# Compounding of rare pathogenic copy-number variants and polygenic background is consistent with assortative mating

Caterina Cevallos[1,2,+], Chiara Auwerx[1,2,3,4,5,6,+], Robin Hofmeister[2,3,7], Théo Cavinato[2,3], Tabea Schoeler[2,3,7], Zoltán Kutalik[2,3,7,*], and Alexandre Reymond[1,8,*]

[1]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland
[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland
[4]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[5]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[6]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[7]University Center for Primary Care and Public Health, Lausanne, Switzerland
[8]Health2030 Genome Center, Foundation Campus Biotech Geneva, Geneva, Switzerland
[+]These authors equally contributed to this work
[*]Corresponding authors

March 18, 2026

**Abstract**

Copy-number variants (CNVs) are linked to a spectrum of outcomes and carriers of the same variant exhibit variable disease severity. We explored the impact of an individual's polygenic score (PGS) on explaining these differences, focusing on 119 established CNV-trait associations involving 43 clinically-relevant phenotypes. We called CNVs among white British UK Biobank participants, then divided samples into a training set (n = 264,372) to derive independent PGS weights, and a CNV-carrier-enriched test set (n = 96,716). Assessing the individual, joint, and synergistic contribution of CNVs and PGSs, we identified a significant additive effect for 45 (38%) CNV-trait pairs, as well as two scale-dependent interactions between PGSs for gamma-glutamyltransferase levels and grip strength, and a 22q11.23 duplication encompassing multiple key genes in glutathione metabolism. A (spurious) *negative* correlation between an individual's CNV carrier status and their PGS would be expected under selective participation-induced collider bias. Instead, we observed a widespread *positive* correlation, which could only be partially accounted for by linkage disequilibrium. Given a non-null inheritance rate for all 17 testable CNVs, we explored whether assortative mating could explain this positive CNV-PGS association. We found strong agreement between this correlation and the one predicted by assortment ($r = 0.45$, $p = 2.0 \times 10^{-7}$). Similar suggestive trends of positive correlation were observed between PGSs and genome-wide burden of CNVs or rare loss-of-function variants. Our results demonstrate that PGSs contribute to the variable expressivity of CNVs and rare variants, and improve the identification of individuals at higher risk of clinically relevant comorbidities. We also highlight pervasive assortative mating as a likely mechanism contributing to the compounding of genetic effects across mutational classes.

**Keywords:** Assortative mating; Copy-number variants; Genomic Disorders; Incomplete penetrance; Polygenic score; Variable expressivity; UK Biobank.

# Introduction

Copy-number variants (CNVs) are characterized by the deletion or duplication of DNA fragments $\geq 50$ bp. In the late 2000s, genomic disorders were defined as a class of conditions caused by large recurrent CNVs, which represent among the strongest individual genetic risk factors for intellectual disability and developmental delay, psychiatric disorders, epilepsy, and congenital anomalies [1–6]. Yet, it was recognized early on that carriers of genomic disorder-associated CNVs express varying constellations of clinical features, implying their incomplete penetrance (i.e., not all carriers present all phenotypes linked with the genomic disorder) and variable expressivity (i.e., carriers show different degrees of phenotype severity) [7, 8]. This insight was strengthened by biobank studies showing that these CNVs are present in the general population – albeit at a lower frequency – and they associate with a broad range of traits [9–27], often paralleling findings from clinical cohorts – but with a lower penetrance and severity – while also revealing novel associations, showcasing the pleiotropy of these CNVs [28–32].

Phenotypic variability among CNV carriers can be explained through the liability threshold model, according to which, for each trait, an individual carries a liability that follows a normal distribution [33]. If that liability exceeds a threshold, the individual is considered a case. Provided that the impact of a CNV on disease liability is large enough to surpass this threshold, each carrier will be affected, and the CNV will be considered to be fully penetrant. Alternatively, the CNV might contribute to disease liability, but contributions of additional genetic, demographic, or environmental risk factors are required for the carrier to be affected, leading to incomplete penetrance and variable expressivity. Importantly, the impact of both the CNV and additional risk factors should be considered in a trait-specific manner, explaining phenotypic heterogeneity across carriers [32].

Early evidence of the contribution of additional genetic factors to disease liability stemmed from family-based studies [34–39]. Yet, often, detailed familial information is not available, making it is almost impossible to pinpoint the exact genetic factors contributing to increased liability, as the effect of rare high-impact variants, and the cumulative (but individually mild) effect of a large number of common variants captured by polygenic scores (PGSs) are indistinguishable [40]. The role of rare high-impact variants has been described through the two-hit model, wherein the CNV sensitizes the genome, but its expressivity is determined by additional rare variants [41]. For instance, clinically ascertained carriers of

a genomic disorder-associated CNV experience more severe and diverse phenotypic consequences if they also carry a second large CNV [42] or a high burden of damaging single-nucleotide variants (SNVs) [43]. More recently, this model was extended to assess the contribution of the polygenic background to the expressivity of rare CNVs by demonstrating their additive contribution [27, 44–56]. For example, carriers of the schizophrenia-associated 22q11.2 deletion are more likely to be diagnosed with schizophrenia if they have a high PGS for the disease [51, 57], while carriers of the body mass index (BMI)-lowering 16p11.2 BP4-BP5 duplication are less likely to be underweight if they carry a high PGS for BMI [49]. Some of these studies further explored the interplay between CNVs and PGS by testing for interaction between the mutational classes or determining whether carriers of risk-increasing CNVs have a different PGS distribution than non-carriers, yielding variable conclusions. Trends for lower disease PGS among cases carrying disease-associated CNVs have been reported, and although a large study identified a significant interaction between schizophrenia PGS and schizophrenia-associated CNVs, others failed to robustly replicate CNV-PGS interactions [27, 44, 45, 47, 49, 50, 54]. Importantly, these studies primarily rely on case-control study designs for neuropsychiatric conditions.

Here, we sought to better understand the contribution of common SNVs to the expressivity of rare CNVs across a broad range of quantitative complex traits in the UK Biobank (UKBB). Starting from 119 CNV-trait associations involving 43 distinct phenotypes and 27 CNV regions [19], we first assessed how these associations are modified by an individual's PGS. Next, to gain mechanistic insights regarding the co-occurrence of these two mutational classes in an individual, we explored three distinct theoretical models of the CNV-PGS relation (Fig. 1). First, participation bias may induce a (spurious) negative CNV-PGS correlation, as the UKBB exhibits a healthy cohort bias [58] (Fig. 1a). As such, carriers of deleterious CNVs who nevertheless take part in UKBB are expected to exhibit a PGS that counteracts the effect of the CNV, reducing expressivity and penetrance through genetic compensation. On the contrary, a positive CNV-PGS correlation is expected if the SNVs included in the PGS are in linkage disequilibrium with the CNV, *de facto* tagging the latter (Fig. 1b). An alternative mechanism explaining a positive CNV-PGS correlation is assortative mating, where the phenotypic similarity caused by one parent's PGS and the other parent's CNV results in a more likely co-inheritance of the CNV and PGS in the offspring, pushing the phenotype in the same direction (Fig. 1c).

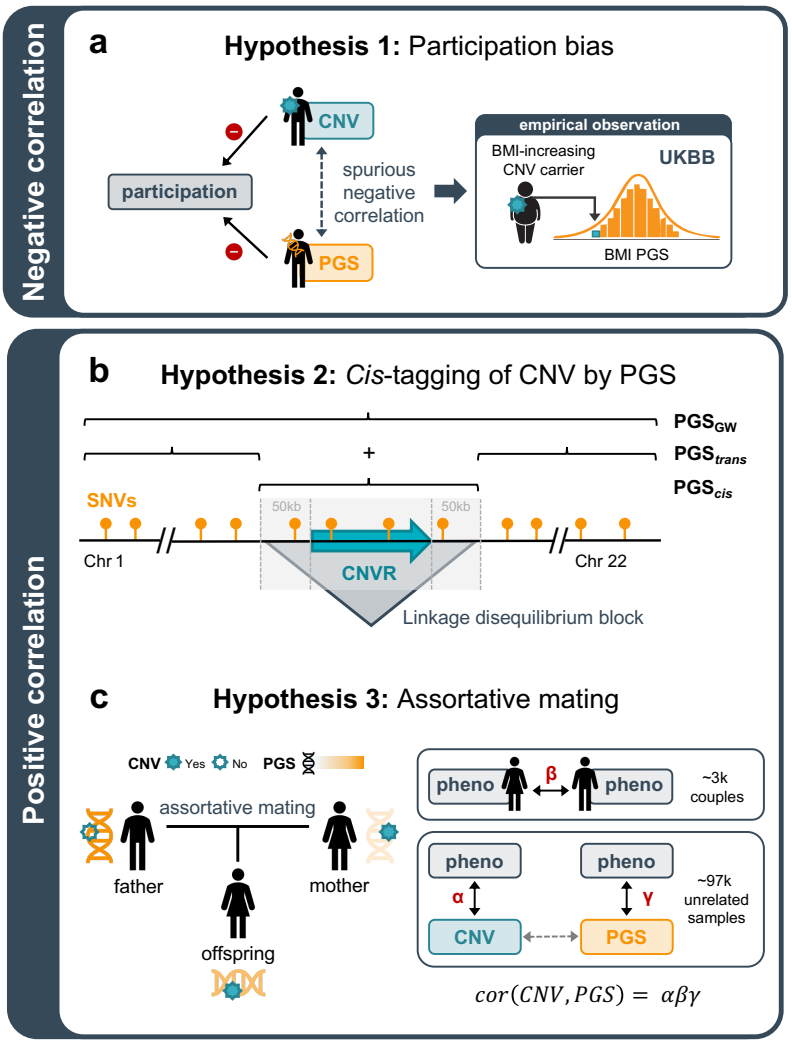


**Figure 1. Mechanisms of CNV-PGS correlation.**
Schematic representation of three possible mechanisms explaining correlation between an individual's CNV carrier status and polygenic score (PGS). **a,** Negative CNV-PGS correlation is expected in the presence of selective UK Biobank (UKBB) participation. Specifically, if both CNV carrier status and PGS predict participation in the same direction (e.g., decreasing or increasing participation probability), collider bias would result in a spurious negative CNV-PGS correlation among UKBB participants. As UKBB participants tend to be healthier than the general population [58], we expect both carriers of deleterious CNVs and individuals with a PGS that phenocopies the impact of the CNV to be less likely to participate. This leads to the observation of a lower PGS among carriers of a trait-increasing CNV compared to non-carriers, as illustrated here with a body mass index (BMI)-increasing CNV. **b,** A first mechanism explaining positive CNV-PGS correlation is tagging of the CNV by the PGS due to the CNV being in linkage disequilibrium with single-nucleotide variants (SNVs) included in the genome-wide PGS ($PGS_{GW}$). This hypothesis can be explored by subseting $PGS_{GW}$ into $PGS_{cis}$, which encompasses SNVs mapping to a padded CNV region (e.g., $\pm$ 50 kb), and $PGS_{trans}$, which encompasses all SNVs that do not map to that region. **c,** An alternative mechanism for positive CNV-PGS correlation is assortative mating, where two individuals exhibiting similar phenotypes due to different underlying genetic etiologies, e.g., a high PGS in the father and a trait-increasing CNV in the mother, produce offspring that carry both genetic risk factors. This hypothesis can be explored by assessing the fraction of empirically observed CNV-PGS correlation that can be explained by an indirect path mediated by assortative mating, derived from multiplying the CNV-trait correlation ($\alpha$), the within-couple phenotypic correlation ($\beta$), and the PGS-trait correlation ($\gamma$).

# Results

## Study material overview

To assess the role of genetic background in shaping the expressivity of CNVs, we selected 119 genome-wide significant autosomal CNV-trait associations that we previously identified in the UKBB [19] (Supplementary Table 1). The 119 signals involved 43 quantitative phenotypes describing a broad range of physiological functions, including musculoskeletal (n = 5), metabolic (n = 11), neuropsychiatric (n = 2), cardiopulmonary (n = 4), hepatic, (n = 7), hematologic (n = 8), endocrine (n = 2), and renal (n = 4) measurements (Supplementary Table 2). Each signal involves a lead CNV-proxy probe whose copy-number most significantly associated with the trait through either of three association models: a mirror model, in which the deletion and duplication impact the phenotype in an opposite direction, and two models considering the effect of the deletion or duplication in isolation. White British UKBB participants carrying a CNV overlapping any of the lead probes were identified and CNVs were encoded according to the previously determined best association model [19]. We then divided samples into a training set (n = 264,372) that we used to derive genome-wide PGS ($PGS_{GW}$) weights, and a CNV-carrier enriched test set of unrelated individuals (n = 96,716) for which $PGS_{GW}$ were computed (see Methods; Supplementary Fig.1). The incremental phenotypic variance explained by $PGS_{GW}$ compared to a baseline model ($\Delta R^2$) ranged between 2.3-24.6% (median = 11.6% [IQR = 7.9%]). Four traits had an $\Delta R^2$ over 20% – total bilirubin (24.6%), mean corpuscular hemoglobin (22.5%), platelet count (23.0%), and alkaline phosphatase (21.3%) – and four traits had a $\Delta R^2$ under 5% – waist-to-hip ratio (WHR; 4.0%), WHR adjusted for BMI (3.9%), neuroticism (2.9%), and grip strength (2.1%) (Supplementary Fig.2; Supplementary Table 2).

## $PGS_{GW}$ and CNVs additively contribute to phenotypic variability

We jointly modeled the phenotypic impact of CNVs and $PGS_{GW}$, finding that for all 119 CNV-trait pairs, both CNV and $PGS_{GW}$ significantly contributed to phenotypic variability ($p \leq 0.05/119 = 4.2 \times 10^{-4}$; Supplementary Table 3). We then estimated the effect of $PGS_{GW}$ among carriers of trait-associated CNVs, detecting 45 (38%) Bonferroni-significant effects and an additional 30 nominally significant effects, distributed across all eight trait categories (Supplementary Fig. 3; Supplementary Table 4). The likelihood of detecting a significant $PGS_{GW}$ effect strongly depended on the number of carriers assessed ($p = 1.1 \times 10^{-6}$) and the predictive performance of the $PGS_{GW}$ ($p = 1.1 \times 10^{-4}$). As such, a single CNV-trait pair

with less than 100 carriers had a Bonferroni-significant effect of $PGS_{GW}$ on the adjusted phenotype. As 55 (46%) CNV-trait pairs have less than 100 CNV carriers, we expect that larger cohorts of CNV carriers will reinforce evidence of the additive contribution of CNVs and PGSs to phenotypic variability. We next illustrate the clinical relevance of these findings through a series of examples (Fig. 2; Extended Data 1).

16p11.2 BP4-BP5 CNVs exhibit genuine pleiotropy directly affecting multiple physiological systems, with highly heterogeneous phenotypic presentation [31, 32]. We found a nominally significant contribution of $PGS_{GW}$ to the variable expressivity of 14 (58%) out of 24 phenotypes associated to CNVs in the region. Example phenotypes include height ($\beta_{DEL}$ = -8.2 cm; $p_{DEL} = 1.0 \times 10^{-31}$; among deletion carriers: $\beta_{PGS}$ = 8.7 cm; $p_{PGS} = 1.3 \times 10^{-10}$; Fig. 2a), BMI ($\beta_{DEL}$ = 6.2 kg/m$^2$; $p_{DEL} = 8.4 \times 10^{-35}$; among deletion carriers: $\beta_{PGS}$ = 4.8 kg/m$^2$; $p_{PGS}$ = 0.044; Fig. 2b), or platelet count ($\beta_{mirror}$ = -38.8 $10^{-9}$ cells/L; $p_{mirror} = 2.5 \times 10^{-21}$; among CNV carriers $\beta_{PGS} = 42.4 \times 10^{-9}$ cells/L; $p_{PGS} = 6.7 \times 10^{-5}$; Fig. 2c). Of note, we previously showed that the effect of the CNV on platelet count was independent of the carrier's BMI and height [31]. Here, we propose that the CNV effect on platelet count might be exacerbated by the carrier's $PGS_{GW\text{-platelet count}}$, positioning the polygenic background as a key determinant of the phenotype constellation expressed by a given 16p11.2 BP4-BP5 CNV carrier.

Laboratory measurements can reach extreme values in individuals carrying CNVs in genes linked to the biomarker's transport or metabolism, if they concurrently have a high $PGS_{GW}$. This is the case for total bilirubin levels among carriers of a 12p12.2-12p12.1 deletion with a high $PGS_{GW\text{-bilrubin}}$ ($\beta_{DEL}$ = 3.0 umol/L; $p_{DEL} = 5.9 \times 10^{-15}$; among deletion carriers: $\beta_{PGS}$ = 7.1 umol/L; $p_{PGS} = 6.5 \times 10^{-7}$; Fig. 2d). The CNV region encompasses *SLCO1B1* and *SLCO1B3*, two genes encoding hepatic bilirubin transporters [59] whose mutational load has been described to form allelic series for bilirubin levels [18, 19]. Another example relates to gamma-glutamyltransferase (GGT) levels among carriers of a 22q11.23 duplication with high $PGS_{GW\text{-GGT}}$ ($\beta_{DUP}$ = 34.7 U/L; $p_{DUP} = 1.9 \times 10^{-31}$; among duplication carriers: $\beta_{PGS}$ = 62.0 U/L; $p_{PGS} = 4.3 \times 10^{-6}$; Fig. 2e). The duplicated region harbors multiple genes involved in glutathione metabolism, including gamma-glutamyltransferase 1 (*GGT1*) and 5 (*GGT5*). As such, a substantially higher fraction of duplication carriers presents abnormally elevated ($>$ 50 U/L) GGT levels, compared to copy-neutral individuals (37% vs 17%; $p = 8.3 \times 10^{-15}$), with a prevalence of 71% among duplication carriers in the top $PGS_{GW\text{-GGT}}$ quintile. These observations indicate that integrating CNV and PGS

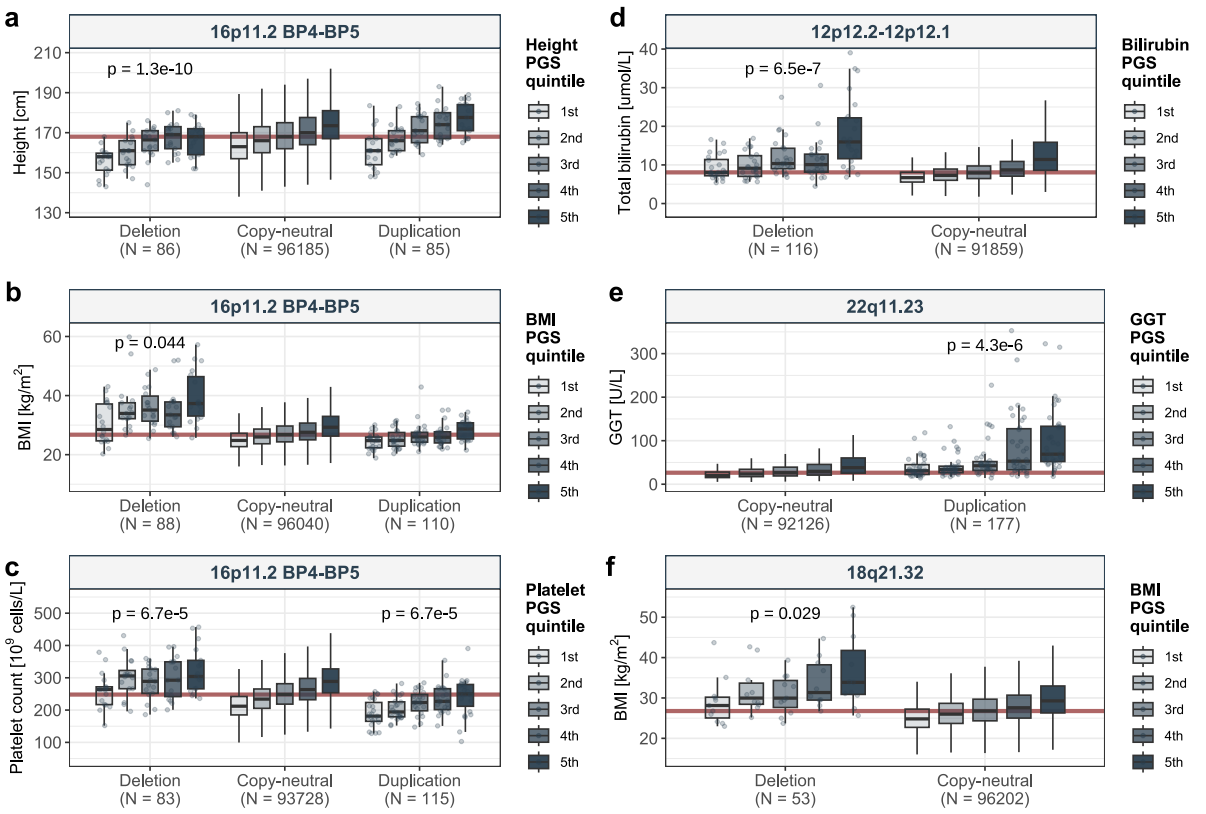


**Figure 2. Additive contribution of $PGS_{GW}$ and CNV to phenotypic variability.**
Boxplots showing the phenotypic distribution (y-axis) of individuals according to their genome-wide polygenic score ($PGS_{GW}$) quintile and CNV carrier status (x-axis) for **a,** height and 16p11.2 BP4-BP5, **b,** body mass index (BMI) and 16p11.2 BP4-BP5, **c,** platelet count and 16p11.2 BP4-BP5, **d,** total bilirubin levels and 12p12.2-12p12.1, **e,** gamma-glutamyl transferase (GGT) and 22q11.23, and **f,** BMI and 18q21.32. Individual data points are shown for CNV carriers as grey dots. The red line represents the median trait value among copy-neutral individuals. Significance of the $PGS_{GW}$'s effect on phenotypes among carriers of the trait-associated CNV is shown above the assessed CNV genotype groups, whose sample size is indicated as N. Only CNV genotype categories with $\geq$ 15 individuals are represented.

information could guide the interpretation of clinical test results by identifying individuals with a genetic predisposition to increased biomarker levels.

Finally, despite showing robust trends and strong literature support, some $PGS_{GW}$ effects fail to reach significance, often because of small sample size and/or low $PGS_{GW}$ predictive performance. For instance, $PGS_{GW\text{-}BMI}$ nominally contributes to adiposity among the 53 carriers of a BMI-associated 18q21.32 deletion ($\beta_{DEL}$ = 4.2 kg/m$^2$; $p_{DEL}$ = 1.6 x $10^{-10}$; among deletion carriers: $\beta_{PGS}$ = 8.4 kg/m$^2$; $p_{PGS}$ = 0.029; Fig. 2f). This intergenic deletion is upstream of *MC4R*, a gene involved in energy homeostasis and appetite regulation, which has been associated with monogenic forms of obesity [60]. Similarly, the extent of muscle wasting among carriers of the 17p12 duplication associated with the degenerative neuropathy of Charcot-Marie-Tooth type 1A [61, 62] is nominally reduced in duplication carriers with a protective $PGS_{GW\text{-grip strength}}$ ($\beta_{DUP}$ = -10.1 kg; $p_{DUP}$ = 3.8 x $10^{-48}$; among duplication carriers: $\beta_{PGS}$ =

13.4 kg; $p_{PGS}$ = 6.3 x $10^{-3}$). Hence, $PGS_{GW}$ can help identify CNV carriers that are at the highest risk of developing severe CNV expression and would benefit the most from lifestyle and/or therapeutic interventions. Together, these examples illustrate how the polygenic background contributes to phenotypic heterogeneity and variable expressivity of pathogenic CNVs across a broad spectrum of traits.

## CNV-$PGS_{GW}$ interactions are rare

After establishing the additive role of CNVs and $PGS_{GW}$, we explored whether the effect of $PGS_{GW}$ depended on an individual's CNV carrier status. While we did not detect any CNV-$PGS_{GW}$ interaction surviving Bonferroni correction ($p \leq 0.05/119 = 4.2 \times 10^{-4}$) using adjusted phenotypes, two Bonferroni-significant interactions were identified when analyzing raw phenotypic values (Supplementary Table 5). These two scale-dependent interactions both mapped to the previously described 22q11.23 region (Fig. 2e) and revealed a stronger contribution of the polygenic background to increased GGT levels ($\beta_{interaction}$ = 35.2 U/L; $p = 2.0 \times 10^{-5}$) and decreased grip strength ($\beta_{interaction}$ = -12.0 kg; $p = 1.4 \times 10^{-4}$) among CNV carriers. Importantly, we confirm that these interactions are not artifacts of poor genotype quality, as both remain Bonferroni-significant when using a $PGS_{trans}$ that excludes SNVs within $\pm$ 250 kb of the CNV region ($p_{GGT} = 3.4 \times 10^{-5}$; $p_{grip\ strength} = 1.4 \times 10^{-4}$). In a second time, we repeated the analysis substituting CNV carrier status for an individual's genome-wide deletion or duplication burden, defined as the number of genes disrupted by deletions or duplications, respectively. We did not detect any Bonferroni-significant interactions on either the normalized or raw phenotypic scale ($p \leq 0.05/(2 \times 43) = 5.8 \times 10^{-4}$) (Supplementary Table 6).

Failure to detect interactions might stem from limited power. Indeed, within the empirically observed range of CNV frequencies ($5 \times 10^{-5}$ to $5 \times 10^{-3}$) and interaction effect sizes on the adjusted phenotype scale ($\leq 0.012$), our power to detect an interaction is virtually null (Supplementary Fig. 4). On the raw scale, we observe larger effect sizes, with half of the CNV-trait pairs having an interaction term with an effect size $\geq 1$ (Supplementary Table 5). This improves detection power, possibly explaining why significant interactions were detected on that scale. We conclude that while we detected two scale-dependent interactions wherein the impact of the polygenic background was exacerbated among CNV carriers, there is, in most cases, no evidence to support that the polygenic background impacts complex traits differently among CNV carriers.

## Carriers of trait-increasing CNVs have higher PGS

Given the healthy cohort bias of UKBB, we hypothesized that CNV carriers have a polygenic background that counteracts the CNV's effect, resulting in carriers having a different $PGS_{GW}$ distribution than copy-neutral individuals (Fig. 1a). Strikingly, 90 (76%) CNV-trait pairs had a CNV effect on $PGS_{GW}$ whose direction *aligned* with the CNV effect on the adjusted phenotype ($p < 1 \times 10^{-3}$, based on 1,000 permutations; Fig. 3a; Supplementary Fig. 5; Extended Data 2; Supplementary Table 7). Focusing on 11 CNV-trait pairs with a nominally significant CNV effect on $PGS_{GW}$, ten were directionally concordant, indicating that CNV carriers tend to have a $PGS_{GW}$ that exacerbates the CNV's phenotypic impact.

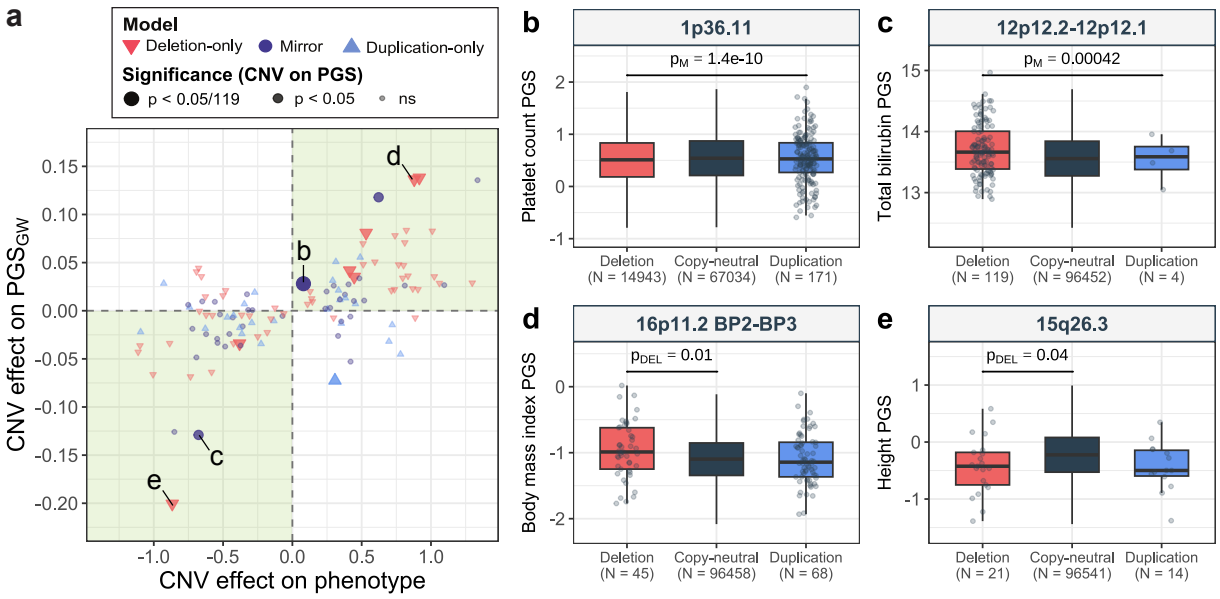


**Figure 3. CNV carriers harbor a $PGS_{GW}$ that reinforces the effect of the CNV.**
**a,** Effect of CNV carrier status on genome-wide PGS ($PGS_{GW}$) (y-axis) against the effect on adjusted phenotypes (x-axis) for 119 CNV-trait pairs. Shape and color indicate the assessed model, while size and transparency reflect the significance of the CNV effect on the $PGS_{GW}$, with "ns" indicating $p > 0.05$. Green areas encompass CNV-trait pairs where carriers of a trait-associated CNV have a $PGS_{GW}$ that is directionally concordant with the phenotypic effect of the CNV. CNV-trait pairs depicted in **b-e** are labeled. **b-e,** Boxplots representing the distribution of $PGS_{GW}$ (y-axis) for **b,** platelet count, **c,** total bilirubin, **d,** body mass index (BMI), and **e,** height according to CNV carrier status at **b,** 1p36.11, **c,** 12p12.2-12p12.1, **d,** 16p11.2 BP2-BP3, and **e,** 15q26.3 (x-axis). The number of individuals is indicated as N. Individual data points are shown for CNV genotype categories with $N \leq 1{,}000$. Significance of the CNV effect on $PGS_{GW}$ encoded according to the mirror (M) or deletion-only (DEL) model is displayed.

The copy-number of the 1p36.11 region encompassing the *RHD* gene – which encodes the Rhesus (Rh) factor antigen key in determining the Rh blood group – showed a Bonferroni-significant positive effect ($p \leq 0.05/119 = 4.2 \times 10^{-4}$) on both platelet count and $PGS_{GW\text{-platelet count}}$ ($p = 1.4 \times 10^{-10}$; Fig. 3b). Other nominally significant examples include the association between the 12p12.2-12p12.1 copy-number and both decreased levels of total bilirubin and lower $PGS_{GW\text{-bilirubin}}$ ($p = 4.205 \times 10^{-4} > 0.05/119$;

Fig. 3c). Similarly, higher $PGS_{GW\text{-}BMI}$ was observed among carriers of the BMI-increasing 16p11.2 BP2-BP3 deletion (p = 0.012; Fig. 3d), which encompasses the obesity-linked gene *SH2B1* [30, 63]. Finally, lower $PGS_{GW\text{-}height}$ was observed among carriers of a 15q26.3 partial deletion of *ADAMTS17* (p = 0.042; Fig. 3e). *ADAMTS17* is linked to autosomal recessive Weill-Marchesani syndrome, which is characterized by short stature [64]. Despite individual effects being weak, the observed data pattern appears conserved across a broad spectrum of CNV-trait pairs. Importantly, it is incompatible with strong selective participation bias, which would induce a negative (spurious) correlation between CNV and $PGS_{GW}$ (Fig. 1a).

## *Cis*-tagging only partially explains the positive correlation between CNV-$PGS_{GW}$

A first hypothesis explaining the positive CNV-$PGS_{GW}$ correlation is tagging of the CNV by $PGS_{GW}$, so that the presence of the CNV is positively correlated through linkage with SNVs included in $PGS_{GW}$ (Fig. 1b). To test this, we decomposed $PGS_{GW}$ into a $PGS_{cis}$ based solely on SNVs mapping to the CNV region $\pm$ 50 kb, $\pm$ 100 kb, $\pm$ 250 kb, and the full chromosome and conversely, a $PGS_{trans}$ excluding SNVs that mapped to these intervals. We then calculate the Pearson correlation between these partial PGSs and the CNV carrier status (Supplementary Table 8). Focusing on ten CNV-trait pairs with a nominally significant effect of the CNV carrier status on $PGS_{GW}$ whose direction aligned with the CNV's effect on the phenotype, the association between the 1p36.11 copy-number and increased platelet count stood out by showing a strong, Bonferroni-significant correlation between CNV status and $PGS_{cis}$, indicating that the CNV-$PGS_{GW}$ was primarily driven by local linkage disequilibrium (Fig. 4). Notably, this was the only signal for which the phenotypic impact of the CNV was significantly diminished upon conditioning on $PGS_{GW}$ (Supplementary Fig. 6; Supplementary Table 3). With a frequency of 3.8%, the *RHD* deletion is the only common CNV in our study. The region was highlighted by the Human Pangenome Reference Consortium for its extreme complexity, with a frequency of 17% for the haplotype carrying the full *RHD* deletion [65], suggesting that our CNV calls underestimate the true deletion frequency. Still, the fact that this locus is the only one with robust evidence of tagging aligns with such a mechanism relying on the CNV being present on (a) specific haplotype(s) – as opposed to appearing recurrently *de novo* – allowing the formation of a correlation pattern between the CNV and SNVs.

We next turned to the remaining nine signals (Fig. 4), of which three showed at least a nominally

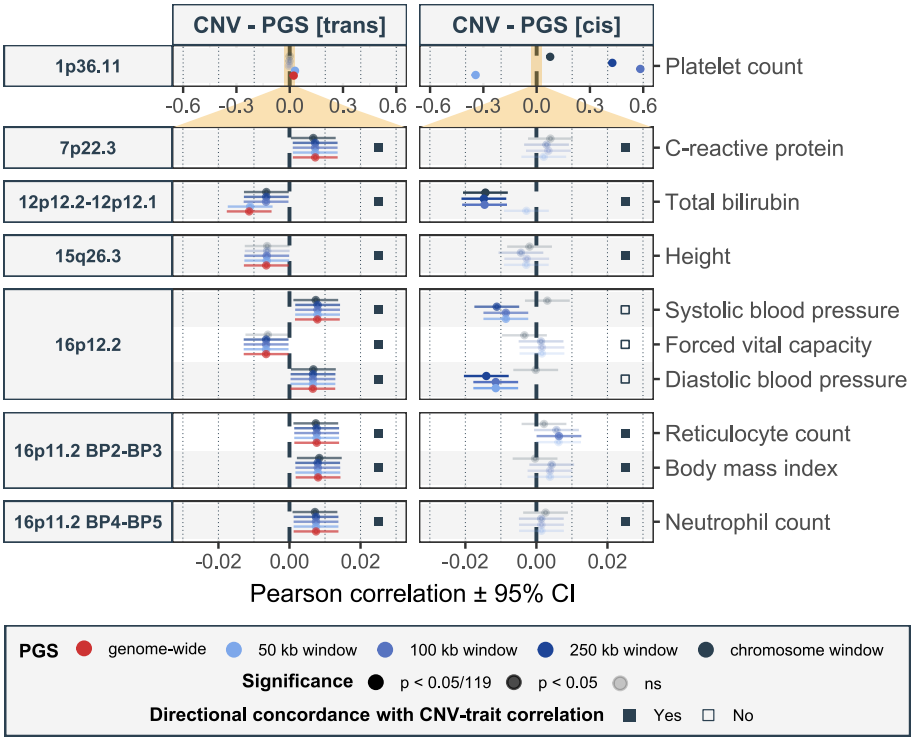


**Figure 4. *Cis*-tagging only partially explains the positive correlation between CNV-PGS$_{GW}$.**
Pearson correlation with 95% confidence interval (CI; x-axis) between polygenic scores (PGSs; y-axis, right) and CNV carrier status (y-axis, left) for ten CNV-trait pairs with a nominally significant effect of the CNV carrier status on PGS$_{GW}$ whose direction aligned with the CNV's effect on adjusted phenotypes. The left panel shows correlation with the genome-wide PGS$_{GW}$ (red) and various PGS$_{trans}$ (blue shades), while the right panel shows correlations with various PGS$_{cis}$ (blue shades). Transparency reflects the significance level, with "ns" indicating $p > 0.05$. The small grey square indicates whether the correlation between the CNV carrier status and PGS$_{trans\text{-}250kb}$ (left panel) or PGS$_{cis\text{-}250kb}$ (right panel) are directionally concordant with the CNV's effect on the phenotype. We use a different x-axis scale for the 1p36.11 CNV region for better visualization (orange area).

significant CNV-PGS$_{cis}$ correlation. Interestingly, 12p12.2-12p12.1 CNVs seem to be tagged by bilirubin-associated SNVs located 50-100 kb from the CNV region, as indicated by the lack of a significant CNV-PGS$_{cis\text{-}50kb}$ correlation and a drop in PGS$_{trans}$ correlation excluding variants within the 50-100 kb range. The 12p12.2-12p12.1 *cis* region contains other genes from the *SLCO* family with overlapping functions in bilirubin transport [66], so that variation in that locus represents a strong determinant of bilirubin levels. Yet, PGS$_{trans\text{-}chr12}$ remained correlated at nominal significance with the CNV carrier status, indicating that linkage disequilibrium cannot fully account for the CNV-PGS$_{GW}$ correlation. The two other pairs with a nominally significant CNV-PGS$_{cis}$ correlation involve 16p11.2 deletions with systolic and diastolic blood pressure, for which we observe local tagging within the $\pm$ 250 kb region. Interestingly, for these pairs, the direction of correlation between the CNV and PGS$_{cis}$ opposes the direction of correlation with PGS$_{trans}$, which itself aligns with the deletion's impact on blood pressure. Hence, local *cis*-tagging masks part of the CNV-PGS$_{GW}$ correlation, indicating that the latter is not driven by *cis*-tagging.

We conclude that among these ten CNV-trait pairs, one is fully accounted for by *cis*-tagging, three show mixed evidence of *cis*-tagging in combination with another mechanism driving CNV-$PGS_{GW}$ correlation, and six show no evidence for *cis*-tagging. We further note that 20 out of 29 CNV-trait pairs with a nominally significant CNV-$PGS_{cis}$ but no significant CNV-$PGS_{GW}$ correlation have opposing CNV-$PGS_{cis}$ and CNV-$PGS_{trans}$ correlations (Supplementary Fig. 7). This creates a scenario wherein local and genome-wide background oppose each other, making it less likely for the CNV-$PGS_{GW}$ correlation to reach significance.

## Assortative mating exacerbates the phenotypic impact of rare CNVs

Next, we explored assortative mating as an alternative mechanism to explain the remainder of the positive CNV-$PGS_{GW}$ correlation. Under assortative mating, phenotypic similarity caused by one parent's high $PGS_{GW}$ and the other parent's risk-increasing CNV results in co-inheritance of these genetic risk factors in the offspring (Fig. 1c). First, we examined correlations between adjusted phenotypes among the 3,396 UKBB couples present in our test set. A total of 33 (77%) same-trait comparisons showed significant positive correlations, with vitamin C ($r$ = 0.48, p = $1.1 \times 10^{-165}$), BMI ($r$ = 0.26, p = $7.8 \times 10^{-53}$), body fat mass ($r$ = 0.25, p = $2.4 \times 10^{-49}$), weight ($r$ = 0.23, p = $9.2 \times 10^{-43}$) and height ($r$ = 0.21, p = $8.0 \times 10^{-36}$) showing the strongest correlation (Supplementary Fig. 8; Supplementary Table 9). Comparing these estimates to those obtained from 51,664 couples present in the full UKBB data set yielded highly consistent estimates ($r$ = 0.98, p = $6.6 \times 10^{-32}$; Supplementary Fig. 9a; Supplementary Table 10). Similarly, using a stricter couples' definition that identified 1,213 couples within our test set yielded highly correlated estimates ($r$ = 0.97, p = $8.5 \times 10^{-27}$; see Methods; Supplementary Fig. 9b; Supplementary Table 10). This confirms widespread phenotypic assortment across diverse traits in our sample, which could channel correlation between CNVs and $PGS_{GW}$.

Next, we leveraged within-couple phenotypic correlation to estimate the fraction of the total CNV-$PGS_{trans\text{-}250kb}$ correlation that can be explained through a path mediated by assortative mating. The latter is defined as the product of the CNV-phenotype correlation ($\alpha$), the within-couple phenotypic correlation ($\beta$), and the $PGS_{trans\text{-}250kb}$-phenotype correlation ($\gamma$) (Fig. 1c). We decided to use $PGS_{trans\text{-}250kb}$, which mimics $PGS_{GW}$ while ensuring that our estimation of assortative mating is not conflated by linkage disequilibrium. Indirect CNV-PGS correlations estimated via the assortative mating path significantly

correlated with total CNV-PGS$_{trans\text{-}250kb}$ correlation ($r$ = 0.45, p = 2.0 $\times$ 10$^{-7}$; Fig. 5a; Supplementary Table 11), and 50,000 permutations of $\alpha$ (p $<$ 2 $\times$ 10$^{-5}$), $\beta$ (p = 4.4 $\times$ 10$^{-4}$), or $\gamma$ (p $<$ 2 $\times$ 10$^{-5}$), showed it was highly unlikely to obtain a total-indirect CNV-PGS$_{trans\text{-}250kb}$ correlation larger than 0.45 (Fig. 5b). As expected, the total CNV-PGS$_{trans\text{-}250kb}$ correlation negatively correlated with the indirect path expected under selective participation bias, discarding this competing scenario as the driving force behind our results.

To ensure the robustness of our findings, we performed a series of sensitivity analyses. Repeating the analysis excluding couples used to calculate $\beta$ to estimate $\alpha$ and $\gamma$ produced similar results ($r$ = 0.42; p = 2.2 $\times$ 10$^{-6}$; Supplementary Fig. 10a; Supplementary Table 12), as did adjusting phenotypes for potential interactions between sex and other covariates ($r$ = 0.46, p = 1.3 $\times$ 10$^{-7}$; Supplementary Fig. 10b; Supplementary Table 13) and for coordinates of residence ($r$ = 0.45, p = 2.1 $\times$ 10$^{-7}$; Supplementary Fig. 10c; Supplementary Table 14). To further ensure that these results are not driven by overestimation of $\alpha$ due to Winner's curse, we repeated the analysis based on 98 CNV-trait pairs that survive a stricter multiple-testing correction threshold for the CNV-trait association strength (see Methods). Despite the lower number of tested pairs (-18%) reducing our power to detect a signal through the assortative mating path, the correlation remained unchanged ($r$ = 0.46, p = 2.0 $\times$ 10$^{-6}$; Supplementary Fig. 10d; Supplementary Table 15). From a theoretical perspective, we show that analytically derived boundaries for $r(CNV, PGS_{GW})$ are wide and thus not expected to correlate with the product of $\alpha \times \beta \times \gamma$, indicating that our results are not due to mere mathematical constraint of correlated variables (Supplementary Note 1). We also show – analytically and through simulations – that under assortment, the correlation between CNV carrier status and PGS$_{GW}$ is approximated by $r(CNV, PGS_{GW}) = \alpha \times \beta \times \gamma$ (Supplementary Notes 2 and 3). Together, these analyses indicate that our finding is unlikely to be driven by artifacts and suggest that a substantial fraction of CNV-PGS$_{GW}$ correlation might be explained by assortative mating.

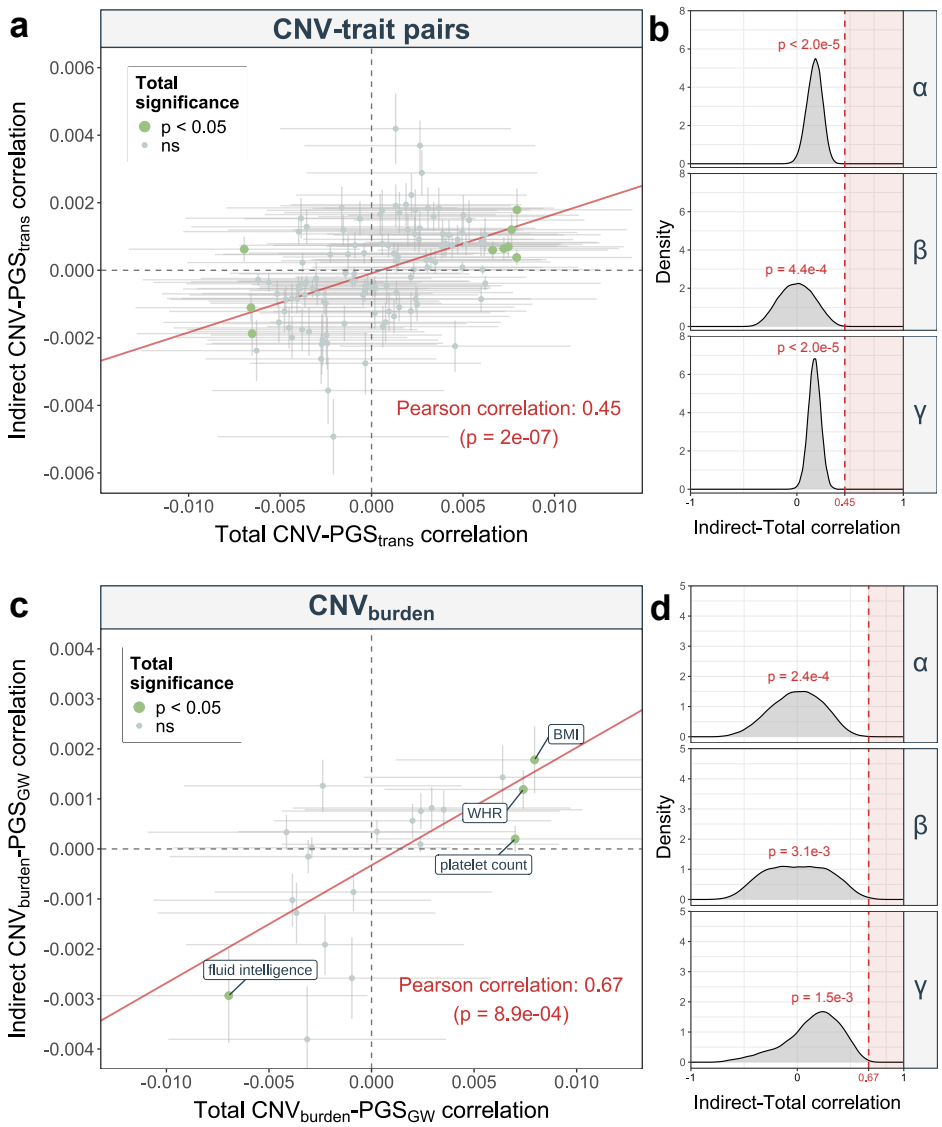


**Figure 5. Assortative mating partially explains the CNV-PGS correlation.**
**a,** Correlation between total (x-axis) and indirect (y-axis; through assortative mating) CNV-PGS$_{trans\text{-}250kb}$ correlation for 119 CNV-trait pairs. Error bars represent 95% confidence intervals. Size and color reflect the significance level of the total CNV-PGS$_{trans\text{-}250kb}$ correlation, with "ns" indicating $p > 0.05$. The red line represents the fitted linear regression of the indirect CNV-PGS$_{trans\text{-}250kb}$ correlation against the total CNV-PGS$_{trans\text{-}250kb}$ correlation. **b,** Density plots displaying the distribution of the total versus indirect CNV-PGS$_{trans\text{-}250kb}$ correlations obtained from 50,000 permutations of $\alpha$ (top), $\beta$ (middle), and $\gamma$ (bottom) across the 119 CNV-trait pairs. The empirical correlation and associated p-value are denoted in red. Values more extreme than the observed correlation are shaded in red. **c,** Correlation between total (x-axis) and indirect (y-axis; through assortative mating) CNV burden-PGS$_{GW}$ correlation for 21 traits significantly ($p \leq 0.05/43 = 1.2 \times 10^{-3}$) associated with the CNV burden and **d,** associated permutation tests' results, following the same legend description as in (**a**) and (**b**), respectively, with nominal-significant effects labeled. BMI = body mass index; WHR = waist-to-hip ratio.

## Rare CNVs are often inherited

Assortative mating can only account for CNV-PGS$_{GW}$ correlation if at least a substantial proportion of CNVs are inherited. Given that many assessed CNVs are recurrent and are expected to have a high *de novo* rate, we sought confirmation of this prerequisite by estimating inheritance rates across the 27 studied CNV regions. Seven deletions and ten duplications met our criteria to assess inheritance rate

(see Methods). All 17 testable CNVs exhibited a non-zero inheritance rate, with a pooled inheritance rate of 23% [95%-CI: 21-25%] and 79% [95%-CI: 73-84%], for deletions and duplications, respectively (Supplementary Table 16). Despite the low number of relatives in UKBB hampering the precision of our estimates, these numbers indicate that the assessed CNVs can be inherited in the general population, supporting our proposed assortative mating model.

## Widespread positive correlation across mutational classes

Finally, we tested whether our proposed mechanism of genetic compounding through assortative mating could be generalized beyond focal CNV-trait pairs. First, we estimated the association of the CNV burden with $PGS_{GW}$ for 21 traits impacted by the CNV burden ($p \leq 0.05/43 = 1.2 \times 10^{-3}$; Supplementary Table 17). While no effect survived multiple testing correction ($p \leq 0.05/21 = 2.4 \times 10^{-3}$), we observed nominally significant effects of the CNV burden on the $PGS_{GW}$ for BMI (p = 0.021), waist-to-hip ratio (p = 0.032), platelet count (p = 0.043), and fluid intelligence (p = 0.044). In line with earlier findings, 18 out of 21 effects were directionally concordant with the effect of the CNV burden on the adjusted phenotypes – including the four nominally significant effects described above – suggesting that carriers of trait-associated CNVs are more prone to carry a phenotype-exacerbating $PGS_{GW}$. We tested whether this pattern could be accounted for by assortative mating by substituting the locus-specific CNV carrier status by the CNV burden in our assortative mating framework (Fig. 1c). We observed a significant correlation between the total CNV burden-$PGS_{GW}$ correlation and the indirect path mediated by assortative mating ($r$ = 0.67, $p = 8.9 \times 10^{-4}$) (Fig. 5c; Supplementary Table 18). Permutation of $\alpha$ ($p = 2.4 \times 10^{-4}$), $\beta$ ($p = 3.1 \times 10^{-3}$), and $\gamma$ ($p = 1.5 \times 10^{-3}$) supported the significance of our results (Fig. 5d).

Second, we examined if the association between PGS and CNVs extended to variants of other mutational classes. A total of 15 traits were significantly modulated by the autosomal predicted loss-of-function (pLoF) SNV burden ($p \leq 0.05/43 = 1.2 \times 10^{-3}$; Supplementary Table 17). For these, we estimated the pLoF burden's impact on $PGS_{GW}$ and identified a single nominally significant effect on $PGS_{GW\text{-}BMI}$ (p = 0.027) whose effect size was directionally concordant with the pLoF's burden effect on adjusted BMI (Supplementary Table 17). Finally, we observed a small but significant correlation between the pLoF and CNV burden ($r$ = 0.005; $p = 3.9 \times 10^{-3}$). We note that the rarity of CNVs in our dataset makes it unlikely to pick up any strong correlation. Indeed, simulating the CNV and pLoF burdens based

on a negative binomial and normal distributions, respectively (Supplementary Fig. 11a), we show that we have ~75% power to detect a correlation of 0.005 (Supplementary Fig. 11b), yet tend to underestimate the correlation's magnitude (Supplementary Fig. 11b-c). Although evidence remains limited and larger studies will be required to confirm these trends, current data align with accumulation of genetic variants with coordinated phenotypic effects across mutational classes as a widespread phenomenon within the general population.

# Discussion

In this study, we explored the interplay between rare deleterious CNVs and common genetic variants, demonstrating that CNV expressivity is influenced by the carrier's polygenic background through additive effects. We also observed a mild but widespread positive CNV-PGS correlation, most of which could not be accounted for by linkage disequilibrium or participation bias. We propose assortative mating as a mechanism that could explain this relationship, supported by a non-null inheritance rate for a substantial fraction of the studied CNVs. A similar trend of positive correlation was observed between CNV or pLoF burdens and PGS, suggesting a pervasive mechanism in which the impact of rare deleterious variants is compounded by common variants through assortative mating.

There has been a surge in interest towards better conceptual understanding of how common and rare variants jointly influence complex traits. Studies have demonstrated that polygenic background contributes to the variable expressivity of rare "monogenic" variants, allowing improved risk estimation [49, 67, 68]. Targeted studies focusing on specific phenotypes such as autism [27, 46, 53], schizophrenia [27, 47, 51, 57], intellectual disability and developmental delay [48, 55, 56, 69, 70], or obesity [71, 72] showed that overall, rare and common variants act additively. Our results reinforce this body of evidence by extending these findings to a more diverse set of quantitative traits, supporting that $PGS_{GW}$ can stratify CNV carriers to identify those who are at increased risk of developing CNV-associated comorbidities. Indeed, CNV carriers with a $PGS_{GW}$ counteracting the effect of the CNV exhibited phenotype values within the range of copy-neutral individuals, while carriers with a phenotype-exacerbating $PGS_{GW}$ tended to exhibit phenotypes outside the normal range. This pertains to the relative contribution of CNVs versus PGS, the latter being reported to have a broader impact on socio-economic outcomes than pathogenic CNVs [73]. We argue that while PGSs might be more relevant at the population scale, the presence of

318 a CNV has important implications at the individual level and future studies should explore how to best
319 integrate CNV and PGS to optimize relevance in terms of personalized medicine.

320 Focusing on more complex CNV-PGS relations, we tested for non-additive effects on two different
321 scales and identified two robust, scale-dependent interactions wherein CNV and polygenic background
322 exacerbate each other's phenotypic impact. The population distribution of GGT levels – and to a lesser
323 extent of grip strength – is skewed and non-normal, possibly explaining why the interactions become ap-
324 parent only on the raw phenotypic scale. To date, only weak CNV-PGS interactions have been reported
325 [27, 44, 45, 49, 54, 67, 68, 71, 72], so that these effects represent among the first robust evidence of
326 epistasis between rare CNVs and polygenic background. Yet, further research will be required to dis-
327 entangle scale from interaction effects [74], and assess whether these findings are of clinical relevance,
328 i.e., whether phenotypic differences among individuals that are outside the normal phenotypic range
329 have clinical relevance. It is also worth noting that we might miss other true interactions due to lack of
330 power or because the interaction only becomes visible with phenotypic transformations that were not
331 tested. Indeed, theoretical work predicts that disease risk variants might exhibit stronger phenotypic
332 consequences in high-risk polygenic background due to the non-linear relationship between disease
333 liability and disease probability [75]. This emphasizes the importance of the choice of the link function
334 for binary traits and the choice of the scale/transformation for continuous outcomes, as well as the key
335 role of ascertainment to accurately characterize rare variant pathogenicity. We conclude that while we
336 could demonstrate interactions for specific CNV-trait pairs, the majority of assessed pairs do not show
337 evidence of deviation from additivity at current sample sizes.

338 While the effect of $PGS_{GW}$ on the phenotype appeared mainly independent of CNV status, the CNV
339 status correlated positively with an individual's $PGS_{GW}$, paralleling a recent report showing that UKBB
340 carriers of CNVs causing neurodevelopmental conditions have lower $PGS_{GW}$ for educational attainment
341 [55]. We found this pattern to be widespread across diverse phenotypes and, after excluding that it
342 was fully driven by linkage disequilibrium, identified assortative mating as a likely pervasive contribu-
343 tor, although we note that it only partially explains the observed CNV-$PGS_{GW}$ correlation. While the
344 imperfect agreement between the CNV-$PGS_{GW}$ correlation and our predictor of it is – at least in part
345 – due to regression dilution bias (as our predictor is a product of three noisy estimates), other factors

346 that we did not explore might also play a role. For instance, while we aimed to account for population
347 stratification by restricting our analyses to a single genetic ancestry and controlling for 40 SNV-derived
348 principal components as well as coordinates of residence, we cannot exclude that part of the positive
349 CNV-$PGS_{GW}$ is driven by residual population stratification. Still, CNV-$PGS_{GW}$ correlation induced by
350 population stratification is not expected to be proportional to assortative mating; while it might account
351 for part of the CNV-$PGS_{GW}$ correlation that is *not* explained by assortative mating, it is unlikely to drive
352 our observations.

353 Positive phenotypic assortment is well-established for neuropsychiatric phenotypes, anthropomet-
354 ric features, and lifestyle factors [39, 76–79], with recent biobank studies reporting partner phenotypic
355 correlation among less conspicuous traits, e.g., blood measurements [76]. Studies of the downstream
356 consequences of assortative mating on genomic architecture have emphasized how ignoring its impact
357 can bias estimation of key genetic metrics [80–89], yet, the mechanisms through which it acts and the
358 full scale of its implications remain underappreciated. Given assortative mating on a given trait, if a
359 rare CNV contributes to that trait and is inherited – which we demonstrate is a plausible assumption
360 – we expect compounding genetic factors in subsequent generations until the number of accumulated
361 variants starts affecting fitness. This induces a positive correlation between CNV (or other rare variants)
362 and $PGS_{GW}$, analogous to what was observed for PGSs between different chromosomes [81]. Previous
363 work has shown that over generations, assortative mating correlates with co-occurrence of psychiatric
364 conditions between parents and offspring [79] and leads to genetic anticipation, wherein disease liability
365 increases over generations among families carrying rare pLoF variants with variable expressivity [39].
366 Another study proposed parental assortment to explain the correlation between PGS and rare variant
367 burden associated with cognitive and educational outcomes both between partners and within individu-
368 als [90]. Our data support this mechanism and suggest that it operates across a broader spectrum of
369 variants and phenotypes.

370 Our assortative mating framework is targeted toward detecting direct assortment, which induces a
371 correlation in genetic variants directly associated with the trait on which assortment takes place. This
372 scenario is conceivable for traits such as height or BMI, but seems improbable for traits such as bilirubin
373 levels. Yet, these phenotypes can act as biomarkers, correlating to visible traits on which assortment

might occur. Such complex mechanisms are evidenced by both direct same-trait and indirect cross-trait assortment having been demonstrated in the UKBB [76]. Acknowledging that mate selection is likely to be based on a large combination of traits, a partner-choice index has recently been proposed [91]. An interesting related consideration is that most CNVs assessed in this study are pleiotropic. As such, polygenic liability for various traits might determine the presence and/or severity of a given phenotype, explaining phenotypic heterogeneity across carriers of the same CNV. One key question, however, is whether PGSs for different CNV-associated traits are independent. Persistent cross-trait assortative mating was shown to induce spurious genetic correlations [84], highlighting the importance of factoring in both pleiotropy and polygenicity [92] – an extension that goes beyond the scope of this study.

Another consideration is sample ascertainment, as cohorts in which healthy participants are over- (e.g., volunteer samples [58, 93, 94]) or under- (e.g., case-control studies [95, 96] or electronic health record-linked biobanks with higher disease burden [97, 98]) represented may exhibit a negative CNV-PGS correlation (Fig. 1a), as documented in previous clinical studies [44, 45, 47, 50]. In UKBB, we observed a weak, yet consistent, positive CNV-PGS correlation. While healthy volunteer bias in UKBB could have attenuated this association, it did not outweigh the positive correlation induced by assortative mating. We conclude that while selective participation in volunteer biobanks may alter the magnitude of observed effects and should therefore be acknowledged, correct inference with respect to directionality of effects remains largely possible [58, 99]. Case-control studies, in contrast, are subject to extreme participant selection, possibly canceling out any positive CNV-PGS correlation induced by assortative mating. This suggests that genetic findings obtained from population biobanks are likely to be more robust against selection bias than those obtained from clinical studies, highlighting the usefulness of the former when interrogating the role of rare and common variants in disease susceptibility.

Finally, our study should be interpreted in light of some limitations. The examined CNVs were detected from genotyping microarrays and span large genomic regions, with the corollary that they have strong phenotypic consequences and are rarer. While these characteristics were purposefully selected for in our study, low CNV frequency limits the power of our analyses and influences its outcomes; we expect that investigating CNVs with higher frequencies and smaller effect sizes would identify more cases in which *cis*-tagging represents a dominant mechanism. Indeed, SNV associations have been found to

402 tag stronger associations with smaller structural variants [22, 100–102]. Accordingly, the only common
403 CNV in our analyses – the *RHD* deletion – was subject to *cis*-tagging. Recent efforts in producing large
404 genome sequencing datasets coupled with the development of methods allowing robust detection of
405 structural variants across the size and allele frequency spectrum will allow exploration of the interplay
406 between different mutation classes at a scale [100, 103]. Furthermore, because PGS portability de-
407 creases with genetic distance [104], we restricted our study to a single genetic ancestry group with a
408 sufficiently large sample size to explore the interplay between PGS and rare CNVs. This stresses the im-
409 portance of continued efforts to develop large and diverse biobanks with rich phenotypic data to enable
410 replication studies and extend the proposed framework to a broader set of phenotypes, including binary
411 disease outcomes. Larger sample sizes and further methodological development will also enable better
412 estimation of CNV inheritance rates, allowing incorporation of the latter in our assortative mating frame-
413 work. Improved collection and integration of genetic and environmental factors will likely reveal a more
414 complex and nuanced picture of the mechanisms shaping variable expressivity. Importantly, because
415 participation bias and assortative mating have opposing effects on the CNV-PGS correlation, we likely
416 underestimate the role of assortative mating. This is exacerbated by our observation that local linkage
417 disequilibrium can oppose the global CNV-$PGS_{GW}$ correlation. Finally, our assortative mating model ig-
418 nores potentially pleiotropic CNV and PGS effects acting on (further) off-target traits under assortment,
419 an aspect that could be further explored in future studies.

420 In conclusion, our study supports the notion that rare deleterious CNVs and common polygenic back-
421 grounds additively shape diverse phenotypes and identifies assortative mating as a candidate mecha-
422 nism promoting the co-occurrence of risk-associated common and rare genetic contributions.

# Materials and methods

## Study material

### *Software*

Analyses were run on the UK Biobank Research Analysis Platform using Python v3.11.5 and R v4.4.0.

### *Cohort description & sample selection*

This study is based on volunteer participants from UKBB, a population cohort encompassing ~500,000 individuals (54% females) from the UK general population aged 40-69 years at recruitment (2006-2010). Genetic data from genotype microarray (UKBB field identifier #22418) and exome sequencing (#23149) are accessible for all participants, along with extensive phenotypic assessment [105, 106]. Presented analyses were performed on various subsets of "white British" individuals (N = 409,703), defined as `in_white_british_ancestry_subset==1` in `ukb_sqc_v2.txt`. UKBB has approval from the NorthWest Multi-centre Research Ethics Committee as a Research Tissue Bank. All participants signed a broad informed consent form. Data were accessed under application 16389.

### *CNV association data*

Focal CNV-traits pairs stem from previous work by our group [19]. Briefly, CNV genome-wide association studies (GWASs) were performed in the UKBB to explore the association between the copy-number of genome-wide CNV-proxy probes derived from microarray-based CNV calls and 57 quantitative traits according to three association models: a mirror model in which deletions and duplications have opposing phenotypic consequences, as well as deletion-only and duplication-only models assessing the isolated impact of the deletion or duplication, respectively. We identified 131 genome-wide significant signals characterized by a lead CNV-proxy probe [19]. These were filtered for autosomal CNV regions and phenotypes measured in both males and females, resulting in 119 associations involving 43 phenotypes and 27 CNV regions.

### *Phenotype transformation*

Raw phenotypic measurements for the 43 analyzed quantitative traits (see CNV association data) were extracted and averaged over available instances. Grip strength was defined as the mean of right and

left hand grip strength (#47 and #46). WHR was defined as the ratio between waist and hip circumference (#48 and #49); WHR adjusted for body mass index (WHRadjBMI), by regressing WHR on BMI (#21001). To develop PGS scoring files, phenotypes were rank-transformed (`rntransform` from the GenABEL v1.8.0 R package) within the training sample set (see Polygenic scores calculation). For association studies with CNVs, measurements were inverse normal transformed across the test set (see Polygenic scores calculation) before correcting traits for sex, age (#21003), age$^2$, genotyping batch, and principal components (PCs) 1-40 [19]. The latter are referred to as adjusted phenotypes throughout the manuscript.

### *CNV carrier identification & CNV burden calculation*

We called genotype microarray-based CNVs (GRCh37/hg19) using standard PennCNV v1.0.5 settings [107] for all UKBB participants [19]. Briefly, each CNV call was attributed a quality score ranging between -1 (likely deletion) and 1 (likely duplication) [108]. Only high-confidence CNVs with absolute value of quality score $> 0.5$ were retained. CNV carrier status was established for each UKBB participant across the 119 assessed CNV-trait pairs (see CNV association data): individuals were considered as deletion or duplication carriers when harboring a high-quality deletion (copy-number = 1) or duplication (copy-number = 3), respectively, overlapping the lead CNV-proxy probe for that signal. Copy-neutral individuals were defined as having no evidence of a CNV (even with low quality score) overlapping the lead CNV-proxy probe. For association studies, CNV carrier status was encoded according to the model yielding the most significant CNV-trait association [19], using the following deletion/copy-neutral/duplication encoding for the mirror (-1/0/1), deletion-only (1/0/NA), and duplication-only (NA/0/1) models. Individual-level CNV, deletion, and duplication burdens were calculated as previously described [19], reflecting the number of hg19 RefSeq genes disrupted by high-confidence autosomal CNVs, deletions, and duplications, respectively ($\geq$ 1 bp overlap with exons, splice sites, non-coding RNAs, or UTRs).

### *Polygenic scores calculation*

Due to the lack of standardized scoring files for all analyzed phenotypes in other cohorts, we divided the “white British” UKBB samples into a training set (n = 264,372; 54% females), used to develop PGS scoring files, and a CNV-carrier-enriched test set (n = 96,716; 54% females) (Supplementary Fig.1).

Individuals carrying a high-confidence CNV (absolute value of quality score: $> 0.5$) overlapping the lead probe of one of the 119 analyzed CNV-trait pairs or an overall high CNV burden ($\geq 10$ genes) (see CNV carrier identification & CNV burden calculation) were pre-selected for the test set and pruned to 38,236 unrelated individuals (kinship $\leq 0.0884$) by iteratively removing the sample with the highest number of relatives. These individuals and all their relatives ($\leq 3^{rd}$ degree, regardless of CNV carrier status) were excluded from the 409,703 "white British" samples, along with individuals carrying a lower-confidence CNV call (absolute value of quality score: 0.2–0.5). Remaining individuals were randomly split into training (80%) and control (20%) sets with matching age and sex distributions. Individuals in the control set were pruned to unrelated individuals and merged with pre-selected CNV carriers to form the complete test set. Non-carriers in the test set related to samples in the training set were excluded from the test set. Finally, individuals whose reported sex did not match their genetically inferred sex (`Submitted.Gender` and `Inferred.Gender` in `ukb_sqc_v2.txt`) were excluded from both sets.

GWASs were conducted in the training set for the 43 rank-transformed phenotypes (see Phenotype selection). Briefly, PLINK v2.00a4.3 [109] was used to filter genotype microarray data, excluding variants with minor allele frequency (MAF) $< 1\%$ (`--maf 0.01`), missingness rate $> 10\%$ (`--geno 0.1`), or Hardy-Weinberg $p < 1 \times 10^{-15}$ (`--hwe 1e-15`). Samples with $> 10\%$ missing genotype calls (`--mind 0.1`) were excluded. For imputed variants, those with MAF $< 0.5\%$ (`--maf 0.005`) or low imputation quality (`--mach-r2-filter 0.3 2`) were removed. GWASs were carried out with Regenie v3.2.9 [110], including age, sex, and PC1-20 as covariates. Only quality-controlled genotyped variants were used for model fitting (i.e., Regenie step 1). We then tested both genotyped and imputed variant dosages for association, using default parameters except for minimum minor allele count (`--minMAC 1`) (step 2). PGS weights were derived from the resulting GWAS summary statistics, filtered for common (MAF $> 1\%$) high imputation quality (INFO $> 0.8$) variants present in the HapMap3+ extended set. LDpred2 (`snp_ldpred2_auto` implemented in the bigsnpr v1.12.18 R package) was run using the pre-computed linkage disequilibrium matrix based on UKBB participants of European ancestry, a heritability estimate from LD score regression (using `snp_ldsc`), and default parameters, except for `num_iter=500`, `shrink_corr=0.95`, `allow_jump_sign=FALSE`, `use_MLE=TRUE`, and `vec_p_init=seq_log(1e-4,0.2,length.out=10)` [111–113]. $PGS_{GW}$, defined as the sum of effect

allele dosages (from imputed genotypes) weighted by their corresponding effect sizes (derived from LDpred2), were computed for the test set using `pgsc_calc` v1.3.2 (output: `SUM`) [114].

To assess $PGS_{GW}$ performance, we fitted two linear regressions that explain the inverse-normal transformed phenotypic values as a function of covariates-only (i.e., sex, age, age$^2$, genotyping batch, and 40 PCs) or covariates in combination with $PGS_{GW}$. Confidence intervals (CI; 95%) were computed for the model's $R^2$ using the `ci.R2()` function from the MBESS v4.9.3 R package [115]. The difference in $R^2$ between the models ($\Delta R^2$) – which represents the contribution of $PGS_{GW}$ to phenotypic variability – was computed by subtracting the $R^2$ of the covariate-only model from the $R^2$ of the model including $PGS_{GW}$. The 95%-CI for $\Delta R^2$ was computed with `ci.R2()`, using a single predictor (`K = 1`).

### *Rare pLoF burden calculation*

A subset of 319,161 unrelated ($>$ 3$^{rd}$ degree) "white British" UKBB participants [25] with available exome sequencing data were used to compute the pLoF burden. pLoF variants were annotated using LOFTEE [116] and SnpEff (retrieved from `ukb23158_500k_OQFE.annotations.txt.gz`) [117]. Only those classified as high-confidence pLoF by LOFTEE and either annotated as pLoF or absent in SnpEff were retained. These were filtered for MAF $\leq$ 1% among non-related "white British" individuals. Rare pLoF burden was defined as the number of autosomal pLoF variants in an individual's genome (`plink2 --score --no-mean-imputation`).

## CNV-PGS modeling

### *Significance threshold*

Analyses modeling the interplay between CNV and PGS are repeated for the 119 selected CNV-trait pairs (see CNV association data). Significance was determined using Bonferroni correction ($p \leq 0.05/119 = 4.2 \times 10^{-4}$), unless specified otherwise. We report nominally significant signals ($p \leq 0.05$).

### *Additive contribution of $PGS_{GW}$ to CNV effect*

We modeled the joint contribution of CNV and $PGS_{GW}$ to phenotypic variability by fitting a linear regression model explaining adjusted phenotypes as a function of CNV carrier status (encoded according to the most significant CNV-GWAS model [19]) and $PGS_{GW}$. To specifically assess the contribution of $PGS_{GW}$ among CNV carriers, a linear regression modeling raw phenotypes as a function of $PGS_{GW}$, age,

age$^2$, sex, and PC1-10 was fitted among CNV, deletion, or duplication carriers, depending on whether the CNV-trait association had previously been found to be most significant under the mirror, deletion-only, or duplication-only model, respectively [19]. The standard error of the $PGS_{GW}$ effect was calculated as the standard deviation of 5,000 bootstrap estimates of the coefficient and was plugged into the Wald test to obtain two-sided p-values. Comparison of effects across phenotypes was done based on standardized effect sizes along with their 95%-CIs. These were obtained by dividing the effect size and the CI endpoints by the standard deviation of the corresponding phenotype in the test set. We used logistic regression to jointly estimate the impact of the number of CNV carriers and the fraction of phenotypic variance explained by $PGS_{GW}$ on the likelihood of a CNV-trait pair to be found Bonferroni significant in the above-described modeling strategy. For visualization, the distribution of the raw phenotypic values is shown for each $PGS_{GW}$ quintile among CNV carriers and copy-neutral individuals. Differences in prevalence of abnormally elevated GGT values ($>$ 50 U/L) among carriers of a 22q11.23 duplication versus copy-neutral individuals was evaluated by two-sided Fisher's exact test.

### *CNV-$PGS_{GW}$ interaction*

For each CNV-trait pair, a linear regression model was fitted that explains the adjusted phenotype as a function of CNV carrier status (encoded according to the most significant CNV-GWAS model [19]), $PGS_{GW}$, and an interaction term between the two. Similarly, we tested for an interaction between the deletion or duplication burden and $PGS_{GW}$, defining significance based on Bonferroni correction ($p \leq 0.05/(43 \times 2) = 5.8 \times 10^{-4}$). As interactions can be scale-dependent, we recomputed linear regressions explaining raw phenotypic values as a function of CNV carrier status, $PGS_{GW}$, an interaction term between the two, and covariates (age, age$^2$, sex, genotype batch, and PCs 1-40). For significant interactions, the regression model was recomputed using a $PGS_{trans}$ that excludes all SNVs within $\pm$ 250 kb of the focal CNV region to ensure that the interaction was not driven by errors in genotyping caused by the presence of the overlapping CNV. Analogously, interactions between the deletion or duplication burden and $PGS_{GW}$ on the raw phenotypic scale were assessed.

To assess our power to detect significant interactions at current sample size ($N$ = 96,716; i.e., equivalent to our test set), we conducted simulations across a grid of CNV frequencies, $q$, and interaction effects sizes, $\beta$, that overlap empirically observed data, i.e., $q$ and $\beta$ varied between $1 \times 10^{-5}$ and $1 \times 10^{-2}$ and

0.001 and 1, respectively. For each combination of $q$ and $\beta$, 1,000 simulations were performed. Specifically, CNV and $PGS_{GW}$ genotypes were simulated for $N$ individuals, following $CNV \sim$ Bernoulli$(q)$ and $PGS_{GW} \sim \mathcal{N}(0,1)$. These were used to model phenotypic values $Y$, as $Y \sim \beta(CNV \times PGS_{GW}) + \epsilon$. We then fitted a linear regression $Y \sim CNV + PGS + (CNV \times PGS)$ and deemed the interaction significant at p $\leq$ 0.05/119. Power was assessed as the fraction of the simulations for which the interaction term reached significance.

### *$PGS_{GW}$ distribution among CNV carriers*

For each CNV-trait pair, a linear regression model was fitted explaining $PGS_{GW}$ as a function of CNV carrier status (encoded according to the most significant CNV-GWAS model [19]). Similarly, we fitted linear regressions explaining $PGS_{GW}$ as a function of the CNV burden for the 21 phenotypes that were significantly (p $\leq$ 0.05/43 = 1.2 $\times$ $10^{-3}$) impacted by the latter. Significant effects were defined based on a Bonferroni correction (p $\leq$ 0.05/21 = 2.4 $\times$ $10^{-3}$). Hypothesizing that selective participation bias would induce a negative (spurious) correlation between CNV carrier status and $PGS_{GW}$ (Fig. 1c), we assessed whether the fraction of CNV-trait pairs with a CNV effect on $PGS_{GW}$ whose direction *aligned* with the CNV effect on the adjusted phenotype deviates from the expectation, accounting for the fact that some phenotypes are correlated. To do so, we simulated PGSs by drawing $N$ samples from a multivariate normal distribution $PGS_{GW} \sim \mathcal{N}(\mu, \mathbf{\Sigma_{PGS}})$ using `rmvnorm` from the mvtnorm v1.3.3 R package [118], with sample size $N$, mean vector $\mu$, and covariance matrix $\mathbf{\Sigma_{PGS}}$ derived from real data. CNV status were sampled as $CNV \sim$ Bernoulli$(q)$ with probabilities $q$ corresponding to the empirically observed CNV frequencies. Phenotypes were generated under a linear model $Y \sim CNV + PGS + \epsilon$, with errors $\epsilon$ coming from a multivariate normal distribution $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_Y})$ with mean $\vec{0}$ and empirical phenotypic covariance $\mathbf{\Sigma_Y}$. Using the simulated data, we fitted the same two models as for the real-data analyses: i) phenotype as a function of CNV status, and ii) $PGS_{GW}$ as a function of CNV status. The proportion of the 119 CNV-trait pairs for which the CNV's effect shows directional agreement in the two models was computed and the procedure was repeated 1,000 times. A p-value was obtained as the proportion of trials with a ratio greater than the empirical one.

### *Cis-tagging of CNVs by PGS*

For each CNV-trait pair, using a similar approach as for $PGS_{GW}$ (see Polygenic scores calculation), a local PGS ($PGS_{cis}$) using only SNVs mapping to the CNV region $\pm$ 50 kb ($PGS_{cis\text{-}50kb}$), $\pm$ 100 kb ($PGS_{cis\text{-}100kb}$), $\pm$ 250 kb ($PGS_{cis\text{-}250kb}$), or the entire chromosome ($PGS_{cis\text{-}chr}$) were calculated. Similarly, a $PGS_{trans}$ was computed excluding all SNVs that map to the CNV region $\pm$ 50 kb ($PGS_{trans\text{-}50kb}$), $\pm$ 100 kb ($PGS_{trans\text{-}100kb}$), $\pm$ 250 kb ($PGS_{trans\text{-}250kb}$), or the entire chromosome ($PGS_{trans\text{-}chr}$). We next computed the Pearson correlation between the CNV carrier status (encoded according to the most significant CNV-GWAS model [19]) and various $PGS_{cis}$ and $PGS_{trans}$. Standard errors for the Pearson correlations were computed through Fisher's z-transformation and correlations were deemed Bonferroni-significant at $p \leq 0.05/119 = 4.2 \times 10^{-4}$.

### *Conditioning of CNV-trait relations on $PGS_{GW}$*

For each CNV-trait pair, a linear regression model was fitted explaining the adjusted phenotype as a function of either CNV carrier status-only ($model_{M1}$) or CNV carrier status and $PGS_{GW}$ ($model_{M2}$). If conditioning the CNV-trait relation on $PGS_{GW}$ results in a significant change in the CNV-trait association, this would indicate that CNV and $PGS_{GW}$ are not independent, supporting tagging of the CNV by $PGS_{GW}$. Differences in effect size of the CNV carrier status were assessed using a t-statistic:

$$t = \frac{\beta_{M1} - \beta_{M2}}{\sqrt{SE_{M1}^2 + SE_{M2}^2 - 2 \times SE_{M1} \times SE_{M2} \times r}}\ , \tag{1}$$

where $\beta$ and $SE$ are the effect size and standard error of the effect of the CNV on the adjusted phenotype estimated through either $model_{M1}$ or $model_{M2}$. We adjust our test statistic for the fact that $\beta_{M1}$ and $\beta_{M2}$ are estimated from the same data, by accounting for $r$, the correlation between 5,000 bootstrap estimates of $\beta_{M1}$ and $\beta_{M2}$. Two-sided p-values were computed in R (`2*pnorm(-abs(t), mean = 0, sd = 1)`).

## Assortative mating

### *Phenotypic similarity within couples*

Couples were defined as "white British" individuals in households with two unrelated, opposite-sex individuals who identified as "husband, partner or wife" (#6141) [76]. Of the 51,664 identified couples,

3,396 have both partners in the test sample set. Focusing on the latter subset, same-trait and cross-trait phenotypic similarity among couples was assessed through Pearson correlation across 43 adjusted phenotypic measurements (see Phenotype selection), with significance defined through Bonferroni correction ($p \leq 0.05/43^2 = 2.7 \times 10^{-5}$). To ensure that downsampling did not impact the accuracy of our estimates, we compared same-trait phenotypic correlations computed from the test set to those obtained from the 51,664 couples. Furthermore, we compared estimates of within-couple phenotypic correlation to those obtained using a stricter definition of couples [81, 119], requiring individuals to have matching home location East and North coordinates (#22702 and #22704), Townsend deprivation index (#22189), number of smokers in household (#1259), number in household (#709), length of time at the current address (#699), average total household income before tax (#738), and living with husband or wife or partner (#6141). Couples were further required to be of opposite sex, of "white British" ancestry, unrelated, and have partners appear in a single couple, resulting in 22,068 couples, among which 1,213 had both partners in our test set.

### *Assortative mating modeling*

We explored assortative mating as a potential mechanism to explain positive CNV-PGS correlation by comparing the total CNV-$PGS_{trans\text{-}250kb}$ correlation to an indirect path mediated by assortment, defined as:

$$r(CNV, PGS_{trans-250kb})_{indirect} = \alpha \times \beta \times \gamma \,, \tag{2}$$

where $\alpha$ is the Pearson correlation between the adjusted phenotype and CNV carrier status (encoded by the most significant CNV-GWAS model [19]), $\beta$ is the within-couple Pearson correlation of adjusted phenotypic values, and $\gamma$ is the Pearson correlation between the adjusted phenotype and $PGS_{trans\text{-}250kb}$ (Fig. 1c). By using $PGS_{trans\text{-}250kb}$ instead of $PGS_{GW}$, we can ensure that our framework does not model local linkage disequilibrium. CIs for the total CNV-$PGS_{trans\text{-}250kb}$ correlation were computed through Fisher's z-transformation. SE of $\alpha$, $\beta$, and $\gamma$ were estimated using Bonett's approximation ($(1-r^2)/\sqrt{N-3}$, where $N$ corresponds to the sample size). For the indirect path, the SE was derived

from the variance of the product of independent variables as:

$$\mathrm{var}(X_1 \cdots X_n) = \prod_{i=1}^{n} \left(\mathrm{var}(X_i) + (E[X_i])^2\right) - \prod_{i=1}^{n} \left(E[X_i]\right)^2 \ , \tag{3}$$

where 95%-CIs were then obtained from propagated errors.

### *Significance assessment*

We first compared the calculated total-indirect CNV-PGS$_{\textit{trans}\text{-250kb}}$ with two competing hypotheses: a null scenario in which the true correlation is zero (`cor.test()` in R), and a scenario in which our data is better explained by selective participation. To test the latter, we computed the correlation between the total CNV-PGS$_{\textit{trans}\text{-250kb}}$ correlation and $-\alpha \times \gamma$, which is proportional to the bias induced by selective participation. In a second time, significance of the indirect path through assortative mating was assessed by comparing the total-indirect Pearson correlation to a set of three null distributions derived from 50,000 permutations, in which one of $\alpha$, $\beta$, or $\gamma$ was randomized. For $\alpha$ and $\gamma$, these were obtained by sampling without replacement, so that each phenotype was paired with a random CNV status or PGS$_{\textit{trans}\text{-250kb}}$, respectively, across the 119 CNV-trait pairs. For $\beta$, within-couple phenotypic correlation was recomputed after randomly re-pairing couples without replacement. The product of $\alpha \times \beta \times \gamma$ was recomputed, maintaining the empirical values for the two other coefficients. Finally, the correlation between the total and indirect CNV-PGS$_{\textit{trans}\text{-250kb}}$ correlations was computed and p-values were calculated as the proportion of permutations yielding as or more extreme correlations than the empirical one.

### *Sensitivity analyses for assortative mating modeling*

We conducted a series of follow-up analyses to ensure that our results were robust to analytical variations. First, we repeated the analysis estimating $\alpha$ and $\gamma$ excluding couples used to calculate $\beta$. Second, we repeated the analysis adjusting phenotypes for sex $\times$ covariates interactions. Third, to account for residual correlation between phenotypes and geographic location, we included participants' home location East and North coordinates at 1 km resolution (#22702 and #22704) as additional covariates when correcting phenotypic values and repeated the analysis. Fourth, Winner's curse could affect the CNV-trait pairs that were retained for this study, leading to an overestimation of the CNV-phenotype correlation, $\alpha$. To ensure this was not the case, we repeated the analysis focusing on 98 CNV-trait pairs

659 that in the original study would survive a stricter significance threshold of p $\leq$ 0.05/(11,804 $\times$ 3), which
660 accounts for the three tested association models [19].

661 From a theoretical perspective, to assess whether the observed CNV-PGS correlation could arise
662 solely from algebraic constraints, we first obtained bounds for $r(CNV, PGS_{GW})$ based on the positive
663 semi-definiteness of the full correlation matrix (comprising phenotype, CNV, and $PGS_{GW}$). These limits
664 define the range of allowed $r(CNV, PGS_{GW})$ values under the assumed causal structure (Supplemen-
665 tary Note 1). We then derived an analytical formula and conducted simulations showing that under
666 the assortative mating model assumptions, the CNV-$PGS_{GW}$ correlation corresponds to the product of
667 $\alpha \times \beta \times \gamma$ (Supplementary Notes 2 and 3).

668 Finally, we repeated the assortative mating framework (see Assortative mating modeling) at the
669 phenotypic level for 21 traits that were significantly (p $\leq$ 0.05/43 = 1.2 $\times$ $10^{-3}$) impacted by the CNV
670 burden (see CNV carrier identification & CNV burden calculation).

## 671 CNV inheritance rate estimation

### 672 *Sample selection*

673 We used KING [120] to identify related UKBB participants ($\leq$ 3rd degree; kinship $>$ 0.044). Among the
674 67,958 pairs of relatives (103,726 individuals) in which at least one individual carried a high-confidence
675 CNV, 12,831 pairs (21,216 individuals) harbored a CNV that overlaps ($\geq$ 1 bp) one of the 27 assessed
676 CNV regions (see CNV carrier identification & CNV burden calculation).

### 677 *Identity-by-descent estimation*

678 For each CNV carrier with a close relative, we estimated the probability that the region harboring the CNV
679 is identity-by-descent (IBD) with its relative, using THORIN [121, 122]. Namely, average IBD probability
680 $IBD_{avg}$ within a 1 Mb window around the CNV was computed by summing per-SNV IBD probability on
681 each haplotype. We classified the IBD status of the CNV carrier and its relative as either $IBD0$ (i.e., no
682 shared haplotypes; $IBD_{avg} < 0.25$), $IBD1$ (i.e., one shared haplotype; $0.75 < IBD_{avg} < 1.25$), $IBD2$
683 (i.e., two shared haplotypes; $IBD_{avg} > 1.75$). IBD status was set to unknown for all other probabilities
684 ($]0.25, 0.75[\cup]1.25, 1.75[$).

### *CNV sharing among relatives*

Accurate CNV breakpoint estimation is notoriously challenging from microarray data, so that even inherited CNVs could be detected with different breakpoints. Relatives were considered to carry the same CNV ($CNV2$) if they harbor a CNV of the same type showing over 50% reciprocal overlap. Otherwise, the pair is considered to have one CNV carrier ($CNV1$).

### *Inheritance rate estimation*

We combined information on IBD and CNV sharing between two relatives to classify our samples:

$$\begin{array}{cc} & \begin{array}{ccc} IBD0 & IBD1 & IBD2 \end{array} \\ \begin{array}{c} CNV2 \\ CNV1 \end{array} & \begin{pmatrix} A_{(unknown)} & B_{(inherited)} & C_{(inherited)} \\ D_{(unknown)} & E_{(unknown)} & F_{(denovo)} \end{pmatrix} \end{array}$$

where $IBD0$, $IBD1$, or $IBD2$ indicate whether none, one, or two haplotypes are shared, respectively; $CNV1$ or $CNV2$ indicate whether one or both individuals in a relative pair carry the same CNV, respectively. Specifically, $B$ and $C$ represent cases where the CNV is likely to be inherited, $F$ represents cases where the CNV is likely to be *de novo* (or a false positive), while $A$, $D$, and $E$, represent cases where we do not have sufficient data to infer the inheritance status. Based on this, we estimated the average CNV inheritance rate as:

$$R = \frac{2 \times B + C}{B + C + E + F} \ . \tag{4}$$

We estimated the standard error $SE(R)$ of the inheritance rate using parametric simulations from binomial distributions. For each locus $i$, we used the counts of correctly assigned $IBD1$ and $IBD2$ segments ($B$ and $C$, respectively) and the total number of individuals in $IBD1$ and $IBD2$ status ($n_1 = B + E$ and $n_2 = C + F$, respectively). We estimated inheritance rates separately for each IBD status:

$$R_1 = \frac{B}{n_1} \ ,$$

and

$$R_2 = \frac{C}{n_2} \ .$$

We then sampled 1,000 replicates of $B$ and $C$ from a binomial distribution:

$$B_{sim} \sim \text{Binomial}(1000, n_1, R_1) \ ,$$

and

$$C_{sim} \sim \text{Binomial}(1000, n_2, R_2) \ .$$

We computed the simulated inheritance rate as:

$$R_{sim} = \begin{cases} \frac{2 \times B_{sim} + C_{sim}}{n_1 + n_2}, & n_1 > 0 \text{ and } n_2 > 0. \\ \frac{2 \times B_{sim}}{n_1}, & n_1 > 0 \text{ and } n_2 = 0. \\ \frac{C_{sim}}{n_2}, & n_1 = 0 \text{ and } n_2 > 0. \end{cases} \tag{5}$$

and estimated the standard error of the inheritance rate $SE(R)$ as the standard deviation of $R_{sim}$. The pooled inheritance rate estimate and its standard error for duplications and deletions was calculated using inverse-variance weighted meta-analysis for testable CNVs (i.e., CNVs with $\geq$ 20 carriers with a $\leq$ 3rd degree relative who shares $IBD1$ or $IBD2$) with a non-null inheritance rate.

## pLoF burden analyses

Linear regression models explaining adjusted phenotypes as a function of pLoF burden identified 15 traits subject to modulation by the latter among unrelated "white British" samples with available exome sequencing data ($p \leq 0.05/43 = 1.2 \times 10^{-3}$). For these traits, linear regression models explaining $PGS_{GW}$ as a function of pLoF burden were fitted, with significance defined based on Bonferroni correction ($p \leq 0.05/15 = 3.3 \times 10^{-3}$). Pearson correlation between the CNV and the pLoF burdens was assessed. To assess our power to detect a correlation between the CNV and pLoF burdens, we performed 1,000 simulation trials across a grid of correlations, $r$, centered on the empirically observed correlation ($r$ = 0.005) and ranging from 0.0005 to 0.01 with a step size of 0.0005. In each trial, two samples of 319,161 individuals (i.e., the size of the unrelated "white British" UKBB participants with available exome sequencing data) were drawn from a bivariate normal distribution with means $(0, 0)$, unit variance and a Pearson correlation $r$. After selecting the most appropriate distribution to represent each burden type,

we rank-transformed the sample describing the CNV burden to follow a negative binomial distribution with mean $\mu_{CNV}$ = 1.51 and size = 0.18. The second sample, representing the pLoF burden, was rank-transformed to match a normal distribution with mean $\mu_{pLoF}$ = 13.71 and standard deviation $\sigma$ = 3.78. Parameters of these distributions were estimated by fitting negative binomial (CNV burden) or normal (pLoF burden) models to our data (filtered for "white British" UKBB participants) using `fitdistr()` from the MASS v7.3.65 R package [123]. The Pearson correlations between these two samples were computed, and the corresponding power was estimated using `wp.correlation()` from the WebPower v0.9.4 R package [124] at a significance level of $\alpha$ = 0.05. For each value of $r$, we averaged correlation and power across 1,000 trials, and reported the average power at the empirically observed correlation ($r$ = 0.005). We further explored how the discrepancy between true and measured correlation was affected by the CNV burden frequency by repeating the previous procedure with a true Pearson correlation fixed at $r = 1$ and varying CNV burden mean ($\mu_{CNV}$; ranging from 0.1 to 1 in 0.1 step increments and 1 to 10 in unit step increments).

# Data and code Availability

Data used in this study are publicly available, as described in the Methods. Code to reproduce analyses is available at `https://github.com/cauwerx/CNV_PGS_interaction`.

# Acknowledgments

We thank UKBB participants for sharing their data. Computations were carried out on the Urblauna High-Performance Computing server from the University of Lausanne and the UK Biobank Research Analysis Platform. This work was supported by funding from the Department of Computational Biology (ZK) and the Center for Integrative Genomics (AR) from the University of Lausanne, as well as grants from the Swiss National Science Foundation (P500-3_235131 to C.A.; 31003A_182632 and IZSTZ0_216615 to AR) and Horizon2020 Twinning projects (ePerMed 692145, AR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

# Author contributions

C.C., C.A., A.R., and Z.K. conceived and designed the study; C.C. and C.A. performed the main analyses; R.H. and T.C. derived the PGS weights; R.H. estimated the *de novo* rate of CNVs; T.S. provided guidance for assortative mating analyses; Z.K. supervised all statistical analyses; C.C. and C.A. drafted, and A.R. and Z.K. made critical revisions to the manuscript; All authors read, approved, and provided feedback on the final manuscript.

# Supplementary information

## Supplementary Information

Supplementary Figures 1-11.

Supplementary Notes 1-3.

Supplementary Tables 1-16.

# References


[1] Jonathan Sebat, B Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, et al. Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449, 2007.

[2] Christian R Marshall, Abdul Noor, John B Vincent, Anath C Lionel, Lars Feuk, Jennifer Skaug, Mary Shago, Rainald Moessner, Dalila Pinto, Yan Ren, et al. Structural variation of chromosomes in autism spectrum disorder. *The American Journal of Human Genetics*, 82(2):477–488, 2008.

[3] Heather C Mefford, Hiltrud Muhle, Philipp Ostertag, Sarah von Spiczak, Karen Buysse, Carl Baker, Andre Franke, Alain Malafosse, Pierre Genton, Pierre Thomas, et al. Genome-wide copy number variation in epilepsy: Novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS Genetics*, 6(5):e1000962, 2010.

[4] Sébastien Jacquemont, Alexandre Reymond, Flore Zufferey, Louise Harewood, Robin G Walters, Zoltán Kutalik, Danielle Martinet, Yiping Shen, Armand Valsesia, Noam D Beckmann, et al. Mirror extreme bmi phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature*, 478(7367):97–102, 2011.

[5] Bradley P Coe, Kali Witherspoon, Jill A Rosenfeld, Bregje WM Van Bon, Anneke T Vulto-van Silfhout, Paolo Bosco, Kathryn L Friend, Carl Baker, Serafino Buono, Lisenka ELM Vissers, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature genetics*, 46(10):1063–1071, 2014.

[6] Christian R Marshall, Daniel P Howrigan, Daniele Merico, Bhooma Thiruvahindrapuram, Wenting Wu, Douglas S Greer, Danny Antaki, Aniket Shetty, Peter A Holmans, Dalila Pinto, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*, 49(1):27–35, 2017.

[7] T Harel and JR Lupski. Genomic disorders 20 years on—mechanisms for clinical manifestations. *Clinical Genetics*, 93(3):439–449, 2018.

[8] Shuxiang Goh, Tracy Dudding-Byth, Mark Pinese, and Edwin P Kirk. Updated penetrance estimates for recurrent copy number variants–an improved definition and formula. *European Journal of Human Genetics*, pages 1–9, 2025.

[9] Hreinn Stefansson, Andreas Meyer-Lindenberg, Stacy Steinberg, Brynja Magnusdottir, Katrin Morgen, Sunna Arnarsdottir, Gyda Bjornsdottir, G Bragi Walters, Gudrun A Jonsdottir, Orla M Doyle, et al. Cnvs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, 505(7483):361–366, 2014.

[10] Katrin Männik, Reedik Mägi, Aurélien Macé, Ben Cole, Anna L Guyatt, Hashem A Shihab, Anne M Maillard, Helene Alavere, Anneli Kolk, Anu Reigo, et al. Copy number variations and cognitive phenotypes in unselected populations. *JAMA*, 313(20):2044–2054, 2015.

[11] Aurélien Macé, Marcus A Tuke, Patrick Deelen, Kati Kristiansson, Hannele Mattsson, Margit Nõukas, Yadav Sapkota, Ursula Schick, Eleonora Porcu, Sina Rüeger, et al. Cnv-association meta-analysis in 191,161 european adults reveals new loci associated with anthropometric traits. *Nature Communications*, 8(1):744, 2017.

[12] David Owen, Mathew Bracher-Smith, Kimberley M Kendall, Elliott Rees, Mark Einon, Valentina Escott-Price, Michael J Owen, Michael C O'Donovan, and George Kirov. Effects of pathogenic cnvs on physical traits in participants of the uk biobank. *BMC Genomics*, 19:1–9, 2018.

[13] Matthew Aguirre, Manuel A Rivas, and James Priest. Phenome-wide burden of copy-number variation in the uk biobank. *The American Journal of Human Genetics*, 105(2):373–383, 2019.

[14] Karen Crawford, Matthew Bracher-Smith, David Owen, Kimberley M Kendall, Elliott Rees, Antonio F Pardiñas, Mark Einon, Valentina Escott-Price, James TR Walters, Michael C O'Donovan, et al. Medical consequences of pathogenic cnvs in adults: Analysis of the uk biobank. *Journal of Medical Genetics*, 56(3):131–138, 2019.

[15] Kimberley Marie Kendall, Elliott Rees, Matthew Bracher-Smith, Sophie Legge, Lucy Riglin, Stanley Zammit, Michael Conlon O'Donovan, Michael John Owen, Ian Jones, George Kirov, et al. Association of rare copy number variants with risk of depression. *JAMA psychiatry*, 76 (8):818–825, 2019.

[16] Miguel Verbitsky, Rik Westland, Alejandra Perez, Krzysztof Kiryluk, Qingxue Liu, Priya Krithivasan, Adele Mitrotti, David A Fasel, Ekaterina Batourina, Matthew G Sampson, et al. The copy number variation landscape of congenital anomalies of the kidney and urinary tract. *Nature genetics*, 51(1):117–127, 2019.

[17] Christa Lese Martin, Karen E Wain, Matthew T Oetjens, Kasia Tolwinski, Emily Palen, Abby Hare-Harris, Lukas Habegger, Evan K Maxwell, Jeffrey G Reid, Lauren Kasparson Walsh, et al. Identification of neuropsychiatric copy number variants in a health care system population. *JAMA Psychiatry*, 77(12):1276–1285, 2020.

[18] Nasa Sinnott-Armstrong, Yosuke Tanigawa, David Amar, Nina Mars, Christian Benner, Matthew Aguirre, Guhan Ram Venkataraman, Michael Wainberg, Hanna M Ollila, Tuomo Kiiskinen, et al. Genetics of 35 blood and urine biomarkers in the uk biobank. *Nature Genetics*, 53(2): 185–194, 2021.

[19] Chiara Auwerx, Maarja Lepamets, Marie C Sadler, Marion Patxot, Miloš Stojanov, David Baud, Reedik Mägi, Tõnu Esko, Andres Metspalu, Lili Milani, et al. The individual and global impact of copy-number variants on complex human traits. *The American Journal of Human Genetics*, 109(4):647–668, 2022.

[20] Rebecca Birnbaum, Behrang Mahjani, Ruth JF Loos, and Andrew J Sharp. Clinical characterization of copy number variants associated with neurodevelopmental disorders in a large-scale multiancestry biobank. *JAMA Psychiatry*, 79(3):250–259, 2022.

[21] Ryan L Collins, Joseph T Glessner, Eleonora Porcu, Maarja Lepamets, Rhonda Brandon, Christopher Lauricella, Lide Han, Theodore Morley, Lisa-Marie Niestroj, Jacob Ulirsch, et al. A cross-disorder dosage sensitivity map of the human genome. *Cell*, 185(16):3041–3055, 2022.

[22] Margaux LA Hujoel, Maxwell A Sherman, Alison R Barton, Ronen E Mukamel, Vijay G Sankaran, Chikashi Terao, and Po-Ru Loh. Influences of rare copy-number variation on human complex traits. *Cell*, 185(22):4233–4248, 2022.

[23] Jakub Kopal, Kuldeep Kumar, Karin Saltoun, Claudia Modenato, Clara A Moreau, Sandra Martin-Brevet, Guillaume Huguet, Martineau Jean-Louis, Charles-Olivier Martin, Zohra Saci, et al. Rare cnvs and phenome-wide profiling highlight brain structural divergence and phenotypical convergence. *Nature Human Behaviour*, pages 1–17, 2023.

[24] Ludovica Montanucci, David Lewis-Smith, Ryan L Collins, Lisa-Marie Niestroj, Shridhar Parthasarathy, Julie Xian, Shiva Ganesan, Marie Macnee, Tobias Brünger, Rhys H Thomas, et al. Genome-wide identification and phenotypic characterization of seizure-associated copy number variations in 741,075 individuals. *Nature communications*, 14(1):4392, 2023.

[25] Chiara Auwerx, Maarja Jõeloo, Marie C Sadler, Nicolò Tesio, Sven Ojavee, Charlie J Clark, Reedik Mägi, Estonian Biobank Research Team Esko Tõnu Metspalu Andres Milani Lili Nelis Mari, Alexandre Reymond, and Zoltán Kutalik. Rare copy-number variants as modulators of common disease susceptibility. *Genome Medicine*, 16(1):5, 2024.

[26] Morteza Vaez, Simone Montalbano, Xabier Calle Sánchez, Kajsa-Lotta Georgii Hellberg, Saeid Rasekhi Dehkordi, Morten Dybdahl Krebs, Joeri Meijsen, John Shorter, Jonas Bybjerg-Grauholm, Preben B Mortensen, et al. Population-based risk of psychiatric disorders associated with recurrent copy number variants. *JAMA psychiatry*, 81(10):957–966, 2024.

[27] Morteza Vaez, Simone Montalbano, Ryan Waples, Morten Dybdahl Krebs, Kajsa-Lotta Georgii Hellberg, Jesper Gådin, Jonas Bybjerg-Grauholm, Preben B Mortensen, Anders D Børglum, Merete Nordentoft, et al. Evaluating the joint effects of recurrent copy number variants and polygenic scores on the risk of psychiatric disorders in the ipsych2015 case-cohort sample. *medRxiv*, pages 2024–09, 2024.

[28] Mikhail Vysotskiy, Xue Zhong, Tyne W Miller-Fleming, Dan Zhou, Autism Working Group of the Psychiatric Genomics Consortium^, Bipolar Disorder Working Group of the Psychiatric Genomics Consortium^, Schizophrenia Working Group of the Psychiatric Genomics Consortium^, Nancy J Cox, and Lauren A Weiss. Integration of genetic, transcriptomic, and clinical data provides insight into 16p11.2 and 22q11.2 cnv genes. *Genome Medicine*, 13:1–26, 2021.

[29] Malú Zamariolli, Chiara Auwerx, Marie C Sadler, Adriaan van der Graaf, Kaido Lepik, Tabea Schoeler, Mariana Moysés-Oliveira, Anelisa G Dantas, Maria Isabel Melaragno, and Zoltán Kutalik. The impact of 22q11.2 copy-number variants on human traits in the general population. *The American Journal of Human Genetics*, 110(2):300–313, 2023.

[30] Ruth Hanssen, Chiara Auwerx, Maarja Jõeloo, Marie C Sadler, Elana Henning, Julia Keogh, Rebecca Bounds, Miriam Smith, Helen V Firth, Zoltán Kutalik, et al. Chromosomal deletions on 16p11.2 encompassing sh2b1 are associated with accelerated metabolic disease. *Cell Reports Medicine*, 4(8):101155, 2023.

[31] Chiara Auwerx, Samuel Moix, Zoltán Kutalik, and Alexandre Reymond. Disentangling mechanisms behind the pleiotropic effects of proximal 16p11. 2 bp4-5 cnvs. *The American Journal of Human Genetics*, 111(11):2347–2361, 2024.

[32] Chiara Auwerx, Zoltán Kutalik, and Alexandre Reymond. The pleiotropic spectrum of proximal 16p11. 2 cnvs. *The American Journal of Human Genetics*, 111(11):2309–2346, 2024.

[33] Douglas S Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76, 1965.

[34] Amy K Olszewski, Petya D Radoeva, Wanda Fremont, Wendy R Kates, and Kevin M Antshel. Is child intelligence associated with parent and sibling intelligence in individuals with developmental disorders? an investigation in youth with 22q11. 2 deletion (velo-cardio-facial) syndrome. *Research in developmental disabilities*, 35(12):3582–3590, 2014.

[35] Andres Moreno-De-Luca, David W Evans, KB Boomer, Ellen Hanson, Raphael Bernier, Robin P Goin-Kochel, Scott M Myers, Thomas D Challman, Daniel Moreno-De-Luca, Mylissa M Slane, et al. The role of parental cognitive, behavioral, and motor profiles in clinical variability in individuals with chromosome 16p11.2 deletions. *JAMA Psychiatry*, 72(2):119–126, 2015.

[36] Debra D'Angelo, Sébastien Lebon, Qixuan Chen, Sandra Martin-Brevet, LeeAnne Green Snyder, Loyse Hippolyte, Ellen Hanson, Anne M Maillard, W Andrew Faucett, Aurélien Macé, et al. Defining the effect of the 16p11.2 duplication on cognition, behavior, and medical comorbidities. *JAMA Psychiatry*, 73(1):20–30, 2016.

[37] Petra Klaassen, Sasja Duijff, Henriëtte Swanenburg de Veye, Frits Beemer, Gerben Sinnema, Elemi Breetvelt, Renske Schappin, and Jacob Vorstman. Explaining the variable penetrance of cnvs: Parental intelligence modulates expression of intellectual impairment caused by the 22q11. 2 deletion. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 171(6):790–796, 2016.

[38] David W Evans and Mirko Uljarević. Parental education accounts for variability in the iqs of probands with down syndrome: A longitudinal study. *American Journal of Medical Genetics Part A*, 176(1):29–33, 2018.

[39] Corrine Smolen, Matthew Jensen, Lisa Dyer, Lucilla Pizzo, Anastasia Tyryshkina, Deepro Banerjee, Laura Rohan, Emily Huber, Laila El Khattabi, Paolo Prontera, et al. Assortative mating and parental genetic relatedness contribute to the pathogenicity of variably expressive variants. *The American Journal of Human Genetics*, 110(12):2015–2028, 2023.

[40] Rebecca Kingdom and Caroline F Wright. Incomplete penetrance and variable expressivity: From clinical studies to population cohorts. *Frontiers in Genetics*, 13:920390, 2022.

[41] Santhosh Girirajan and Evan E Eichler. Phenotypic variability and genetic susceptibility to genomic disorders. *Human Molecular Genetics*, 19 (R2):R176–R187, 2010.

[42] Santhosh Girirajan, Jill A Rosenfeld, Bradley P Coe, Sumit Parikh, Neil Friedman, Amy Goldstein, Robyn A Filipink, Juliann S McConnell, Brad Angle, Wendy S Meschino, et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *New England Journal of Medicine*, 367(14):1321–1331, 2012.

[43] Lucilla Pizzo, Matthew Jensen, Andrew Polyak, Jill A Rosenfeld, Katrin Mannik, Arjun Krishnan, Elizabeth McCready, Olivier Pichon, Cedric Le Caignec, Anke Van Dijck, et al. Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. *Genetics in Medicine*, 21(4):816–825, 2019.

[44] J Martin, MC O'donovan, A Thapar, K Langley, and N Williams. The relative contribution of common and rare genetic variants to adhd. *Translational psychiatry*, 5(2):e506–e506, 2015.

[45] Katherine E Tansey, Elliott Rees, DE Linden, S Ripke, KD Chambert, JL Moran, SA McCarroll, P Holmans, George Kirov, J Walters, et al. Common alleles contribute to schizophrenia in cnv carriers. *Molecular psychiatry*, 21(8):1085–1089, 2016.

[46] Daniel J Weiner, Emilie M Wigdor, Stephan Ripke, Raymond K Walters, Jack A Kosmicki, Jakob Grove, Kaitlin E Samocha, Jacqueline I Goldstein, Aysu Okbay, Jonas Bybjerg-Grauholm, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature genetics*, 49(7):978–985, 2017.

[47] Sarah E Bergen, Alexander Ploner, Daniel Howrigan, CNV Analysis Group, the Schizophrenia Working Group of the Psychiatric Genomics Consortium, Michael C O'Donovan, Jordan W Smoller, Patrick F Sullivan, Jonathan Sebat, Benjamin Neale, and Kenneth S Kendler. Joint contributions of rare copy number variants and common snps to risk for schizophrenia. *American Journal of Psychiatry*, 176(1):29–35, 2019.

[48] Mitja I Kurki, Elmo Saarentaus, Olli Pietiläinen, Padhraig Gormley, Dennis Lal, Sini Kerminen, Minna Torniainen-Holm, Eija Hämäläinen, Elisa Rahikkala, Riikka Keski-Filppula, et al. Contribution of rare and common variants to intellectual disability in a sub-isolate of northern finland. *Nature communications*, 10(1):410, 2019.

[49] MT Oetjens, MA Kelly, AC Sturm, CL Martin, and DH Ledbetter. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. *Nature Communications*, 10(1):4897, 2019.

[50] Satoru Taniguchi, Kohei Ninomiya, Itaru Kushima, Takeo Saito, Ayu Shimasaki, Takaya Sakusabe, Yukihide Momozawa, Michiaki Kubo, Yoichiro Kamatani, Norio Ozaki, et al. Polygenic risk scores in schizophrenia with clinically significant copy number variants. *Psychiatry and clinical neurosciences*, 74(1):35–39, 2020.

[51] Robert W Davies, Ania M Fiksinski, Elemi J Breetvelt, Nigel M Williams, Stephen R Hooper, Thomas Monfeuga, Anne S Bassett, Michael J Owen, Raquel E Gur, Bernice E Morrow, et al. Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nature Medicine*, 26(12):1912–1918, 2020.

[52] Sonja LaBianca, Jette LaBianca, Anne Katrine Pagsberg, Klaus Damgaard Jakobsen, Vivek Appadurai, Alfonso Buil, and Thomas Werge. Copy number variants and polygenic risk scores predict need of care in autism and/or adhd families. *Journal of autism and developmental disorders*, 51:276–285, 2021.

[53] Lambertus Klei, Lora Lee McClain, Behrang Mahjani, Klea Panayidou, Silvia De Rubeis, Anna-Carin Säll Grahnat, Gun Karlsson, Yangyi Lu, Nadine Melhem, Xinyi Xu, et al. How rare and common risk variation jointly affect liability for autism spectrum disorder. *Molecular autism*, 12: 1–13, 2021.

[54] Josephine Mollon, Laura M Schultz, Guillaume Huguet, Emma EM Knowles, Samuel R Mathias, Amanda Rodrigue, Aaron Alexander-Bloch, Zohra Saci, Martineau Jean-Louis, Kuldeep Kumar, et al. Impact of copy number variants and polygenic risk scores on psychopathology in the uk biobank. *Biological Psychiatry*, 94(7):591–600, 2023.

[55] Thomas J Dinneen, Fiana Ní Ghrálaigh, Cathal Ormond, Elizabeth A Heron, George Kirov, Lorna M Lopez, and Louise Gallagher. Polygenic scores stratify neurodevelopmental copy number variant carrier cognitive outcomes in the uk biobank. *npj Genomic Medicine*, 9(1):43, 2024.

[56] Rebecca Kingdom, Robin N Beaumont, Andrew R Wood, Michael N Weedon, and Caroline F Wright. Genetic modifiers of rare variants in monogenic developmental disorder loci. *Nature Genetics*, 56(5):861–868, 2024.

[57] Maris Alver, Valentina Mancini, Kristi Läll, Maude Schneider, Luciana Romano, Estonian Biobank Research Team Milani Lili 6 Nelis Mari 6 Mägi Reedik 6 Esko Tõnu 6 Metspalu Andres 6, Reedik Mägi, Emmanouil T Dermitzakis, Stephan Eliez, and Alexandre Reymond. Contribution of schizophrenia polygenic burden to longitudinal phenotypic variance in 22q11. 2 deletion syndrome. *Molecular psychiatry*, 27(10):4191–4200, 2022.

[58] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American Journal of Epidemiology*, 186(9):1026–1034, 2017.

[59] Nicola F Smith, William D Figg, and Alex Sparreboom. Role of the liver-specific transporters oatp1b1 and oatp1b3 in governing drug elimination. *Expert Opinion on Drug Metabolism & Toxicology*, 1(3):429–445, 2005.

[60] I Sadaf Farooqi, Julia M Keogh, Giles SH Yeo, Emma J Lank, Tim Cheetham, and Stephen O'Rahilly. Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. *New England Journal of Medicine*, 348(12):1085–1095, 2003.

[61] James R Lupski, Roberto Montes de Oca-Luna, Susan Slaugenhaupt, Liu Pentao, Vito Guzzetta, Barbara J Trask, Odila Saucedo-Cardenas, David F Barker, James M Killian, Carlos A Garcia, et al. Dna duplication associated with charcot-marie-tooth disease type 1a. *Cell*, 66(2): 219–232, 1991.

[62] Barbara W van Paassen, Anneke J van der Kooi, Karin Y van Spaendonck-Zwarts, Camiel Verhamme, Frank Baas, and Marianne de Visser. Pmp22 related neuropathies: Charcot-marie-tooth disease type 1a and hereditary neuropathy with liability to pressure palsies. *Orphanet Journal of Rare Diseases*, 9:1–15, 2014.

[63] Michael E Doche, Elena G Bochukova, Hsiao-Wen Su, Laura R Pearce, Julia M Keogh, Elana Henning, Joel M Cline, Anne Dale, Tim Cheetham, Inês Barroso, et al. Human sh2b1 mutations are associated with maladaptive behaviors and obesity. *The Journal of clinical investigation*, 122(12):4732–4736, 2012.

[64] Jose Morales, Latifa Al-Sharif, Dania S Khalil, Jameela MA Shinwari, Prashant Bavi, Rahima A Al-Mahrouqi, Ali Al-Rajhi, Fowzan S Alkuraya, Brian F Meyer, and Nada Al Tassan. Homozygous mutations in adamts10 and adamts17 cause lenticular myopia, ectopia lentis, glaucoma, spherophakia, and short stature. *The American Journal of Human Genetics*, 85(5):558–568, 2009.

[65] Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K Lucas, Jean Monlong, Haley J Abel, et al. A draft human pangenome reference. *Nature*, 617(7960):312–324, 2023.

[66] Bruno Hagenbuch and Bruno Stieger. The slco (former slc21) superfamily of transporters. *Molecular aspects of medicine*, 34(2-3):396–412, 2013.

[67] Akl C Fahed, Minxian Wang, Julian R Homburger, Aniruddh P Patel, Alexander G Bick, Cynthia L Neben, Carmen Lai, Deanna Brockman, Anthony Philippakis, Patrick T Ellinor, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nature communications*, 11(1):3635, 2020.

[68] Julia K Goodrich, Moriel Singer-Berk, Rachel Son, Abigail Sveden, Jordan Wood, Eleina England, Joanne B Cole, Ben Weisburd, Nick Watts, Lizz Caulkins, et al. Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. *Nature communications*, 12(1):3505, 2021.

[69] Mari EK Niemi, Hilary C Martin, Daniel L Rice, Giuseppe Gallone, Scott Gordon, Martin Kelemen, Kerrie McAloney, Jeremy McRae, Elizabeth J Radford, Sui Yu, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*, 562(7726):268–271, 2018.

[70] Lea Urpa, Mitja I Kurki, Elisa Rahikkala, Eija Hämäläinen, Veikko Salomaa, Jaana Suvisaari, Riikka Keski-Filppula, Merja Rauhala, Satu Korpi-Heikkilä, Jonna Komulainen-Ebrahim, et al. Evidence for the additivity of rare and common variant burden throughout the spectrum of intellectual disability. *European Journal of Human Genetics*, 32(5):576–583, 2024.

[71] Parsa Akbari, Ankit Gilani, Olukayode Sosina, Jack A Kosmicki, Lori Khrimian, Yi-Ya Fang, Trikaldarshi Persaud, Victor Garcia, Dylan Sun, Alexander Li, et al. Sequencing of 640,000 exomes identifies gpr75 variants associated with protection from obesity. *Science*, 373(6550): eabf8683, 2021.

[72] Deepro Banerjee and Santhosh Girirajan. Discovery of novel obesity genes through cross-ancestry analysis. *medRxiv*, 2025.

[73] Elmo Christian Saarentaus, Aki Samuli Havulinna, Nina Mars, Ari Ahola-Olli, Tuomo Tapio Johannes Kiiskinen, Juulia Partanen, Sanni Ruotsalainen, Mitja Kurki, Lea Martta Urpa, Lei Chen, et al. Polygenic burden has broader impact on health, cognition, and socioeconomic outcomes than most rare and high-risk copy number variants. *Molecular Psychiatry*, 26(9):4884–4895, 2021.

[74] Jonathan Sulc, Ninon Mounier, Felix Günther, Thomas Winkler, Andrew R Wood, Timothy M Frayling, Iris M Heid, Matthew R Robinson, and Zoltán Kutalik. Quantification of the overall contribution of gene-environment interaction for obesity-related traits. *Nature communications*, 11 (1):1385, 2020.

[75] Madison Chapel, Jessica Dennis, and Carl G de Boer. Variant effects depend on polygenic background: experimental, clinical, and evolutionary implications. *bioRxiv*, pages 2025–01, 2025.

[76] Jennifer Sjaarda and Zoltán Kutalik. Partner choice, confounding and trait convergence all contribute to phenotypic partner similarity. *Nature Human Behaviour*, 7(5):776–789, 2023.

[77] Ashley E Nordsletten, Henrik Larsson, James J Crowley, Catarina Almqvist, Paul Lichtenstein, and David Mataix-Cols. Patterns of nonrandom mating within and across 11 major psychiatric disorders. *JAMA psychiatry*, 73(4):354–361, 2016.

[78] Tanya B Horwitz, Jared V Balbona, Katie N Paulich, and Matthew C Keller. Evidence of correlations between human partners based on systematic reviews and meta-analyses of 22 traits and uk biobank analysis of 133 traits. *Nature Human Behaviour*, 7(9):1568–1583, 2023.

[79] Chun Chieh Fan, Saeid Rasekhi Dehkordi, Richard Border, Lucy Shao, Bohan Xu, Robert Loughnan, Wesley K Thompson, Le-Yin Hsu, Mei-Chen Lin, Chi-Fung Cheng, et al. Assortative mating across nine psychiatric disorders is consistent and persistent over cultures and generations. *medRxiv*, pages 2024–09, 2024.

[80] Matthew R Robinson, Aaron Kleinman, Mariaelisa Graff, Anna AE Vinkhuyzen, David Couper, Michael B Miller, Wouter J Peyrot, Abdel Abdellaoui, Brendan P Zietsch, Ilja M Nolte, et al. Genetic evidence of assortative mating in humans. *Nature Human Behaviour*, 1(1):0016, 2017.

[81] Loic Yengo, Matthew R Robinson, Matthew C Keller, Kathryn E Kemper, Yuanhao Yang, Maciej Trzaskowski, Jacob Gratten, Patrick Turley, David Cesarini, Daniel J Benjamin, et al. Imprint of assortative mating on the human genome. *Nature human behaviour*, 2(12):948–954, 2018.

[82] Laurence J Howe, Daniel J Lawson, Neil M Davies, Beate St. Pourcain, Sarah J Lewis, George Davey Smith, and Gibran Hemani. Genetic evidence for assortative mating on alcohol consumption in the uk biobank. *Nature communications*, 10(1):5039, 2019.

[83] Richard Border, Sean O'Rourke, Teresa de Candia, Michael E Goddard, Peter M Visscher, Loic Yengo, Matt Jones, and Matthew C Keller. Assortative mating biases marker-based heritability estimators. *Nature communications*, 13(1):660, 2022.

[84] Richard Border, Georgios Athanasiadis, Alfonso Buil, Andrew J Schork, Na Cai, Alexander I Young, Thomas Werge, Jonathan Flint, Kenneth S Kendler, Sriram Sankararaman, et al. Cross-trait assortative mating is widespread and inflates genetic correlation estimates. *Science*, 378 (6621):754–761, 2022.

[85] Fartein Ask Torvik, Espen Moen Eilertsen, Laurie J Hannigan, Rosa Cheesman, Laurence J Howe, Per Magnus, Ted Reichborn-Kjennerud, Ole A Andreassen, Pål R Njølstad, Alexandra Havdahl, et al. Modeling assortative mating and genetic similarities between partners, siblings, and in-laws. *Nature Communications*, 13(1):1108, 2022.

[86] Brenda Cabrera-Mendoza, Frank R Wendt, Gita A Pathak, Loic Yengo, and Renato Polimanti. The impact of assortative mating, participation bias and socioeconomic status on the polygenic risk of behavioural and psychiatric traits. *Nature Human Behaviour*, 8(5):976–987, 2024.

[87] Carl Veller and Graham M Coop. Interpreting population-and family-based genome-wide association studies in the presence of confounding. *PLoS Biology*, 22(4):e3002511, 2024.

[88] Hans Fredrik Sunde, Nikolai Haahjem Eftedal, Rosa Cheesman, Elizabeth C Corfield, Thomas H Kleppesto, Anne Caroline Seierstad, Eivind

Ystrom, Espen Moen Eilertsen, and Fartein Ask Torvik. Genetic similarity between relatives provides evidence on the presence and history of assortative mating. *Nature Communications*, 15(1):2641, 2024.

[89] Hans Fredrik Sunde, Espen Moen Eilertsen, and Fartein Ask Torvik. Understanding indirect assortative mating and its intergenerational consequences for educational attainment. *Nature Communications*, 16(1):1–14, 2025.

[90] Qin Qin Huang, Emilie M Wigdor, Daniel S Malawsky, Patrick Campbell, Kaitlin E Samocha, V Kartik Chundru, Petr Danecek, Sarah Lindsay, Thomas Marchant, Mahmoud Koko, et al. Examining the role of common variants in rare neurodevelopmental conditions. *Nature*, pages 1–8, 2024.

[91] Qinwen Zheng, Sjoerd van Alten, Torkild Hovde Lyngstad, Edoardo Ciscato, Zhongxuan Sun, Jiacheng Miao, Yuchang Wu, Stephen Dorn, Boyan Zheng, Alexandra Havdahl, et al. Genetic basis of partner choice. *bioRxiv*, pages 2025–02, 2025.

[92] Michael G Levin, Satoshi Koyama, Jakob Woerner, David Y Zhang, Alexis Rodriguez, Tarak Nandi, Buu Truong, Sarah A Abramowitz, Hritvik Gupta, Himani Kamineni, et al. Genome-wide assessment of pleiotropy across¿ 1000 traits from global biobanks. *medRxiv*, pages 2025–04, 2025.

[93] Emmanuel Stamatakis, Katherine B Owen, Leah Shepherd, Bradley Drayton, Mark Hamer, and Adrian E Bauman. Is cohort representativeness passé? poststratified associations of lifestyle risk factors with mortality in the uk biobank. *Epidemiology*, 32(2):179–188, 2021.

[94] Sjoerd van Alten, Benjamin W Domingue, Jessica Faul, Titus Galama, and Andries T Marees. Reweighting uk biobank corrects for pervasive selection bias due to volunteering. *International Journal of Epidemiology*, 53(3):dyae054, 2024.

[95] J Alexander Cole, John S Taylor, Thomas N Hangartner, Neal J Weinreb, Pramod K Mistry, and Aneal Khan. Reducing selection bias in case-control studies from rare disease registries. *Orphanet Journal of Rare Diseases*, 6:1–7, 2011.

[96] Katherine Jonas, Amna Asim, Yuan Yang, Urs Heilbronner, Thomas Schulze, Todd Lencz, Evangelos Vassos, Sean Clouston, and Roman Kotov. The impact of selection bias on genetic prediction using the bipolar polygenic risk score in first-admission psychosis. *ResearchSquare*, 2024.

[97] Younga Heather Lee, Tanayott Thaweethai, Yi-Han Sheu, Yen-Chen Anne Feng, Elizabeth W Karlson, Tian Ge, Peter Kraft, and Jordan W Smoller. Impact of selection bias on polygenic risk score estimates in healthcare settings. *Psychological Medicine*, 53(15):7435–7445, 2023.

[98] Aditya Pimplaskar, Junqiong Qiu, Sandra Lapinska, Veronica Tozzo, Jeffrey N Chiang, Bogdan Pasaniuc, and Loes M Olde Loohuis. Inclusion bias affects common variant discovery and replication in a health-system linked biobank. *medRxiv*, pages 2025–04, 2025.

[99] Tabea Schoeler, Doug Speed, Eleonora Porcu, Nicola Pirastu, Jean-Baptiste Pingault, and Zoltán Kutalik. Participation bias in the uk biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour*, 7(7):1216–1227, 2023.

[100] Ryan L Collins, Harrison Brand, Konrad J Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C Francioli, Amit V Khera, Chelsea Lowther, Laura D Gauthier, Harold Wang, et al. A structural variation reference for medical and population genetics. *Nature*, 581(7809):444–451, 2020.

[101] Tomas Fitzgerald and Ewan Birney. Cnest: A novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the uk biobank. *Cell Genomics*, 2(8), 2022.

[102] Boris Noyvert, A Mesut Erzurumluoglu, Dmitriy Drichel, Steffen Omland, Till FM Andlauer, Stefanie Mueller, Lau Sennels, Christian Becker, Aleksandr Kantorovich, Boris A Bartholdy, et al. Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. *medRxiv*, pages 2023–12, 2023.

[103] Shuwei Li, Keren J Carss, Bjarni V Halldorsson, Adrian Cortes, and UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of half-a-million uk biobank participants. *medRxiv*, pages 2023–12, 2023.

[104] Yi Ding, Kangcheng Hou, Ziqi Xu, Aditya Pimplaskar, Ella Petter, Kristin Boulier, Florian Privé, Bjarni J Vilhjálmsson, Loes M Olde Loohuis, and Bogdan Pasaniuc. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, 618(7966):774–781, 2023.

[105] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[106] Joshua D Backman, Alexander H Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D Kessler, Christian Benner, Daren Liu, Adam E Locke, Suganthi Balasubramanian, et al. Exome sequencing and analysis of 454,787 uk biobank participants. *Nature*, 599(7886):628–634, 2021.

[107] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan FA Grant, Hakon Hakonarson, and Maja Bucan. Penncnv: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. *Genome Research*, 17(11):1665–1674, 2007.

[108] A Mace, MA Tuke, JS Beckmann, L Lin, S Jacquemont, MN Weedon, A Reymond, and Z Kutalik. New quality measure for snp array based cnv detection. *Bioinformatics*, 32(21):3298–3305, 2016.

[109] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

[110] Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature genetics*, 53(7):1097–1103, 2021.

[111] Florian Privé, Julyan Arbel, and Bjarni J Vilhjálmsson. Ldpred2: better, faster, stronger. *Bioinformatics*, 36(22-23):5424–5431, 2020.

[112] Florian Privé. LD reference for HapMap3+. 10 2022. doi: $10.6084/m9.figshare.21305061.v2$. URL `https://figshare.com/articles/dataset/LD_reference_for_HapMap3_/21305061`.

[113] Florian Privé, Clara Albiñana, Julyan Arbel, Bogdan Pasaniuc, and Bjarni J Vilhjálmsson. Inferring disease architecture and predictive ability with ldpred2-auto. *The American Journal of Human Genetics*, 110(12):2042–2055, 2023.

[114] Samuel A Lambert, Benjamin Wingfield, Joel T Gibson, Laurent Gil, Santhi Ramachandran, Florent Yvon, Shirin Saverimuttu, Emily Tinsley, Elizabeth Lewis, Scott C Ritchie, et al. Enhancing the polygenic score catalog with tools for score calculation and ancestry normalization. *Nature Genetics*, 56(10):1989–1994, 2024.

[115] Ken Kelley and Keke Lai. Mbess: The mbess r package, 2025. URL `https://CRAN.R-project.org/package=MBESS`. R package version 4.9.41.

[116] Konrad J Karczewski, Matthew Solomonson, Katherine R Chao, Julia K Goodrich, Grace Tiao, Wenhan Lu, Bridget M Riley-Gillis, Ellen A Tsai, Hye In Kim, Xiuwen Zheng, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 uk biobank exomes. *Cell genomics*, 2(9), 2022.

[117] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[118] Alan Genz and Frank Bretz. mvtnorm: Multivariate normal and t distributions, 2025. URL `https://CRAN.R-project.org/package=mvtnorm`. R package version 1.3.3.

[119] Robin J Hofmeister, Davide Marnetto, Théo Cavinato, Leona Knüsel, Giorgia Modenini, Erwan Pennarun, Estonian Biobank research team, Nicola Barban, Reedik Mägi, Tabea Schoeler, et al. Parental haplotypes reconstruction in up to 440,209 individuals reveals recent assortative mating dynamics. *bioRxiv*, pages 2025–09, 2025.

[120] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[121] Robin J Hofmeister, Simone Rubinacci, Diogo M Ribeiro, Alfonso Buil, Zoltán Kutalik, and Olivier Delaneau. Parent-of-origin inference for biobanks. *Nature Communications*, 13(1):6668, 2022.

[122] Robin J Hofmeister, Théo Cavinato, Roya Karimi, Adriaan van der Graaf, Fanny-Dhelia Pajuste, Jaanika Kronberg, Nele Taba, Estonian Biobank research team Metspalu Andres 4 Esko Tõnu 4 Nelis Mari 4 Hudjashov Georgi 4, Reedik Mägi, Marc Vaudel, et al. Parent-of-origin effects on complex traits in up to 236,781 individuals. *Nature*, pages 1–10, 2025.

[123] W. N. Venables and B. D. Ripley. Mass: Support functions and datasets for venables and ripley's mass, 2025. URL `https://CRAN.R-project.org/package=MASS`. R package version 7.3.65.

[124] Miao Yang Ziqian Xu Zhiyong Zhang, Yujiao Mai and Conor McNamara. Webpower: Basic and advanced statistical power analysis, 2023. URL `https://CRAN.R-project.org/package=WebPower`. R package version 0.9.4.

[125] Julian C Stanley and Marilyn D Wang. Restrictions on the possible values of r12, given r13 and r23. *Educational and Psychological Measurement*, 29(3):579–581, 1969.

# Supplementary Figures

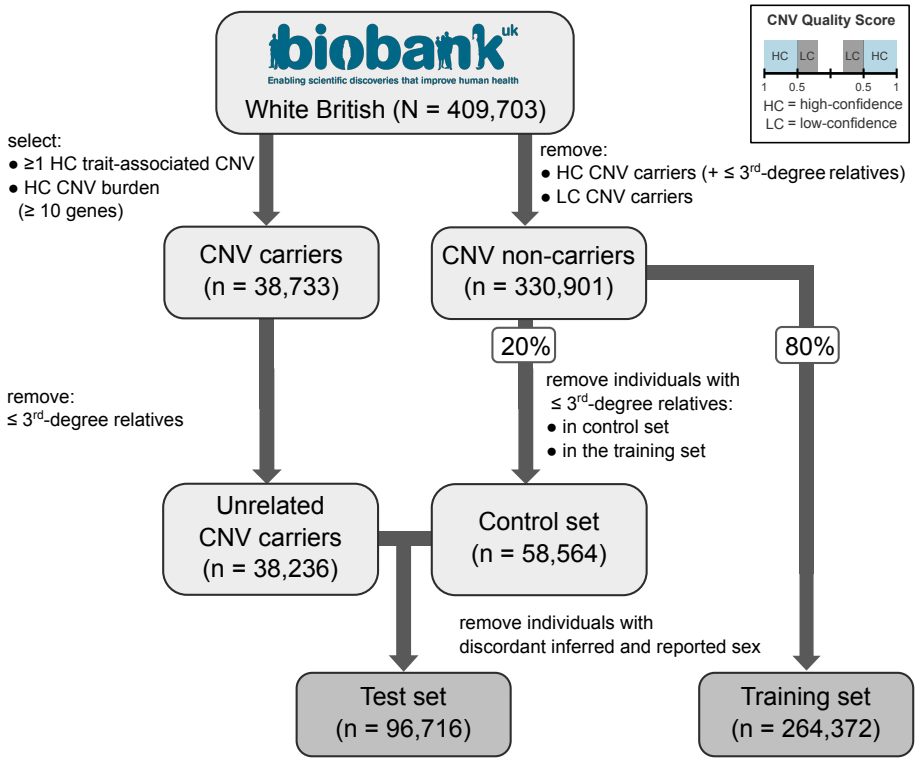


**Figure S1. Study design.**
Schematic representation of the study design. We split 409,703 "white British" UK Biobank samples into a training set used to develop polygenic score (PGS) weights, and a CNV-carrier-enriched test set used to perform analyses. Individuals carrying a high-confidence (HC) CNV call (absolute value of quality score: $\geq$ 0.5) overlapping the lead probe of one of the assessed CNV-trait pairs or a high CNV burden ($\geq$ 10 genes) were pre-selected for the test set and related individuals were pruned. These individuals and all their relatives were excluded from the 409,703 "white British" samples, along with individuals carrying a lower-confidence (LC) CNV call (absolute value of quality score: 0.2-0.5). Remaining individuals were randomly split into training (80%) and control (20%) sets with matching age and sex distributions. The control set was pruned to unrelated individuals, samples related to samples in the training set were excluded, and remaining individuals were merged with pre-selected CNV carriers to form the test set. Finally, individuals with mismatched reported versus genetically inferred sex were excluded.

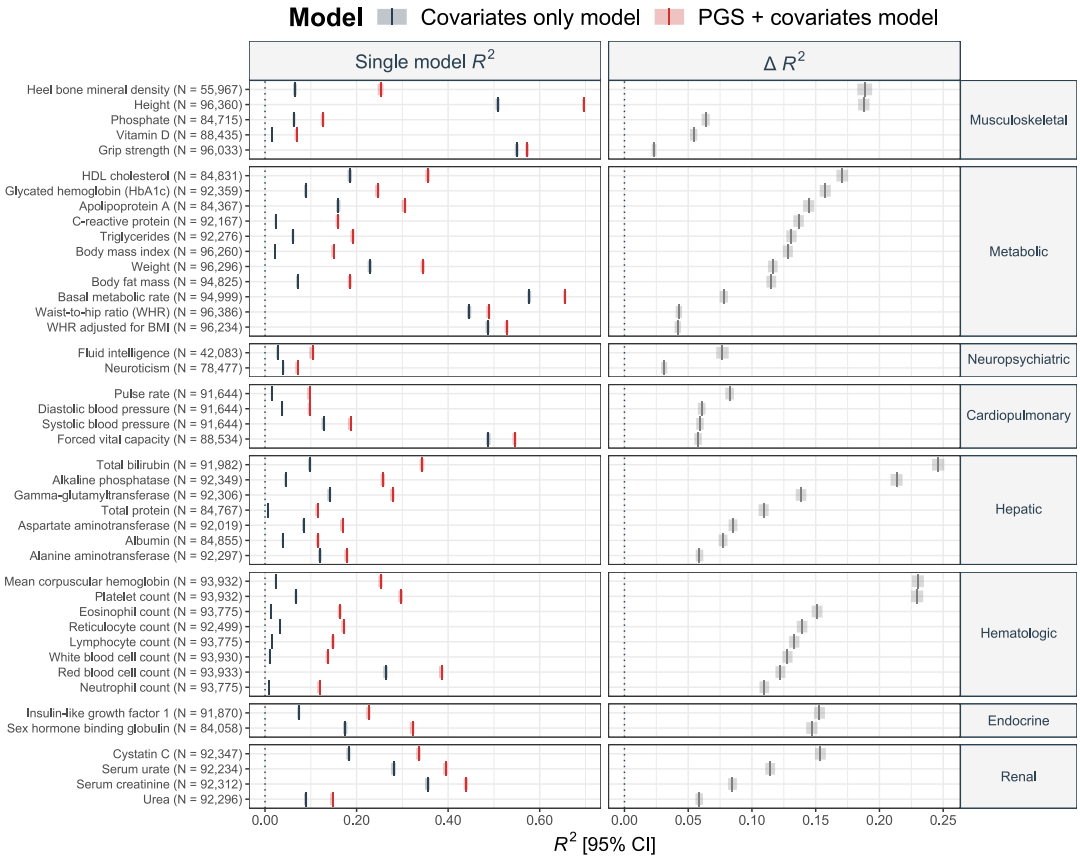


**Figure S2. Phenotypic variance explained by $PGS_{GW}$.**
The left panel shows the $R^2$ (x-axis) with 95% confidence interval (CI) of a linear regression explaining adjusted phenotypic values for the 43 traits assessed in this study (y-axis, left) as a function of covariates only (i.e., sex, age, age$^2$, genotyping batch, and 40 principal components; in grey) or covariates and genome-wide polygenic score ($PGS_{GW}$; in red). The right panel depicts the difference in $R^2$ between these two models ($\Delta R^2$), with 95%-CI. Phenotypes are ordered by decreasing $\Delta R^2$ within each phenotype category (y-axis, right) and the number of samples in the test set used to fit the model is indicated as N (y-axis, left).

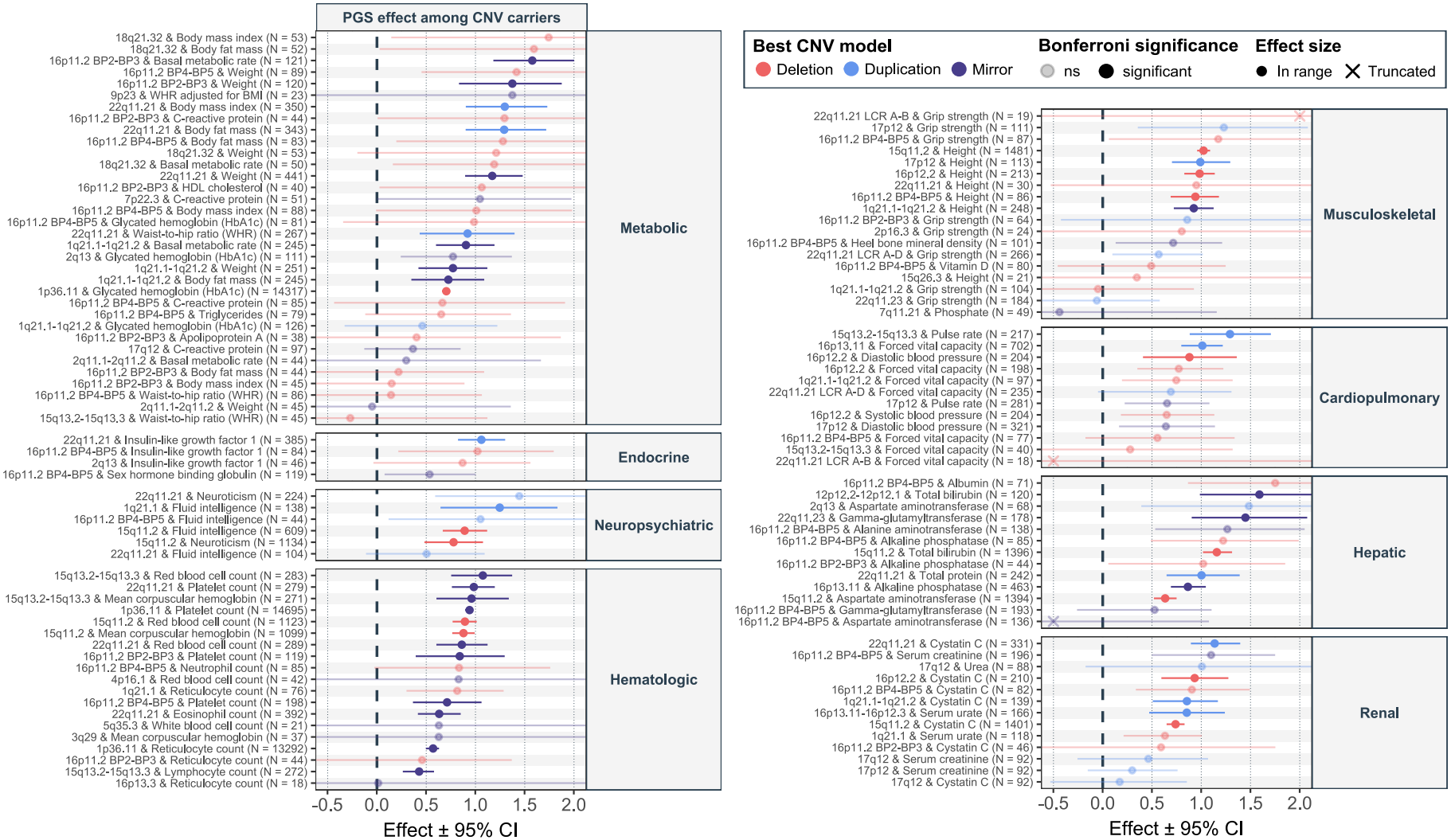


**Figure S3. $PGS_{GW}$ effect among CNV carriers.**
Standardized effect sizes with 95% confidence intervals (CI) of the $PGS_{GW}$ effect on the phenotype (x-axis) among CNV carriers for the 119 unique CNV-trait pairs assessed in this study (y-axis, left) ordered by decreasing effect size within each trait categories (y-axis, right). $PGS_{GW}$ effects were assessed among deletion (red), duplication (blue), or CNV (purple) carriers depending on whether the CNV-trait association was previously found to be most significant according to a deletion-only, duplication-only, or mirror model, respectively in Auwerx *et al.*, 2022. The number of considered carriers is indicated as N (y-axis, left). Transparency indicates whether the effect is significant under Bonferroni correction ($p \leq 0.05/119 = 4.2 \times 10^{-4}$) or not. Some non-significant effects were truncated to improve readability and are indicated with a cross instead of a circle.

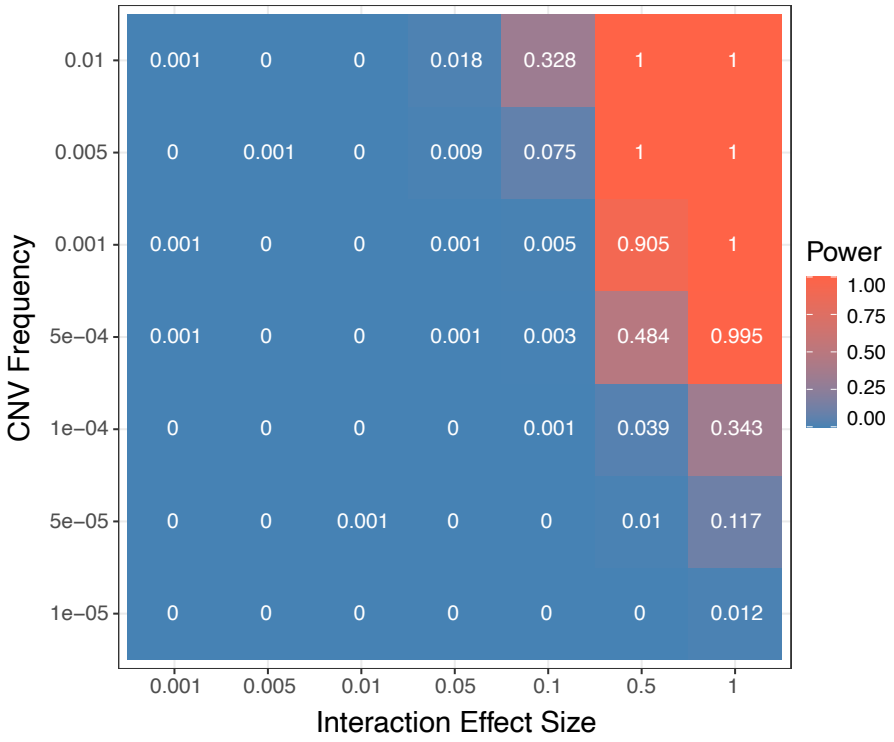


**Figure S4. Power to detect CNV × PGS interactions.**
Simulations were conducted across a grid of CNV frequencies (y-axis) and interaction effect sizes (x-axis; absolute values) using a sample size equivalent to our test set (N = 96,716). Color represents the power to detect an interaction of a given effect size at a specific CNV frequency based on 1,000 simulations.

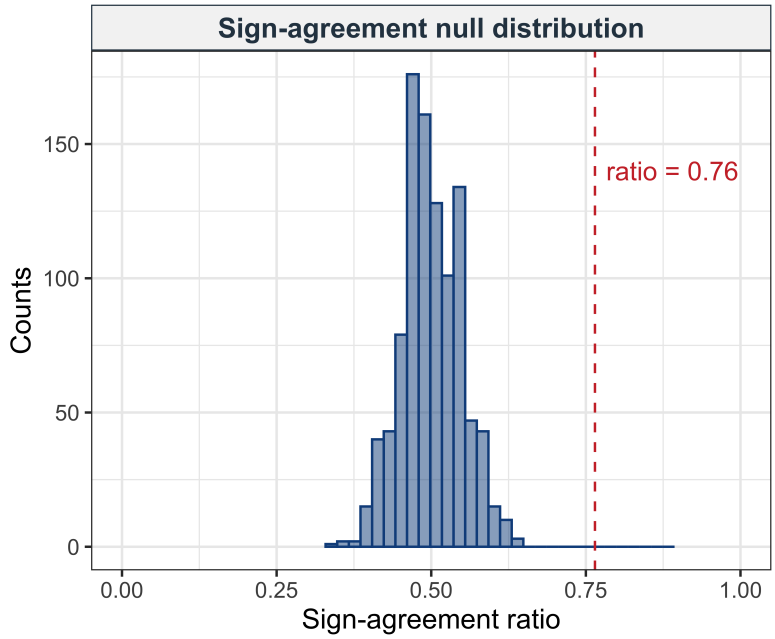


**Figure S5. Directional concordance of CNV effects on phenotypes and $PGS_{GW}$.**
Histogram showing the null distribution of CNV effects on phenotype and $PGS_{GW}$ sign agreement, accounting for the $PGS_{GW}$ and phenotype covariance structure across all 43 traits. The empirical sign-agreement ratio is indicated by the dashed red vertical dashed line.

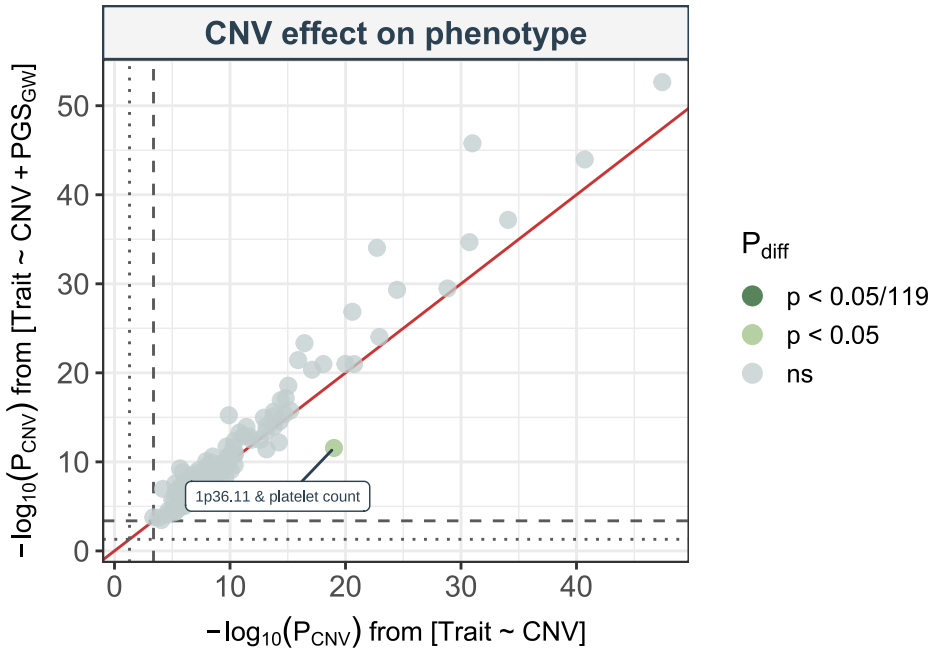


**Figure S6. Conditioning of the CNV-phenotype relation on $PGS_{GW}$.**
**c,** Negative logarithm of the p-value of the effect of the CNV on adjusted phenotypes, when conditioning on $PGS_{GW}$ (y-axis) or not (x-axis). Data points are colored according to the significance degree of a t-test assessing whether the CNV effect on the adjusted phenotype is different in the two compared models, with "ns" indicating $p > 0.05$. The single nominally significant effect is labeled. The red diagonal represents the identity, while the grey dashed and dotted lines mark the threshold for Bonferroni ($-\log_{10}(0.05/119) = 3.38$) and nominal ($-\log_{10}(0.05) = 1.30$) significance, respectively.

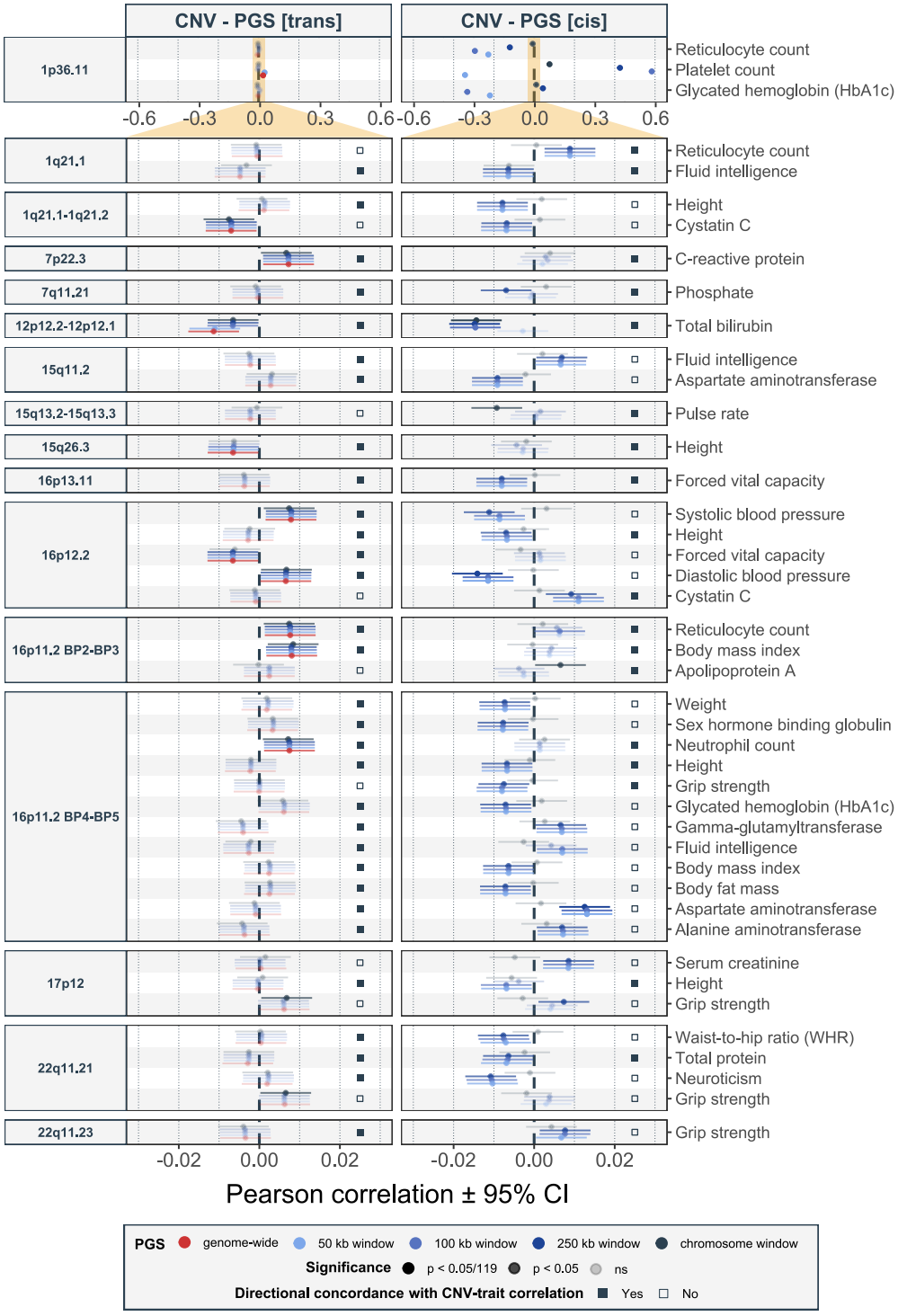


**Figure S7. Partial PGS-CNV correlations.**
Pearson correlation with 95% confidence interval (CI; x-axis) between polygenic scores (PGSs; y-axis, right) and CNV carrier status (y-axis, left) for all CNV-trait pairs with at least one nominally significant CNV-$PGS_{GW}$, CNV-$PGS_{trans}$, or CNV-$PGS_{cis}$ correlation. The left panel shows correlation with the genome-wide $PGS_{GW}$ (red) and various $PGS_{trans}$ (blue shades), while the right panel shows correlations with various $PGS_{cis}$ (blue shades). Transparency reflects the significance level, with "ns" indicating $p > 0.05$. The small grey square indicates whether the correlation between the CNV carrier status and $PGS_{trans\text{-}250kb}$ (left panel) or $PGS_{cis\text{-}250kb}$ (right panel) are directionally concordant with the CNV's effect on the trait. We use a different x-axis scale for the 1p36.11 CNV region for better visualization (orange area).

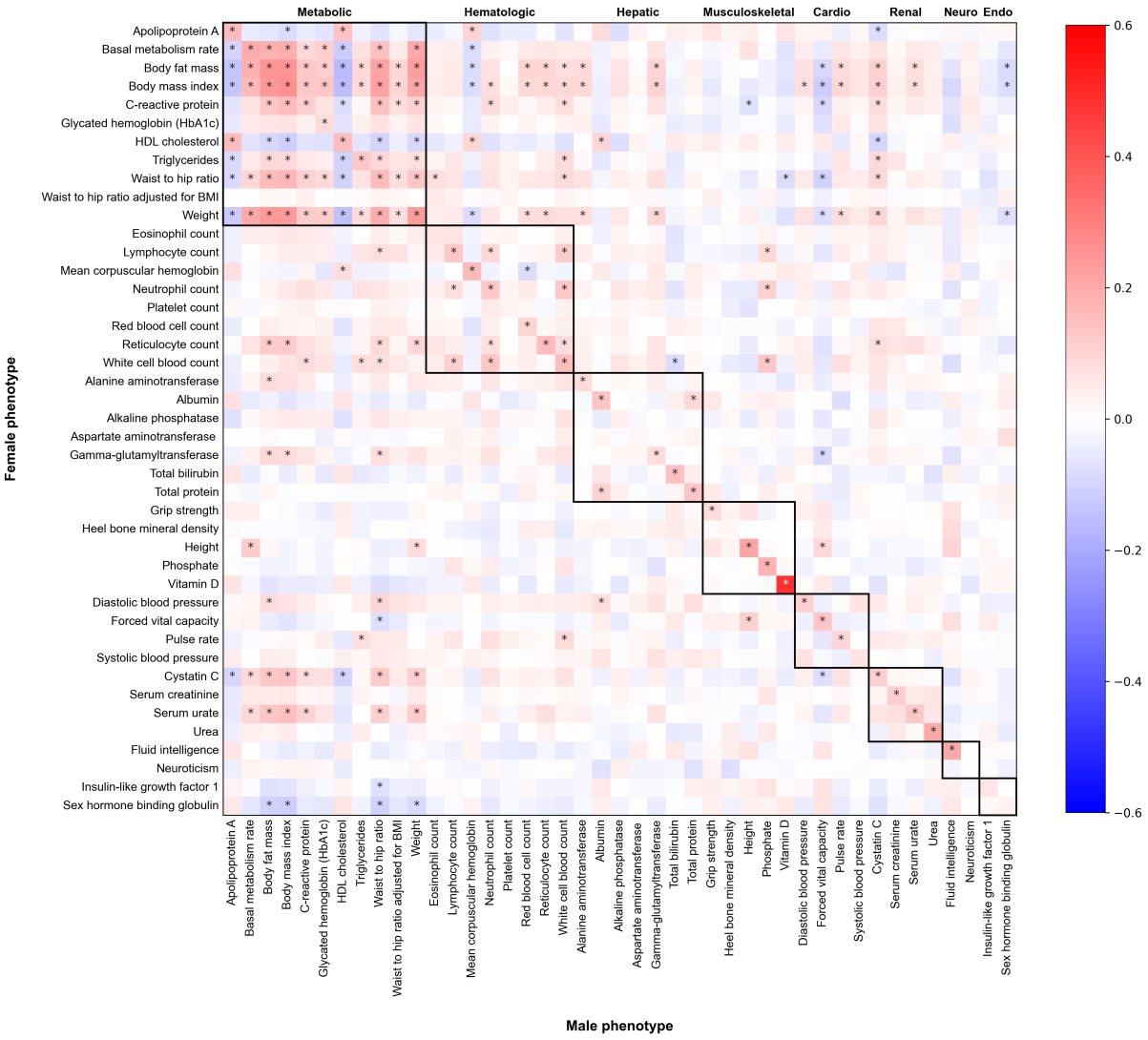


**Figure S8. Pairwise partner phenotypic correlation.**
Heatmap of Pearson correlation coefficients between 43 adjusted phenotypes in males (x-axis) versus the corresponding phenotype in their female partner (y-axis). Stars indicate Bonferroni-significant correlations ($p \leq 0.05/43^2 = 2.7 \times 10^{-5}$). Phenotypes are clustered by trait categories, as depicted by bolded squares and top labels. Cardio = Cardiometabolic; Neuro = Neuropsychiatric; Endo = Endocrine.

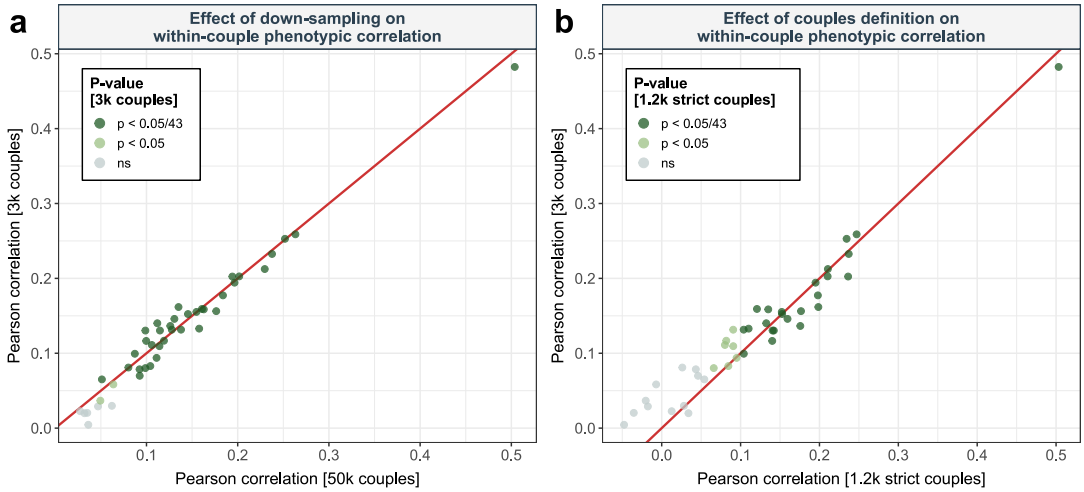


**Figure S9. Impact of couples selection on same-trait within-couple phenotypic correlation.**
**a,** Impact of down-sampling: Same-trait, within-couple Pearson phenotypic correlation among the 3,396 couples present in our testing subset (y-axis, [3k]) against same-trait, within-couple Pearson phenotypic correlation among the 51,664 couples present in the entire "white British" subset of the UKBB (x-axis, [50k]). All phenotypes show Bonferroni-significant ($p \leq 0.05/43 = 1.2 \times 10^{-3}$) within-couple correlation among the 50k couple subset and data points are colored according to the significance level of the within-couple phenotypic correlation among the 3k subset, with "ns" indicating $p > 0.05$. The red diagonal represents the identity line. **b,** Impact of stricter couples definition: Same-trait, within-couple Pearson phenotypic correlation among the 3,396 couples defined based on our original couples definition (y-axis, [3k]) against same-trait, within-couple Pearson phenotypic correlation among the 1,213 couples defined according to a strict couples definition present in our test subset (x-axis, [1.2k strict]). Data points are colored according to the significance level of the within-couple phenotypic correlation based on the strict couples definition, with "ns" indicating $p > 0.05$. The red diagonal represents the identity line.

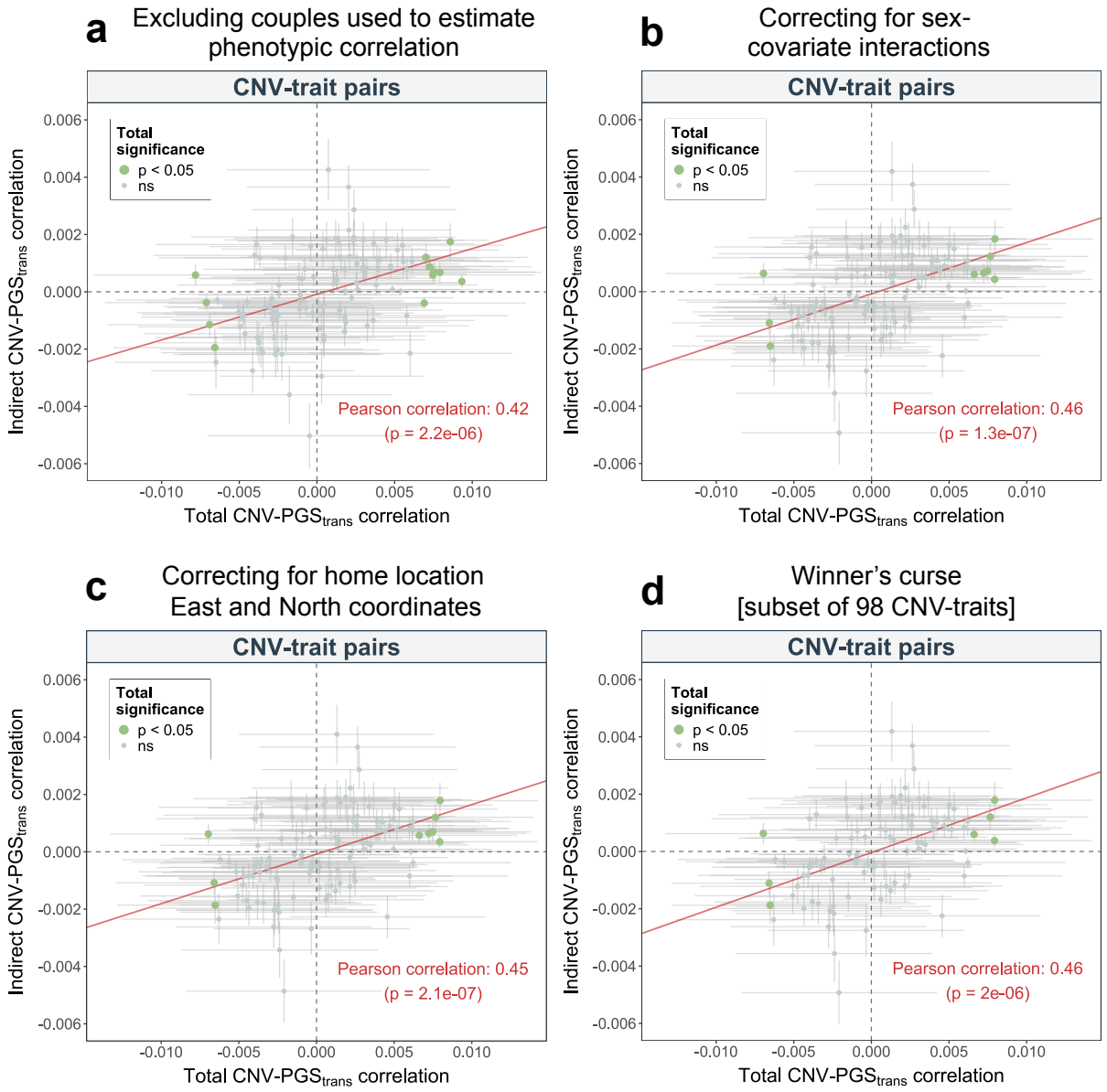


**Figure S10. Assortative mating sensitivity analyses.**
Sensitivity analyses probing the robustness of the correlation between total (x-axis) and indirect (y-axis; through assortative mating) CNV-PGS$_{trans\text{-}250kb}$ correlation. **a,** Excluding the couples used for $\beta$ estimation (i.e., within-couple phenotype correlation) from the estimation of $\alpha$ and $\gamma$. **b,** Using phenotypic values corrected for sex $\times$ covariate interactions. **c,** Using phenotypic values corrected for home location East and North coordinates (1 km resolution). **d,** Restricting the analysis to 98 CNV-trait pairs with $p \leq 0.05/(11{,}804 \times 3)$ in the original CNV-GWAS study (Auwerx *et al.*, 2022) to assess the impact of Winner's curse. Error bars represent 95% confidence intervals. Size and color reflect the significance level of the total CNV-PGS$_{trans\text{-}250kb}$ correlation, with "ns" indicating $p > 0.05$. The red line represents the fitted linear regression of the indirect CNV-PGS$_{trans\text{-}250kb}$ correlation against the total CNV-PGS$_{trans\text{-}250kb}$ correlation.

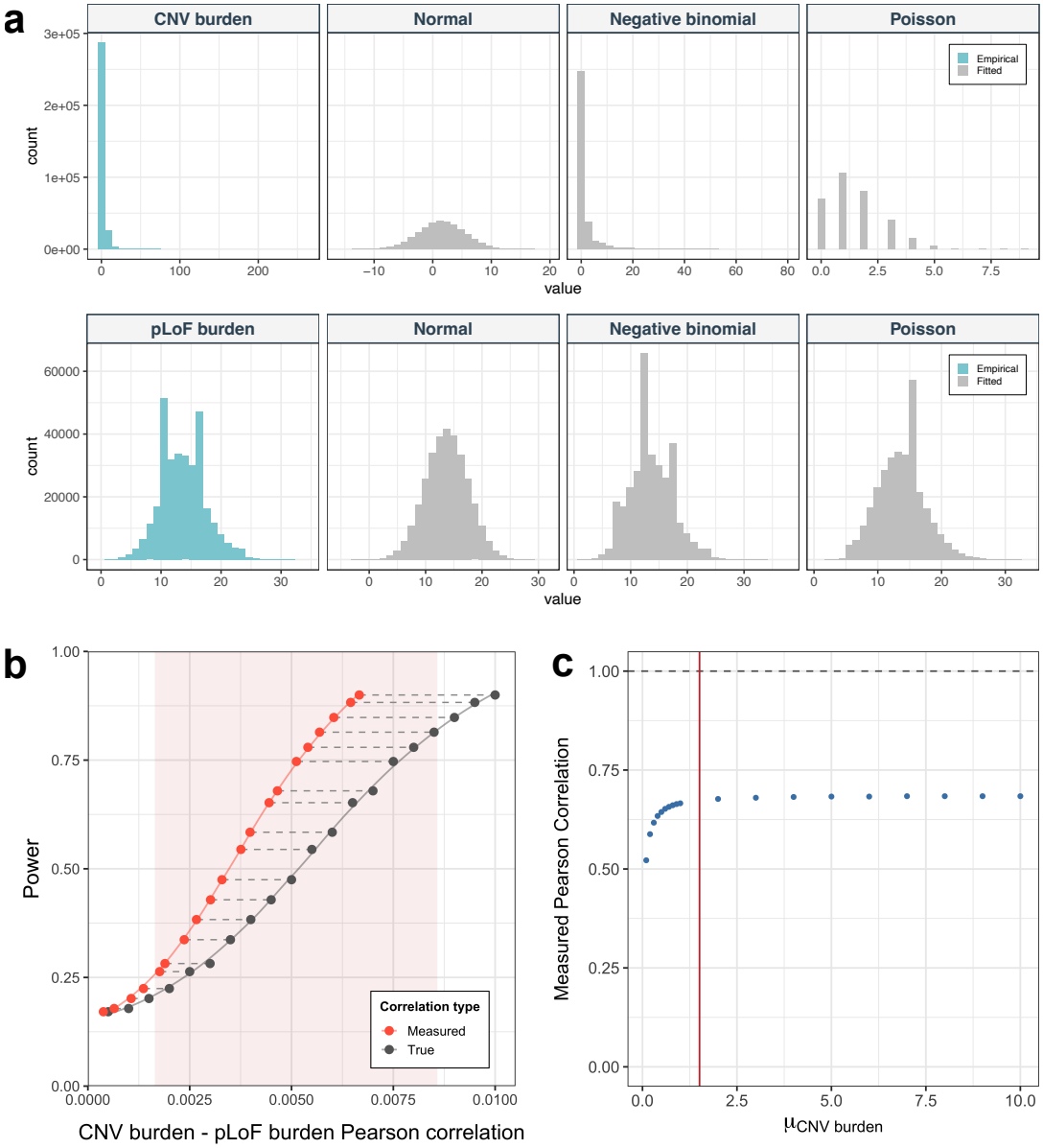


**Figure S11. Power to detect correlation between CNV burden and pLoF burden.**
**a,** Normal, negative binomial, and Poisson distributions (grey) were fitted to the observed (blue) CNV burden (top row) and pLoF burden (bottom row) by maximum likelihood parameter estimation. **b,** Power (y-axis) as a function of the Pearson correlation (x-axis) between simulated CNV burden and pLoF burden. True correlations are shown in grey and measured correlations in red, with dashed lines connecting corresponding values. CNV burden and pLoF burden were generated from a bivariate distribution with a given correlation $r$ (i.e., "true correlation") and rank-transformed to match empirical negative binomial and normal distributions, respectively. The shaded region corresponds to the 95% confidence interval centered on the empirically observed correlation in our data ($r$ = 0.005). **c,** CNV burden and pLoF burden were generated from a bivariate normal distribution with fixed correlation $r = 1$ and rank-transformed to approximate negative binomial and normal distributions with means $\mu_{CNV}$ and $\mu_{pLoF}$, respectively. Correlation strength was measured across a grid of $\mu_{CNV}$ values, with remaining parameters $size_{CNV}$ and $\sigma_{pLoF}$ held constant at their empirical values. The red vertical line depicts the empirical $\mu_{CNV}$ = 1.51.

# Supplementary Notes

## Supplementary Note 1

### Theoretical limits of the CNV-PGS$_{GW}$ correlation

Given the following causal structure between CNV, PGS$_{GW}$, and phenotype (Fig. N1.1):

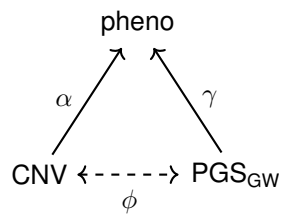


**Figure N1.1. Assumed CNV-PGS$_{GW}$ relationship via their influence on the phenotype.**
Directed acyclic graph (DAG) in which the solid arrows depict the causal effects of a copy-number variant (CNV) and the genome-wide polygenic score (PGS$_{GW}$) on phenotype (pheno) for a single individual, represented by correlations $\alpha = r(CNV, pheno)$ and $\gamma = r(PGS_{GW}, pheno)$, respectively. The bidirectional dashed arrow denotes the observed correlation $\phi = r(CNV, PGS_{GW})$ between CNV and PGS$_{GW}$.

Since correlation matrices are positive semi-definite (PSD, i.e., they are symmetric and all principal minors are non-negative), we can obtain an expression for the boundaries of the correlation coefficients. Consider the associated correlation matrix:

$$\begin{array}{c} \\ pheno \\ CNV \\ PGS \end{array} \begin{array}{c} \begin{array}{ccc} pheno & CNV & PGS \end{array} \\ \begin{pmatrix} 1 & \alpha & \gamma \\ \alpha & 1 & \phi \\ \gamma & \phi & 1 \end{pmatrix} \end{array}$$

where $\alpha$, $\gamma$, and $\phi$ are bounded between [-1,1]. Since all eigenvalues must be $\geq 0$, the determinant is also $\geq 0$:

$$1 - \phi^2 - \alpha(\alpha - \phi\gamma) + \gamma(\alpha\phi - \gamma) \geq 0$$

$$1 + 2\alpha\gamma\phi - (\alpha^2 + \gamma^2 + \phi^2) \geq 0$$

Solving the inequality for $\phi$ gives:

$$\alpha\gamma - \sqrt{(\alpha\gamma)^2 + 1 - \alpha^2 - \gamma^2} \leq \phi \leq \alpha\gamma + \sqrt{(\alpha\gamma)^2 + 1 - \alpha^2 - \gamma^2}$$

This relationship, as established by [125], imposes a lower and upper bound for $\phi$, when both PGS$_{GW}$ and CNV impact the trait of a single individual.

We next extend this logic to an assortative mating scenario, where CNV affects the phenotype of one partner and $PGS_{GW}$ the phenotype of the other, as depicted in Fig. N1.2:

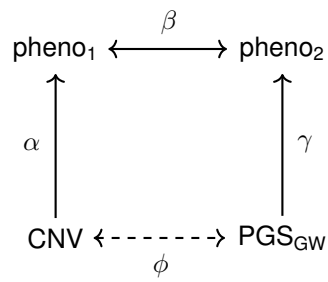


**Figure N1.2. Assumed CNV-$PGS_{GW}$ relationship via assortative mating.**
DAG in which the solid arrows depict the causal effects of copy-number variants (CNV) and the genome-wide polygenic score ($PGS_{GW}$) on the phenotypes of the two partners ($pheno_1$ and $pheno_2$), represented by correlations $\alpha = r(CNV, pheno_1)$ and $\gamma = r(PGS_{GW}, pheno_2)$, respectively. The solid bidirectional arrow between the phenotypes indicates the within-couple, same-trait, phenotypic correlation, $\beta$. The dashed bidirectional arrow denotes the observed correlation $\phi = r(CNV, PGS_{GW})$.

Assuming that CNV-$pheno_2$ and $PGS_{GW}$-$pheno_1$ have no direct relationship, only through $pheno_1$ and $pheno_2$, their correlation can be approximated using path-tracing rules, leading to the correlation matrix $A$:

$$\begin{array}{c} \\ pheno1 \\ pheno2 \\ CNV \\ PGS \end{array} \begin{array}{c} \begin{array}{cccc} pheno1 & pheno2 & CNV & PGS \end{array} \\ \begin{pmatrix} 1 & \beta & \alpha & \gamma\beta \\ \beta & 1 & \alpha\beta & \gamma \\ \alpha & \alpha\beta & 1 & \phi \\ \gamma\beta & \gamma & \phi & 1 \end{pmatrix} \end{array}$$

Computing all 15 principal minors $M = \det(A_I)$, where $A_I$ is generated from $A$ by selecting the rows and columns in each subset $I \subseteq \{1, 2, 3, 4\}$ (Table N1), we can once again obtain boundaries for $\phi$ from the restriction that $M_{10}$, $M_{13}$, $M_{14}$ and $M_{15}$ should all be $\geq 0$.

We thus obtain the following bounds for the total CNV-$PGS_{GW}$ correlation ($\phi$):

$$-1 \leq \phi \leq 1$$

$$\alpha\beta\gamma - \sqrt{(\alpha\beta\gamma)^2 + 1 - \alpha^2 - \gamma^2\beta^2} \leq \phi \leq \alpha\beta\gamma + \sqrt{(\alpha\beta\gamma)^2 + 1 - \alpha^2 - \gamma^2\beta^2}$$

$$\alpha\beta\gamma - \sqrt{(\alpha\beta\gamma)^2 + 1 - \alpha^2\beta^2 - \gamma^2} \leq \phi \leq \alpha\beta\gamma + \sqrt{(\alpha\beta\gamma)^2 + 1 - \alpha^2\beta^2 - \gamma^2}$$

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a} \leq \phi \leq \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

Where $a = \beta^2 - 1$, $b = 2\alpha\beta\gamma(1-\beta^2)$, and $c = 1 - (\alpha^2 + \beta^2 + \gamma^2) + (\alpha^2\beta^2 + \alpha^2\gamma^2 + \beta^2\gamma^2) - 2\alpha^2\beta^2\gamma^2 + \alpha^2\beta^4\gamma^2$.

The most strict bounds are given by: $\phi \in [max(L_{10}, L_{13}, L_{14}, L_{15}), min(U_{10}, U_{13}, U_{14}, U_{15})]$ where $L_i$ and $U_i$ correspond to the the lower and upper bounds obtained from principal minor $M_i$. We can then assess whether the correlation between the indirect and total CNV-PGS$_{\text{GW}}$ correlation is expected within this mathematical constraint.

**Table N1.** Principal minors of the $4 \times 4$ correlation matrix A

| **Minor** | **Subset** $I$ | **Principal Minor** $M = \det(A_I)$ |
|---|---|---|
| $M_1$ | $\{1\}$ | $1$ |
| $M_2$ | $\{2\}$ | $1$ |
| $M_3$ | $\{3\}$ | $1$ |
| $M_4$ | $\{4\}$ | $1$ |
| $M_5$ | $\{1,2\}$ | $1-\beta^2$ |
| $M_6$ | $\{1,3\}$ | $1-\alpha^2$ |
| $M_7$ | $\{1,4\}$ | $1-\gamma^2\beta^2$ |
| $M_8$ | $\{2,3\}$ | $1-\alpha^2\beta^2$ |
| $M_9$ | $\{2,4\}$ | $1-\gamma^2$ |
| $M_{10}$ | $\{3,4\}$ | $1-\phi^2$ |
| $M_{11}$ | $\{1,2,3\}$ | $1-\alpha^2-\beta^2+\alpha^2\beta^2$ |
| $M_{12}$ | $\{1,2,4\}$ | $1-\gamma^2-\beta^2+\gamma^2\beta^2$ |
| $M_{13}$ | $\{1,3,4\}$ | $1-\alpha^2-\phi^2-\gamma^2\beta^2+2\alpha\beta\gamma\phi$ |
| $M_{14}$ | $\{2,3,4\}$ | $1-\gamma^2-\phi^2-\alpha^2\beta^2+2\alpha\beta\gamma\phi$ |
| $M_{15}$ | $\{1,2,3,4\}$ | $1-(\alpha^2+\beta^2+\gamma^2+\phi^2)$ $+\alpha^2\beta^2+\alpha^2\gamma^2+\beta^2\gamma^2+\beta^2\phi^2$ $+2\alpha\beta\gamma\phi-2\alpha^2\beta^2\gamma^2-2\alpha\beta^3\gamma\phi$ $+\alpha^2\beta^4\gamma^2$ |

As shown in Fig. N1.3a, the predicted range for $r(CNV, PGS_{GW})$ for each trait is broad, with the narrowest interval across the 119 assessed CNV-trait pairs spanning -0.79 to 0.78. Similar minimum ranges are observed for the 21 phenotypes affected by the CNV burden (Fig. N1.3b). This supports that the observed correlation between indirect and total paths is not trivially dictated by PSD bounds alone.

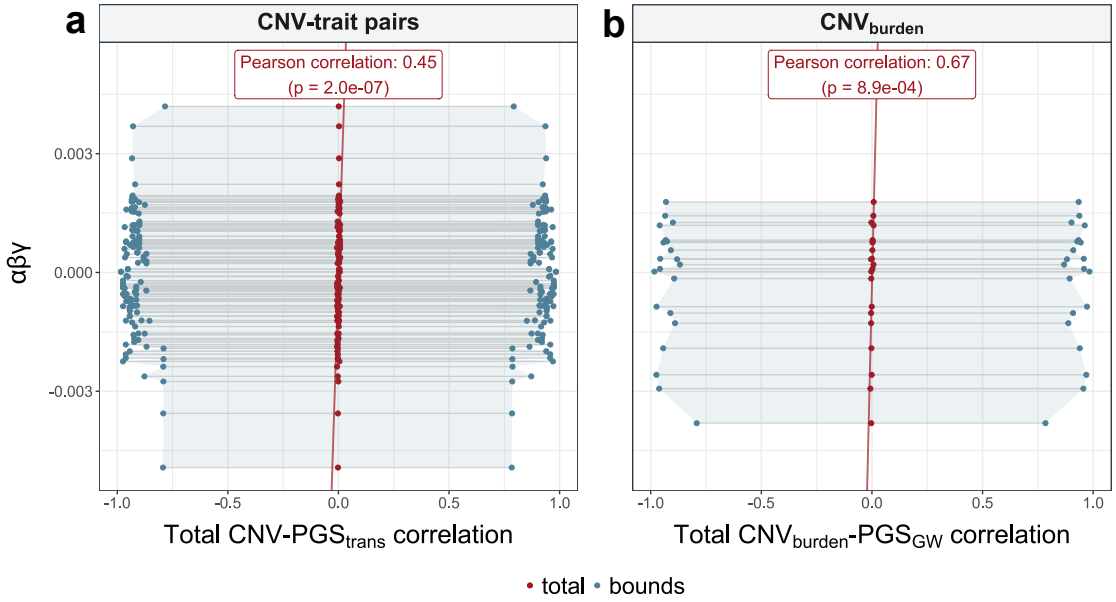


**Figure N1.3. Theoretical bounds for the CNV-PGS$_{GW}$ correlation.**
**a,** $\alpha \times \beta \times \gamma$ (y-axis) against the total observed CNV-PGS$_{trans}$ correlation (x-axis) are shown as red dots. Lower ($max(L_{10}, L_{13}, L_{14}, L_{15})$) and upper ($min(U_{10}, U_{13}, U_{14}, U_{15})$) bounds for total CNV-PGS$_{GW}$ correlation are shown in blue. Line segments connect values corresponding to each of the 119 CNV-trait pairs. The Pearson correlation coefficient between $\alpha \times \beta \times \gamma$ and the total CNV-PGS$_{trans}$ correlation is displayed. **b,** same analysis as in **a,** using the CNV burden instead of individual CNV-trait pairs, and PGS$_{GW}$ instead of PGS$_{trans}$. Each connected segment corresponds to one of the 21 phenotypes significantly affected by CNV burden.

## Supplementary Note 2

### Analytical approximation of the CNV-PGS correlation.

We define the random variables $X$ and $Y$ as the phenotypes of each partner with $C$ representing the CNV carrier status of the first partner, and $G$ the $\text{PGS}_{\text{GW}}$ of the second partner (Fig. N2).

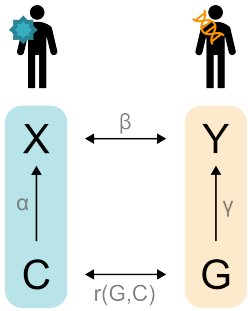


**Figure N2. Schematic representation of assortative mating modeling.**
$X$ and $Y$ represent the phenotypic values for the same trait among two partners, with $X$ being influenced by the first individual's CNV (C; correlation between the variables is denoted $\alpha$) and $Y$ being influenced by the second individual's PGS ($G$; correlation between the variables is denoted $\gamma$). The phenotypic correlation strength between $X$ and $Y$ is denoted $\beta$.

Each variable is approximated by a standard normal distribution:

$$X \sim \mathcal{N}(0,1)$$

$$Y \sim \mathcal{N}(0,1)$$

$$C \sim \mathcal{N}(0,1)$$

$$G \sim \mathcal{N}(0,1)$$

We then assume a causal model in which the CNV directly influences the phenotype of one partner ($C \rightarrow X$) and the $\text{PGS}_{\text{GW}}$ directly influences the phenotype of the other partner ($G \rightarrow Y$) so that:

$$X = \alpha \times C + \varepsilon_X, \text{ where } \varepsilon_X \sim \mathcal{N}(0, 1-\alpha^2)$$

$$Y = \gamma \times G + \varepsilon_Y, \text{ where } \varepsilon_Y \sim \mathcal{N}(0, 1-\gamma^2)$$

$$X \mid C = c \sim \mathcal{N}(\alpha c, 1-\alpha^2)$$

$$Y \mid G = g \sim \mathcal{N}(\gamma g, 1-\gamma^2)$$

Partners' phenotypes $X$ and $Y$ are correlated due to assortative mating, such that $r(X,Y) = \beta$. To

compute the joint probability density function (PDF) of $C$ and $G$, we apply the law of total probability by conditioning on the couple's phenotypes and integrating over all possible values of $X$ and $Y$. This yields:

$$f_{CG}(c,g) = \iint f_{CG|XY}(c,g \mid x,y) f_{XY}(x,y)\, dx\, dy$$

Since $C$ and $G$ are conditionally independent, the joint conditional probability can be factorized as:

$$f_{CG}(c,g) = \iint f_{C|XY}(c \mid x,y) f_{G|XY}(g \mid x,y) f_{XY}(x,y)\, dx\, dy$$

Given that $C$ is independent of $Y$, and $G$ is independent of $X$, we can further simplify the expression:

$$f_{CG}(c,g) = \iint f_{C|X}(c \mid x) f_{G|Y}(g \mid y) f_{XY}(x,y)\, dx\, dy$$

Using Bayes' theorem to express the PDFs $f_{C|X}(c \mid x)$ and $f_{G|Y}(g \mid y)$ in terms of their corresponding inverse forms, we obtain:

$$f_{CG}(c,g) = \iint f_{X|C}(x \mid c) \frac{f_C(c)}{f_X(x)} f_{Y|G}(y \mid g) \frac{f_G(g)}{f_Y(y)} f_{XY}(x,y)\, dx\, dy$$

Finally, under the assumption that $f_{XY}$ is bivariate normal:

$$(X,Y) \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \beta \\ \beta & 1 \end{bmatrix} \right)$$

The integral becomes:

$$f_{CG}(c,g) = \iint \frac{1}{\sqrt{2\pi(1-\alpha^2)}} e^{\frac{-(x-\alpha c)^2}{2(1-\alpha^2)}} \frac{\frac{1}{\sqrt{2\pi}} e^{-c^2/2}}{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}} \frac{1}{\sqrt{2\pi(1-\gamma^2)}} e^{\frac{-(y-\gamma g)^2}{2(1-\gamma^2)}} \frac{\frac{1}{\sqrt{2\pi}} e^{-g^2/2}}{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}} \frac{1}{2\pi} \frac{1}{\sqrt{1-\beta^2}} e^{\frac{-(x^2+y^2-2\beta xy)}{2(1-\beta^2)}}\, dx\, dy$$

$$f_{CG}(c,g) = \iint \frac{1}{(2\pi)^2} \frac{1}{\sqrt{1-\alpha^2}} \frac{1}{\sqrt{1-\gamma^2}} \frac{1}{\sqrt{1-\beta^2}} e^{-(c^2+g^2)/2} e^{(x^2+y^2)/2} e^{\frac{-(x-\alpha c)^2}{2(1-\alpha^2)}} e^{\frac{-(y-\gamma g)^2}{2(1-\gamma^2)}} e^{\frac{-(x^2+y^2-2\beta xy)}{2(1-\beta^2)}}\, dx\, dy$$

Symbolic integration with Maple yields the following closed-form solution for the PDF of $G$ and $C$:

$$f_{CG}(c,g) = \frac{e^{\frac{-(c^2-2\alpha\beta\gamma cg+g^2)}{2(1-(\alpha\beta\gamma)^2)}}}{2\pi\sqrt{1-(\alpha^2\beta^2\gamma^2)}}$$

This expression corresponds to the joint PDF of two random variables that follow a bivariate normal distribution. The general form of such distribution is given by:

$$f_{XY}(x,y) = \frac{e^{\frac{-(x^2-2\rho xy+y^2)}{2(1-\rho^2)}}}{2\pi\sqrt{1-\rho^2}}$$

Where $\rho$ denotes the correlation between $X$ and $Y$. Comparing the two expressions, it follows that under the specified assumptions:

$$r(C,G) = \alpha \times \beta \times \gamma$$

## Supplementary Note 3

**Total vs indirect CNV–PGS$_{GW}$ correlations in mating simulations.**

To assess whether $\alpha \times \beta \times \gamma$ reliably predicts the CNV-PGS$_{GW}$ correlation under assortative mating, we performed several forward simulations. In each of the 100 runs, 10,000 individuals with CNV-determined phenotype selected a partner from a pool of 500 candidates, each with a PGS$_{GW}$-determined phenotype. The matching criterion was to pair individuals whose phenotypic correlation was the closest to a fixed value $\beta$, which defined the strength of assortative mating. We ran four replicates for $\beta = 0$ (no phenotypic correlation), $\beta = 0.25$ (moderate correlation), $\beta = 0.5$ (strong correlation), and $\beta = 1$ (perfect correlation). We computed both the predicted, indirect ($\alpha \times \beta \times \gamma$) and empirical, total ($r(CNV, PGS_{GW})$) correlations for four different values of $\beta$, across a grid of $\alpha$ and $\gamma$ values. For each coordinate, the total correlation was averaged across simulation replicates. Finally, we fit the simulation results to non-linear models (specifically, rotated hyperbolic paraboloids) using the `nlsLM` function from the `minpack.lm` R package, except for $\beta = 0$ results, which were fitted with a linear model, using the `lm` function. Figure N3 shows the predictions of these models for a larger grid of $\alpha$ and $\gamma$ values. Simulations showed that $\alpha \times \beta \times \gamma$ predicts well the CNV-PGS$_{GW}$ correlation for each value of $\beta$. Total and indirect correlations exhibit the largest positive correlation when both $\alpha$ and $\gamma$ are simultaneously 1 or -1. Conversely, they are maximally anticorrelated when they present opposite signs (total = 1, indirect = -1, or vice versa). We can also observe that, in the absence of assortment, we expect no CNV-PGS$_{GW}$ correlation, as correctly predicted by the indirect path.

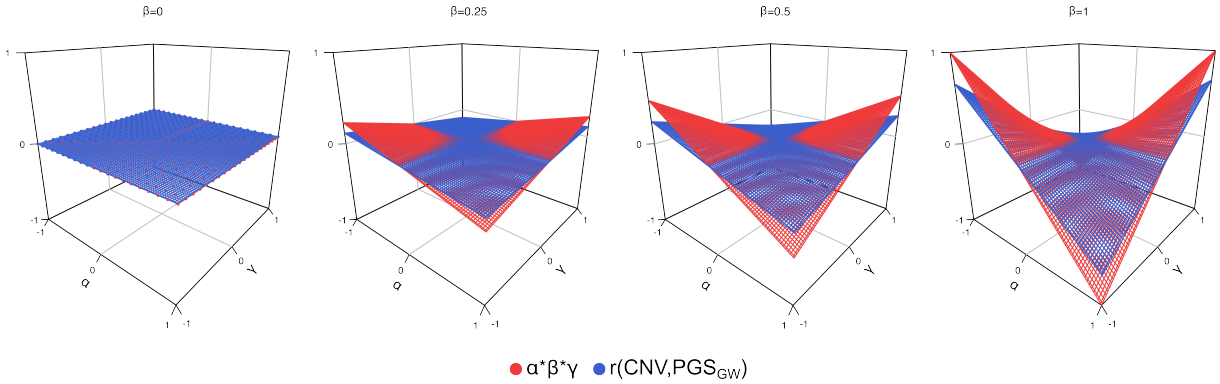


**Figure N3. Simulated total versus indirect CNV-PGS$_{GW}$ under various degrees of assortative mating.**
Mean of the empirical, total CNV–PGS$_{GW}$ correlations (red) and corresponding $\alpha \times \beta \times \gamma$ predicted indirect correlation (blue) (z-axis) across four levels of assortative mating strength ($\beta = 0, 0.25, 0.5$, and $1$) plotted over a range of $\alpha$ (x-axis) and $\gamma$ (y-axis) values.