CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

# ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/158483/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Kopal, Jakub, Kumar, Kuldeep, Saltoun, Karin, Modenato, Claudia, Moreau, Clara A., Martin-Brevet, Sandra, Huguet, Guillaume, Jean-Louis, Martineau, Martin, Charles-Olivier, Saci, Zohra, Younis, Nadine, Tamer, Petra, Douard, Elise, Maillard, Anne M., Rodriguez-Herreros, Borja, Pain, Aurèlie, Richetin, Sonia, Kushan, Leila, Silva, Ana I. , van den Bree, Marianne B. M. , Linden, David E. J. , Owen, Michael J. , Hall, Jeremy , Lippé, Sarah, Draganski, Bogdan, Sønderby, Ida E., Andreassen, Ole A., Glahn, David C., Thompson, Paul M., Bearden, Carrie E., Jacquemont, Sébastien and Bzdok, Danilo 2023. Rare CNVs and phenome-wide profiling highlight brain structural divergence and phenotypical convergence. Nature Human Behaviour 7 , pp. 1001-1007. 10.1038/s41562-023-01541-9

Publishers page: http://dx.doi.org/10.1038/s41562-023-01541-9

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

# Rare CNVs and phenome-wide profiling highlight brain-structural divergence and phenotypical convergence

Jakub Kopal[1,2], Kuldeep Kumar[3], Karin Saltoun[1,2], Claudia Modenato[5], Clara A. Moreau[4], Sandra Martin-Brevet[5], Guillaume Huguet[3], Martineau Jean-Louis[3], Charles-Olivier Martin[3], Zohra Saci[3], Nadine Younis[3], Petra Tamer[3], Elise Douard[3], Anne M. Maillard[6], Borja Rodriguez-Herreros[6], Aurèlie Pain[6], Sonia Richetin[6], Leila Kushan[7], Ana I. Silva[8,9], Marianne B. M. van den Bree[9,10,11], David E. J. Linden[8,9,11], Michael J. Owen[9,10], Jeremy Hall[9,10], Sarah Lippé[3], Bogdan Draganski[5,12], Ida E. Sønderby[13,14,15], Ole A. Andreassen[13,15], David C. Glahn[16], Paul M. Thompson[17], Carrie E. Bearden[7], Sébastien Jacquemont[3], *Danilo Bzdok[1,2,18]

[1]Department of Biomedical Engineering, Faculty of Medicine, McGill University, Montreal, Canada

[2]Mila - Quebec Artificial Intelligence Institute, Montréal, QC, Canada

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

[4]Human Genetics and Cognitive Functions, CNRS UMR 3571: Genes, Synapses and Cognition, Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France

[5]LREN - Department of Clinical Neurosciences, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland

[6]Service des Troubles du Spectre de l'Autisme et apparentés, Centre Hospitalier Universitaire Vaudois and University of Lausanne, Lausanne, Switzerland

[7]Semel Institute for Neuroscience and Human Behavior, Departments of Psychiatry and Biobehavioral Sciences and Psychology, UCLA, Los Angeles, USA

[8]School for Mental Health and Neuroscience, Maastricht University, Maastricht, Netherlands

[9]MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK

[10]Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK

[11]Neuroscience and Mental Health Research Institute, Cardiff University, Cardiff, UK

[12]Neurology Department, Max-Planck-Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[13]NORMENT, Division of Mental Health and Addiction, Oslo University Hospital and University of Oslo, Oslo, Norway

[14]Department of Medical Genetics, Oslo University Hospital, Oslo, Norway

[15]KG Jebsen Centre for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway.

[16]Department of Psychiatry, Boston Children's Hospital and Harvard Medical School, Boston, Massachusetts, USA

[17]Imaging Genetics Center, Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, Marina del Rey, California, USA

[18]TheNeuro - Montreal Neurological Institute (MNI), McConnell Brain Imaging Centre, Faculty of Medicine, McGill University, Montreal, QC, Canada

## Corresponding author

Danilo Bzdok, danilobzdok@gmail.com

## Abstract

Copy number variations (CNVs) are rare genomic deletions and duplications that can affect brain and behavior. Previous reports of CNV pleiotropy imply that they converge on shared mechanisms at some level of pathway cascades, from genes to large-scale neural circuits to the phenome. However, existing studies have primarily examined single CNV loci in small clinical cohorts. It remains unknown how distinct CNVs escalate vulnerability for the same developmental and psychiatric disorders. Here, we quantitatively dissect the associations between brain organization and behavioral differentiation across eight key CNVs. In 534 CNV carriers, we explored CNV-specific brain morphology patterns. CNVs were characteristic of disparate morphological changes involving multiple large-scale networks. We extensively annotated these CNV-associated patterns with ~1000 lifestyle indicators through the UK Biobank resource. The resulting phenotypic profiles largely overlap and have body-wide implications, including the cardiovascular, endocrine, skeletal, and nervous systems. Our population-level investigation established brain structural divergences and phenotypical convergences of CNVs, with direct relevance to major brain disorders.

# Introduction

A chief goal of modern neuroscience is understanding how genetic variation impacts brain organization and inter-individual differences in behavior. Advances in genomic microarray technology streamlined the detection of copy number variations (CNVs) – deletions or duplications of chromosomal segments of >1000 base pairs[1,2]. This class of genetic mutations opens a unique window into the investigation of how neurogenetic determinants shape human behavior, cognition, and development[3,4]. Pathogenic CNVs that reoccur across individuals provide opportunities to study groups of individuals who carry the same deletion or duplication of a well-defined set of genes[5]. Moreover, CNVs have larger effects on phenotype than the low effect-size single-nucleotide polymorphisms often identified by genome-wide association studies[6]. Concretely, CNVs overall have been shown to detrimentally affect cognition and raise the risk for psychiatric conditions[4,7]. Nevertheless, it remains unexplained why many different CNVs escalate vulnerability for the same developmental and psychiatric disorders[4,8,9].

The vast majority of large recurrent CNVs have been linked to more than one clinical diagnosis, including intellectual disability, autism spectrum disorders, and schizophrenia[10–12]. These findings make a case that circumscribed genetic changes are rarely exclusively associated with a single clinical diagnosis[13]. Further, CNVs have demonstrable consequences even in seemingly unaffected middle and old age carriers, who show no overt signs of early-onset neuropsychiatric disorders. Recent evidence points to a broader spectrum of impacts from CNV status, ranging from physical traits to diabetes to hypertension to obesity to renal dysfunction[3,14,15], as well as psychopathology[16]. Understudied body-wide CNV effects may contribute to the links of schizophrenia-associated CNVs with diminished academic qualifications, occupation, or household income[17]. In summary, this class of genetic variants affecting distant parts of the genome can be associated with various behavioral and clinical phenotypes[18,19].

Despite many advances in genomic profiling, investigations into the corresponding brain signatures have only been performed for a few CNVs and mostly focused on a single variant at a time[20,21]. These parallel approaches to catalog CNVs highlighted a wide spectrum of robust effects on brain structure[22,23]. Although distinct rare CNVs are associated with a range of brain alterations, they have been suggested to lead to a degree of similarity in associated behavioral phenotypes[9,24]. However, it remains unknown how similar CNVs are in terms of their effects on the brain and the phenome. Since deleterious CNVs are rare, such as 1 in 3,000 for 22q11.2 deletion[25], previous investigations suffered from small samples of subjects and a lack of phenotypic depth. Therefore, previous studies were chronically underpowered to paint a complete picture of CNVs in medicine. There is a need for a systematic investigation of intermediate brain measures and their phenotypic associations across several CNVs by means of a large well-phenotyped patient pool. The recent advent of population cohorts with rich phenotypic assessment batteries represents an untapped opportunity to conjointly examine a set of CNVs and characterize them at an unprecedented scale.

In the present study, we interrogated the largest existing biomedical data resource, the UK Biobank[26], which allowed a head-to-head comparison of an envelope of CNVs. As a first step, we leveraged tools from machine learning, including linear discriminant analysis (LDA), to isolate CNV-specific brain morphology signatures from a multisite clinical cohort. These individuals carried one of eight recurrent CNVs that are among the most widely studied CNV loci to

date[11,23,27]. Deletions and duplications at the loci 1q21.1, 15q11.2, 16p11.2, and 22q11.2 strike a balance between being frequent and having a significant impact on brain and behavior[23]. Subsequently, the advantageous properties of LDA allowed us to carry over the CNV-specific whole-brain signatures from the clinical cohort to the large-scale UK Biobank cohort. The UK Biobank is ideally suited to tease apart the commonalities in phenotypic indicators across CNV alterations due to the breadth of available phenotypic annotations. We directly linked a rich portfolio of phenotypes to eight CNV brain signatures in ~40,000 UK Biobank participants. Specifically, we performed separate phenome-wide association studies (PheWAS) for the eight CNV-brain-imaging signatures across 977 phenotypes from eleven categories. In this way, we provide a population-level characterization of what unites and divides the eight CNVs by detailing convergences and divergences from genomic variants to brain morphology to phenome. In an attempt to establish cornerstone evidence for the community, such a study can illuminate fundamental links between genetic variation and brain organization, with their consequences to bodily systems.

# Results

## Dissecting different CNV effects on whole-brain morphology

We systematically analyzed volumetric measures derived from brain-imaging scans in the clinical cohort comprising 846 total subjects: 534 carried one of eight recurrent CNVs (deletion and duplications of 1q21.1 distal, 15q11.2 BP1-BP2, 16p11.2 proximal, or 22q11.2 proximal), while 312 controls did not carry a CNV (Table 1). We parsed volume measures from these structural brain scans using a 400-region anatomical definition (Schaefer-Yeo reference atlas; see Online methods). To account for variation outside of our current primary scientific interest, each brain region volume was adjusted for intracranial volume, age, age$^2$, sex, and acquisition site for all downstream analysis steps. A schematic flow of all analysis steps is depicted in Supplementary Figure 1.

As a first step, we compared the effects on brain region volume measures for the eight CNVs. Specifically, after normalizing (z-scoring) brain volumes across groups, that is, across the respective CNV carriers and controls, we examined the extent of volumetric divergence between carriers of each single CNV and controls by computing Cohen's *d* (giving an effect size for the group difference) for each individual brain region (Fig. 1a). In doing so, for each examined CNV, we obtained a brain map of Cohen's *d* effect sizes that summarize magnitudes of CNV-induced structural abnormalities across the brain's gray matter. We noted widespread smaller volumes in the majority of the examined atlas regions for the 1q21.1 deletion, 15q11.2 duplication, 16p11.2 duplication, and 22q11.2 deletion. Conversely, a preferential increase in most regional volumes became apparent for the 1q21.1 duplication, 15q11.2 deletion, 16p11.2 deletion, and 22q11.2 duplication. These findings align with well-known regional alterations identified in cohorts with patients carrying neurodevelopmental disorders[22].

Each target CNV locus was characterized by an overall constellation of gray matter changes – a brain-wide CNV map of how particular CNV carriership results in systematic brain deviations from controls. To delineate the similarity among effect-size brain maps, we computed Pearson's correlation between all 400 regional Cohen's *d* values corresponding to each pair of CNVs (Fig. 1b). Statistical significance was assessed using a spin-permutation test across the whole brain surface. We found a large disparity between Cohen's *d* maps evidenced by the wide spectrum of Pearson's correlations ranging from -0.51 to 0.63. We noted certain similarities, such as for deletions of 22q11.2 and 15q11.2 ($r = 0.66$, $p_{FDR\text{-}adj} = 0.03$). Further, we observed a strong mirroring effect with significant anti-correlations between deletions and duplication of

the same locus. Mirroring effects were strongest for 22q11.2 (r = -0.51, $p_{FDR\text{-}adj}$ = 0.03), followed by 16p11.2 (r = -0.39, $p_{FDR\text{-}adj}$ = 0.03). The average volumetric similarity measured by the average absolute Pearson's correlations was r = 0.23. Taken together, this cursory analysis indicated that spatial distributions of mutation-induced changes in brain morphology differed considerably across CNVs.

## Visualizing CNV differences in low-dimensional signatures

A drawback of the approach based on Cohen's *d* lies in its univariate character, which considers each region separately, ignoring the respective remaining atlas regions. Hence, next, we used dimensionality reduction techniques to obtain holistic summaries of the CNV carriers' morphological profiles. We set out from the possibility that CNVs cause coordinated volume changes distributed across the entire brain. Therefore, we expected an intrinsically brain-spanning pattern could be extracted that faithfully captures the induced morphological differences. Principal component analysis (PCA) is the most commonly used multivariate tool that is demonstrably most effective at representing linear latent factors. PCA can be interpreted as computing a new coordinate system such that the axes are oriented in the directions of the largest variation across the 400 region volume measures. We thus used PCA to project all CNV carriers' regional volumes onto the two dominant directions of coherent whole-cortex variation (Fig. 1c). In the ensuing two-dimensional subject embedding, CNV carriers were scattered randomly without an apparent systematic relationship with each other. In other words, the results suggested that CNVs were not the primary source of the interindividual variation in whole-cortex morphology in our cohort. Hence, a method without access to CNV-carriership status, such as PCA, could not provide a satisfying overall description of what drives structural brain deviations induced by specific CNVs.

Therefore, we turned to linear discriminant analysis (LDA) as a pattern classification algorithm that is naturally capable of recovering a low-dimensional representation explicitly aimed at maximizing the separation between the eight CNVs based on the individuals' brain morphometry measures. We then re-expressed the brain-wide regional volumes as the two primary dimensions of structural variation under the LDA model (Fig. 1d). In particular, the leading dimension of the LDA-derived subject embedding captured the differences between 16p11.2 deletion and duplications. The second most explanatory dimension of the LDA-derived embedding mainly captured the differences between 22q11.2 deletion and duplications. This distribution of a single CNV locus along a single dimension points again at similar structural effects with opposite directions. In summary, LDA formed a new low-dimensional space in which the brain morphology of CNV carriers could be effectively identified, quantified, and, subsequently, examined in further detail.

## Deriving CNV-specific intermediate phenotypes

To supplement the multi-CNV classification model, which explored differences between CNVs (described above), our next analysis step was to extract robust whole-brain signatures specific for each CNV that we could then use to study unseen participants in any number of external cohorts. Therefore, we constructed eight LDA models of order one dedicated to the eight CNVs. Notably, there was a considerable imbalance between the number of controls and CNV carriers (from 2-fold for 15q11.2 duplication to 22-fold for 1q21.1 duplication). Moreover, the number of model parameters to be estimated (at least 400 parameters associated with the 400 atlas regions) was larger than the number of subjects. To remedy the challenges of this data scenario, our analysis pipeline combined bagging and regularization to prevent overfitting the

model hyperparameters (see details in Online methods). We evaluated the model performance indexed by out-of-sample prediction in brain scans unseen by the model using the Matthews correlation coefficient. All CNVs were successfully classified with a consistent above-chance accuracy (Fig. 2a). Chance level accuracy was defined as the performance of an empirical null model obtained by label shuffling. High classification performance provides empirical evidence that these CNVs are characteristic of robust volumetric signatures.

After extracting predictive principles of structural brain deviations by means of LDA, each model included a collection of 400 coefficients associated with the atlas regions (Fig. 2b). These coefficients encapsulated a multivariate prediction rule which maximized the difference between controls and CNV carriers. In other words, each CNV's LDA model encapsulated an intermediate phenotype – a brain-wide volumetric signature that characterizes each CNV. To quantify the similarity between the derived intermediate phenotype representations, we compared them using Pearson's correlation coefficient. Again, we observed certain similarities across the eight CNVs, as well as mirroring effects between reciprocal CNVs (Fig. 2c). However, the wide range and low strength (average similarity $r = 0.2$) of obtained CNV-CNV similarities indicated that LDA models reflected the sizable diverging effects of CNVs on brain morphometry. The identified intermediate phenotypes bore a degree of similarity to the Cohen's *d* brain maps (Fig. 2d). The strong positive Pearson's correlation between the intermediate phenotypes and Cohen's *d* brain maps was significant for all CNVs. In other words, LDA-derived (brain-global) patterns capture certain volumetric effects highlighted by previous (region-local) Cohen's *d* analysis. Along with the high prediction accuracy, a degree of similarity with estimated region-wise Cohen's *d* maps is an important step on the path toward characterizing derived signatures in another dataset.

We further inspected the 400 region coefficients of each LDA model that captured the influence of each CNV on each brain region. By carrying out a one-sample bootstrap hypothesis test independently for each CNV, we assessed which region-specific model coefficients are robustly different from zero and, thus, robustly affected by CNVs. Specifically, during the learning of the coefficients of one of the 8 CNV-specific LDA models, in 100 resampling iterations; we drew a different set of subjects based on drawing subjects with replacement from the control subjects and corresponding CNV carriers. Statistically relevant coefficients were robustly different from zero if their two-sided confidence interval - according to the 2.5/97.5% intervals of the bootstrap-derived distribution - did not include zero. Different CNVs affected (displayed statistically relevant coefficients) different cortical parcels that correspond to the seven large-scale brain networks populating the cortex, as defined by our atlas (Fig. 3a). For example, while 16p11.2 proximal duplication primarily affects 20% of all regions in the limbic network, 22q11.2 deletion affects 20% of regions in the salience ventral attentional network as well as more than 10% of regions in the limbic, dorsal attentional, and default-mode networks. Across all examined CNVs and target brain networks, the 16p11.2 deletion affected the largest number of brain regions, while 15q11.2 duplication affected the lowest number of regions. Higher-order network circuits showed, on average, the relatively highest number of significant coefficients. Concretely, the limbic network had the highest relative number of affected regions, followed by the salience and default-mode networks (Fig. 3b). Together, the wide range of effects on the large-scale networks again highlights the diverging consequences of CNVs on brain morphometry.

To further explore characteristic relationships between the eight CNVs, we probed for a linear relationship of the number of salient LDA coefficients with LDA classifier performance and average brain-wide Cohen's *d*. We found a significant positive Pearson's correlation with classifier performance ($r = 0.74$, $p = 0.04$) (Fig. 3c) and mean absolute effect size ($r = 0.75$, $p = 0.03$). Furthermore, when we included sample size in the testing scheme, we found only a

negative linear association with average Cohen's *d* (r = -0.80, p = 0.02), calling for careful interpretation of effect sizes, owing to the estimation of population mean in small samples. In sum, our collective findings highlighted how LDA models reflect CNV-specific changes in large-scale brain networks to form distinctive intermediate phenotypes.

## Lifting over phenotypes patterns from the clinical cohort

We built eight separate LDA models that encapsulated CNV-specific intermediate phenotypes. By doing so, we could quantify the presence of each intermediate CNV phenotype for each subject. Hence, as an illustrative example, we compared the expression level of 16p11.2 proximal duplication intermediate phenotype between the carriers of that CNV and controls. Based on a two-sample bootstrap hypothesis test for the difference of means with 10,000 bootstrap iterations (Online methods), the ensuing means of the intermediate phenotype expressions differed significantly between CNV carriers in the clinical sample and controls (p-value < $10^{-4}$) (Fig. 4a). As a critical step in our analysis, the CNV-specific volumetric signatures derived from our clinical population using LDA could be used in a phenotypically richer population data repository.

To carry over the intermediate phenotypes from the clinical cohort to the UK Biobank, we quantified the expression of each intermediate CNV phenotype for all 39,085 UK Biobank participants (Table 2). We first extracted brain volume measures from the 400 atlas regions, adjusting for several confound variables (see Online methods). We then calculated the subject-specific expression for all intermediate CNV phenotypes in the UK Biobank. It is important to stress that the intermediate phenotypes were derived in the clinical cohort. However, UK Biobank also contains several carriers of the analyzed mutations. The generalizability of the derived intermediate phenotypes was indicated by the difference between the intermediate phenotype expression level of non-carriers and CNV carriers in the UK Biobank (for 16p11.2 duplication p-value < $10^{-4}$ using an identical test to that above, Fig. 4a). Notably, we obtained similar results for all other seven intermediate phenotypes (Supp. Fig. 2). Carrying over CNV-associated MRI profiles computed in the clinical cohort to the UK Biobank was a critical step that allowed us to identify phenotype correlates of the CNV-associated MRI profiles in a population >500 times larger than our median CNV cohort.

## Charting phenome-wide associations of CNV signatures

The UK Biobank is the largest existing uniform brain-imaging dataset in terms of subject sample size and the breadth of available phenotypic annotations. It provides 977 unique phenotypes spanning eleven different categories (Supp. Fig. 3). We performed an exploratory phenome-wide association study (PheWAS) for the purpose of generating new candidate hypotheses. PheWAS allows investigation of the overall patterns of connections by charting associations between hundreds of non-imaging phenotypes and imaging-derived phenotypes. Specifically, we calculated Pearson's correlation between the derived subject-specific expressions of the eight intermediate CNV phenotypes and each of the 977 phenotypes provided by the UK Biobank resource (Fig. 4b). In our recurring example of the 16p11.2 duplication intermediate phenotype, 55 associations surpassed Bonferroni correction for multiple testing (including comparative body size at age 10, education score, hemoglobin concentration, or physically abused by family as a child), while 145 associations surpassed FDR correction. In other words, individuals with greater similarity to the 16p11.2 duplication MRI profiles showed a stronger association with levels of education or blood assays biomarkers.

To gain additional insight, we summarized the phenotypic association profiles by domain. To this end, we calculated the relative number of association hits for each of the eleven phenotypic domains (using the more stringent Bonferroni correction) as a ratio between the number of significant associations and the number of phenotypes in each category. The highest relative number of associations were in categories detailing physical measures, blood assays, and early life factors categories (Fig. 4c). Among all examined CNVs, the 22q11.2 deletion intermediate phenotype displayed the highest number of phenome-wide hits, with 90 robust associations after Bonferroni's correction for multiple comparisons (Fig. 4c; for further details, see Supp. Fig. 4-11). The collective results showed that CNVs are associated with numerous rich and diverse phenotypes across all eleven categories.

Analogous to comparing volumetric signatures (cf. above), we examined the similarity of phenotypic profiles across CNVs. To this end, we calculated a correlation between the association strengths (Pearson's correlations) from each PheWAS analysis (Fig. 5a). The definitive collection of brain signature-phenotype links reflected the linear association strength between CNV phenotypical profiles across 977 indicators (Fig. 5b). We found a strong resemblance (average similarity r = 0.62) between the eight phenotypical profiles with positive as well as negative correlations (Pearson's correlations from r = -0.84 to 0.82). We subsequently zoomed in on the strong convergence across the phenotypic profiles characterizing each CNV by computing the correlation between CNV-phenotypic associations within each of the eleven considered categories (Fig. 5c). In particular, we found the bone density and sizes along with blood assays categories showed strong associations across CNV intermediate phenotypes, suggesting similar behavior within these categories. Altogether, the strong correspondences among CNV pairs suggest that CNV brain profiles are linked to similar phenotypes across a rich portfolio of ~1000 curated lifestyle indicators.

## Detailing shared and distinct phenotypic associations

To shed light on which particular phenotypes are most strongly associated with CNV-specific brain signatures, we calculated the mean absolute Pearson's correlations across the eight PheWAS analyses. Across all CNVs, diastolic blood pressure, alkaline phosphatase, and red blood cell count showed the strongest associations (Fig. 6a). Moreover, we examined which phenotypes are most consistently associated with CNV brain profiles. We found eight phenotypes associated with six CNV intermediate phenotypes and eleven phenotypes shared by five CNV intermediate phenotypes (Supp. Fig. 3b, c). The most consistently overlapping phenotype hits were from the blood assays category (e.g., mean corpuscular volume, SHBG, IGF-1), along with weight or home population density. In total, these robust and shared phenotypic associations point to the fact that CNV brain profiles are associated with similar systemic phenotypes.

Comparisons of the phenotypical profiles associated with each CNV intermediate phenotype revealed that there remains unexplained residual variance, as suggested by a maximum absolute association strength of r = 0.81. To access this remaining part of the variance, we computed new brain profiles adjusting for the other CNVs. Specifically, for each CNV-specific intermediate phenotype, we singled out the variation explained by the remaining seven. Thus, we obtained a set of eight unique intermediate phenotypes, each with the variation shared with other intermediate phenotypes removed. Subsequently, we used this new set to perform the PheWAS analysis and counted the relative number of associations surpassing the Bonferroni correction in each category. We still observed significant associations across CNVs and categories even after conditioning out on the shared associations. In particular, 22q11.2 deletion

showed a high relative number of associations in the physical measures category (Fig. 6c). As such, next to the substantial phenotypic similarity, CNVs also displayed some unique characteristic phenotypic associations relative to other CNVs.

### Quantifying the path toward converging phenotypical profile

The observed magnitude of similarity between the phenotypic profiles of the CNV intermediate phenotypes reaching Pearson's r = 0.84 demonstrated a strong relationship between phenotypic profiles across the 977 indicators. In general, the phenotypic similarity (absolute Pearson's correlation of PheWAS outcomes) between CNVs exceeded their morphological similarity (absolute Pearson's correlation between Cohen's *d* maps) (Fig. 6d). The dissonance between the two similarity measures was highlighted by Lin's concordance correlation coefficient equal to -0.23, suggesting poor concordance. More specifically, 22 of 28 CNV pairs showed stronger phenotypical similarity compared to volumetric similarity. Thus, CNVs were characteristic of stronger phenotypic signature associations compared to associations among volumetric signatures or intermediate phenotypes (Fig. 6e).

Our collective analyses demonstrated that although each CNV displays largely distinct whole-brain morphometric signatures, they converged on similar phenotypic profiles. In proving this, we transferred the intermediate phenotypes derived in the clinical cohort to the UK Biobank population cohort with 39,085 subjects. Using the subject-specific expression levels of eight intermediate phenotypes from eight rare CNVs allowed us to characterize complex phenotypical profiles of each CNV, providing a detailed portrait of their commonalities and idiosyncrasies.

## Discussion

CNVs offer a unique window of opportunity into the consequences of localized genetic variation on human traits. This is especially the case, given their known genetic architecture and typically high penetrance. In the present study, we built computational bridges between eight key CNVs in a multisite clinical dataset, on the one hand, and their deep phenotypic profiling in 39,085 subjects from the wider population, on the other hand. To this end, we designed an analytic framework that can quantitatively dissect the impact of distinct genetic mutations on brain organization and behavioral differentiation. Bringing over derived CNV-specific intermediate phenotypes to the population cohort revealed that the CNVs are tied to pleiotropic associations beyond physical and cognitive domains. This phenome-wide analysis across ~1000 phenotypes revealed many ramifications for several body systems. Our collective analyses also reveal wide-ranging similarities between the PheWAS profiles of the eight CNVs. Therefore, the phenotypic level appears to be the point of alignment for distinct long-segment genetic variants that we show to cause diverging morphological changes in brain morphology. Such late convergence in phenotypic consequences speaks to profound basic science questions regarding the organization of genetic influences on human brain and behavior.

For a long time, inquiries targeting genetic influences have been limited by the lack of longitudinal and deep multimodal measures of brain and behavior in large subject samples[24]. Studies aimed at elucidating genotype-phenotype links were challenged by several obstacles, including ascertainment bias, limited statistical power, and patchy phenotypic coverage[22]. We are unlikely to have access to large enough clinical datasets soon – a condition *sine qua non* for definitive tests of phenotypic overlaps and differences between genetic variants. As a concrete example, Marek and colleagues (2022) highlighted the need for thousands of participants to

obtain reproducible and reliable brain-wide associations. Therefore, to overcome several of these hurdles, we here put forward solutions that take advantage of intermediate CNV phenotypes, a term coined in research on psychiatric disorders[29]. These refer to biological traits that lie in between an individual's external phenotype and innate genetic blueprint[30,31]. We captured CNV-specific intermediate phenotype representations as "genetics-first" whole-brain signatures derived from our clinical boutique dataset. These signatures recapitulated previous findings on morphology alterations, such as the predominant decrease in regional volumes for deletions of 1q21.2 or 22q11.2, as well as the increase for 16p11.2 deletion[23,32,33]. We also observed reported mirror dose responses, especially strong in 22q11.2 locus[22]. Therefore, the validity of LDA-derived intermediate phenotypes is corroborated by recapitulating key findings from clinical studies.

The eight analyzed CNVs are known to differ in the ensuing effects on brain architecture[11,12]. The magnitude of their effects has previously been associated with the number of affected genes and clinical outcomes. In concordance, we found 16p11.2 deletion to affect the largest number of regions. This CNV contains 29 genes and is associated with an almost 40-fold increase in the odds of autism spectrum disorder[27]. Conversely, we found 15q11.2 duplication, which contains only four genes and is not formally associated with any disease, to affect the fewest number of regions. In addition, we provide a fresh look into the diverging CNV effects on brain morphology by summarizing the effects with respect to seven large-scale Schaefer-Yeo networks. The network effects revealed a degree of similarity to functional connectivity alterations in CNV carriers[21]. We also observed effects in the default mode and limbic network for 22q11.2 deletion, as well as for ventral attention and motor network for 16p11.2 deletion. Together, the structural and functional alterations showed significant overlap with alterations of idiopathic autism spectrum disorder and schizophrenia[34]. The resemblance suggests that the risk conferred by genetic variants, structural alterations, and the associated functional connectivity patterns represent important dimensions that are coupled with diseases.

In the present work, we demonstrate the added value of how intermediate phenotypes can be transferred for direct usage in other cohorts, including large-scale populational datasets. By transferring these brain-wide representations over to the UK Biobank and carrying out PheWAS, we obtained systemic phenotypic associations across eleven rich phenotypic categories that go beyond mere cognitive domains. The reported PheWAS associations of the intermediate CNV phenotypes were concordant with previous studies investigating more circumscribed links between CNV status and indicators of cognitive performance, including fluid intelligence score[17], physical measurements like weight or height[3,15], common medical conditions like hypertension or obesity, and blood biomarkers like indicators of cholesterol fat metabolism pathways[35]. As one of many examples, we demonstrated how intermediate phenotypes tied to 22q11.2 deletion relate to an array of phenotypes in blood assays as well as cardiac and blood vessels categories. It is important to stress that PheWAS only charts associations between imaging and non-imaging measures to generate testable hypotheses without providing causal links[36]. Even though there may not be a causative link between a brain phenotype and cardiac biomarkers, the thus revealed association suggests a hidden causal effect of the CNV on both traits (e.g., brain morphology and artery wall thickening[37]).

Similar to Auwerx and colleagues (2022), six of our eight examined CNVs were associated with body weight, insulin-like growth factor 1, alkaline phosphatase, or mean red blood cell volume. Therefore, these bodily alterations may not be mere secondary effects[38]. Instead, systemic manifestations could be a fundamental aspect of the primary biology of CNVs and brain disorders in general. Critically, they might also lead to a reduced life span, as suggested

by the 63% probability of survival to age 50 in adult carriers of 22q11.2 deletion[14,39]. Similar to 22q11.2 deletion, psychotic disorders have been linked with 15–20 years shorter life expectancy[40]. Most of this premature mortality is predominantly due to elevated cardiovascular risk factors[41,42] – causes that belong to the phenotype category among the most consistent associations in our phenome-wide assays. Detected associations speak in favor of CNVs as a complex disorder with several manifestations outside the brain that have considerable deleterious impacts on various parts of everyday lives.

By combining hand-crafted analytic solutions with recently emerged data resources, our computational assays lay out pleiotropic associations in CNV carriers. These consequences include systemic associations outside the central nervous system. This underappreciated insight is reflected in our results, including strong brain-behavior associations of the CNV profile in the UK Biobank population with blood pressure, cholesterol, and weight. Since CNVs do not show complete penetrance in all cases[44], such associations portray a necessary picture of a broad spectrum of outcomes later in life. Hence, the constellation of results advocates rebalancing the medical care of CNV carriers towards more comprehensive medical monitoring in a broader patient pool[45].

In a similar way, previous clinical research has provided evidence that schizophrenia and related psychotic disorders often affect multiple body systems (e.g., nervous, immune, or endocrine), even from illness onset[46,47]. Pillinger and colleagues (2019) reported robust alterations in immune and cardiometabolic systems of a comparable magnitude to alterations in the central nervous system. Further examples of major brain disorders accompanied by problems outside the brain include gastrointestinal disorders in autism[48], loss of bone density in depression[49], or cardiovascular symptoms in bipolar disorder[50]. Finally, a recent study showed that genetic liabilities for five major psychiatric disorders are associated with long-term outcomes in adult life, including sociodemographic factors and physical health[51]. Our findings thus add pieces of knowledge that illuminate how the nervous system is interlocked with the rest of the body in a way that affects general well-being.

More broadly, understanding pathophysiological disease mechanisms will be propelled by further disentangling the perplexing link between genes, brain, and behavior[52]. There is an active debate on the extent to which distinct gene dosage disorders can lead to different non-overlapping phenotypical profiles[24]. This discourse was sparked from the observations that many SNPs and CNVs increase the risk for schizophrenia or autism[11,53]. Polygenicity and pleiotropy, key features of the genetics underpinning psychiatric disorders[13,54], imply that genetic mutations can converge on shared mechanisms at some level of pathway cascades, from genes to large-scale brain networks to the phenome. Here, we report a low similarity of intermediate phenotypes representing morphological CNV-specific brain signatures, in line with a documented broad diversity of regional morphometry patterns across genomic loci[22,55,56]. Conversely, the ramifications of carrying distinct CNV variants for cognition and behavior have previously been hypothesized to be more similar than those on brain anatomy[9,24]. We here find evidence for substantial convergence of phenotypic measures across CNVs quantified by increased phenotypical similarity. Specifically, we observed a high degree of similarity between the phenotypical profiles (mean similarity r = 0.46 as measured by Pearson's correlation across the CNV's corresponding PheWAS profiles), which largely exceeded the similarity of brain morphometry profiles (mean similarity r = 0.2 as measured by the correlation of volumetric Cohen's *d* maps). Based on the presented strong resemblance of phenotypic profiles of the examined eight CNVs, we speculate that the polygenic architecture of human phenotypic traits

may be related to genotype-phenotype convergence that occurs later than on molecular pathways or macroscopic brain networks.

This study has several limitations. One is that we did not investigate the effects of medication on derived CNV-specific brain signatures since medication information was not available for the whole clinical dataset. Nevertheless, previous studies have reported no significant effects of psychiatric comorbidities (e.g. psychosis, ASD, ADHD, anxiety and mood disorder) and psychotropic medication on neuroimaging patterns[34,57]. We also did not study causal relationships between brain patterns and non-imaging indicators. Making causal inference requires proposing and defending a plausible causal structure by spelling out the assumed (directional) dependencies among the outcome, input variables, and relevant confounding variables[58]. Future studies can start off from hypotheses generated by our catalog of PheWAS links to find causal links between variables, for example, using structural equation modeling. Finally, given our data scenario, we resorted to linear models in combination with bagging and shrinkage to safeguard from overfitting.

In conclusion, we have triangulated i) a purpose-designed analytical strategy, ii) a roadmap for investigating rare brain pathologies employing intermediate phenotypes derived from smaller clinical datasets, and iii) a framework for application in population-scale cohorts. Our results highlight the potential of using intermediate phenotypes as a device to study a wide variety of rare conditions and thus accelerate the pace of neurogenetic innovation. By building bridges between the broad population of the UK Biobank and carefully collected clinical datasets, we derived prediction models for CNV-specific brain phenotype expressions that can be used in other hospitals and healthcare institutions. Deep phenotypic profiling of these models clearly demonstrates that CNVs may have whole-body manifestations. Therefore, our study shows that CNV effects go beyond relevance for childcare and psychiatry by potentially extending to other areas of medical care and treatment, which are blind spotted today. In addition, detected overlapping system-wide phenotype associations across multiple CNVs advance our understanding of genotype-phenotype correspondences. Specifically, the observed phenotypic convergence sheds light on why so many CNVs increase the risk for the same developmental, psychiatric disorders.

# Methods

## Multisite clinical cohort

Signed consents were obtained from all clinical participants or legal representatives prior to the investigation. The current study, which is purely analytical, was approved by the IRB (Project 4165) of the Sainte Justine Hospital. UK Biobank participants gave written, informed consent for the study, which was approved by the Research Ethics Committee. The present analyses were conducted under UK Biobank application number 25163. Further information on the consent procedure can be found online (biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=200).

Our clinical dataset consisted of volumetric measurements derived from magnetic resonance imaging (MRI) brain scans of 860 subjects: 548 CNV carriers and 312 controls not carrying any CNV (Table 1). The here examined CNVs are among the most commonly studied CNVs[59]. Deletions and duplications of 1q21.1, 15q11.2, 16p11.2, and 22q11.2 represent some of the most frequent risk factors for neuropsychiatric disorders identified in pediatric clinics[19,20]. That is why the target CNV loci were also selected by The Enhancing NeuroImaging Genetics through Meta-Analysis copy number variant (ENIGMA-CNV) in a study on their cognitive, psychiatric, and behavioral manifestations[23]. These deletions and duplications strike a balance between occurrence in the population and their effect size. In other words, the selected CNVs are frequent enough so that we can start studying large enough sample sizes that allow for across-CNV comparison in the first place. At the same time, this class of CNVs has been shown to detrimentally affect cognition and raise the risk for psychiatric conditions[23,25,33]. Our CNV carriers did not carry any other large CNV.

An extensive description of methods and analyses is available in an already published study with an identical dataset[60]. In short, PennCNV and QuantiSNP were used, with standard quality control metrics, to identify CNVs. CNV carriers were selected based on the following breakpoints according to the reference genome GRCh37/hg19: 16p11.2 proximal (BP4-5, 29.6-30.2MB), 1q21.1 distal (Class I, 146.4-147.5MB & II, 145.3-147.5MB), 22q11.2 proximal (BPA-D, 18.8-21.7MB) and 15q11.2 (BP1-2, 22.8–23.0MB). Control individuals did not carry any CNV at these loci. The CNV carriers were either probands referred to the genetic clinic for the investigation of neurodevelopmental and psychiatric disorders or their relatives (parents, siblings, and other relatives).

UK Biobank might represent the largest dataset of carriers affected by 15q11.2 deletions and duplications. Therefore, after identifying 15q11.2 deletions and duplications in the UK Biobank, we added the respective carriers to our clinical cohort. In other words, we excluded these subjects from the UK Biobank and treated them as part of our clinical dataset. Sensitivity analysis concluded that including this CNV locus does not change our main findings (Supp. Fig. 12). Controls were either non-carriers within the same families or individuals from the general population. Furthermore, controls were carefully matched for sex and age to CNV carriers.

## Clinical MRI data recording and processing

We analyzed a data sample of T1-weighted (T1w) images at 0.8–1 mm isotropic resolution. All T1w included in the analysis were quality checked by a domain expert[60]. Data for Voxel-Based Morphometry were preprocessed and analyzed with SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/)[61–63] running under MATLAB R2018b (https://www.mathworks.com/products/new_products/release2018b.html). Further quality control was performed using standardized ENIGMA quality control procedures (http://enigma.ini.usc.edu/protocols/imaging-protocols/). Finally, neurobiologically interpretable measures of gray matter volume were extracted in all participants by summarizing

whole-brain MRI maps in the MNI reference space. This feature-generation step was guided by the topographical brain region definitions of the commonly used Schaefer-Yeo atlas with 400 parcels[64]. The derived quantities of local gray matter volumetry resulted in 400 volume measures for each participant. As a data-cleaning step, derived regional brain volumes were adjusted for intracranial volume, age, age$^2$, and sex as fixed effects and scanning site as a random factor, following previous research on this dataset[60]. In particular, we have previously demonstrated that CNVs show independent effects on regional and total brain volumes[33]. Our current investigation is focused on how CNVs induce regional brain effects. Note that ancillary analyses revealed additional adjustments for total gray matter volume not to have any appreciable effect on subsequent analyses.

## Population data source

The UK Biobank is the largest biomedical resource that offers extensive behavioral and demographic assessments, medical and cognitive measures, as well as biological samples in a cohort of ~500,000 participants recruited from across Great Britain (https://www.ukbiobank.ac.uk/). The present study was based on the recent brain-imaging data release from February/March 2020. Our data sample included measurements from 39,085 participants with brain-imaging measures and expert-curated image-derived phenotypes of gray matter morphology (T1-weighted MRI) (Table 2). Among the participants, 48% were men and were 52% women with age between 40 and 69 y.o. when recruited [mean age 55 y.o., standard deviation (SD) 7.5 y.]). We benefited from the uniform data preprocessing pipelines designed and implemented by the FMRIB, Oxford University, Oxford, UK[65], to improve comparability and reproducibility.

MRI scanners (3T Siemens Skyra) at several dedicated data collection sites used matching acquisition protocols and standard Siemens 32-channel radiofrequency receiver head coils. Brain-imaging measures were defaced to protect the study participants' anonymity, and any sensitive meta-information was removed. Automated processing and quality control pipelines were deployed[36,65]. To improve the homogeneity of the brain-imaging scans, the noise was removed using 190 sensitivity features. This approach allowed for the reliable identification and exclusion of problematic brain scans, such as due to excessive head motion.

The structural MRI data were acquired as high-resolution T1-weighted images of brain anatomy using a 3D MPRAGE sequence at 1mm isotropic resolution. It was preprocessing included gradient distortion correction, the field of view reduction using the Brain Extraction Tool[66] and FLIRT[67], as well as non-linear registration to MNI152 standard space at 1 mm resolution using FNIRT[68]. To avoid unnecessary interpolation, all image transformations were estimated, combined, and applied by a single interpolation step. Tissue-type segmentation into the cerebrospinal fluid, gray matter and white matter to generate full bias-field-corrected images was achieved using FAST (FMRIB's Automated Segmentation Tool[69]). Finally, gray matter images were used to extract gray matter volumes in parcels according to the Schaefer-Yeo atlas with 400 regions[64]. Following previous work on the UKBB[70,71], inter-individual variations in brain region volumes that could be explained by nuisance variables of no interest were adjusted for by regressing out: body mass index, head size, head motion during task-related brain scans, head motion during resting-state fMRI scanning, head position and receiver coil in the scanner (x, y, and z), position of scanner table, as well as the data acquisition site.

## Statistical analysis for volumetric brain measures

All subsequent analyses were performed in Python v3.8 as a scientific computing engine (https://www.python.org/downloads/release/python-380/). We used Cohen's *d* to quantify the effect size of the CNVs on individual regional volumes. For a given region, Cohen's *d* is defined as:

$$d = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2 + s_2^2}{2}}},$$

where $\overline{x_1}$ corresponds to the mean region volume across CNV carriers, $\overline{x_2}$ corresponds to the mean region volume across controls. Similarly, $s_1$ and $s_2$ correspond to standard deviations of CNV carriers and controls.

Results from Cohen's *d* analyses were confirmed by a non-parametric effect size measure (Supp. Fig. 13).

We compared Cohen's *d* volumetric brain maps (and intermediate phenotypes brain maps) between different CNVs using Pearson's correlation. Furthermore, we used spin permutation testing to calculate empirical p-values for the ensuing correlation coefficient [72].

Finally, we calculated Lin's concordance correlation coefficient to quantify the agreement of similarities between volumetric Cohen's *d* maps, intermediate phenotypes, and PheWAS profiles. The degree of concordance between the two measures is thus calculated as:

$$\rho_C = \frac{2s_{1,2}}{s_1^2 + s_2^2 + (\overline{x_1} - \overline{x_2})^2},$$

where $s_{1,2}$ corresponds to the covariance between $x_1$ and $x_2$.

## Charting complex association using phenom-wide association study

We performed a rich annotation of the derived intermediate phenotypes by means of a phenome-wide association analysis benefitting from a wide variety of almost 1,000 lifestyle factors. For a detailed description of phenotype extraction and analysis, refer to our previously published studies[73]. Feature extraction was carried out using two utilities designed to obtain, clean, and normalize UKBB phenotype data according to predefined rules. In short, we collected a raw set of ~15,000 phenotypes that we further processed by the FMRIB UKB Normalisation, Parsing And Cleaning Kit (FUNPACK version 2.5.0; https://zenodo.org/record/4762700#.YQrpui2caJ8). FUNPACK is designed to perform automatic refinement on the UKB data, which includes removing 'do not know' responses and filling the blank left by unanswered sub-questions. The FUNPACK-derived phenotype information covered 11 major categories, including cognitive and physiological assessments, physical and mental health records, blood assays, as well as sociodemographic and lifestyle factors. The output consisted of a collection of 3,330 curated phenotypes which were then fed into PHEnome Scan ANalysis Tool (PHESANT[74], https://github.com/MRCIEU/PHESANT) for further refinement in an automated fashion. PHESANT performs further data cleaning and normalization along with labeling data as one of four data types: categorical ordered, categorical unordered, binary, and numerical. Categorical unordered variables were one-hot encoded, such that each possible response was represented by a binary column (true or false). The final curated inventory comprised 977 phenotypes spanning 11 FUNPACK-defined categories. Furthermore, we used Pearson's correlation to quantify the association strength between these 977 phenotypes with subject-specific expressions of our eight intermediate phenotypes (cf. below). To ensure that the correlations are not driven by a few outlying intermediate phenotype expressions, we first discarded 551 subjects based on Tukey's interquartile range rule for outlier detection.

## Multi-class prediction model and intermediate phenotype extraction

Technically, our core aim was to derive robust CNV-specific representations of intermediate phenotypes from a clinical sample that could be transferred to a large population resource for deep profiling. We derived the intermediate phenotypes as systematic brain morphometric co-deviations attributable to each of our eight target CNVs. To this end, we capitalized on linear discriminant analysis to extract separating rules between CNV carriers and controls based on whole-brain volume measurements. LDA can be viewed as a generative approach to classifying CNV carriers, which requires fitting multivariate Gaussian distribution to regional brain volumes and producing a linear decision boundary[75]. In particular, LDA-derived discriminant vectors/functions represented CNV-specific intermediate phenotypes. Using a linear model represents a data-efficient and directly biologically interpretable approach to our analysis, especially in our boutique datasets with limited subject samples[76]. These datasets are characteristic of the low sample size regularly encountered in biology and medicine, which typically impedes the application of more complex non-linear models that require high numbers of parameters to be estimated.

As another key model property of direct relevance to our present analysis goals, LDA can also be viewed as a dimensionality technique because this modeling framework enables the extraction of underlying coherent principles among our anatomical target regions that are most informative in telling apart CNV carriers from controls. To do so, LDA has access to class labels (CNV status in our case) and thus belongs to supervised techniques[75]. Specifically, LDA projects the input subjects' set of brain morphology measurements into a linear subspace, consisting of the directions which maximally separate our classes[77]. This dimensionality reduction quality of LDA was a necessary prerequisite for extracting intermediate phenotypes from one dataset and transferring them to other datasets.

In our study, we used LDA models to classify between CNV carriers and controls. Specifically, we derived a single LDA prototype for each CNV status, which yielded eight CNV-specific models. The dimensionality reduction capability of the LDA framework provides biologically interpretable compact views on distinguishing the CNV carriers and controls based on a linear combination of brain region volumes. As a general rule, the maximum number of dimensions equals the number of classes -1. Since each LDA model instance discriminated between two classes at hand (e.g., controls and 22q11.2 deletion), we obtained a one-dimensional vector encapsulating the 22q11.2 deletion intermediate phenotype. This vector of coefficients revealed the concomitant contribution of each brain region volume towards the separability of the CNV carriers based on whole-brain morphology measurements. Therefore, the coefficients provided quantitative information on the relative importance of the collective brain regions for CNV-health separation. Moreover, the LDA coefficients were estimated hand-in-hand with the other brain region volume effects, in contrast to the estimation of marginal or partial variable effects as in linear regression. Furthermore, to embed each subject's brain morphology in a low-rank subspace that maximally separates 22q11.2 deletion carriers and controls, we used the LDA coefficient vector to re-express (i.e., more formally, project) the set of 400 regional volumes of a given subject onto a single dimension representing 22q11.2 deletion intermediate expression level signature. Finally, as a step from dimensionality reduction to classification, these expressions of predictive subject brain morphology indicators were then used to construct a discriminant function.

## Building and validating robust prediction models

To recapitulate, our goal was to derive eight CNV-specific intermediate phenotypes using LDA. Therefore, we built separate CNV-specific LDA models designated to learn predictive principles to tell apart between CNV carriers and controls. However, we faced the challenge of the low number of CNV carriers. This challenge is inherent to various boutique datasets of rare medical conditions. Consequently, our number of measured features (regional brain volumes) was higher than the number of observation samples (subjects). Concretely, we disposed on average of 67 subjects per CNV class (cf. Table 1), while each subject was described by 400 regional volumes. Such a high-dimensional data scenario can lead to overfitting[78], where the model learns the detail and noise in the training samples and performs poorly in group classification on unseen test samples[75]. Hence, we used bootstrap aggregation (bagging), an ensemble learning method that can be used to reduce overfitting[79]. Bagging gains its value by profiting from a wisdom-of-crowds strategy. Concretely, we used a set of trained LDA models to obtain a more robust and better predictive performance than could be obtained from a single trained LDA model in isolation[79]. Such a model-averaging design improves classification performance by reducing variance[75].

We performed bagging during the derivation of LDA models separately for all eight CNV classes. Specifically, we used the following analytical strategy for a set of subjects consisting of a single CNV type and controls. In the first phase, a randomly perturbed version of the dataset is created by sampling the subject cohort with replacement. This bootstrap resampling served as the "in-the-bag" set of samples (i.e., subjects). The number of "in-the-bag" CNV carriers and controls equals their number in the dataset. Furthermore, the LDA model was trained on this training "in-the-bag" dataset. Model performance was then evaluated on all subjects from the dataset that were not selected for the "in-the-bag" dataset. These subject samples formed a testing "out-of-bag" dataset. The performance (i.e., classification accuracy) was based on the Mathews correlation coefficient, which has been reported to produce a more informative and truthful score than accuracy and F1 score[80]. The coefficient ranges between −1 and +1, where a coefficient of +1 represents a perfect prediction, 0 random prediction, and −1 indicates total disagreement between prediction and observation.

We repeated the bootstrap resampling procedure with 100 iterations. In so doing, we obtained different realizations of the entire analysis process and ensuing LDA model estimate. Concretely, the bagging algorithm resulted in 100 trained LDA models used to obtain 100 out-of-bag predictions in unseen subjects. We calculated the final prediction accuracy as a mean across the 100 performance estimates. Critically, the average over the collection of separately estimated LDA discriminant functions served as our CNV-specific intermediate phenotype that provided the basis for downstream analysis steps. Finally, we characterized each subject by the intermediate phenotype expression level, which we calculated as the average one-dimensional LDA projection of regional volume sets across the 100 replications. In summary, the variance of local information in the 100 redraws of our original clinical subject cohort promoted diversity among the obtained candidate predictive rules, thus strengthening the fidelity of our ultimate predictions.

To further safeguard against the risk of overfitting, we optimized the shrinkage parameter of each LDA model. Shrinkage corresponds to regularization used to stabilize the estimation of model parameters, such as in covariance matrices during model training. The empirical sample covariance is a poor estimator when the number of samples is small compared to the number of features. The covariance matrix estimation involved an interpolation between the sample covariance matrix based on the maximum likelihood estimator and a weighted identity matrix,

which amounted to the l2-penalization of the covariance matrix that then provided the basis for deriving a robust LDA solution.

Indeed, our sample covariance matrix held 80,200 unique entries, almost 1200 times more than the average number of CNV carriers available. Therefore, the vanilla estimation of the covariance matrix is singular and thus degenerate for downstream analysis steps, such as matrix inversion. To avoid such an inversion problem, we applied a dedicated shrinkage approach for the covariance matrix estimation step within LDA (ShrunkCovariance function from *sklearn*). Using a nested cross-validation architecture, we performed a rigorous search over 11 shrinkage hyper-parameter choices between 0 and 1, in steps of 0.1, in each "in-the-bag" bootstrap iteration (GridSearchCV function from *sklearn*). The optimal hyperparameter choice was based on a leave-one-out strategy. In this cross-validation technique, each sample of the "in-the-bag" dataset was used once as a test set of unseen subjects, while the remaining subject samples formed the training set.

Finally, we evaluated the significance of a cross-validated score and thus assessed whether our ensemble LDA model displayed above-chance classification performance. Specifically, we carried out a label permutation test to quantify whether our LDA model outperforms the empirical null model. The null distribution was generated by calculating the prediction accuracy of our LDA classifier on 100 different permutations of the dataset. In these, features remained unchanged, but class labels (i.e., CNV carriers or controls) were randomly shuffled. Such a shuffling corresponded to the null hypothesis, which states no dependency between the features and labels. LDA model displayed above-chance classification performance if its prediction accuracy was higher than the 97.5th percentile of prediction accuracy coefficient distribution derived from 100 permuted models.

## Performing model inspection using feature importance

After deriving robust LDA classifiers, we inspected which brain regions were the most informative in telling apart CNV carriers and controls. In other words, we aimed to contextualize and unpack the prediction rules of our ensemble LDA model. The bagging algorithm led to obtaining a collection of LDA models, resulting in a collection of estimates for each LDA coefficient and subject-specific intermediate phenotype expressions. Since each LDA model is trained on a different bootstrap population, it might happen that two distinct LDA models' coefficients would carry opposite signs due to the sign invariance of LDA dimensionality reduction. Therefore, we aligned all LDA models by multiplying them with -1 or 1 to produce a positive correlation between LDA coefficients and a corresponding Cohen's *d* map.

Furthermore, we designed a criterion to test which LDA coefficients are significant, meaning which features significantly contribute to the classification. Significant coefficients had the distribution of 100 LDA coefficients significantly different from 0. Specifically, they were robustly different from zero if their two-sided confidence interval according to the 2.5/97.5% bootstrap-derived distribution did not include zero.

## Carrying intermediate phenotype expressions over for deep characterization in other data resources

One of the aims of this study is to use a population dataset to investigate derived intermediate phenotypes. To do so, we transferred the CNV-specific intermediate phenotypes carefully derived in our boutique dataset and quantified their expression in the general population (i.e., UK Biobank). It is important to note that the derived intermediate phenotypes were not influenced by ASD or schizophrenia diagnosis (Supp. Fig. 14).

UK Biobank itself contains CNV carriers. Therefore, we aimed to validate the transferability of intermediate phenotypes by testing the difference in intermediate phenotype expression between CNV carriers and controls in both the clinical dataset and UK Biobank. Specifically, we tested the null hypothesis of no difference in the mean expression of intermediate phenotype in CNV carriers and controls. We adopted a two-sample bootstrap hypothesis test for means difference with 1,000 bootstrap replicates [81].

## Data availability

The majority of 16p11.2 data are publicly available (https://www.sfari.org/). For the 22q11.2 sample, raw data are available upon request from the PI (CB). All derived measures used in this study are available upon request (SJ). The rest of the CNV carriers' data cannot be shared as participants did not provide consent. All data from UK Biobank are available to other investigators online (ukbiobank.ac.uk). The Schaefer-Yeo atlas is accessible online (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal).

## Code availability

The processing scripts and custom analysis software used in this work are available in a publicly accessible GitHub repository along with examples of key visualizations in the paper: https://github.com/dblabs-mcgill-mila/CNV-convergence.

## Acknowledgements

DB was supported by the Brain Canada Foundation, through the Canada Brain Research Fund, with the financial support of Health Canada, National Institutes of Health (NIH R01 AG068563A, NIH R01 R01DA053301-01A1), the Canadian Institute of Health Research (CIHR 438531, CIHR 470425), the Healthy Brains Healthy Lives initiative (Canada First Research Excellence fund), Google (Research Award, Teaching Award), and by the CIFAR Artificial Intelligence Chairs program (Canada Institute for Advanced Research). This research was supported by Calcul Quebec (http://www.calculquebec.ca) and Compute Canada (http://www.computecanada.ca), the Brain Canada Multi-Investigator initiative, 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. SJ is a recipient of a Canada Research Chair in neurodevelopmental disorders, and a chair from the Jeanne et Jean Louis Levesque Foundation. The Cardiff CNV cohort was supported by the Wellcome Trust Strategic Award "DEFINE" and the National Centre for Mental Health with funds from Health and Care Research Wales (code 100202/Z/12/Z). The CHUV cohort was supported by the SNF (Maillard Anne, Project, PMPDP3 171331). Data from the UCLA cohort provided by CEB (participants with 22q11.2 deletions or duplications and controls) was supported through grants from the NIH (U54EB020403), NIMH (R01MH085953, R01MH100900, R03MH105808), and the Simons Foundation (SFARI Explorer Award). KK was supported by The Institute of Data Valorization (IVADO) Postdoctoral Fellowship program through the Canada First Research Excellence Fund. IES is supported by the Research Council of Norway (#223273), South-Eastern Norway Regional Health Authority (#2020060), European Union's Horizon2020 Research and Innovation Programme (CoMorMent project; Grant #847776) and Kristian Gerhard Jebsen Stiftelsen (SKGJ-MED-021). We thank all of the families participating at the Simons Searchlight sites and 16p11.2 European Consortium, Simons Searchlight Consortium. We appreciate obtaining access to brain-imaging and phenotypic data

on SFARI Base. We are grateful to all families who participated in the 16p11.2 European Consortium. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Author Contributions Statement

JK, DB, and SJ designed the study, analyzed imaging data, and drafted the manuscript. JK, CMod and KK did all the preprocessing and analysis of neuroimaging data. KS provided scripts for the PheWAS analysis. DB and SJ contributed to the interpretation of the results and in the editing of the manuscript. CMod, AM, AP, SR, and SM-B recruited and scanned participants in the 16p11.2 European Consortium. SL, COM, NY, PT, and ED recruited and scanned participants in the Brain Canada cohort. LK collected and provided the data for the UCLA cohort. DEJL, MJO, MBMVdB, JH, and AIS provided the data for the Cardiff cohort. All authors provided feedback on the manuscript. DB led data analytics.

## Competing Interests Statement

The authors declare no competing interests.

## Tables

**Table 1**.

**Clinical dataset demographics**.

CNV loci chromosome coordinates are provided with the number of genes encompassed in each CNV and with a well-known gene for each locus to help recognize the CNV. Other diagnoses included: language disorder, major depressive disorder, posttraumatic stress disorder, unspecified disruptive and impulse-control and conduct disorder, social anxiety disorder, social phobia disorder, speech sound disorder, moderate intellectual disability, specific learning disorder, gambling disorder, bipolar disorder, conduct disorder, attention-deficit/hyperactivity disorder, substance abuse disorder, global developmental delay, motor disorder, obsessive-compulsive disorder, sleep disorder, Tourette's disorder, mood disorder, eating disorders, transient tic disorder, trichotillomania, pervasive developmental disorder, specific phobia, body dysmorphic disorder, mathematics disorder, and dysthymic disorder. Abbreviations, Del: deletion; Dup: duplication; ASD: autism spectrum disorder; SZ: schizophrenia; chr: chromosome; Age: mean age; SD: standard deviation; nGenes: number of genes.

| Loci | Chr (hg19) start-stop | nGenes (Gene) | Type | Subjects | Age (SD) | Sex (M/ F) | ASD \| SZ diagnosis | Other diagnoses |
|---|---|---|---|---|---|---|---|---|
| 1q21.1 | chr1 | 7 | Del | 24 | 31 (18) | 9 / 15 | 0 \| 0 | 4 |
| | 146.53-147.39 | CHDIL | Dup | 15 | 33 (17) | 7 / 8 | 3 \| 0 | 2 |
| 15q11.2 | chr15 | 4 | Del | 112 | 55 (7) | 51 / 61 | 0 \| 0 | 2 |
| | 22.81-23.09 | CYFIP1 | Dup | 146 | 54 (7) | 69 / 77 | 0 \| 0 | 6 |
| 16p11.2 | chr16 | 27 | Del | 80 | 17 (12) | 46 / 34 | 10 \| 0 | 10 |
| | 29.65-30.20 | KCTD13 | Dup | 69 | 31 (14) | 37 / 32 | 7 \| 1 | 10 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 22q11.2 | chr22 | 49 | Del | 69 | 17 (9) | 33 / 36 | 8 \| 2 | 29 |
| | 19.04-21.47 | AIFM3 | Dup | 19 | 19 (14) | 12 / 7 | 2 \| 0 | 5 |
| | Controls | | | 312 | 26 (14) | 179 / 133 | 1 \| 0 | 12 |

862

**Table 2.**

**UK Biobank Imaging demographics.**

Our data sample included measurements from 39,085 participants with brain-imaging measures and expert-curated image-derived phenotype. Based on the cohort's sociodemographic, physical, lifestyle, and health-related characteristics, UK Biobank participants are known to be close to the general population[43] (Fry et al., 2017). CNVs were identified using PennCNV and QuantiSNP. UK Biobank might represent the largest dataset of carriers affected by 15q11.2 deletions and duplications. Therefore, we excluded these subjects from the UK Biobank and treated them as part of our clinical dataset. The remaining CNV carriers served for validation of derived LDA prediction patterns. [1]ICD10 code, including diagnoses of schizophrenia, schizotypal and delusional disorders (F20-F29). [2]ICD10 code, including diagnoses of childhood autism (F84.0), atypical autism (F84.1), Asperger's syndrome (F84.5), other pervasive developmental disorders (F84.8), and pervasive developmental disorder, unspecified (F84.9). Mean age is depicted along with standard deviation (SD).

| | Non-carriers | 1q21.1 | | 15p11.2 | | 16p11.2 | | 22q11.2 | |
|---|---|---|---|---|---|---|---|---|---|
| | | del | dup | del | dup | del | dup | del | dup |
| Subjects | 38731 | 12 | 14 | 117 | 155 | 4 | 7 | 5 | 47 |
| Percent female | 52 | 42 | 64 | 54 | 53 | 25 | 43 | 60 | 43 |
| Age (SD) | 55 (8) | 51 (6) | 54 (7) | 55 (7) | 54 (7) | 58 (3) | 55 (6) | 53 (8) | 54(8) |
| ASD[1] \| SCZ[2] diagnosis | 68 \| 18 | 0 \| 0 | 0 \| 0 | 0 \| 0 | 0 \| 0 | 0 \| 0 | 0 \| 0 | 0 \| 0 | 0 \| 0 |

# Figure legends

**Figure 1**

**Eight CNVs lead to largely distinct spatial patterns of abnormalities in brain morphology.**

We analyzed gray matter region volumes in 534 subjects carrying one of eight CNVs and 312 controls. Regional volumes were adjusted for intracranial volume, age, age$^2$, sex, and acquisition site. a) Cohen's *d* brain map quantifies the magnitude of structural change for each CNV. We have computed Cohen's *d* between CNV carriers and controls separately for each of the 400 brain regions (Schaefer-Yeo reference atlas). Our analysis reveals increased (red) and decreased (blue) brain volumes depending on the variation type. The uncovered patterns of volumetric changes confirm established knowledge on the regional increase and decrease across CNV loci[22,23]. b) Examining associations between Cohen's *d* brain maps rendered on brain surface from each pair of CNVs. The wide range and low magnitude of Pearson's correlations show that CNVs have distinct effects on brain volumes (more red=more similar, more blue=more dissimilar). Average similarity stands for the mean absolute Pearson's correlations across all CNVs. 22q11.2 and 16p11.2 deletions and duplications show strong mirroring (opposing) effects. Asterisk denotes FDR-corrected spin permutation p-values. c) Projecting brain volumes onto two dominant dimensions of variation using principal component analysis (PCA). Although the first two dominant PCA components explain 18 % of the variance, they are unrelated to differences between CNVs. The light and dark symbols represent deletions and duplication, respectively. The gray hexagonal bin plot represents the frequency of controls. Controls were not used to calculate the PCA and were projected post hoc. d) Projections of brain volumes to two dimensions using linear discriminant analysis (LDA). The first LDA dimension ($LD_1$) mainly captures differences between 16p11.2 proximal deletion and duplication, while the second LDA dimension ($LD_2$) mainly captures differences between 22q11.2 deletions and duplications. Symbols and hexagonal binning plots were constructed in the same way as for the PCA approach. CNVs lead to distinct changes

often represented by a predominant increase or decrease in the gray matter cortex that could effectively be described using low dimensional representations derived by LDA models.

**Figure 2**

**Pattern-learning models extract distinct intermediate brain phenotypes from CNV status.**

We estimated eight LDA models to classify between controls and each of the eight different CNVs. a) Classification performance of eight distinct LDA models when telling apart controls and CNV carriers, given as Matthews correlation coefficient. All eight CNVs are successfully classified based on brain structure at above-chance accuracy as their performance exceeds that of an empirical null model (black line depicts upper 2.5 percentile threshold of the null distribution obtained by label shuffling). b) Prediction rule derived for each of the eight CNV-specific LDA models projected on the brain (red/blue = positive/negative weight). The prediction rule is a CNV-specific brain signature and can be treated as an intermediate phenotype. C) Similarity between CNV-specific intermediate phenotypes. The wide range and low magnitudes of ensuing Pearson's correlations reflect the disparity in the captured intermediate phenotypes. Average similarity represents the mean absolute correlation across all CNVs. Asterisk denotes FDR-corrected spin permutation p-values. d) Relationship between Cohen's *d* brain maps and intermediate phenotypes. Based on FDR-corrected Pearson's correlations, all eight intermediate phenotypes appear to largely follow the respective Cohen's *d* brain maps. LDA models identified and quantified CNV-specific intermediate phenotypes that effectively captured distinct morphometric differences between CNV carriers and the general population.

**Figure 3**

**Intermediate brain phenotypes track structural changes with distinct impacts in large-scale networks.**

We identified which aspects of the LDA-derived prediction rule robustly contributed to classification success by calculating 100 bootstrapped LDA models for each CNV while sampling CNV carriers randomly. A) Percentage of statistically relevant LDA coefficients in a CNV carrier group among all the regions that belong to each brain network (one-sample bootstrap hypothesis test for non-zero mean with 10,000 replicates). For example, 16p11.2 proximal deletion strongly affects most large-scale networks except the frontoparietal network. Altogether, the estimated LDA coefficients represent the backbone of each intermediate phenotype. Large-scale networks correspond to seven SchaeferYeo networks; Vis: Visual, FrontPar: Frontoparietal, SomMot: Somatomotor, DorsAttn: Dorsal attention, SalVenAttn: Salience ventral attention, Limbic, Frontoparietal, Default: Default mode. b) Significant LDA coefficients grouped by the large-scale networks. The highest relative number of affected regions is in the limbic network. Conversely, regions in the frontoparietal network are targeted less frequently. c) Relationship between CNV effects and LDA performance. There is a significant positive correlation between the number of significant LDA coefficients and classifier performance, unlike for the sample size of the cohort (marker size). According to the eight specific LDA models, CNVs affected predominantly high-level networks such as the limbic, salience, and default-mode networks.

**Figure 4**

**Using intermediate CNV phenotypes as a basis for phenome-wide association analysis.**

We performed a phenome-wide association study (PheWAS) by computing Pearson's correlation between the expression of each of the eight intermediate CNV phenotypes and 977 phenotypes spanning 11 categories in 39,085 UK Biobank subjects. a) Letter-value (boxen) plot for the expression of 16p11.2 proximal duplication intermediate phenotype is shown for the sake of illustration. The boxen plot depicts the distribution of quantiles for the expression scores computed by quantifying the presence of derived 16p11.2 proximal duplication intermediate phenotype in both the clinical cohort (left) and the UK Biobank (right). Based on a two-sample bootstrap hypothesis test for difference of means with 10,000 bootstrap replicates, the 16p11.2 proximal duplication carriers significantly differed in the expression level from controls both in the clinical cohort ($p < 10^{-4}$) and UK Biobank dataset ($p < 10^{-4}$). b) PheWAS study using the CNV-specific intermediate phenotype. We calculated the Pearson's correlation between the expression of 16p11.2 proximal duplication intermediate phenotype and each of 977 phenotypes. After the Bonferroni correction for multiple comparisons (BON), there were 55 significant associations, such as education score, hemoglobin concentration, or physically abused by family as a child. There were 145 significant associations exceeding false discovery rate correction (FDR) c) The relative number of significant correlations summarized for each of the eleven categories for each CNV in the UK Biobank. Most CNVs are strongly associated with multiple categories and their respective phenotypes. For example, up to 35% of phenotypes in general physical measures show a significant correlation with four CNV brain signatures. The light and dark symbols represent deletions and duplication, respectively. As an insight from the performed phenome-wide association analysis, CNV brain signatures are linked with multiple phenotypes across most categories but mainly in the general physical measures, blood assays, and early life factors categories.

**Figure 5**

**Eight different CNVs converge on similar phenome-wide association profiles.**
We carried out the PheWAS analysis for each intermediate phenotype to quantify the differences and commonalities in phenotypical consequences due to the eight CNVs. a) Pearson's correlations from PheWAS analysis for each CNV status. Among those, 22q11.2 deletion shows the strongest associations with numerous phenotypes across categories. Colors indicate the eleven categories. b) Linear association strength between PheWAS outcomes across all CNVs. Strong Pearson's correlations suggest that CNVs are linked with similar phenotypes. Average similarity exceeds those of volumetric Cohen's *d* maps and intermediate phenotypes. Asterisk denotes FDR-corrected significant correlations. c) Linear association strength between category-specific Pearson's correlations from the PheWAS analysis across all CNVs. Detailed visualization depicts the similarity of the impact of CNVs on all phenotype categories. The direction of the linear relationship tends to be identical across categories for a given CNV pair (strong negative or strong positive), unlike across CNV pairs for a given category. The eight CNVs exhibited similar PheWAS profiles, especially in bone density, blood assays, and general physical measures categories.

**Figure 6**
**Detailing aspects convergence in phenome-wide portfolios across different CNVs.**
For all eight CNVs, we delineate the most prominent as well as distinctive associations among their PheWAS profiles in 39,085 UK biobank participants. We also compare CNVs based on their brain and behavior similarities. a) Phenotypes from the PheWAS analysis most strongly associated with the eight CNVs. We show ten phenotypes with the strongest average Pearson's correlations across all CNVs. The most prominent association across CNVs is with diastolic blood pressure. The box plot displays the first quartile, median, third quartile, and whiskers corresponding to the appropriate quartile plus 1.5 times the interquartile range. b) Phenotypes most consistently associated with the eight CNVs. We find eight phenotypes associated with most (six) of the CNVs. Phenotypes are ordered according to the mean strength of the association. Most of the phenotypes are from the blood assays category. c) Number of significant hits per category for each intermediate phenotype conditioned on the shared phenotypical profile. For each of the eight intermediate phenotype expressions, we regressed out the remaining seven. Even after conditioning on the shared phenotypical associations, each particular CNV still shows a specific set of distinct phenome-wide associations across various categories. For example, 22q11.2 deletion still displays a high number of associations in physical measures - general category. d) Concordance between brain volume effects and PheWAS effects. The absolute value of correlation between Cohen's *d* brain maps (Fig. 1a) is plotted against the absolute value of correlation between PheWAS profiles. Negative Lin's concordance correlation hints at the disparity between volumetric and phenotypical similarity. Moreover, the majority of points lie above the 45°-degree line suggesting that PheWAS similarities are more substantial than volumetric similarities. e) From diverging brain patterns to converging portfolios. Each line represents a similarity of Cohen's *d* map, intermediate phenotype, and PheWAS profile for a given CNV pair. Convergence on PheWAS profiles is demonstrated by the increase in similarity in 22 of 28 CNV pairs. Hence, the similarity of CNV portfolios exceeded that of volumetric intermediate phenotypes.

# References


1. Freeman, J. L. *et al.* Copy number variation: New insights in genome diversity. *Genome Res.* **16**, 949–961 (2006).
2. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
3. Auwerx, C. *et al.* The individual and global impact of copy-number variants on complex human traits. *Am. J. Hum. Genet.* S0002-9297(22)00061–1 (2022) doi:10.1016/j.ajhg.2022.02.010.
4. 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).
5. Rutkowski, T. P. *et al.* Unraveling the genetic architecture of copy number variants associated with schizophrenia and other neuropsychiatric disorders. *J. Neurosci. Res.* **95**, 1144–1160 (2017).
6. Lauer, S. & Gresham, D. An evolving view of copy number variants. *Curr. Genet.* **65**, 1287–1295 (2019).
7. Huguet, G. *et al.* Genome-wide analysis of gene dosage in 24,092 individuals estimates that 10,000 genes modulate cognitive ability. *Mol. Psychiatry* **26**, 2663–2676 (2021).
8. Moberg, P. J. *et al.* Neurocognitive Functioning in Patients with 22q11.2 Deletion Syndrome: A Meta-Analytic Review. *Behav. Genet.* **48**, 259–270 (2018).
9. Silva, A. I. *et al.* Neuroimaging findings in neurodevelopmental copy number variants: identifying molecular pathways to convergent phenotypes. *Biol. Psychiatry* (2022) doi:10.1016/j.biopsych.2022.03.018.

10. Chawner, S. J. R. A. *et al.* Genotype-phenotype associations in children with copy number variants associated with high neuropsychiatric risk in the UK (IMAGINE-ID): a case-control cohort study. *Lancet Psychiatry* **6**, 493–505 (2019).
11. 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).
12. Sanders, S. J. *et al.* A framework for the investigation of rare genetic disorders in neuropsychiatry. *Nat. Med.* **25**, 1477–1487 (2019).
13. Moreno-De-Luca, D. & Martin, C. L. All for one and one for all: heterogeneity of genetic etiologies in neurodevelopmental psychiatric disorders. *Curr. Opin. Genet. Dev.* **68**, 71–78 (2021).
14. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J. Med. Genet.* **56**, 131–138 (2019).
15. Owen, D. *et al.* Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. *BMC Genomics* **19**, 867 (2018).
16. Adams, R. L. *et al.* Psychopathology in adults with copy number variants. *Psychol. Med.* 1–8 (2022) doi:10.1017/S0033291721005201.
17. Kendall, K. M. *et al.* Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank. *Br. J. Psychiatry J. Ment. Sci.* **214**, 297–304 (2019).
18. Doelken, S. C. *et al.* Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. *Dis. Model. Mech.* **6**, 358–372 (2013).
19. Viñas-Jornet, M. *et al.* High Incidence of Copy Number Variants in Adults with Intellectual Disability and Co-morbid Psychiatric Disorders. *Behav. Genet.* **48**, 323–336 (2018).

20. Moreau, C. A. *et al.* Dissecting autism and schizophrenia through neuroimaging genomics. *Brain* **144**, 1943–1957 (2021).
21. Moreau, C. A. *et al.* Mutations associated with neuropsychiatric conditions delineate functional brain connectivity dimensions contributing to autism and schizophrenia. *Nat. Commun.* **11**, 5272 (2020).
22. Modenato, C. *et al.* Lessons Learned From Neuroimaging Studies of Copy Number Variants: A Systematic Review. *Biol. Psychiatry* **90**, 596–610 (2021).
23. Sønderby, I. E. *et al.* Effects of copy number variations on brain structure and risk for psychiatric illness: Large-scale studies from the ENIGMA working groups on CNVs. *Hum. Brain Mapp.* **43**, 300–328 (2022).
24. Raznahan, A., Won, H., Glahn, D. C. & Jacquemont, S. Convergence and divergence of rare genetic disorders on brain phenotypes. *JAMA Psychiatry* (2022).
25. Zinkstok, J. R. *et al.* Neurobiological perspective of 22q11.2 deletion syndrome. *Lancet Psychiatry* **6**, 951–960 (2019).
26. Littlejohns, T. J. *et al.* The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat. Commun.* **11**, 2624 (2020).
27. Weiss, L. A. *et al.* Association between Microdeletion and Microduplication at 16p11.2 and Autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
28. Marek, S. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
29. Gottesman, I. I. & Gould, T. D. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am. J. Psychiatry* **160**, 636–645 (2003).

30. Mark, W. & Toulopoulou, T. Cognitive intermediate phenotype and genetic risk for psychosis. *Curr. Opin. Neurobiol.* **36**, 23–30 (2016).
31. Meyer-Lindenberg, A. & Weinberger, D. R. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nat. Rev. Neurosci.* **7**, 818–827 (2006).
32. Lin, A. *et al.* Mapping 22q11.2 Gene Dosage Effects on Brain Morphometry. *J. Neurosci.* **37**, 6183–6199 (2017).
33. Martin-Brevet, S. *et al.* Quantifying the Effects of 16p11.2 Copy Number Variants on Brain Structure: A Multisite Genetic-First Study. *Biol. Psychiatry* **84**, 253–264 (2018).
34. Sun, D. *et al.* Large-scale mapping of cortical alterations in 22q11.2 deletion syndrome: Convergence with idiopathic psychosis and effects of deletion size. *Mol. Psychiatry* **25**, 1822–1834 (2020).
35. Bracher-Smith, M. *et al.* Effects of pathogenic CNVs on biochemical markers: a study on the UK Biobank. *bioRxiv* 723270 (2019) doi:10.1101/723270.
36. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).
37. Momma, K. Cardiovascular Anomalies Associated With Chromosome 22q11.2 Deletion Syndrome. *Am. J. Cardiol.* **105**, 1617–1624 (2010).
38. Pillinger, T., D'Ambrosio, E., McCutcheon, R. & Howes, O. D. Is psychosis a multisystem disorder? A meta-review of central nervous system, immune, cardiometabolic, and endocrine alterations in first-episode psychosis and perspective on potential models. *Mol. Psychiatry* **24**, 776–794 (2019).
39. Van, L. *et al.* All-cause mortality and survival in adults with 22q11.2 deletion syndrome. *Genet. Med.* **21**, 2328–2335 (2019).

40. Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry Off. J. World Psychiatr. Assoc. WPA* **6**, 168–176 (2007).
41. Correll, C. U. *et al.* Prevalence, incidence and mortality from cardiovascular disease in patients with pooled and specific severe mental illness: a large-scale meta-analysis of 3,211,768 patients and 113,383,368 controls. *World Psychiatry Off. J. World Psychiatr. Assoc. WPA* **16**, 163–180 (2017).
42. Hoang, U., Goldacre, M. J. & Stewart, R. Avoidable mortality in people with schizophrenia or bipolar disorder in England. *Acta Psychiatr. Scand.* **127**, 195–201 (2013).
43. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
44. Kirov, G. *et al.* The penetrance of copy number variations for schizophrenia and developmental delay. *Biol. Psychiatry* **75**, 378–385 (2014).
45. Chawner, S. J., Watson, C. J. & Owen, M. J. Clinical evaluation of patients with a neuropsychiatric risk copy number variant. *Curr. Opin. Genet. Dev.* **68**, 26–34 (2021).
46. Alessi, M. G. & Bennett, J. M. Mental health is the health of the whole body: How psychoneuroimmunology & health psychology can inform & improve treatment. *J. Eval. Clin. Pract.* **26**, 1539–1547 (2020).
47. Guest, P. C. Psychiatric Disorders as "Whole Body" Diseases. in *Biomarkers and Mental Illness: It's Not All in the Mind* (ed. Guest, P. C.) 3–16 (Springer International Publishing, 2017). doi:10.1007/978-3-319-46088-8_1.
48. Kohane, I. S. *et al.* The Co-Morbidity Burden of Children and Young Adults with Autism Spectrum Disorders. *PLOS ONE* **7**, e33224 (2012).

49. Sotelo, J. L. & Nemeroff, C. B. Depression as a systemic disease. *Pers. Med. Psychiatry* **1–2**, 11–25 (2017).

50. Leboyer, M. *et al.* Can bipolar disorder be viewed as a multi-system inflammatory disease? *J. Affect. Disord.* **141**, 1–10 (2012).

51. Leppert, B. *et al.* A cross-disorder PRS-pheWAS of 5 major psychiatric disorders in UK Biobank. *PLOS Genet.* **16**, e1008185 (2020).

52. Thompson, P. M. *et al.* Genetic influences on brain structure. *Nat. Neurosci.* **4**, 1253–1258 (2001).

53. Bacchelli, E. *et al.* An integrated analysis of rare CNV and exome variation in Autism Spectrum Disorder using the Infinium PsychArray. *Sci. Rep.* **10**, 3198 (2020).

54. Gratten, J. & Visscher, P. M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med.* **8**, 78 (2016).

55. Seidlitz, J. *et al.* Transcriptomic and cellular decoding of regional brain vulnerability to neurogenetic disorders. *Nat. Commun.* **11**, 3358 (2020).

56. Moreau, C. A., Ching, C. R., Kumar, K., Jacquemont, S. & Bearden, C. E. Structural and functional brain alterations revealed by neuroimaging in CNV carriers. *Curr. Opin. Genet. Dev.* **68**, 88–98 (2021).

57. Rogdaki, M. *et al.* Magnitude and heterogeneity of brain structural abnormalities in 22q11.2 deletion syndrome: a meta-analysis. *Mol. Psychiatry* **25**, 1704–1717 (2020).

58. Wysocki, A. C., Lawson, K. M. & Rhemtulla, M. Statistical Control Requires Causal Justification. *Adv. Methods Pract. Psychol. Sci.* **5**, 25152459221095824 (2022).

59. Modenato, C. *et al.* Effects of eight neuropsychiatric copy number variants on human brain structure. *Transl. Psychiatry* **11**, 1–10 (2021).

60. Modenato, C. *et al.* Effects of eight neuropsychiatric copy number variants on human brain structure. *Transl. Psychiatry* **11**, 399 (2021).
61. Ashburner, J. A fast diffeomorphic image registration algorithm. *NeuroImage* **38**, 95–113 (2007).
62. Ashburner, J. & Friston, K. J. Unified segmentation. *NeuroImage* **26**, 839–851 (2005).
63. Lorio, S. *et al.* New tissue priors for improved automated classification of subcortical brain structures on MRI. *NeuroImage* **130**, 157–166 (2016).
64. Schaefer, A. *et al.* Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb. Cortex N. Y. N 1991* **28**, 3095–3114 (2018).
65. Alfaro-Almagro, F. *et al.* Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* **166**, 400–424 (2018).
66. Smith, S. M. Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**, 143–155 (2002).
67. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17**, 825–841 (2002).
68. Andersson, J. L., Jenkinson, M. & Smith, S. Non-linear registration, aka Spatial normalisation FMRIB technical report TR07JA2. *FMRIB Anal. Group Univ. Oxf.* **2**, e21 (2007).
69. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).
70. Schurz, M. *et al.* Variability in Brain Structure and Function Reflects Lack of Peer Support. *Cereb. Cortex* **31**, 4612–4627 (2021).

71. Spreng, R. N. *et al.* The default network of the human brain is associated with perceived social isolation. *Nat. Commun.* **11**, 6393 (2020).
72. Alexander-Bloch, A., Giedd, J. N. & Bullmore, E. Imaging structural co-variance between human brain regions. *Nat. Rev. Neurosci.* **14**, 322–336 (2013).
73. Savignac, C. *et al.* APOE ε2 vs APOE ε4 dosage shows sex-specific links to hippocampus-default network subregion co-variation. 2022.03.15.484482 Preprint at https://doi.org/10.1101/2022.03.15.484482 (2022).
74. 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).
75. Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning: data mining, inference, and prediciton. (2009).
76. Bzdok, D. & Ioannidis, J. P. A. Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends Neurosci.* **42**, 251–262 (2019).
77. Hart, P. E., Stork, D. G. & Duda, R. O. *Pattern classification*. (Wiley Hoboken, 2000).
78. Bzdok, D., Nichols, T. E. & Smith, S. M. Towards Algorithmic Analytics for Large-scale Datasets. *Nat. Mach. Intell.* **1**, 296–306 (2019).
79. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
80. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
81. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (CRC Press, 1994).