

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

1 **The impact of 22q11.2 copy number variants on human traits in the general**  
2 **population.**

3 Malú Zamariolli<sup>1,2</sup>, Chiara Auwerx<sup>2,3,4,5</sup>, Marie C Sadler<sup>2,3,4</sup>, Adriaan van der Graaf<sup>2</sup>,  
4 Kaido Lepik<sup>2</sup>, Mariana Moysés-Oliveira<sup>1</sup>, Anelisa G Dantas<sup>1</sup>, Maria Isabel Melaragno<sup>1</sup>,  
5 Zoltán Kutalik<sup>2,3,4</sup>

6 <sup>1</sup>Genetics Division, Universidade Federal de São Paulo, São Paulo, Brazil

7 <sup>2</sup>Department of Computational Biology, University of Lausanne, Lausanne,  
8 Switzerland

9 <sup>3</sup>Swiss Institute of Bioinformatics, Lausanne, Switzerland

10 <sup>4</sup>University Center for Primary Care and Public Health, University of Lausanne,  
11 Lausanne, Switzerland

12 <sup>5</sup>Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

13

14

15

16

17

18

19

20

21

22 **Corresponding author:**

23 Zoltán Kutalik

24 Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.

25 Swiss Institute of Bioinformatics, Lausanne, Switzerland.

26 University Center for Primary Care and Public Health, University of Lausanne,  
27 Lausanne, Switzerland.

28 e-mail: zoltan.kutalik@unil.ch

29

30

31

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

32

## ABSTRACT

33 While extensively studied in clinical cohorts, the phenotypic consequences of 22q11.2  
34 copy number variants (CNVs) in the general population remain understudied. To  
35 address this gap, we performed a phenome-wide association scan in 405'324 unrelated  
36 UK Biobank (UKBB) participants using CNV calls from genotyping array. We mapped  
37 236 Human Phenotype Ontology terms linked to any of the 90 genes encompassed by  
38 the region to 170 UKBB traits and assessed the association between these traits and the  
39 copy-number state of 504 SNP-array probes in the region. We found significant  
40 associations for eight continuous and nine binary traits associated under different  
41 models (duplication-only, deletion-only, U-shape and mirror model). The causal effect  
42 of the expression level of 22q11.2 genes on associated traits was assessed through  
43 transcriptome-wide mendelian randomization (TWMR), revealing that increased  
44 expression of *ARVCF* increased BMI. Similarly, increased *DGCR6* expression causally  
45 reduced mean platelet volume, in line with the corresponding CNV effect. Furthermore,  
46 cross-trait multivariable mendelian randomization (MVMR) suggested a predominant  
47 role of genuine (horizontal) pleiotropy in the CNV region. Our findings show that  
48 within the general population, 22q11.2 CNVs are associated with traits previously  
49 linked to genes in the region, with duplications and deletions acting upon traits in  
50 different fashion. We also showed that gain or loss of distinct segments within 22q11.2  
51 may impact a trait under different association models. Our results have provided new  
52 insights to help further the understanding of the complex 22q11.2 region.

53

54

55

56

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

## 57 INTRODUCTION

58 The 22q11.2 region is a structurally complex region of the genome due to the presence  
59 of segmental duplications or low copy repeats (LCRs), named LCRA to LCRH, which  
60 predispose the region to genomic rearrangements, resulting in deletions or duplications  
61 of different segments. Specifically, deletions within the ~3 Mb segment from LCRA to  
62 LCRD represent the main cause of the 22q11.2 deletion syndrome (22q11.2DS), the  
63 most frequent microdeletion syndrome in humans, with an estimated incidence between  
64 1: 3000 and 1: 6000 live births <sup>1</sup>.

65 Studies in clinical cohorts have investigated the phenotypic consequences of the  
66 22q11.2 deletion, which include cardiac defects, facial and palate alterations,  
67 immunodeficiencies, endocrine, genitourinary, and gastrointestinal alterations <sup>1,2</sup>,  
68 developmental delay, cognitive deficits, and psychiatric disorders, such as schizophrenia  
69 <sup>1</sup>. In contrast, the phenotypic consequences of the region's duplication remain more  
70 elusive. Most of what is known is based on studies of a few individuals or families, but  
71 the findings indicate pleiotropy and variable consequences, similarly to the deletion.  
72 Some features, such as heart defects, velopharyngeal insufficiency, and  
73 neurodevelopmental and psychiatric disorders are shared with the 22q11.2DS <sup>3,4</sup>. Other  
74 22q11.2 duplication carriers exhibit very mild or unnoticeable phenotypes <sup>5</sup>, suggesting  
75 variable expressivity and/or reduced penetrance. Finally, rare single nucleotide variants  
76 (SNVs) in genes encompassed by the region have been linked to various disorders, such  
77 as Bernard–Soulier syndrome, caused by SNVs in *GP1BB* <sup>6</sup>, or CEDNIK syndrome,  
78 caused by SNVs in *SNAP29* <sup>7</sup>. Overall, the multitude of variants and phenotypes that  
79 have been linked to the 22q11.2 LCRA to LCRD region highlights its clinical relevance.  
80 Because of their highly deleterious impact, 22q11.2 variants are often investigated in  
81 clinical settings. Studied cohorts are thus heavily biased towards individuals with severe  
82 phenotypic manifestation, leading to an incomplete and biased understanding of these  
83 variants' role in the human population. This is particularly relevant considering recent

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

84 studies that have shown variable expressivity and incomplete penetrance of SNVs<sup>8,9</sup>  
85 and CNVs<sup>10</sup> that were previously believed to be highly pathogenic, including at the  
86 22q11.2 LCRA-LCRD locus<sup>11</sup>. To address this gap, we performed a phenome-wide  
87 analysis in the UK Biobank (UKBB) (Bycroft et al., 2018), a populational cohort of  
88 ~500,000 individuals, to identify associations of 22q11.2 CNVs with traits previously  
89 implicated by their genetic content.

## 90 **MATERIAL AND METHODS**

### 91 **Cohort description**

92 Analyses were performed in the UK Biobank (UKBB), a volunteer-based cohort from  
93 the general UK adult population (Bycroft et al., 2018). Gender mismatched, related and  
94 retracted samples (by 09/08/2021), as well as CNV outliers (see **CNV calling**) were  
95 excluded, resulting in a total of 405'324 participants (218'719 females and 186'605  
96 males) used for the analyses. Individuals were aged between 40 and 69 years at  
97 recruitment. All participants signed a broad informed consent form and data was  
98 accessed through a UKBB application (#16389).

### 99 **22q11.2 Region Definition**

100 The 22q11.2 region was defined as chr22:18,630,000\_21,910,000 based on the human  
101 genome reference build GRCh37/hg19 in order to encompass LCRs from A to D. The  
102 90 NCBI RefSeq genes contained in the region were downloaded from the UCSC Table  
103 Browser (<http://genome.ucsc.edu/cgi-bin/hgTables?command=start>).

### 104 **Trait Selection**

105 Phenotypes linked to the 22q11.2 region's genetic content were identified using the  
106 Human Phenotype Ontology (HPO) mapping<sup>13</sup>, an ontology-based system that uses  
107 information from different medical sources including OMIM and Orphanet. Genes and  
108 their most specific associated HPO term (i.e., not all ancestors) were downloaded from  
109 the HPO database ([http://purl.obolibrary.org/obo/hp/hpoa/genes\\_to\\_phenotype.txt](http://purl.obolibrary.org/obo/hp/hpoa/genes_to_phenotype.txt) -

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

110 Accessed on 22/10/2021). Overall, 24 out of 90 genes in the 22q11.2 region were  
111 associated to at least one HPO term, yielding 631 associated HPO terms.

### 112 **Mapping of HPO terms to UKBB binary traits**

113 To map HPO terms to binary UKBB traits, two complementary approaches were used.  
114 First, the online tool EMBL-EBI Ontology Xref Service (OxO)  
115 (<https://www.ebi.ac.uk/spot/oxo/>) was used to map HPO terms to International  
116 Classification of Diseases, 10<sup>th</sup> Revision [ICD-10] codes, followed by manual curation  
117 and grouping of ICD10 codes into broader phenotypes when appropriate according to  
118 the Phecode map<sup>14</sup>. Remaining HPO terms were mapped to Phecode definitions using  
119 manual curation by Bastarache et al. (2018)<sup>15</sup>. Mapping was manually curated and only  
120 phenotypes with at least 500 cases were retained. In addition, individuals with a related  
121 ICD10 code or self-reported disease to the one studied were excluded from controls in a  
122 phenotype-specific fashion (**Table S1**). Overall, 218 HPO terms were mapped to 152  
123 UKBB binary traits (**Table S2**). The number of individuals by phenotype is reported in  
124 **Table S3**.

### 125 **Mapping of HPO terms to UKBB continuous traits**

126 An in-house web scraping approach was developed to map HPO terms to UKBB  
127 continuous traits. A list of 1'769 continuous UKBB measures was used as input on the  
128 HPO database (<https://hpo.jax.org/app/>) to obtain the web page's results for each query.  
129 Results were filtered for HPO terms of interest i.e., 631 terms linked to 22q11.2 genes.  
130 With this approach, 18 UKBB continuous traits were obtained from 18 HPO terms  
131 (**Table S4**). The number of individuals by trait is reported in **Table S5**.

### 132 **22q11.2 CNV association scan**

#### 133 *CNV calling*

134 CNVs were called with PennCNV v1.0.5 and underwent quality control as described in  
135 Auwerx et al., 2022. Briefly, a quality score reflecting the probability for the CNV to be

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

136 a true positive was assigned to each call and used for filtering ( $|QS| \geq 0.5$ )<sup>16</sup>. CNVs  
137 from samples genotyped on plates with a mean CNV count per sample  $> 100$  or from  
138 samples with  $> 200$  CNVs or a single CNV  $> 10$  Mb were excluded to minimize batch  
139 effects, genotyping errors, or extreme chromosomal abnormalities.  
140 CNV calls were transformed into probe-by-sample matrices with copy-number state for  
141 each probe (deletion = -1; copy-neutral = 0; duplication = 1).

#### 142 ***Plink encoding and association models***

143 Probe-level matrices were converted to PLINK binary file sets, with copy-number states  
144 being encoded to accommodate analysis according to four different association models:  
145 duplication-only, deletion-only, mirror, and U-shape model (**Table 1**). The duplication-  
146 only model assessed the impact of duplications disregarding deletions; the deletion-only  
147 model assessed the impact of deletions disregarding duplications; the mirror model  
148 assessed the additive effect of each additional copy of a probe (i.e., duplications and  
149 deletions have opposing effects); while the U-shape model assumes that duplications  
150 and deletions have the same effect direction<sup>10</sup>.

#### 151 ***CNV probe selection and number of effective tests***

152 Probes with high genotype missingness ( $> 5\%$ ) were excluded, resulting in 864 CNV  
153 proxy probes spanning chr22:18,630,000\_21,910,000. We retained 504 CNV proxy  
154 probes that are highly correlated ( $r^2 \geq 0.999$ ) to at least ten other probes, allowing to  
155 reduce the multiple testing burden while ensuring that selected probes adequately  
156 capture the CNV landscape of the region.

157 The number of effective probes (i.e., number of probes required to capture 99.5% of the  
158 variance in the probe-by-sample CNV matrices) was calculated according to Gao et al.  
159 (2008) based on the 504 CNV proxy probes ( $N_{\text{eff-probes}} = 6$ ). The same approach was  
160 used to account for correlation among 18 continuous ( $N_{\text{eff-continuous}} = 16$ ) and 152 binary  
161 traits ( $N_{\text{eff-binary}} = 113$ ). This resulted in 774 effective tests ( $N_{\text{eff}} = N_{\text{eff-probes}} * (N_{\text{eff-continuous}}$   
162  $+ N_{\text{eff-binary}})$ ), setting the threshold for significance at  $p \leq 0.05/774 = 6.5 \times 10^{-5}$ .

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

163 ***Continuous traits***

164 The 18 selected continuous traits were inverse normal transformed and corrected for  
165 covariates: age, age<sup>2</sup>, sex, genotyping batch, and principal components (PCs) 1–40.  
166 Associations between the copy number (CN) of selected probes and normalized  
167 covariate-corrected traits were performed in PLINK v2.0 according to all four  
168 association models using linear regression, as previously described <sup>10</sup>. Significant  
169 associations ( $p \leq 6.5 \times 10^{-5}$ ) were retained.

170 ***Binary traits***

171 For each trait, covariates among age, age<sup>2</sup>, sex, genotyping batch, and principal  
172 components (PCs) 1–40 that were significantly associated with the trait ( $p \leq 0.05$ ) were  
173 selected with logistic regression in R. Associations between the copy number (CN) of  
174 selected probes and 152 binary selected traits were performed in PLINK v2.0 according  
175 to all four association models using logistic regression and correcting for trait-specific  
176 selected covariates. Significant associations ( $p \leq 6.5 \times 10^{-5}$ ) were retained.

177 ***Stepwise conditional analysis***

178 The number of independent signals per trait and association model was determined by  
179 stepwise conditional analysis <sup>10</sup>, i.e., CNV status of the lead probe was regressed out  
180 from the trait and association scan was conducted again until no more significantly  
181 associated probes remained.

182 ***Sensitivity analysis***

183 Due to the low frequency of CNVs within the 22q11.2 region, alternative tests were  
184 performed to ensure the confidence of significant associations. For significant  
185 associations with continuous traits, a Wilcoxon rank-sum test was performed as a  
186 sensitivity analysis to assess agreement with linear regression. Significant associations  
187 with binary traits were retained only when confirmed by at least one of two approaches:  
188 1) Fisher's exact test ( $p \leq 0.005$ ) for the duplication-only, deletion-only and U-shape

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

189 models and Cochran-Armitage test ( $p \leq 0.0005$ ) for the mirror model; 2) linear  
190 regression ( $p \leq 0.005$ ) of the inverse normal quantile transformed trait residuals  
191 obtained from the logistic regression model of the binary outcome on the selected  
192 covariates.

### 193 **Enrichment analysis**

194 For each gene, two groups of traits were defined: traits linked to the focal gene  
195 implicated by HPO *versus* other traits related to other genes in the 22q11.2 region but  
196 not to the focal gene. Association p-values for each probe within the gene (+/- 10 kb)  
197 and each association model were compared between traits in the two groups using a  
198 one-sided Wilcoxon rank-sum test (i.e.,  $H_a$ : unrelated traits have lower association p-  
199 values with the focal gene than related ones). The number of effective tests (see **CNV**  
200 **probe selection and number of effective tests**) for each gene was calculated and used  
201 to define gene-specific significance thresholds. Genes were considered significant when  
202 the probe with the smallest p-value reached that threshold. The comparison was only  
203 performed for genes with at least four continuous traits and ten binary traits in each  
204 group. A binominal enrichment was performed to establish whether the number of  
205 genes significant in the Wilcoxon rank-sum test was higher than expected by chance  
206 with *pbinom* function in R.

### 207 **Transcriptome-wide Mendelian Randomization (TWMR)**

208 TWMR was conducted as previously described<sup>18</sup> to identify changes in transcript levels  
209 of genes in the 22q11.2 region that causally modulate traits found to be associated to  
210 22q11.2 CNVs by our association scan and, if this was the case, in which direction (i.e.,  
211 whether increased gene expression associates with increased or decreased phenotype  
212 value). Briefly, the exposure (i.e., transcript level) and outcome (i.e., trait) are  
213 instrumented using independent genetic variants (instrumental variables (IVs);  $r^2 <$   
214  $0.01$ ). Given their genetic effect sizes on these two quantities, a causal effect of the  
215 exposure on the outcome can be estimated using two-sample MR. Genetic effect sizes



It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

216 on transcript levels originate from whole blood expression quantitative trait loci (eQTL)  
217 provided by the eQTLGen consortium (cis-eQTLs at FDR < 0.05, 2-cohort filter)<sup>19</sup>.  
218 Effect sizes on the traits stem from genome-wide association study (GWAS) summary  
219 statistics conducted on the UK Biobank (Neale's lab: [http://www.nealelab.is/uk-](http://www.nealelab.is/uk-biobank/)  
220 [biobank/](http://www.nealelab.is/uk-biobank/); Pan-UKBB team: <https://pan.ukbb.broadinstitute.org>) (Table S6). Prior to the  
221 analysis, eQTL and GWAS data were harmonized, palindromic SNPs were removed, as  
222 well as SNPs with an allele frequency difference > 0.05 between datasets. For increased  
223 robustness of the estimated causal effects,  $\geq 5$  (independent) IVs were required. MR  
224 estimates were considered significant when  $p \leq 0.05/17 = 0.003$  to account for the  
225 testing of 17 transcripts with at least 5 IVs and only significant genes overlapped by the  
226 CNV association signal were reported.  
227 TWMR results were used for validation of the mirror model associations. It is expected  
228 that TWMR and mirror model effects are directionally concordant, i.e.,  
229 increase/decrease in copy-number has the same direction of effect on a trait as an  
230 increase/decrease in gene expression. For this purpose, nominally significant ( $p < 0.05$ )  
231 TWMR effects were retained and their direction was compared to the direction of the  
232 probe with the smallest nominally significant p-value ( $p < 0.05$ ) in the mirror  
233 association model for the corresponding gene ( $\pm 10$  kb) and trait.

#### 234 **Multivariable Mendelian Randomization (MVMR)**

235 MVMR was performed to assess the causal relationship between significantly  
236 associated traits and compute a phenotype network. IVs were obtained from Neale Lab  
237 UKBB (<http://www.nealelab.is/uk-biobank>) and Pan-UKBB  
238 (<https://pan.ukbb.broadinstitute.org>) (Table S6) GWAS summary statistics for all eight  
239 significant continuous traits and nine significant binary traits. Data were harmonized  
240 with genetic variants in the UK10K reference dataset and variants with minor allele  
241 frequency (MAF)  $\leq 0.01$  were filtered out. Genetic variants were clumped at  $r^2 = 0.001$   
242 using UK10K as a reference panel in PLINK v1.9. Mendelian randomization (MR)

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

243 analysis was performed in two steps. First, potentially causal effects were identified  
244 with a univariable inverse variance weighted (IVW) MR for all exposure-outcome  
245 combinations (i.e. pairs of associated traits). Second, all exposures with nominally  
246 significant IVW causal effect estimates for a given outcome were included in an  
247 MVMR analysis as exposures. To reduce bias due to potential reverse causation, Steiger  
248 filtering was performed in all MR analyses ( $p < 5 \times 10^{-3}$ ).

249 MVMR established the causal relationships among assessed traits using genetic variants  
250 as IVs. To infer if the pleiotropic effect of CNVs is vertical (indirect) or horizontal  
251 (genuine), we estimated what would be the expected CNV effect on the outcome trait  
252 ( $\beta_{\text{expected outcome}}$ ) if that outcome is a downstream result of the exposure trait as suggested  
253 by the MVMR analysis (vertical pleiotropy).  $\beta_{\text{expected outcome}}$  was determined as  $\beta_{\text{exposure}} \times$   
254  $\beta_{\text{IVW}}$ , with  $\beta_{\text{exposure}}$  the effect size of the best probe in the mirror model for each exposure  
255 (i.e. observed CNV - exposure trait association) and  $\beta_{\text{IVW}}$  the causal estimate for each  
256 exposure-outcome pair obtained from IVW MR. We then compared  $\beta_{\text{expected outcome}}$  with  
257 the observed CNV effect on the outcome trait ( $\beta_{\text{observed outcome}}$ ) obtained from the mirror  
258 association model.

## 259 **Software versions**

260 Genetic analyses were conducted with PLINK v1.9 and PLINK v2.0. Statistical  
261 analyses were performed with R v3.6.1 and figures were generated with R v4.2.0.

## 262 **RESULTS**

### 263 **22q11.2 CNVs in the UKBB**

264 After CNV calling and quality control in 405'324 unrelated individuals of the UKBB,  
265 we identified 1'127 individuals with a duplication and 694 individuals with a deletion  
266 overlapping the 22q11.2 LCRA-D region (**Figure 1A**). CNVs varied in size:  
267 duplication length ranged between 71 kb and 8.8 Mb (i.e., breakpoints extending

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

268 beyond the defined region) with a median of 132 kb, while deletion length ranged

269 between 80 kb and 2.8 Mb also with a median of 132 kb.

270 To assess whether individuals with these CNVs (mean number of diagnoses = 8.6) had a

271 higher disease burden than individuals that are copy-neutral within this region (mean

272 number of diagnoses = 8), we compared the reported number of ICD-10 codes and

273 identified no statistical difference (two-sided Wilcoxon rank-sum test:  $p_{\text{del}} = 0.44$ ;  $p_{\text{dup}} =$

274 0.053) (**Figure 1B**).

275 CNVs were classified according to their localization as defined by LCRA-D. Between

276 LCRs A and B, duplications were identified at a frequency of 0.01% and deletions at

277 0.002%; CNVs from LCR A to D, had a frequency of 0.06% and 0.001% for

278 duplications and deletions respectively; from LCR B to D, duplications had a frequency

279 of 0.002% and no deletions were identified; between LCRs C and D, duplications were

280 identified at a frequency of 0.04% while deletions at 0.008%. CNVs that did not fall

281 into these categories were considered as atypical and had a frequency of 0.16% for both

282 duplications and deletions (**Figure 1A**).

283 To account for all CNVs and bypass issues related to breakpoint variability, CNV calls

284 were converted into probe-by-sample matrices for the CNV association scan. Probe-

285 level CNV frequency after excluding LCRA probes (mean duplication frequency:

286 0.07%; mean deletion frequency: 0.004%) ranged between 0.004-0.1% and 0.001-

287 0.01% for duplications and deletions, respectively (**Figure 1C**).

## 288 **Associated Traits**

289 CNV association scan revealed significant links for eight continuous (**Table 2, Figure**

290 **S1**) and nine binary traits (**Table 3, Figure S2**), which were associated under different

291 association models. Eight traits (four binary and four continuous) were associated most

292 significantly under the U-shape model, three continuous traits did so under the mirror

293 model, four binary traits were associated more significantly under the duplication-only

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

294 model and two traits under the deletion-only model (one continuous and one binary),  
295 highlighting the importance of testing models mimicking different dosage mechanisms.  
296 Among the identified continuous traits, body mass index (BMI) was found associated  
297 under the U-shape model ( $\beta = 1.56 \text{ kg/m}^2$ ,  $p = 4.9 \times 10^{-10}$ ) throughout LCRA to LCRD  
298 (**Figure 2A**) indicating that both duplications and deletions increase BMI level (**Figure**  
299 **2B**). TWMR analysis showed that increased expression of *ARVCF* increases BMI ( $\beta =$   
300  $0.05$ ,  $p = 10^{-4}$ ), concordantly with the positive association found by the mirror CNV  
301 association scan (**Figure 2C**).

302 Mean platelet volume (MPV) was found associated under the mirror model ( $\beta = -0.58$   
303 femtolitres,  $p = 1.3 \times 10^{-18}$ ) with the strongest association occurring in the LCRA to  
304 LCRB region (**Figure 3A**). The signal replicated in both the duplication-only ( $\beta = -0.54$   
305 femtolitres,  $p = 1.16 \times 10^{-15}$ ) and deletion-only ( $\beta = 1.66$  femtolitres,  $p = 1.13 \times 10^{-6}$ )  
306 model providing further evidence of a “true mirror” effect, despite the deletion effect  
307 being slightly stronger than the duplication one (**Figure 3B**). In line with this effect,  
308 TWMR revealed that increased *DGCR6* expression causally reduces MPV ( $\beta = -0.03$ ,  $p$   
309  $= 0.001$ ) (**Figure 3C**). It is worth noting that this trait is negatively correlated with  
310 platelet count (also significant under the mirror model,  $\beta = 19.86 \times 10^9$  cells/L,  $p = 2.5 \times$   
311  $10^{-8}$ ). As expected, MVMR showed bidirectional causality between both traits,  
312 highlighting the challenges on interpreting their association separately.

313 Unlike other phenotypes, height was associated under different models in distinct  
314 regions.

315 The U-shape model appeared as the most significant model in the region spanning  
316 LCRA to LCRB ( $\beta = -2.09 \text{ cm}$ ,  $p = 1.1 \times 10^{-7}$ ), while the deletion-only model was the  
317 only significant one at the distal portion between LCRC and LCRD ( $\beta = -4.86 \text{ cm}$ ,  $p =$   
318  $5.5 \times 10^{-6}$ ) (**Figure 4A**). Given this unexpected pattern, we stratified CNVs according to  
319 LCR categories (**Figure 1A**) to inspect their impact on height. Within LCRA-LCRB  
320 and LCRA-LCRD (**Figures 4B and 4C**), both duplications and deletions were

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

321 associated with a height decrease in concordance with the U-shape model. However,  
322 duplications and deletions within LCRC and LCRD had opposing effects on height, in  
323 line with a mirror model which was confirmed by linear regression ( $\beta = 0.17$  cm,  $p =$   
324  $0.0003$ ) (**Figure 4D**).

325 Given the low number of deletion carriers affected by binary outcomes (0 to 3 carriers)  
326 (**Table S7**), associations found under the U-shape or mirror models often reflect the  
327 effect of duplications (i.e., the most common CNV type) in these phenotypes. One  
328 example is gastroesophageal reflux disease which was found to be associated under the  
329 duplication-only model ( $OR = 2.72$ ,  $p = 2.53 \times 10^{-8}$ ) with a stronger association  
330 occurring in the LCRA to LCRB region (**Figure 5A**), indicating an increased  
331 prevalence of gastroesophageal reflux disease among duplication carriers (**Figure 5B**).

### 332 **Enrichment analysis**

333 For continuous traits, 6 out of 8 assessed genes were found to have significantly greater  
334 association p-values for the group of unrelated traits compared to the group of linked  
335 traits for all association models (see Method section: Enrichment analysis for the  
336 definition of these groups). Binomial enrichment analysis indicated that CNV probes in  
337 genes linked to a given HPO term are 15 times more likely ( $p < 6 \times 10^{-9}$ ) to show  
338 stronger association with the corresponding UKBB continuous trait. For the binary  
339 traits, however, only 2 out of 19 assessed genes were significant in the mirror model  
340 which does not indicate an enrichment ( $p = 0.07$ ).

### 341 **Concordance in the direction of effect between association scan and TWMR**

342 Besides showing that differential expression of two 22q11.2 genes (*ARVCF* and  
343 *DGCR6*) causally affects two associated traits (BMI and MPV), TWMR results were  
344 also used to reinforce reliability of CNV associations. We evaluated concordance in the  
345 direction of effect sizes from nominally significant ( $p < 0.05$ ) results of the mirror CNV  
346 association scan and nominally significant ( $p < 0.05$ ) TWMR results (**Table S8**). As

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

347 expected, we observed a significant agreement in effect size directions between both  
348 when fitting a linear regression line ( $\beta = 1.6$ ,  $p = 0.01$ ; **Figure 6**).

### 349 **Causal links between traits and CNV pleiotropy**

350 Cross-trait MVMR was performed for all 17 significantly associated traits. Out of a  
351 total of 289 trait-pair combinations, we identified 48 pairs that are causally linked to  
352 each other at nominal significance ( $p < 0.05$ ) using the IVW MR method. MVMR was  
353 then applied on these 48 combinations and 17 trait-pairs were significant after  
354 Bonferroni correction ( $p < 0.05/289 = 0.0002$ ) (**Figure 7A**). Most traits were associated  
355 in a bidirectional manner, indicating that many (closely related) traits are mutually  
356 related to each other likely due to high genetic correlation. To distinguish between  
357 horizontal and vertical pleiotropy, we plotted the CNV effect on the outcome expected  
358 under vertical pleiotropy ( $\beta_{\text{expected outcome}}$ ) against the effect observed in the association  
359 scan ( $\beta_{\text{observed outcome}}$ ) to examine the concordance in effect direction (**Figure 7B**;  
360 MVMR method section). This analysis revealed agreement only for very closely related  
361 trait pairs (driven by strong genetic correlation) such as platelet count – mean platelet  
362 volume, and indicated that, in general, pleiotropic CNV association are not due to  
363 vertical, but rather due to genuine horizontal pleiotropy.

## 364 **DISCUSSION**

365 Most of our knowledge regarding the impact of CNVs in the 22q11.2 region in the  
366 general population stems from genome-wide studies<sup>10,20–24</sup>. Here, we focused on this  
367 region specifically and developed a tailored set of analyses with more lenient, yet  
368 appropriate, significance threshold and in-depth follow-up analyses that allowed to  
369 detect plausible associations missed by genome-wide studies (e.g., hearing loss,  
370 cardiomegaly, diplopia, and disorders of binocular vision). Our findings show that  
371 22q11.2 CNV carriers in the general population, that are likely on the milder end of the  
372 phenotypic spectrum, are associated with traits previously implicated by genes in the

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

373 region, shedding light on the variable expressivity and penetrance of CNVs impacting  
374 this complex genomic region.

375 Assessed traits linked to 22q11.2 genes have been previously identified in different  
376 contexts including the 22q11.2 deletion and duplication syndromes, clinical conditions  
377 caused by variants in a single gene, and complex conditions associated with the locus  
378 (**Figure S3**). Therefore, using the HPO database to select investigated traits allowed us  
379 to leverage information from different genetic variants in a clinical context <sup>13</sup>, to  
380 identify associations in the general population. Our enrichment analysis showed that for  
381 continuous traits this was an effective approach. We also show that CNVs can impact  
382 traits previously known to be associated with individual genes in the region, such as  
383 cardiomegaly (*LZTR1*, OMIM:616564) and other venous embolism and thrombosis  
384 (*SERPIND1*, OMIM:612356), that were both associated under the duplication-model in  
385 the distal region between LCRC and LCRD, which harbors these genes.

386 Our results validated several known associations and furthermore shed light on traits  
387 that have not yet been extensively studied in the context of 22q11.2 CNVs. For instance,  
388 gastroesophageal reflux disease is not a vastly explored clinical feature in 22q11.2  
389 deletion or duplication syndromes. While LCRA to LCRD duplications have been  
390 previously associated with this trait in the UKBB cohort <sup>20</sup>, replication of the  
391 association in our study emphasizes its relevance in 22q11.2 CNV carriers. Another  
392 relevant association identified in our study is with BMI. Obesity (BMI > 30) is a well-  
393 known phenotype in individuals with 22q11.2DS <sup>25</sup>. Even though this phenotype is not  
394 well described in clinical studies characterizing the 22q11.2 duplication syndrome, an  
395 increase in BMI has been associated with duplications in other studies assessing the  
396 UKBB cohort <sup>10,21</sup>. We have further shown a causal effect of differential expression of  
397 *ARVCF* – a gene whose product is part of the catenin family and is involved in protein-  
398 protein interactions at adherent junctions (OMIM: 602269) – on BMI. Recently, a  
399 rare *ARVCF* missense variant of unknown significance has been identified in an



It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

400 individual with early-onset severe obesity <sup>26</sup>, suggesting that *ARVCF* may play an  
401 important role in the etiology of obesity.

402 Besides validating the link between CNVs in the 22q11.2 region and platelet count <sup>10</sup>,  
403 we revealed a new association with mean platelet volume which exhibits a “true mirror”  
404 effect, reinforcing the role of this genomic region in phenotypes such as  
405 thrombocytopenia. Thrombocytopenia is a well-known clinical hallmark in 22q11.2DS <sup>1</sup>  
406 but is not yet recognized as a clinical feature of the 22q11.2 duplication syndrome.  
407 *GPIBB* represents a top candidate to explain the observed platelet phenotypes as  
408 biallelic loss of function variants in the gene are responsible for Bernard-Soulier  
409 Syndrome, a platelet disorder (OMIM: 231200), and inclusion of *GPIBB* in the deleted  
410 region has been implicated in reduction of platelet count levels in 22q11.2DS patients  
411 <sup>27</sup>. Due to lack of sufficient IVs, *GPIBB* could not be assessed by TWMR analysis,  
412 which instead revealed a causal effect of *DGCR6* differential expression on MPV.  
413 While *DGCR6*'s function is not yet clearly defined (OMIM: 601279), it has been  
414 implicated in regulating other genes in the 22q11.2 region <sup>28</sup>, suggesting that multiple  
415 genes in the region influence platelet phenotypes.

416 Usage of four different association models allowed for the identification of deletion-  
417 specific effects (e.g., calcium level), as well as traits in which duplications and deletions  
418 act in the same or in opposite directions. By performing association scans at the probe  
419 level, we also showed that gain or loss of distinct segments within 22q11.2 may impact  
420 a trait following different association models, as was seen for height. Short stature has  
421 been identified for the 22q11.2DS <sup>1</sup> but variable height measures have been described  
422 for the 22q11.1 duplication syndrome <sup>29-31</sup>. In concordance with our study, both  
423 duplications and deletions (LCRA to LCRD) have been previously associated with a  
424 decrease in height in the UKBB cohort <sup>21</sup>. However, our study is the first to show a  
425 mirror behavior involving the LCRC to LCRD region. The impact of CNVs in the  
426 LCRC-D region is often overlooked or considered in combination with LCRA to



It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

427 LCRB. However, the unexpectedly distinct impact of CNVs in this region on height, as  
428 well as certain traits that were only significant in this region (such as weight,  
429 cardiomegaly, other venous embolism and thrombosis, dental caries), reveal the value of  
430 a more refined study of CNVs overlapping this complex region.

431 A drawback of studying pathogenic CNVs in a general population such as the UKBB is  
432 that the number of affected participants is low as carriers of 22q11.2 CNVs with larger  
433 phenotypic impact are less likely to participate, a phenomenon often described as the  
434 “healthy volunteer” selection bias<sup>32</sup>. As such, frequencies of the 22q11.2 deletions and  
435 duplications have not been precisely estimated outside of clinical cohorts. In the general  
436 population, frequency of deletions and duplications encompassing the LCRA to LCRB  
437 region have been estimated at 0.02% and 0.08%, respectively (Kirov et al. 2014).  
438 Another study estimated a frequency of 0.03% for deletions and 0.07% for duplications  
439 considering the typical 3 Mb and 1.5 Mb CNVs<sup>34</sup>. In our work, the frequency of CNVs  
440 in LCRA to LCRB and LCRA to LCRD is 0.07% for duplications and 0.003% for  
441 deletions. It is worth noting that we consider smaller nested CNVs between LCRA and  
442 LCRB that were not appreciated in previous studies, indicating that if we applied  
443 similar definitions to these works, our frequency estimates would be lower.

444 While the absolute number of CNV carriers considered in our study is still larger than  
445 the sample size of some clinical cohorts, these individuals tend to exhibit milder  
446 phenotypes. This hampers statistical power to detect associations, especially for binary  
447 outcomes for which trait definition through grouping of ICD-10 codes is imperfect and  
448 arbitrary and case number can be extremely low. We offer corroborating evidence of  
449 our findings’ reliability by performing sensitivity analyses and examining the  
450 concordance of CNV findings with TWMR effects. Importantly, effects observed in our  
451 study are potentially smaller than the ones observed in clinical cohorts<sup>35</sup> as they are  
452 mainly derived from CNV carriers with sub-clinical phenotypes and thus represent  
453 lower bound estimates. While in theory estimates from clinical cohorts might offer

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

454 upper bound estimates, their poor and unstandardized reporting makes it difficult to  
455 establish accurate comparisons. Still, we hope that our study offers a better  
456 understanding on the spectrum of phenotypic consequences exerted by 22q11.2 and will  
457 improve diagnostic rates in individuals with low expressed phenotypes as molecular  
458 diagnostic of genomic syndromes still often relies on recognition of characteristic signs  
459 to guide genetic testing.

## 460 **CONCLUSION**

461 We found that 22q11.2 CNVs affect traits compatible with clinical manifestations seen  
462 in the genomic disorders within the general population. The probe-level association  
463 scan revealed that dosage of different segments within the 22q11.2 region may impact a  
464 trait through different mechanisms, as illustrated with height. Besides, yielding further  
465 insights into the complex 22q11.2 region, our study provides a framework that can be  
466 adapted to study the phenotypic consequences of other clinically relevant genomic  
467 regions.

## 468 **DECLARATION OF INTERESTS**

469 The authors declare no competing interests.

## 470 **ACKNOWLEDGMENTS**

471 This study was conducted with the UK Biobank Resource (under application number  
472 16389), we thank all biobank participants for sharing their data. This work was  
473 supported by funding from the Department of Computational Biology (Z.K.), the Swiss  
474 National Science Foundation (310030-189147) as well as financial support from  
475 Fundação de Amparo à Pesquisa do Estado de São Paulo [2020/11241-2, M.Z.;  
476 2019/21644-0, M.I.M] and from the Coordenação de Aperfeiçoamento de Pessoal de  
477 Nível Superior □ Brasil (CAPES).

## 478 **AUTHOR CONTRIBUTIONS**

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

479 M.Z. contributed to study design, conducted analysis, and interpretation of the data and  
480 wrote the article. C.A. contributed to study design and interpretation of the data. M.C.S  
481 performed TWMR analysis. A.G. performed MVMR analysis. K.L. designed the web  
482 scraping approach used for mapping of HPO terms to UKBB traits. M.M.O., A.G.D.  
483 and M.I.M. contributed to study design and interpretation. Z.K. supervised the study,  
484 contributed to study design and interpretation of the data. All authors critically revised  
485 the manuscript and approved the final version.

486

## 487 REFERENCES

- 488 1. McDonald-McGinn, D.M., Sullivan, K.E., Marino, B., Philip, N., Swillen, A.,  
489 Vorstman, J.A.S., Zackai, E.H., Emanuel, B.S., Vermeesch, J.R., Morrow, B.E., et al.  
490 (2015). 22q11.2 deletion syndrome. *Nat Rev Dis Primers* 15071.
- 491 2. Monteiro, F.P., Vieira, T.P., Sgardiolli, I.C., Molck, M.C., Damiano, A.P., Souza, J.,  
492 Monlleó, I.L., Fontes, M.I.B., Fett-Conte, A.C., Félix, T.M., et al. (2013). Defining new  
493 guidelines for screening the 22q11.2 deletion based on a clinical and dysmorphic  
494 evaluation of 194 individuals and review of the literature. *Eur J Pediatr* 172, 927–945.
- 495 3. Portnoi, M.-F. (2009). Microduplication 22q11.2: A new chromosomal syndrome.  
496 *Eur J Med Genet* 52, 88–93.
- 497 4. Verbesselt, J., Zink, I., Breckpot, J., and Swillen, A. (2022). Cross-sectional and  
498 longitudinal findings in patients with proximal 22q11.2 duplication: A retrospective  
499 chart study. *Am J Med Genet A* 188, 46–57.
- 500 5. Yobb, T.M., Somerville, M.J., Willatt, L., Firth, H. v., Harrison, K., MacKenzie, J.,  
501 Gallo, N., Morrow, B.E., Shaffer, L.G., Babcock, M., et al. (2005). Microduplication  
502 and Triplication of 22q11.2: A Highly Variable Syndrome. *The American Journal of*  
503 *Human Genetics* 76, 865–876.

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

- 504 6. Savoia, A., Kunishima, S., de Rocco, D., Zieger, B., Rand, M.L., Pujol-Moix, N.,  
505 Caliskan, U., Tokgoz, H., Pecci, A., Noris, P., et al. (2014). Spectrum of the Mutations  
506 in Bernard-Soulier Syndrome. *Hum Mutat* 35, 1033–1045.
- 507 7. Nunes, N., Zamariolli, M., Dantas, A.G., Cola, P., de Agostinho Júnior, F., Piazzon,  
508 F.B., Meloni, V.A., and Melaragno, M.I. (2022). CEDNIK syndrome in a Brazilian  
509 patient with compound heterozygous pathogenic variants. *Eur J Med Genet* 65, 104440.
- 510 8. Kingdom, R., and Wright, C.F. (2022). Incomplete Penetrance and Variable  
511 Expressivity: From Clinical Studies to Population Cohorts. *Front Genet* 13,.
- 512 9. Wright, C.F., West, B., Tuke, M., Jones, S.E., Patel, K., Laver, T.W., Beaumont,  
513 R.N., Tyrrell, J., Wood, A.R., Frayling, T.M., et al. (2019). Assessing the Pathogenicity,  
514 Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population  
515 Setting. *The American Journal of Human Genetics* 104, 275–286.
- 516 10. Auwerx, C., Lepamets, M., Sadler, M.C., Patxot, M., Stojanov, M., Baud, D., Mägi,  
517 R., Porcu, E., Reymond, A., Kutalik, Z., et al. (2022). The individual and global impact  
518 of copy-number variants on complex human traits. *The American Journal of Human*  
519 *Genetics* 109, 647–668.
- 520 11. Davies, R.W., Fiksinski, A.M., Breetvelt, E.J., Williams, N.M., Hooper, S.R.,  
521 Monfeuga, T., Bassett, A.S., Owen, M.J., Gur, R.E., Morrow, B.E., et al. (2020). Using  
522 common genetic variation to examine phenotypic expression and risk prediction in  
523 22q11.2 deletion syndrome. *Nat Med* 26, 1912–1918.
- 524 12. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A.,  
525 Vukcevic, D., Delaneau, O., O’Connell, J., et al. (2018). The UK Biobank resource with  
526 deep phenotyping and genomic data. *Nature* 562, 203–209.
- 527 13. Köhler, S., Gargano, M., Matentzoglou, N., Carmody, L.C., Lewis-Smith, D.,  
528 Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., et al. (2021).  
529 The Human Phenotype Ontology in 2021. *Nucleic Acids Res* 49, D1207–D1217.

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

- 530 14. Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll,  
531 R., Bastarache, L., Denny, J.C., et al. (2019). Mapping ICD-10 and ICD-10-CM Codes  
532 to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* 7,  
533 e14325.
- 534 15. Bastarache, L., Hughey, J.J., Hebring, S., Marlo, J., Zhao, W., Ho, W.T., van  
535 Driest, S.L., McGregor, T.L., Mosley, J.D., Wells, Q.S., et al. (2018). Phenotype risk  
536 scores identify patients with unrecognized Mendelian disease patterns. *Science* (1979)  
537 359, 1233–1239.
- 538 16. Macé, A., Tuke, M.A., Beckmann, J.S., Lin, L., Jacquemont, S., Weedon, M.N.,  
539 Reymond, A., and Kutalik, Z. (2016). New quality measure for SNP array based CNV  
540 detection. *Bioinformatics* 32, 3298–3305.
- 541 17. Gao, X., Starmer, J., and Martin, E.R. (2008). A multiple testing correction method  
542 for genetic association studies using correlated single nucleotide polymorphisms. *Genet*  
543 *Epidemiol* 32, 361–369.
- 544 18. Porcu, E., Rüeger, S., Lepik, K., Santoni, F.A., Reymond, A., and Kutalik, Z.  
545 (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic  
546 determinants of complex and clinical traits. *Nat Commun* 10, 3300.
- 547 19. Vösa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B.,  
548 Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and  
549 trans-eQTL analyses identify thousands of genetic loci and polygenic scores that  
550 regulate blood gene expression. *Nat Genet* 53, 1300–1310.
- 551 20. Crawford, K., Bracher-Smith, M., Owen, D., Kendall, K.M., Rees, E., Pardiñas,  
552 A.F., Einon, M., Escott-Price, V., Walters, J.T.R., O’Donovan, M.C., et al. (2019).  
553 Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J*  
554 *Med Genet* 56, 131–138.

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

- 555 21. Owen, D., Bracher-Smith, M., Kendall, K.M., Rees, E., Einon, M., Escott-Price, V.,  
556 Owen, M.J., O'Donovan, M.C., and Kirov, G. (2018). Effects of pathogenic CNVs on  
557 physical traits in participants of the UK Biobank. *BMC Genomics* *19*, 867.
- 558 22. Aguirre, M., Rivas, M.A., and Priest, J. (2019). Phenome-wide Burden of Copy-  
559 Number Variation in the UK Biobank. *The American Journal of Human Genetics* *105*,  
560 373–383.
- 561 23. Kendall, K.M., Bracher-Smith, M., Fitzpatrick, H., Lynham, A., Rees, E., Escott-  
562 Price, V., Owen, M.J., O'Donovan, M.C., Walters, J.T.R., and Kirov, G. (2019).  
563 Cognitive performance and functional outcomes of carriers of pathogenic copy number  
564 variants: analysis of the UK Biobank. *The British Journal of Psychiatry* *214*, 297–304.
- 565 24. Kendall, K.M., Rees, E., Bracher-Smith, M., Legge, S., Riglin, L., Zammit, S.,  
566 O'Donovan, M.C., Owen, M.J., Jones, I., Kirov, G., et al. (2019). Association of Rare  
567 Copy Number Variants With Risk of Depression. *JAMA Psychiatry* *76*, 818.
- 568 25. Voll, S.L., Boot, E., Butcher, N.J., Cooper, S., Heung, T., Chow, E.W.C.,  
569 Silversides, C.K., and Bassett, A.S. (2017). Obesity in adults with 22q11.2 deletion  
570 syndrome. *Genetics in Medicine* *19*, 204–208.
- 571 26. Loid, P., Pekkinen, M., Mustila, T., Tossavainen, P., Viljakainen, H., Lindstrand,  
572 A., and Mäkitie, O. (2022). Targeted Exome Sequencing of Genes Involved in Rare  
573 CNVs in Early-Onset Severe Obesity. *Front Genet* *13*,.
- 574 27. Campbell, I.M., Crowley, T.B., Jobaliya, C., Bailey, A., McGinn, D.E., Gaiser, K.,  
575 Bassett, A., Gur, R.E., Morrow, B., Emanuel, B.S., et al. (2022). Platelet findings in  
576 22q11.2 Deletion Syndrome correlate with disease manifestations but do not correlate  
577 with GPIb surface expression. *MedRxiv* 2022.06.10.22276258.
- 578 28. Hierck, B.P., Molin, D.G.M., Boot, M.J., Poelmann, R.E., and Gittenberger-De  
579 Groot, A.C. (2004). A Chicken Model for DGCR6 as a Modifier Gene in the DiGeorge  
580 Critical Region. *Pediatr Res* *56*, 440–448.

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

- 581 29. Yu, A., Turbiville, D., Xu, F., Ray, J.W., Britt, A.D., Lupo, P.J., Jain, S.K.,  
582 Shattuck, K.E., Robinson, S.S., and Dong, J. (2019). Genotypic and phenotypic  
583 variability of 22q11.2 microduplications: An institutional experience. *Am J Med Genet*  
584 *A 179*, 2178–2189.
- 585 30. Courtens, W., Schramme, I., and Laridon, A. (2008). Microduplication 22q11.2: A  
586 benign polymorphism or a syndrome with a very large clinical variability and reduced  
587 penetrance?—Report of two families. *Am J Med Genet A 146A*, 758–763.
- 588 31. Verbesselt, J., Zink, I., Breckpot, J., and Swillen, A. (2022). Cross-sectional and  
589 longitudinal findings in patients with proximal 22q11.2 duplication: A retrospective  
590 chart study. *Am J Med Genet A 188*, 46–57.
- 591 32. Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T.,  
592 Collins, R., and Allen, N.E. (2017). Comparison of Sociodemographic and Health-  
593 Related Characteristics of UK Biobank Participants With Those of the General  
594 Population. *Am J Epidemiol 186*, 1026–1034.
- 595 33. Kirov, G., Rees, E., Walters, J.T.R., Escott-Price, V., Georgieva, L., Richards, A.L.,  
596 Chambert, K.D., Davies, G., Legge, S.E., Moran, J.L., et al. (2014). The Penetrance of  
597 Copy Number Variations for Schizophrenia and Developmental Delay. *Biol Psychiatry*  
598 *75*, 378–385.
- 599 34. Olsen, L., Sparsø, T., Weinsheimer, S.M., dos Santos, M.B.Q., Mazin, W.,  
600 Rosengren, A., Sanchez, X.C., Hoeffding, L.K., Schmock, H., Baekvad-Hansen, M., et  
601 al. (2018). Prevalence of rearrangements in the 22q11.2 region and population-based  
602 risk of neuropsychiatric and developmental disorders in a Danish population: a case-  
603 cohort study. *Lancet Psychiatry 5*, 573–580.
- 604 35. Kingdom, R., Tuke, M., Wood, A., Beaumont, R.N., Frayling, T.M., Weedon, M.N.,  
605 and Wright, C.F. (2022). Rare genetic variants in genes and loci linked to dominant  
606 monogenic developmental disorders cause milder related phenotypes in the general  
607 population. *The American Journal of Human Genetics 109*, 1308–1316.

608

609

## 610 **FIGURE TITLE AND LEGENDS**

611 **Figure 1 | 22q11.2 CNVs landscape.** **A)** Each UKBB CNV carrier is displayed through  
612 a segment than spans the genomic coordinates of the CNV. Duplications are represented  
613 in the top part of the graph, while deletions at the bottom. Shades of blue and red  
614 represent different duplication and deletion categories, respectively, according to their  
615 localization in reference to the LCRA to LCRD. The number of duplications and  
616 deletions for each category is displayed in the boxes. **B)** Boxplot representing the  
617 number of ICD-10 codes reported in individuals grouped according to their copy-  
618 number state in the 22q11.2 region. N indicates the sample size for each category; dots  
619 show the mean; boxes show the first (Q1), second (median, thick line), and third (Q3)  
620 quartiles; lower and upper whiskers show the most extreme value within Q1 minus and  
621 Q3 plus 1.5× the interquartile range; outliers are not shown. **C)** Probe-level duplication  
622 (top, blue) and deletion (bottom, red) frequencies [%] for 864 probes plotted against the  
623 22q11.2 genomic region. Frequency was calculated as the number of duplications or  
624 deletions divided by the total number of individuals assessed for the probe.

625

626 **Figure 2 | 22q11.2 CNVs and body mass index (BMI).** **A) Top:** The negative  
627 logarithm of the association p-value for the U-shape CNV-BMI association scan is  
628 plotted against the 22q11.2 genomic region. Each point represents a CNV proxy probe  
629 and the lead signal (chr22:20,765,989) is shown in black. The red dashed line indicates  
630 significance threshold ( $p < 6.5 \times 10^{-5}$ ). **Bottom:** Low copy-repeat region (LCR) A-D, as  
631 well as the 90 genes contained in the region. The 24 genes linked to traits according to  
632 HPO are labeled and genes linked to BMI through HPO are labeled in black. *ARVCF*  
633 expression was found to causally influence BMI through TWMR and is shown in green.  
634 **B)** Boxplot representing BMI in individuals grouped according to their copy-number



It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

635 state of the lead signal probe (chr22:20,765,989). N indicates the sample size for each  
636 category; dots show the mean; boxes show the first (Q1), second (median, thick line),  
637 and third (Q3) quartiles; lower and upper whiskers show the most extreme value within  
638 Q1 minus and Q3 plus  $1.5 \times$  the interquartile range; outliers are not shown. C) Representation of the TWMR analysis showing SNPs as instrumental variables (IVs),  
639 *ARVCF* gene expression as exposure, and its causal effect size ( $\beta = 0.05$ ) on BMI.

641

642 **Figure 3 | 22q11.2 CNVs and mean platelet volume (MPV).** A) **Top:** The negative  
643 logarithm of the mirror association p-value for the CNV-MPV association is plotted  
644 against the 22q11.2 genomic region. Each point represents a CNV proxy probe and the  
645 lead signal (chr22:19,639,383) is shown in black. The red dashed line indicates  
646 significance threshold ( $p < 6.5 \times 10^{-5}$ ). **Bottom:** Low copy-repeat region (LCR) A-D, as  
647 well as the 90 genes contained in the region. The 24 genes linked to traits according to  
648 HPO are labeled and genes linked to mean platelet volume through HPO are labeled in  
649 black. *DGCR6* expression was found to causally influence mean platelet volume  
650 through TWMR and is shown in orange. B) Boxplot representing mean platelet volume  
651 in individuals grouped according to their copy-number state for the lead signal probe  
652 (chr22:19,639,383). N indicates the sample size for each category; dots show the mean;  
653 boxes show the first (Q1), second (median, thick line), and third (Q3) quartiles; lower  
654 and upper whiskers show the most extreme value within Q1 minus and Q3 plus  $1.5 \times$  the  
655 interquartile range; outliers are not shown. C) Representation of the TWMR analysis  
656 showing SNPs as instrumental variables (IVs), *DGCR6* gene and its causal effect size ( $\beta$   
657 = - 0.03) on MPV.

658

659 **Figure 4 | 22q11.2 CNVs and height.** A) **Top:** The negative logarithm of the  
660 association p-value for the CNV-height association according to a deletion-only (red),  
661 duplication-only (blue), mirror (orange), and U-shape (green) is plotted against the

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

662 22q11.2 genomic region. The red dashed line indicates significance threshold ( $p < 6.5 \times$   
663  $10^{-5}$ ). **Bottom:** Low copy-repeat region (LCR) A-D, as well as the 90 genes contained  
664 in the region. The 24 genes linked to traits according to HPO are labeled and genes  
665 linked to height through HPO are labeled in black. **B)** Boxplots representing height in  
666 individuals with **(B)** LCRA-B, **(C)** LCRA-D, and **(D)** LCRC-D CNVs grouped  
667 according to their copy-number state. N indicates the sample size for each category;  
668 dots show the mean; boxes show the first (Q1), second (median, thick line), and third  
669 (Q3) quartiles; lower and upper whiskers show the most extreme value within Q1 minus  
670 and Q3 plus  $1.5 \times$  the interquartile range; outliers are not shown.

671

672 **Figure 5 | 22q11.2 CNVs and gastroesophageal reflux disease (GERD).** **A) Top:** The  
673 negative logarithm of the duplication-only association p-value for the CNV-GERD  
674 association is plotted against the 22q11.2 genomic region. Each point represents a CNV  
675 proxy probe and the lead signal (chr22:19,998,655). The red dashed line indicates  
676 significance threshold ( $p < 6.5 \times 10^{-5}$ ). **Bottom:** Low copy-repeat region (LCR) A-D, as  
677 well as the 90 genes contained in the region. The 24 genes linked to traits according to  
678 HPO are labeled and genes linked to mean platelet volume through HPO are labeled in  
679 black. **B)** Barplot representing prevalence (cases/total) of GERD grouped according to  
680 copy-number state for the lead signal probe (chr22:19,998,655). 95% confidence  
681 interval for deletion is truncated at zero.

682

683 **Figure 6 | Concordance between TWMR and CNV association scan effect sizes:**  
684 Scatter plot depicting mirror association scan (y-axis) versus TWMR (x-axis) effect  
685 sizes. Vertical and horizontal bars represent the 95% confidence intervals. The (zero-  
686 intercept) regression line and the corresponding slope are in black. For association scan  
687 effect sizes, the probe with the smallest p-value in the mirror model located in the  
688 TWMR gene was selected. Trait-gene pairs with agreeing direction between TWMR

It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/) .

689 and CNV association scan are in green and trait-gene pairs with opposite directions are  
690 in pink. Labels indicate: (1) hypotension - *GNBIL*; (2) cardiomegaly - *P2RX6*; (3) mean  
691 platelet volume - *DGCR6*; (4) gastroesophageal reflux disease - *GNBIL*; (5) weight -  
692 *P2RX6*; (6) mean platelet volume - *CLDN5*; (7) height - *TANGO2*; (8) height - *CLDN5*;  
693 (9) weight - *CLDN5*; (10) height - *GNBIL*; (11) weight - *GNBIL*; (12) body mass index  
694 - *SLC25A1*; (13) calcium levels - *CLTCL1*; (14) platelet count - *CLDN5*; (15) platelet  
695 count - *P2RX6*; (16) whole body fat mass - *ARVCF*; (17) body mass index - *ARVCF*;  
696 (18) weight - *ARVCF*; (19) nausea and vomiting - *DGCR6*; (20) diplopia and disorders  
697 of binocular vision - *DGCR6*; (21) cardiomegaly - *ARVCF*; (22) hearing loss -  
698 *SLC25A1*; (23) hypotension - *DGCR2*.

699

#### 700 **Figure 7 | Concordance between CNV expected and observed effect on outcome**

701 **trait: (A)** Causal links identified in the MVMR analysis. Colored shapes indicate  
702 clusters of traits grouped based on their correlation ( $r > |0.45|$ ). **(B)** Scatter plot  
703 depicting estimated CNV expected effect on the outcome (y-axis) versus CNV observed  
704 effect on outcome (x-axis) for each trait pair. Trait pairs from the same cluster **(A)** are in  
705 green and trait pairs from different clusters are in pink. The vertical and horizontal bars  
706 represent the 95% confidence intervals. Labels indicate exposure – outcome pairs: (1)  
707 platelet count – mean platelet volume; (2) body mass index – height; (3) weight –  
708 height; (4) platelet count - calcium levels; (5) height – weight; (6) fat mass - platelet  
709 count; (7) weight - platelet count; (8) body mass index – weight; (9) fat mass – weight;  
710 (10) platelet count - fat mass; (11) mean platelet volume - fat mass; (12) height – body  
711 mass index; (13) mean platelet volume - platelet count; (14) BMI - fat mass; (15) weight  
712 - fat mass; (16) weight – body mass index; (17) fat mass – body mass index.

713

714

#### 715 **Table 1. PLINK encoding of CNVs into association models**

It is made available under a [CC-BY-ND 4.0 International license](#) .

716

<b>Association Model</b>	<b>Deletion-only</b>	<b>Duplication-only</b>	<b>Mirror</b>	<b>U-shape<sup>a</sup></b>
<b>Deletion</b>	TT	00	AA	AA
<b>Copy-Neutral</b>	AT	AT	AT	AT
<b>Duplication</b>	00	TT	TT	TT

717 <sup>a</sup>For the U-shape model, the “hetonly” modifier in Plink was used.

718

719

**Table 2. Continuous traits associated to CNVs in the 22q11.2 region with different models.**

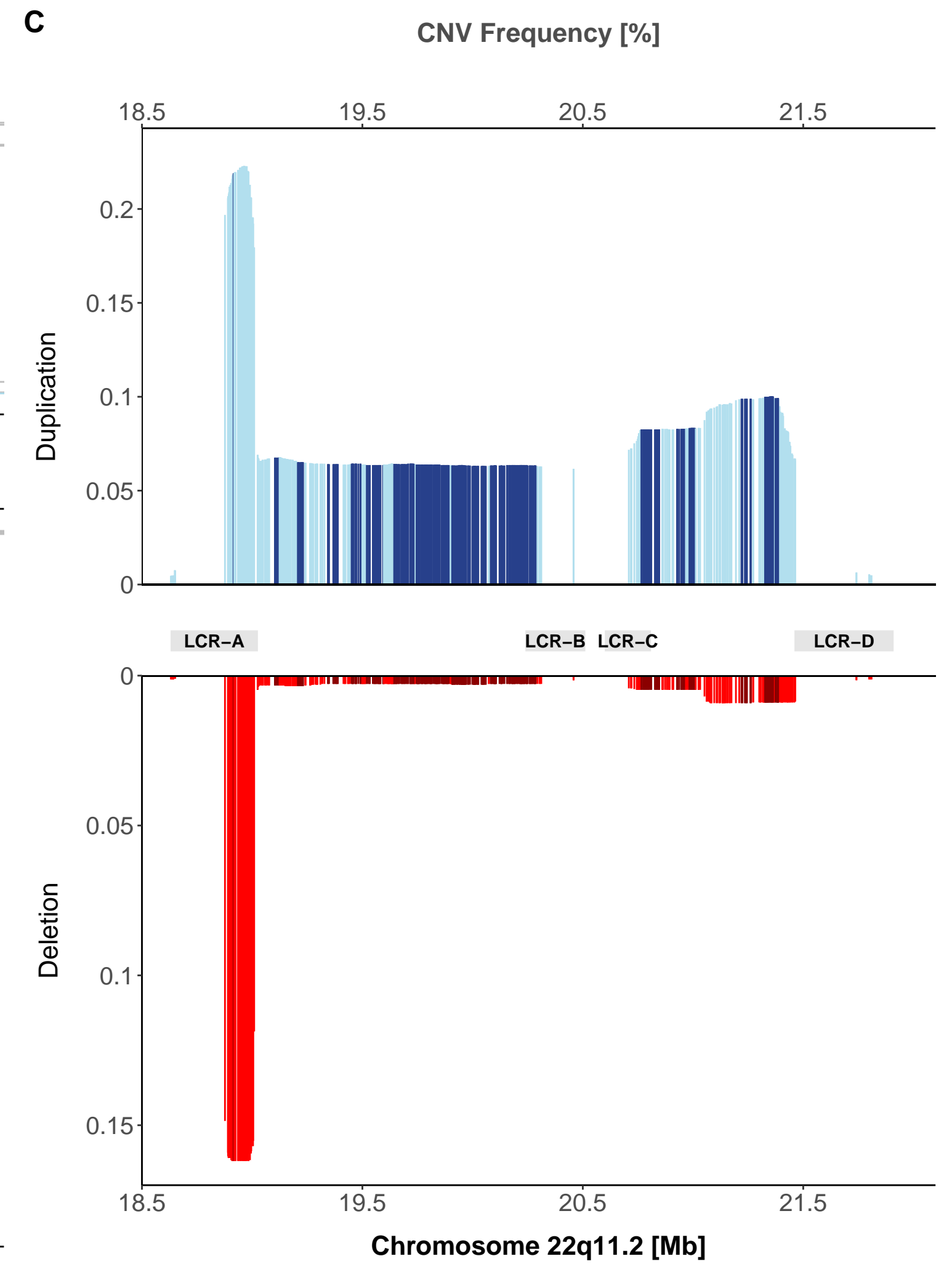
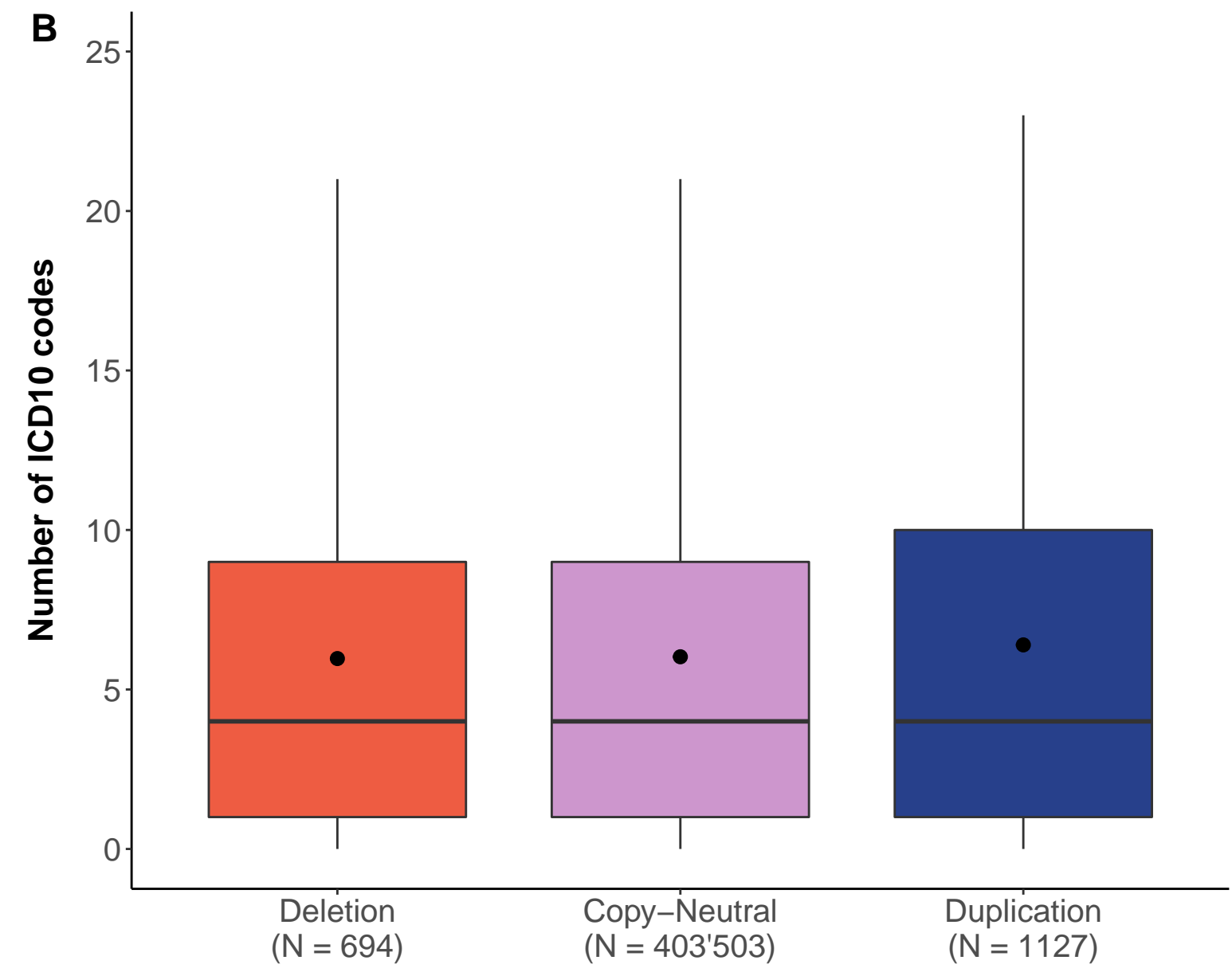
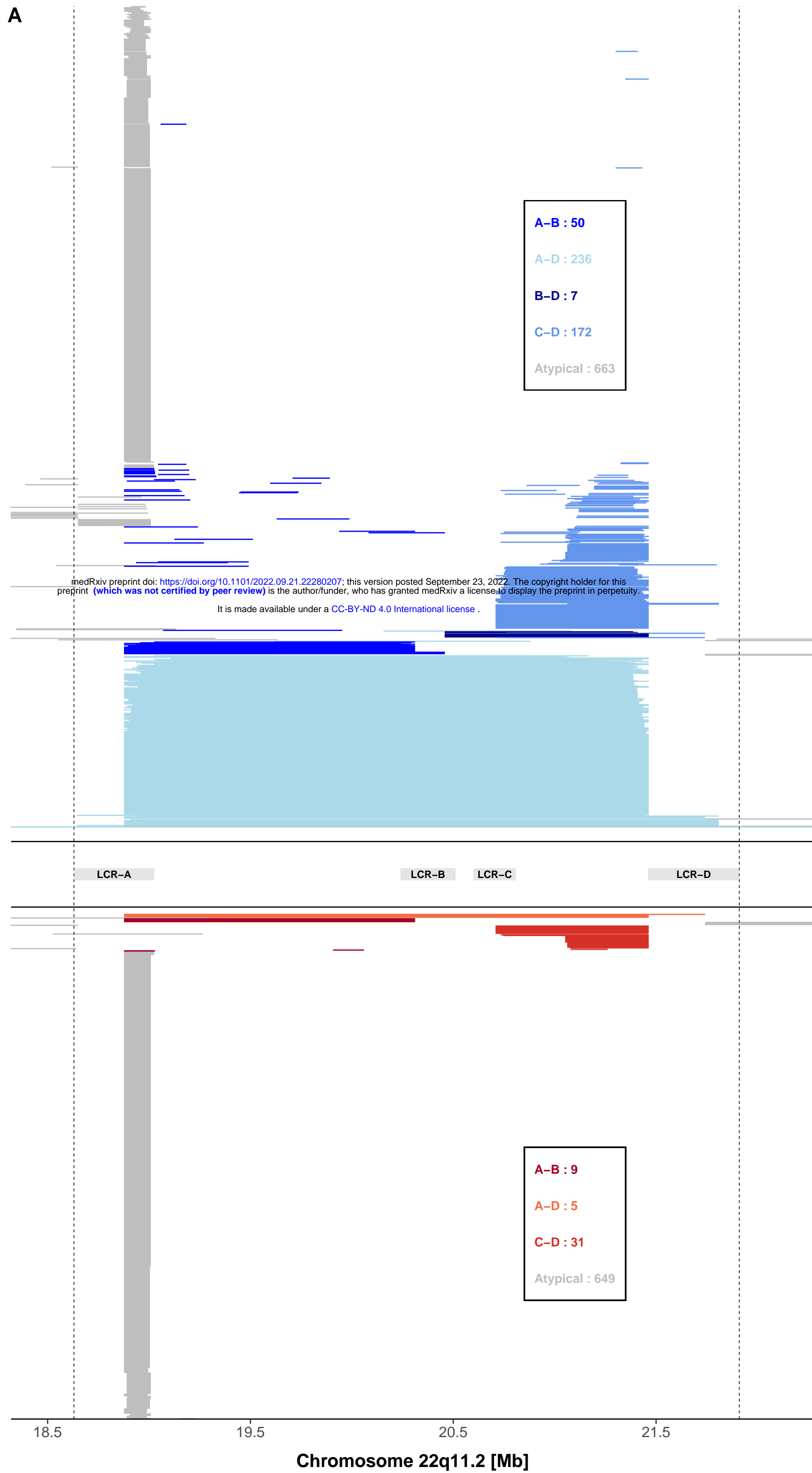
Phenotype	Genomic Position	Duplication-only			Deletion-only			U-shape			Mirror		
		$\beta$	95 % CI	p-value	$\beta$	95 % CI	p-value	$\beta$	95 % CI	p-value	$\beta$	95 % CI	p-value
Mean Platelet Volume (femtolitres)	chr22:19639383	-0.54	[-0.67,-0.41]	$1.16 \times 10^{-15}$	1.66	[0.99,2.32]	$1.13 \times 10^{-6}$	-0.46	[-0.59,-0.33]	$4.97 \times 10^{-12}$	<b>-0.58</b>	<b>[-0.71,-0.45]</b>	$1.31 \times 10^{-18}$
Body mass index (Kg/m <sup>2</sup> )	chr22:20765989	1.65	[1.15,2.16]	$1.55 \times 10^{-10}$	-0.06	[-2.23,2.12]	0.96	<b>1.56</b>	<b>[1.07,2.06]</b>	$4.9 \times 10^{-10}$	1.57	[1.08,2.06]	$4.23 \times 10^{-10}$
Whole body fat mass (kg)	chr22:20765989	3.17	[2.18,4.16]	$3.70 \times 10^{-10}$	-1.74	[-6.37,2.88]	0.46	2.95	[1.98,3.92]	$2.33 \times 10^{-9}$	<b>3.11</b>	<b>[2.14,4.07]</b>	$3.35 \times 10^{-10}$
Fluid intelligence score	chr22:19343881	-1.21	[-1.64,-0.79]	$2.25 \times 10^{-8}$	-3.76	[-6.04,-1.49]	0.001	<b>-1.3</b>	<b>[-1.72,-0.88]</b>	$1.12 \times 10^{-9}$	-1.04	[-1.46,-0.63]	$9.54 \times 10^{-7}$
Weight (kg)	chr22:20765989	3.83	[2.33,5.32]	$5.63 \times 10^{-7}$	-4.28	[-11.28,2.73]	0.231	3.47	[2.01,4.94]	$3.44 \times 10^{-6}$	<b>3.85</b>	<b>[2.38,5.31]</b>	$2.70 \times 10^{-7}$
Height (cm)	chr22:21219710	-0.6	[-1.23,0.03]	0.064	<b>-4.86</b>	<b>[-6.96,-2.77]</b>	$5.51 \times 10^{-6}$	-0.95	[-1.56,-0.35]	0.002	-0.14	[-0.75,0.46]	0.64
Height (cm)	chr22:19518079	-1.94	[-2.72,-1.15]	$1.43 \times 10^{-6}$	-6.02	[-10,-2.04]	0.003	<b>-2.09</b>	<b>[-2.86,-1.32]</b>	$1.14 \times 10^{-7}$	-1.64	[-2.41,-0.86]	$3.26 \times 10^{-5}$
Platelet Count (10 <sup>9</sup> cells/L)	chr22:19738355	16.68	[9.56,23.8]	$4.43 \times 10^{-6}$	-100.09	[-135.83,-64.35]	$4.05 \times 10^{-8}$	12.22	[5.24,19.21]	0.0006	<b>19.86</b>	<b>[12.88,26.85]</b>	$2.48 \times 10^{-8}$
Calcium level (mmol/L)	chr22:19207491	0.01	[0,0.02]	0.089	<b>-0.13</b>	<b>[-0.18,-0.08]</b>	$2.86 \times 10^{-7}$	0.003	[-0.01,0.01]	0.64	0.02	[0.01,0.03]	0.004

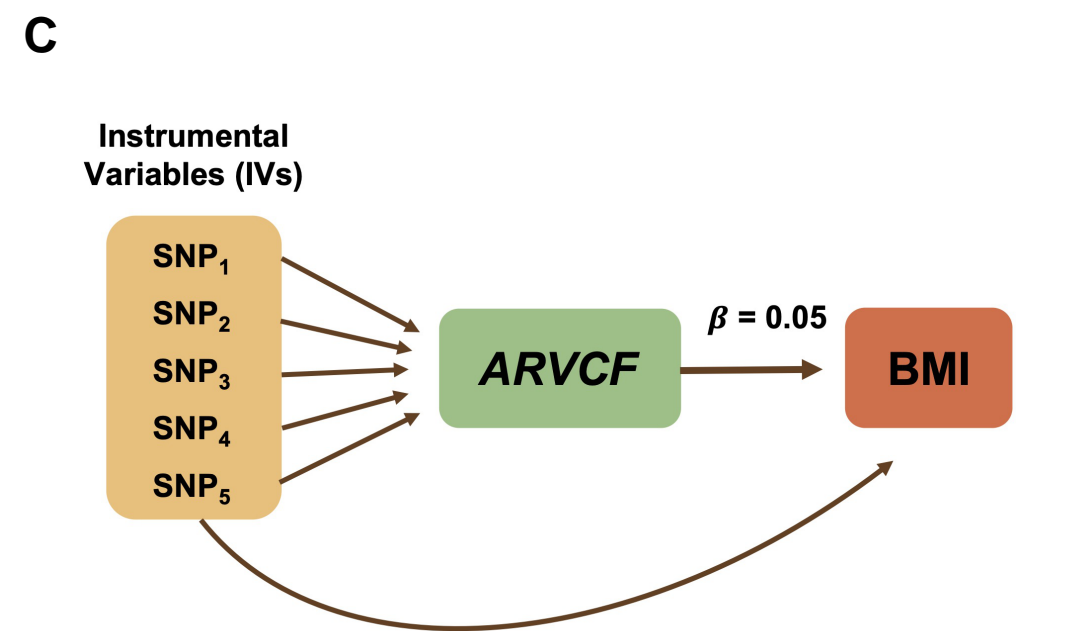
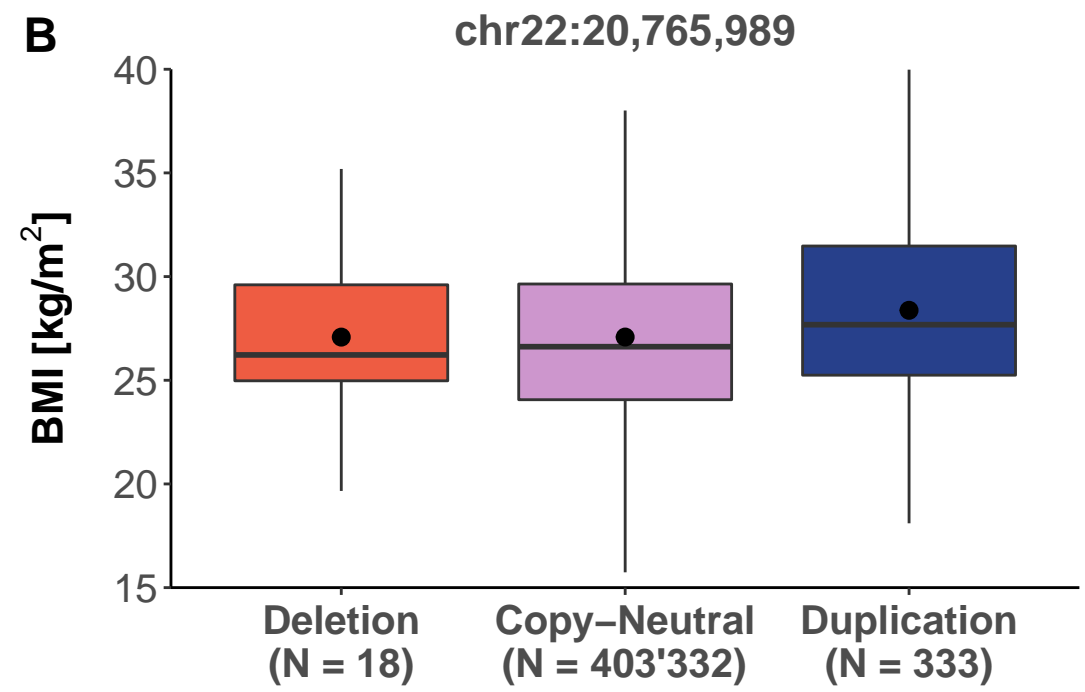
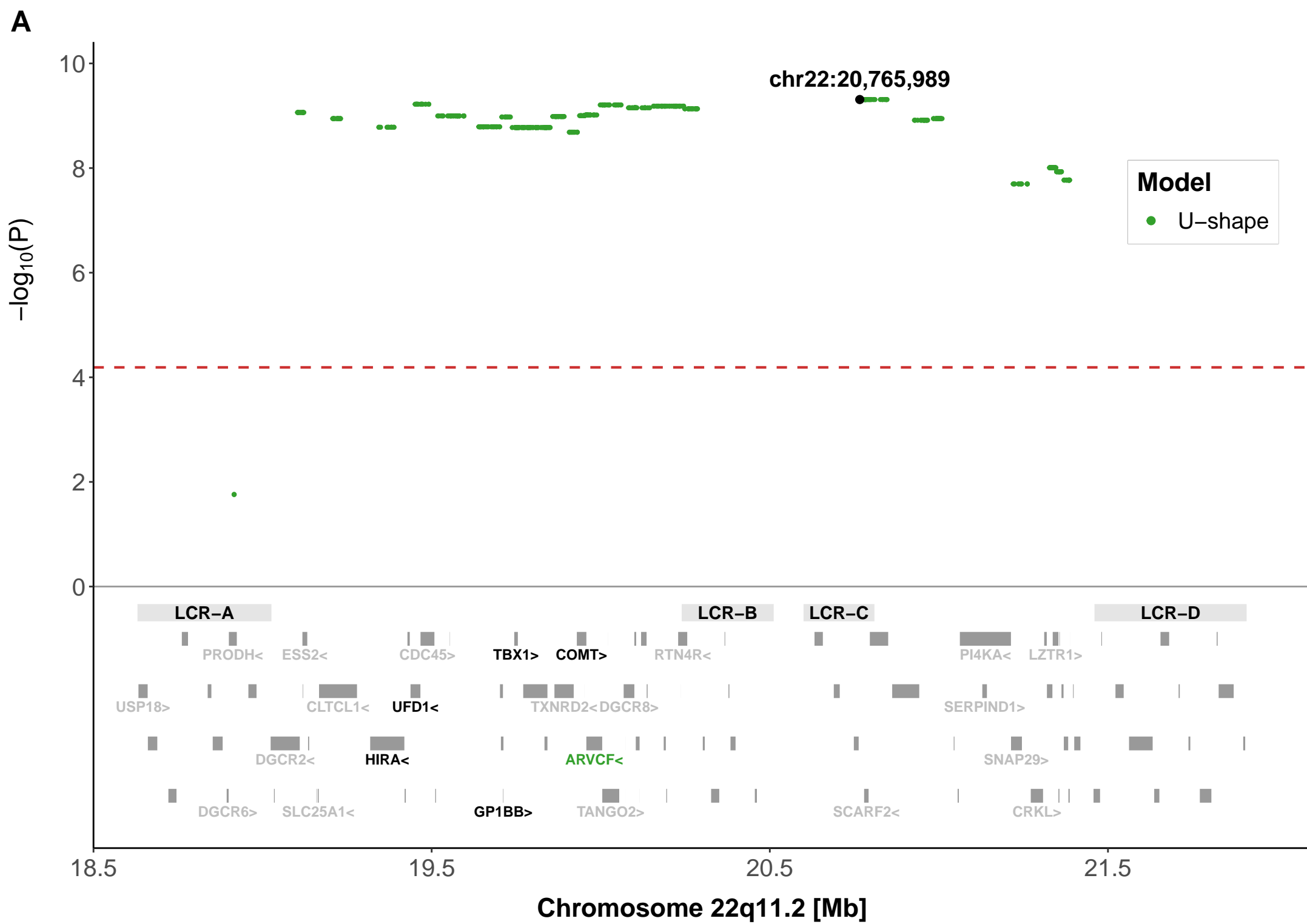
Reported effect sizes and p-values for each model are referring to the lead signal of the most relevant model for each phenotype (in bold).

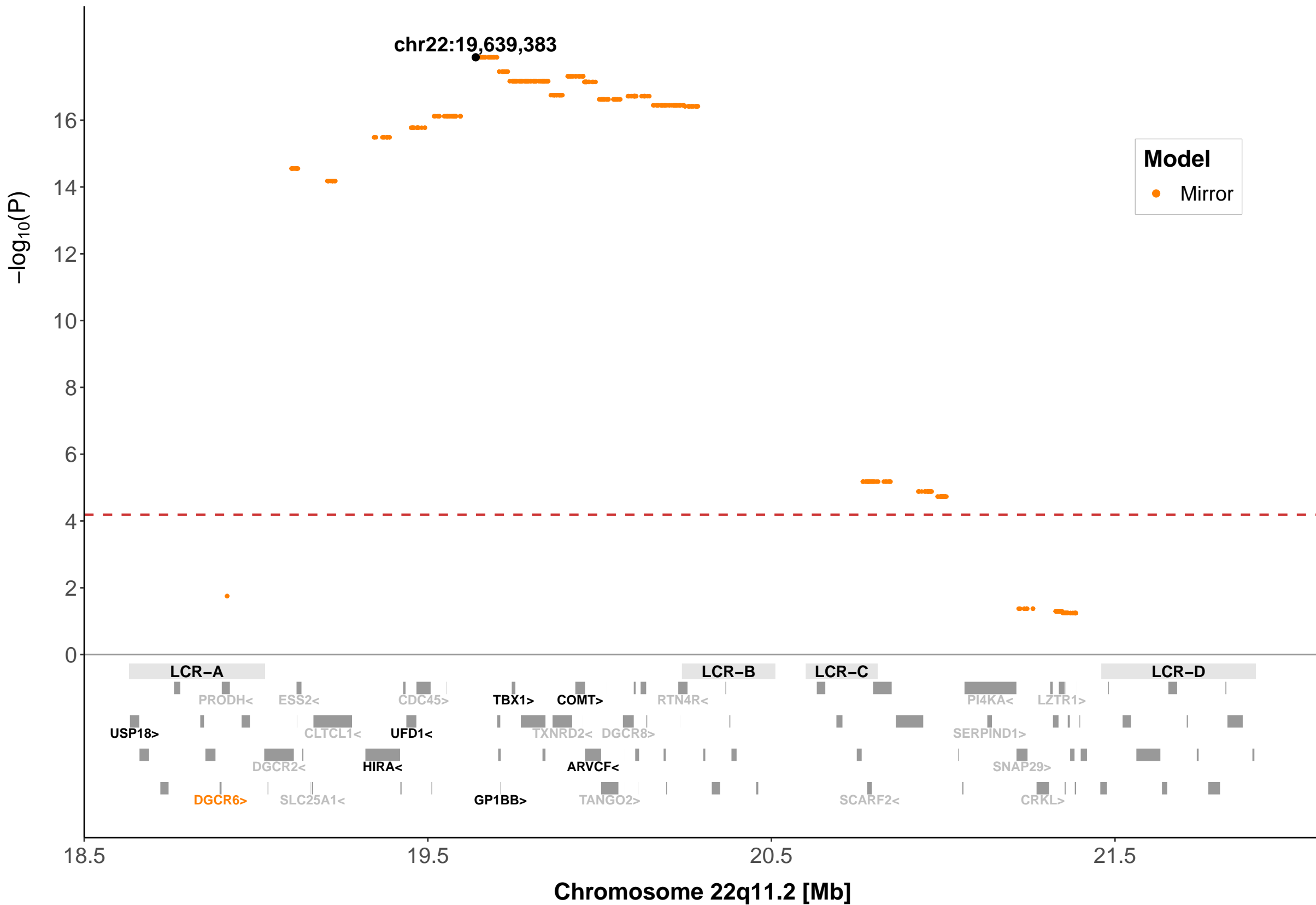
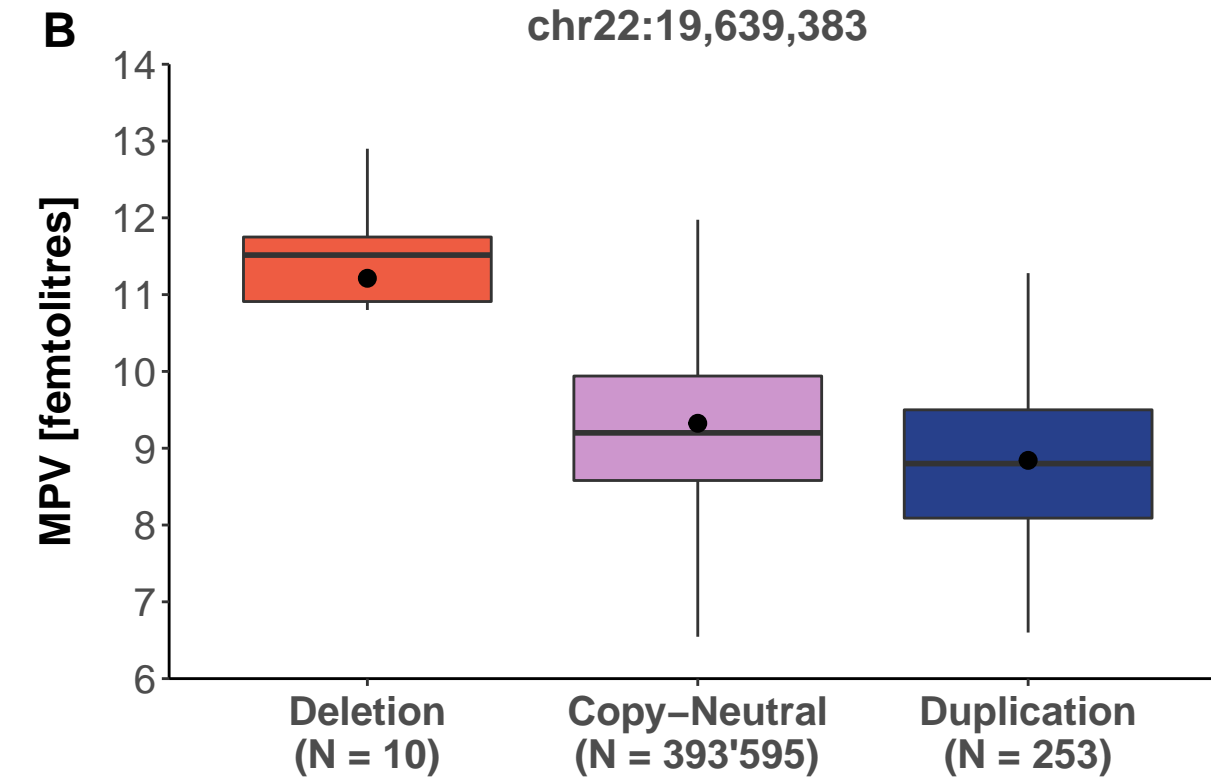
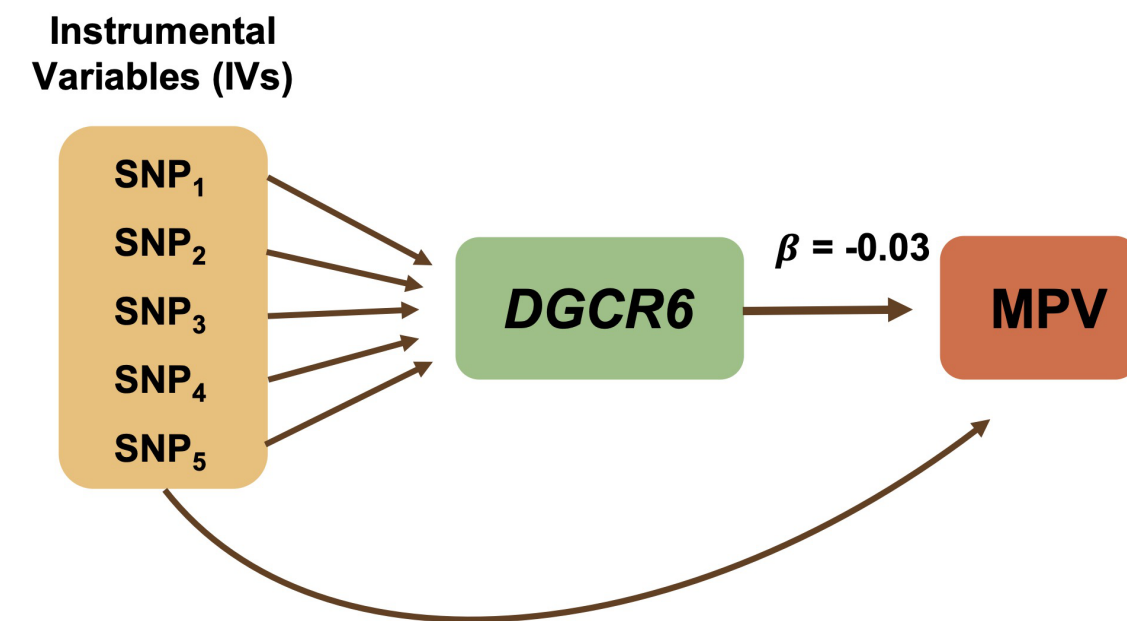
**Table 3. Binary traits associated to CNVs in the 22q11.2 region with different models.**

Phenotype	Genomic Position	Duplication-only			Deletion-only			U-shape			Mirror		
		OR	95 % CI	p-value	OR	95 % CI	p-value	OR	95 % CI	p-value	OR	95 % CI	p-value
Gastroesophageal reflux disease	chr22:19998655	<b>2.72</b>	<b>[1.91,3.88]</b>	$2.53 \times 10^{-8}$	1.65	[0.21,13.01]	0.63	2.68	[1.89,3.79]	$6.69 \times 10^{-8}$	2.66	[1.87,3.79]	$6.23 \times 10^{-8}$
Hearing loss	chr22:20082293	4.47	[2.49,8.02]	$5.32 \times 10^{-7}$	12.9	[1.58,105.24]	0.017	<b>4.71</b>	<b>[2.68,8.27]</b>	$6.95 \times 10^{-8}$	4.08	[2.22,7.5]	$5.87 \times 10^{-6}$
Cardiomegaly	chr22:21370246	<b>3.53</b>	<b>[1.92,6.47]</b>	$4.69 \times 10^{-5}$	4.78	[0.64,35.95]	0.13	3.6	[2.02,6.45]	$1.53 \times 10^{-5}$	3.21	[1.71,6.03]	0.0003
Dental caries	chr22:21370246	3.29	[1.85,5.85]	$5.21 \times 10^{-5}$	5.94	[1.4,25.12]	0.015	<b>3.51</b>	<b>[2.06,5.99]</b>	$4.21 \times 10^{-6}$	2.76	[1.48,5.13]	0.001
Diplopia and disorders of binocular vision	chr22:21219710	<b>6.23</b>	<b>[2.57,15.09]</b>	$5.18 \times 10^{-5}$	7.31	[0.43,124.7]	0.17	5.74	[2.37,13.92]	0.0001	6.24	[2.58,15.1]	$4.89 \times 10^{-5}$
Other venous embolism and thrombosis	chr22:20765989	<b>7.6</b>	<b>[2.82,20.46]</b>	$6 \times 10^{-5}$	18.57	[1.01,340.01]	0.049	7.24	[2.69,19.49]	$8.9 \times 10^{-5}$	7.61	[2.83,20.46]	$5.86 \times 10^{-5}$
Other cerebral degenerations	chr22:20927716	1.76	[0.44,7.07]	0.428	<b>45</b>	<b>[10.05,201.43]</b>	$6.43 \times 10^{-7}$	3.38	[1.26,9.1]	0.016	0.21	[0.01,3.57]	0.28
Hypotension	chr22:20927716	3.16	[1.79,5.6]	$7.7 \times 10^{-5}$	7.39	[0.89,61.65]	0.065	<b>3.3</b>	<b>[1.9,5.73]</b>	$2.16 \times 10^{-5}$	2.91	[1.61,5.25]	0.0004
Nausea and vomiting	chr22:21370246	2.17	[1.42,3.31]	0.0003	3.67	[1.09,12.37]	0.036	<b>2.28</b>	<b>[1.53,3.39]</b>	$1.34 \times 10^{-5}$	1.92	[1.24,2.98]	0.003

Reported OR and p-values for each model are referring to the lead signal of the most relevant model for each phenotype (in bold)

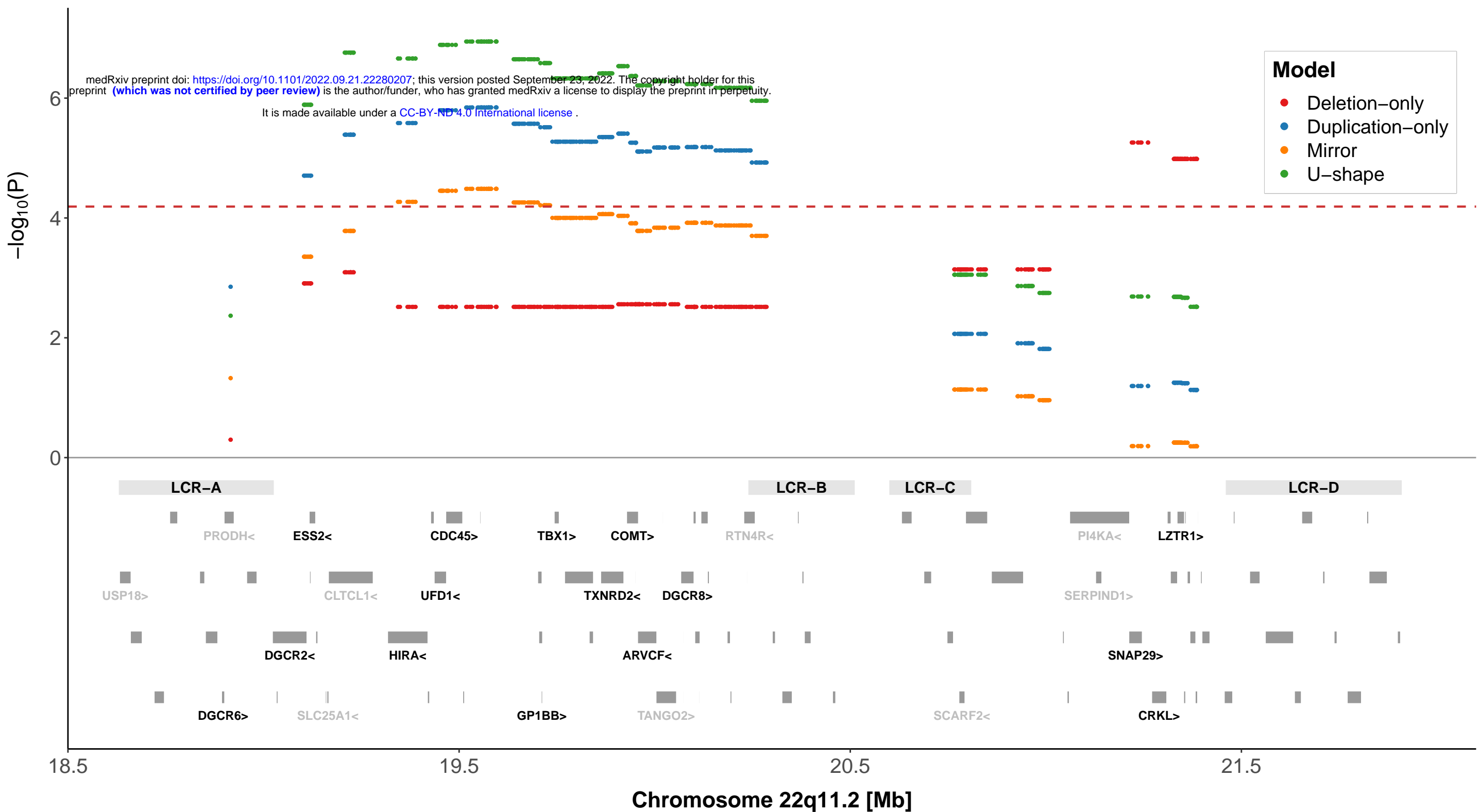




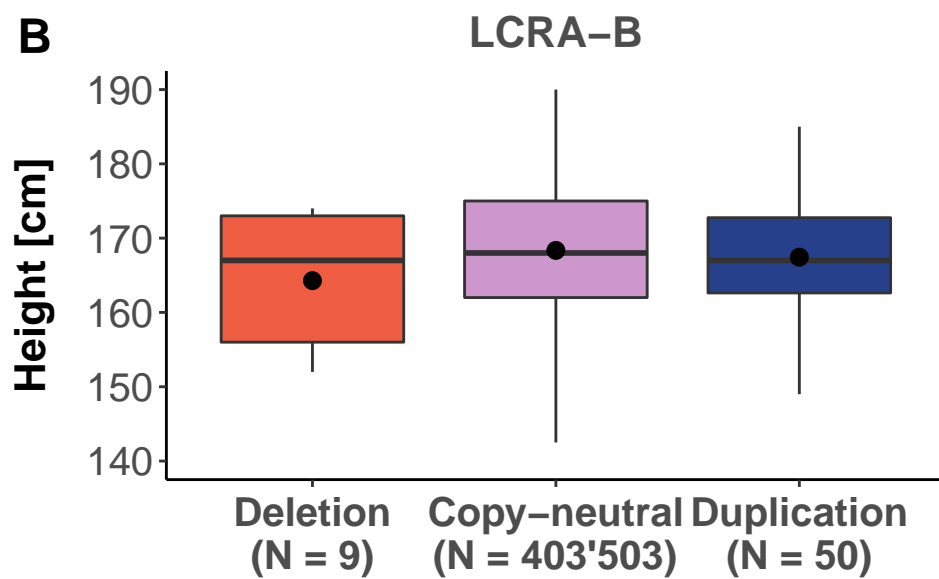
**A****B****C**



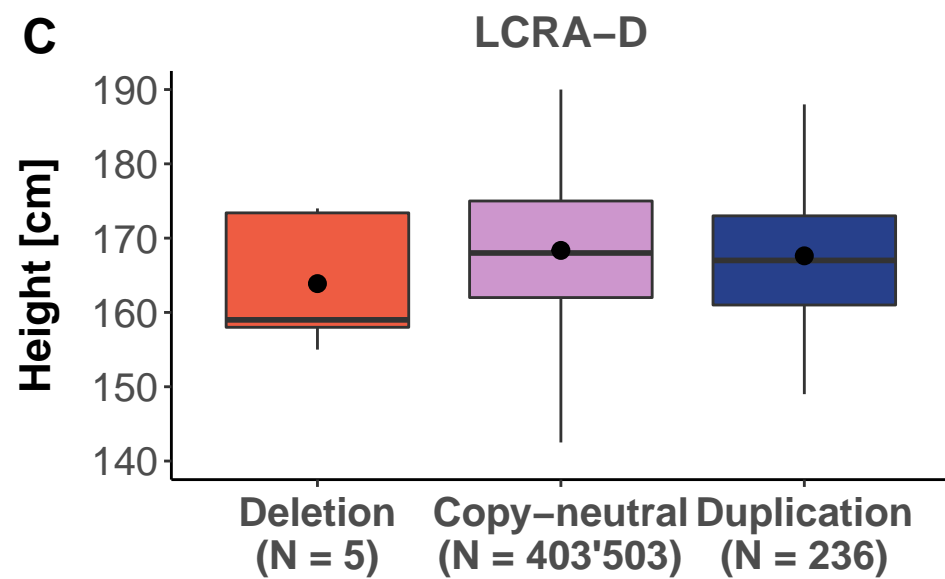
A



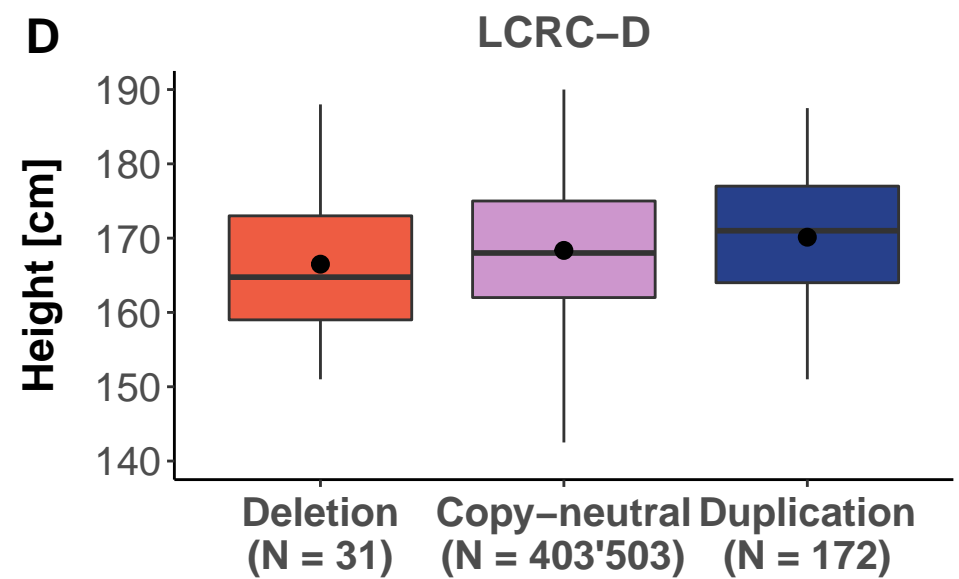
B

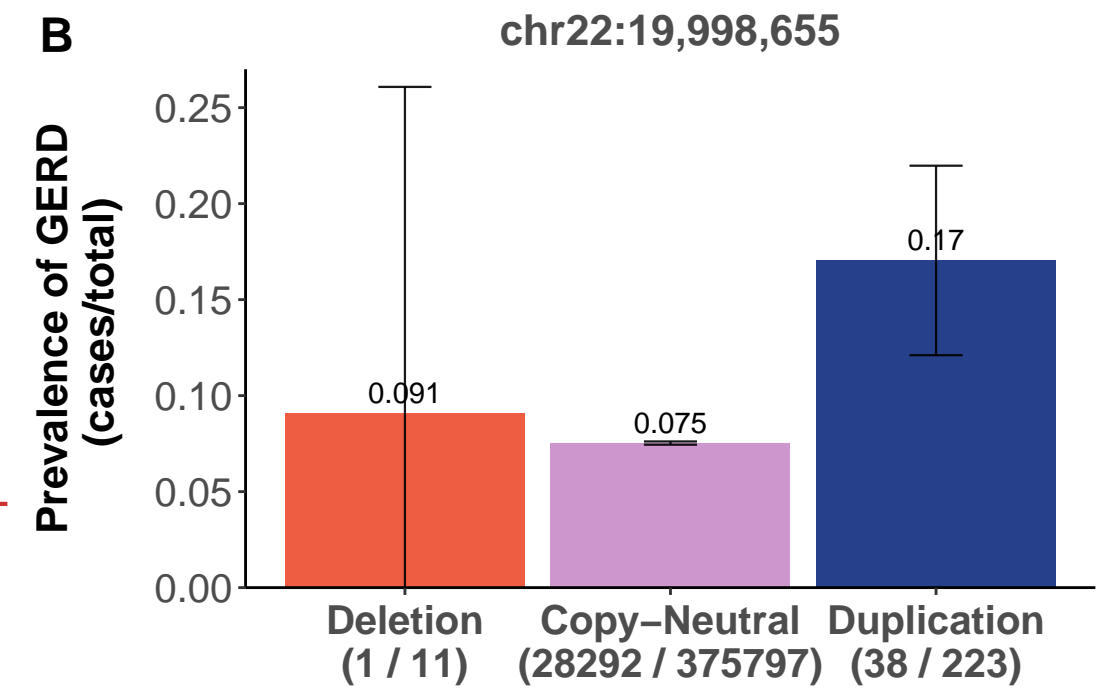
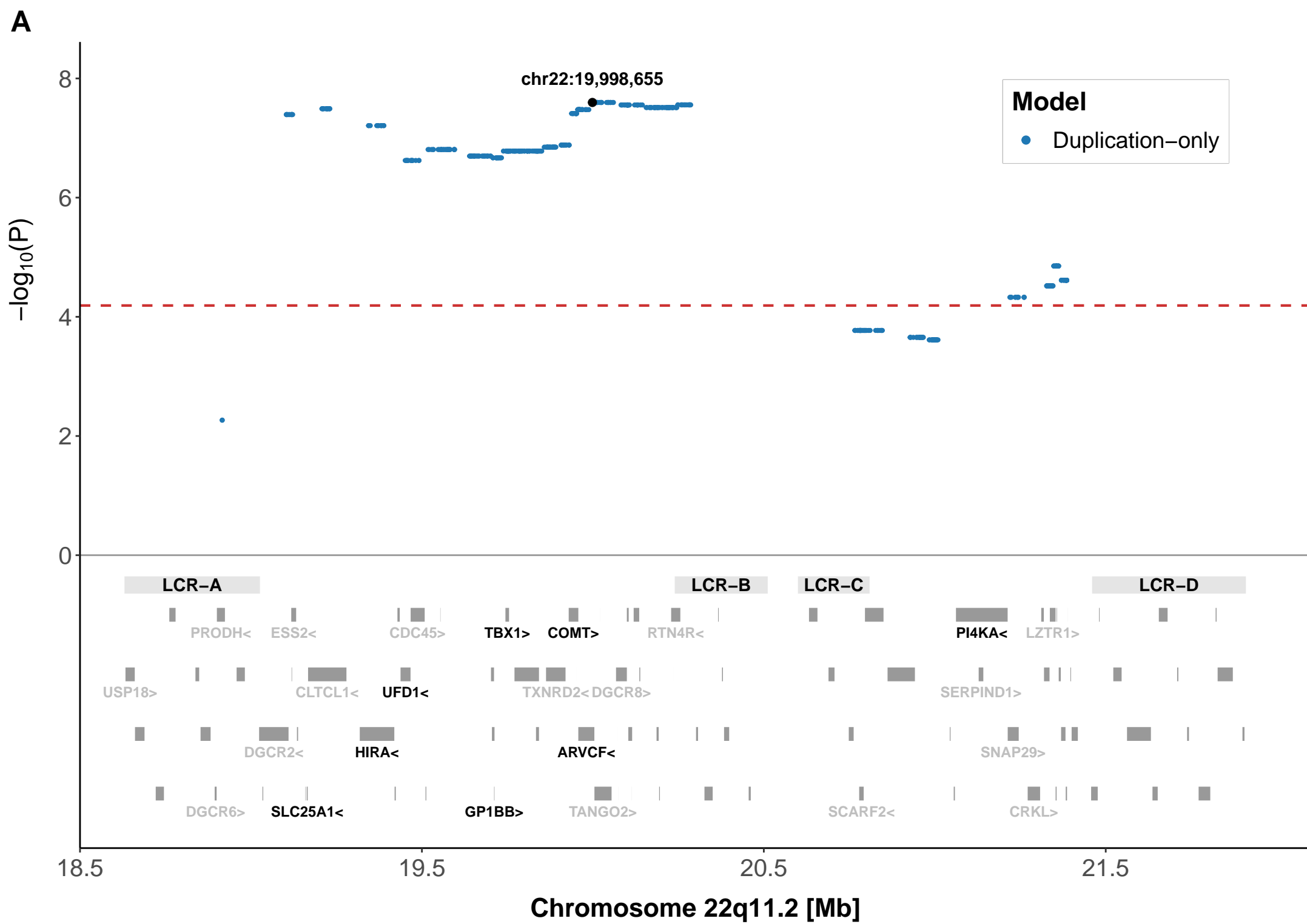


C

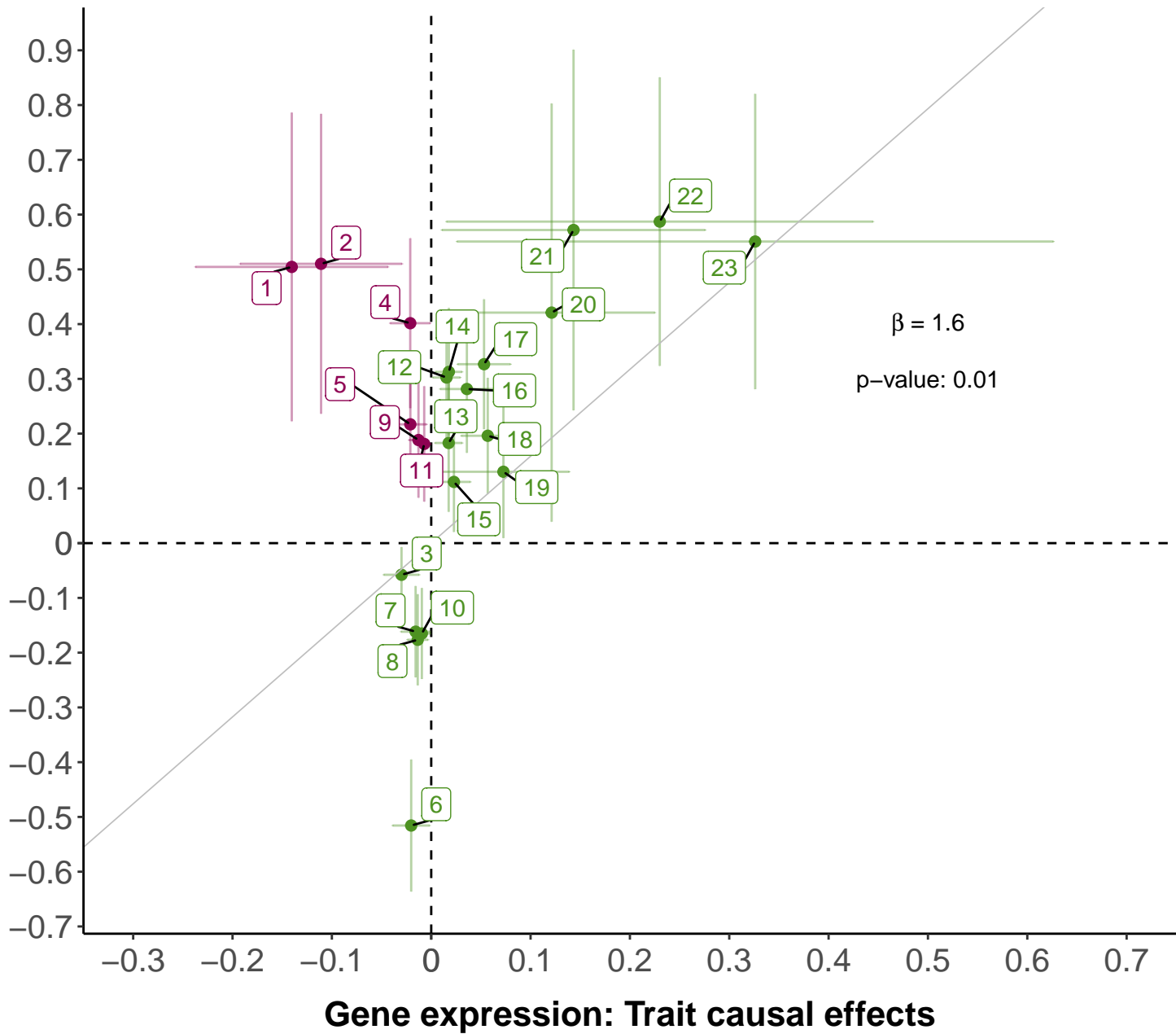


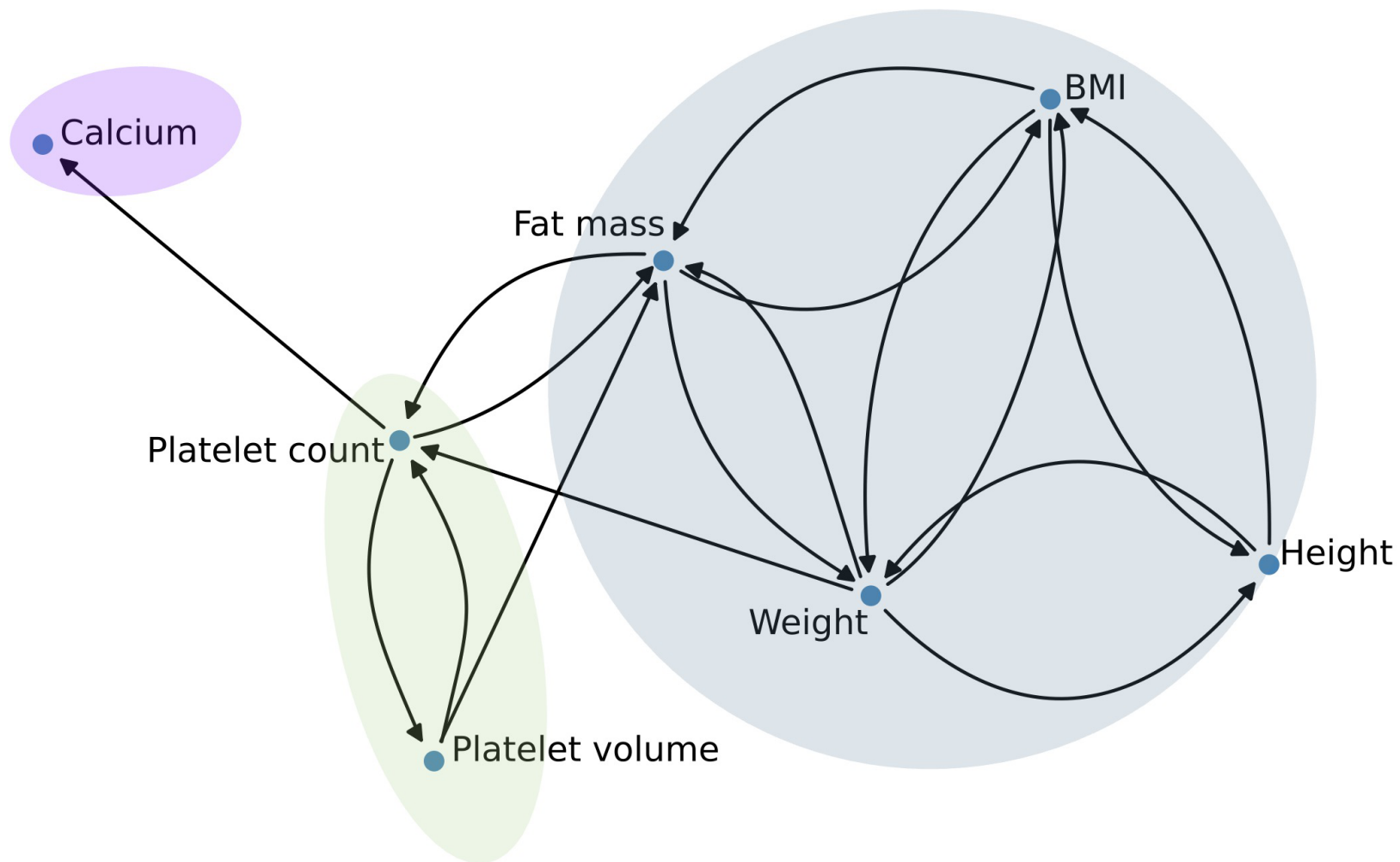
D





**Copy-number state: Trait effect**



**A****B**