Determinants of pleiotropy and monotonic gene dosage responses across human traits

Authors: Sayeh Kazem[1]*, Kuldeep Kumar[1]*, Guillaume Huguet[1], Josephine Mollon[2,3], Thomas Renne[1], Laura M. Schultz[4,5,6], Emma E.M. Knowles[2,3], Worrawat Engchuan[7], Omar Shanta[8], Bhooma Thiruvahindrapuram[7], Jeffrey R. MacDonald[7], Celia M. T. Greenwood[9,10], Stephen W. Scherer[7,11], Laura Almasy[4,5,6], Jonathan Sebat[8], David C. Glahn[2,3], Guillaume Dumas[1,12]**, Sébastien Jacquemont[1]**.

Affiliations:

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

[2] Harvard Medical School, Department of Psychiatry, 25 Shattuck St, Boston, MA, USA

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

[4] Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, PA, USA

[5] Lifespan Brain Institute, Children's Hospital of Philadelphia, and Penn Medicine, PA, USA

[6] Department of Genetics, University of Pennsylvania, PA, USA

[7] The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada

[8] Department of Psychiatry, University of California San Diego, La Jolla, CA, USA

[9] Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada

[10] Gerald Bronfman Department of Oncology, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada

[11] Department of Molecular Genetics, University of Toronto, Ontario, Canada

[12] Mila – Quebec AI Institute, University of Montréal, Canada

*** authors contributed equally**

**** Shared senior authorship**

Contact information:

Lead contact: Sayeh Kazem: sayeh.kazem@umontreal.ca

Kuldeep Kumar: kuldeep.kumar@umontreal.ca

Guillaume Dumas: guillaume.dumas@umontreal.ca

Lead PI: Sebastien Jacquemont: sebastien.jacquemont@umontreal.ca

ABSTRACT

Pleiotropic effects of gene dosage are central to understanding comorbidities in developmental pediatric and psychiatric disorders, yet the underlying biological processes are unknown. We developed Functional Burden analysis (FunBurd) to investigate the association of all protein-coding copy-number-variants (CNVs), genome-wide, with 43 complex traits in ~500,000 UK-Biobank participants. We tested CNV associations disrupting 172 tissue or cell-type gene-sets, observing associations across all traits. Pleiotropy was correlated with genetic constraint and was higher in the brain compared to non-brain functions, even after normalizing for genetic constraint. Cognition and mental health traits showed specific gene-dosage effects across cortical/sub-cortical and neuronal/glial functional categories. The levels of pleiotropy, measured by burden correlation, were similar in deletions and loss-of-function SNVs, and higher compared to common variants and duplications. Gene sets under high genetic constraint showed less monotonic gene dosage responses across traits. Across most traits, we observed a negative deletion-duplication effect size correlation, indicating that functional gene sets are preferentially sensitive to either deletion or duplication, but rarely both. Our results highlight the key role of genetic constraint and brain-specific mechanisms in shaping CNV-driven pleiotropy, providing a mechanistic basis for the whole-body multimorbidity observed in neurodevelopmental and psychiatric conditions.

MAIN

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[1,4–8].

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 [9–12]. 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 [11,13,14] 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 developmental pediatric clinics for complex neurodevelopmental disorders and congenital malformations[15–17]. 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[10–12,18,19], 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 recurrent CNVs[17,19,20], which collectively affect only approximately 2% of the coding genome[21]. 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[22]) explain a significant proportion of the effects of coding CNVs on cognitive and behavioral traits[23–26]. 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 (both brain and non-brain) across different levels of biological function, encompassing tissue and cell types, and compare it 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)[26], which aggregated CNVs that disrupt genes with similar functions, defined through shared expression patterns across tissues and cell types (**Fig. 1**).

We quantified the effect size of biological functions disrupted by CNVs for each of 43 complex traits—categorized into brain-related (e.g., cognitive metrics, mental health) and non-brain-related (e.g., blood assays, physical measures, reproductive factors)—in ~500,000 UK Biobank participants. We compared these CNV burden associations to those of common variants and loss-of-function single nucleotide variants (SNVs)[18,20,27–30]. Furthermore, we

systematically characterize the differences between deletions and duplications by quantifying gene dosage responses and performing CNV-burden correlations across traits. This is, to our knowledge, the first large-scale investigation of gene dosage across multiple categories of traits and biological functions.

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, physical measures, cognitive, mental health measures, and reproductive traits) and 60 gene sets assigned to 60 whole-body tissues which on average showed an overlap of 7.9% (**Methods, Fig. S1, Table ST5**). We chose these 43 phenotypes (**Table ST1**) relevant to a broad spectrum of medical conditions, because their associations with common variants and rare SNVs have been previously studied [20,27,29]. Out of the 2580 associations (43 traits by 60 gene sets for deletions and duplications separately, **Table ST2),** CNVs showed FDR-significant association in all 5 trait categories (**Figs. 2A, S2-S4**). Deletions showed a higher proportion of FDR significant associations than duplications (n-del=568 [22%], n-dup=352 [13.6%]; proportion test p-value=1.2e-14, **Fig. 2B, Table ST8**).

Because CNVs are strong contributors to neurodevelopmental and psychiatric disorders[15,16,31], we stratified functions into brain and non-brain gene sets. Brain tissue gene sets showed higher pleiotropy—a greater proportion of significant associations—compared to non-brain

gene sets. This was true for brain and non-brain traits, (FDR-significant proportion test, Del p-value<1e-16, Dup p-value=2.2e-10, **Fig. 2C, Table ST9**). These findings remained unchanged after removing correlated traits (**Fig. S11**). Sensitivity analyses (**Fig. S17, Table ST12)** showed that results were not influenced by i) different RNA-seq datasets (GTEx[32] and Human Protein Atlas[33]; ii) phenotype measures methods, iii) age (<60 versus > 60 years), and sex (Female and Male); and iv) ancestries**.**

We observed specific dosage-dependent effects across distinct brain structures and functions (functional gene set categories/annotations, **Figs. 2D, S4**). The effects of deletions on cognitive measures were primarily driven by sensorimotor cortical gene sets, while duplication effects were preferentially linked to association cortical regions. Notably, the sensorimotor and association cortical regions represent the two ends of cortical brain organization hierarchy, previously characterized at the molecular, cellular, and functional levels[34–36], suggesting an interplay between gene dosage and cortical organization hierachy[35,37]. At the subcortical level, duplications and deletions showed effects on brain traits through the input and output (respectively) of the basal ganglia system[38], a group of nuclei that modulate cortical activity to influence motor control, as well as decision-making and executive functions. Finally, dosage specificity was also evident in regions implicated in neuromodulation (neurotransmitter release, such as locus coeruleus (norepinephrine)[39], substantia nigra (dopamine)[40], and brain stem structures of pons and medulla oblongata (acetylcholine)[41]) that were predominantly sensitive to deletions.

Sex differences impact the genetic architecture of the complex traits[42], including mental

health[43]. To assess this, we ran sex-stratified CNV burden associations across functional gene sets and trait categories (**Figs. S5-S10**). Although overall male and female CNV burden association patterns were similar to the main analysis, for mental health measures, a higher proportion of deletion compared to duplication associations were observed for females only (proportion test, FDR corrected p-value=2.4e-8, for females, **Fig. S7, S10)**. These female-specific gene dosage sensitive associations for mental health were notable for sensorimotor, association, and basal-ganglia gene-sets.

Finally, we asked if these observations were similar for common variants, by performing S-LDSC[5,28] using the same traits in the same dataset with the same gene sets **(Method, Table ST3, Fig. S18)**. While the proportion of significant gene sets was lower (**Fig. 2B**), common variants showed higher levels of association/pleiotropy for the brain compared to non-brain gene sets (**Fig. 2C**), and this remained true when accounting for correlated traits **(Fig. S11B)**.

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. To do so, we defined 81 gene sets assigned to 81 whole-body single-cell clusters (from Human Protein Atlas, HPA)[44], with an average overlap of 4.6% (**Fig. S19, Table ST5**). Functional burden analyses showed that 13% of the 6,966 trait / cell-type associations were FDR significant for deletions and duplications (**Figs. 3A-3B, S20-S21,** and **Table ST2**). Significant associations were observed across HPA cell type groupings and all 5 trait categories (**Fig. S22**). 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-3, **Fig. 3B**).

Similar to tissue gene sets, the brain, compared to non-brain cell type gene sets, exhibited a higher proportion of significant associations (Deletion, p-value=5.5e-03, **Fig. 3C, Table ST8**). Beyond pleiotropy, we also observe some level of specificity with brain and non-brain cell type gene sets showing a higher proportion of associations with brain (Deletion, FDR corrected p-value=1.8e-2) and non-brain (Deletion, FDR corrected p-value=2.2e-7) traits, respectively (**Fig. S12B, Table ST13**). Extending the systematic comparison to common variants using the same cell-type gene-sets showed higher levels of association/pleiotropy for the brain compared to non-brain gene sets (**Figs. 3C**, **S29, Table ST9**).

We also observed distinct dosage-dependent effects across cell-type groupings (functional gene set categories/annotations, **Figs. 3D, S22**). Specifically, deletions and duplications effects on cognitive measures were primarily driven by glial and neuronal gene sets, respectively. For mental health measures, while deletions showed higher overall associations compared to duplications for glial gene sets, duplications exhibited an overall higher proportion of significant associations than deletions across all cell type gene sets (**Fig. S22**). Finally, when assessing the impact of sex, sex-stratified CNV burden associations across functional gene sets and trait categories showed that overall male and female CNV burden association patterns were similar to the main analysis (**Figs. S23-S28**).

Building on these observations, we focused on human brain cell types, using single-cell RNA-seq data delineating 31 adult brain cell type clusters[45] with an average overlap of 7.5% (**Fig. S30, Table ST5**). Both neuronal and non-neuronal cell types exhibited pleiotropic effects

when disrupted by CNVs (**Figs. S31-S32, Table ST8**). These associations extended to both brain and non-brain traits, with neuronal cell types showing a higher proportion of associations with brain traits than non-neuronal ones, specifically for deletions, suggesting stronger brain-specific relevance of neuronal CNV burden **(Fig. S12C)**. For common variants, neuronal gene sets show higher levels of pleiotropy than non-neuronal ones (**Fig. S33)**.

Dissecting pleiotropy, gene function, and genetic constraint

Previous studies have shown that common and rare variant heritability enrichments are strongest in constrained genes across complex traits [22,27,29,46,47]. We, therefore, asked if the levels of pleiotropy observed above were related to genetic constraint. We observed that functional gene sets with higher proportions of genes under constraint (LOEUF top-decile, **Methods**) showed a higher level of pleiotropy for deletion (r=0.39, FDR corrected p-Jaccard=7.5e-3, **Fig. 4A**), and duplication (r=0.44, FDR corrected p-Jaccard=3e-3, **Fig. 4A**). 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, FDR corrected p-Jaccard=2.4e-2, **Fig. 4A**). This remained true regardless of the constraint metric (**Fig. 4C, Table ST10)**. Overall, brain gene sets showed significantly higher proportions of genes under constraint compared to non-brain gene sets (Wilcox Ranksum test p-value=3.5e-12, **Fig. 4B**).

Using an approach normalizing functional burden association for genetic constraint (**Methods**, normative constraint modeling), we demonstrate that the most functional gene sets (~90%) exhibited trait associations within an expected range of 5th to 95th percentiles (**Fig. 4D, Table**

ST4). However, several gene sets exhibited significantly higher levels of pleiotropy than expected based on the proportion of genes under constraint (Wilcox Ranksum test, FDR corrected p-value < 0.05, **Figs. 4D-4E**). Overall, even after accounting for genetic constraint, the brain compared to non-brain gene sets showed higher levels of pleiotropy (Wilcox Ranksum test, FDR corrected p-value=2.1e-11, for Tissues, **Fig. 4D**).

Quantifying functional pleiotropy by CNV-burden correlations

Studies often quantify pleiotropy and shared biological mechanisms[27,48] through genetic correlation between pairs of traits. This approach has been widely used for common variants and recently extended to Loss of Function single-nucleotide variants (LoF SNVs)[27], but has not yet been investigated for CNVs. We estimated the CNV-burden correlation between pairs of traits, separately for deletions and duplications, and tested their concordance with each other, as well as against previously published single-nucleotide polymorphism (SNP)-based and rare LoF SNV-based genetic correlations[27] (**Fig. 5A, Table ST7).**

Our CNV-burden correlation approach showed widespread between-trait correlations, with distinct patterns of correlations for deletions and duplications (**Fig. 5B)**. Deletions exhibited stronger between-trait correlations across the five trait categories compared to duplications. Specifically, for deletions, the mean absolute burden correlations (between trait categories) exceeded 0.2 for the majority of trait category pairs. In contrast, this strong correlation (mean absolute burden correlation > 0.2) was observed for duplications only between mental health

and reproductive traits **(Fig. 5C)**. Notably, the between-trait correlations for mental health traits showed the highest mean absolute correlations with other trait categories for deletions, suggesting a pleiotropic impact of deletions that extends beyond the brain to encompass whole-body traits (**Figs. 5B-5C**).

Overall, the between-trait burden correlations were concordant across all classes of variants (**Fig. 5**). The sign concordance between types of variants ranged from 65% to 74% across all traits, which shows the consistent directionality of the between trait genetic correlations across types of variants **(Table ST11)**. Deletion burden correlations were higher than those observed for duplications (4-fold, **Figs. 5D, S34-S35**), and SNPs (2.9-fold, **Fig. 5F**), and comparable to those observed for loss-of-function SNVs (0.82-fold, **Fig. 5H**). In contrast, duplication burden correlations were the lowest among all classes of variants (SNPs: 0.66-fold, **Fig. 5G,** LoF SNVs: 0.15-fold, **Fig. 5I, Table ST11**). These results suggest that deletions show high levels of pleiotropy, consistent with previous observations for loss-of-function SNVs[27]. Of note, variance explained estimates for CNVs were small (**Fig. 5E, Table ST6)** and could impact the accuracy of CNV burden correlation estimates.

Gene dosage responses and deletion-duplication correlations across traits

Deletions and duplications represent a unique paradigm for comparing two classes of variants with opposing effects on gene expression[9]. While monotonic gene dosage responses (i.e., deletions and duplications have opposing effects on a trait, **Fig. 6A**) have been observed for some traits at specific genomic loci[49], it has been challenging to characterize those at the

phenome-wide and genome-wide level [50,51]. We used functional burden to characterize 7396 gene dosage responses (172 functional gene sets by 43 traits). We considered 1916 responses that showed an FDR significant association for either deletion, duplication, or both. The proportion of monotonic (i.e., opposing effect sizes for deletions and duplications, **Table ST6**) and non-monotonic responses was 41.7% and 58.3%, respectively.

We show that gene sets under high genetic constraint show less monotonic gene dosage responses (r=-0.39, p-Jaccard=1e-3, **Fig. 6H**). Brain gene sets showed less monotonic gene dosage responses compared to non-brain gene sets (Wilcox Ranksum test p-value=1.1e-2, **Fig. 6I**), and reciprocally, brain traits showed less monotonic gene dosage responses compared to non-brain traits (Wilcox Ranksum test p-value=3.5e-2, **Figs. 6B-6C**).

However, only 4% and 11% of responses were truly monotonic or non-monotonic (i.e., both deletions and duplications showing FDR significant associations). Most gene dosage responses (85.5%) were undetermined, with significant associations between functional gene sets and traits observed for either deletions (60%) or duplications (40%) and rarely both (**Figs. 6D-6E, Table ST6**). Most traits (84%) showed negative deletion-duplication effect size correlations, suggesting that associations between functional gene sets and traits are, in most cases either sensitive to deletions or duplications (**Figs. 6F-6G, Table ST6**).

We reasoned that the differential sensitivity of functional gene sets to either deletions or duplications may extend to other classes of variants. We therefore quantified the overlap between gene sets with CNV associations and common variant enrichments across the same 43 traits. Six brain traits showed significant functional gene sets overlap between common

variants (S-LDSC enrichment analysis[5,28]) and duplications (permutation-based, **Figs. S36-S37)**. We did not observe any significant overlap between deletions and common variants **(Figs. S36-S37**), despite the fact that deletions show a higher number of significant associations with traits compared to duplications (**Figs. 2B-3D,** FDR significant proportion test p-value<4.8e-3).

DISCUSSION

This study represents the first comprehensive investigation into the biological determinants of pleiotropic effects of CNVs and monotonic gene dosage responses. Genome-wide CNV functional burden tests revealed that pleiotropy was correlated to levels of genetic constraint. We also demonstrate higher levels of pleiotropy for the brain, compared to non-brain tissue and cell type gene sets, even after normalizing for genetic constraint. We observed distinct dosage-dependent effects, with deletions preferentially impacting sensorimotor and neuromodulation-related gene sets, while duplications were linked to association cortical regions . At the cellular level, these specific effects were observed across glial and neuronal functional categories. Levels of pleiotropy and shared genetic contributions between pairs of traits were highest for deletions and loss-of-function SNVs compared to duplications and common variants. A minority of gene dosage responses were monotonic, and this proportion was negatively correlated with genetic constraint. While brain traits showed less monotonic gene dosage responses compared to non-brain traits, a negative deletion-duplication effect size correlation was observed across the majority of traits. Overall, brain-related gene sets were under higher genetic constraint, showed higher levels of pleiotropy, and had less

monotonic gene dosage responses.

The observation of distinct dosage-dependent effects across brain regions and cell types provides key insights into the functional architecture of CNV pleiotropy. The preferential impact of deletions on sensorimotor cortical gene sets and neuromodulation regions (e.g., locus coeruleus, substantia nigra) suggests that loss of function is particularly detrimental to fundamental motor, sensory, and regulatory systems. Conversely, duplications were primarily linked to association cortical regions, which are central to complex cognitive and executive functions. At the cellular level, these differences persist, as both glial and neuronal functional categories showed distinct contributions to cognitive and mental health traits depending on the variant type. These fine-grained functional distinctions demonstrate that CNV effects are not merely a broad dosage perturbation but are highly dependent on the affected functional module, potentially explaining the diverse clinical outcomes observed in patients.

Funburd aggregating CNVs across biological functions demonstrates that rare CNVs genome-wide are associated with a broad spectrum of mental health, cognitive, physical, reproductive, and medical traits (such as blood assays, physical measures, and reproductive factors). Pleiotropic effects were higher in the brain compared to non-brain-related functional gene sets, even with respect to non-brain traits. This provides molecular-level support for the whole body multimorbid presentations in individuals with psychiatric and neurodevelopmental disorders, underscoring that psychiatric conditions share a substantial imprint of poor body health affecting unrelated bodily functions[52–54]. This has important implications for clinical interpretation and management. CNVs are commonly screened in

individuals, often with a primary focus on neurodevelopmental disorders[55], particularly within pediatric populations. Our results suggest that clinical management should be potentially reassessed when individuals transition from pediatric to adult care. The finding of a higher proportion of deletion associations for mental health traits in females suggests that sex-specific models of CNV risk propagation may be necessary for precise clinical stratification.

Disentangling genetic constraint from biological function has been challenging, and many studies have highlighted the predominant role of genetic constraint in shaping the genetic architecture of complex traits[22,27,28,46]. Using a novel approach to normalize for genetic constraint, we demonstrate that brain-related gene sets showed higher levels of pleiotropy compared to non-brain gene sets beyond what would be expected for gene sets with similar levels of genetic constraint.

Between-trait CNV burden correlations were the highest for deletions compared to duplications and common variants. This is in line with a previous study showing higher levels of genetic correlation for LoF SNVs compared to common variants[27]. The high pleiotropy observed for deletions, comparable to LoF SNVs, likely reflects the fact that both variant classes disrupt constrained, coding genes. This constraint sensitivity may explain why genetic correlations were higher for gene-disrupting variants compared to common variants affecting non-coding regions, as well as a much broader range of genes under less genetic constraint[27,56].

For deletions and duplications with opposing effects on transcription [9], it has been difficult to

observe monotonic effects on complex traits[50]. Our FunBurd approach allowed us to quantify opposing effects of CNVs, showing that non-brain traits have more monotonic gene dosage responses compared to brain traits. This may reflect the fact that the average population value is often the optimal one for non-brain traits, while it is rarely the optimal score for complex brain traits (e.g., adaptive functioning, cognition, educational attainment, mental health, etc.) [50].

We also demonstrate that gene sets under higher genetic constraint show lower proportions of monotonic gene dosage response. To our knowledge, this has never been reported and may reflect the fact that such genes are intolerant to deletions and duplications (any deviation from normal gene dosage), leading to worse outcomes and non-monotonic responses.

Finally, we showed limited functional overlap between deletions and duplications, with the majority of functional gene sets being sensitive to either deletions or duplications, but rarely both, for a given trait. This suggests that different classes of genetic variation may exert their effects on traits through different biological processes[1,6,8,27] This led to a negative deletion-duplication burden correlation for the majority of traits, which is more extreme than findings from previous studies showing discordant burden correlations between different classes of rare variant (e.g., missense and LoF variants) for a given trait, also implying that different classes of variants in the same genes often have divergent phenotypic effects[27].

While the functional burden analysis adjusts for multigenic CNVs by statistically controlling for the count of disrupted genes outside the gene set of interest, we acknowledge that the full complexity of CNV architecture—including total CNV length, CNV count per individual, and

underlying genomic context (e.g., segmental duplications)—presents a continuous challenge for rare variant association studies. To mitigate the risks of residual confounding and inflated associations due to gene set overlap (similar to LD structure), we incorporated several robust methodological safeguards. Specifically, we utilized a new permutation preserving Jaccard distance (P-Jaccard) method[57] and LASSO/ridge regression to account for overlapping gene sets. The UK Biobank, as a relatively healthy population cohort[58], poses limitations for disease analysis due to the under-representation of individuals with severe conditions and deleterious CNVs. This bias reduces our ability to detect associations between traits, conditions, and CNVs. However, using the UK Biobank's highly harmonized genetic dataset, does offer some benefits including standardized genotyping and a consistent CNV calling pipeline, thereby minimizing technical artifacts that often plague heterogeneous CNV cohorts.

Our functional burden association test provides critical insight into the interplay of genetic constraint, gene function, pleiotropy, and gene dosage response across human traits. We showed that the associations between gene functions and phenotypes are strongly modulated by classes of variants. Overall, these findings underscore the necessity of moving beyond locus-specific analyses to embrace functional and variant-specific models, thereby accelerating the deciphering of disease etiology and guiding interventions targeting the mechanistic basis of whole-body multimorbidity.

SUPPLEMENTAL INFORMATION

Supplementary Figs. S1–S39, Supplementary Tables S1-S13.

RESOURCE AVAILABILITY

MATERIALS & CORRESPONDENCE

Requests for further information and resources should be directed to and will be fulfilled by the lead PI, Sebastien Jacquemont (sebastien.jacquemont@umontreal.ca).

DATA 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 Weiner et al.[27]. Gene-level constraint

data are available at https://gnomad.broadinstitute.org and other publications listed in methods.

CODE AVAILABILITY

The code for generating all the Figures reported in the main analysis and supplement material can be found at the following Github link : https://github.com/SayehKazem/FunBurd

ACKNOWLEDGEMENTS

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.

COMPETING INTERESTS

The authors declare no competing interests.

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.

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. Ghatan, S. *et al.* CRISPRi perturbation screens and eQTLs provide complementary and

distinct insights into GWAS target genes. *bioRxiv* 2025.05.05.651929 (2025) doi:10.1101/2025.05.05.651929.

9. Han, L. *et al.* Functional annotation of rare structural variation in the human brain. *Nat. Commun.* **11**, 2990 (2020).
10. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
11. Hujoel, M. L. A. *et al.* Influences of rare copy-number variation on human complex traits. *Cell* **185**, 4233–4248.e27 (2022).
12. 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.
13. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J. Med. Genet.* **56**, 131–138 (2019).
14. 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).
15. 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).
16. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223–1241 (2012).
17. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
18. Auwerx, C., Kutalik, Z. & Reymond, A. The pleiotropic spectrum of proximal 16p11.2 CNVs.

Am. J. Hum. Genet. **0**, (2024).

19. Auwerx, C. *et al.* Rare copy-number variants as modulators of common disease susceptibility. *Genome Med.* **16**, 5 (2024).
20. 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).
21. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
22. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
23. 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).
24. 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.
25. 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).
26. 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.

27. Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* 1–8 (2023).

28. 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.

29. Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nat. Commun.* **12**, 1164 (2021).

30. 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).

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

32. 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).

33. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

34. Sydnor, V. J. *et al.* Neurodevelopment of the association cortices: Patterns, mechanisms, and implications for psychopathology. *Neuron* Preprint at https://doi.org/10.1016/j.neuron.2021.06.016 (2021).

35. 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).

36. Markello, R. D. *et al.* neuromaps: structural and functional interpretation of brain maps. *Nat. Methods* **19**, 1472–1479 (2022).

37. Engchuan, W. *et al.* Psychiatric disorders converge on common pathways but diverge in cellular context, spatial distribution, and directionality of genetic effects. *medRxiv* 2025.07.11.25331381 (2025) doi:10.1101/2025.07.11.25331381.

38. Lanciego, J. L., Luquin, N. & Obeso, J. A. Functional neuroanatomy of the basal ganglia. *Cold Spring Harb. Perspect. Med.* **2**, a009621 (2012).

39. Berridge, C. W. & Waterhouse, B. D. The locus coeruleus-noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain Res. Brain Res. Rev.* **42**, 33–84 (2003).

40. Moore, R. Y. & Bloom, F. E. Central catecholamine neuron systems: anatomy and physiology of the dopamine systems. *Annu. Rev. Neurosci.* **1**, 129–169 (1978).

41. Jones, B. E. From waking to sleeping: neuronal and chemical substrates. *Trends Pharmacol. Sci.* **26**, 578–586 (2005).

42. Bernabeu, E. *et al.* Sex differences in genetic architecture in the UK Biobank. *Nat. Genet.* **53**, 1283–1289 (2021).

43. Silveira, P. P., Pokhvisneva, I., Howard, D. M. & Meaney, M. J. A sex-specific genome-wide association study of depression phenotypes in UK Biobank. *Mol. Psychiatry* **28**, 2469–2479 (2023).

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

47. 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.

48. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

49. Jacquemont, S. *et al.* Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* **478**, 97–102 (2011).

50. 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.

51. Macé, A. *et al.* CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nat. Commun.* **8**, 744 (2017).

52. Tian, Y. E. *et al.* Evaluation of Brain-Body Health in Individuals With Common Neuropsychiatric Disorders. *JAMA Psychiatry* (2023) doi:10.1001/jamapsychiatry.2023.0791.

53. Tian, Y. E. *et al.* Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nat. Med.* **29**, 1221–1231 (2023).

54. Launders, N., Kirsh, L., Osborn, D. P. J. & Hayes, J. F. The temporal relationship between severe mental illness diagnosis and chronic physical comorbidity: a UK primary care cohort study of disease burden over 10 years. *Lancet Psychiatry* **9**, 725–735 (2022).

55. Vorstman, J., Sebat, J., Bourque, V.-R. & Jacquemont, S. Integrative genetic analysis: cornerstone of precision psychiatry. *Mol. Psychiatry* **30**, 229–236 (2024).

56. O'Connor, L. J. *et al.* Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).

57. Burt, J. B., Helmer, M., Shinn, M., Anticevic, A. & Murray, J. D. Generative modeling of brain maps with spatial autocorrelation. *Neuroimage* **220**, 117038 (2020).

58. Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.* **7**, 1216–1227 (2023).

59. 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).

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

61. 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.

62. 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).

63. 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).

64. 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.

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

66. Millard, L. A. C., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).

67. 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).

68. Kumar, K. *et al.* 295. Rare variant genetic architecture of the human cortical MRI phenotypes in general population. *Biol. Psychiatry* **95**, S220–S221 (2024).

69. Kumar, K. *et al.* P397. Genomic deletions and duplications show mirror effects on cognitive ability according to spatial patterns of gene expression in the human brain. *Biol. Psychiatry* **91**, S247–S248 (2022).

70. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).

71. 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).

72. Clifford, P., Richardson, S. & Hémon, D. Assessing the significance of the correlation between two spatial processes. *Biometrics* **45**, 123–134 (1989).

73. Collins, R. L. *et al.* A cross-disorder dosage sensitivity map of the human genome. *Cell* **185**, 3041–3055.e25 (2022).

74. 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).

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[59]. 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 was 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)[60]. Quality control and processing followed our previously published pipeline[24,26,61]. 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[62] and QuantiSNP[63], combined with CNVision[31], and concatenated with CNVs INHERITANCE ANALYSIS (CIAv.2.0), following our published pipeline (https://martineaujeanlouis.github.io/MIND-GENESPARALLELCNV/#/)[24,26,61,64]. CNVs were called using both algorithms with the following parameters: number of 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[61]. We kept CNVs passing inclusion criteria such as confidence score ≥ 30 (with at least one of the 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[24,26,61].

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)[65] – 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 [32] 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)[33], 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 [28,30] to create tissue and cell-type gene sets. 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, **Table ST1**). 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 [66], 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[67].

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 for 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 scaled trait(IRNT) 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[26,35,68,69].

Multiple comparison correction

All analyses were corrected for multiple testing using the False Discovery Rate (FDR)[70]. 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 a plausible null distribution [57,71,72] (the corresponding Jupyter notebook is available on GitHub: https://github.com/SayehKazem/FunBurd).

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)[5]. S-LDSC assesses whether certain biological functions are enriched for common variant GWAS signals [28].

Functional overlap between CNVs and SNPs was defined as the proportion of gene sets enriched jointly in CNVs (FunBurd) and common variant 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- burden correlations

To compute CNV-burden correlations, we used the geneset-trait association profiles of deletions and duplications. Prior to this, we addressed two key issues: the overlap between gene sets and the heritability (variance explained) attributable to functional gene sets. To address gene set redundancy, we integrated two complementary analytical approaches: pairwise Jaccard distances (**Figs. S1-S19-S30**) to quantify overlap among gene sets, and LASSO (Least Absolute Shrinkage and Selection Operator) regression to perform variable selection and regularization (**Fig. S38**). The LASSO model ran on scaled (IRNT) residuals after regressing out the covariates (Sex, Age, and ancestry (PCs) from the scaled trait:

$$Trait = \alpha_0 + \sum_{i=1}^{10} \alpha_i \cdot PC_i + \alpha_{11} \cdot Age + \alpha_{12} \cdot Sex + \varepsilon$$

$$r = Trait - \widehat{Trait}$$

$$r = \beta_0 + \sum_{j=1}^{172} \beta_j \cdot G_j + \varepsilon'$$

By considering these two measures, we excluded the gene sets that had a Jaccard similarity percentile greater than 0.2, that are also removed by LASSO in at least 80% of traits, from further analysis (n = 13; see **Fig. S38**). We use selected gene sets (159 out of 172) to calculate the variance explained by these gene sets for each trait (the model is applied on residuals after regressing out the covariance):

$$r = \gamma_0 + \sum_{k=1}^{159} \gamma_k \cdot G_k + \varepsilon''$$

We then extracted the coefficient of determination (R^2) from each regression model as a proxy for trait heritability (h^2 or variance explained).

Finally, we computed deletion and duplication burden correlations between trait pairs using a genetic correlation formula:

$$r_g(T_1, T_2) = \frac{Cov(T_1, T_2)}{\sqrt{h^2_{T_1} \cdot h^2_{T_2}}}$$

The numerator was defined as the covariance between shrinkage adjusted effect size profiles (excluding the 13 filtered gene sets), while the denominator represented trait specific heritability, estimated as the variance explained by effective gene sets using a linear regression model.

For binary traits, we first applied logistic regression to regress out covariates. The resulting

residuals were then treated as continuous outcomes in downstream analyses to compute variance explained (the denominator). Moreover, to enable meaningful comparison of effect sizes and their covariances across binary–binary, binary–continuous, and continuous–continuous trait pairs (the numerator), we transformed the binary trait effect sizes to the liability scale. This transformation ensures that all effect sizes are on a comparable scale:

$$\beta_{liability} = \beta_{logistic} \cdot \left(\frac{\phi(t)}{K(1-K)} \right)$$

$\beta_{logistic}$ is the logistic regression coefficient, K is the population prevalence of the binary trait, t is the threshold on the standard normal distribution corresponding to the prevalence, and $\phi(t)$ is the standard normal density evaluated at t.

LoF SNV summary statistics of burden correlations

SNV loss of function (LoF) summary stats and burden correlations were downloaded from https://pmc.ncbi.nlm.nih.gov/articles/instance/10614218/bin/NIHMS1933335-supplement-Supplementary_tables.xlsx[27]. Notably, these results were derived from the same set of UK Biobank participants [27].

Genetic constraint metrics

We used several measures of constraint, including: i) the loss-of-function observed/expected upper bound fraction (LOEUF) score[22]; ii) probability of being loss-of-function intolerant (pLI)[22]; iii) missense Z-score[22]; iv) CDS: percentage of coding sequence (CDS) under evolutionary constraint[46]; v) two measures of dosage sensitivity [73]: probability of haploinsufficiency (pHaplo), and probability of triplosensitivity (pTriplo); vi) s-het [74], a constraint metric predicted from multiple transcriptomic and gene features using machine learning; and vii) gene-length [22,73] in base pairs. We calculated a mean statistic of these measures of constraint for each of our 172 gene sets. To do this, we used genes that overlapped with the top decile (10%) of genes for each constraint measure[22].

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 (**Fig. S39**). 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.

FIGURES

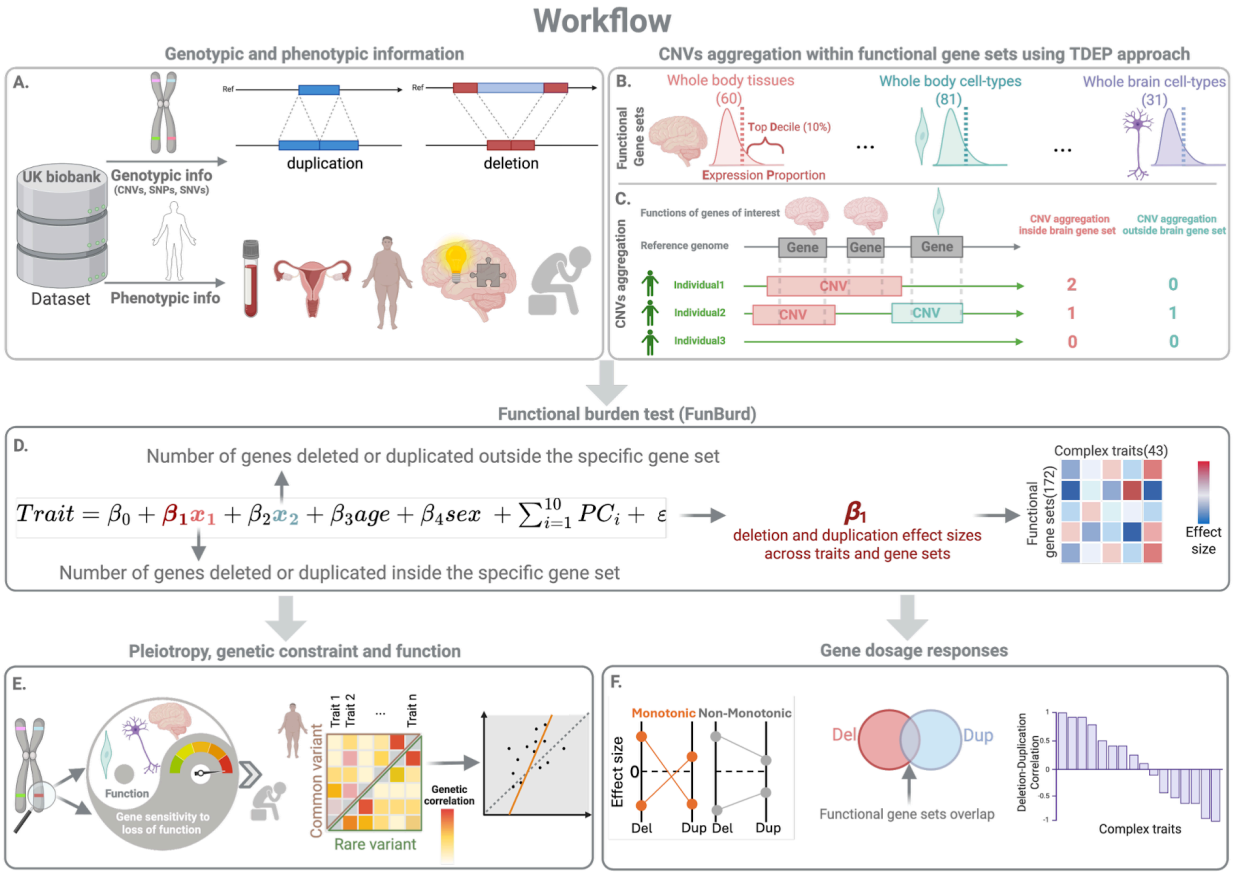

Fig. 1: Estimating CNV effects on complex traits and downstream comparative analysis.

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 (IRNT scaled). Tests were performed separately for deletions and duplications. Downstream analyses systematically investigated **(E)** the relationship between gene constraint and function, and rare and common variant

architectures of complex traits, and **(F)** monotonicity, correlation, and functional overlap in deletion and duplication effect size profiles. *Figure created with BioRender.*

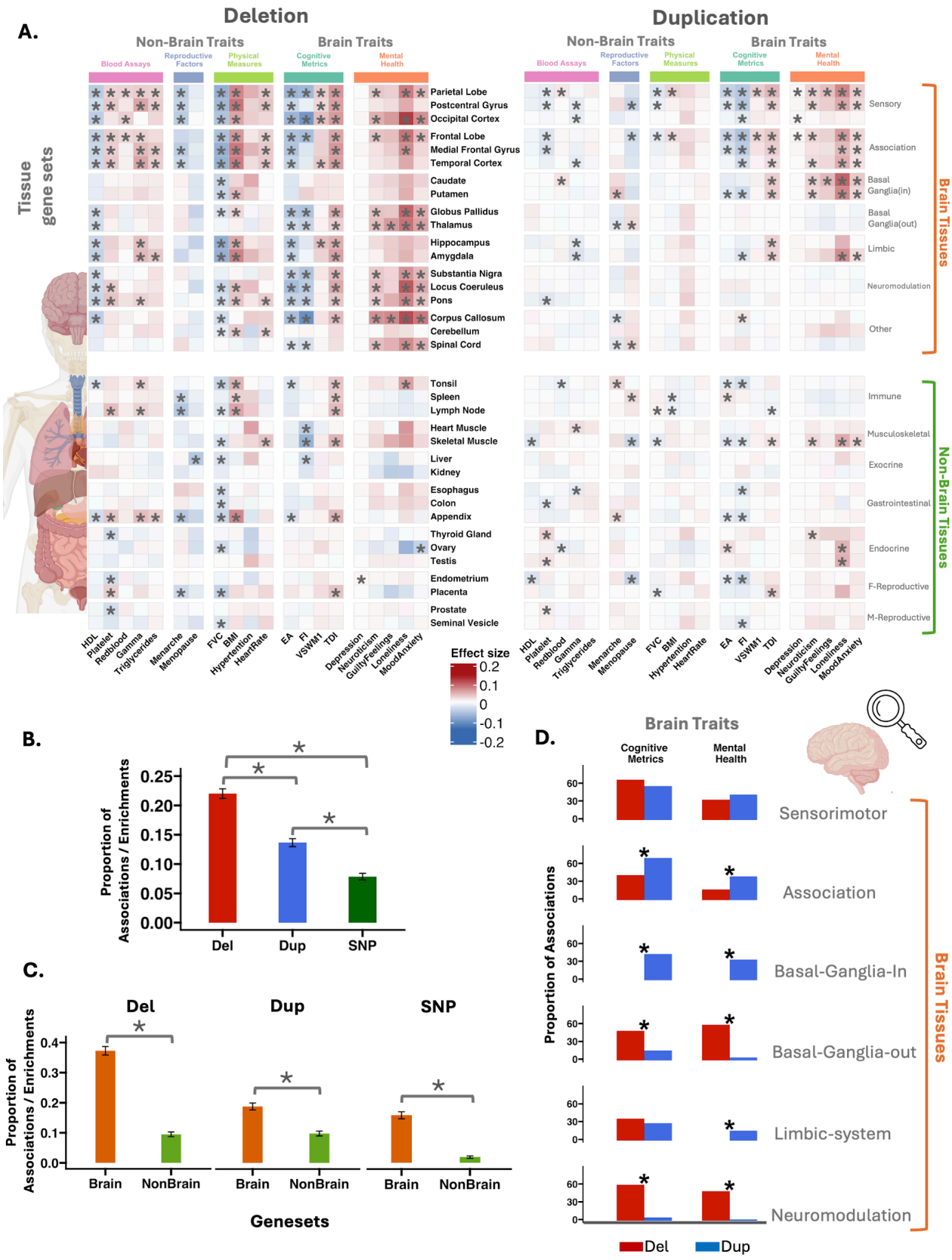
A.
Deletion
Duplication
Non-Brain Traits
Brain Traits
Blood Assays
Reproductive Factors
Physical Measures
Cognitive Metrics
Mental Health
Tissue gene sets
Parietal Lobe
Postcentral Gyrus
Occipital Cortex
Frontal Lobe
Medial Frontal Gyrus
Temporal Cortex
Caudate
Putamen
Globus Pallidus
Thalamus
Hippocampus
Amygdala
Substantia Nigra
Locus Coeruleus
Pons
Corpus Callosum
Cerebellum
Spinal Cord
Tonsil
Spleen
Lymph Node
Heart Muscle
Skeletal Muscle
Liver
Kidney
Esophagus
Colon
Appendix
Thyroid Gland
Ovary
Testis
Endometrium
Placenta
Prostate
Seminal Vesicle
Sensory
Association
Basal Ganglia(in)
Basal Ganglia(out)
Limbic
Neuromodulation
Other
Brain Tissues
Immune
Musculoskeletal
Exocrine
Gastrointestinal
Endocrine
F-Reproductive
M-Reproductive
Non-Brain Tissues
HDL
Platelet
Redblood
Gamma
Triglycerides
Menarche
Menopause
FVC
BMI
Hypertention
HeartRate
EA
FI
VSWM1
TDI
Depression
Neuroticism
GuiltyFeelings
Loneliness
MoodAnxiety
Effect size
0.2
0.1
0
-0.1
-0.2
B.
Proportion of Associations / Enrichments
0.25
0.20
0.15
0.10
0.05
0.00
Del
Dup
SNP
C.
Del
Dup
SNP
Proportion of Associations / Enrichments
0.4
0.3
0.2
0.1
0.0
Brain
NonBrain
Genesets
D.
Brain Traits
Cognitive Metrics
Mental Health
Proportion of Associations
60
30
0
Sensorimotor
Association
Basal-Ganglia-In
Basal-Ganglia-out
Limbic-system
Neuromodulation
Brain Tissues
Del
Dup

Fig. 2: Tissue-specific associations of CNVs with complex traits.

A) The heatmap displays a representative set of association effect sizes between five categories of traits (x-axis) and tissue-specific gene sets (y-axis), separated by variant type (deletions and duplications). The five trait categories are shown along the x-axis (top), and gene sets are listed on the y-axis, with their categories/annotations indicated on the right side. The color intensity reflects the direction and magnitude of the association (blue = negative effect size; red = positive effect size). Black asterisks (*) indicate statistically significant associations between traits and genes (FDR correction across all gene sets and traits, 172 × 43 × 2 =14,792 tests). **B-D)** Summary of CNV association differences. Bar plots (with standard error) summarizing the differences in the level of association/enrichments between the type of variants for **(B)** Whole-Body Comparison: Differences in association/enrichments level across type of variants for all tissue gene sets; **(C)** Tissue-Specific Comparison: Differences in association level comparing brain and non-brain functional gene sets. **D)** Trait-Specific Brain Category Comparison: Differences for cognitive and mental health traits specifically within Brain gene set categories/annotations. Black asterisks (*) indicate statistically significant proportion differences (FDR-adjusted).

Abbreviations: HDL: high-density lipoprotein; HbA1c: glycated haemoglobin; BMD: bone mineral density; BMI: body mass index; EA: educational attainment; FI: fluid intelligence; TDI: townsend deprivation index; VSWM: visuospatial working memory; FVC: Forced vital capacity; Del: deletion; Dup: duplication; SNP: single-nucleotide polymorphism.

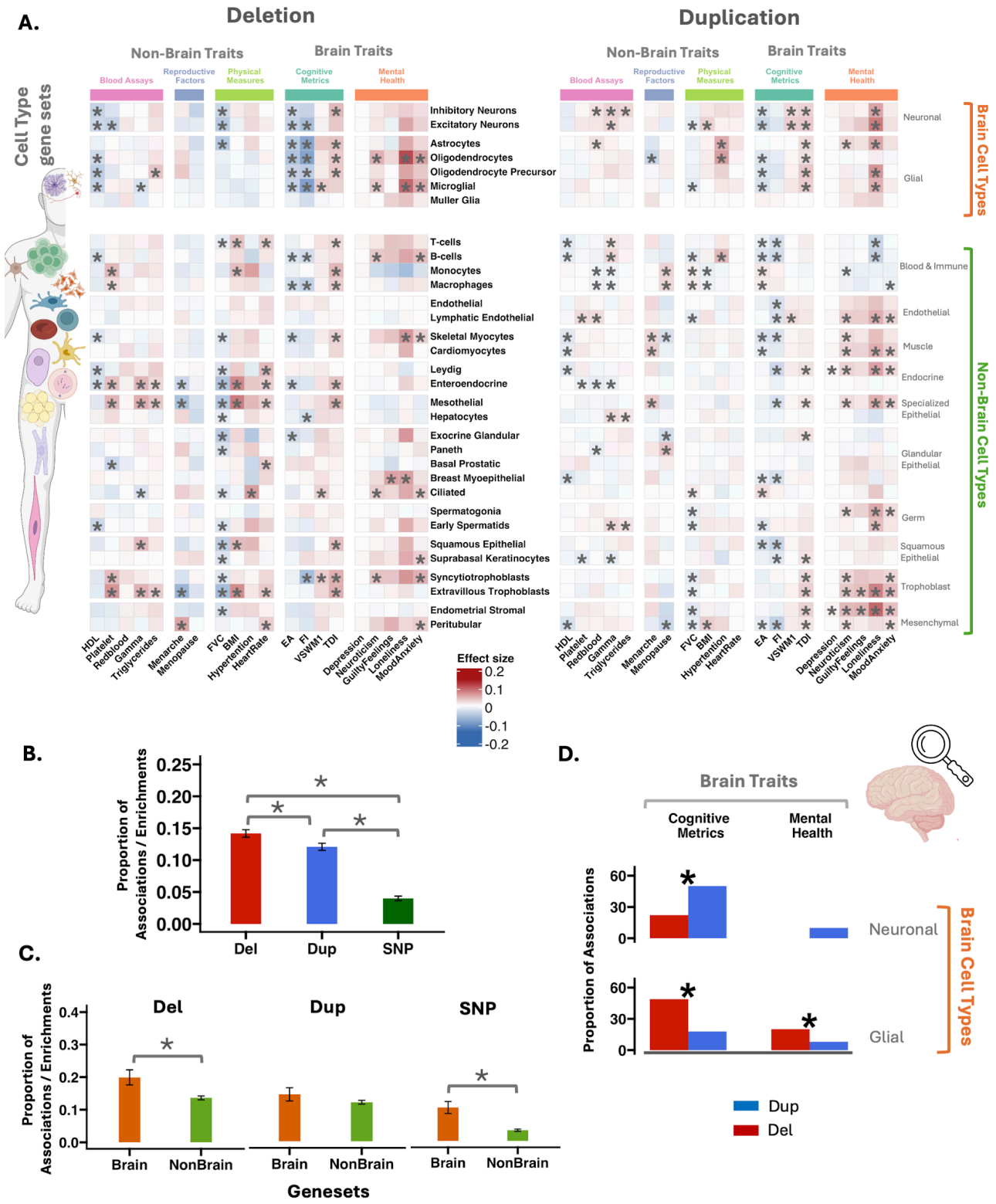
A.
Deletion
Duplication
Non-Brain Traits
Brain Traits
Blood Assays
Reproductive Factors
Physical Measures
Cognitive Metrics
Mental Health
Cell Type gene sets
Inhibitory Neurons
Excitatory Neurons
Astrocytes
Oligodendrocytes
Oligodendrocyte Precursor
Microglial
Muller Glia
T-cells
B-cells
Monocytes
Macrophages
Endothelial
Lymphatic Endothelial
Skeletal Myocytes
Cardiomyocytes
Leydig
Enteroendocrine
Mesothelial
Hepatocytes
Exocrine Glandular
Paneth
Basal Prostatic
Breast Myoepithelial
Ciliated
Spermatogonia
Early Spermatids
Squamous Epithelial
Suprabasal Keratinocytes
Syncytiotrophoblasts
Extravillous Trophoblasts
Endometrial Stromal
Peritubular
Neuronal
Glial
Blood & Immune
Endothelial
Muscle
Endocrine
Specialized Epithelial
Glandular Epithelial
Germ
Squamous Epithelial
Trophoblast
Mesenchymal
Brain Cell Types
Non-Brain Cell Types
HDL
Platelet
Redblood
Gamma
Triglycerides
Menarche
Menopause
FVC
BMI
Hypertention
HeartRate
EA
FI
VSWM1
TDI
Depression
Neuroticism
GuiltyFeelings
Loneliness
MoodAnxiety
Effect size
0.2
0.1
0
-0.1
-0.2
B.
Proportion of Associations / Enrichments
0.25
0.20
0.15
0.10
0.05
0.00
Del
Dup
SNP
C.
Proportion of Associations / Enrichments
0.4
0.3
0.2
0.1
0.0
Del
Dup
SNP
Brain
NonBrain
Genesets
D.
Brain Traits
Cognitive Metrics
Mental Health
Proportion of Associations
60
30
0
Neuronal
Glial
Brain Cell Types
Dup
Del

Fig. 3: Whole-body cell type-specific associations of CNVs with complex traits.

A) Heatmap displays a representative set of association effect sizes between five categories of traits (x-axis) and deletions and duplications aggregated by whole-body cell types (y-axis). Traits are categorized and shown along the x-axis, while the y-axis lists gene sets, with their categories/annotations indicated on the right side. 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 $\times$ 43 $\times$ 2 =14,792 tests). **B-D)** Summary of CNV association differences. Bar plots (with standard error) summarizing the differences in the level of association/enrichments between the type of variants for **(B)** comparison across type of variants for all whole-body cell type gene sets; **(C)** comparison between brain and non-brain functional gene sets; and **(D)** differences for cognitive and mental health traits specifically within brain gene set categories of neuronal and glia. Black asterisks (*) indicate statistically significant proportion differences (FDR-adjusted).

Abbreviations: HDL: high-density lipoprotein; HbA1c: glycated haemoglobin; BMD: bone mineral density; BMI: body mass index; EA: educational attainment; FI: fluid intelligence; TDI: townsend deprivation index; VSWM: visuospatial working memory; FVC: Forced vital capacity; Del: deletion; Dup: duplication; SNP: single nucleotide polymorphism.

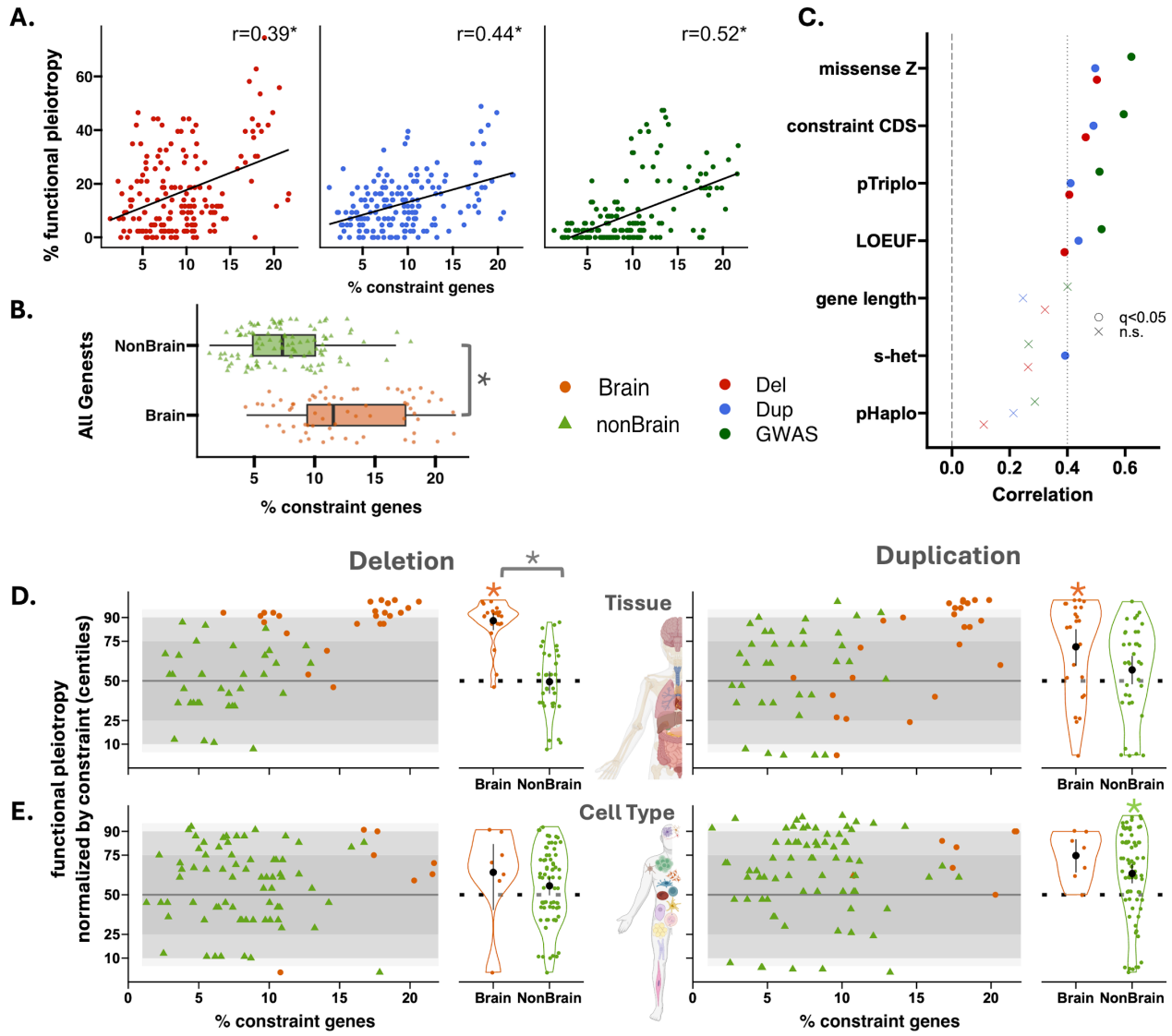

Fig. 4: Dissecting pleiotropy, gene function, and genetic constraint.

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 (functional pleiotropy); X-axis: the percentage of top decile LOEUF genes within the gene set. * indicate FDR-adjusted P-value Jaccard (1000 nulls). **B)** Box plots show the distribution of constraint gene percentages for all Brain and Non-Brain gene sets (172). Each point represents a gene set, and asterisks (*) indicate statistically significant differences

between groups. **C)** Shows the same information on correlations in panel **(A)** for different constraint metrics. q<0.05: FDR-adjusted P-value Jaccard (1000 nulls) are shown with circles; n.s.: not significant. **D)** Functional pleiotropy, normalized by the fraction of genetic constraint at the tissue level, is shown for 27 brain and 33 non-brain gene sets – deletions are presented on the left and duplications on the right. X-axis: represents the proportion of intolerant genes for the different gene sets. Y-axis: functional pleiotropy normalized by genetic constraint (centile). I.e., the 50th centile shows median functional pleiotropy computed across 100 randomly sampled gene sets. Circles and triangles represent brain and non-brain gene sets, respectively. Gray shaded ribbons indicating 25th–75th (ribbon 1), 10th–90th (ribbon 2), and 5th–95th (ribbon 3) centiles. This is followed by violin plots showing the distribution of normalized functional pleiotropy across brain and non-brain traits. Orange and green asterisks demonstrate significantly (FDR-corrected, q < 0.05) increased functional pleiotropy compared to what is expected for a gene set with a comparable fraction of genetic constraint. The grey star indicates a significant difference between the brain and non-brain gene sets, normalized functional pleiotropy. **E)** The same analyses at the whole-body cell type level, including 7 brain and 74 non-brain gene sets.

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 a loss-of-function in any given gene; SNP: single nucleotide polymorphism.

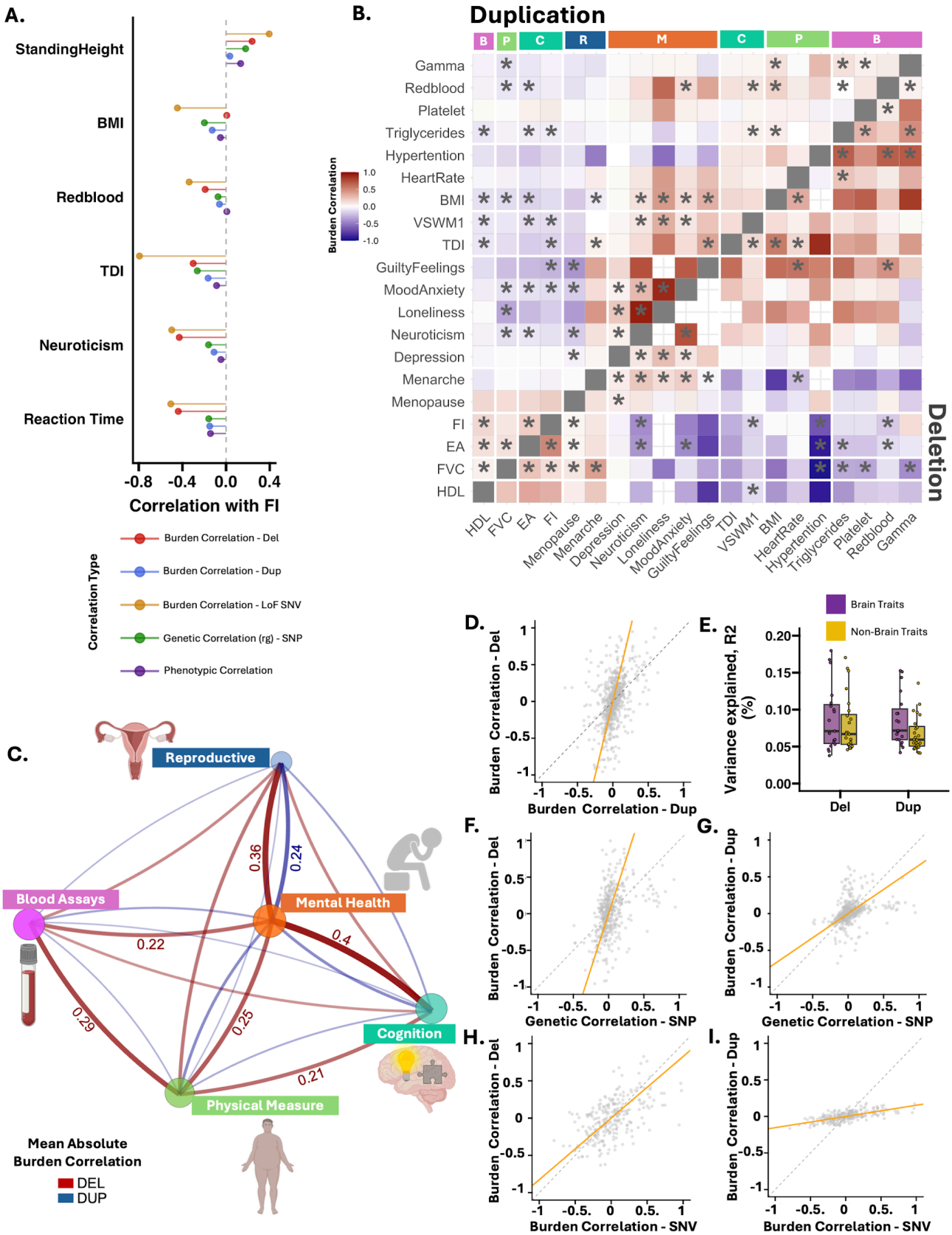
A.
StandingHeight
BMI
Redblood
TDI
Neuroticism
Reaction Time
-0.8 -0.4 0.0 0.4
Correlation with FI
Correlation Type
Burden Correlation - Del
Burden Correlation - Dup
Burden Correlation - LoF SNV
Genetic Correlation (rg) - SNP
Phenotypic Correlation
B.
Duplication
Deletion
Burden Correlation
1.0
0.5
0.0
-0.5
-1.0
B P C R M C P B
Gamma
Redblood
Platelet
Triglycerides
Hypertention
HeartRate
BMI
VSWM1
TDI
GuiltyFeelings
MoodAnxiety
Loneliness
Neuroticism
Depression
Menarche
Menopause
FI
EA
FVC
HDL
HDL FVC EA FI Menopause Menarche Depression Neuroticism Loneliness MoodAnxiety GuiltyFeelings TDI VSWM1 BMI HeartRate Hypertention Triglycerides Platelet Redblood Gamma
C.
Reproductive
Blood Assays
Mental Health
Cognition
Physical Measure
0.36
0.24
0.22
0.4
0.29
0.25
0.21
Mean Absolute
Burden Correlation
DEL
DUP
D.
Burden Correlation - Del
Burden Correlation - Dup
E.
Brain Traits
Non-Brain Traits
Variance explained, R2 (%)
0.20
0.15
0.10
0.05
0.00
Del
Dup
F.
Burden Correlation - Del
Genetic Correlation - SNP
G.
Burden Correlation - Dup
Genetic Correlation - SNP
H.
Burden Correlation - Del
Burden Correlation - SNV
I.
Burden Correlation - Dup
Burden Correlation - SNV

Fig. 5: Comparison of rare and common variant architectures across complex traits.

A) Dot plot illustrating the genetic and phenotypic correlations between fluid intelligence (FI) and six other traits., with each point color-coded by correlation type: SNP-based genetic correlations, deletion, duplication burden correlations, loss-of-function SNV burden correlation, and phenotypic correlations. **B)** The heatmap displaying CNV burden correlations for trait pairs, where the lower triangle matrix shows correlations based on deletions and the upper triangle matrix shows correlations based on duplications. Significant correlations are marked by asterisks (*), based on FDR corrected p-values (p-Jaccard). The color bar on the top (x-axis) represents the five different trait categories: B: Blood Assays; P: Physical Measures; C: Cognitive Metrics; M: Mental Health; R: Reproductive Factors. **C)** A network plot visualizing the the Mean Absolute CNV Burden Correlation between the five trait categories (Blood Assays, Physical Measures, Cognitive Metrics, Mental Health, and Reproductive Factors. Red and blue edges indicate deletion and duplication burden correlations, respectively. Edge thickness represents the strength of the mean absolute CNV burden correlation, with thicker lines showing a higher mean absolute burden correlation. Edge labels indicate the magnitudes of the strongest correlations (only values > 0.2 are shown). **D-I)** Comparison of between-trait correlations across variant types. **D)** A scatter plot comparing the CNV burden correlations between deletions and duplications. The orange line represents the Total Least Squares regression, and the dashed line indicates perfect concordance (y = x). **E)** Boxplots summarizing the variance explained by deletions and duplications aggregated across functional gene sets for brain and non-brain traits, determined via regression analysis (**Methods**). Concordance between CNV burden correlations and **(F-G)** SNP-based between trait genetic correlations and **(H-I)** loss-of-function (LoF) SNV burden correlations (phenotypic correlations are shown in **Fig. S34)**.

Abbreviations: BMI: body mass index; EA: educational attainment; FI: fluid intelligence; TDI: townsend deprivation index; VSWM: visuospatial working memory; FVC: Forced vital capacity; Del: deletion; Dup: duplication; SNP: single nucleotide polymorphism; SNV: single nucleotide variants; LoF: loss of function;

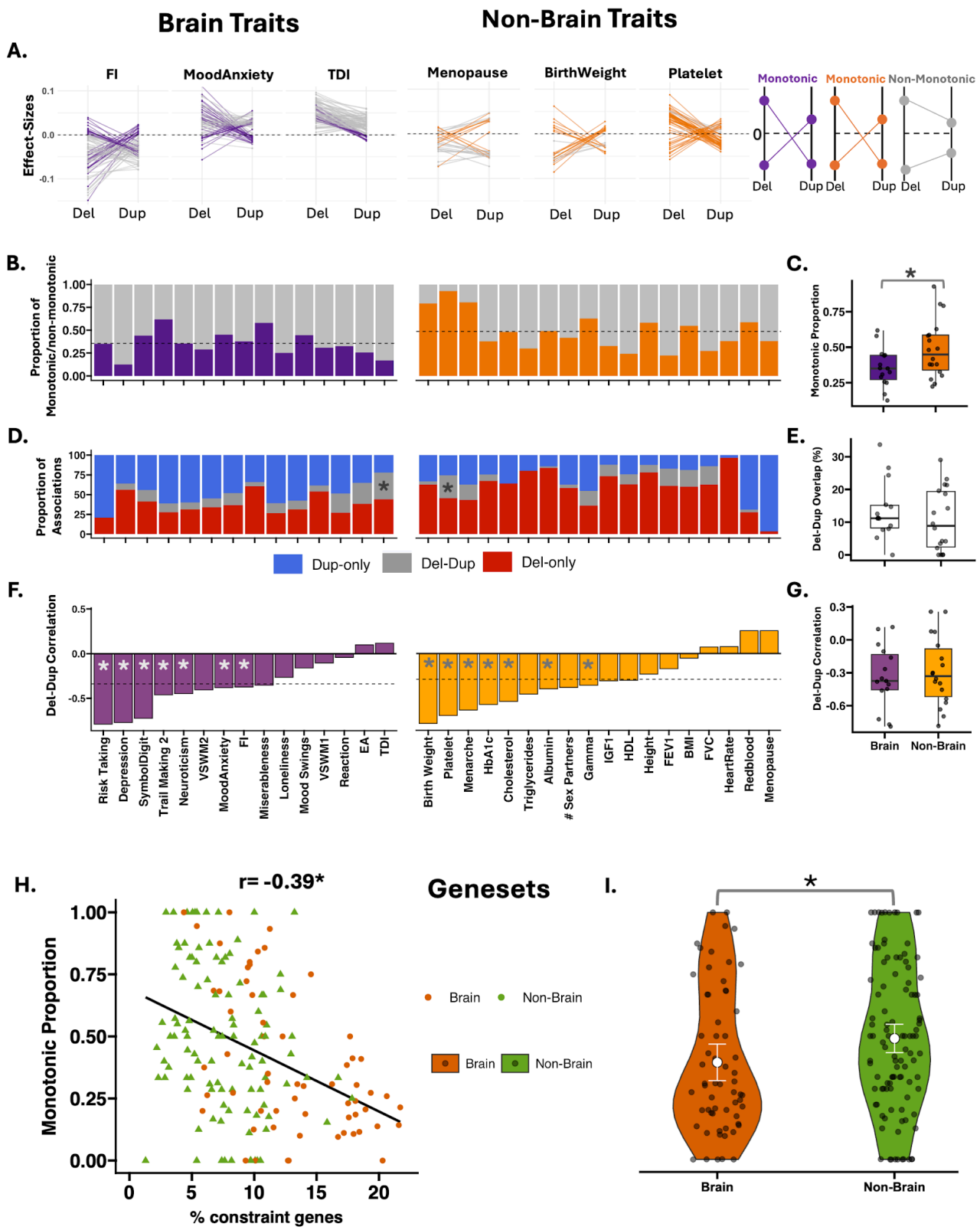
Brain Traits
Non-Brain Traits
A.
FI
MoodAnxiety
TDI
Menopause
BirthWeight
Platelet
Effect-Sizes
Del
Dup
Monotonic
Monotonic
Non-Monotonic
B.
Proportion of Monotonic/non-monotonic
C.
Monotonic Proportion
D.
Proportion of Associations
Dup-only
Del-Dup
Del-only
E.
Del-Dup Overlap (%)
F.
Del-Dup Correlation
Risk Taking
Depression
SymbolDigit
Trail Making 2
Neuroticism
VSWM2
MoodAnxiety
FI
Miserableness
Loneliness
Mood Swings
VSWM1
Reaction
EA
TDI
Birth Weight
Platelet
Menarche
HbA1c
Cholesterol
Triglycerides
Albumin
Sex Partners
Gamma
IGF1
HDL
Height
FEV1
BMI
FVC
HeartRate
Redblood
Menopause
G.
Del-Dup Correlation
Brain
Non-Brain
H.
r= -0.39*
Monotonic Proportion
% constraint genes
Genesets
Brain
Non-Brain
I.
Brain
Non-Brain

Fig. 6: Gene dosage responses across traits and functional gene sets.

A) Illustration of gene dosage responses across 3 brain and non-brain traits. Each line connects the effect sizes of deletion (left) and duplication (right) associations for the same gene set and trait. Dark purple lines represent monotonic responses (defined as significant effect sizes with opposing directionality for deletion and duplication associations) for a given functional gene set and a given brain trait. Dark orange shows the same information for non-brain traits. Grey lines represent a non-monotonic response (same effect size directionality for deletion and duplication associations). Only FDR significant effect sizes (for either deletions, duplications, or both) are represented. Y-axis: effect sizes. **B-E)** Summary of gene dosage responses across traits. **B)** The proportions of monotonic and non-monotonic responses are shown across brain and non-brain traits. The total number of responses includes only FDR-significant effects (either deletions, duplications, or both), and the proportions of monotonic versus non-monotonic responses are calculated within this significant subset. **C**) Box plots depict the distribution of monotonic responses for brain and non-brain traits; * indicates statistically significant (Wilcox Ranksum) differences between groups at the trait level. **D**) Proportion of gene dosage responses significant for deletions-only, duplication-only, and both deletion-duplication, across brain and non-brain traits. **E**) Brain and non-brain traits showed similar levels of deletion-duplication association overlap. **F-G**) Summary of deletion-duplication effect size correlations across traits. **F**) Barplots showing the deletion-duplication effect size correlations across traits. The * indicates significant correlation (FDR corrected p-Jaccard). **G**) Brain and non-brain traits show similar levels of deletion-duplication effect size correlations. **H-I**) Summary of gene dosage responses across functional gene-sets. **H)** Correlation between the genetic constraint (fraction of constraint genes, LOEUF top-decile) within a gene set and the proportion of monotonic gene dosage responses for that gene set (r=-0.39, p-Jaccard=1e-3). For a given gene-set, the proportion of monotonic responses are calculated across all traits, and using the total number of responses that are FDR-significant (for either deletions, duplications, or both). Data points correspond to individual gene sets, with brain and non-brain gene sets shown in orange, and green,

respectively. Y-axis: the proportion of monotonic responses within the gene set; X-axis: the percentage of constraint genes within the gene set. **I)** Box plots showing the distribution of monotonic proportions for brain and non-brain gene sets. Each point represents a gene set, and asterisks (*) indicate statistically significant (Wilcox Ranksum) differences between brain and non-brain gene sets. Y-axis: the proportion of monotonic responses within the gene set.