

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 phenotypic convergence

Jakub Kopal^{1,2}, Kuldeep Kumar³, Karin Saltoun^{1,2}, Claudia Modenato⁵, Clara A. Moreau⁴, Sandra Martin-Brevet⁵, Guillaume Huguet³, Martineau Jean-Louis³, Charles-Olivier Martin³, Zohra Saci³, Nadine Younis³, Petra Tamer³, Elise Douard³, Anne M. Maillard⁶, Borja Rodriguez-Herreros⁶, Aurèlie Pain⁶, Sonia Richetin⁶, Leila Kushan⁷, 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 Lippé³, Bogdan Draganski^{5,12}, Ida E. Søndersby^{13,14,15}, Ole A. Andreassen^{13,15}, David C. Glahn¹⁶, Paul M. Thompson¹⁷, Carrie E. Bearden⁷, Sébastien Jacquemont³, *Danilo Bzdok^{1,2,18}

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

41 Corresponding author

42 Danilo Bzdok, danilobzdok@gmail.com

43

44 Abstract

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

59 Introduction

60 A chief goal of modern neuroscience is understanding how genetic variation impacts brain
61 organization and inter-individual differences in behavior. Advances in genomic microarray
62 technology streamlined the detection of copy number variations (CNVs) – deletions or
63 duplications of chromosomal segments of >1000 base pairs^{1,2}. This class of genetic mutations
64 opens a unique window into the investigation of how neurogenetic determinants shape human
65 behavior, cognition, and development^{3,4}. Pathogenic CNVs that reoccur across individuals
66 provide opportunities to study groups of individuals who carry the same deletion or duplication
67 of a well-defined set of genes⁵. Moreover, CNVs have larger effects on phenotype than the low
68 effect-size single-nucleotide polymorphisms often identified by genome-wide association
69 studies⁶. Concretely, CNVs overall have been shown to detrimentally affect cognition and raise
70 the risk for psychiatric conditions^{4,7}. Nevertheless, it remains unexplained why many different
71 CNVs escalate vulnerability for the same developmental and psychiatric disorders^{4,8,9}.

72
73 The vast majority of large recurrent CNVs have been linked to more than one clinical
74 diagnosis, including intellectual disability, autism spectrum disorders, and schizophrenia¹⁰⁻¹².
75 These findings make a case that circumscribed genetic changes are rarely exclusively associated
76 with a single clinical diagnosis¹³. Further, CNVs have demonstrable consequences even in
77 seemingly unaffected middle and old age carriers, who show no overt signs of early-onset
78 neuropsychiatric disorders. Recent evidence points to a broader spectrum of impacts from CNV
79 status, ranging from physical traits to diabetes to hypertension to obesity to renal
80 dysfunction^{3,14,15}, as well as psychopathology¹⁶. Understudied body-wide CNV effects may
81 contribute to the links of schizophrenia-associated CNVs with diminished academic
82 qualifications, occupation, or household income¹⁷. In summary, this class of genetic variants
83 affecting distant parts of the genome can be associated with various behavioral and clinical
84 phenotypes^{18,19}.

85
86 Despite many advances in genomic profiling, investigations into the corresponding brain
87 signatures have only been performed for a few CNVs and mostly focused on a single variant at
88 a time^{20,21}. These parallel approaches to catalog CNVs highlighted a wide spectrum of robust
89 effects on brain structure^{22,23}. Although distinct rare CNVs are associated with a range of brain
90 alterations, they have been suggested to lead to a degree of similarity in associated behavioral
91 phenotypes^{9,24}. However, it remains unknown how similar CNVs are in terms of their effects on
92 the brain and the phenome. Since deleterious CNVs are rare, such as 1 in 3,000 for 22q11.2
93 deletion²⁵, previous investigations suffered from small samples of subjects and a lack of
94 phenotypic depth. Therefore, previous studies were chronically underpowered to paint a
95 complete picture of CNVs in medicine. There is a need for a systematic investigation of
96 intermediate brain measures and their phenotypic associations across several CNVs by means
97 of a large well-phenotyped patient pool. The recent advent of population cohorts with rich
98 phenotypic assessment batteries represents an untapped opportunity to conjointly examine a
99 set of CNVs and characterize them at an unprecedented scale.

100
101 In the present study, we interrogated the largest existing biomedical data resource, the
102 UK Biobank²⁶, which allowed a head-to-head comparison of an envelope of CNVs. As a first step,
103 we leveraged tools from machine learning, including linear discriminant analysis (LDA), to isolate
104 CNV-specific brain morphology signatures from a multisite clinical cohort. These individuals
105 carried one of eight recurrent CNVs that are among the most widely studied CNV loci to

106 date^{11,23,27}. Deletions and duplications at the loci 1q21.1, 15q11.2, 16p11.2, and 22q11.2 strike
107 a balance between being frequent and having a significant impact on brain and behavior²³.
108 Subsequently, the advantageous properties of LDA allowed us to carry over the CNV-specific
109 whole-brain signatures from the clinical cohort to the large-scale UK Biobank cohort. The UK
110 Biobank is ideally suited to tease apart the commonalities in phenotypic indicators across CNV
111 alterations due to the breadth of available phenotypic annotations. We directly linked a rich
112 portfolio of phenotypes to eight CNV brain signatures in ~40,000 UK Biobank participants.
113 Specifically, we performed separate phenome-wide association studies (PheWAS) for the eight
114 CNV-brain-imaging signatures across 977 phenotypes from eleven categories. In this way, we
115 provide a population-level characterization of what unites and divides the eight CNVs by
116 detailing convergences and divergences from genomic variants to brain morphology to
117 phenome. In an attempt to establish cornerstone evidence for the community, such a study can
118 illuminate fundamental links between genetic variation and brain organization, with their
119 consequences to bodily systems.

120 Results

121 Dissecting different CNV effects on whole-brain morphology

122 We systematically analyzed volumetric measures derived from brain-imaging scans in the
123 clinical cohort comprising 846 total subjects: 534 carried one of eight recurrent CNVs (deletion
124 and duplications of 1q21.1 distal, 15q11.2 BP1-BP2, 16p11.2 proximal, or 22q11.2 proximal),
125 while 312 controls did not carry a CNV (Table 1). We parsed volume measures from these
126 structural brain scans using a 400-region anatomical definition (Schaefer-Yeo reference atlas;
127 see Online methods). To account for variation outside of our current primary scientific interest,
128 each brain region volume was adjusted for intracranial volume, age, age², sex, and acquisition
129 site for all downstream analysis steps. A schematic flow of all analysis steps is depicted in
130 Supplementary Figure 1.

131 As a first step, we compared the effects on brain region volume measures for the eight
132 CNVs. Specifically, after normalizing (z-scoring) brain volumes across groups, that is, across the
133 respective CNV carriers and controls, we examined the extent of volumetric divergence between
134 carriers of each single CNV and controls by computing Cohen's *d* (giving an effect size for the
135 group difference) for each individual brain region (Fig. 1a). In doing so, for each examined CNV,
136 we obtained a brain map of Cohen's *d* effect sizes that summarize magnitudes of CNV-induced
137 structural abnormalities across the brain's gray matter. We noted widespread smaller volumes
138 in the majority of the examined atlas regions for the 1q21.1 deletion, 15q11.2 duplication,
139 16p11.2 duplication, and 22q11.2 deletion. Conversely, a preferential increase in most regional
140 volumes became apparent for the 1q21.1 duplication, 15q11.2 deletion, 16p11.2 deletion, and
141 22q11.2 duplication. These findings align with well-known regional alterations identified in
142 cohorts with patients carrying neurodevelopmental disorders²².

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

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

157

158 Visualizing CNV differences in low-dimensional signatures

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

176 Therefore, we turned to linear discriminant analysis (LDA) as a pattern classification
177 algorithm that is naturally capable of recovering a low-dimensional representation explicitly
178 aimed at maximizing the separation between the eight CNVs based on the individuals' brain
179 morphometry measures. We then re-expressed the brain-wide regional volumes as the two
180 primary dimensions of structural variation under the LDA model (Fig. 1d). In particular, the
181 leading dimension of the LDA-derived subject embedding captured the differences between
182 16p11.2 deletion and duplications. The second most explanatory dimension of the LDA-derived
183 embedding mainly captured the differences between 22q11.2 deletion and duplications. This
184 distribution of a single CNV locus along a single dimension points again at similar structural
185 effects with opposite directions. In summary, LDA formed a new low-dimensional space in which
186 the brain morphology of CNV carriers could be effectively identified, quantified, and,
187 subsequently, examined in further detail.

188

189 Deriving CNV-specific intermediate phenotypes

190 To supplement the multi-CNV classification model, which explored differences between
191 CNVs (described above), our next analysis step was to extract robust whole-brain signatures
192 specific for each CNV that we could then use to study unseen participants in any number of
193 external cohorts. Therefore, we constructed eight LDA models of order one dedicated to the
194 eight CNVs. Notably, there was a considerable imbalance between the number of controls and
195 CNV carriers (from 2-fold for 15q11.2 duplication to 22-fold for 1q21.1 duplication). Moreover,
196 the number of model parameters to be estimated (at least 400 parameters associated with the
197 400 atlas regions) was larger than the number of subjects. To remedy the challenges of this data
198 scenario, our analysis pipeline combined bagging and regularization to prevent overfitting the

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

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

222 We further inspected the 400 region coefficients of each LDA model that captured the
223 influence of each CNV on each brain region. By carrying out a one-sample bootstrap hypothesis
224 test independently for each CNV, we assessed which region-specific model coefficients are
225 robustly different from zero and, thus, robustly affected by CNVs. Specifically, during the
226 learning of the coefficients of one of the 8 CNV-specific LDA models, in 100 resampling iterations;
227 we drew a different set of subjects based on drawing subjects with replacement from the control
228 subjects and corresponding CNV carriers. Statistically relevant coefficients were robustly
229 different from zero if their two-sided confidence interval - according to the 2.5/97.5% intervals
230 of the bootstrap-derived distribution - did not include zero. Different CNVs affected (displayed
231 statistically relevant coefficients) different cortical parcels that correspond to the seven large-
232 scale brain networks populating the cortex, as defined by our atlas (Fig. 3a). For example, while
233 16p11.2 proximal duplication primarily affects 20% of all regions in the limbic network, 22q11.2
234 deletion affects 20% of regions in the salience ventral attentional network as well as more than
235 10% of regions in the limbic, dorsal attentional, and default-mode networks. Across all examined
236 CNVs and target brain networks, the 16p11.2 deletion affected the largest number of brain
237 regions, while 15q11.2 duplication affected the lowest number of regions. Higher-order network
238 circuits showed, on average, the relatively highest number of significant coefficients. Concretely,
239 the limbic network had the highest relative number of affected regions, followed by the salience
240 and default-mode networks (Fig. 3b). Together, the wide range of effects on the large-scale
241 networks again highlights the diverging consequences of CNVs on brain morphometry.

242 To further explore characteristic relationships between the eight CNVs, we probed for
243 a linear relationship of the number of salient LDA coefficients with LDA classifier performance
244 and average brain-wide Cohen's d . We found a significant positive Pearson's correlation with
245 classifier performance ($r = 0.74$, $p = 0.04$) (Fig. 3c) and mean absolute effect size ($r = 0.75$, $p =$
246 0.03). Furthermore, when we included sample size in the testing scheme, we found only a

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

251

252 [Lifting over phenotypes patterns from the clinical cohort](#)

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

263 To carry over the intermediate phenotypes from the clinical cohort to the UK Biobank,
264 we quantified the expression of each intermediate CNV phenotype for all 39,085 UK Biobank
265 participants (Table 2). We first extracted brain volume measures from the 400 atlas regions,
266 adjusting for several confound variables (see Online methods). We then calculated the subject-
267 specific expression for all intermediate CNV phenotypes in the UK Biobank. It is important to
268 stress that the intermediate phenotypes were derived in the clinical cohort. However, UK
269 Biobank also contains several carriers of the analyzed mutations. The generalizability of the
270 derived intermediate phenotypes was indicated by the difference between the intermediate
271 phenotype expression level of non-carriers and CNV carriers in the UK Biobank (for 16p11.2
272 duplication p -value $< 10^{-4}$ using an identical test to that above, Fig. 4a). Notably, we obtained
273 similar results for all other seven intermediate phenotypes (Supp. Fig. 2). Carrying over CNV-
274 associated MRI profiles computed in the clinical cohort to the UK Biobank was a critical step that
275 allowed us to identify phenotype correlates of the CNV-associated MRI profiles in a population
276 >500 times larger than our median CNV cohort.

277

278 [Charting phenome-wide associations of CNV signatures](#)

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

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

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

317

318 [Detailing shared and distinct phenotypic associations](#)

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

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

340 showed a high relative number of associations in the physical measures category (Fig. 6c). As
341 such, next to the substantial phenotypic similarity, CNVs also displayed some unique
342 characteristic phenotypic associations relative to other CNVs.
343

344 Quantifying the path toward converging phenotypical profile

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

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

362 Discussion

363 CNVs offer a unique window of opportunity into the consequences of localized genetic
364 variation on human traits. This is especially the case, given their known genetic architecture and
365 typically high penetrance. In the present study, we built computational bridges between eight
366 key CNVs in a multisite clinical dataset, on the one hand, and their deep phenotypic profiling in
367 39,085 subjects from the wider population, on the other hand. To this end, we designed an
368 analytic framework that can quantitatively dissect the impact of distinct genetic mutations on
369 brain organization and behavioral differentiation. Bringing over derived CNV-specific
370 intermediate phenotypes to the population cohort revealed that the CNVs are tied to pleiotropic
371 associations beyond physical and cognitive domains. This phenome-wide analysis across ~ 1000
372 phenotypes revealed many ramifications for several body systems. Our collective analyses also
373 reveal wide-ranging similarities between the PheWAS profiles of the eight CNVs. Therefore, the
374 phenotypic level appears to be the point of alignment for distinct long-segment genetic variants
375 that we show to cause diverging morphological changes in brain morphology. Such late
376 convergence in phenotypic consequences speaks to profound basic science questions regarding
377 the organization of genetic influences on human brain and behavior.
378

379 For a long time, inquiries targeting genetic influences have been limited by the lack of
380 longitudinal and deep multimodal measures of brain and behavior in large subject samples²⁴.
381 Studies aimed at elucidating genotype-phenotype links were challenged by several obstacles,
382 including ascertainment bias, limited statistical power, and patchy phenotypic coverage²². We
383 are unlikely to have access to large enough clinical datasets soon – a condition *sine qua non* for
384 definitive tests of phenotypic overlaps and differences between genetic variants. As a concrete
385 example, Marek and colleagues (2022) highlighted the need for thousands of participants to

386 obtain reproducible and reliable brain-wide associations. Therefore, to overcome several of
387 these hurdles, we here put forward solutions that take advantage of intermediate CNV
388 phenotypes, a term coined in research on psychiatric disorders²⁹. These refer to biological traits
389 that lie in between an individual's external phenotype and innate genetic blueprint^{30,31}. We
390 captured CNV-specific intermediate phenotype representations as "genetics-first" whole-brain
391 signatures derived from our clinical boutique dataset. These signatures recapitulated previous
392 findings on morphology alterations, such as the predominant decrease in regional volumes for
393 deletions of 1q21.2 or 22q11.2, as well as the increase for 16p11.2 deletion^{23,32,33}. We also
394 observed reported mirror dose responses, especially strong in 22q11.2 locus²². Therefore, the
395 validity of LDA-derived intermediate phenotypes is corroborated by recapitulating key findings
396 from clinical studies.

397 The eight analyzed CNVs are known to differ in the ensuing effects on brain
398 architecture^{11,12}. The magnitude of their effects has previously been associated with the number
399 of affected genes and clinical outcomes. In concordance, we found 16p11.2 deletion to affect
400 the largest number of regions. This CNV contains 29 genes and is associated with an almost 40-
401 fold increase in the odds of autism spectrum disorder²⁷. Conversely, we found 15q11.2
402 duplication, which contains only four genes and is not formally associated with any disease, to
403 affect the fewest number of regions. In addition, we provide a fresh look into the diverging CNV
404 effects on brain morphology by summarizing the effects with respect to seven large-scale
405 Schaefer-Yeo networks. The network effects revealed a degree of similarity to functional
406 connectivity alterations in CNV carriers²¹. We also observed effects in the default mode and
407 limbic network for 22q11.2 deletion, as well as for ventral attention and motor network for
408 16p11.2 deletion. Together, the structural and functional alterations showed significant overlap
409 with alterations of idiopathic autism spectrum disorder and schizophrenia³⁴. The resemblance
410 suggests that the risk conferred by genetic variants, structural alterations, and the associated
411 functional connectivity patterns represent important dimensions that are coupled with diseases.
412

413 In the present work, we demonstrate the added value of how intermediate phenotypes
414 can be transferred for direct usage in other cohorts, including large-scale populational datasets.
415 By transferring these brain-wide representations over to the UK Biobank and carrying out
416 PheWAS, we obtained systemic phenotypic associations across eleven rich phenotypic
417 categories that go beyond mere cognitive domains. The reported PheWAS associations of the
418 intermediate CNV phenotypes were concordant with previous studies investigating more
419 circumscribed links between CNV status and indicators of cognitive performance, including fluid
420 intelligence score¹⁷, physical measurements like weight or height^{3,15}, common medical
421 conditions like hypertension or obesity, and blood biomarkers like indicators of cholesterol fat
422 metabolism pathways³⁵. As one of many examples, we demonstrated how intermediate
423 phenotypes tied to 22q11.2 deletion relate to an array of phenotypes in blood assays as well as
424 cardiac and blood vessels categories. It is important to stress that PheWAS only charts
425 associations between imaging and non-imaging measures to generate testable hypotheses
426 without providing causal links³⁶. Even though there may not be a causative link between a brain
427 phenotype and cardiac biomarkers, the thus revealed association suggests a hidden causal effect
428 of the CNV on both traits (e.g., brain morphology and artery wall thickening³⁷).

429 Similar to Auwerx and colleagues (2022), six of our eight examined CNVs were
430 associated with body weight, insulin-like growth factor 1, alkaline phosphatase, or mean red
431 blood cell volume. Therefore, these bodily alterations may not be mere secondary effects³⁸.
432 Instead, systemic manifestations could be a fundamental aspect of the primary biology of CNVs
433 and brain disorders in general. Critically, they might also lead to a reduced life span, as suggested

434 by the 63% probability of survival to age 50 in adult carriers of 22q11.2 deletion^{14,39}. Similar to
435 22q11.2 deletion, psychotic disorders have been linked with 15–20 years shorter life
436 expectancy⁴⁰. Most of this premature mortality is predominantly due to elevated cardiovascular
437 risk factors^{41,42} – causes that belong to the phenotype category among the most consistent
438 associations in our phenome-wide assays. Detected associations speak in favor of CNVs as a
439 complex disorder with several manifestations outside the brain that have considerable
440 deleterious impacts on various parts of everyday lives.

441

442 By combining hand-crafted analytic solutions with recently emerged data resources, our
443 computational assays lay out pleiotropic associations in CNV carriers. These consequences
444 include systemic associations outside the central nervous system. This underappreciated insight
445 is reflected in our results, including strong brain-behavior associations of the CNV profile in the
446 UK Biobank population with blood pressure, cholesterol, and weight. Since CNVs do not show
447 complete penetrance in all cases⁴⁴, such associations portray a necessary picture of a broad
448 spectrum of outcomes later in life. Hence, the constellation of results advocates rebalancing the
449 medical care of CNV carriers towards more comprehensive medical monitoring in a broader
450 patient pool⁴⁵.

451

452 In a similar way, previous clinical research has provided evidence that schizophrenia and
453 related psychotic disorders often affect multiple body systems (e.g., nervous, immune, or
454 endocrine), even from illness onset^{46,47}. Pillinger and colleagues (2019) reported robust
455 alterations in immune and cardiometabolic systems of a comparable magnitude to alterations
456 in the central nervous system. Further examples of major brain disorders accompanied by
457 problems outside the brain include gastrointestinal disorders in autism⁴⁸, loss of bone density in
458 depression⁴⁹, or cardiovascular symptoms in bipolar disorder⁵⁰. Finally, a recent study showed
459 that genetic liabilities for five major psychiatric disorders are associated with long-term
460 outcomes in adult life, including sociodemographic factors and physical health⁵¹. Our findings
461 thus add pieces of knowledge that illuminate how the nervous system is interlocked with the
462 rest of the body in a way that affects general well-being.

463

464 More broadly, understanding pathophysiological disease mechanisms will be propelled
465 by further disentangling the perplexing link between genes, brain, and behavior⁵². There is an
466 active debate on the extent to which distinct gene dosage disorders can lead to different non-
467 overlapping phenotypical profiles²⁴. This discourse was sparked from the observations that
468 many SNPs and CNVs increase the risk for schizophrenia or autism^{11,53}. Polygenicity and
469 pleiotropy, key features of the genetics underpinning psychiatric disorders^{13,54}, imply that
470 genetic mutations can converge on shared mechanisms at some level of pathway cascades, from
471 genes to large-scale brain networks to the phenome. Here, we report a low similarity of
472 intermediate phenotypes representing morphological CNV-specific brain signatures, in line with
473 a documented broad diversity of regional morphometry patterns across genomic loci^{22,55,56}.
474 Conversely, the ramifications of carrying distinct CNV variants for cognition and behavior have
475 previously been hypothesized to be more similar than those on brain anatomy^{9,24}. We here find
476 evidence for substantial convergence of phenotypic measures across CNVs quantified by
477 increased phenotypical similarity. Specifically, we observed a high degree of similarity between
478 the phenotypical profiles (mean similarity $r = 0.46$ as measured by Pearson's correlation across
479 the CNV's corresponding PheWAS profiles), which largely exceeded the similarity of brain
480 morphometry profiles (mean similarity $r = 0.2$ as measured by the correlation of volumetric
481 Cohen's d maps). Based on the presented strong resemblance of phenotypic profiles of the
482 examined eight CNVs, we speculate that the polygenic architecture of human phenotypic traits

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

484

485 This study has several limitations. One is that we did not investigate the effects of
486 medication on derived CNV-specific brain signatures since medication information was not
487 available for the whole clinical dataset. Nevertheless, previous studies have reported no
488 significant effects of psychiatric comorbidities (e.g. psychosis, ASD, ADHD, anxiety and mood
489 disorder) and psychotropic medication on neuroimaging patterns^{34,57}. We also did not study
490 causal relationships between brain patterns and non-imaging indicators. Making causal
491 inference requires proposing and defending a plausible causal structure by spelling out the
492 assumed (directional) dependencies among the outcome, input variables, and relevant
493 confounding variables⁵⁸. Future studies can start off from hypotheses generated by our catalog
494 of PheWAS links to find causal links between variables, for example, using structural equation
495 modeling. Finally, given our data scenario, we resorted to linear models in combination with
496 bagging and shrinkage to safeguard from overfitting.

497

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

512

513 Methods

514 Multisite clinical cohort

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

521 Our clinical dataset consisted of volumetric measurements derived from magnetic
522 resonance imaging (MRI) brain scans of 860 subjects: 548 CNV carriers and 312 controls not
523 carrying any CNV (Table 1). The here examined CNVs are among the most commonly studied
524 CNVs⁵⁹. Deletions and duplications of 1q21.1, 15q11.2, 16p11.2, and 22q11.2 represent some of
525 the most frequent risk factors for neuropsychiatric disorders identified in pediatric clinics^{19,20}.
526 That is why the target CNV loci were also selected by The Enhancing NeuroImaging Genetics
527 through Meta-Analysis copy number variant (ENIGMA-CNV) in a study on their cognitive,
528 psychiatric, and behavioral manifestations²³. These deletions and duplications strike a balance
529 between occurrence in the population and their effect size. In other words, the selected CNVs
530 are frequent enough so that we can start studying large enough sample sizes that allow for
531 across-CNV comparison in the first place. At the same time, this class of CNVs has been shown
532 to detrimentally affect cognition and raise the risk for psychiatric conditions^{23,25,33}. Our CNV
533 carriers did not carry any other large CNV.

534 An extensive description of methods and analyses is available in an already published
535 study with an identical dataset⁶⁰. In short, PennCNV and QuantiSNP were used, with standard
536 quality control metrics, to identify CNVs. CNV carriers were selected based on the following
537 breakpoints according to the reference genome GRCh37/hg19: 16p11.2 proximal (BP4-5, 29.6-
538 30.2MB), 1q21.1 distal (Class I, 146.4-147.5MB & II, 145.3-147.5MB), 22q11.2 proximal (BPA-D,
539 18.8-21.7MB) and 15q11.2 (BP1-2, 22.8–23.0MB). Control individuals did not carry any CNV at
540 these loci. The CNV carriers were either probands referred to the genetic clinic for the
541 investigation of neurodevelopmental and psychiatric disorders or their relatives (parents,
542 siblings, and other relatives).

543 UK Biobank might represent the largest dataset of carriers affected by 15q11.2 deletions
544 and duplications. Therefore, after identifying 15q11.2 deletions and duplications in the UK
545 Biobank, we added the respective carriers to our clinical cohort. In other words, we excluded
546 these subjects from the UK Biobank and treated them as part of our clinical dataset. Sensitivity
547 analysis concluded that including this CNV locus does not change our main findings (Supp. Fig.
548 12). Controls were either non-carriers within the same families or individuals from the general
549 population. Furthermore, controls were carefully matched for sex and age to CNV carriers.
550

551 Clinical MRI data recording and processing

552 We analyzed a data sample of T1-weighted (T1w) images at 0.8–1 mm isotropic
553 resolution. All T1w included in the analysis were quality checked by a domain expert⁶⁰. Data for
554 Voxel-Based Morphometry were preprocessed and analyzed with SPM12
555 (<http://www.fil.ion.ucl.ac.uk/spm/software/spm12/>)⁶¹⁻⁶³ running under MATLAB R2018b
556 (https://www.mathworks.com/products/new_products/release2018b.html). Further quality
557 control was performed using standardized ENIGMA quality control procedures
558 (<http://enigma.ini.usc.edu/protocols/imaging-protocols/>). Finally, neurobiologically
559 interpretable measures of gray matter volume were extracted in all participants by summarizing

560 whole-brain MRI maps in the MNI reference space. This feature-generation step was guided by
561 the topographical brain region definitions of the commonly used Schaefer-Yeo atlas with 400
562 parcels⁶⁴. The derived quantities of local gray matter volumetry resulted in 400 volume measures
563 for each participant. As a data-cleaning step, derived regional brain volumes were adjusted for
564 intracranial volume, age, age², and sex as fixed effects and scanning site as a random factor,
565 following previous research on this dataset⁶⁰. In particular, we have previously demonstrated
566 that CNVs show independent effects on regional and total brain volumes³³. Our current
567 investigation is focused on how CNVs induce regional brain effects. Note that ancillary analyses
568 revealed additional adjustments for total gray matter volume not to have any appreciable effect
569 on subsequent analyses.

570

571 Population data source

572 The UK Biobank is the largest biomedical resource that offers extensive behavioral and
573 demographic assessments, medical and cognitive measures, as well as biological samples in a
574 cohort of ~500,000 participants recruited from across Great Britain
575 (<https://www.ukbiobank.ac.uk/>). The present study was based on the recent brain-imaging data
576 release from February/March 2020. Our data sample included measurements from 39,085
577 participants with brain-imaging measures and expert-curated image-derived phenotypes of gray
578 matter morphology (T1-weighted MRI) (Table 2). Among the participants, 48% were men and
579 were 52% women with age between 40 and 69 y.o. when recruited [mean age 55 y.o., standard
580 deviation (SD) 7.5 y.]. We benefited from the uniform data preprocessing pipelines designed
581 and implemented by the FMRIB, Oxford University, Oxford, UK⁶⁵, to improve comparability and
582 reproducibility.

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

590 The structural MRI data were acquired as high-resolution T1-weighted images of brain
591 anatomy using a 3D MPRAGE sequence at 1mm isotropic resolution. It was preprocessing
592 included gradient distortion correction, the field of view reduction using the Brain Extraction
593 Tool⁶⁶ and FLIRT⁶⁷, as well as non-linear registration to MNI152 standard space at 1 mm
594 resolution using FNIRT⁶⁸. To avoid unnecessary interpolation, all image transformations were
595 estimated, combined, and applied by a single interpolation step. Tissue-type segmentation into
596 the cerebrospinal fluid, gray matter and white matter to generate full bias-field-corrected
597 images was achieved using FAST (FMRIB's Automated Segmentation Tool⁶⁹). Finally, gray matter
598 images were used to extract gray matter volumes in parcels according to the Schaefer-Yeo atlas
599 with 400 regions⁶⁴. Following previous work on the UKBB^{70,71}, inter-individual variations in brain
600 region volumes that could be explained by nuisance variables of no interest were adjusted for
601 by regressing out: body mass index, head size, head motion during task-related brain scans, head
602 motion during resting-state fMRI scanning, head position and receiver coil in the scanner (x, y,
603 and z), position of scanner table, as well as the data acquisition site.

604

605 Statistical analysis for volumetric brain measures

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

$$610 \quad d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}$$

611 where \bar{x}_1 corresponds to the mean region volume across CNV carriers, \bar{x}_2 corresponds to the
612 mean region volume across controls. Similarly, s_1 and s_2 correspond to standard deviations of
613 CNV carriers and controls.

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

616 We compared Cohen's d volumetric brain maps (and intermediate phenotypes brain
617 maps) between different CNVs using Pearson's correlation. Furthermore, we used spin
618 permutation testing to calculate empirical p-values for the ensuing correlation coefficient⁷².

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

$$622 \quad \rho_c = \frac{2s_{1,2}}{s_1^2 + s_2^2 + (\bar{x}_1 - \bar{x}_2)^2}$$

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

624

625 Charting complex association using phenom-wide association study

626 We performed a rich annotation of the derived intermediate phenotypes by means of a
627 phenome-wide association analysis benefitting from a wide variety of almost 1,000 lifestyle
628 factors. For a detailed description of phenotype extraction and analysis, refer to our previously
629 published studies⁷³. Feature extraction was carried out using two utilities designed to obtain,
630 clean, and normalize UKBB phenotype data according to predefined rules. In short, we collected
631 a raw set of ~15,000 phenotypes that we further processed by the FMRIB UKB Normalisation,
632 Parsing And Cleaning Kit (FUNPACK version 2.5.0;
633 <https://zenodo.org/record/4762700#.YQrpuj2caJ8>). FUNPACK is designed to perform automatic
634 refinement on the UKB data, which includes removing 'do not know' responses and filling the
635 blank left by unanswered sub-questions. The FUNPACK-derived phenotype information covered
636 11 major categories, including cognitive and physiological assessments, physical and mental
637 health records, blood assays, as well as sociodemographic and lifestyle factors. The output
638 consisted of a collection of 3,330 curated phenotypes which were then fed into PHENome Scan
639 ANalysis Tool (PHESANT⁷⁴, <https://github.com/MRCIEU/PHESANT>) for further refinement in an
640 automated fashion. PHESANT performs further data cleaning and normalization along with
641 labeling data as one of four data types: categorical ordered, categorical unordered, binary, and
642 numerical. Categorical unordered variables were one-hot encoded, such that each possible
643 response was represented by a binary column (true or false). The final curated inventory
644 comprised 977 phenotypes spanning 11 FUNPACK-defined categories. Furthermore, we used
645 Pearson's correlation to quantify the association strength between these 977 phenotypes with
646 subject-specific expressions of our eight intermediate phenotypes (cf. below). To ensure that
647 the correlations are not driven by a few outlying intermediate phenotype expressions, we first
648 discarded 551 subjects based on Tukey's interquartile range rule for outlier detection.

649 [Multi-class prediction model and intermediate phenotype extraction](#)

650 Technically, our core aim was to derive robust CNV-specific representations of
651 intermediate phenotypes from a clinical sample that could be transferred to a large population
652 resource for deep profiling. We derived the intermediate phenotypes as systematic brain
653 morphometric co-deviations attributable to each of our eight target CNVs. To this end, we
654 capitalized on linear discriminant analysis to extract separating rules between CNV carriers and
655 controls based on whole-brain volume measurements. LDA can be viewed as a generative
656 approach to classifying CNV carriers, which requires fitting multivariate Gaussian distribution to
657 regional brain volumes and producing a linear decision boundary⁷⁵. In particular, LDA-derived
658 discriminant vectors/functions represented CNV-specific intermediate phenotypes. Using a
659 linear model represents a data-efficient and directly biologically interpretable approach to our
660 analysis, especially in our boutique datasets with limited subject samples⁷⁶. These datasets are
661 characteristic of the low sample size regularly encountered in biology and medicine, which
662 typically impedes the application of more complex non-linear models that require high numbers
663 of parameters to be estimated.

664 As another key model property of direct relevance to our present analysis goals, LDA can
665 also be viewed as a dimensionality technique because this modeling framework enables the
666 extraction of underlying coherent principles among our anatomical target regions that are most
667 informative in telling apart CNV carriers from controls. To do so, LDA has access to class labels
668 (CNV status in our case) and thus belongs to supervised techniques⁷⁵. Specifically, LDA projects
669 the input subjects' set of brain morphology measurements into a linear subspace, consisting of
670 the directions which maximally separate our classes⁷⁷. This dimensionality reduction quality of
671 LDA was a necessary prerequisite for extracting intermediate phenotypes from one dataset and
672 transferring them to other datasets.

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

693

694 Building and validating robust prediction models

695 To recapitulate, our goal was to derive eight CNV-specific intermediate phenotypes using
696 LDA. Therefore, we built separate CNV-specific LDA models designated to learn predictive
697 principles to tell apart between CNV carriers and controls. However, we faced the challenge of
698 the low number of CNV carriers. This challenge is inherent to various boutique datasets of rare
699 medical conditions. Consequently, our number of measured features (regional brain volumes)
700 was higher than the number of observation samples (subjects). Concretely, we disposed on
701 average of 67 subjects per CNV class (cf. Table 1), while each subject was described by 400
702 regional volumes. Such a high-dimensional data scenario can lead to overfitting⁷⁸, where the
703 model learns the detail and noise in the training samples and performs poorly in group
704 classification on unseen test samples⁷⁵. Hence, we used bootstrap aggregation (bagging), an
705 ensemble learning method that can be used to reduce overfitting⁷⁹. Bagging gains its value by
706 profiting from a wisdom-of-crowds strategy. Concretely, we used a set of trained LDA models to
707 obtain a more robust and better predictive performance than could be obtained from a single
708 trained LDA model in isolation⁷⁹. Such a model-averaging design improves classification
709 performance by reducing variance⁷⁵.

710 We performed bagging during the derivation of LDA models separately for all eight CNV
711 classes. Specifically, we used the following analytical strategy for a set of subjects consisting of
712 a single CNV type and controls. In the first phase, a randomly perturbed version of the dataset
713 is created by sampling the subject cohort with replacement. This bootstrap resampling served
714 as the “in-the-bag” set of samples (i.e., subjects). The number of “in-the-bag” CNV carriers and
715 controls equals their number in the dataset. Furthermore, the LDA model was trained on this
716 training “in-the-bag” dataset. Model performance was then evaluated on all subjects from the
717 dataset that were not selected for the “in-the-bag” dataset. These subject samples formed a
718 testing “out-of-bag” dataset. The performance (i.e., classification accuracy) was based on the
719 Mathews correlation coefficient, which has been reported to produce a more informative and
720 truthful score than accuracy and F1 score⁸⁰. The coefficient ranges between -1 and $+1$, where a
721 coefficient of $+1$ represents a perfect prediction, 0 random prediction, and -1 indicates total
722 disagreement between prediction and observation.

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

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

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

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

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

764

765 [Performing model inspection using feature importance](#)

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

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

780

781 [Carrying intermediate phenotype expressions over for deep characterization in 782 other data resources](#)

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

788 UK Biobank itself contains CNV carriers. Therefore, we aimed to validate the
789 transferability of intermediate phenotypes by testing the difference in intermediate phenotype
790 expression between CNV carriers and controls in both the clinical dataset and UK Biobank.
791 Specifically, we tested the null hypothesis of no difference in the mean expression of
792 intermediate phenotype in CNV carriers and controls. We adopted a two-sample bootstrap
793 hypothesis test for means difference with 1,000 bootstrap replicates⁸¹.

794 Data availability

795 The majority of 16p11.2 data are publicly available (<https://www.sfari.org/>). For the
796 22q11.2 sample, raw data are available upon request from the PI (CB). All derived measures
797 used in this study are available upon request (SJ). The rest of the CNV carriers' data cannot be
798 shared as participants did not provide consent. All data from UK Biobank are available to other
799 investigators online (ukbiobank.ac.uk). The Schaefer-Yeo atlas is accessible online
800 (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal).
801

802 Code availability

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

806 Acknowledgements

807 DB was supported by the Brain Canada Foundation, through the Canada Brain Research
808 Fund, with the financial support of Health Canada, National Institutes of Health (NIH R01
809 AG068563A, NIH R01 R01DA053301-01A1), the Canadian Institute of Health Research (CIHR
810 438531, CIHR 470425), the Healthy Brains Healthy Lives initiative (Canada First Research
811 Excellence fund), Google (Research Award, Teaching Award), and by the CIFAR Artificial
812 Intelligence Chairs program (Canada Institute for Advanced Research). This research was
813 supported by Calcul Quebec (<http://www.calculquebec.ca>) and Compute Canada
814 (<http://www.computecanada.ca>), the Brain Canada Multi-Investigator initiative, the Canadian
815 Institutes of Health Research, CIHR_400528, The Institute of Data Valorization (IVADO) through
816 the Canada First Research Excellence Fund, Healthy Brains for Healthy Lives through the Canada
817 First Research Excellence Fund. SJ is a recipient of a Canada Research Chair in
818 neurodevelopmental disorders, and a chair from the Jeanne et Jean Louis Levesque Foundation.
819 The Cardiff CNV cohort was supported by the Wellcome Trust Strategic Award "DEFINE" and the
820 National Centre for Mental Health with funds from Health and Care Research Wales (code
821 100202/Z/12/Z). The CHUV cohort was supported by the SNF (Maillard Anne, Project, PMPDP3
822 171331). Data from the UCLA cohort provided by CEB (participants with 22q11.2 deletions or
823 duplications and controls) was supported through grants from the NIH (U54EB020403), NIMH
824 (R01MH085953, R01MH100900, R03MH105808), and the Simons Foundation (SFARI Explorer
825 Award). KK was supported by The Institute of Data Valorization (IVADO) Postdoctoral Fellowship
826 program through the Canada First Research Excellence Fund. IES is supported by the Research
827 Council of Norway (#223273), South-Eastern Norway Regional Health Authority (#2020060),
828 European Union's Horizon2020 Research and Innovation Programme (CoMorMent project;
829 Grant #847776) and Kristian Gerhard Jebsen Stiftelsen (SKGJ-MED-021). We thank all of the
830 families participating at the Simons Searchlight sites and 16p11.2 European Consortium, Simons
831 Searchlight Consortium. We appreciate obtaining access to brain-imaging and phenotypic data

832 on SFARI Base. We are grateful to all families who participated in the 16p11.2 European
 833 Consortium. The funders had no role in study design, data collection and analysis, decision to
 834 publish or preparation of the manuscript.
 835

836 Author Contributions Statement

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

845 Competing Interests Statement

846 The authors declare no competing interests.

847 Tables

848 **Table 1.**

849 **Clinical dataset demographics.**

850 CNV loci chromosome coordinates are provided with the number of genes encompassed in each
 851 CNV and with a well-known gene for each locus to help recognize the CNV. Other diagnoses
 852 included: language disorder, major depressive disorder, posttraumatic stress disorder,
 853 unspecified disruptive and impulse-control and conduct disorder, social anxiety disorder, social
 854 phobia disorder, speech sound disorder, moderate intellectual disability, specific learning
 855 disorder, gambling disorder, bipolar disorder, conduct disorder, attention-deficit/hyperactivity
 856 disorder, substance abuse disorder, global developmental delay, motor disorder, obsessive-
 857 compulsive disorder, sleep disorder, Tourette’s disorder, mood disorder, eating disorders,
 858 transient tic disorder, trichotillomania, pervasive developmental disorder, specific phobia, body
 859 dysmorphic disorder, mathematics disorder, and dysthymic disorder. Abbreviations, Del:
 860 deletion; Dup: duplication; ASD: autism spectrum disorder; SZ: schizophrenia; chr: chromosome;
 861 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

863
864
865
866
867
868
869
870
871
872
873
874
875
876

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 population⁴³ (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. ¹ICD10 code, including diagnoses of schizophrenia, schizotypal and delusional disorders (F20-F29). ²ICD10 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

877

Figure legends

878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898

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, age², 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 (LD₁) mainly captures differences between 16p11.2 proximal deletion and duplication, while the second LDA dimension (LD₂) 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

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

901

902

Figure 2

903

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

904

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

917

918

Figure 3

919

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

920

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

933

934

Figure 4

935

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

936

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

955

956

Figure 5

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

969

970 **Figure 6**

971 **Detailing aspects convergence in phenome-wide portfolios across different CNVs.**

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

992

993 **References**

- 994 1. Freeman, J. L. *et al.* Copy number variation: New insights in genome diversity. *Genome*
995 *Res.* **16**, 949–961 (2006).
- 996 2. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human
997 genome. *Nature* **464**, 704–712 (2010).
- 998 3. Auwerx, C. *et al.* The individual and global impact of copy-number variants on complex
999 human traits. *Am. J. Hum. Genet.* S0002-9297(22)00061–1 (2022)
1000 doi:10.1016/j.ajhg.2022.02.010.
- 1001 4. Jacquemont, S. *et al.* Genes To Mental Health (G2MH): A framework to map the combined
1002 effects of rare and common variants on dimensions of cognition and psychopathology.
1003 *Am. J. Psychiatry* (2021).
- 1004 5. Rutkowski, T. P. *et al.* Unraveling the genetic architecture of copy number variants
1005 associated with schizophrenia and other neuropsychiatric disorders. *J. Neurosci. Res.* **95**,
1006 1144–1160 (2017).
- 1007 6. Lauer, S. & Gresham, D. An evolving view of copy number variants. *Curr. Genet.* **65**, 1287–
1008 1295 (2019).
- 1009 7. Huguet, G. *et al.* Genome-wide analysis of gene dosage in 24,092 individuals estimates
1010 that 10,000 genes modulate cognitive ability. *Mol. Psychiatry* **26**, 2663–2676 (2021).
- 1011 8. Moberg, P. J. *et al.* Neurocognitive Functioning in Patients with 22q11.2 Deletion
1012 Syndrome: A Meta-Analytic Review. *Behav. Genet.* **48**, 259–270 (2018).
- 1013 9. Silva, A. I. *et al.* Neuroimaging findings in neurodevelopmental copy number variants:
1014 identifying molecular pathways to convergent phenotypes. *Biol. Psychiatry* (2022)
1015 doi:10.1016/j.biopsych.2022.03.018.

- 1016 10. Chawner, S. J. R. A. *et al.* Genotype-phenotype associations in children with copy number
1017 variants associated with high neuropsychiatric risk in the UK (IMAGINE-ID): a case-control
1018 cohort study. *Lancet Psychiatry* **6**, 493–505 (2019).
- 1019 11. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a
1020 genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).
- 1021 12. Sanders, S. J. *et al.* A framework for the investigation of rare genetic disorders in
1022 neuropsychiatry. *Nat. Med.* **25**, 1477–1487 (2019).
- 1023 13. Moreno-De-Luca, D. & Martin, C. L. All for one and one for all: heterogeneity of genetic
1024 etiologies in neurodevelopmental psychiatric disorders. *Curr. Opin. Genet. Dev.* **68**, 71–78
1025 (2021).
- 1026 14. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK
1027 Biobank. *J. Med. Genet.* **56**, 131–138 (2019).
- 1028 15. Owen, D. *et al.* Effects of pathogenic CNVs on physical traits in participants of the UK
1029 Biobank. *BMC Genomics* **19**, 867 (2018).
- 1030 16. Adams, R. L. *et al.* Psychopathology in adults with copy number variants. *Psychol. Med.* 1–8
1031 (2022) doi:10.1017/S0033291721005201.
- 1032 17. Kendall, K. M. *et al.* Cognitive performance and functional outcomes of carriers of
1033 pathogenic copy number variants: analysis of the UK Biobank. *Br. J. Psychiatry J. Ment. Sci.*
1034 **214**, 297–304 (2019).
- 1035 18. Doelken, S. C. *et al.* Phenotypic overlap in the contribution of individual genes to CNV
1036 pathogenicity revealed by cross-species computational analysis of single-gene mutations in
1037 humans, mice and zebrafish. *Dis. Model. Mech.* **6**, 358–372 (2013).
- 1038 19. Viñas-Jornet, M. *et al.* High Incidence of Copy Number Variants in Adults with Intellectual
1039 Disability and Co-morbid Psychiatric Disorders. *Behav. Genet.* **48**, 323–336 (2018).

- 1040 20. Moreau, C. A. *et al.* Dissecting autism and schizophrenia through neuroimaging genomics.
1041 *Brain* **144**, 1943–1957 (2021).
- 1042 21. Moreau, C. A. *et al.* Mutations associated with neuropsychiatric conditions delineate
1043 functional brain connectivity dimensions contributing to autism and schizophrenia. *Nat.*
1044 *Commun.* **11**, 5272 (2020).
- 1045 22. Modenato, C. *et al.* Lessons Learned From Neuroimaging Studies of Copy Number Variants:
1046 A Systematic Review. *Biol. Psychiatry* **90**, 596–610 (2021).
- 1047 23. Sønderby, I. E. *et al.* Effects of copy number variations on brain structure and risk for
1048 psychiatric illness: Large-scale studies from the ENIGMA working groups on CNVs. *Hum.*
1049 *Brain Mapp.* **43**, 300–328 (2022).
- 1050 24. Raznahan, A., Won, H., Glahn, D. C. & Jacquemont, S. Convergence and divergence of rare
1051 genetic disorders on brain phenotypes. *JAMA Psychiatry* (2022).
- 1052 25. Zinkstok, J. R. *et al.* Neurobiological perspective of 22q11.2 deletion syndrome. *Lancet*
1053 *Psychiatry* **6**, 951–960 (2019).
- 1054 26. Littlejohns, T. J. *et al.* The UK Biobank imaging enhancement of 100,000 participants:
1055 rationale, data collection, management and future directions. *Nat. Commun.* **11**, 2624
1056 (2020).
- 1057 27. Weiss, L. A. *et al.* Association between Microdeletion and Microduplication at 16p11.2 and
1058 Autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
- 1059 28. Marek, S. *et al.* Reproducible brain-wide association studies require thousands of
1060 individuals. *Nature* **603**, 654–660 (2022).
- 1061 29. Gottesman, I. I. & Gould, T. D. The endophenotype concept in psychiatry: etymology and
1062 strategic intentions. *Am. J. Psychiatry* **160**, 636–645 (2003).

- 1063 30. Mark, W. & Toulopoulou, T. Cognitive intermediate phenotype and genetic risk for
1064 psychosis. *Curr. Opin. Neurobiol.* **36**, 23–30 (2016).
- 1065 31. Meyer-Lindenberg, A. & Weinberger, D. R. Intermediate phenotypes and genetic
1066 mechanisms of psychiatric disorders. *Nat. Rev. Neurosci.* **7**, 818–827 (2006).
- 1067 32. Lin, A. *et al.* Mapping 22q11.2 Gene Dosage Effects on Brain Morphometry. *J. Neurosci.* **37**,
1068 6183–6199 (2017).
- 1069 33. Martin-Brevet, S. *et al.* Quantifying the Effects of 16p11.2 Copy Number Variants on Brain
1070 Structure: A Multisite Genetic-First Study. *Biol. Psychiatry* **84**, 253–264 (2018).
- 1071 34. Sun, D. *et al.* Large-scale mapping of cortical alterations in 22q11.2 deletion syndrome:
1072 Convergence with idiopathic psychosis and effects of deletion size. *Mol. Psychiatry* **25**,
1073 1822–1834 (2020).
- 1074 35. Bracher-Smith, M. *et al.* Effects of pathogenic CNVs on biochemical markers: a study on
1075 the UK Biobank. *bioRxiv* 723270 (2019) doi:10.1101/723270.
- 1076 36. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective
1077 epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).
- 1078 37. Momma, K. Cardiovascular Anomalies Associated With Chromosome 22q11.2 Deletion
1079 Syndrome. *Am. J. Cardiol.* **105**, 1617–1624 (2010).
- 1080 38. Pillinger, T., D’Ambrosio, E., McCutcheon, R. & Howes, O. D. Is psychosis a multisystem
1081 disorder? A meta-review of central nervous system, immune, cardiometabolic, and
1082 endocrine alterations in first-episode psychosis and perspective on potential models. *Mol.*
1083 *Psychiatry* **24**, 776–794 (2019).
- 1084 39. Van, L. *et al.* All-cause mortality and survival in adults with 22q11.2 deletion syndrome.
1085 *Genet. Med.* **21**, 2328–2335 (2019).

- 1086 40. Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of mental disorders
1087 in the World Health Organization’s World Mental Health Survey Initiative. *World*
1088 *Psychiatry Off. J. World Psychiatr. Assoc. WPA* **6**, 168–176 (2007).
- 1089 41. Correll, C. U. *et al.* Prevalence, incidence and mortality from cardiovascular disease in
1090 patients with pooled and specific severe mental illness: a large-scale meta-analysis of
1091 3,211,768 patients and 113,383,368 controls. *World Psychiatry Off. J. World Psychiatr.*
1092 *Assoc. WPA* **16**, 163–180 (2017).
- 1093 42. Hoang, U., Goldacre, M. J. & Stewart, R. Avoidable mortality in people with schizophrenia
1094 or bipolar disorder in England. *Acta Psychiatr. Scand.* **127**, 195–201 (2013).
- 1095 43. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK
1096 Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–
1097 1034 (2017).
- 1098 44. Kirov, G. *et al.* The penetrance of copy number variations for schizophrenia and
1099 developmental delay. *Biol. Psychiatry* **75**, 378–385 (2014).
- 1100 45. Chawner, S. J., Watson, C. J. & Owen, M. J. Clinical evaluation of patients with a
1101 neuropsychiatric risk copy number variant. *Curr. Opin. Genet. Dev.* **68**, 26–34 (2021).
- 1102 46. Alessi, M. G. & Bennett, J. M. Mental health is the health of the whole body: How
1103 psychoneuroimmunology & health psychology can inform & improve treatment. *J. Eval.*
1104 *Clin. Pract.* **26**, 1539–1547 (2020).
- 1105 47. Guest, P. C. Psychiatric Disorders as “Whole Body” Diseases. in *Biomarkers and Mental*
1106 *Illness: It’s Not All in the Mind* (ed. Guest, P. C.) 3–16 (Springer International Publishing,
1107 2017). doi:10.1007/978-3-319-46088-8_1.
- 1108 48. Kohane, I. S. *et al.* The Co-Morbidity Burden of Children and Young Adults with Autism
1109 Spectrum Disorders. *PLOS ONE* **7**, e33224 (2012).

- 1110 49. Sotelo, J. L. & Nemeroff, C. B. Depression as a systemic disease. *Pers. Med. Psychiatry* **1–2**,
1111 11–25 (2017).
- 1112 50. Leboyer, M. *et al.* Can bipolar disorder be viewed as a multi-system inflammatory disease?
1113 *J. Affect. Disord.* **141**, 1–10 (2012).
- 1114 51. Leppert, B. *et al.* A cross-disorder PRS-pheWAS of 5 major psychiatric disorders in UK
1115 Biobank. *PLOS Genet.* **16**, e1008185 (2020).
- 1116 52. Thompson, P. M. *et al.* Genetic influences on brain structure. *Nat. Neurosci.* **4**, 1253–1258
1117 (2001).
- 1118 53. Bacchelli, E. *et al.* An integrated analysis of rare CNV and exome variation in Autism
1119 Spectrum Disorder using the Infinium PsychArray. *Sci. Rep.* **10**, 3198 (2020).
- 1120 54. Gratten, J. & Visscher, P. M. Genetic pleiotropy in complex traits and diseases: implications
1121 for genomic medicine. *Genome Med.* **8**, 78 (2016).
- 1122 55. Seidlitz, J. *et al.* Transcriptomic and cellular decoding of regional brain vulnerability to
1123 neurogenetic disorders. *Nat. Commun.* **11**, 3358 (2020).
- 1124 56. Moreau, C. A., Ching, C. R., Kumar, K., Jacquemont, S. & Bearden, C. E. Structural and
1125 functional brain alterations revealed by neuroimaging in CNV carriers. *Curr. Opin. Genet.*
1126 *Dev.* **68**, 88–98 (2021).
- 1127 57. Rogdaki, M. *et al.* Magnitude and heterogeneity of brain structural abnormalities in
1128 22q11.2 deletion syndrome: a meta-analysis. *Mol. Psychiatry* **25**, 1704–1717 (2020).
- 1129 58. Wysocki, A. C., Lawson, K. M. & Rhemtulla, M. Statistical Control Requires Causal
1130 Justification. *Adv. Methods Pract. Psychol. Sci.* **5**, 25152459221095824 (2022).
- 1131 59. Modenato, C. *et al.* Effects of eight neuropsychiatric copy number variants on human brain
1132 structure. *Transl. Psychiatry* **11**, 1–10 (2021).

- 1133 60. Modenato, C. *et al.* Effects of eight neuropsychiatric copy number variants on human brain
1134 structure. *Transl. Psychiatry* **11**, 399 (2021).
- 1135 61. Ashburner, J. A fast diffeomorphic image registration algorithm. *NeuroImage* **38**, 95–113
1136 (2007).
- 1137 62. Ashburner, J. & Friston, K. J. Unified segmentation. *NeuroImage* **26**, 839–851 (2005).
- 1138 63. Lorio, S. *et al.* New tissue priors for improved automated classification of subcortical brain
1139 structures on MRI. *NeuroImage* **130**, 157–166 (2016).
- 1140 64. Schaefer, A. *et al.* Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic
1141 Functional Connectivity MRI. *Cereb. Cortex N. Y. N 1991* **28**, 3095–3114 (2018).
- 1142 65. Alfaro-Almagro, F. *et al.* Image processing and Quality Control for the first 10,000 brain
1143 imaging datasets from UK Biobank. *NeuroImage* **166**, 400–424 (2018).
- 1144 66. Smith, S. M. Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**, 143–155
1145 (2002).
- 1146 67. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust
1147 and accurate linear registration and motion correction of brain images. *NeuroImage* **17**,
1148 825–841 (2002).
- 1149 68. Andersson, J. L., Jenkinson, M. & Smith, S. Non-linear registration, aka Spatial
1150 normalisation FMRIB technical report TR07JA2. *FMRIB Anal. Group Univ. Oxf.* **2**, e21
1151 (2007).
- 1152 69. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden
1153 Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med.*
1154 *Imaging* **20**, 45–57 (2001).
- 1155 70. Schurz, M. *et al.* Variability in Brain Structure and Function Reflects Lack of Peer Support.
1156 *Cereb. Cortex* **31**, 4612–4627 (2021).

- 1157 71. Spreng, R. N. *et al.* The default network of the human brain is associated with perceived
1158 social isolation. *Nat. Commun.* **11**, 6393 (2020).
- 1159 72. Alexander-Bloch, A., Giedd, J. N. & Bullmore, E. Imaging structural co-variance between
1160 human brain regions. *Nat. Rev. Neurosci.* **14**, 322–336 (2013).
- 1161 73. Savignac, C. *et al.* APOE ϵ 2 vs APOE ϵ 4 dosage shows sex-specific links to hippocampus-
1162 default network subregion co-variation. 2022.03.15.484482 Preprint at
1163 <https://doi.org/10.1101/2022.03.15.484482> (2022).
- 1164 74. Millard, L. A. C., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software
1165 Application Profile: PHESANT: a tool for performing automated phenome scans in UK
1166 Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).
- 1167 75. Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning: data mining,
1168 inference, and prediction. (2009).
- 1169 76. Bzdok, D. & Ioannidis, J. P. A. Exploration, Inference, and Prediction in Neuroscience and
1170 Biomedicine. *Trends Neurosci.* **42**, 251–262 (2019).
- 1171 77. Hart, P. E., Stork, D. G. & Duda, R. O. *Pattern classification*. (Wiley Hoboken, 2000).
- 1172 78. Bzdok, D., Nichols, T. E. & Smith, S. M. Towards Algorithmic Analytics for Large-scale
1173 Datasets. *Nat. Mach. Intell.* **1**, 296–306 (2019).
- 1174 79. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
- 1175 80. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC)
1176 over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
- 1177 81. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (CRC Press, 1994).
- 1178