a gene's total expression attributed to a specific tissue or cell type. We then selected the top decile of these expression proportion measures to create each gene set.

Traits preprocessing and normalization

We extracted 43 binary and continuous complex traits from the UK Biobank (UKBB). For traits with multiple measures over time, we only considered data from the initial visit and assessment. Individuals with missing trait measures were removed from the dataset. We excluded outliers by considering values beyond ±6 standard deviations (SD). To be consistent with previous GWAS for the same UKBB continuous traits, we used PHESANT (ref), an inverse rank normalization transformation (IRNT). This method involves ranking the continuous data and then transforming these ranks into quantiles of the standard normal distribution[46].

Functional Burden Association Test (FunBurd)

FunBurd is designed to test the association between variants aggregated across a gene set and a given trait. The traits of interest were considered as a function of the number of genes within the gene set disrupted by CNVs. To avoid effect size inflation, due to multigenic CNVs, we adjusted for the number of genes (not members of the gene set) disrupted by the same CNV. We also adjusted with age, sex, and ancestry.

$$Trait = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 age + \beta_4 sex \ + \sum_{i=1}^{10} PC_i + \ \varepsilon$$

The coefficient B_1 represents the effect size of a given gene set on the trait of interest. In this analysis, we examined 14,792 associations, derived from 172 gene sets, 43 traits, and 2 types of variations (deletions and duplications). Our group has recently explored similar burden analysis models for the analysis of a single trait[25,47,48].

Multiple comparison correction

All analyses were corrected for multiple testing using the False Discovery Rate (FDR). The FunBurd considered the entire matrix, which includes all merged deletion-duplication events and gene sets, totaling 14,792 associations.

Functional Pleiotropy

"Functional pleiotropy" was defined as the number of traits significantly associated with a functional gene set.

Permutation preserving Jaccard distance (P-Jaccard)

Because gene sets partially overlap, correlations between gene sets may be inflated. To adjust the p-values of such correlations, we developed a method based on the Jaccard distance matrix. This approach performs permutations conditioned on the Jaccard distance of the gene sets and thus avoids inflated p-values by generating plausible null distribution [49–51].

GWAS summary statistics, heritability estimates, and genetic correlations

We included the GWAS summary statistics for the same traits derived from nearly the same set of participants from the UK Biobank. GWAS summary statistics, heritability estimates, and genetic correlations, for all the UK Biobank traits, are available at https://www.nealelab.is/uk-biobank.

Comparing functional overlap between CNVs and common variants

To systematically compare functional enrichments between CNVs and common variants, we used the same gene sets and UKBB traits to run functional burden association tests

(FunBurd) and Stratified LD Score Regression (S-LDSC). S-LDSC assesses whether certain biological functions are enriched for common variant GWAS signals [27].

Functional overlap between CNVs and SNPs was defined as the proportion of gene sets enriched jointly in CNVs (FunBurd) and common variants associations (S-LDSC). The Functional overlap between deletions and duplications was the proportion of gene sets associated with both types of CNVs. Significance was tested by fixing the gene sets enriched in common variants and permuting CNV-associated gene sets 1000 times. We then tested whether the observed functional overlap was greater than expected by chance (null distribution).

CNV effect size correlations

We computed effect size correlations between pairs of traits by correlating the effect size profiles using Spearman correlation(Fig 3B). For each trait, we included only the gene sets that were FDR significant for either deletion or duplication. We only considered traits that had at least 10% significant genesets (Fig 4E).

LoF SNV summary statistics of burden correlations

SNV loss of function (LoF) summary stats and burden correlations were downloaded from [26]. Notably, these results were derived from the same set of UK Biobank participants [26].

Genetic constraint metrics

We used several measures of constraint including: i) the loss-of-function observed/expected upper bound fraction (LOEUF) score[21]; ii) probability of being loss-of-function intolerant (pLI)[21]; iii) CDS: percentage of coding sequence (CDS) under evolutionary constraint[34]; iv) two measures of dosage sensitivity [52]: probability of haploinsufficiency (pHaplo), and probability of triplosensitivity (pTriplo). v) s-het [53], a constraint metric predicted from multiple transcriptomic and gene features using machine learning; and vi) gene-length [21,52] in base pairs.

Normative constraint modeling

We developed a normative model to evaluate whether the association between functional

gene sets and traits is explained by genetic constraint. We generated a null distribution by randomly sampling gene sets 100 times (n = 1,298 genes; without replacement) and using different proportions within constraint (top decile of LOEUF scores) and remaining (other LOEUF score deciles), then calculated the corresponding number of trait associations (see **Supplementary Figure 25** for a summary of the implementation). This approach allowed us to define the expected relationship between the number of trait associations and the proportion of genes under constraint. Our normative model revealed a sharp increase in trait associations for gene sets with a constraint fraction exceeding 0.25, highlighting a nonlinear relationship between genetic constraint and pleiotropy. By providing a constraint-normalized effect size for each gene set, this approach offers a more refined interpretation of functional pleiotropy.

References

1. Lappalainen, T., Li, Y. I., Ramachandran, S. & Gusev, A. Genetic and molecular architecture of complex traits. *Cell* **187**, 1059–1075 (2024).
2. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).
3. Sullivan, P. F., Yao, S. & Hjerling-Leffler, J. Schizophrenia genomics: genetic complexity and functional insights. *Nature Reviews Neuroscience* 1–14 (2024).
4. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
5. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
6. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* 1–10 (2023).
7. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).

8. Han, L. *et al.* Functional annotation of rare structural variation in the human brain. *Nat. Commun.* **11**, 2990 (2020).
9. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
10. Hujoel, M. L. A. *et al.* Influences of rare copy-number variation on human complex traits. *Cell* **185**, 4233–4248.e27 (2022).
11. Hujoel, M. L. A. *et al.* Protein-altering variants at copy number-variable regions influence diverse human phenotypes. *Nat. Genet.* (2024) doi:10.1038/s41588-024-01684-z.
12. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J. Med. Genet.* **56**, 131–138 (2019).
13. Kendall, K. M. *et al.* Cognitive Performance Among Carriers of Pathogenic Copy Number Variants: Analysis of 152,000 UK Biobank Subjects. *Biol. Psychiatry* **82**, 103–110 (2017).
14. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).
15. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).
16. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
17. Auwerx, C., Kutalik, Z. & Reymond, A. The pleiotropic spectrum of proximal 16p11.2 CNVs. *Am. J. Hum. Genet.* **0**, (2024).
18. Auwerx, C. *et al.* Rare copy-number variants as modulators of common disease susceptibility. *Genome Med.* **16**, 5 (2024).
19. Auwerx, C. *et al.* The individual and global impact of copy-number variants on complex human traits. *Am. J. Hum. Genet.* **109**, 647–668 (2022).
20. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
21. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456

humans. *Nature* **581**, 434–443 (2020).

22. Mollon, J., Almasy, L., Jacquemont, S. & Glahn, D. C. The contribution of copy number variants to psychiatric symptoms and cognitive ability. *Mol. Psychiatry* **28**, 1480–1493 (2023).
23. Alexander-Bloch, A. *et al.* Copy Number Variant Risk Scores Associated With Cognition, Psychopathology, and Brain Structure in Youths in the Philadelphia Neurodevelopmental Cohort. *JAMA Psychiatry* (2022) doi:10.1001/jamapsychiatry.2022.1017.
24. Jacquemont, S. *et al.* Genes To Mental Health (G2MH): A framework to map the combined effects of rare and common variants on dimensions of cognition and psychopathology. *Am. J. Psychiatry* (2021).
25. Huguet, G. *et al.* Effects of gene dosage on cognitive ability: A function-based association study across brain and non-brain processes. *bioRxiv* 2024.04.16.589618 (2024) doi:10.1101/2024.04.16.589618.
26. Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* 1–8 (2023).
27. Yao, S. *et al.* Connecting genomic results for psychiatric disorders to human brain cell types and regions reveals convergence with functional connectivity. *bioRxiv* (2024) doi:10.1101/2024.01.18.24301478.
28. Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nat. Commun.* **12**, 1164 (2021).
29. Bryois, J. *et al.* Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat. Genet.* **52**, 482–493 (2020).
30. Consortium, G. & GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* vol. 550 204–213 Preprint at https://doi.org/10.1038/nature24277 (2017).
31. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
32. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

33. Milind, N., Smith, C. J., Zhu, H., Spence, J. P. & Pritchard, J. K. Buffering and non-monotonic behavior of gene dosage response curves for human complex traits. *medRxiv* 2024.11.11.24317065 (2024) doi:10.1101/2024.11.11.24317065.

34. Sullivan, P. F. *et al.* Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science* **380**, eabn2937 (2023).

35. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *bioRxiv* (2024) doi:10.1101/2023.05.19.541520.

36. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

37. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

38. Mollon, J. *et al.* Impact of Copy Number Variants and Polygenic Risk Scores on Psychopathology in the UK Biobank. *Biol. Psychiatry* (2023) doi:10.1016/j.biopsych.2023.01.028.

39. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).

40. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013–2025 (2007).

41. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).

42. Huguet, G. *et al.* Genome-wide analysis of gene dosage in 24,092 individuals estimates that 10,000 genes modulate cognitive ability. *Mol. Psychiatry* (2021) doi:10.1038/s41380-020-00985-z.

43. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

44. Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, (2020).

45. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382**,

eadd7046 (2023).

46. Jonsdottir, G. A. *et al.* Genetic propensities for verbal and spatial ability have opposite effects on body mass index and risk of schizophrenia. *Intelligence* **88**, 101565 (2021).
47. Kumar, K. *et al.* Title: Mirror effect of genomic deletions and duplications on cognitive ability across the human cerebral cortex. *bioRxiv* 2025.01.06.631492 Preprint at https://doi.org/10.1101/2025.01.06.631492 (2025).
48. Kumar, K. *et al.* 295. Rare variant genetic architecture of the human cortical MRI phenotypes in general population. *Biol. Psychiatry* **95**, S220–S221 (2024).
49. Burt, J. B., Helmer, M., Shinn, M., Anticevic, A. & Murray, J. D. Generative modeling of brain maps with spatial autocorrelation. *Neuroimage* **220**, 117038 (2020).
50. Viladomat, J., Mazumder, R., McInturff, A., McCauley, D. J. & Hastie, T. Assessing the significance of global and local correlations under spatial autocorrelation: a nonparametric approach. *Biometrics* **70**, 409–418 (2014).
51. Clifford, P., Richardson, S. & Hémon, D. Assessing the significance of the correlation between two spatial processes. *Biometrics* **45**, 123–134 (1989).
52. Collins, R. L. *et al.* A cross-disorder dosage sensitivity map of the human genome. *Cell* **185**, 3041–3055.e25 (2022).
53. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat. Genet.* **56**, 1632–1643 (2024).

Figures

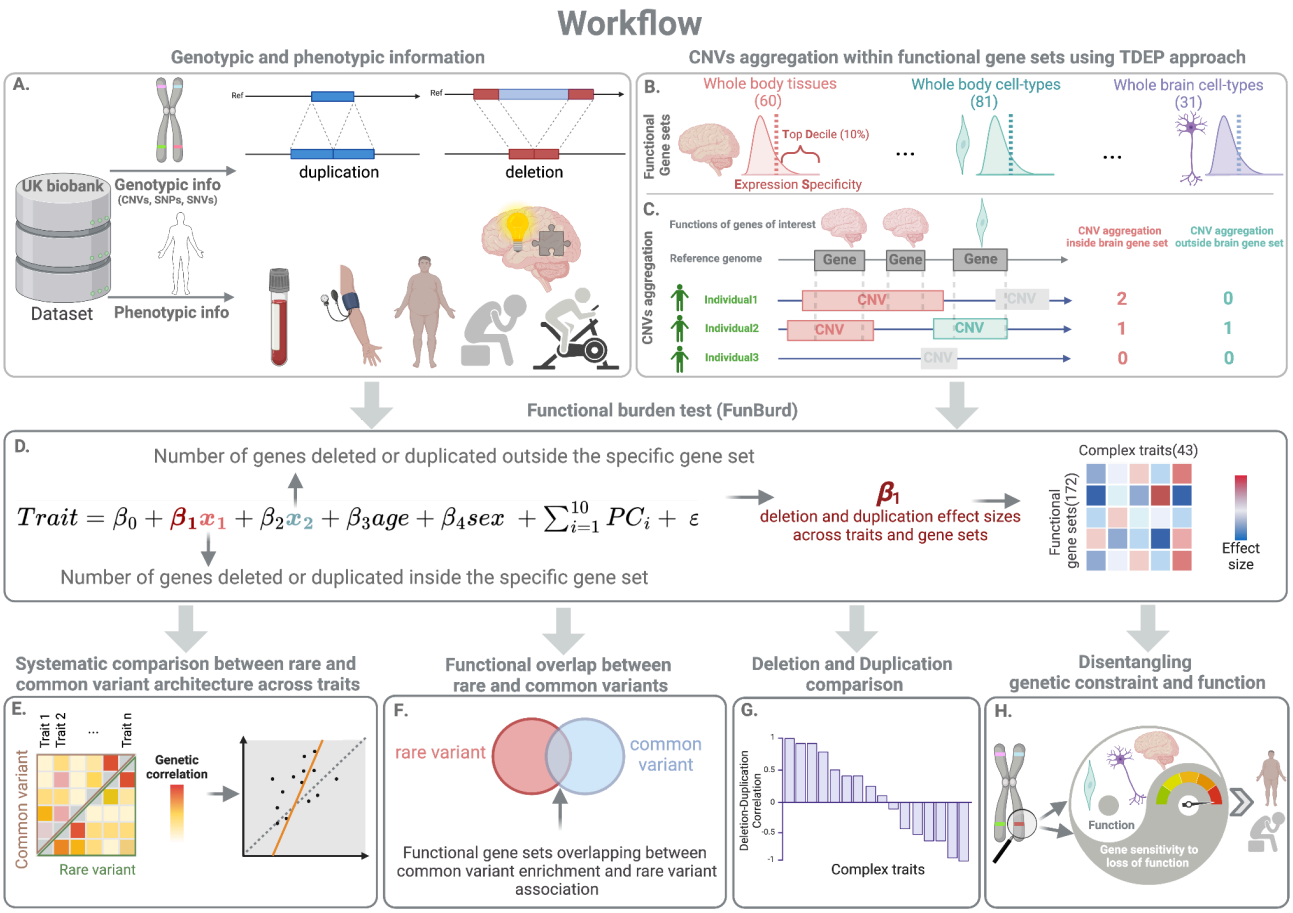

Figure 1 | Estimating CNV effects on complex traits and downstream comparative analysis. **A)** We used CNVs and 43 phenotypes from the UK Biobank. **B)** We defined partially overlapping functional gene sets for 60 tissues, 81 and 31 whole body and brain cell types using TDEP (Top Decile Expression Proportion). **C)** CNVs were aggregated within each of the 172 gene sets. **D)** We performed functional burden tests to estimate the associations between CNVs disrupting 172 gene sets and 43 complex traits. Tests were performed separately for deletions and duplications. Downstream analyses, systematically investigated (**E)** rare and common variant architectures of complex traits, and the functional overlap between rare and common variants **(F)**. Effect size of deletion and duplication across traits **(G)**, and the relationship between gene constraint and function.

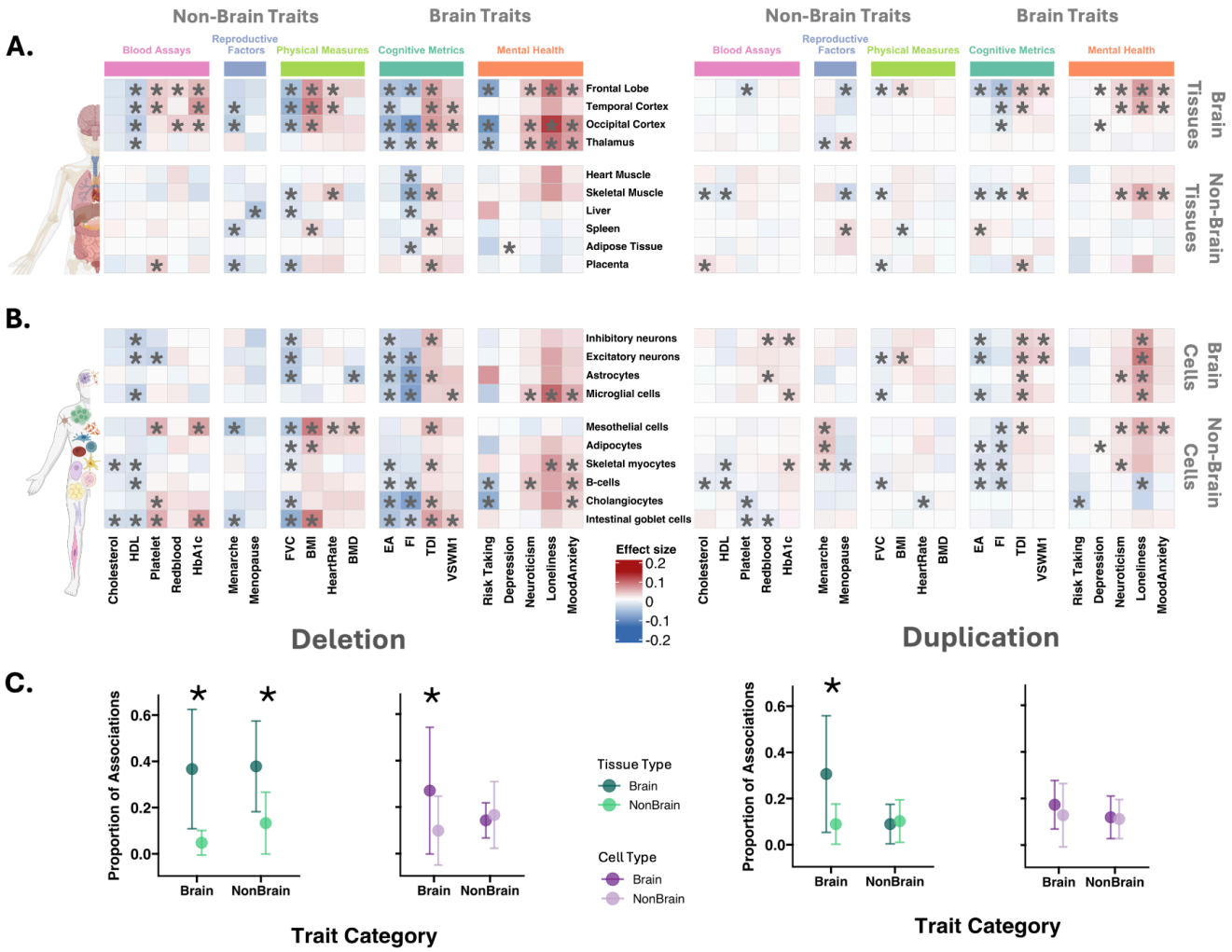

Figure 2 | Heatmap of effect sizes for whole body tissue, cell type gene sets across traits. Heatmap displays a representative set of the most significant associations between 5 categories of traits and deletions and duplications aggregated across **(A)** tissues and **(B)** whole body cell types. Traits are categorized and shown along the x-axis, while the y-axis lists gene sets, tissues, and cell types. The blue and red intensity color scale reflects the negative and positive effect sizes. Black asterisks (*) indicate statistically significant associations between traits and genes (FDR correction across 172 × 43 × 2 =14,792 tests). **C)** Point-range plots (mean and standard error of the mean) summarizing the differences in the level of association between brain and non-brain functional gene sets.

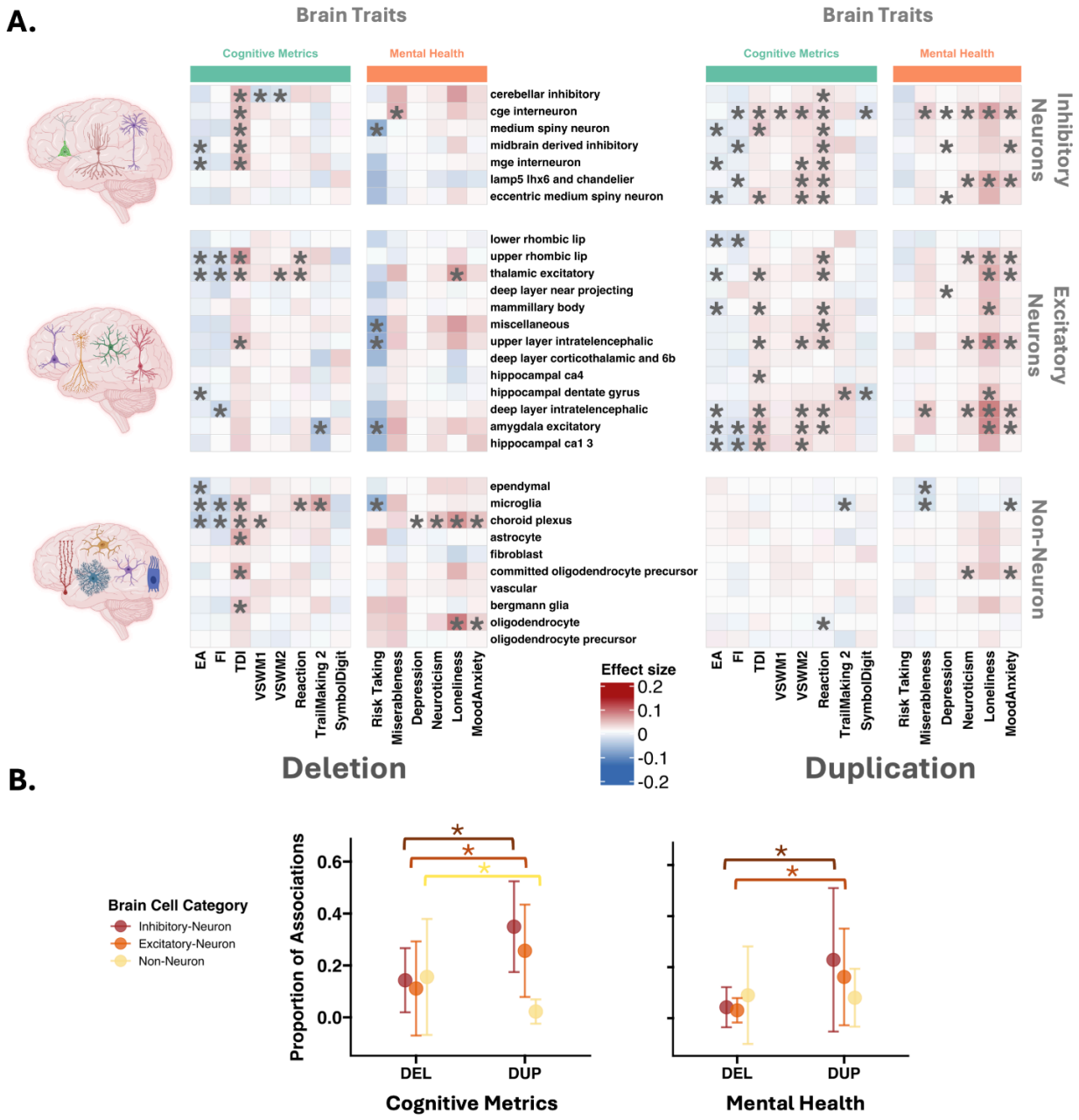

Figure 3 | Heatmap of effect sizes for whole brain cell type gene sets across traits.

Heatmap displaying associations between traits and deletions and duplications aggregated across whole brain cell types. Traits are categorized and shown along the x-axis, while the brain cell-type gene sets categorized as neuronal and non-neuronal, are on the y-axis. Colors indicate effect sizes, with asterisks (*) marking significant associations. (B) Point-range plots summarize association differences between neuronal and non-neuronal functional gene sets.

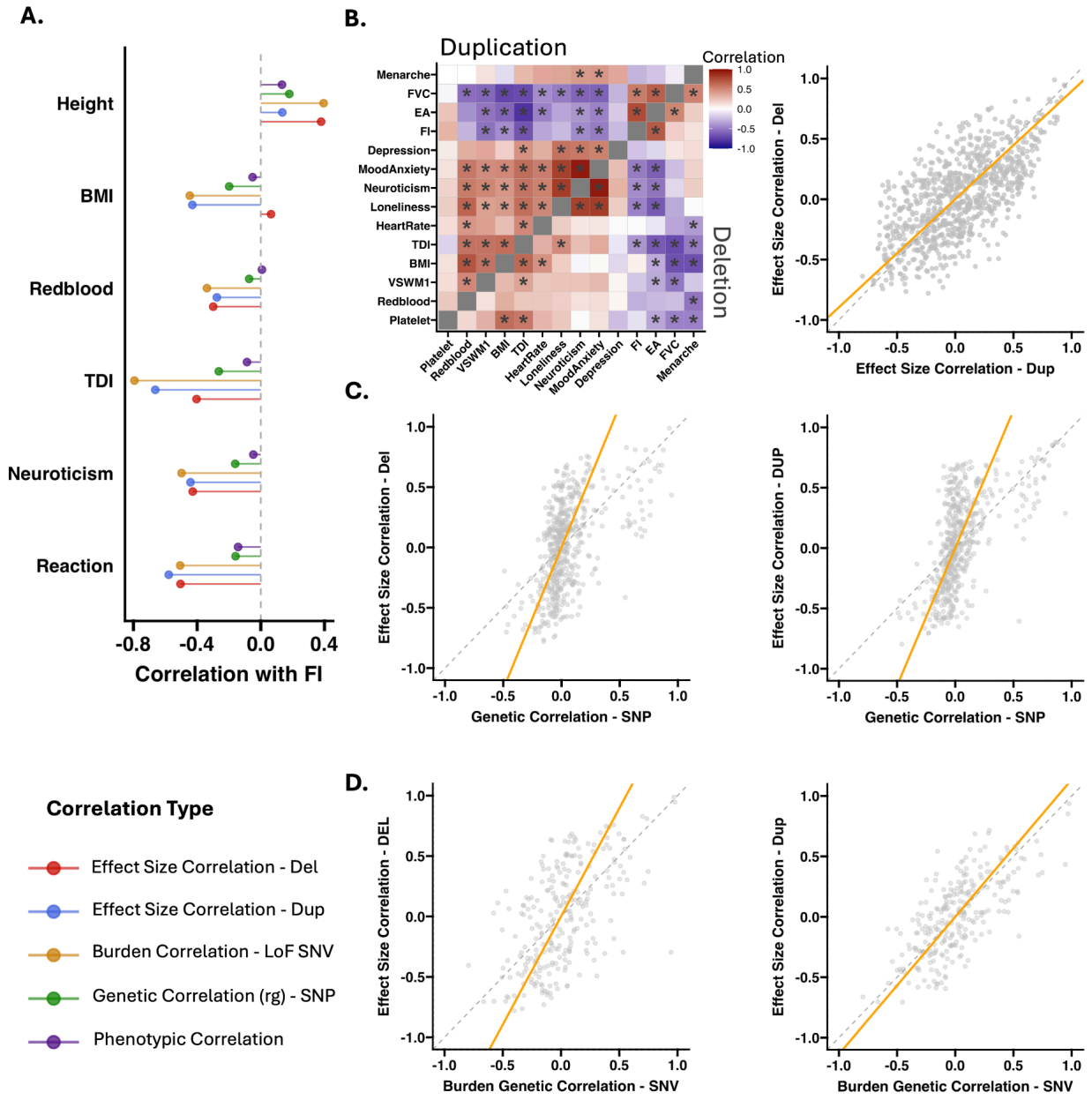

Figure 4 | Rare and common variant architectures across complex traits

A) illustrates the genetic and phenotypic correlations between FI and a subset of traits, with each point color-coded by correlation type: SNP-based genetic correlations, deletion, duplication functional burden effect size correlations, loss-of-function SNV burden genetic correlation, and phenotypic correlations. **B)** on the left, presents a combined heatmap of genetic correlations, showing functional burden effect size correlations for deletions and duplications in the lower and upper triangle matrices, respectively, with

significant correlations marked by asterisks. The scatter plot on the right compares effect size correlations between deletions and duplications, where orange lines represent the TLS (Total Least Squares) regression. The dashed line indicates perfect concordance (y = x). We provide the same comparison between CNVs and **(C)** SNP-based genetic correlations (rg) and **(D)** SNV burden correlations (SNV).

Abbreviations, BMI: body mass index; Del: deletion; Dup: duplication; EA: educational attainment; FI: fluid intelligence; SNP: single nucleotide polymorphism; TDI: townsend deprivation index;

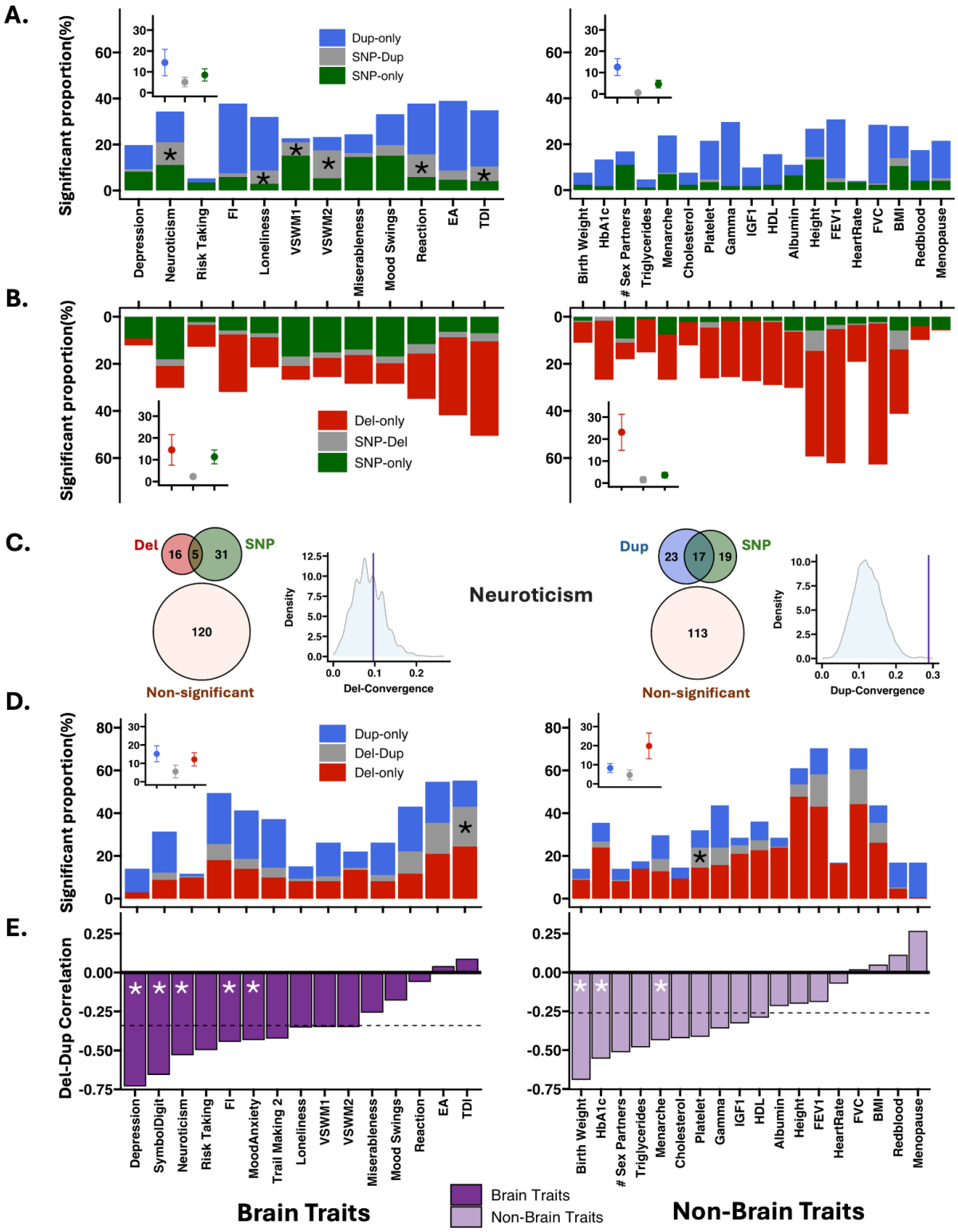
A.
Significant proportion(%)
Dup-only
SNP-Dup
SNP-only
Depression
Neuroticism
Risk Taking
FI
Loneliness
VSWM1
VSWM2
Miserableness
Mood Swings
Reaction
EA
TDI
Birth Weight
HbA1c
Sex Partners
Triglycerides
Menarche
Cholesterol
Platelet
Gamma
IGF1
HDL
Albumin
Height
FEV1
HeartRate
FVC
BMI
Redblood
Menopause
B.
Del-only
SNP-Del
SNP-only
C.
Del
SNP
16
5
31
120
Non-significant
Density
Del-Convergence
Neuroticism
Dup
SNP
23
17
19
113
Non-significant
Dup-Convergence
D.
Dup-only
Del-Dup
Del-only
E.
Del-Dup Correlation
SymbolDigit
MoodAnxiety
Trail Making 2
Brain Traits
Non-Brain Traits

Figure 5 | Functional overlap between common variants and CNVs

A) The stack bar plot represents the proportion of functional genes set overlapping between common variants and duplications for each trait. Overlapping functional gene sets (enriched in common variant and CNV signal) are gray, and non-overlapping functional gene sets are represented in blue (associations only significant for duplications or deletions). Stars represent significant functional overlap (FDR permutation p-value < 0.05). Y-axis: proportion of associated gene sets. X-axis: Brain and non-brain traits. The inserts represent the average convergence (between common variants and duplications) and average duplication-specific associations across traits.

The same information on functional convergence is shown in **(B)** between deletions and common variants (SNPs); and **(D)** between deletions and duplications. **C)** illustrates the example of the functional overlap between SNPs and CNVs for Neuroticism using a Venn diagram. The density plot represents the null distribution of overlap significance (1000 permutations), with the observed overlap shown by the purple vertical line.

E) represents the correlation between deletion and duplication effect sizes across the brain and non-brain traits. Stars represent significant correlations accounting for overlapping gene sets (FDR-corrected permutation preserving Jaccard distance (P-Jaccard)). The Y-axis represents correlation (Spearman r value). X-axis: Brain and non-brain traits.

Abbreviations, BMI: body mass index; Del: deletion; Dup: duplication; EA: educational attainment; FI: fluid intelligence; SNP: single nucleotide polymorphism; TDI: townsend deprivation index;

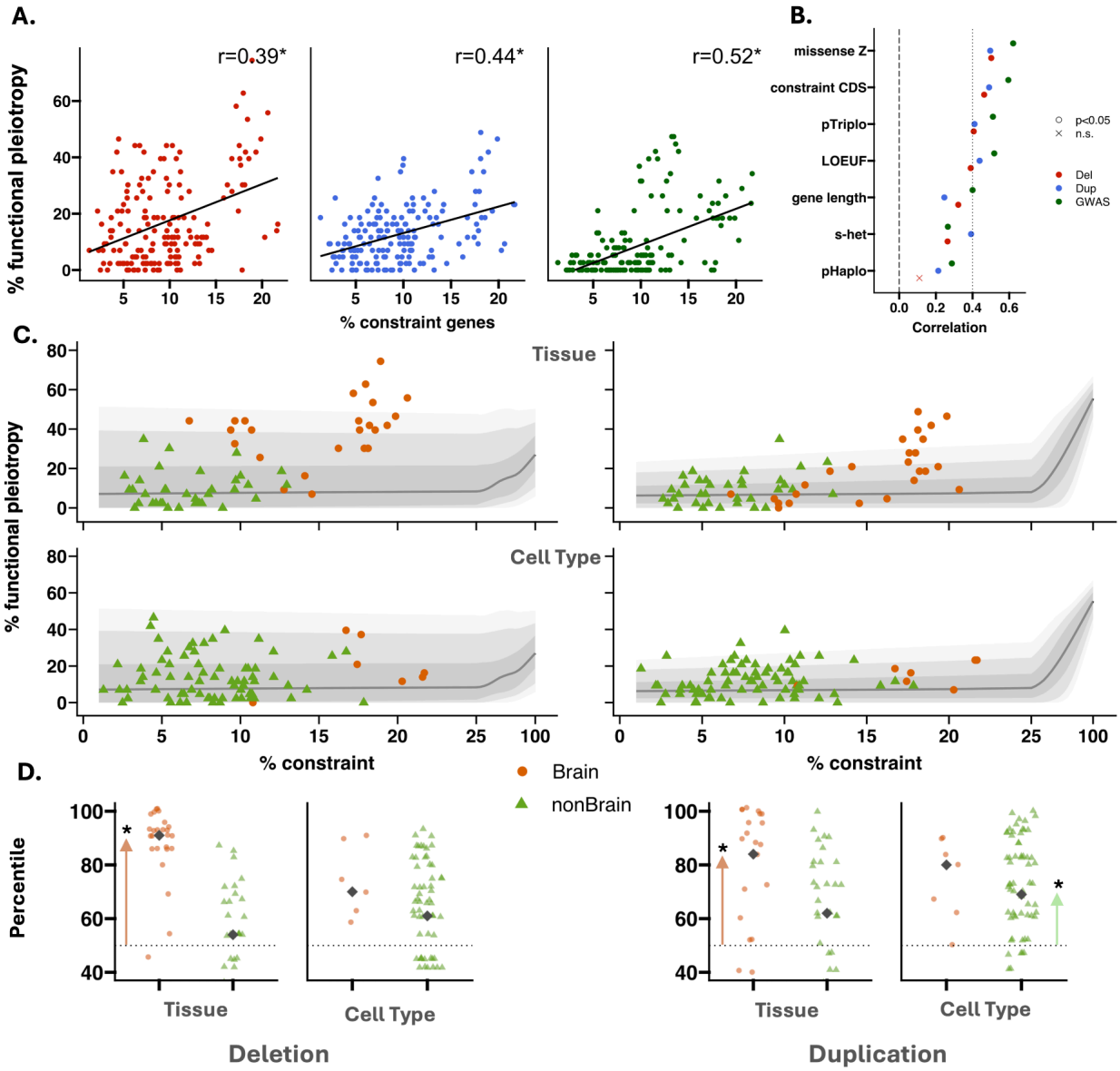

Figure 6 | Dissecting genetic constraint and function

A) Correlation between the fraction of constraint genes (LOEUF top-decile) within a gene set and the number of traits showing significant associations with that gene set for deletions and duplications. Each data point is a gene set. Y-axis: the percentage of significantly associated traits for a variant; X-axis: the percentage of top decile LOEUF genes within the gene set. **B)** shows the same information on correlations in panel **(A)** for different constraint metrics. **C)** depicts the normative relationship between the proportion of intolerant genes (constraint fraction) in randomly sampled gene sets and the proportion of significant trait associations

across different gene set types (rows) and CNV types (columns). The x-axis represents the proportion of intolerant genes, while the y-axis shows the proportion of significant associations across traits (functional pleiotropy). Each shape denotes observed gene sets, categorized into two distinct groups (Brain; and nonbrain). The gray line indicates the expected trend (normative model, 50th percentile), with the shaded gradient showing the range between 25/75 (ribbon 1), 10/90 (ribbon 2), and 3/97 (ribbon 3) centiles. **D)** Stip charts show centiles (normalized functional pleiotropy with respect to the null model) for the 2 groups (Brain; and nonbrain). Arrows and * represent FDR-significant ($q<0.05$) deviation of the median from the 50th centile.

Abbreviations, CDS: coding sequence; Del: deletion; Dup: duplication; GWAS: genome wide significant association; LOEUF: the loss-of-function observed/expected upper bound fraction; n.s.: non significant; pHaplo: the probability of haploinsufficiency; pTriplo: the probability of triplosensitivity; SC: single cell; s-het: the fitness reduction for heterozygous carriers of an loss-of-function in any given gene; SNP: single nucleotide polymorphism;

Additional Information

Data and materials availability

UK Biobank data was downloaded under the application 40980, and may be accessed via their standard data access procedure (see http://www.ukbiobank.ac.uk/register-apply). UK Biobank CNVs were called using the pipeline developed in the Jacquemont Lab, as described at https://github.com/labjacquemont/MIND-GENESPARALLELCNV. The final CNV calls are available for download from the UK Biobank returned datasets (Return ID: 3104, https://biobank.ndph.ox.ac.uk/ukb/dset.cgi?id=3104). References to the processing pipeline and R package versions used for analysis are listed in the methods. GWAS summary stats, heritability estimates, and genetic correlations, for all the UK Biobank traits, are publicly available and were downloaded from the NealeLab: https://www.nealelab.is/uk-biobank. SNV loss of function summary stats and burden correlations are publicly available (Genebass: https://app.genebass.org) and were downloaded from[26]. Gene-level constraint data are available at https://gnomad.broadinstitute.org and other publications listed in methods.

Funding

This research was supported by Calcul Quebec (http://www.calculquebec.ca) and Compute Canada (http://www.computecanada.ca), the Brain Canada Multi-Investigator initiative, NIH U01 grant for CAMP (1U01MH119690-01), the Canadian Institutes of Health Research, CIHR_400528, The Institute of Data Valorization (IVADO) through the Canada First Research Excellence Fund, Healthy Brains for Healthy Lives through the Canada First Research Excellence Fund. Dr Jacquemont is a recipient of a Canada Research Chair in neurodevelopmental disorders and a chair from the Jeanne et Jean Louis Levesque Foundation.

Disclosures

All other authors reported no biomedical financial interests or potential conflicts of interest.

Author contributions

S.K., K.K., G.D., and S.J. designed the study, analyzed data, and drafted the manuscript. M.J.L., Z.S., and G.H. performed CNV calling and quality control. All authors contributed to the interpretation of the results and the editing of the manuscript.