

Copy-number variants as modulators of common disease susceptibility

Chiara Auwerx^{1,2,3,4,*}, Maarja Jõeloo^{5,6}, Marie C. Sadler^{2,3,4}, Nicolò Tesio¹, Sven Ojavee^{2,3},
Charlie J. Clark¹, Reedik Mägi⁶, Estonian Biobank Research Team^{6,§}, Alexandre
Reymond^{1,#,*} & Zoltán Kutalik^{2,3,4,#,*}

¹ Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

² Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

³ Swiss Institute of Bioinformatics, Lausanne, Switzerland

⁴ University Center for Primary Care and Public Health, Lausanne, Switzerland

⁵ Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

⁶ Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia

§ Estonian Biobank Research Team: Tõnu Esko, Andres Metspalu, Lili Milani, Reedik Mägi, Mari Nelis

These authors jointly supervised this work.

*** Correspondence:**

Chiara Auwerx: chiara.auwerx@unil.ch; Alexandre Reymond : alexandre.reymond@unil.ch; Zoltán
Kutalik: zoltan.kutalik@unil.ch.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

21 ABSTRACT

22 **Background:** Copy-number variations (CNVs) have been associated with rare and
23 debilitating genomic syndromes but their impact on health later in life in the general population
24 remains poorly described.

25 **Methods:** Assessing four modes of CNV action, we performed genome-wide association
26 scans (GWASs) between the copy-number of CNV-proxy probes and 60 curated ICD-10
27 based clinical diagnoses in 331,522 unrelated white UK Biobank participants with replication
28 in the Estonian Biobank.

29 **Results:** We identified 73 signals involving 40 diseases, all of which indicating that CNVs
30 increased disease risk and caused earlier onset. Even after correcting for these signals, a
31 higher CNV burden increased risk for 18 disorders, mainly through the number of deleted
32 genes, suggesting a polygenic CNV architecture. Number and identity of genes disturbed by
33 CNVs affected their pathogenicity, with many associations being supported by colocalization
34 with both common and rare single nucleotide variant association signals. Dissection of
35 association signals provided insights into the epidemiology of known gene-disease pairs (e.g.,
36 deletions in *BRCA1* and *LDLR* increased risk for ovarian cancer and ischemic heart disease,
37 respectively), clarified dosage mechanisms of action (e.g., both increased and decreased
38 dosage of 17q12 impacts renal health), and identified putative causal genes (e.g., *ABCC6* for
39 kidney stones). Characterization of the pleiotropic pathological consequences of recurrent
40 CNVs at 15q13, 16p13.11, 16p12.2, and 22q11.2 in adulthood indicated variable expressivity
41 of these regions and the involvement of multiple genes.

42 **Conclusions:** Our results shed light on the prominent role of CNVs in determining common
43 disease susceptibility within the general population and provide actionable insights allowing
44 to anticipate later-onset comorbidities in carriers of recurrent CNVs.

45

46 KEYWORDS

47 Structural variation; CNV; GWAS; common diseases; pleiotropy; genomic disorders.

48 BACKGROUND

49 Copy-number variants (CNVs) refer to duplicated or deleted DNA fragments ($\geq 50\text{bp}$) and
50 represent an important source of inter-individual variation [1,2]. As a highly diverse mutational
51 class, they can alter the copy-number of dosage sensitive genes, induce gain- or loss-of-
52 function (LoF) through gene fusion or truncation, unmask recessive alleles, or disrupt
53 regulatory sequences, thereby representing potent phenotypic modifiers [3]. As such, their
54 role in human disease has mainly been studied in clinically ascertained cohorts often
55 presenting with congenital anomalies and/or severe neurological (e.g., developmental delay
56 and intellectual disability, epilepsy) or psychiatric (e.g., autism or schizophrenia) symptoms
57 [4–7] and today, close to 100 genomic disorders (i.e., disease caused by genomic
58 rearrangements) have been described [8,9]. Despite their deleteriousness, some of these
59 CNVs flanked by repeats recurrently appear and remain at a low but stable frequency in the
60 population [10].

61

62 The emergence of large biobanks coupling genotype to phenotype data has fostered the study
63 of CNVs in the general population. Whole genome sequencing represents the best approach
64 to characterize the full human CNV landscape [1,11,12] but current long- and short-read
65 sequencing association studies have limited samples size [13–15]. Alternatively, larger
66 sample sizes are available for exome sequencing data, offering the possibility to assess the
67 phenotypic consequence of small CNVs [16,17], while microarray-based CNV calls are better-
68 suited for the study of large CNVs and have been successfully used in association studies
69 [8,18–28]. Performing a CNV genome-wide association study (GWAS) on 57 medically
70 relevant continuous traits in the UK Biobank (UKBB) [29], we previously identified 131
71 independent associations, including allelic series wherein carriers of CNVs at loci previously
72 associated with rare Mendelian disorders exhibited subtle changes in disease-associated
73 phenotypes but lacked the corresponding clinical diagnosis [23]. Paralleling findings for point
74 mutations [30–33], this supports a model of variable expressivity, where CNVs can cause a

75 wide spectrum of phenotypic alteration ranging from severe, early-onset diseases to mild
76 subclinical symptoms, opening the question as to whether these loci are also associated with
77 common diseases.

78

79 Unlike continuous traits that can be objectively measured in all participants, population
80 cohorts, such as UKBB, have low numbers of diseased individuals [34]. Moreover, defining
81 cases relies on the arbitrary dichotomization of complex underlying pathophysiological
82 processes [35]. Beyond the inherent loss of power associated to usage of binary variables
83 [36], cases might be missed because an individual did not consult a physician, was
84 misdiagnosed due to atypical clinical presentation, or is in a prodromal disease phase. Studies
85 investigating CNV-disease associations in the general population have either focused on only
86 few diseases [27,37] or well-established recurrent CNVs [20,38,39]. Alternatively, high-
87 throughput studies have assessed a broad range of continuous and binary traits
88 simultaneously [17,24,25] without any precautions to accommodate the aforementioned
89 challenges. To date, the largest disease CNV-GWAS meta-analyzed ~1,000,000 individuals
90 [8]. While boosting power through increased sample size, it comes at the cost of extensive
91 data harmonization, resulting in the exclusion of smaller CNVs ($\leq 100\text{kb}$) and broader disease
92 categories (e.g., “immune abnormality”). Moreover, as this study includes several clinical
93 cohorts, phenotypes are biased towards neuropsychiatric disorders (24/54 phenotypes) for
94 which the role of CNVs is well-established [4–7].

95

96 Using tailored CNV-GWAS models mimicking four mechanisms of CNV action and time-to-
97 event analysis, we investigate the relationship between CNVs and 60 carefully defined
98 common diseases affecting a broad range of physiological systems in 331,522 unrelated white
99 UKBB participants. Extensively validating our results, we report associations according to
100 confidence tiers and take advantage of rich individual-level phenotypic data to demonstrate
101 the contribution of CNVs to the common disease burden in the general population.

102 **METHODS**

103 **1. Study material**

104 *Discovery cohort: UK Biobank*

105 The UK Biobank (UKBB) is composed of ~500,000 volunteers (54% females) from the general
106 UK population for which microarray-based genotyping and extensive phenotyping data –
107 including hospital based International Classification of Diseases, 10th Revision (ICD-10)
108 codes (up to September 2021) and self-reported conditions – are available [29]. Participants
109 signed a broad informed consent form and data were accessed through application #16389.

110

111 *Replication cohort: Estonian Biobank*

112 The Estonian Biobank (EstBB) is a population-based cohort of ~208,000 Estonian individuals
113 (65% females; data freeze 2022v01 [12/04/2022]) for which microarray-based genotyping data
114 and ICD-10 codes from crosslinking with national and hospital databases (up to end 2021) are
115 available [40]. The activities of the EstBB are regulated by the Human Genes Research Act,
116 which was adopted in 2000 specifically for the operations of the EstBB. Individual level data
117 analysis in the EstBB was carried out under ethical approval 1.1-12/624 from the Estonian
118 Committee on Bioethics and Human Research (Estonian Ministry of Social Affairs), using data
119 according to release application 3-10/GI/34668 [20/12/2022] from the EstBB. All participants
120 signed a broad informed consent form.

121

122 *Other public resources*

123 The PheCode Map 1.2 (beta) (https://phewascatalog.org/phecodes_icd10) was used for ICD-
124 10 code classification [41]. Genomic regions were annotated with the NHGRI-EBI GWAS
125 Catalog (<https://www.ebi.ac.uk/gwas/>; 24/11/2022) [42], and the Online Mendelian Inheritance
126 in Man (OMIM; <https://www.omim.org/>; 27/07/2022) [43]. Recurrent CNV coordinates were
127 retrieved from DECIPHER (<https://www.deciphergenomics.org/>) [9]. Unless specified
128 otherwise, Neale UK Biobank single-nucleotide polymorphism (SNP)-GWASs summary

129 statistics were used (<http://www.nealelab.is/uk-biobank>). Allele frequencies and genomic
130 constraint scores (probability of LoF Intolerance (pLI); LoF Observed over Expected Upper
131 bound Fraction (LOEUF)) originate from the Genome Aggregation Database (GnomAD;
132 <https://gnomad.broadinstitute.org/>) [44]; pHaplo and pTriplo scores from [8]. Tissue-specific
133 gene expression was assessed in the Genotype-Tissue Expression project (GTEx;
134 <https://gtexportal.org/home/>) [45].

135

136 *Software versions*

137 CNVs were called with PennCNV v1.0.5 [46] using PennCNV-Affy (27/08/2009) and filtered
138 based on a quality scoring pipeline [47]. Genetic analyses were conducted with PLINK v1.9
139 and v2.0 [48]. ANNOVAR (24/10/2019) was used to map genes to genetic regions [49]. The
140 UCSC Genome Browser was used to determine the human genome size (GRCh37/hg19) and
141 the LiftOver tool was used to lift over genomic coordinates [50]. Statistical analyses were
142 performed with R v3.6.1 and graphs were generated with R v4.1.3.

143

144 **2. CNV association studies in the UK Biobank**

145 *Microarray-based CNV calling*

146 UKBB genotype microarray data were acquired from two arrays with 95% probe overlap
147 (Applied Biosystems UK Biobank Axiom Array: 438,427 samples; Applied Biosystems UK
148 BiLEVE Axiom Array by Affymetrix: 49,950 samples) [29] and used to call CNVs as previously
149 described [23]. Briefly, CNVs were called using standard PennCNV settings and samples on
150 genotyping plates with a mean CNV count per sample > 100 and samples with > 200 CNVs
151 or single CNV > 10 Mb were excluded. Remaining CNVs were attributed a probabilistic quality
152 score (QS) ranging from -1 (likely deletion) to 1 (likely duplication) [47]. High confidence CNVs,
153 stringently defined by $|QS| > 0.5$, were retained and encoded in chromosome-wide probe-by-
154 sample matrices (i.e., entries of 1, -1, or 0 indicate probes overlapping high confidence
155 duplication, deletion, or no/low quality CNV, respectively) [47], which were converted into three
156 PLINK binary file sets to accommodate association analysis according to four modes of CNV

157 action. Details about the CNV encoding and handling of chromosome X are provided in
158 [Supplemental Note 1](#). Probe-level CNV frequency was calculated [23]. All results in this study
159 are based on the human genome reference build GRCh37/hg19.

160

161 *Case-control definition and age-at-disease onset calculation*

162 A pool of 331,522 unrelated white British UKBB participants (54% females) was considered
163 after excluding retracted (up to August 2020), as well as related, high missingness and non-
164 white British samples (*used.in.pca.calculation* = 0 and *in.white.British.ancestry.subset* = 0 in
165 Sample-QC v2 file). CNV outliers (*Microarray-based CNV calling*) and individuals reporting
166 blood malignancies (i.e., possibly harboring somatic CNVs; UKBB field #20001: 10047, 1048,
167 1050, 1051, 1052, 1053, 1055, 1056, 1056; #41270: ICD-10 codes mapping to PheCode
168 exclusion range “*cancer of lymphatic and hematopoietic tissue*”), were further excluded.

169

170 Cases and controls were assigned for 60 ICD-10-based clinical diagnoses using *diagnosis –*
171 *ICD10* (#41270), *cancer code, self-reported* (#20001), and *non-cancer illness code, self-*
172 *reported* (#20002) to build exclusion and inclusion lists. For each disease, we first defined all
173 331,522 individuals as controls. We excluded individuals with a self-reported or hospital
174 diagnosis of a broad set of conditions that include the disease of interest, as well as related
175 disorders/ICD-10 codes (e.g., other cancers or radio-/chemotherapy for breast cancer; mood
176 or personality disorders for schizophrenia). We then re-introduce as cases individuals having
177 received a restricted range of ICD-10 codes matching our disease definition. For second level
178 ICD-10 codes, all subcodes are considered, otherwise only the specified ones. The disease
179 burden was calculated as the number of diagnoses (out of the 60 assessed) an individual has
180 received. For male- (prostate cancer) and female- (menstruation disorders, endometriosis,
181 breast cancer, ovarian cancer) specific diseases, downstream analyses were conducted
182 excluding individuals from the opposite sex.

183

184 Based on the *date at first in-patient diagnosis – ICD10 (#41280)* and the individual's *month*
185 *(#52)* and *year (#34) of birth* (birthday assumed on average to be the 15th), the age at diagnosis
186 was calculated by subtracting the earliest diagnosis date for codes on the inclusion list from
187 the birth date and converting it to years by dividing by 365.25 to account for leap years.

188

189 *Probe and covariate selection*

190 Relevant covariates and probes were pre-selected to fit tailored main CNV-GWAS models and
191 reduce computation time. For each disease, a logistic regression was fitted to explain disease
192 probability as a function of age (#21003), sex, genotyping array, and the 40 first principal
193 components (PCs). Nominally significantly associated covariates ($p \leq 0.05$) were retained for
194 the main analysis. CNV-proxy probes with a CNV frequency $\geq 0.01\%$ were pruned at $r^2 >$
195 0.9999 in PLINK_{CNV} (`--indep-pairwise 500 250 0.9999` PLINK v2.0) to group probes
196 at the core of CNV regions while retaining resolution at breakpoints (BPs), resulting in 18,725
197 probes. For each disease, 2-by-3 genotypic Fisher tests assessed dependence between
198 disease status and probe copy-number (rows: control versus case; columns: deletion versus
199 copy-neutral versus duplication; `--model fisher` PLINK v1.9; TEST column GENO). Probes
200 with $p \leq 0.001$ and a minimum of two disease cases among CNV, duplication, or deletion
201 carriers were retained for assessment through the mirror/U-shaped, duplication-only, or
202 deletion-only model, respectively.

203

204 *Genome-wide significance threshold*

205 Due to the recurrent nature of CNVs, the 18,725 probes retained after frequency filter and
206 pruning remain highly correlated and are thus not independent. Accounting for all of them
207 would result in an overly strict multiple testing correction. Using an established protocol
208 [22,23,51], we estimated the number of effective tests performed to $N_{\text{eff}} = 6,633$, setting the
209 genome-wide (GW) threshold for significance at $p \leq 0.05/6,633 = 7.5 \times 10^{-6}$. This threshold is
210 of the same order of magnitude as what others have estimated for disease CNV-GWAS [8].

211

212 *Main CNV-GWAS model*

213 Association between disease risk and copy-number of CNV-proxy probes was assessed
214 through logistic regression with Firth fallback (`--covar-variance-standardize --glm`
215 `firth-fallback omit-ref no-x-sex hide-covar --ci 0.95 PLINK v2.0`), using
216 disease- and model-specific probes and covariates (*Probe and covariate selection*). Four
217 association models were assessed: the mirror model assesses the additive effect of each
218 additional copy (`PLINKCNV`); the U-shape model assesses a consistent effect of any deviation
219 from the copy neutral state (`PLINKCNV`, using the `hetonly` option in `glm PLINK v2.0`); the
220 duplication-only model (`PLINKDUP`) assesses the impact of a duplication while disregarding
221 deletions; the deletion-only model (`PLINKDEL`) assesses the impact of a deletion while
222 disregarding duplications. Odds ratios (OR) and their 95% confidence interval (CI) were
223 harmonized (A1 to “T”; [Supplemental Note 1 – Table 1](#)), i.e., $OR_{new} = \frac{1}{OR_{old}}$ and
224 $CI_{new} = e^{\log(OR_{new}) \pm 1.96 * SE_{\log(OR_{old})}}$, respectively. GW-significant associations ($p \leq 7.5 \times 10^{-6}$;
225 *Genome-wide significance threshold*) were pruned at $r^2 > 0.8$ (`--indep-pairwise 3000`
226 `500 0.8 PLINK v2.0`), giving priority to probes with the strongest association signal by
227 inputting a scaled negative logarithm of association p-value as frequency (`--read-freq`
228 `PLINK v2.0`). For the U-shape model, pruning was performed using custom code by extracting
229 probes from `PLINKCNV` and recoding them to match U-shape numerical encoding. Number of
230 independent signals per disease was determined by stepwise conditional analysis. Briefly, for
231 each disease and association model, the numerical CNV genotype of the lead probe was
232 included along selected covariates in the logistic regression model. This process was repeated
233 until no more GW-significant signal remained.

234

235 Due to its continuous nature, the disease burden CNV-GWAS was based on linear
236 regressions between copy-number of selected probes and the disease burden (`--glm omit-`
237 `ref no-x-sex hide-covar allow-covars PLINK v2.0`), correcting for selected
238 covariates. Post-GWAS processing was performed as previously described [23].

239

240 *CNV region definition and annotation*

241 CNV region (CNVR) boundaries were defined by the most distant probe within $\pm 3\text{Mb}$ and $r^2 \geq$
242 0.5 of independent lead probes (`--show-tags --tag-kb 3000 --tag-r2 0.5 PLINK`
243 v1.9; U-shape model: custom code, as described previously for pruning). Signals from
244 different models were merged when overlapping ($\geq 1\text{bp}$) and involving the same disease, with
245 CNVR boundaries defined as the maximal CNVR. Characteristics of the most significant
246 model (i.e., “best model”) are reported. The “main model” indicates which CNV type mainly
247 drives the association, i.e., when associations were found through multiple models, priority
248 was given to either the duplication-only or deletion-only models, otherwise to the model
249 yielding the lowest p-value. CNVRs were annotated with hg19 HGNC and ENSEMBL gene
250 names using `annotate_variation.pl` from ANNOVAR (`--geneanno`). Number of genes
251 mapped to a CNVR was calculated and set to zero for CNVRs with `REGION` not equaling
252 “exonic”. Resulting count was used as a predictor for pleiotropy (i.e., number of associations)
253 through linear regression.

254

255 *Statistical confidence tiers*

256 Following primary assessment through logistic regression (*Main CNV-GWAS model*), three
257 statistical approaches were implemented to gauge robustness of the lead probe’s association
258 signal. First, we assessed *post hoc* the p-value of 2-by-3 genotypic Fisher tests (*Probe and*
259 *covariate selection*). Second, we transformed the binary disease status into a continuous
260 variable by computing the response residuals of the logistic regression of disease status on
261 disease-relevant covariates. This allowed the usage of linear regressions to estimate the effect
262 of the CNV genotype (encoded according to all significantly associated models in the primary
263 analysis) on disease risk. The model generating the lowest p-value for the CNV encoding is
264 reported. Third, time-to-event analysis was used to assess whether CNVs influence age-at-
265 disease onset. Age at last healthy measurement was calculated as age-at-disease onset for

266 cases and date of last recorded diagnosis (30/09/2021) minus birth date converted to years
267 for controls (*Case-control definition and age-at-disease onset calculation*). Cox proportional-
268 hazards (CoxPH) models were fitted including disease-relevant covariates and numerically
269 encoded CNV genotype for either of the four association models as predictors, using `coxph()`
270 function from the R `survival` package [52]. The model with the lowest CNV genotype p-
271 value is reported. CNV-disease associations were classified in confidence tiers depending on
272 whether they were confirmed by 3 (tier 1), 2 (tier 2), or 1 (tier 3) of the above-described
273 approaches at the arbitrary validation significance threshold of $p \leq 1 \times 10^{-4}$. Validation
274 approaches not being suited for continuous variables – which do not suffer from the same
275 caveats as binary traits – all disease burden associations were classified as tier 1.

276

277 *Literature-based supporting evidence*

278 Using three literature-based approaches, we examined whether disease-associated CNVRs
279 had previously been linked to relevant phenotypes. First, we investigated the colocalization of
280 autosomal CNVRs with SNP-GWAS signals. GRCh38/hg38 lifted CNVR coordinates were
281 inputted in the GWAS Catalog and associations ($p \leq 1 \times 10^{-7}$) relevant to the investigated
282 disease (i.e., synonym, continuous proxy, or major risk factor) were identified through manual
283 curation. Second, we overlapped OMIM morbid genes (i.e., linked to an OMIM disorder;
284 `morbidmap.txt`) with disease-associated CNVRs. Through manual curation, we flagged
285 OMIM genes associated to Mendelian disorders sharing clinical features with the common
286 disease associated through CNV-GWAS. Third, we examined if implicated CNVRs overlapped
287 regions at which CNVs were found to modulate continuous traits [23] or disease risk [20,25].

288

289 **3. Replication in the Estonian Biobank**

290 *CNV calling and sample selection*

291 Autosomal CNVs were called from Illumina Global Screening Array (GSA) genotype data for
292 193,844 individuals that survived general quality control and had matching genotype-

293 phenotype identifiers, matching inferred versus reported sex, a SNP-call rate $\geq 98\%$, and were
294 included in the EstBB SNP imputation pipeline. CNV outliers and individuals with a reported
295 blood malignancy were excluded, as previously described. High confidence CNV calls ($|QS|$
296 > 0.5) of the 156,254 remaining individuals were encoded into three PLINK binary file sets,
297 following the procedure described for the UKBB ([CNV association studies in the UK Biobank](#)).

298

299 *EstBB disease definition*

300 Disease cases and disease burden were defined similarly than in the UKBB. To account for
301 differences in recording practices between the countries, Z12 (routine preventive screens for
302 cancer), and D22-23 (benign skin lesions) subcodes, were removed from the exclusion list of
303 cancer traits as they were much more frequent than in the UKBB and strongly reduced the
304 number of controls. Due to lack of matching data in the EstBB, no self-reported diseases and
305 cancers were used as an exclusion criterion for disease definition.

306

307 *EstBB replication analysis*

308 Related individuals with available CNV calls were pruned (KING kinship coefficient > 0.0884),
309 prioritizing individuals whose disease status was least often missing, leaving 90,211 unrelated
310 samples for the replication study. Disease-relevant covariates were selected among sex, year
311 of birth, genotyping batch (1-11), and PC1-20. For each of the 73 UKBB signals, probes
312 overlapping the CNVR and with an EstBB CNV, duplication, or deletion frequency $\geq 0.01\%$,
313 were retained, depending on whether the mirror/U-shape, duplication-only, or deletion-only
314 was the best UKBB model, respectively. Association studies were performed on remaining
315 probes using disease-specific covariates and the best UKBB model, following the previously
316 described procedure. Forty signals (55%) could not be assessed due to null/low CNV
317 frequency, failure of the regression to converge, absence of at least one case CNV carrier, or
318 because not mapping on the autosomes. For the remaining 33 signals, summary statistics of
319 the probe showing the strongest association within the CNVR were retained and p-values
320 were adjusted to account for directional concordance with UKBB effects by rewarding and

321 penalizing signals with matching and non-matching effect size signs, respectively. Specifically,
322 one-sided p-values were obtained as $p_{new} = \frac{p_{old}}{2}$ and $p_{new} = 1 - (\frac{p_{old}}{2})$ for 24 concordant and
323 9 non-concordant signals, respectively. Accounting for 33 testable signals, the replication
324 threshold for significance was set at $p \leq 0.05/33 = 1.5 \times 10^{-3}$. One-sided binomial tests
325 (`binom.test()`) were used to assess enrichment of observed versus expected significant
326 associations at various thresholds ($\alpha = 0.1$ to 0.005 by steps of 0.005), with the R function
327 arguments: x the number of observed signals at α , n the number of testable signals (i.e., 33),
328 and p the expected probability of signals meeting α (i.e., α).

329

330 4. CNV region constraint analysis

331 Evolutionary constraint of genes overlapping disease-associated CNVRs, i.e., “disease
332 genes” (*CNV region definition and annotation*), was assessed by comparing their pLI, LOEUF,
333 pHaplo, and pTriplo scores to the ones of “background genes”. The latter were identified by
334 annotating ranges of one or multiple consecutive probes with CNV frequency $\geq 0.01\%$ with
335 ANNOVAR (hg19 HGNC gene names) and excluding disease genes. For pLi and LOEUF, all
336 disease genes were considered together. For pHaplo and pTriplo, two disease gene groups
337 were considered: genes overlapping CNVRs with at least one association through the
338 duplication-only model and genes overlapping CNVRs with at least one association through
339 the deletion-only model. As many CNVRs associated through both models, the analysis was
340 repeated considering genes overlapping CNVRs with at least one association through the
341 duplication-only and none through the deletion-only model and vice-versa. Comparison with
342 background genes was done through two-sided Wilcoxon rank-sum test.

343

344 5. Extended phenotypic assessment

345 To elaborate on specific associations, we made use of the rich phenotypic data available for
346 UKBB participants, as detailed in [Supplemental Note 2](#). For fine-mapping of association

347 signals, CNV carriers were divided in subgroups based on visual inspection of CNV
348 breakpoints and segmental duplications, as detailed in [Supplemental Note 3](#).

349

350 *CNV versus copy-neutral comparisons*

351 Comparisons between groups of CNV carriers and copy-neutral individuals always exclude
352 low quality CNV ($|QS| \leq 0.5$) carriers altogether. For diseases, prevalence is estimated as $q =$

353 $\frac{c}{n}$, with c and n are the number of cases and total number of individuals in a group, and

354 $SE(q) = \sqrt{\frac{q*(1-q)}{n}}$. Differences in prevalence compared to copy-neutral individuals were

355 assessed by two-sided Fisher test. For continuous traits, comparisons are based on two-sided
356 t-tests.

357

358 **6. CNV burden analyses**

359 *CNV burden association studies*

360 In the UKBB, individual-level CNV, duplication, and deletion burden were calculated as the
361 number of Mb or genes affected by high-confidence ($|QS| > 0.5$) autosomal CNVs,
362 duplications, and deletions, respectively, as previously described [23]. Association between
363 burden values and the 60 diseases (logistic regression) or the disease burden (linear
364 regression), was assessed including disease-relevant covariates in the model. Accounting for
365 the 61 evaluated traits, significance was defined at $p \leq 0.05/61 = 8.2 \times 10^{-4}$. We next corrected
366 burden values for CNV-GWAS signals. For each disease, CNVs, duplications, and deletions
367 overlapping (≥ 1 bp) a CNVR significantly associated with the disease of interest through CNV-
368 GWAS were omitted from the CNV, duplication, and deletion burden calculations if the CNVR
369 had been found to associate with the disease through the mirror/U-shape, duplication-only, or
370 deletion-only model, respectively. Association studies were repeated using corrected burden
371 values. Only the most significant burden types are reported in the text.

372

373 *Relative importance of protein coding regions in mediating the burden's effect*

374 The average genome-wide gene density (GD_{GW}) was estimated to 8.4 genes/Mb based on
375 26,289 genes (i.e., unique HGNC gene names in hg19 RefSeq, excluding microRNAs but
376 including genes of uncertain function (i.e., “LOC”)) and a human genome length of
377 3,137,161,264 bp. If CNVs affecting the coding and non-coding DNA have similar effects, we
378 expect that the impact of 1Mb affected by CNVs to be equivalent to 8.4 genes being affected
379 by CNVs. Hence, the association effect size of the CNV burden measured in Mb (β_{Mb}) is
380 expected to be 8.4-times larger than the one measured in number of affected genes (β_{gene}),
381 i.e., $\frac{\beta_{Mb}}{\beta_{gene}} = 8.4$. This hypothesis was tested independently for the deletion and duplication
382 burdens for traits with at least one significant uncorrected burden association (i.e., 20 diseases
383 + disease burden). Significant deviations from the expected ratio were assessed by t-statistic:

$$384 \quad t = \frac{8.4 - \frac{\hat{\beta}_{Mb}}{\hat{\beta}_{gene}}}{\widehat{SD}}$$

385 where $\hat{\beta}_{Mb}$ and $\hat{\beta}_{gene}$ are the estimated effects of the burden measured in Mb or number of
386 genes impacted by CNVs, respectively, on the assessed trait. \widehat{SD} is the empirically observed
387 standard deviation of the $\frac{\tilde{\beta}_{Mb}}{\tilde{\beta}_{gene}}$ ratio, estimated based on 10,000 simulations of $\tilde{\beta}_{Mb} \sim N(\hat{\beta}_{Mb},$
388 $\widehat{Var}(\hat{\beta}_{Mb}))$ and $\tilde{\beta}_{gene} \sim N(\hat{\beta}_{gene}, \widehat{Var}(\hat{\beta}_{gene}))$. P-values were computed based on a two-sided
389 one sample t-test and deemed significant at $p \leq 0.05/21 = 2.4 \times 10^{-3}$. This analysis was
390 repeated for the modified burden definitions not accounting for CNVs overlapping disease-
391 associated CNVRs.

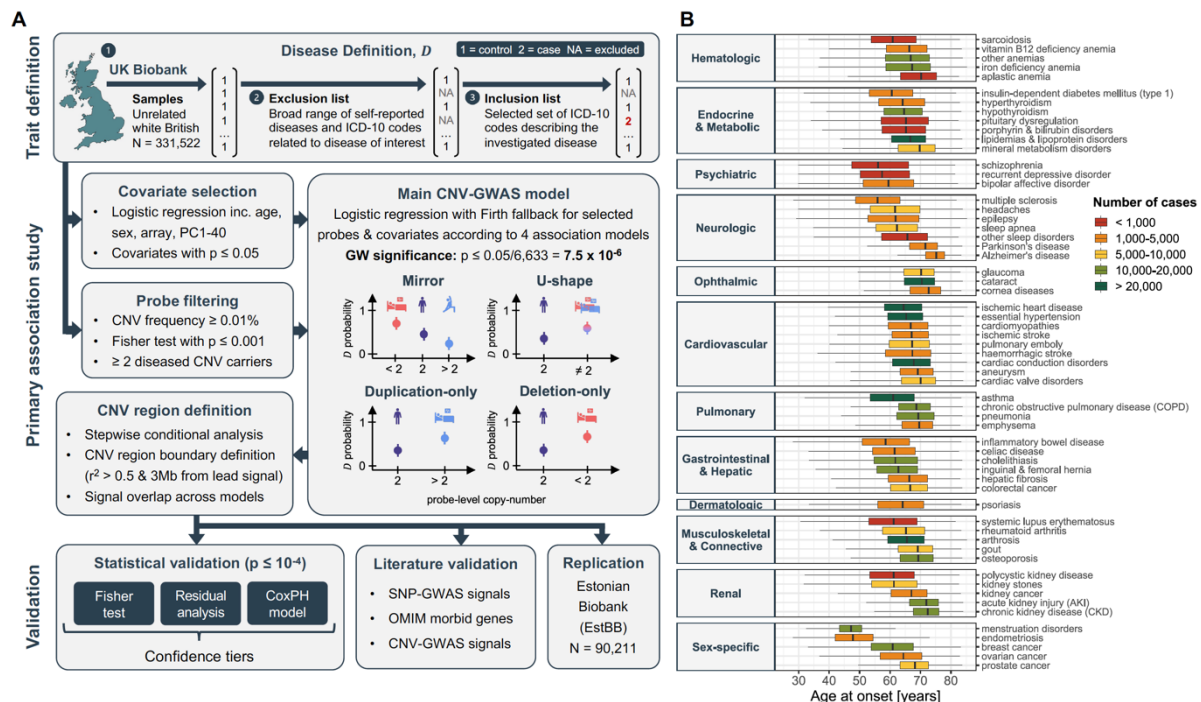
392

393 RESULTS

394 *The spectrum of common diseases in the UK Biobank*

395 To mitigate issues related to disease definition, we used a three-step approach to designate
396 cases and controls in the UKBB (Figure 1A; top). Starting from 331,522 unrelated white British
397 individuals, we defined cases based on a narrow list of hospital-based diagnoses (i.e., ICD-
398 10 codes) and excluded self-reported cases, as well as self-reported and hospital diagnoses

399 of related conditions. Sixty disorders spanning 12 ICD-10 chapters were selected to cover a
 400 wide range of physiological systems, favoring conditions with sufficiently large sample size
 401 and a likely genetic basis (Figure 1B; Figure S1A; Table S1). Except for systemic lupus
 402 erythematosus (N = 422) and polycystic kidney disease (N = 454), all diseases had over 500
 403 cases. Nineteen diseases had a case count > 10,000, with osteoarthritis (N = 62,175) and
 404 essential hypertension (N = 97,860) being the most frequent. Seven diseases had a median
 405 age of onset ≤ 60 years, predominantly female reproductive disorders, autoimmune
 406 conditions, and psychiatric disorders. Conversely, the nine diseases with a median age at
 407 onset ≥ 70 years were mainly degenerative disorders of the brain, eye, and kidney, overall
 408 aligning align with epidemiological knowledge of the respective diseases.



409

410 **Figure 1. Overview of the study**

411 **(A)** Schematic representation of the analysis workflow. Trait definition: For each of the 60 investigated
 412 diseases, unrelated white British UK Biobank participants were assigned as controls, individuals self-
 413 reporting or diagnosed with the disease of interest or a broader set of related conditions were excluded
 414 and set as missing, and individuals with a hospital-based ICD-10 diagnosis of the condition of interest
 415 were re-introduced as cases. Primary association study: Disease-specific relevant covariates were
 416 selected. Probes were pre-filtered based on copy-number variant (CNV) frequency, required to
 417 associate with the disease, and a minimum of two diseased carriers was required for the probe to be
 418 carried forward. Disease- and model-specific covariates and probes were used to generate tailored
 419 CNV genome-wide association studies (GWASs) based on Firth fallback logistic regression according
 420 to a mirror, U-shape, duplication-only (i.e., considering only duplications), and deletion-only (i.e.,
 421 considering only deletions) models. Independent lead signals were identified through stepwise

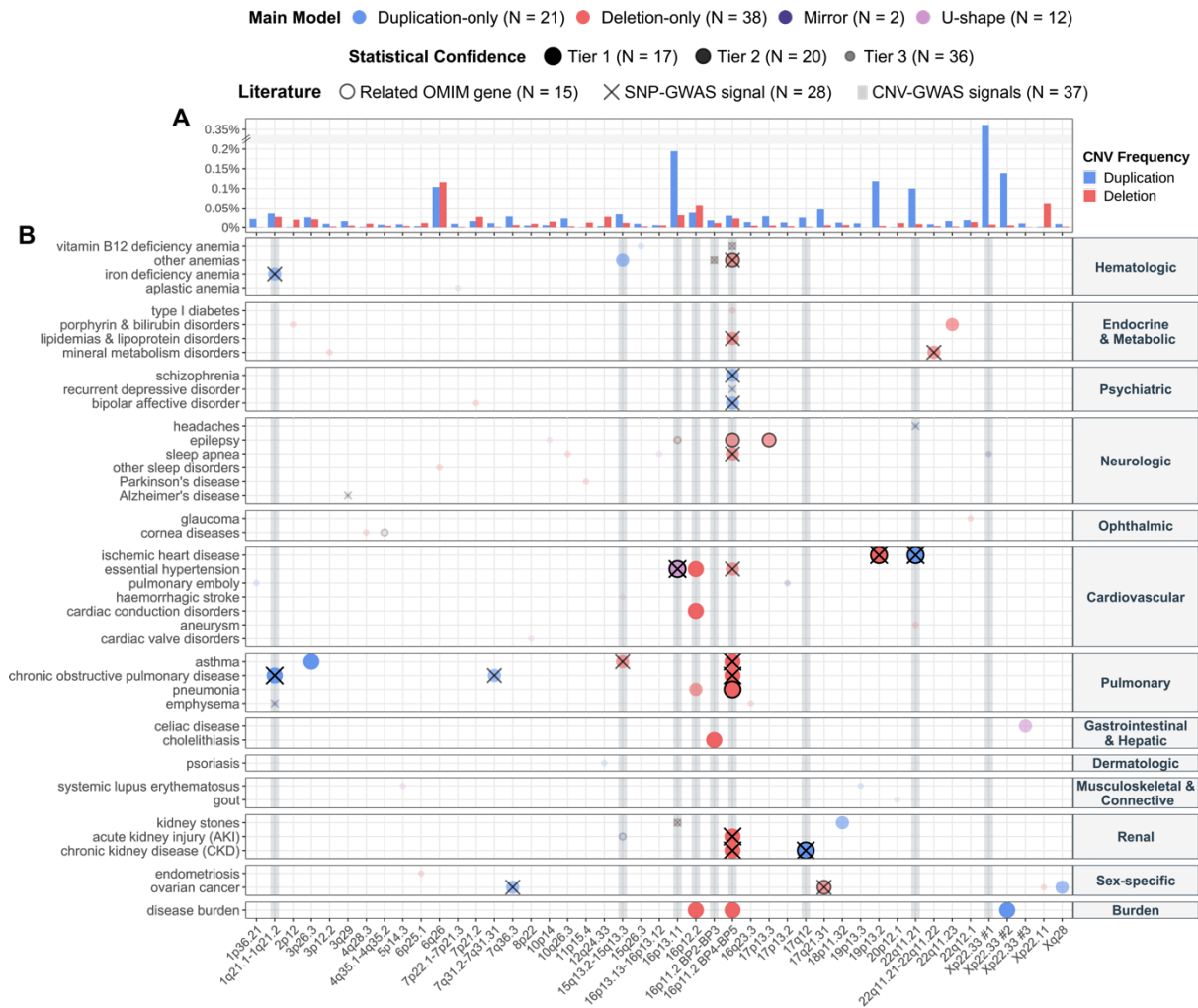
422 conditional analysis and CNV regions were defined based on probe correlation and merged across
423 models. Validation: Statistical validation methods (i.e., Fisher test, residuals regression, and Cox
424 proportional hazards model (CoxPH)) were used to rank associations in confidence tiers. Literature
425 validation approaches leverage data from independent studies to corroborate that genetic perturbation
426 (i.e., single-nucleotide polymorphisms (SNP), rare variants from the OMIM database, and CNVs) in the
427 region are linked to the disease. Independent replication in the Estonian Biobank. **(B)** Age at onset for
428 the 60 assessed diseases, categorized based on ICD-10 chapters and colored according to case count.
429 Data are represented as boxplots; outliers are not shown.

430 *Copy-number variant genome-wide association study*

431 To assess whether disease susceptibility is modulated by CNVs, we performed CNV genome-
432 wide association studies (GWASs), i.e., test if the copy-number of selected probes influence
433 the probability to develop a disease or an individual's disease burden (i.e., number of
434 diagnoses among the 60 studied diseases) (see [Methods](#); [Figure 1A](#); middle). Briefly,
435 microarray-called CNVs for 331,522 unrelated white British individuals were transformed to
436 the probe level after quality-control [23]. As CNVs can act through different gene dosage
437 mechanisms, four association models were assessed: mirror and U-shape models consider
438 deletions and duplications simultaneously, assuming that they impact disease risk in opposite
439 or identical direction, respectively, while the CNV type-specific duplication- and deletion-only
440 models assess independently the effect of duplications and deletions, respectively. To reduce
441 the number and complexity of implemented logistic regressions, pre-processing steps
442 selected relevant covariates and probes for each disease and model combination, thereby
443 lowering computation time and decreasing the multiple testing burden ([Figure S2](#); [Table S2](#)).
444 All summary statistics are available ([Data Availability](#)).

445
446 Stepwise conditional analysis narrowed GW significant associations ($p \leq 7.5 \times 10^{-6}$; see
447 [Methods](#) for threshold calculation) to 40, 41, 21, and 38 independent signals for the mirror, U-
448 shape, duplication-only, and deletion-only models, respectively. These were combined into 70
449 risk-increasing (i.e., no disease-protecting CNV) associations and 3 disease burden
450 associations ([Figure 2](#); [Table S3](#)). Forty-five associations (45/73 = 62%) were supported at
451 GW significance by multiple models, the lowest p-value (i.e., "best model") being obtained
452 through the mirror, deletion-only, U-shape, and duplication-only models for 24, 23, 21, and 5

453 of the signals, respectively. No association was detected at GW significance by both the
 454 duplication-only and deletion-only models, so that each signal was attributed a “main model”
 455 that indicates whether the association is primarily driven by duplications or deletions (see
 456 [Methods; Figure 2](#)). The main model should be interpreted with caution as both deletions and
 457 duplications might influence disease risk but only one CNV type-specific model might reach
 458 GW significance (e.g., due to higher frequency). This is particularly relevant as 73% (33/45)
 459 of disease-associated CNV regions (CNVRs) have a higher duplication than deletion
 460 frequency ([Figure 2A](#)). Hence, 95% (20/21) of signals mainly driven by duplications were also
 461 identified by the mirror/U-shape model(s) and contribution of deletions cannot be excluded.



462 **Figure 2. CNV-disease association map**

463 (A) Duplication and deletion frequencies ([%]; y-axis; break: //) of the lead probe for each unique and
 464 non-overlapping disease-associated CNV region (CNVR), labeled with corresponding cytogenetic band
 465 (x-axis; 16p11.2 is split to distinguish the distal 220kb breakpoint BP2-3 and proximal 600kb BP4-5
 466 CNVRs; non-overlapping CNVRs on the same cytogenetic band are numbered). If signals mapping to
 467 the same CNVR have different lead probes, the maximal frequency was plotted. (B) Associations

468 between CNVRs (x-axis) and diseases (y-axis) identified through CNV-GWAS. Color indicates the main
469 association model. Size and transparency reflect the statistical confidence tier. Black contours indicate
470 overlap with OMIM gene causing a disease with shared phenotypic features. Black crosses indicate
471 overlap with SNP-GWAS signal for a related trait. Grey shaded lines indicate CNVRs with continuous
472 trait associations [23]. N provides count for various features.

473 *Validation of identified CNV-GWAS signals*

474 Across the 45 CNVRs, CNV frequencies were low, ranging between 0.01% (our frequency
475 cutoff) and 0.36%, with 87% (39/45) of CNVRs having a frequency $\leq 0.1\%$ (Figure 2A).
476 Consequently, associations rely on a low number of diseased CNV carriers and require
477 validation (see Methods; Figure 1A; bottom; Figure 2B; Table S3). We used three statistical
478 approaches to assess the robustness of CNV-diseases associations: i) Fisher test, ii) residual
479 regression, and iii) time-to-event analysis through CoxPH modeling. We replicated 28/70
480 (40%), 23/70 (33%), and 70/70 (100%) of the associations with the respective methods at the
481 arbitrary validation threshold of $p \leq 10^{-4}$. This allowed to stratify associations in confidence
482 tiers, with 17 signals replicating with all methods (tier 1), 20 with two (tier 2), and 36 only
483 through time-to-event analysis (tier 3). Importantly, time-to-event analysis showed that CNVs
484 always contributed to an earlier age of disease onset, in line with the paradigm that diseases
485 with a strong genetic etiology have earlier onset [53].

486

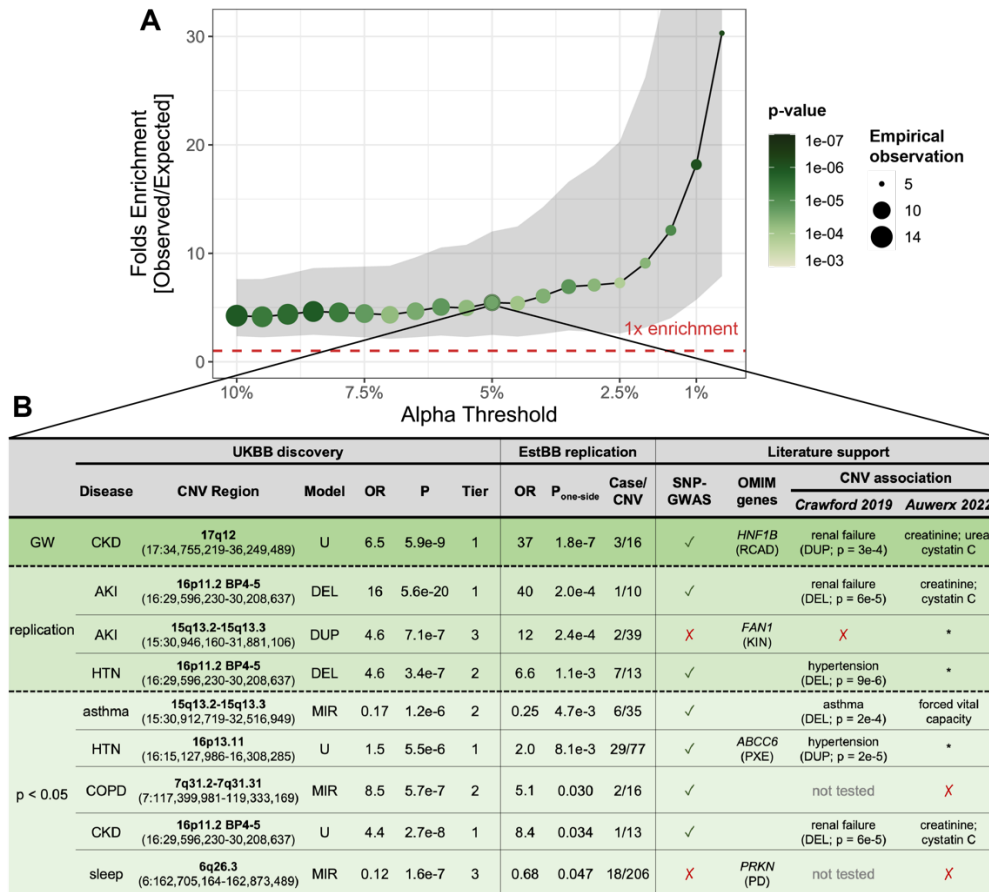
487 In parallel, we gathered literature evidence linking genetic variation at CNVRs with relevant
488 phenotypes (Table S3). Forty-eight signals (48/73 = 64%) mapped to a CNVR harboring a
489 least one OMIM morbid gene and in 15 cases, the gene was linked to a Mendelian disorder
490 sharing phenotypic features with the associated common disease. For instance, association
491 between 4q35 CNVs and corneal conditions (chr4:186,687,554-187,182,384; $OR_{U-shape} = 18.2$;
492 95%-CI [5.2; 63.1]; $p = 5.0 \times 10^{-6}$) encompasses *CYP4V2* [MIM: 608614], a gene associated
493 with autosomal recessive Bietti crystalline corneoretinal dystrophy [MIM: 210370], a disorder
494 that impairs vision and progresses to blindness by age 50-60 years [54]. We next assessed
495 whether SNPs overlapping disease-associated CNVRs were reported to associate with the
496 implicated disease or a biomarker thereof in the GWAS Catalog. This was the case for 28

497 (28/66 = 42%) autosomal signals, a similar proportion (38%) than for continuous trait CNV-
498 GWAS [23]. For instance, distal 22q11.2 CNVs increased risk for disorders of mineral
499 metabolism (chr22:21,797,101-22,661,627; $OR_{\text{mirror}} = 0.02$; 95%-CI [0.006; 0.083]; $p = 9.9 \times$
500 10^{-9}) and overlapped heel bone mineral density SNP-GWASs signals, while 3q29 CNVs
501 increased Alzheimer's disease risk (chr3:196,953,177-197,331,898; $OR_{\text{U-shape}} = 11.8$; 95%-CI
502 [4.0; 34.7]; $p = 6.6 \times 10^{-6}$) and overlapped with SNP-GWAS signal for PHF-tau levels, and
503 suggestive signals ($p < 5 \times 10^{-6}$) for frontotemporal dementia and cognitive decline in
504 Alzheimer's disease. Finally, 37 signals (37/73 = 51%) mapped to nine CNVRs previously
505 found to be associated with complex traits [23].

506

507 We also set out to replicate association signals in 90,211 unrelated EstBB individuals [40],
508 using similarly case definition than for the UKBB (see [Methods](#); [Figure S1B](#)). Requesting at
509 least one diseased CNV carrier, 33 of 73 associations could be evaluated, among which four
510 were strictly replicated ($p \leq 0.05/33 = 1.5 \times 10^{-3}$) and five additional ones reached nominal
511 significance ($p \leq 0.05$) ([Table S3](#)). Compared to what would be expected by chance, this
512 corresponds to a 5.5-fold ($p_{\text{binomial}} = 2.5 \times 10^{-5}$) and 30.3-fold ($p_{\text{binomial}} = 6.6 \times 10^{-7}$) enrichment
513 for replication at $p \leq 0.05$ and $p \leq 5 \times 10^{-3}$, respectively ([Figure 3A](#)). Despite low power, these
514 results support validity of the primary UKBB association signals. Signals replicating at nominal
515 significance are detailed in [Figure 3B](#). Many harbor SNP-GWAS signals for related
516 phenotypes (7/9), relevant morbid OMIM genes (4/9), or map to CNVRs previously associated
517 with similar diseases (6/7) or biomarkers (4/9). Among them, two are in the lowest UKBB
518 confidence tier. 15q13 duplications were linked to increased risk for acute kidney injury (AKI;
519 chr15:30,946,160-31,881,106 | UKBB: $OR_{\text{dup}} = 4.6$; 95%-CI [2.5; 8.4]; $p = 7.1 \times 10^{-7}$ | EstBB:
520 $p = 2.4 \times 10^{-4}$). Homozygous mutations in *FAN1* [MIM: 613534], one of the five genes mapping
521 to this CNVR, have been linked to karyomegalic interstitial nephritis [MIM: 614817], a
522 progressive renal condition that leads to CKD [55]. The second example links CNVs affecting
523 exon 2 and intron 2-3 of *PRKN* ([MIM: 602544]) – a gene causing juvenile autosomal recessive
524 Parkinson's disease [MIM: 600116] – to sleep disorders such as insomnia and hypersomnia

525 (chr6:162,705,164-162,873,489 | UKBB: $OR_{\text{mirror}} = 0.12$; 95%-CI [0.05; 0.26]; $p = 1.6 \times 10^{-7}$ |
 526 EstBB: $p = 0.047$). This finding is particularly relevant given the region's high CNV frequency
 527 (0.22%) and the fact that sleep disturbances are among the earliest symptoms of Parkinson's
 528 disease [56]. Follow-up studies should determine whether these individuals are more prone
 529 to develop Parkinson's disease in the future.



530 **Figure 3. Replication of CNV-disease associations in the Estonian Biobank**
 531 **(A)** Enrichment for signal replication (y-axis; 95% confidence interval as grey ribbon) at different levels
 532 of significance (alpha; x-axis) in the Estonian Biobank (EstBB). Color and size indicate the p-value of
 533 the enrichment (one-sided binomial test) and the number of observed associations, respectively.
 534 Dashed red line indicates 1x enrichment, i.e., the number of observed associations matches the number
 535 of expected ones. **(B)** Associations replicated at nominal significance in the EstBB, color-stratified
 536 according to whether they meet the discovery genome-wide (GW; $p \leq 7.5 \times 10^{-6}$; dark green), replication
 537 ($p \leq 1.5 \times 10^{-3}$; green), or nominal ($p \leq 0.05$; light green) significance threshold. Disease (CKD = chronic
 538 kidney disease; AKI = acute kidney injury; HTN = hypertension; COPD = chronic obstructive lung
 539 disease), cytogenic band and coordinates, best model (MIR = mirror; U = U-shape; DUP = duplication-
 540 only; DEL = deletion-only), odds ratio (OR), p-value (P) and statistical confidence tier are given for the
 541 UK Biobank (UKBB) discovery analysis. OR, one-sided p-values, and number of cases among CNV
 542 carriers are provided for the EstBB replication. Overlap with SNP-GWAS signals for a related trait (✓ =
 543 yes; X = no) or a relevant OMIM gene (RCAD = renal cyst and diabetes; KIN = karyomegalic interstitial
 544 nephritis; PXE = pseudoxanthoma elasticum; PD = Parkinson's disease) is indicated. Previous
 545 association with diseases[20] (duplication (DUP) or deletion (DEL) was associated with indicated

546 disease; no association (X); some CNVRs were not tested) and continuous traits [23] (disease-relevant
547 biomarkers are specified; other traits (*); no association (X)) are listed.

548 Evidence provided by statistical, literature-based, or independent replication help prioritizing
549 the most promising associations for follow-up studies and pinpoint plausible candidate genes.
550 We highlight several examples where deviations by one copy-number are linked to common
551 diseases sharing clinical features with rare Mendelian conditions caused by homozygous
552 perturbations of the same genetic region. This argues against a dichotomic view on dominant
553 versus recessive modes of inheritance and analogously to allelic series, suggest that
554 Mendelian and common diseases represent different ends of the phenotypic spectrum caused
555 by genetic variation at a given locus.

556

557 *Global characterization of disease-associated CNV regions*

558 We sought to identify global characteristics that distinguish disease-associated CNVRs ([Table](#)
559 [S4](#)). Number of protein-coding genes embedded in disease-associated CNVRs, hereafter
560 referred to as “disease genes”, ranged from 0 to over 30 and generally correlated with the
561 number of encompassed probes ($\rho_{\text{Pearson}} = 0.50$; $p = 4.2 \times 10^{-4}$; [Figure S3A](#)). Exceptions
562 include single-gene CNVRs overlapping known pathogenic genes captured thanks to high
563 probe coverage (e.g., *BRCA1*). While only seven CNVRs (16%) associated with multiple
564 diseases, propensity for pleiotropy depended on CNV length (+0.16 association/disease gene;
565 $p = 1.5 \times 10^{-5}$). Accordingly, CNVRs containing more than five genes were also more likely to
566 associate with continuous traits ($\text{OR}_{\text{Fisher}} = 53.2$; $p = 8.5 \times 10^{-6}$) [23]. One CNVR that stood out
567 is the 600kb 16p11.2 BP4-5 region ([Figure 2B](#)). Originally identified as a major risk factor for
568 autism, schizophrenia, developmental delay and intellectual disability, macro-/microcephaly,
569 epilepsy, and obesity [57–63], we previously found the region to associate with 26 continuous
570 complex traits [23]. Here, we show that 16p11.2 BP4-5 deletions increase the risk of 12
571 diseases – including both new and previously reported associations across multiple organ
572 systems – as well as the disease burden (+3 diseases/deletion; $p = 1.2 \times 10^{-26}$), while the

573 region's duplication drove increased risk for psychiatric conditions (i.e., bipolar disorder,
574 schizophrenia, and depression), in line with previous findings [62].

575

576 Next, we assessed whether disease genes were under stronger evolutionary constraint (i.e.,
577 less tolerant to mutations) than genes affected by CNVs at the same frequency but not
578 associated with any disease (i.e., “background genes”). Compared to background genes, the
579 231 disease genes had more constrained pLI ($p_{\text{Wilcoxon}} = 1.3 \times 10^{-4}$; [Figure S3B](#)) and LOEUF
580 ($p_{\text{Wilcoxon}} = 1.9 \times 10^{-7}$; [Figure S3C](#)) scores, suggesting stronger intolerance to LoF mutations.
581 Splitting CNVRs depending on whether they have at least one association through either the
582 duplication-only or deletion-only model, we evaluated whether embedded disease genes were
583 sensitive to having less (i.e, haploinsufficiency; [Figure S3D](#)) or more (i.e., triplosensitivity;
584 [Figure S3E](#)) than two functional copies. No significant difference in pHaplo scores were
585 observed but genes overlapping regions whose duplication ($p_{\text{Wilcoxon}} = 9.0 \times 10^{-19}$) and deletion
586 ($p_{\text{Wilcoxon}} = 1.0 \times 10^{-23}$) have been linked to diseases were more likely to be triplosensitive than
587 background genes. Similar trends were observed considering genes overlapping CNVRs
588 involved uniquely through the duplication-only and deletion-only models and not the other
589 CNV type-specific model ([Figure S3F-G](#)). Overall, our results indicate that a CNVR's
590 pathogenicity is determined both by the number and characteristics of affected genes.

591

592 *New insights in known disease genes*

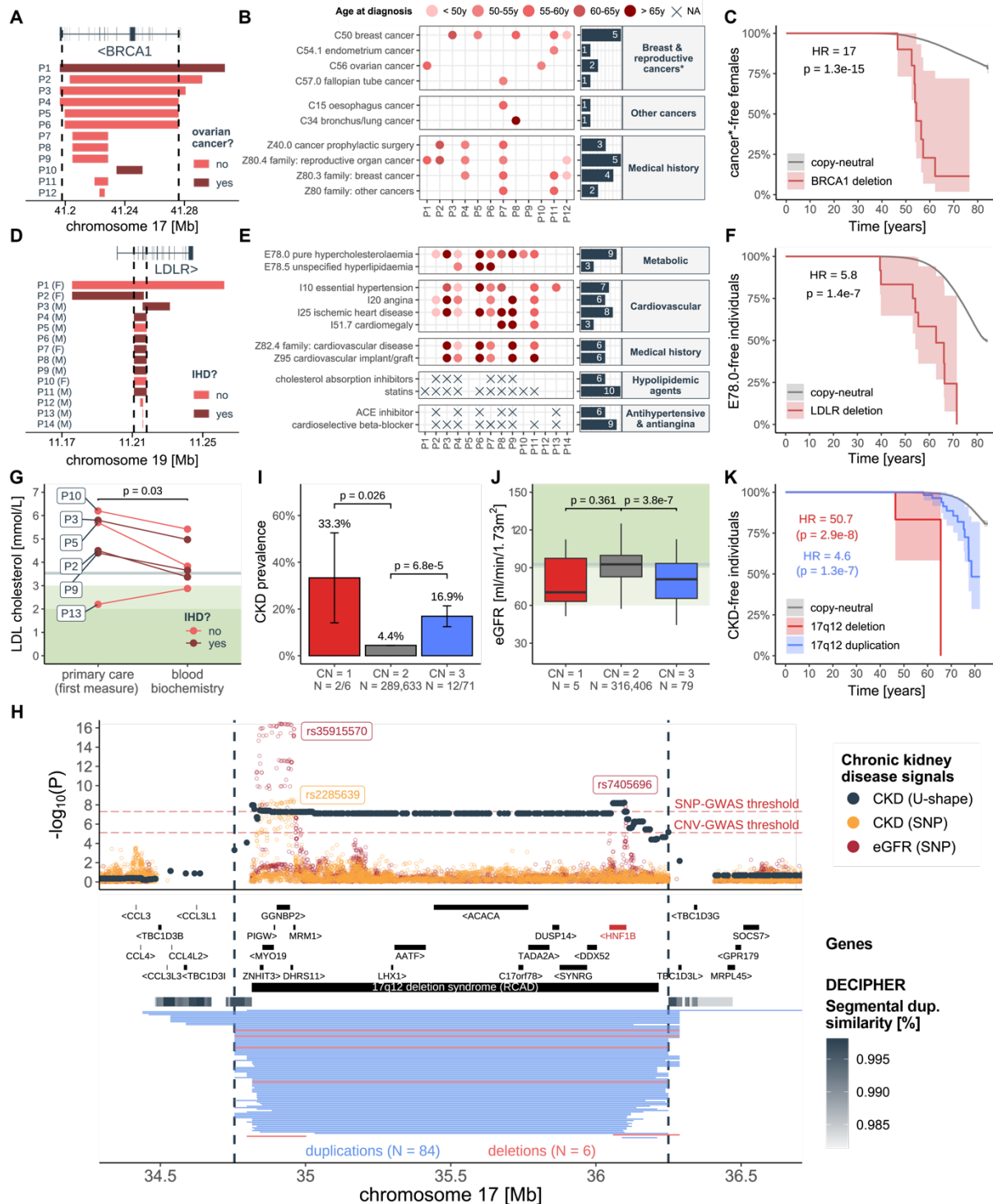
593 Two out of 12 female *BRCA1* deletion carriers were diagnosed with ovarian cancer
594 (chr17:41,197,733-41,276,111; $\text{OR}_{\text{del}} = 284.3$; 95%-CI [24.6; 3290.8]; $p = 6.1 \times 10^{-6}$; [Figure](#)
595 [4A](#)). *BRCA1* [MIM: 113705] is a tumor suppressor gene whose LoF represents a major genetic
596 risk factor for the development of hereditary breast and ovarian cancer (HBOC) [MIM: 604370]
597 [64]. Exploring the clinical records of the 12 deletion carriers, we found five diagnoses of breast
598 cancer (a trait assessed by CNV-GWAS but that did not yield a GW-significant association),
599 one of endometrial cancer, and one of Fallopian tube cancer, so that eight carriers (67%) had
600 received a HBOC diagnosis ([Figure 4B](#)). Not only was prevalence of HBOC higher among

601 *BRCA1* deletion carriers ($OR_{\text{Fisher}} = 31.0$; $p = 1.1 \times 10^{-6}$), but disease onset was earlier ($HR =$
602 17.0 ; $p = 1.3 \times 10^{-15}$; [Figure 4C](#)). Among the four carriers with no HBOC, two had received
603 cancer prophylactic surgery, de facto reducing the penetrance of the deletion. Surgeries were
604 likely carried out based on family history of HBOC, which was reported for 6 carriers (50%),
605 suggesting that these deletions are inherited. We did not observe higher prevalence of other
606 cancer types ([Figure 4B](#)).

607

608 High abundance of *Alu* repeats make the low-density lipoprotein (LDL) receptor (*LDLR*) [MIM:
609 606945] susceptible to CNVs [65]. We found that deletion of exon 2-6 increased risk for
610 ischemic heart disease (IHD; chr19:11,210,904-11,218,188; $OR_{\text{del}} = 31.2$; 95%-CI [7.1; 137.8];
611 $p = 5.6 \times 10^{-6}$), a condition present in 8 of 14 deletion carriers ([Figure 4D](#)). Heterozygous - and
612 less frequently homozygous – mutations in *LDLR* represent the main genetic etiology for
613 familial hypercholesterolemia [66], which is characterized by elevated LDL cholesterol and
614 predisposition for adverse cardiovascular outcomes [67]. Previously identified in clinical
615 studies of familial hypercholesterolemia [68], the CNVR implicated by our analysis specifically
616 encompasses the ligand-binding domain of *LDLR* [66]. Confirming widespread prevalence and
617 family history (43%) of cardiovascular diseases ([Figure 4E](#)), medical records of deletion
618 carriers further revealed higher prevalence ($OR_{\text{Fisher}} = 11.6$; $p = 7.9 \times 10^{-5}$) and earlier onset
619 ($HR = 5.8$; $p = 1.4 \times 10^{-7}$; [Figure 4F](#)) of pure hypercholesterolemia (E78.0), a code included in
620 our lipidemia definition but that did not yield a signal pick-up by the CNV-GWAS. As we
621 previously did not find the CNVR to associate with standardized blood biochemistry LDL levels
622 [23], we hypothesized that the latter were lowered by hypolipidemic agents. Ten (71%)
623 deletion carriers were on statins and six (43%) were additionally using cholesterol absorption
624 inhibitors, while the remaining four did not receive a dyslipidemia or IHD diagnosis and
625 harbored smaller deletions (i.e., P12-14; [Figure 4E](#)). We concluded that drugs likely masked
626 genetically determined LDL levels, as shown by higher LDL levels in the first primary care
627 measurement on record, measured prior to the standardized LDL measurement ($p_{\text{t-test}} = 0.03$;
628 [Figure 4G](#)). Despite this, the recommended target of ≤ 1.8 mmol/L for high-risk individuals [69]

629 was never met. By recovering known gene-disease pairs typically studied in clinical cohorts,
 630 we showcase how the rich phenotypic data from biobanks can generate insights into the
 631 mechanisms, epidemiology, and comorbidities of these diseases, implicating CNVs as
 632 important genetic risk factors.



633 **Figure 4. Refining contribution of CNVs to gene-disease pairs**

634 (A) Genomic coordinates of the 12 females (P1-12) carrying a *BRCA1* deletion (CNVR delimited by
635 vertical dashed lines), colored according to ovarian cancer diagnosis. (B) Left: Cancer and related
636 family/personal diagnoses received by individuals in (A). Color indicates age at diagnosis. Right: Counts
637 per ICD-10 code. (C) Kaplan-Meier curve depicting the percentage, with 95% confidence interval, of
638 females free of female-specific cancers over time among copy-neutral and *BRCA1* deletion carriers.
639 Hazard ratio (HR) and p-value for the *BRCA1* deletion are given (CoxPH model). (D) Genomic
640 coordinates of the 14 individuals (P1-14) carrying an *LDLR* deletion (CNVR delimited by vertical dashed
641 lines), colored according to ischemic heart disease (IHD) diagnosis. (E) Left: Medical conditions and
642 family/personal diagnoses and medication received by ≥ 3 *LDLR* deletion carriers in (D), following
643 legend in (B). (F) Kaplan-Meier curve for pure hypercholesterolemia (E78.0) among copy-neutral and
644 *LDLR* deletion carriers, following legend as in (C). (G) Low density lipoprotein (LDL)-cholesterol levels
645 (y-axis) from primary care data (first available measurement) and blood biochemistry (average over
646 instances) for six deletion carriers in (D) with at least one antecedent primary care LDL-cholesterol
647 measurement, colored according to IHD diagnosis. P-value compares the two data sources (paired
648 one-sided t-test). Grey horizontal line represents median LDL-cholesterol value (from blood
649 biochemistry) in non-carriers. Light and darker green background represent recommended target
650 values for low (≤ 3 mmol/L) and high (≤ 1.8 mmol/L) risk individuals, respectively. (H) 17q12 association
651 landscape. Top: Negative logarithm of association p-values of CNVs (dark grey; CNVR delimited by
652 vertical dashed lines) and SNPs (orange) [70] with chronic kidney disease (CKD) and SNPs with
653 estimated glomerular filtration rate (eGFR; red) [71]. Lead SNPs are labeled. Red horizontal dashed
654 lines represent the genome-wide threshold for significance for CNV-GWAS ($p \leq 7.5 \times 10^{-6}$) and SNP-
655 GWAS ($p \leq 5 \times 10^{-8}$). Middle: Genomic coordinates of genes and DECIPHER CNV, with *HNF1B*, the
656 putative causal gene in red. Segmental duplications are represented as a gray gradient proportional to
657 the degree of similarity. Bottom: Genomic coordinates of duplications (blue) and deletions (red) of UK
658 Biobank participants overlapping the region. (I) CKD prevalence (\pm standard error) according to 17q12
659 copy-number (CN). P-values compare deletion (CN = 1) and duplication (CN = 3) carriers to copy-
660 neutral (CN = 2) individuals (two-sided Fisher test). Number of cases and sample sizes are indicated
661 (N = cases/sample size). (J) eGFR levels according to 17q12 CN, shown as boxplots; outliers are not
662 shown. P-values comparisons as in (I) (two-sided t-test). Grey horizontal line represents median eGFR
663 in non-carriers. Light and darker green background represent mildly decreased (60-90 ml/min/1.73m²)
664 and normal (≥ 90 ml/min/1.73m²) kidney function, respectively. (K) Kaplan-Meier curve for CKD among
665 17q12 deletion and duplication carriers, following legend as in (C).

666 *Biomarker CNV associations tag pathophysiological processes*

667 Integration of biomarker and disease CNV-GWAS signals can identify high-confidence,
668 clinically relevant associations. Heterozygous LoF of *HNF1B* [MIM: 189907] and 17q12
669 deletions cause renal cyst and diabetes (RCAD) [MIM: 137920], a severe disorder
670 characterized by renal abnormalities and maturity-onset diabetes by the young [72,73]. While
671 we previously showed that renal biomarkers were increased in duplication carriers [23], here,
672 we demonstrate that both 17q12 deletions and duplications increase CKD risk
673 (chr17:34,755,219-36,249,489; $OR_{U\text{-shape}} = 6.5$; 95%-CI [3.4; 12.1]; $p = 5.9 \times 10^{-9}$; Figure 4H),
674 with a prevalence of 33.3% ($p_{t\text{-test}} = 0.026$) and 16.9% ($p_{t\text{-test}} = 6.8 \times 10^{-5}$) among deletion and

675 duplication carriers respectively, versus 4.4% in copy-neutral individuals (Figure 4I). Results
676 replicated in the EstBB ($p = 1.8 \times 10^{-7}$; Figure 3B) and are supported by 20% of CNV carriers
677 showing signs of kidney disease based on estimated glomerular filtration rate ($eGFR < 60$
678 $ml/min/1.73m^2$), compared to 2.2% in copy-neutral individuals (Figure 4J). Importantly, both
679 17q12 deletion and duplication lower age of CKD onset ($HR \geq 4.6$; $p \geq 1.3 \times 10^{-7}$; Figure 4K),
680 providing strong evidence of the deleterious consequences on kidney health of altered dosage
681 of 17q12.

682

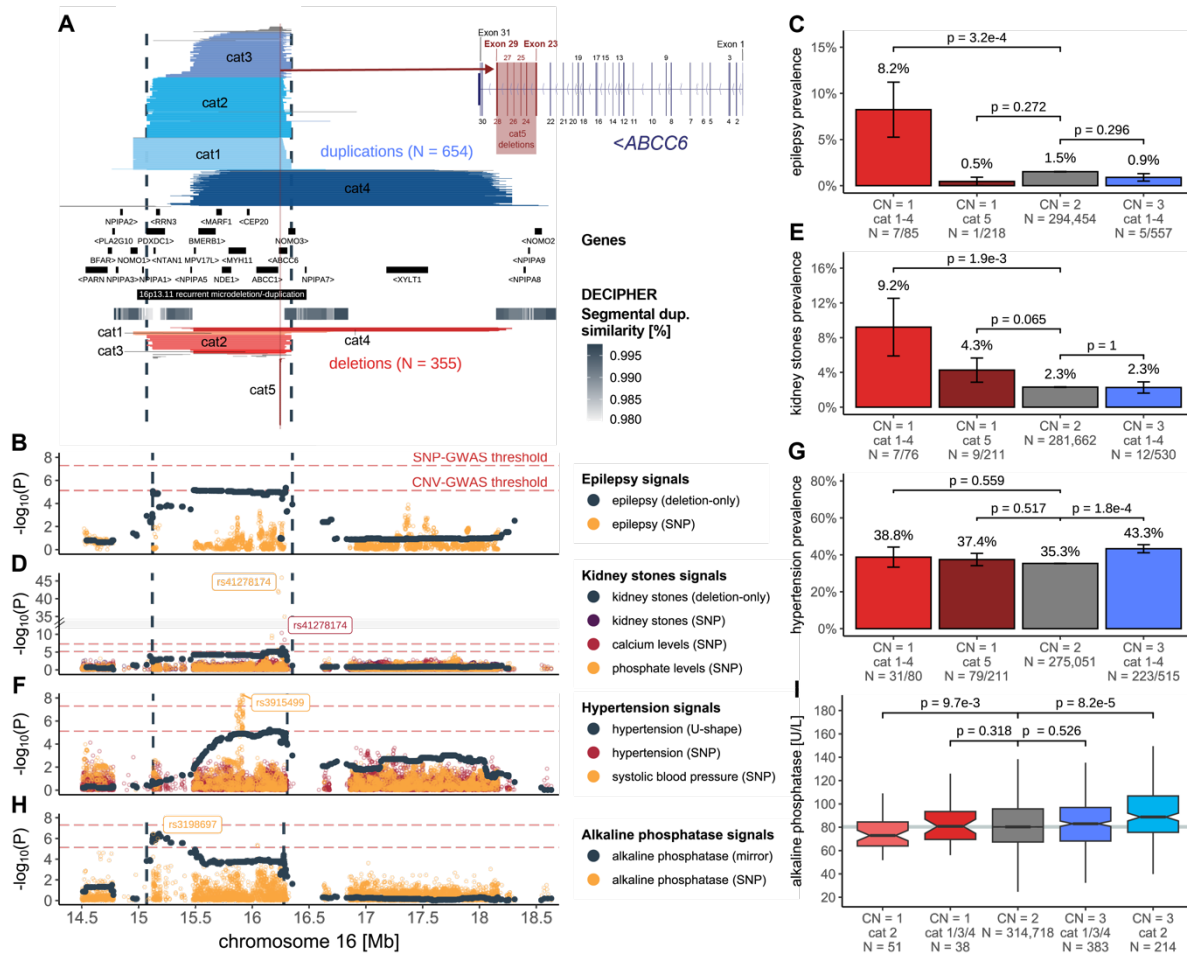
683 Similarly, the blood pressure-increasing 16p12.2 deletion (chr16:21,946,523-22,440,319)
684 [19,23] increased risk for hypertension ($OR_{del} = 2.7$; 95%-CI [1.9; 3.8]; $p = 1.3 \times 10^{-8}$) and
685 cardiac conduction disorders ($OR_{del} = 3.3$; 95%-CI [2.2; 4.9]; $p = 1.1 \times 10^{-8}$), suggesting a role
686 in cardiovascular health (Figure S4A-D). Primarily associated with developmental delay and
687 intellectual disability [74,75] – proxied by decreased fluid intelligence ($p_{t-test} = 8.7 \times 10^{-5}$) and
688 income ($p_{t-test} = 1.4 \times 10^{-12}$) in the UKBB (Figure S4E-F) – cardiac malformations are reported
689 in ~38% of clinically ascertained cases [76]. Among 193 UKBB deletion carriers, two (1%) had
690 congenital insufficiency of the aortic valve (Q23.1), corresponding to a higher but not
691 significantly different prevalence of cardiovascular malformations (Q20-28) than in copy-
692 neutral individuals ($OR_{Fisher} = 2.1$; $p = 0.251$). The deletion also associated with pneumonia
693 ($OR_{del} = 3.0$; 95%-CI [1.9; 4.6]; $p = 5.4 \times 10^{-7}$), coinciding with associations with decreased
694 forced vital capacity [23] (Figure S4G-H) and peak expiratory flow [19], together demonstrating
695 the clinical relevance of CNV-biomarker associations.

696

697 *Dissecting complex pleiotropic CNV regions*

698 While some CNV signals converge onto the same underlying physiological processes, others
699 tie apparently unrelated traits to the same genetic region, suggesting genuine pleiotropy.
700 16p13.11 harbors multiple, partially overlapping recurrent groups of CNVs allowing fine-
701 mapping of signals to different subregions of the CNVR (Figure 5). Through different
702 association models, the CNVR was linked to uncorrelated traits including epilepsy, kidney

703 stones, hypertension, alkaline phosphatase (ALP), forced vital capacity, and age at
 704 menopause and menarche. We previously proposed *MARF1* as a candidate gene for the
 705 female reproductive phenotypes [23] and will focus here on the remaining traits.



706 **Figure 5. Dissection of complex pleiotropic patterns of recurrent CNVs at 16p13.11**

707 **(A)** 16p13.11 genetic landscape. Coordinates of UK Biobank duplications (shades of blue; top) and
 708 deletions (shades of red; bottom) overlapping the maximal CNV region (CNVR delimited by vertical
 709 dashed lines) associated with epilepsy, kidney stones, hypertension, and alkaline phosphatase (ALP).
 710 CNVs are divided and colored according to five categories (cat1-5) to reflect recurrent breakpoints, with
 711 atypical CNVs in grey. Breakpoints reflect segmental duplications, represented with a grey gradient
 712 proportional to the degree of similarity. Middle: genomic coordinates of genes and DECIPHER CNV.
 713 Inset: Overlap between *ABCC6*'s exonic structure and cat5 deletions. **(B, D, F, H)** Negative logarithm
 714 of association p-values of CNVs with **(B)** epilepsy; **(D)** kidney stones; **(F)** hypertension; and **(H)** ALP
 715 (dark grey; model in parenthesis; CNVR delimited by vertical dashed lines) and SNPs with **(B)** epilepsy
 716 [77]; **(D)** kidney stones [78], calcium levels, and phosphate levels (y-axis; break: //); **(F)** essential
 717 hypertension and systolic blood pressure [79]; and **(H)** ALP. Lead SNPs are labeled. Red horizontal
 718 dashed lines represent genome-wide thresholds for significance for CNV-GWAS ($p \leq 7.5 \times 10^{-6}$) and
 719 SNP-GWAS ($p \leq 5 \times 10^{-8}$). **(C, E, G)** Prevalence (\pm standard error) of **(C)** epilepsy, **(E)** kidney stones,
 720 and **(G)** hypertension according to 16p13.11 copy-number (CN) and CNV categories from (A). P-values
 721 compare deletion (CN = 1) and duplication (CN = 3) carriers to copy-neutral (CN = 2) individuals (two-
 722 sided Fisher test). Number of cases and sample sizes are indicated (N = cases/sample size). **(I)** ALP
 723 levels according to 16p13.11 CN and CNV category, shown as boxplots; outliers are not shown. P-

724 values compare deletion (CN = 1) and duplication (CN = 3) carriers to copy-neutral (CN = 2) individuals
725 (two-sided t-test). Grey horizontal line represents median ALP value in non-carriers.

726 The 654 duplications and 355 deletions overlapping the maximal CNVR (chr16:15,070,916-
727 16,353,166) were grouped into 5 categories (cat1-5) based on their breakpoints ([Figure 5A](#);
728 [Supplemental Note 3](#)). Risk for epilepsy was increased in deletion carriers (chr16:15,122,801-
729 16,353,166; $OR_{del} = 6.2$; 95%-CI [2.8; 13.4]; $p = 4.4 \times 10^{-6}$; [Figure 5B](#)), with a prevalence of
730 8.3% among cat1-4 deletion carriers compared to less than 1.5% among copy-neutral and
731 duplication carriers ([Figure 5C](#)). Previously associated with epilepsy in clinical cohorts
732 [6,80,81], the region harbors *NDE1* [MIM: 609449], a gene associated with autosomal
733 recessive lissencephaly [MIM: 614019] and microhydranencephaly [MIM: 605013] and whose
734 mutation has been linked to epilepsy [82,83]. Deletions also increased risk for kidney stones
735 (chr16:15,120,501-16,353,166; $OR_{del} = 5.9$; 95%-CI [2.9; 11.9]; $p = 7.3 \times 10^{-7}$), with the CNV-
736 GWAS signal peaking close to a missense variant (rs41278174 G>A; Frequency_A: 2.1%) in
737 exon 23 of *ABCC6* [MIM: 603234] associating with calcium and phosphate levels through
738 SNP-GWAS ([Figure 5D](#)). These signals coincide with the recurrent cat5 deletion that covers
739 29 probes spanning exons 23-29 from *ABCC6* ([Figure 5A](#)). Kidney stones prevalence reaches
740 4.3% among cat5 deletion carriers, in-between estimates for larger cat1-4 deletion carriers
741 (9.3%) and copy-neutral individuals (2.3%) ([Figure 5E](#)). A wide range of variants affecting
742 *ABCC6* have been identified and linked to the calcification disorder pseudoxanthoma
743 elasticum through recessive [MIM: 264800] – and more rarely dominant [MIM: 177850] –
744 inheritance [84–87], with the *Alu*-mediated cat5 deletion representing one of the most frequent
745 variants [88,89]. *ABCC6* is expressed in the kidney and recent estimates from clinical cohorts
746 suggested that kidney stones are an unrecognized (i.e., not used to establish clinical
747 diagnosis) but prevalent (11-40%) feature of pseudoxanthoma elasticum [90–92]. Our data
748 support kidney stones as a clinical outcome of *ABCC6* disruption with partial gene deletions
749 leading to lower penetrance than large 16p13.11 deletions. Unlike epilepsy and kidney stones,
750 both deletion (39.2%) and duplication (42%) carriers are at increased risk for hypertension
751 (chr16:15,127,986-16,308,285; $OR_{U-shape} = 1.5$; 95%-CI [1.3; 1.8]; $p = 5.5 \times 10^{-6}$; [Figure 5F](#)),

752 compared to copy-neutral individuals (35.3%) (Figure 5G). The CNVR overlaps a SNP-GWAS
753 signal for systolic blood pressure mapping to *MYH11* [MIM: 160745] (Figure 5F). Expressed
754 in arteries, *MYH11* encodes for smooth muscle myosin heavy chains and has been linked to
755 dominant familial thoracic aortic aneurysm [MIM: 132900], for which hypertension represents
756 a leading risk factor. Increased prevalence (37.4%) of hypertension among cat5 deletions
757 implicates *ABCC6*, suggesting *cis*-epistasis for hypertension risk at 16p13.11. Consistent with
758 this model, *ABCC6* plays a role in vascular calcification as the causal gene for generalized
759 arterial calcification of infancy [MIM: 614473] [93,94], typically diagnosed by hypertension in
760 newborns. Interestingly, the previously described mirror association with ALP
761 (chr16:15,070,916-16,276,964; $\beta_{\text{mirror}} = 6.6$ U/L; $p = 3.5 \times 10^{-7}$) peaks at the distal end of the
762 CNVR [23], nearby a suggestive SNP-GWAS signal for ALP levels (Figure 5H). Splitting ALP
763 levels by CNV category revealed that this mirroring effect is driven by individuals with cat2
764 deletion (mean = 76.4 U/L; $p_{\text{t-test}} = 9.7 \times 10^{-3}$) and duplication (mean = 92.9 U/L; $p_{\text{t-test}} = 8.2 \times$
765 10^{-5}), as other CNV carriers had ALP levels indistinguishable from those of copy-neutral
766 individuals (mean = 83.6 U/L) (Figure 5I). Hence, we propose the distal region of the CNVR
767 to harbor the critical region regulating ALP levels, even if no obvious candidate gene could be
768 identified through literature review.

769

770 The proximal 22q11.2 region, previously linked to DiGeorge [MIM: 188400] and
771 velocardiofacial [MIM: 192430] syndromes, harbors four low-copy repeat (LCR; labeled A-to-
772 D) [95]. Building on evidence of complex association patterns with this CNVR [38], we report
773 novel associations between CNVs spanning LCR A-D and IHD (chr22:19,024,651-
774 21,463,545; $OR_{\text{U-shape}} = 2.1$; 95%-CI [1.6; 2.8]; $p = 1.5 \times 10^{-7}$), LCR B-D and aneurysm
775 (chr22:20,708,685-21,460,008; $OR_{\text{del}} = 41.8$; 95%-CI [10.0; 175.1]; $p = 3.2 \times 10^{-7}$), and LCR
776 A-C and headaches (chr22:19,024,651-21,110,240; $OR_{\text{mirror}} = 3.7$; 95%-CI [2.1; 6.5]; $p = 4.8$
777 $\times 10^{-6}$) (Figure S5A; Supplemental Note 3). Based on 3 LCR B-D deletion carriers with
778 aneurysm, this corresponds to a 22-times higher prevalence than in copy-neutral individuals
779 (Figure S5B). Association with IHD is better powered, with a prevalence of 12%, 21%, 16%,

780 and 20% among copy-neutral individuals and carriers of LCR C-D, B-D, and A-D CNVs,
781 respectively (Figure S5C). This suggests that IHD risk scales with the amount of affected
782 genetic content, supporting the presence of genetic driver(s) and/or modifier(s) in the C-D
783 interval, beyond the prime candidate *TBX1* [MIM: 602054] [95]. Collectively, our data indicates
784 that altered 22q11.2 dosage can result in a spectrum of cardiovascular afflictions of various
785 degrees of severity, ranging from well-described congenital malformation [95,96] to adult-
786 onset cardiovascular disorders.

787

788 15q13 deletions spanning BP4-5 [MIM: 612001] – and to a lesser extent duplications – have
789 been associated with neuropsychiatric and developmental conditions [97,98], with the nicotinic
790 acetylcholine receptor ion channel *CHRNA7* being proposed as the driver gene based on the
791 presence of similar phenotypes in individuals with a smaller deletion (D-*CHRNA7*-BP5) only
792 affecting *CHRNA7* [99] (Figure S6A). While BP4-5 duplication carriers showed higher
793 prevalence of AKI (EstBB-replicated: Figures 3B, S6B), hemorrhagic stroke
794 (chr15:30,912,719-31,982,408; $OR_{U\text{-shape}} = 7.5$; 95%-CI [3.2; 17.9]; $p = 4.3 \times 10^{-6}$; Figure S6C),
795 and anemia (chr15:30,912,719-31,094,479; $OR_{\text{dup}} = 4.9$; 95%-CI [2.5; 9.7]; $p = 3.2 \times 10^{-6}$;
796 Figure S6D), reminiscent of associations with pulse rate, mean corpuscular hemoglobin, and
797 red blood cell count [19,23], this was not the case for the ~10-times more numerous D-
798 *CHRNA7*-BP5 duplication carriers. Replicating an association with asthma [20]
799 (chr15:30,912,719-32,516,949; $OR_{\text{mirror}} = 0.17$; 95%-CI [0.08; 0.35]; $p = 1.2 \times 10^{-6}$) which
800 parallels decreased forced vital capacity [23] and peak expiratory flow [19], this was the only
801 deletion-driven signal for which the CNVR encompassed the entire BP4-5 region. However,
802 only BP4-5 (46.2%; $p_{\text{t-test}} = 1.8 \times 10^{-5}$) deletion carriers had higher asthma prevalence than
803 copy-neutral individuals (12.1%) (Figure S6E). Overall, the non-neurological disorders we
804 associate with 15q13 CNVs appear to specifically involve dosage of the genes within BP4-D-
805 *CHRNA7* and not *CHRNA7*.

806

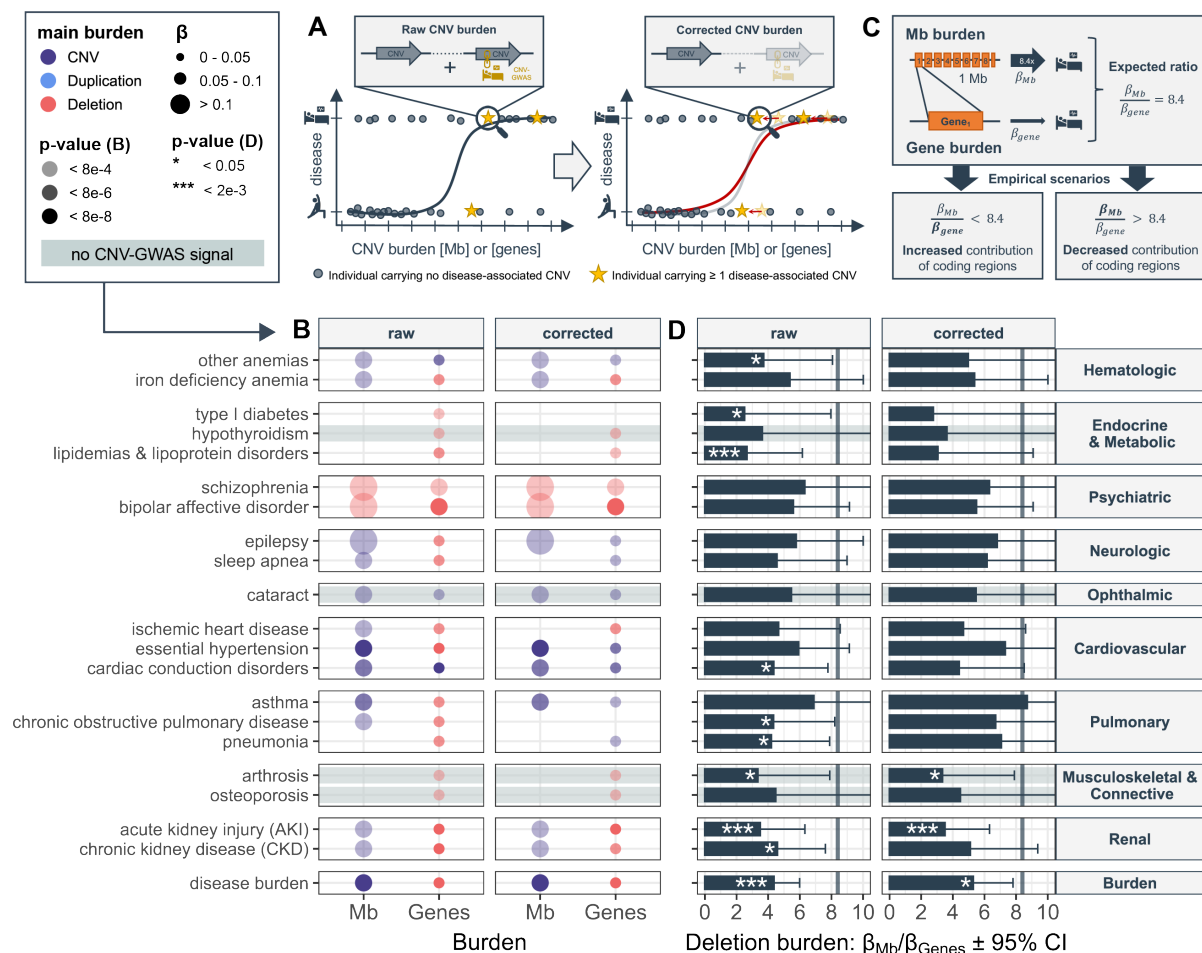
807 *Pathological consequences of an increased CNV burden*

808 By assessing the global pathogenic impact of CNVs we can capture the effect of ultra-rare
809 variants (frequency $\leq 0.01\%$), as well as those whose effect is not strong enough to reach GW
810 significance under current settings. Individual-level autosomal CNV (duplication + deletion),
811 duplication, and deletion burdens were calculated as the number of Mb or genes affected by
812 the considered type of variation and their predictive value on the same 60 diseases (and the
813 disease burden) previously assessed through CNV-GWAS was estimated (Figure 6A; left).
814 Disease burden strongly associated with a high CNV load (Mb: $\beta_{\text{CNV}} = +0.08$ disease/Mb; $p =$
815 3.7×10^{-27} | Gene: $\beta_{\text{del}} = +0.03$ disease/gene; $p = 3.7 \times 10^{-27}$) and risk for 20 individual disorders
816 was increased by at least one type of CNV burden ($p \leq 0.05/61 = 8.2 \times 10^{-4}$; Figure 6B; left;
817 Table S5). Strongest effect sizes were observed for the Mb burden for psychiatric disorders,
818 such as bipolar disorder ($\text{OR}_{\text{del}} = 1.4$; $p = 6.9 \times 10^{-4}$), schizophrenia ($\text{OR}_{\text{del}} = 1.4$; $p = 4.1 \times 10^{-$
819 5), or epilepsy ($\text{OR}_{\text{CNV}} = 1.1$; $p = 8.3 \times 10^{-5}$), in agreement with CNVs representing important
820 risk factors for these complex and polygenic disorders. To ensure that we do not merely
821 capture the effect of individual CNV-disease associations previously isolated by CNV-GWAS,
822 we excluded CNVs overlapping disease-associated CNVRs from the burden calculation and
823 re-estimated their predictive value (Figure 6A; right). Overall association strength dropped but
824 signal was lost only for type 1 diabetes and chronic obstructive pulmonary disease (Figure 6B;
825 right; Table S5), suggesting that CNVs not captured by the CNV-GWAS contribute to
826 modulating disease risk. Establishing the polygenic CNV architecture of a substantial number
827 of common diseases, this implies that increased power will likely lead to the discovery of
828 further CNV-disease associations.

829

830 As the “gene burden” captures more associations than the “Mb burden” and as the deletion
831 burden yields stronger associations than the duplication one (Figure 6B), we hypothesized
832 that the pathogenic effect of CNVs is mainly driven by the number of deleted genes. Based
833 on an average GW gene density of 8.4 genes/Mb, we assume the deletion burden measured
834 in Mb (β_{Mb}) to be 8.4 times larger than the one measured in number of affected genes (β_{gene}),

835 expecting a ratio of 8.4 if the type of affected genetic content does not impact disease risk
 836 (see [Methods](#); [Figure 6C](#)). Ratios significantly ($p \leq 0.05/21 = 2.4 \times 10^{-3}$) smaller or larger than
 837 8.4 indicate disproportionately large or small contribution of protein coding regions,
 838 respectively. We detected a larger than expected contribution of deleted genes to the disease
 839 burden (ratio = 4.4; $p = 1.3 \times 10^{-6}$), and AKI (ratio = 3.5; $p = 7.0 \times 10^{-4}$) and lipidemia (ratio =
 840 2.7; $p = 1.4 \times 10^{-3}$) risk, along with seven additional nominally decreased ratios ([Figure 6D](#);
 841 left; [Table S6](#)), confirming that haploinsufficient genes mediate the burden's effect on disease
 842 risk. Repeating the analysis with the duplication burden did not reveal any significant
 843 deviations ([Table S6](#)). After correcting for CNV-GWAS signals, only AKI ($p = 7.0 \times 10^{-4}$),
 844 arthritis ($p = 0.030$), and disease burden ($p = 0.016$) showed increased contribution of deleted
 845 genes on disease risk, indicating that a substantial fraction of effects capture by the CNV-
 846 GWAS are mediated by deletion of coding regions.



847 **Figure 6. The coding deletion burden increases overall disease risk**
848 (A) Burden calculation. Left: Raw CNV burden is calculated by summing up the length (in Mb or number
849 of affected genes) of all CNVs in an individual. Burden values are used as a predictor for disease risk.
850 Right: Corrected CNV burden is calculated by summing up the length of all CNVs that do not overlap
851 with a CNV region (CNVR) significantly associated with the investigated disease through CNV-GWAS.
852 Corrected burden values are used to re-estimate contribution of the CNV burden to disease risk (red
853 curve). (B) Contribution of raw and corrected CNV burdens in Mb or affected genes (x-axis) to disease
854 risk (y-axis). Only the most significantly associated burden, providing $p \leq 0.05/61 = 8.2 \times 10^{-4}$, is shown.
855 Color indicates whether the CNV (duplication + deletion), duplication, or deletion burden was most
856 significantly associated, with size and transparency being proportional to the effect size (beta) and p-
857 value, respectively. Grey horizontal bands mark traits with no CNV-GWAS signal. (C) Top: Based on
858 an average genome-wide gene density of 8.4 genes/Mb, we expect the effect of the CNV burden in Mb
859 to be 8.4x larger than the one measured in number of affected genes. Bottom: If the ratio is smaller
860 than 8.4, it indicates an increased contribution of coding regions. Alternatively, if larger than 8.4, it
861 indicates a decreased contribution of coding regions. (D) Ratio, with 95% confidence interval (CI), of
862 the raw and corrected Mb/genes deletion burdens (x-axis) for all diseases showing at least one
863 significant burden association. Significant deviation from the expected ratio of 8.4 (grey vertical line) is
864 indicated as *** for $p \leq 0.05/21 = 2.4e-3$ and * for $p \leq 0.05$ (two-sided one sample t-test).
865

866 DISCUSSION

867 Using an adapted GWAS framework, we provide a detailed investigation of the contribution of
868 CNVs to the genetic architecture of 60 common diseases and showcase how the rich
869 phenotypic data of the UKBB can be leveraged to gain new biological insights, highlighting the
870 role of CNVs as modulators of common disease susceptibility in the general population.

871
872 Various strategies have been used to study CNV-disease associations in the UKBB. Focusing
873 on diseases related to the ones assessed in the current study, we replicate 10 out of the 24
874 detected associations (at $FDR \leq 0.1$) with 54 likely pathogenic CNVs [20] and all four
875 associations (at $p \leq 1 \times 10^{-9}$) in a recent CNV-GWAS investigating 757 diseases [25]. Despite
876 analyzing the same dataset, we often obtained p-values orders of magnitude smaller (e.g.,
877 16p11.2 BP4-5 deletion and AKI: $p = 5.6 \times 10^{-20}$; $p_{Crawford} = 6.0 \times 10^{-5}$; $p_{Hujjoel} = 3.3 \times 10^{-15}$).
878 Besides accruing case count from updated hospital records, careful case-control definition
879 and statistical handling of the binary outcomes, probe-level association analysis, and usage
880 of different association models to mimic various dosage mechanisms could explain the
881 increased power of our study. We consequently identified previously unreported CNV-disease
882 associations whose relevance was asserted by follow-up analyses. Only one signal – 17q12

883 CNVs increasing CKD risk – was backed by all approaches, while others were supported by
884 some analysis but not by others – e.g., two associations in the lowest statistical confidence
885 tier were replicated in the EstBB and harbored plausible candidate genes – emphasizing the
886 importance of considering diverse lines of corroborative evidence. Substantial overlap with
887 relevant SNP-GWAS signals and OMIM genes indicates shared genetic mechanisms with
888 convergent effects on the phenotype. Another line of evidence is coincidence with a disease-
889 relevant biomarker CNV signals, e.g., four (1q21.1-2, 15q13, 16p12.2, 16p11.2 BP4-5) out of
890 six CNVRs decreasing forced vital capacity [23] were found to increase risk for pulmonary
891 diseases. This demonstrates that biomarkers are efficient proxies underlying (CNV-driven)
892 pathological processes, often increasing the statistical power to detect associations due to
893 their continuous nature. To render binary outcomes continuous, we regressed covariates out
894 of the disease status. With the same goal, more sophisticated approaches have recently been
895 developed that transform binary outcomes in continuous liability scores while borrowing
896 information from age-at-disease onset, sex, and familial history [100]. Successfully applied to
897 SNP-based GWASs, future exploration is warranted to assess their benefit in the context of
898 CNV-GWASs. By coupling a CNV-GWAS framework designed to account for challenges
899 linked to disease CNV association studies in population cohorts to extensive validation, we
900 generated a list of 73 CNV-disease pairs with various levels of supporting evidence that can
901 inform follow-up studies.

902

903 Disease-associated CNVRs harbored genes under stronger evolutionary constraint than
904 those lacking associations and their length correlated with their propensity for pleiotropy,
905 indicating that as previously observed [8], both the number and the nature of genes affected
906 by CNVs influence their pathogenicity. Consequently, large multi-gene recurrent CNVs
907 exhibited the strongest pleiotropy. A longstanding question relates to the identification of
908 causal genes whose altered dosage drives the phenotypic alterations observed in carriers.
909 Models with various levels of complexity have been proposed, ranging from a single driver
910 gene to multiple driver genes modulated by epistatic interactions with other genes in the CNVR

911 [101]. By analyzing disease prevalence in subsets of CNV carriers, association signals could
912 be fine-mapped to narrower regions, pinpointing candidate drivers - such as *ABCC6* for kidney
913 stones. In other cases, our data suggests that multiple subregions of the CNVR contribute to
914 increased risk for a given disease by *cis*-epistasis, as observed for 22q11.2 and ischemic
915 heart disease or 16p13.11 and hypertension. Interestingly, the putative driver for phenotypes
916 originally associated with a CNVR might not be driving our newly identified associations, as
917 shown for the 15q13 CNVR, whose non-neurological phenotypes do not appear to be linked
918 to altered dosage of *CHRNA7*. Beyond characterizing the pleiotropic pathological
919 consequences of recurrent CNVRs, we demonstrate that dissection of CNV-GWAS signals
920 can fine-map associations and provide mechanistic insights into their phenotypic expression.

921

922 All CNVs increased disease risk and led to an earlier age at onset. Incorporating age at onset
923 information has been shown to improve power to detect associations [100], and more
924 importantly, represents the ultimate proof of clinical relevance. Indeed, many signals mapped
925 to regions whose genetic perturbation has been reported to be pathogenic. These include
926 associations between well-described, clinically relevant gene-disease pairs – such as *BRCA1*
927 and *LDLR* deletions increasing the risk for ovarian cancer and IHD, respectively – but for which
928 the role of CNVs in a large population cohort had not been previously investigated. CNVs in
929 these genes have high penetrance but are extremely rare in the UKBB, so that association
930 barely reached GW significance. Follow-up analyses based on the medical records, family
931 history, medication use, and biomarkers could recapitulate additional clinical associations and
932 establish that these deletions were most likely inherited, thereby generating insights into their
933 role in the general population. We further show that many CNVRs previously linked to pediatric
934 genomic disorders also increased risk for a broad spectrum of adult-onset common diseases.
935 These associations were probably overlooked as the medical consequences in adulthood of
936 these etiologies is often poorly characterized owing to ascertainment bias and difficulty to
937 gather large cohorts. While awaiting validation in clinical cohorts of CNV carriers, we hope
938 that these findings will improve clinical characterization of genomic disorders, thereby

939 facilitating diagnosis and allowing physicians to anticipate later-onset comorbidities. For
940 instance, we found carriers of 16p13.11 deletions affecting *ABCC6*, the causal gene for
941 pseudoxanthoma elasticum, to be at increased risk for kidney stones, paralleling reports from
942 clinical cohorts showing that kidney stones represent an unrecognized feature of the disease
943 [90–92]. Awareness of this disease feature can mitigate kidney stone risk through adapted
944 diet and sufficient water intake. Together, our results advocate for a complex model of variable
945 CNV expressivity and penetrance that can result in a broad range of phenotypes along the
946 rare-to-common disease spectrum and represent fertile ground for in-depth, phenome-wide
947 studies aiming at better characterizing specific CNV regions [38].

948

949 Corroborating the deleterious impact of CNVs on an individual's health parameters, socio-
950 economic status, and lifespan [18,22,23,25,31,102–105], we here speculate that it acts by
951 increasing risk for a broad range of common diseases beyond their known role in
952 neuropsychiatric disorders [4–7]. While both duplications and deletions contributed to
953 increased disease risk, the deletion burden's impact was much stronger – especially for
954 metabolic, psychiatric, and musculoskeletal diseases – and for about half the diseases,
955 predominantly driven by deletion of coding region, in line with the commonly accepted view
956 that deletions tend to be more deleterious than duplications and that genetic variation affecting
957 coding regions have stronger phenotypic consequences. Only a marginal fraction of the CNV
958 burden's contribution to disease risk was captured by disease-associated CNVRs and risk for
959 four diseases, i.e., hypothyroidism, cataract, arthrosis, and osteoporosis, associated with the
960 burden despite lacking CNV-GWAS signals. With over 10,000 cases, these diseases were not
961 underpowered compared to others, suggesting genuine differences in the genetic architecture,
962 and illustrating the added value of burden tests. Collectively, these results predict that better
963 powered CNV-GWAS are likely to isolate further associations currently captured by the CNV
964 burden.

965

966 A major limitation of our study is the reliance on microarray CNV calls, which allows to assess
967 only a fraction of the CNV landscape, i.e., mostly large CNVs or in regions with high probe
968 coverage. We speculate that small and/or multiallelic CNVs that can only be uncovered by
969 sequencing, will have a genetic architecture closer to the one of SNPs and indels, with higher
970 frequencies and more subtle effect sizes, resembling those of SNPs and indels. These effects,
971 however, are more likely tagged by common variants, limiting novel discoveries. Furthermore,
972 by detecting more events, sequencing-based studies require adapted and more stringent
973 significance thresholds. Still, having improved breakpoint resolution, such CNV calls are also
974 likely to enhance fine-mapping strategies. Microarray CNV calls also exhibit high false positive
975 rates [47]. By using stringent CNV selection criteria, we decrease the latter at the cost of
976 decreasing power to detect true associations. This aspect is particularly relevant given that
977 the type of CNVs we assess are rare and that the UKBB is depleted for disease cases [34],
978 resulting in low powered GWASs. While we adopt strategies to counter the lack of power, our
979 results are likely subject to Winner's curse, only capturing a fraction of the strongest, possibly
980 overestimated effects. This phenomenon might be compensated by UKBB CNV carriers being
981 at the milder end of the clinical spectrum, leading to effect underestimation. An interesting
982 question will be to compare effect sizes from population-based studies to those emerging from
983 clinical cohort. In the future, longitudinal follow-up of UKBB participants will increase the
984 number of cases – especially for late-onset diseases such as Alzheimer's or Parkinson'
985 diseases – allowing better powered CNV-GWASs. Alternatively, meta-analyses can boost
986 case number through inclusion of clinical cohorts, at the cost of poorer disease definition due
987 to imperfect data harmonization [8]. Larger and more diverse biobanks linking genotype to
988 phenotype data [106–108] should both validate reported associations and identify new ones.

989

990 **CONCLUSIONS**

991 Our study provides in-depth analysis of the role of CNVs in modulating susceptibility to 60
992 common diseases in the general population, broadening our view on how this mutational class

993 impacts human health. Besides describing clinically relevant and actionable associations, we
994 illustrate how complex pleiotropic patterns can be dissected to gain new insights into the
995 pathological mechanisms of large recurrent CNVs, providing a framework that can be applied
996 to an even larger spectrum of diseases.

997

998 SUPPLEMENTAL DATA

999 Supplemental data include 6 figures, 6 tables, and 3 notes.

1000

1001 DECLARATIONS

1002 **Availability of data and materials:** All data used in this study are publicly available, as
1003 described in the methods. CNV-GWAS summary statistics (UKBB) will be deposited on the
1004 GWAS Catalog upon publication and are available upon request until then.

1005

1006 **Competing interests:** SO is an employee of MSD at the time of the submission; contribution
1007 to the research occurred during the affiliation at the University of Lausanne.

1008

1009 **Funding:** The study was funded by the Swiss National Science Foundation (31003A_182632,
1010 AR; 310030_189147, ZK), Horizon2020 Twinning projects (ePerMed 692145, AR), the
1011 Estonian Research Council (PRG687, MJ and RM), and the Department of Computational
1012 Biology (ZK) and the Center for Integrative Genomics (AR) from the University of Lausanne.

1013

1014 **Author's contributions:** CA, AR and ZK conceived the study; CA carried out the analyses
1015 with contributions from MCS, NT, and CJC; The Estonian Biobank Research Team
1016 coordinated genotyping and sequencing data acquisition in the EstBB; MJ performed the
1017 replication study in the EstBB under the supervision of RM; ZK supervised statistical analyses;
1018 SO provided guidance for time-to-event analysis; CA drafted the manuscript and generated

1019 the figures; AR and ZK made critical revisions; All authors read, approved, and provided
1020 feedback on the final manuscript.

1021

1022 **Acknowledgments:** We thank all biobank participants for sharing their data. UKBB and
1023 EstBB computations were performed on the JURA server (University of Lausanne) and the
1024 High-Performance Computing Center (University of Tartu), respectively.

REFERENCES

- 1025 1. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of
1026 structural variation in 2,504 human genomes. *Nature*. 2015;526:75–81.
- 1027 2. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy
1028 number variation in the human genome. *Nature*. 2010;464:704–12.
- 1029 3. Zhang F, Gu W, Hurler ME, Lupski JR. Copy Number Variation in Human Health, Disease, and Evolution. *Annu*
1030 *Rev Genomics Hum Genet*. 2009;10:451–81.
- 1031 4. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-martin C, Walsh T, et al. Strong Association of De Novo Copy
1032 Number Mutations with Autism. *Science* (1979). 2007;316:445–9.
- 1033 5. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, et al. Rare structural variants
1034 disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* (1979). 2008;320:539–43.
- 1035 6. Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, Baker C, et al. Genome-Wide Copy Number
1036 Variation in Epilepsy: Novel Susceptibility Loci in Idiopathic Generalized and Focal Epilepsies. *PLoS Genet*.
1037 2010;6:e1000962.
- 1038 7. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map
1039 of developmental delay. *Nat Genet*. 2011;43:838–46.
- 1040 8. Collins RL, Glessner JT, Porcu E, Lepamets M, Brandon R, Lauricella C, et al. A cross-disorder dosage
1041 sensitivity map of the human genome. *Cell*. 2022;185:3041-3055.e25.
- 1042 9. Firth H V., Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal
1043 Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet*. 2009;84:524–33.
- 1044 10. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev*
1045 *Genet*. 2016;17:224–38.
- 1046 11. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for
1047 medical and population genetics. *Nature*. 2020;581:444–51.
- 1048 12. Abel HJ, Larson DE, Chiang C, Das I, Kanchi KL, Layer RM, et al. Mapping and characterization of structural
1049 variation in 17,795 deeply sequenced human genomes. *Nature*. 2020;583:83–9.
- 1050 13. Halvorsen M, Huh R, Oskolkov N, Wen J, Netotea S, Giusti-Rodriguez P, et al. Increased burden of ultra-rare
1051 structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nat Commun*.
1052 2020;11:1–13.
- 1053 14. Chen L, Abel HJ, Das I, Larson DE, Ganel L, Kanchi KL, et al. Association of structural variation with
1054 cardiometabolic traits in Finns. *Am J Hum Genet*. 2021;108:583–96.
- 1055 15. Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, et al. Long-read
1056 sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other
1057 traits. *Nat Genet*. 2021;53:779–86.
- 1058 16. Babadi M, Fu JM, Lee SK, Smirnov AN, Gauthier LD, Walker M, et al. GATK-gCNV: A Rare Copy Number
1059 Variant Discovery Algorithm and Its Application to Exome Sequencing in the UK Biobank. *bioRxiv*.
1060 2022;2022.08.25.504851.
- 1061 17. Fitzgerald T, Birney E. CNest: A novel copy number association discovery method uncovers 862 new
1062 associations from 200,629 whole-exome sequence datasets in the UK Biobank. *Cell Genomics*.
1063 2022;2:100167.

- 1064 18. Kendall KM, Rees E, Escott-Price V, Einon M, Thomas R, Hewitt J, et al. Cognitive Performance Among
1065 Carriers of Pathogenic Copy Number Variants: Analysis of 152,000 UK Biobank Subjects. *Biol Psychiatry*.
1066 2017;82:103–10.
- 1067 19. Owen D, Bracher-Smith M, Kendall KM, Rees E, Einon M, Escott-Price V, et al. Effects of pathogenic CNVs on
1068 physical traits in participants of the UK Biobank. *BMC Genomics*. 2018;19:1–9.
- 1069 20. Crawford K, Bracher-Smith M, Owen D, Kendall KM, Rees E, Pardiñas AF, et al. Medical consequences of
1070 pathogenic CNVs in adults: Analysis of the UK Biobank. *J Med Genet*. 2019;56:131–8.
- 1071 21. Kendall KM, Rees E, Bracher-Smith M, Legge S, Riglin L, Zammit S, et al. Association of Rare Copy Number
1072 Variants with Risk of Depression. *JAMA Psychiatry*. 2019;76:818–25.
- 1073 22. Macé A, Tuke MA, Deelen P, Kristiansson K, Mattsson H, Nõukas M, et al. CNV-association meta-analysis in
1074 191,161 European adults reveals new loci associated with anthropometric traits. *Nat Commun*. 2017;8:1–11.
- 1075 23. Auwerx C, Lepamets M, Sadler MC, Patxot M, Stojanov M, Baud D, et al. The individual and global impact of
1076 copy-number variants on complex human traits. *Am J Hum Genet*. 2022;109:647–68.
- 1077 24. Aguirre M, Rivas MA, Priest J. Phenome-wide Burden of Copy-Number Variation in the UK Biobank. *Am J Hum*
1078 *Genet*. 2019;105:373–83.
- 1079 25. Hujoel MLA, Sherman MA, Barton AR, Mukamel RE, Sankaran VG, Terao C, et al. Influences of rare copy-
1080 number variation on human complex traits. *Cell*. 2022;185:4233–4248.e27.
- 1081 26. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, et al. Genetics of 35 blood and urine
1082 biomarkers in the UK Biobank. *Nat Genet*. 2021;53:185–94.
- 1083 27. Li YR, Glessner JT, Coe BP, Li J, Mohebnasab M, Chang X, et al. Rare copy number variants in over 100,000
1084 European ancestry subjects reveal multiple disease associations. *Nat Commun*. 2020;11:1–9.
- 1085 28. Kopal J, Kumar K, Saltoun K, Modenato C, Moreau CA, Martin-Brevet S, et al. Rare CNVs and phenome-wide
1086 profiling highlight brain structural divergence and phenotypical convergence. *Nat Hum Behav*. 2023;
- 1087 29. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep
1088 phenotyping and genomic data. *Nature*. 2018;562:203–9.
- 1089 30. Wright CF, West B, Tuke M, Jones SE, Patel K, Laver TW, et al. Assessing the Pathogenicity, Penetrance, and
1090 Expressivity of Putative Disease-Causing Variants in a Population Setting. *Am J Hum Genet*. 2019;104:275–
1091 86.
- 1092 31. Kingdom R, Tuke M, Wood A, Beaumont RN, Frayling TM, Weedon MN, et al. Rare genetic variants in genes
1093 and loci linked to dominant monogenic developmental disorders cause milder related phenotypes in the general
1094 population. *Am J Hum Genet*. 2022;109:1308–16.
- 1095 32. Chen R, Shi L, Hakenberg J, Naughton B, Sklar P, Zhang J, et al. Analysis of 589,306 genomes identifies
1096 individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol*. 2016;34:531–8.
- 1097 33. Goodrich JK, Singer-Berk M, Son R, Sveden A, Wood J, England E, et al. Determinants of penetrance and
1098 variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nat Commun*. 2021;12:1–15.
- 1099 34. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic
1100 and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *Am J*
1101 *Epidemiol*. 2017;186:1026–34.
- 1102 35. Falconer D. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann*
1103 *Hum Genet*. 1965;29:51–76.
- 1104 36. Senn Stephen S. *Statistical issues in drug development*. Wiley; 2021.
- 1105 37. Mollon J, Schultz LM, Hugué G, Knowles EEM, Mathias SR, Rodrigue A, et al. Impact of Copy Number Variants
1106 and Polygenic Risk Scores on Psychopathology in the UK Biobank. *Biol Psychiatry*. 2023;
- 1107 38. Zamarioli M, Auwerx C, Sadler MC, Graaf A Van Der, Lepik K, Schoeler T, et al. The impact of 22q11.2 copy-
1108 number variants on human traits in the general population. *Am J Hum Genet*. 2023;110:1–14.
- 1109 39. Giannuzzi G, Schmidt PJ, Porcu E, Willemin G, Munson KM, Nuttle X, et al. The Human-Specific BOLA2
1110 Duplication Modifies Iron Homeostasis and Anemia Predisposition in Chromosome 16p11.2 Autism Individuals.
1111 *Am J Hum Genet*. 2019;105:947–58.
- 1112 40. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, et al. Cohort profile: Estonian biobank of
1113 the Estonian genome center, university of Tartu. *Int J Epidemiol*. 2015;44:1137–47.
- 1114 41. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM Codes to
1115 Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform*. 2019;7:e14325.
- 1116 42. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS Catalog: knowledgebase
1117 and deposition resource. *Nucleic Acids Res*. 2023;51:D977–85.
- 1118 43. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM),
1119 a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33.
- 1120 44. Karczewski KJ., Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint
1121 spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434–43.

- 1122 45. Aguet F, Barbeira A, Bonazzola R, Brown A, Castel S, Jo B, et al. The GTEx Consortium atlas of genetic
1123 regulatory effects across human tissues. *Science* (1979). 2020;369:1318–30.
- 1124 46. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, et al. PennCNV: An integrated hidden Markov model
1125 designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome*
1126 *Res.* 2007;17:1665–74.
- 1127 47. Macé A, Tuke MA, Beckmann JS, Lin L, Jacquemont S, Weedon MN, et al. New quality measure for SNP array
1128 based CNV detection. *Bioinformatics*. 2016;32:3298–305.
- 1129 48. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the
1130 challenge of larger and richer datasets. *Gigascience*. 2015;4:1–16.
- 1131 49. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput
1132 sequencing data. *Nucleic Acids Res.* 2010;38:e164–e164.
- 1133 50. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at
1134 UCSC. *Genome Res.* 2002;12:996–1006.
- 1135 51. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using
1136 correlated single nucleotide polymorphisms. *Genet Epidemiol.* 2008;32:361–9.
- 1137 52. Therneau TM. Survival Analysis [R package survival version 3.5-3]. 2022 [cited 2023 Apr 17]; Available from:
1138 <https://CRAN.R-project.org/package=survival>
- 1139 53. Berman JJ. Rare Diseases and Orphan Drugs: Keys to Understanding and Treating the Common Diseases.
1140 Elsevier Inc.; 2014.
- 1141 54. Li A, Jiao X, Munier FL, Schorderet DF, Yao W, Iwata F, et al. Bietti Crystalline Corneoretinal Dystrophy Is
1142 Caused by Mutations in the Novel Gene CYP4V2. *Am J Hum Genet.* 2004;74:817–26.
- 1143 55. Zhou W, Otto EA, Cluckey A, Airik R, Hurd TW, Chaki M, et al. FAN1 mutations cause karyomegalic interstitial
1144 nephritis, linking chronic kidney failure to defective DNA damage repair. *Nat Genet.* 2012;44:910–5.
- 1145 56. Kryger MH, Roth T, Goldstein CA. Principles and Practice of Sleep Medicine. Elsevier Health Sciences; 2021.
- 1146 57. Shinawi M, Liu P, Kang SHL, Shen J, Belmont JW, Scott DA, et al. Recurrent reciprocal 16p11.2
1147 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy,
1148 and abnormal head size. *J Med Genet.* 2010;47:332–41.
- 1149 58. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, et al. Association between Microdeletion and
1150 Microduplication at 16p11.2 and Autism. *N Engl J Med.* 2008;358:667–75.
- 1151 59. D'Angelo D, Lebon S, Chen Q, Martin-Brevet S, Snyder LAG, Hippolyte L, et al. Defining the Effect of the
1152 16p11.2 Duplication on Cognition, Behavior, and Medical Comorbidities. *JAMA Psychiatry.* 2016;73:20–30.
- 1153 60. Reinthaler EM, Lal D, Lebon S, Hildebrand MS, Dahl HHM, Regan BM, et al. 16p11.2 600 kb Duplications
1154 confer risk for typical and atypical Rolandic epilepsy. *Hum Mol Genet.* 2014;23:6069–80.
- 1155 61. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, et al. Mirror extreme BMI
1156 phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature.* 2011;478:97–102.
- 1157 62. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, et al. Microduplications of 16p11.2 are
1158 associated with schizophrenia. *Nat Genet.* 2009;41:1223–7.
- 1159 63. Walters RG, Jacquemont S, Valsesia A, De Smith AJ, Martinet D, Andersson J, et al. A new highly penetrant
1160 form of obesity due to deletions on chromosome 16p11.2. *Nature.* 2010;463:671–5.
- 1161 64. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the
1162 breast and ovarian cancer susceptibility gene BRCA1. *Science* (1979). 1994;266:66–71.
- 1163 65. Iacocca MA, Hegele RA. Role of DNA copy number variation in dyslipidemias. *Curr Opin Lipidol.* 2018;29:125–
1164 32.
- 1165 66. Hobbs HH, Russell DW, Brown MS, Goldstein JL. The LDL receptor locus in familial hypercholesterolemia:
1166 mutational analysis of a membrane protein. *Annu Rev Genet.* 1990;24:133–70.
- 1167 67. Defesche JC, Gidding SS, Harada-Shiba M, Hegele RA, Santos RD, Wierzbicki AS. Familial
1168 hypercholesterolaemia. *Nat Rev Dis Primers.* 2017;3:17093.
- 1169 68. Iacocca MA, Wang J, Dron JS, Robinson JF, McIntyre AD, Cao H, et al. Use of next-generation sequencing to
1170 detect LDLR gene copy number variation in familial hypercholesterolemia. *J Lipid Res.* 2017;58:2202–9.
- 1171 69. Mach F, Baigent C, Catapano AL, Koskinas KC, Casula M, Badimon L, et al. 2019 ESC/EAS Guidelines for
1172 the management of dyslipidaemias: lipid modification to reduce cardiovascular riskThe Task Force for the
1173 management of dyslipidaemias of the European Society of Cardiology (ESC) and European Atherosclerosis
1174 Society (EAS). *Eur Heart J.* 2020;41:111–88.
- 1175 70. Wuttke M, Li Y, Li M, Sieber KB, Feitosa MF, Gorski M, et al. A catalog of genetic loci associated with kidney
1176 function from analyses of a million individuals. *Nat Genet.* 2019;51:957–72.
- 1177 71. Stanzick KJ, Li Y, Schlosser P, Gorski M, Wuttke M, Thomas LF, et al. Discovery and prioritization of variants
1178 and genes for kidney function in >1.2 million individuals. *Nat Commun.* 2021;12:1–17.

- 1179 72. Fajans SS, Bell GI, Polonsky KS. Molecular mechanisms and clinical pathophysiology of maturity-onset
1180 diabetes of the young. *N Engl J Med*. 2001;345:971-80.
- 1181 73. Mefford HC, Clauin S, Sharp AJ, Moller RS, Ullmann R, Kapur R, et al. Recurrent Reciprocal Genomic
1182 Rearrangements of 17q12 Are Associated with Renal Disease, Diabetes, and Epilepsy. *Am J Hum Genet*.
1183 2007;81:1057-69.
- 1184 74. Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, et al. A recurrent 16p12.1
1185 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet*. 2010;42:203-9.
- 1186 75. Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S, et al. CNVs
1187 conferring risk of autism or schizophrenia affect cognition in controls. *Nature*. 2013;505:361-6.
- 1188 76. Girirajan S, Pizzo L, Moeschler J, Rosenfeld J. 16p12.2 Recurrent Deletion. *GeneReviews*® [Internet]. 2018
1189 [cited 2023 Apr 4]; Available from: <https://www.ncbi.nlm.nih.gov/books/NBK274565/>
- 1190 77. Stevelink R, Campbell C, Chen S, The International League Against Epilepsy Consortium on Complex
1191 Epilepsies. Genome-wide meta-analysis of over 29,000 people with epilepsy reveals 26 loci and subtype-
1192 specific genetic architecture. *medRxiv*. 2022;2022.06.08.22276120.
- 1193 78. Howles SA, Wiberg A, Goldsworthy M, Bayliss AL, Gluck AK, Ng M, et al. Genetic variants of calcium and
1194 vitamin D metabolism in kidney stone disease. *Nat Commun*. 2019;10:1-10.
- 1195 79. Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, et al. Genetic analysis of over 1
1196 million people identifies 535 new loci associated with blood pressure traits. *Nat Genet*. 2018;50:1412-25.
- 1197 80. De Kovel CGF, Trucks H, Helbig I, Mefford HC, Baker C, Leu C, et al. Recurrent microdeletions at 15q11.2 and
1198 16p13.11 predispose to idiopathic generalized epilepsies. *Brain*. 2010;133:23-32.
- 1199 81. Heinzen EL, Radtke RA, Urban TJ, Cavalleri GL, Depondt C, Need AC, et al. Rare Deletions at 16p13.11
1200 Predispose to a Diverse Spectrum of Sporadic Epilepsy Syndromes. *Am J Hum Genet*. 2010;86:707-18.
- 1201 82. Alkuraya FS, Cai X, Emery C, Mochida GH, Al-Dosari MS, Felie JM, et al. Human Mutations in NDE1 Cause
1202 Extreme Microcephaly with Lissencephaly. *Am J Hum Genet*. 2011;88:536-47.
- 1203 83. Bakircioglu M, Carvalho OP, Khurshid M, Cox JJ, Tuysuz B, Barak T, et al. The Essential Role of Centrosomal
1204 NDE1 in Human Cerebral Cortex Neurogenesis. *Am J Hum Genet*. 2011;88:523-35.
- 1205 84. Ringpfeil F, Lebwohl MG, Christiano AM, Uitto J. Pseudoxanthoma elasticum: Mutations in the MRP6 gene
1206 encoding a transmembrane ATP-binding cassette (ABC) transporter. *Proc Natl Acad Sci U S A*. 2000;97:6001-
1207 6.
- 1208 85. Struk B, Cai L, Zäch S, Ji W, Chung J, Lumsden A, et al. Mutations of the gene encoding the transmembrane
1209 transporter protein ABC-C6 cause pseudoxanthoma elasticum. *J Mol Med*. 2000;78:282-6.
- 1210 86. Le Saux O, Urban Z, Tschuch C, Csiszar K, Bacchelli B, Quaglino D, et al. Mutations in a gene encoding an
1211 ABC transporter cause pseudoxanthoma elasticum. *Nat Genet*. 2000;25:223-7.
- 1212 87. Bergen AAB, Plomp AS, Schuurman EJ, Terry S, Breuning M, Dauwerse H, et al. Mutations in ABCC6 cause
1213 pseudoxanthoma elasticum. *Nat Genet*. 2000;25:228-31.
- 1214 88. Le Saux O, Beck K, Sachsinger C, Silvestri C, Treiber C, Goöring HHH, et al. A Spectrum of ABCC6 Mutations
1215 Is Responsible for Pseudoxanthoma Elasticum. *Am J Hum Genet*. 2001;69:749-64.
- 1216 89. Ringpfeil F, Nakano A, Uitto J, Pulkkinen L. Compound Heterozygosity for a Recurrent 16.5-kb Alu-Mediated
1217 Deletion Mutation and Single-Base-Pair Substitutions in the ABCC6 Gene Results in Pseudoxanthoma
1218 Elasticum. *Am J Hum Genet*. 2001;68:642-52.
- 1219 90. Ralph D, Allawh R, Terry IF, Terry SF, Uitto J, Li QL. Kidney Stones Are Prevalent in Individuals with
1220 Pseudoxanthoma Elasticum, a Genetic Ectopic Mineralization Disorder. *Int J Dermatol Venereol*. 2020;3:198-
1221 204.
- 1222 91. Legrand A, Cornez L, Samkari W, Mazzella JM, Venisse A, Boccio V, et al. Mutation spectrum in the ABCC6
1223 gene and genotype-phenotype correlations in a French cohort with pseudoxanthoma elasticum. *Genetics in
1224 Medicine*. 2017;19:909-17.
- 1225 92. Letavernier E, Kauffenstein G, Huguet L, Navasiolava N, Boudierlique E, Tang E, et al. ABCC6 deficiency
1226 promotes development of randall plaque. *Journal of the American Society of Nephrology*. 2018;29:2337-47.
- 1227 93. Nitschke Y, Baujat G, Botschen U, Wittkamp T, Du Moulin M, Stella J, et al. Generalized Arterial Calcification
1228 of Infancy and Pseudoxanthoma Elasticum Can Be Caused by Mutations in Either ENPP1 or ABCC6. *Am J
1229 Hum Genet*. 2012;90:25-39.
- 1230 94. Le Boulanger G, Labrèze C, Croué A, Schurgers LJ, Chassaing N, Wittkamp T, et al. An unusual severe
1231 vascular case of pseudoxanthoma elasticum presenting as generalized arterial calcification of infancy. *Am J
1232 Med Genet A*. 2010;152A:118-23.
- 1233 95. McDonald-McGinn DM, Sullivan KE, Marino B, Philip N, Swillen A, Vorstman JAS, et al. 22q11.2 deletion
1234 syndrome. *Nat Rev Dis Primers*. 2015;1:1-19.
- 1235 96. Bartik LE, Hughes SS, Tracy M, Feldt MM, Zhang L, Arganbright J, et al. 22q11.2 duplications: Expanding the
1236 clinical presentation. *Am J Med Genet A*. 2022;188:779-87.

- 1237 97. Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, et al. A recurrent 15q13.3 microdeletion
1238 syndrome associated with mental retardation and seizures. *Nat Genet.* 2008;40:322–8.
- 1239 98. Lowther C, Costain G, Stavropoulos DJ, Melvin R, Silversides CK, Andrade DM, et al. Delineating the 15q13.3
1240 microdeletion phenotype: a case series and comprehensive review of the literature. *Genetics in Medicine.*
1241 2015;17:149–57.
- 1242 99. Gillentine MA, Schaaf CP. The Human Clinical Phenotypes of Altered CHRNA7 Copy Number. *Biochem*
1243 *Pharmacol.* 2015;97:352.
- 1244 100. Pedersen EM, Agerbo E, Plana-Ripoll O, Grove J, Dreier JW, Musliner KL, et al. Accounting for age of onset
1245 and family history improves power in genome-wide association studies. *Am J Hum Genet.* 2022;109:417–32.
- 1246 101. Golzio C, Katsanis N. Genetic architecture of reciprocal CNVs. *Curr Opin Genet Dev.* 2013;23:240–8.
- 1247 102. Männik K, Mägi R, Macé A, Cole B, Guyatt AL, Shihab HA, et al. Copy number variations and cognitive
1248 phenotypes in unselected populations. *JAMA.* 2015;313:2044–54.
- 1249 103. Wheeler E, Huang N, Bochukova EG, Keogh JM, Lindsay S, Garg S, et al. Genome-wide SNP and CNV
1250 analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nat Genet.*
1251 2013;45:513–7.
- 1252 104. Dauber A, Yu Y, Turchin MC, Chiang CW, Meng YA, Demerath EW, et al. Genome-wide association of copy-
1253 number variation reveals an association between short stature and the presence of low-frequency genomic
1254 deletions. *Am J Hum Genet.* 2011;89:751–9.
- 1255 105. Saarentaus EC, Havulinna AS, Mars N, Ahola-Olli A, Kiiskinen TTJ, Partanen J, et al. Polygenic burden has
1256 broader impact on health, cognition, and socioeconomic outcomes than most rare and high-risk copy number
1257 variants. *Mol Psychiatry.* 2021;26:4884–95.
- 1258 106. Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, et al. The ‘All of Us’ Research
1259 Program. *N Engl J Med.* 2019;381:668–76.
- 1260 107. Hunter-Zinck H, Shi Y, Li M, Gorman BR, Ji SG, Sun N, et al. Genotyping Array Design and Data Quality
1261 Control in the Million Veteran Program. *Am J Hum Genet.* 2020;106:535–48.
- 1262 108. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights
1263 from a well-phenotyped isolated population. *Nature.* 2023;613:508–18.