The impact of rare pathogenic CNVs is exacerbated by 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

September 8, 2025

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) for which PGSs were evaluated. Assessing the individual, joint, and synergistic contribution of CNVs and PGS, we identified a significant additive effect for 45 (38%) CNV-trait pairs but no evidence for interactions. 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 the positive CNV-PGS association. We found strong agreement between this correlation and the one predicted by assortment ($r = 0.45$, $p = 3.9 \times 10^{-7}$). Similar trends of positive correlation were observed between PGS and genome-wide burden of CNVs or rare loss-of-function variants. Our results suggest that PGSs contribute to the variable expressivity of CNVs and rare variants, and improve the identification of those 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, neurodevelopmental and psychiatric disorders, epilepsy, congenital anomalies, and metabolic diseases [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]. 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 [8–26]. Associations recovered from population studies often parallel findings from clinical cohorts – but with a lower penetrance and severity – while also revealing novel associations, showcasing the pleiotropy of these CNVs [27–31].

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 [32]. 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 [31].

Early evidence of the contribution of additional genetic factors to disease liability stems from family-based studies [33–38]. Yet, often, detailed information on family members 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 [39]. 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 [40]. 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 [41] or a high burden of damaging single nucleotide variants (SNVs) [42]. 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 [26, 43–55]. 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 [50, 56], 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 [48]. 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 [26, 43, 44, 46, 48, 49, 53]. 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 [18], 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 [57] (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).

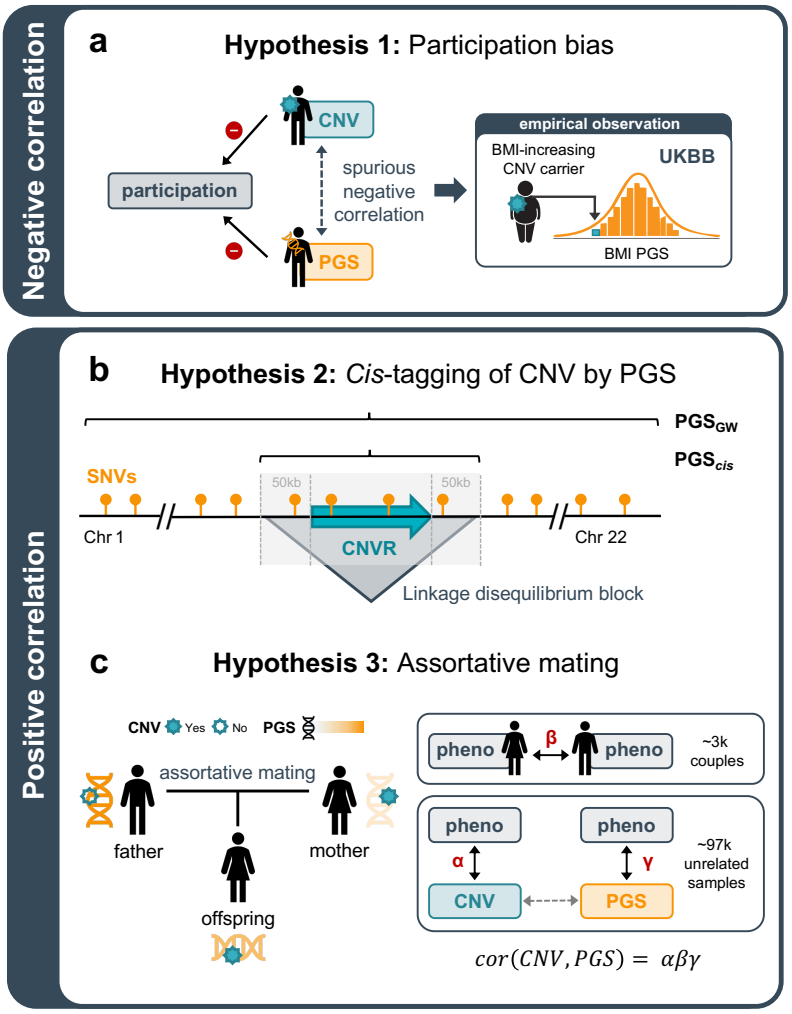

Figure 1. Mechanisms of CNV-PGS correlation.
Schematic representation of three possible mechanisms explaining correlation between an individual's CNV carrier status and polygenic score (PGS). **a,** Negative CNV-PGS correlation is expected in the presence of selective UK Biobank (UKBB) participation. 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 [57], 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 a PGS_{cis} that is restricted to single nucleotide variants (SNVs) mapping to the CNV region (CNVR) $\pm$ 50 kb. **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 (α), the within-couple phenotypic correlation (β), and the PGS-trait correlation (γ).

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 (UKBB field identifier #23149) are accessible for all participants, along with extensive phenotypic assessment [58, 59]. 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 [18]. 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 [18]. 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 selection

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 the right and left hand grip strength (UKBB field identifiers #47 and #46, respectively). Waist-to-hip ratio (WHR)

was defined as the ratio between waist and hip circumference (UKBB field identifiers #48 and #49, respectively); WHR adjusted for body mass index (WHRadjBMI), by regressing WHR on BMI (UKBB field identifiers #21001). To develop PGS scoring files, phenotypes were rank-transformed (`rntransform` from the `GenABEL` R package) among 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 (UKBB field identifier #21003), age^2, genotyping batch, and principal components (PCs) 1-40 [18]. 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 [60] for all UKBB participants [18]. Briefly, each CNV call was attributed a quality score ranging between -1 (likely deletion) and 1 (likely duplication) [61]. 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, 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 [18], 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, 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). Individuals carrying a high-confidence CNV

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 ($\leq$ 3rd 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 test (20%) sets with matching age and sex distributions. Individuals in the test set were pruned to unrelated individuals and merged with pre-selected CNV carriers to form the complete test set. Non-carriers in the test 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 [62] was used to filter genotype microarray data, excluding variants with minor allele frequency (MAF) $<$ 1%, missingness rate $>$ 10% (`--geno 0.1`), or Hardy-Weinberg p $<$ 1×10^{-15}. Samples with $>$ 10% missing genotype calls (`--mind 0.1`) were excluded. GWASs were carried out with Regenie v3.2.9 [63], 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 in step 2, using default parameters. 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` function) 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)` [64–66]. Genome-wide PGSs (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 the `pgsc_calc` v1.3.2 workflow (output: `SUM`) [67]. Phenotypic variance explained by PGS_{GW} was computed as the squared Pearson correlation between PGS_{GW} and adjusted phenotypes (Y, see Phe-

notype selection) as $r(Y, PGS_{GW})^2$.

Rare pLoF burden calculation

A subset of 319,161 unrelated ($>$ 3rd degree) "white British" UKBB participants [24] with available exome sequencing data were used to compute the predicted loss-of-function variants (pLoF) burden. High-confidence pLoF carriers were identified using LOFTEE variant annotations [68] and PLINK v2.00a3.1 [62] was used to recode the alleles, ensuring that the reference allele corresponds to the major allele (`plink2 --maj-ref 'force'`). Variants 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 the Bonferroni correction ($p \leq 0.05/119 = 4.2 \times 10^{-4}$). We also 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) 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 [18]. 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% confidence intervals (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 gamma-glutamyl transferase (GGT) values ($>$ 50 U/L) among carriers of a 22q11.23 duplication, as compared to copy-neutral individuals, was evaluated with a 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), 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 ($p \leq 0.05/(43 \times 2) = 5.8 \times 10^{-4}$).

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). 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), enrichment for discordant effect size direction between the CNV's impact on the adjusted phenotype (either individual CNV region or burden metric) and its impact on PGS_{GW} was estimated with a one-sided binomial test with a hypothesized probability of concordance of 0.5.

Cis-tagging of CNVs by PGS

For each CNV-trait pair, a local PGS (PGS_{cis}) using only SNVs mapping to the CNV region $\pm$ 50 kb, was calculated using a similar approach as for PGS_{GW} (see Polygenic scores calculation). We determined how CNV status predicts PGS_{cis} (see PGS_{GW} distribution among CNV carriers).

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 (M1) or CNV carrier status and PGS_{GW} (M2). 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 β and SE are the effect size and standard error of the effect of the CNV on the adjusted phenotype estimated through M1 or M2, respectively; r is the correlation between 5,000 bootstrap estimates of β_{M1} and β_{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” (UKBB field identifier #6141) [69]. 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 ($p \leq 0.05/43^2 = 2.7 \times 10^{-5}$). Correlation coefficients were re-computed for 1,000 random couple assignments in which each female was paired with a random male without replacement. 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.

Assortative mating modeling

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

$$r(CNV, PGS_{GW})_{indirect} = \alpha \times \beta \times \gamma \ , \tag{2}$$

where α is the Pearson correlation between the adjusted phenotype and CNV carrier status encoded

by the most significant CNV-GWAS model, β is the within-couple Pearson correlation of adjusted phenotypic values, and γ is the Pearson correlation between the adjusted phenotype and PGS_{GW} (Fig. 1c). Coefficients were derived from the test sample set, after excluding CNV-trait pairs for which the driver of positive correlation was local linkage disequilibrium and not assortative mating. This was done by comparing the magnitude of the CNV-PGS_{cis} Pearson correlation to the maximum correlation that can be induced by assortative mating if partner phenotypic correlation is perfect (β = 1). The latter was calculated as the product of the Pearson correlation between CNV status and adjusted phenotypes (α) and the Pearson correlation between PGS_{cis} and adjusted phenotypes (γ for local PGS_{cis}).

Confidence intervals for the total CNV-PGS_{GW} correlation were computed through Fisher's z-transformation. SE of α, β, and γ 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:

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

where 95% confidence intervals were then obtained from propagated errors.

Significance threshold

Significance of the indirect path through assortative mating was assessed by comparing the total-indirect Pearson correlation to a set of null distributions derived from 50,000 permutations. The latter were obtained by shuffling (with a 50% probability of sign flipping) the observed values of either one of α, β, or γ, while maintaining the empirical values for the two other coefficients. 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

First, to assess whether sample overlap could induce a pseudo-correlation between indirect and total CNV-PGS_{GW} (even in the absence of assortative mating), we removed samples that are part of the couples used to estimate β from the computation of α, γ, and total CNV-PGS_{GW} correlation.

Second, we demonstrate that the observed CNV–PGS_{GW} correlation does not reflect a necessary statistical consequence by deriving analytical bounds based on the requirement that the full correlation

matrix (comprising phenotype, CNV, and PGS_{GW}) is positive semi-definite. These bounds define the range of allowed $r(PGS_{GW})$ values, which if narrow, would suggest that the observed correlation is expected under the specified assumptions (Supplementary Note 1). We then derived an analytical formula and conducted simulations showing that under the assortative mating model assumptions, the CNV-PGS_{GW} correlation corresponds to the product of $\alpha \times \beta \times \gamma$ (Supplementary Notes 2 and 3).

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

CNV inheritance rate estimation

Sample selection

We used KING [70] to identify related UKBB participants ($\leq$ 3rd degree; kinship $>$ 0.044). Among the 67,958 pairs of relatives (103,726 individuals) in which at least one individual carried a high-confidence CNV, 12,831 pairs (21,216 individuals) harbored a CNV that overlaps ($\geq$ 1 bp) one of the 27 assessed CNV regions (see CNV carrier identification & CNV burden calculation).

Identity-by-descent estimation

For each CNV carrier with a close relative, we estimated the probability that the region harboring the CNV is identity-by-descent (IBD) with its relative, using THORIN [71, 72]. Namely, average IBD probability IBD_{avg} within a 1 Mb window around the CNV was computed by summing per-SNV IBD probability on each haplotype. We classified the IBD status of the CNV carrier and its relative as either $IBD0$ (i.e., no shared haplotypes; $IBD_{avg} < 0.25$), $IBD1$ (i.e., one shared haplotype; $0.75 < IBD_{avg} < 1.25$), $IBD2$ (i.e., two shared haplotypes; $IBD_{avg} > 1.75$). IBD status was set to unknown for all other probabilities ($]0.25, 0.75[\cup]1.25, 1.75[$).

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 lower bound for the 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.

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 UKBB [18] (Supplementary Table 1). The 119 signals involved 43 quantitative phenotypes describing a broad range of physiological functions, including musculoskeletal (n = 5), metabolic (11), neuropsychiatric (2), cardiopulmonary (4), hepatic, (7), hematologic (8), endocrine (2), and renal (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 tested 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 and duplication in isolation, respectively. White British UKBB participants carrying a CNV overlapping any of the lead probes were identified and CNVs were encoded according to the model yielding the most significant association. 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). Phenotypic variance explained by PGS_{GW} ranged between 5-25%, with two traits showing estimates over 25% – height (38%) and total bilirubin (27%) – and neuroticism having an estimated explained variance of only 3% (Supplementary Fig.1; Supplementary Table 2).

PGS_{GW} drives variable phenotypic expressivity among CNV carriers

We jointly modeled the phenotypic impact of CNV 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 (total nominally significant signals: 63%), distributed across all eight trait categories (Supplementary Fig. 2; Supplementary Table 4). Importantly, 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 and 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 affecting multiple physiological systems, with highly heterogeneous phenotypic presentation [30, 31]. 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 (β_{DEL} = -8.2 cm; $p_{DEL} = 1.0 \times 10^{-31}$; among deletion carriers: β_{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: β_{PGS} = 4.8 kg/m^2; p_{PGS} = 0.044; Fig. 2b), or platelet count (β_{Mirror} = -38.8 10^{-9} cells/L; p_{Mirror} =

330 2.5×10^{-21}; among CNV carriers β_{PGS} = 42.4 $\times$ 10^{-9} cells/L; p_{PGS} = 6.7 $\times$ 10^{-5}; Fig. 2c). We previously
331 showed that the effect of the CNV on platelet count was independent of the carrier's BMI and height
332 [30]. Here, we propose that the CNV effect on platelet count might be exacerbated by the carrier's
333 $PGS_{GW\text{-platelet count}}$, positioning PGS_{GW} as key determinants of the phenotype constellation expressed by
334 a given CNV carrier.

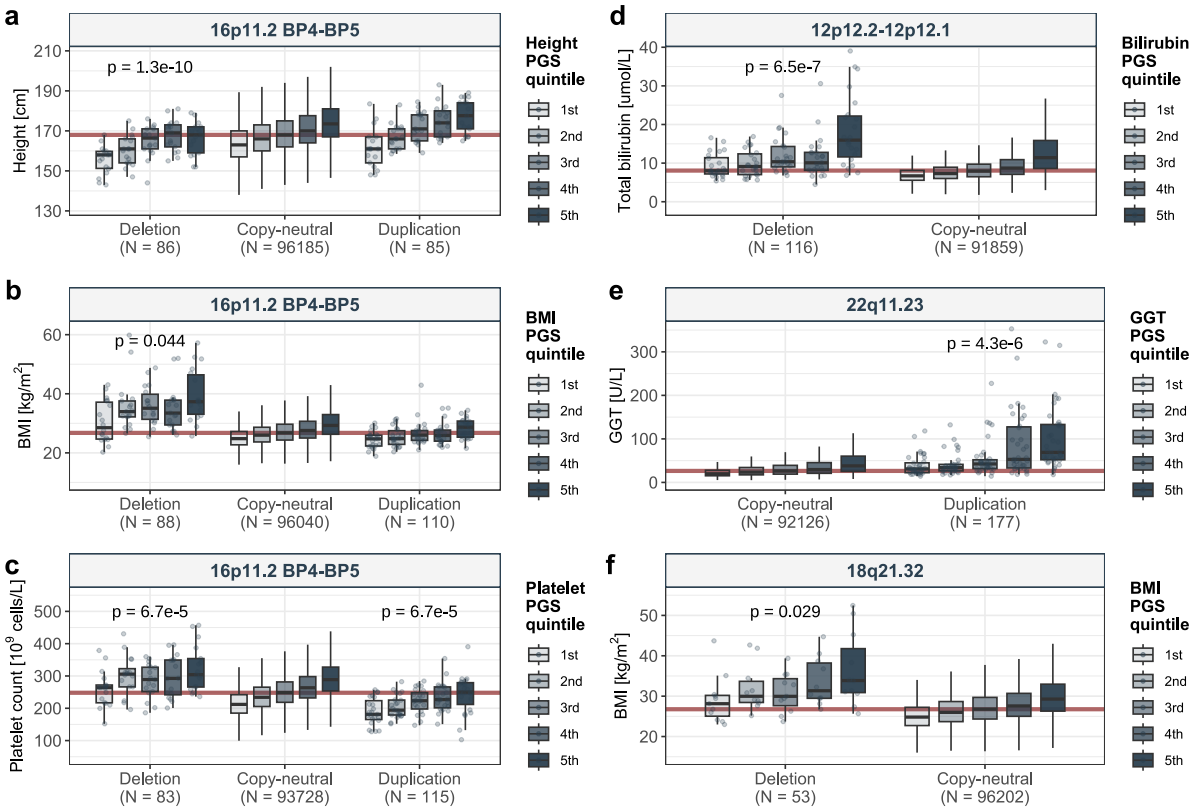

Figure 2. Additive contribution of PGS_{GW} and CNV to phenotypic variability.
Boxplots showing the phenotypic distribution (y-axis) of individuals according to their genome-wide polygenic score (PGS_{GW}) quintile for the specified trait and CNV carrier status (x-axis) for **a,** height and 16p11.2 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. P-values of the effect of PGS_{GW} on phenotypes among carriers of the trait-associated CNV are shown above the assessed CNV genotype groups, whose sample size is indicated as N. Only CNV genotype categories with $\geq$ 15 individuals are represented.

335 Other examples illustrate how laboratory measurements can reach extreme values in individuals
336 carrying CNVs in genes linked to the biomarker's transport or metabolism, if they present with a high
337 PGS_{GW}. This is the case for total bilirubin levels among carriers of a 12p12.2-12p12.1 deletion with a
338 high $PGS_{GW\text{-bilirubin}}$ (β_{DEL} = 3.0 umol/L; p_{DEL} = 5.9 x 10^{-15}; among deletion carriers: β_{PGS} = 7.1 umol/L;
339 p_{PGS} = 6.5 x 10^{-7}; Fig. 2d). The CNV region encompasses *SLCO1B1* and *SLCO1B3*, two genes

encoding hepatic bilirubin transporters [73] and for which mutational load has been described to form allelic series for bilirubin levels [17, 18]. Another example relates to gamma-glutamyltransferase (GGT) levels among carriers of a 22q11.23 duplication which also present with a high $PGS_{GW\text{-}GGT}$ (β_{DUP} = 34.7 U/L; p_{DUP} = 1.9 x 10^{-31}; among duplication carriers: β_{PGS} = 62.0 U/L; p_{PGS} = 4.3 x 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 x 10^{-15}), with a prevalence of 71% among duplication carriers in the top $PGS_{GW\text{-}GGT}$ quintile. Hence, integrating CNV and PGS information could guide 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 (β_{DEL} = 4.2 kg/m^2; p_{DEL} = 1.6 x 10^{-10}; among deletion carriers: β_{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 [74]. Similarly, the extent of muscle wasting among carriers of the 17p12 duplication associated with the degenerative neuropathy of Charcot-Marie-Tooth type 1A [75, 76] is nominally reduced in duplication carriers with a protective $PGS_{GW\text{-grip strength}}$ (β_{DUP} = -10.1 kg; p_{DUP} = 3.8 x 10^{-48}; among duplication carriers: β_{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 explains part of the variable expressivity of pathogenic CNVs across a broad spectrum of traits and likely contributes to phenotypic heterogeneity.

No evidence for CNV-PGS_{GW} interactions

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. We did not detect any CNV-PGS_{GW} interaction surviving Bonferroni correction ($p \leq 0.05/119 = 4.2 \times 10^{-4}$) (Supplementary Table 3). Failure to detect interac-

tions might stem from limited power, so 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. Again, we did not detect any Bonferroni-significant interaction ($p \leq 0.05/(2 \times 43) = 5.8 \times 10^{-4}$) (Supplementary Table 5). We conclude that there is no evidence to support that the polygenic background impacts complex traits differently among CNV carriers.

PGS_{GW} of CNV carriers reinforces the pathogenic impact of CNVs

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 ($p = 9.3 \times 10^{-9}$; Fig. 3a; Extended Data 2; Supplementary Table 6). Restricting to the 11 CNV-trait pairs with a nominally significant effect of the CNV carrier status on PGS_{GW}, ten had a directionally concordant effect ($p = 5.9 \times 10^{-3}$), indicating that CNV carriers tend to have a PGS_{GW} that exacerbates the CNV's phenotypic impact.

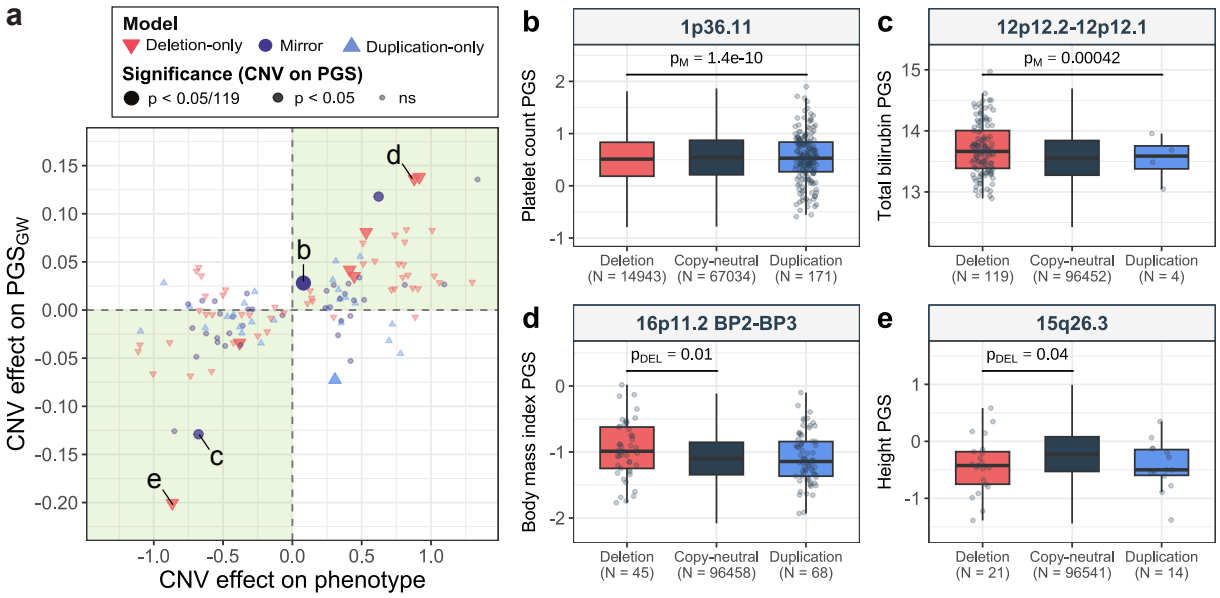

Figure 3. PGS_{GW} exacerbates the impact of rare CNVs.
a, Effect of CNV on genome-wide PGS (PGS_{GW}) (y-axis) against the effect on adjusted phenotypes (x-axis) for 119 CNV-trait pairs. Shape and color indicate the assessed model (i.e., model yielding the most significant CNV effect on the adjusted phenotype), while size and transparency reflect the significance of the CNV effect on the PGS_{GW}, with "ns" indicating $p > 0.05$. Data points in green areas correspond to CNV-trait pairs where carriers of a trait-associated CNV have a PGS_{GW} that reinforces 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 considered individuals is indicated as N. Individual data points are shown for CNV genotype categories with $N \leq 1{,}000$. P-values of the effect of the CNV status on PGS_{GW} encoded according to the mirror (M) or deletion-only (DEL) model are displayed.

381 Only the copy-number of the 1p36.11 region encompassing the *RHD* gene – which encodes the
382 Rhesus (Rh) factor antigen key in determining the Rh blood group – showed a Bonferroni-significant
383 positive effect ($p \leq 0.05/119 = 4.2 \times 10^{-4}$) on both platelet count and $PGS_{GW\text{-platelet count}}$ ($p = 1.4 \times$
384 10^{-10}; Fig. 3b). Other nominally significant examples include the association between the 12p12.2-
385 12p12.1 copy-number (*SLCO1B1* and *SLCO1B3*) and both decreased levels of total bilirubin and lower
386 $PGS_{GW\text{-bilirubin}}$ ($p = 4.2 \times 10^{-4}$; Fig. 3c). Similarly, higher $PGS_{GW\text{-BMI}}$ was observed among carriers of
387 the BMI-increasing 16p11.2 BP2-BP3 deletion ($p = 0.012$; Fig. 3d), which encompasses the obesity-
388 linked gene *SH2B1* [29, 77]. Finally, lower $PGS_{GW\text{-height}}$ was observed among carriers of a 15q26.3
389 partial deletion of *ADAMTS17* ($p = 0.042$; Fig. 3e). *ADAMTS17* is linked to autosomal recessive
390 Weill-Marchesani syndrome, which is characterized by short stature [78]. Despite individual effects
391 being weak, the observed data pattern appears conserved across a broad spectrum of CNV-trait pairs.
392 Importantly, it is incompatible with strong selective participation bias (Fig. 1a), which would induce a
393 *negative* (spurious) correlation between CNV and PGS_{GW} (Fig. 1a).

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

395 A first hypothesis explaining the positive CNV-PGS_{GW} correlation is tagging of the CNV by PGS_{GW}, so
396 that the presence of the CNV is positively correlated through linkage with SNVs included in PGS_{GW}
397 (Fig. 1b). To test this, we calculated a local PGS_{cis} based on SNVs mapping to the CNV region $\pm$
398 50 kb and fitted a linear regression model to estimate if the latter could be explained as a function of
399 the CNV carrier status (Supplementary Fig. 3a; Supplementary Table 6). Five CNVs were associated
400 with PGS_{cis} ($p \leq 0.05/119 = 4.2 \times 10^{-4}$), with a further 26 nominally significant effects. We next asked
401 if the observed *cis*-tagging by linkage could explain the positive correlation between CNV status and
402 PGS_{GW}. Among the ten CNV-trait pairs for which CNV carrier status had a nominally significant and
403 directionally concordant effect of both phenotype and PGS_{GW}, only three showed a nominally significant
404 effect on PGS_{cis} and only the 1p36.11 CNV (*RHD*) had a Bonferroni-significant effect on PGS_{GW} and
405 PGS_{cis} for platelet count (Supplementary Fig. 3b). Accordingly, the 1p36.11-platelet count was the only
406 signal for which the phenotypic impact of the CNV was significantly diminished upon conditioning on
407 PGS_{GW} (Supplementary Fig. 3c; Supplementary Table 3). With a frequency of 3.8%, the *RHD* deletion
408 is the only common CNV considered in our study. The region has been described as highly complex

by the human pangenome, which reports a frequency of 17% for the haplotype with a full *RHD* deletion [79], 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.

Assortative mating exacerbates the phenotypic impact of rare CNVs

Next, we explored assortative mating as an alternative mechanism to explain 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 excluded three CNV-trait pairs mapping to 1p36.11 for which the magnitude of the total CNV-PGS_{cis} correlation was much larger than the theoretical maximum observed under assortative mating (see Methods; Supplementary Fig. 4; Supplementary Table 7), suggesting that the positive CNV-PGS_{GW} correlation is primarily driven by linkage. We then examined correlations between adjusted phenotypes among the 3,396 UKBB couples present in our test set. Thirty-three (77%) same-trait comparisons showed significant positive correlations, with vitamin C (r = 0.48, p = 1.1×10^{-165}), BMI (r = 0.26, p = 7.8×10^{-53}), body fat mass (r = 0.25, p = 2.4×10^{-49}), weight (r = 0.23, p = 9.2×10^{-43}) and height (r = 0.21, p = 8.0×10^{-36}) showing the strongest correlation (Supplementary Fig. 5; Supplementary Table 8). 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 = 6.6×10^{-32}; Supplementary Fig. 6; Supplementary Table 9). Recomputing within-couple correlations under random pairing, 99.8% of the permutations did not exhibit any significant same-trait correlation. This confirms widespread phenotypic assortment across diverse traits in our sample, which could channel correlation between CNVs and PGS_{GW}.

We leveraged within-couple phenotypic correlation to estimate the fraction of the total CNV-PGS_{GW} correlation that can be explained through a path mediated by assortative mating (Fig. 1c). The latter is defined as the product of the CNV-phenotype correlation (α), the within-couple phenotypic correlation (β), and the PGS_{GW}-phenotype correlation (γ) (Fig. 1c). CNV-PGS correlations estimated via the assortative mating path strongly correlated with total CNV-PGS_{GW} correlation (r = 0.45, p = 3.9×10^{-7}) (Fig.

437 4a; Supplementary Table 10). Fifty-thousand permutations of α (p = 0), β (p = 0), or γ (p = 2.0 $\times$ 10^{-5})

438 only once (for γ) resulted in a total-indirect CNV-PGS_{GW} correlation larger than 0.45 (Fig. 4b).

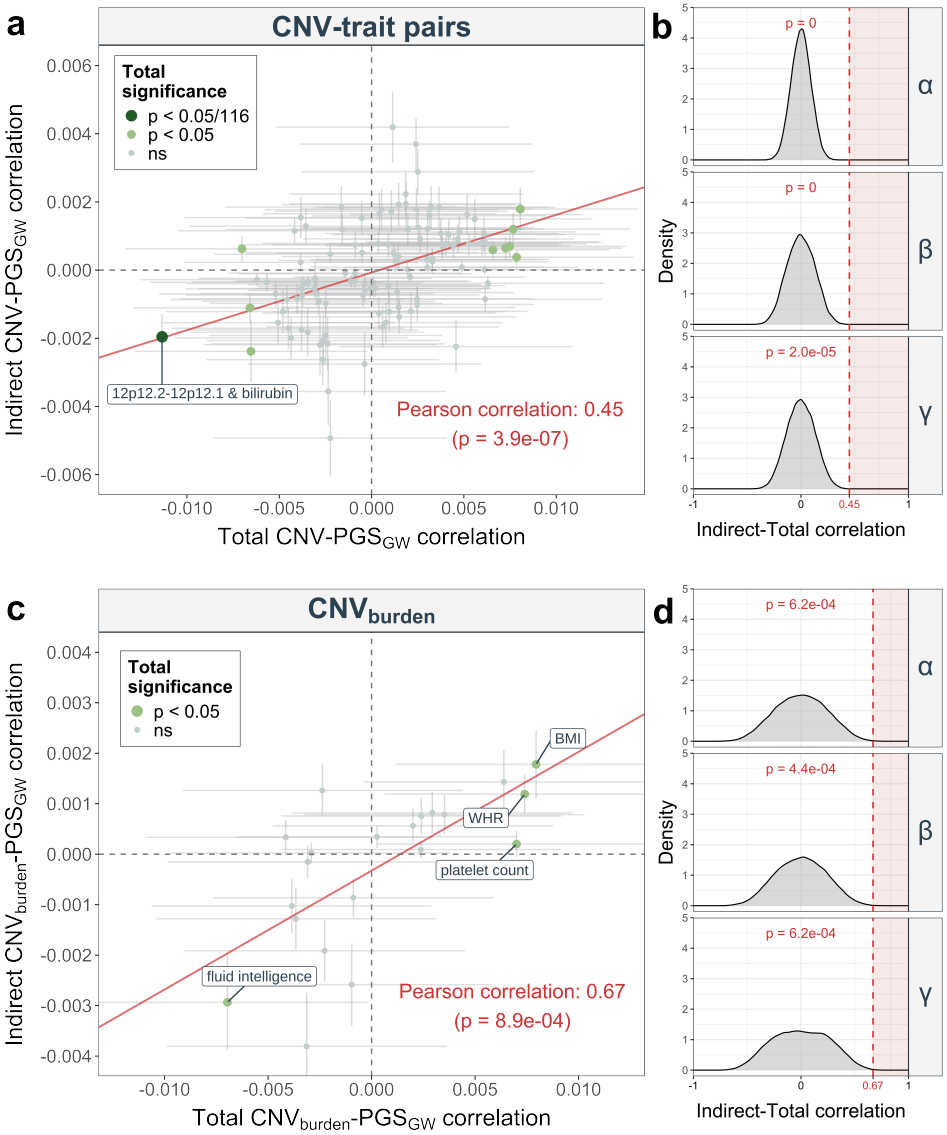

Figure 4. Assortative mating explains CNV-PGS_{GW} correlation.
a, Correlation between total (x-axis) and indirect (y-axis; through assortative mating) CNV-PGS_{GW} correlation for 116 CNV-trait pairs with no evidence of *cis*-tagging by PGS_{GW}. Error bars represent 95% confidence intervals. Size and color reflect the significance level of the total CNV-PGS_{GW} correlation, with "ns" indicating p $>$ 0.05. Bonferroni-significant effects are labeled. The red line represents the fitted linear regression of the indirect against the total CNV-PGS correlation. **b,** Density plots displaying the distribution of the total versus indirect CNV-PGS_{GW} correlations obtained from 500,000 permutations of the α (top), β (middle), and γ (bottom) coefficients across the 116 CNV-trait pairs. The empirical correlation and the associated p-value are denoted in red through a dashed vertical line. 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.

439 Repeating the analysis excluding couples used to calculate β to estimate α, γ, and the total CNV-PGS

440 correlation produced similar results (r = 0.41; p = 3.8 $\times$ 10^{-6}; Supplementary Table 11). Furthermore,

analytically derived boundaries for $r(CNV, PGS_{GW})$ demonstrated that the product of $\alpha \times \beta \times \gamma$ is not expected to correlate with the total CNV-PGS_{GW} correlation, 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). This indicates that our finding is unlikely to be driven by artifacts and that a substantial fraction of CNV-PGS_{GW} correlation can be explained by assortative mating.

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 12). 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

We next tested whether our findings could be generalized beyond our 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 13). 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 score ($p = 0.044$). In line with earlier findings, 18 out of 21 effects were directionally concordant with the effect of the CNV burden on the adjusted phenotype ($p = 7.4 \times 10^{-4}$), including the four nominally significant effects. This supports carriers of trait-associated CNVs being more prone to carry a phenotype-exacerbating PGS_{GW}. We investigated whether this pattern could be accounted for by assor-

tative mating by substituting the locus-specific CNV carrier status by the CNV burden in our assortative mating framework. Across the 21 assessed phenotypes, we observed a significant correlation between the total CNV burden-PGS_{GW} correlation and the path mediated by assortative mating (r = 0.67, p = 8.9 $\times$ 10^{-4}) (Fig. 4c; Supplementary Table 14). Permutation of α (p = 6.2 $\times$ 10^{-4}), β (p = 4.4 $\times$ 10^{-4}), and γ (p = 6.2 $\times$ 10^{-4}) supported the significance of our results (Fig. 4d), indicating that the observed phenomenon is not tied to specific genomic loci but reflects a general mechanism. Excluding couples used to calculate β did not substantially impact the results (r = 0.68, p = 7.3 $\times$ 10^{-4}; Supplementary Table 15).

Finally, 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 pLoF burden (p $\leq$ 0.05/43 = 1.2 $\times$ 10^{-3}; Supplementary Table 13). 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 13). Finally, we observed a small but significant correlation between the pLoF and CNV burden (r = 0.005; p = 3.9 $\times$ 10^{-3}), even though the rarity of both pLoF and CNV events prevents assessing this correlation in a phenotype-specific manner. Together, this provides early evidence for accumulation of genetic variants with coordinated phenotypic effects across mutational classes representing a widespread phenomenon within the general population that contributes to variable expressivity.

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 contributing to this relationship, supported by a non-null inheritance rate for a substantial fraction of the studied CNVs. A similar pattern 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 a better conceptual understanding of how common and

496 rare variants jointly influence complex traits. Studies have demonstrated that polygenic background
497 contributes to the variable expressivity of rare "monogenic" variants, allowing improved risk estima-
498 tion among these individuals [48, 80, 81]. Targeted studies focusing on specific phenotypes such
499 as autism [26, 45, 52], schizophrenia [26, 46, 50, 56], intellectual disability and developmental delay
500 [47, 54, 55, 82, 83], or obesity [84, 85] showed that overall, rare and common variants additively con-
501 tribute to the phenotype. Our results reinforce this body of evidence by extending these findings to a
502 varied set of quantitative traits, supporting that PGS_{GW} can stratify CNV carriers to identify those who are
503 at increased risk of developing clinically relevant associated comorbidities. Indeed, CNV carriers with a
504 PGS_{GW} counteracting the effect of the CNV exhibited phenotype values within the range of copy-neutral
505 individuals, while carriers with a phenotype-exacerbating PGS_{GW} tended to exhibit phenotypes outside
506 the normal range. This pertains to the relative contribution of CNVs versus PGS. It was reported that
507 the PGS for educational attainment has a broader impact on socio-economic outcomes than pathogenic
508 CNVs [86]. We argue that while PGSs might be more relevant at the population scale, the presence of
509 a CNV has important implications at the individual level and future studies should explore how to best
510 integrate CNV and PGS to optimize relevance in terms of personalized medicine.

511 Focusing on more complex CNV-PGS relations, we tested for non-additive effects but did not iden-
512 tify any surviving multiple-testing correction. Weak CNV-PGS interactions have been reported but most
513 studies testing for non-additivity between PGS and rare CNVs or pLoF variants failed to detect robust
514 interactions [26, 43, 44, 48, 53, 80, 81, 84, 85]. We conclude that if weak interactions exist, we are
515 currently underpowered to detect them, and they are likely of moderate clinical relevance. Still, theoreti-
516 cal work predicts that disease risk variants might exhibit stronger phenotypic consequences in high-risk
517 polygenic background due to the non-linear relationship between disease liability and disease probabil-
518 ity [87], emphasizing the importance of the choice of the link function for binary traits and the choice
519 of the scale/transformation for continuous outcomes. This iterates the key role of properly selecting the
520 polygenic background of the studied sample to accurately characterize rare variant pathogenicity.

521 While the effect of PGS_{GW} on the phenotype appeared mainly independent of CNV status, the CNV
522 status correlated positively with an individual's PGS_{GW}, paralleling a recent report showing that UKBB
523 carriers of CNVs causing neurodevelopmental conditions have lower PGS_{GW} for educational attainment

[54]. We found this pattern to be widespread across diverse phenotypes and identified assortative mating as a likely pervasive contributor. Positive phenotypic assortment is well-established for neuropsychiatric phenotypes, anthropometric features, and lifestyle factors [38, 69, 88–90], with recent biobank studies reporting partner phenotypic correlation among less conspicuous traits, e.g., blood measurements [69]. Studies of the downstream consequences of assortative mating on genomic architecture have emphasized how ignoring its impact can bias estimation of key genetic metrics [91–100], yet, the mechanisms through which it acts and the full scale of its implications remain underappreciated. Given assortative mating on a given trait, if a rare CNV contributes to that phenotype and that CNV is inherited – which we demonstrate is a plausible assumption – we expect compounding genetic factors in subsequent generations until the number of accumulated variants starts affecting fitness. This induces a positive correlation between CNV (or other rare variants) and PGS_{GW}, analogous to what was observed for PGSs on different chromosomes [92]. Previous work has shown that over generations, assortative mating correlates with co-occurrence of psychiatric conditions between parents and offspring [90] and leads to genetic anticipation, wherein disease liability increases over generations among families carrying rare LoF with variable expressivity [38]. Another study proposed parental assortment to explain the correlation between PGS and rare variant burden associated with cognitive and educational outcomes both between partners and within individuals [101]. Our data support this mechanism and suggest that it operates across a broader spectrum of variants and phenotypes.

Our assortative mating framework is targeted toward detecting direct assortment, which induces a correlation in genetic variants directly associated with the trait on which assortment takes place. This scenario is conceivable for traits such as height or BMI, but seems improbable for traits such as bilirubin 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 [69]. Acknowledging that mate selection is likely to be based on a large combination of traits, a partner-choice index has recently been proposed [102]. 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

552 whether PGSs for different CNV-associated traits are independent. Persistent cross-trait assortative
553 mating was shown to induce spurious genetic correlations [95], highlighting the importance of factoring
554 in both pleiotropy and polygenicity [103] – an extension that goes beyond the scope of this study.

555 Another consideration is sample ascertainment, as cohorts in which healthy participants are over-
556 (e.g., volunteer samples [57, 104, 105]) or under- (e.g., case-control studies [106, 107] or electronic
557 health record-linked biobanks with higher disease burden [108, 109]) represented may exhibit a negative
558 CNV-PGS correlation (Fig. 1a), as documented in previous clinical studies [43, 44, 46, 49]. In UKBB, we
559 observed a weak, yet consistent, positive CNV-PGS correlation. While healthy volunteer bias in UKBB
560 could have attenuated this association, it did not outweigh the positive correlation induced by assortative
561 mating. We conclude that while selective participation in volunteer biobanks may alter the magnitude of
562 observed effects and should therefore be acknowledged, correct inference with respect to directionality
563 of effects remains largely possible [57, 110]. Case-control studies, in contrast, are subject to extreme
564 participant selection, possibly canceling out any positive CNV-PGS correlation induced by assortative
565 mating. This suggests that genetic findings obtained from population biobanks are likely to be more
566 robust against selection bias than those obtained from clinical studies, highlighting the usefulness of the
567 former when interrogating the role of rare and common variants in disease susceptibility.

568 Our study should be interpreted in light of some limitations. The examined CNVs were detected
569 from genotyping microarrays and span large genomic regions, a corollary of which being that they have
570 strong phenotypic consequences and are rarer. While these characteristics were purposefully selected
571 for in our study, low CNV frequency limits the power of our analyses and influences its outcomes; we ex-
572 pect that investigating CNVs with higher frequencies and smaller effect sizes would identify more cases
573 in which *cis*-tagging represents a dominant mechanism. Indeed, SNV associations have been found to
574 tag stronger associations with smaller structural variants [21, 111–113]. Accordingly, the only common
575 CNV in our analyses – the *RHD* deletion – was subject to *cis*-tagging. Recent efforts in producing large
576 genome sequencing datasets coupled with the development of methods allowing robust detection of
577 structural variants across the size and allele frequency spectrum will allow exploration of the interplay
578 between different mutation classes at a scale [111, 114]. Furthermore, because PGS portability de-
579 creases with genetic distance [115], we restricted our study to a single genetic ancestry group with a

sufficiently large sample size to explore the interplay between PGS and rare CNVs. This stresses the importance of continued efforts to develop large and diverse biobanks with rich phenotypic data to enable replication studies. In the future, collection and integration of environmental factors is likely to reveal a more complex and nuanced picture of the mechanisms shaping variable expressivity. Importantly, because participation bias and assortative mating have opposing effects on the CNV-PGS correlation, we likely underestimate the role of assortative mating. Yet, our assortative mating model ignores potentially pleiotropic CNV and PGS effects acting on (further) off-target traits under assortment, an aspect that could be further explored in the future.

In conclusion, our study supports evidence that rare deleterious variants and common polygenic backgrounds additively shape diverse phenotypes and identifies assortative mating as a mechanism promoting the co-occurrence of risk-associated common and rare genetic contributions.

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 (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 1-6.

Supplementary Notes 1-3.

Supplementary Tables 1-15.

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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Mikhail Vysotskiy, Xue Zhong, Tyne W Miller-Fleming, Dan Zhou, Autism Working Group of the Psychiatric Genomics Consortiumˆ, Bipolar Disorder Working Group of the Psychiatric Genomics Consortiumˆ, Schizophrenia Working Group of the Psychiatric Genomics Consortiumˆ, 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

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

[65] 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`.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

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

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.

[115] 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.

[116] 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

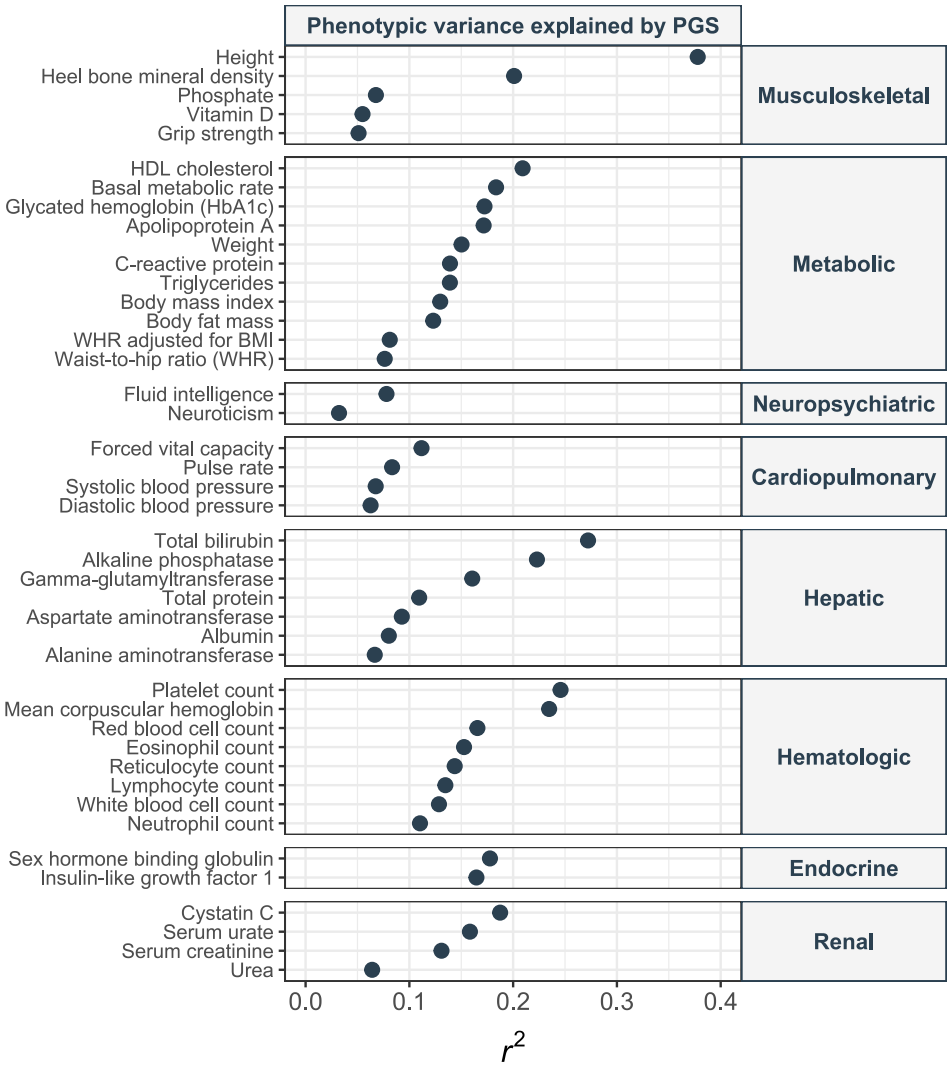

Figure S1. Phenotypic variance explained by PGS_{GW}.
Squared correlation between an individual's adjusted phenotypic values and PGS_{GW} (x-axis) for the 43 traits assessed in this study (y-axis). Phenotypes are ordered by decreasing explained phenotypic variance within each phenotype category.

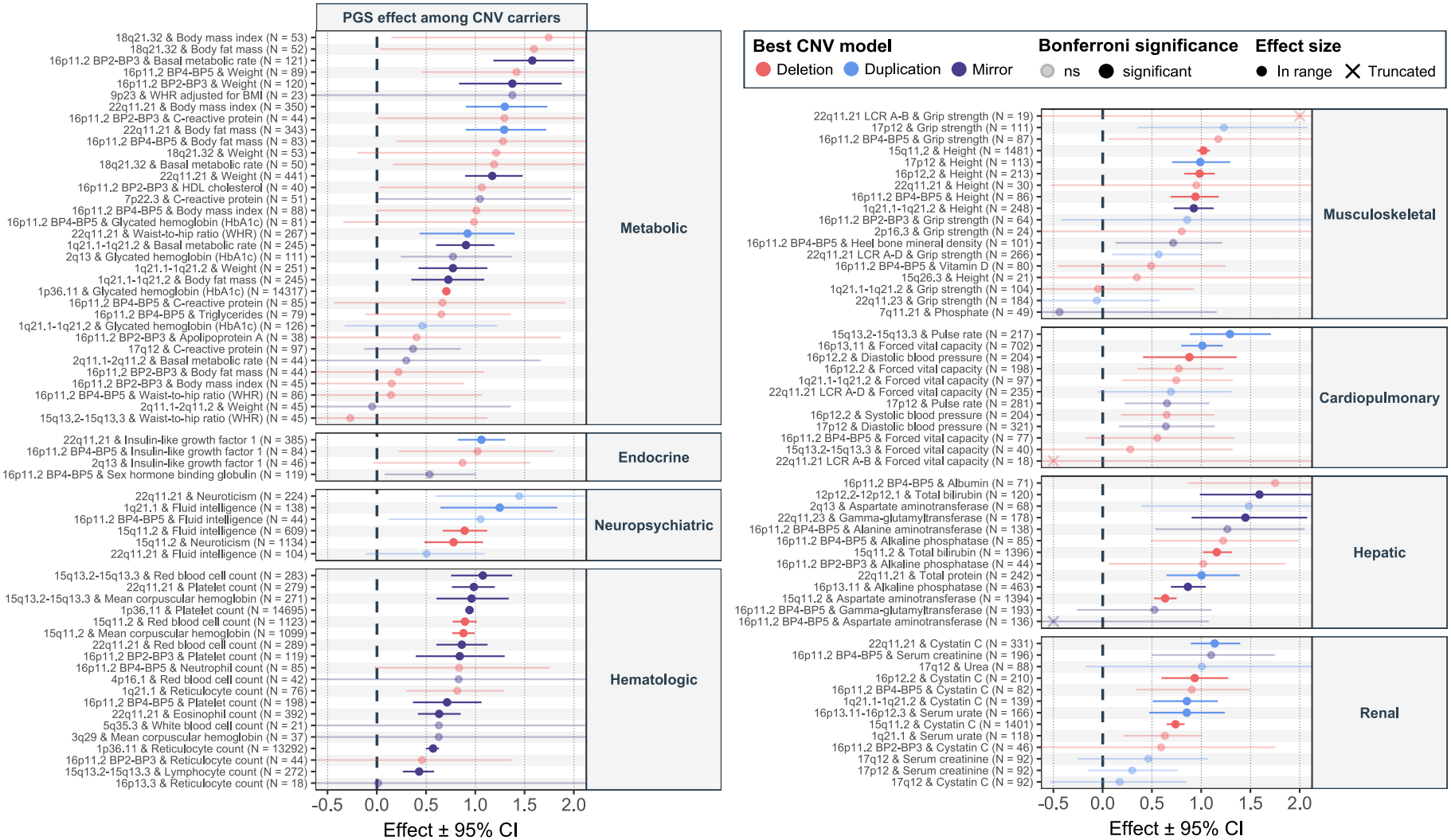

Figure S2. PGS_{GW} effect among CNV carriers.
Standardized effect sizes with 95% confidence intervals (CI) of the PGS_{GW} effect on the phenotype (x-axis) among CNV carriers for the 119 unique CNV-trait pairs assessed in this study (y-axis, left) ordered by decreasing effect size within each trait 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 [18]. The number of considered carriers is indicated as N. 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.

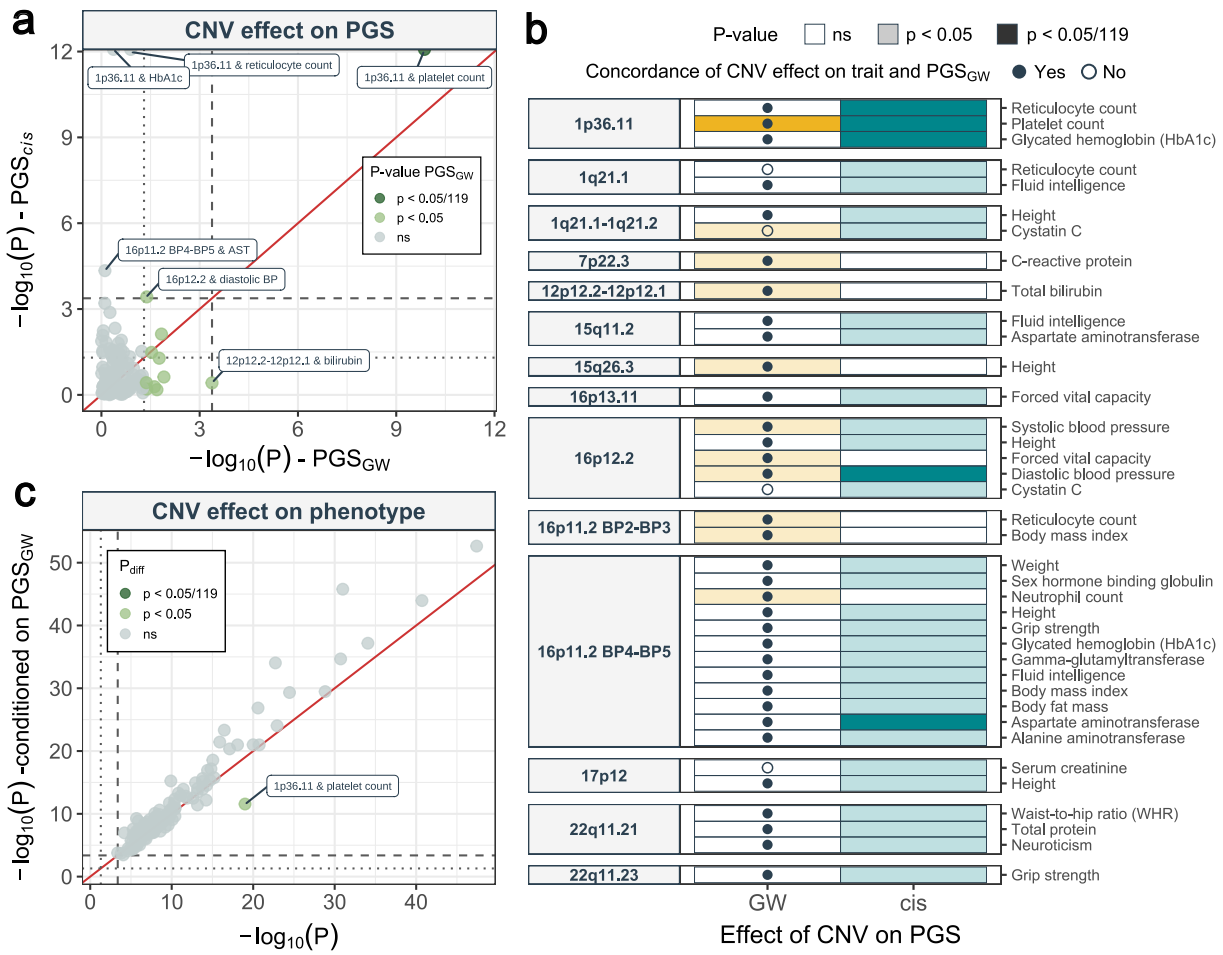

Figure S3. *Cis*-tagging of CNVs by PGS.

a, Negative logarithm of the p-value of the effect of the CNV on the local PGS (PGS_{cis}; y-axis) against the CNV effect on the genome-wide PGS (PGS_{GW}; x-axis). Data points are colored according to the significance of the CNV effect on PGS_{GW}, with "ns" indicating $p > 0.05$. 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. CNV-trait pairs with either CNV effects reaching Bonferroni significance are labeled, as well as a near significant effect of 12p12.2-12p12.1 CNVs on $PGS_{GW\text{-}bilirubin}$. AST = aspartate aminotransferase; diastolic BP = diastolic blood pressure; HbA1c = glycated hemoglobin. **b,** Significance level of the effect of the CNV (left y-axis, ordered by genomic position) on PGS_{GW} (orange) or PGS_{cis} (turquoise) (x-axis) for CNV-associated phenotypes (right y-axis). Transparency indicates whether the effect is non-significant (ns; white), nominally significant ($p < 0.05$, semi-transparent), or Bonferroni-significant ($p \leq 0.05/119 = 4.2 \times 10^{-4}$; opaque). Full or empty circle in the PGS_{GW} column indicates whether the effect of the CNV on the adjusted phenotype and PGS_{GW} for that phenotype are directionally concordant or not, respectively. **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.

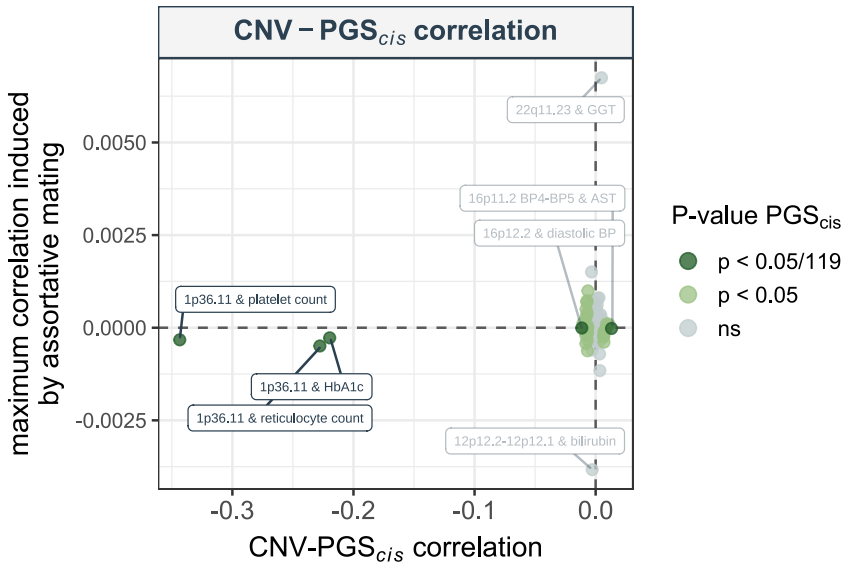

Figure S4. Maximum CNV-PGS$_{cis}$ correlation through assortative mating.
Pearson correlation between the CNV carrier status, encoded according to the model yielding the most significant CNV-trait association, and PGS$_{cis}$ (x-axis) against the product of the correlation between the CNV carrier status and adjusted phenotype (α) and PGS$_{cis}$ and the adjusted phenotype (γ for PGS$_{cis}$) (y-axis). The latter corresponds to the maximum correlation that can be induced by assortative mating, when β is 1. Data points are colored according to the significance of the CNV-PGS$_{cis}$ correlation, with "ns" indicating p $>$ 0.05. The five CNV-trait pairs with a Bonferroni-significant effect of the CNV carrier status on PGS$_{cis}$ are labeled, along with two other outliers. Dark grey labels are used for the three CNV-trait pairs for which the positive CNV-PGS$_{GW}$ correlation can primarily be explained by linkage disequilibrium and not assortative mating.

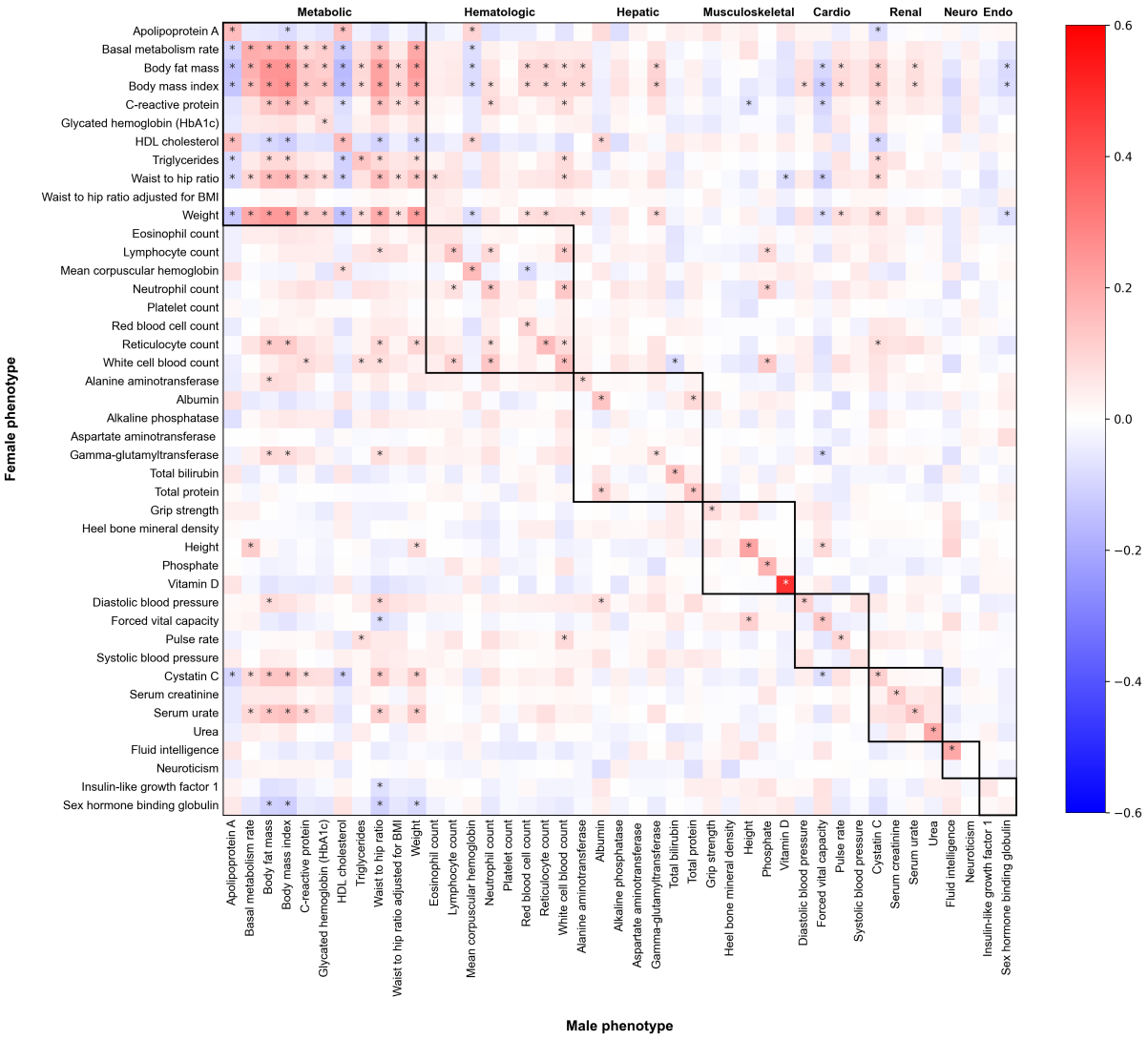

Figure S5. Pairwise partner phenotypic correlation.
Heatmap of Pearson correlation coefficients between 43 adjusted phenotypes in males (x-axis) versus the corresponding phenotype in their female partner (y-axis). Stars indicate Bonferroni-significant correlations ($p \leq 0.05/43^2 = 2.7 \times 10^{-5}$). Phenotypes are clustered by trait categories, as depicted by bolded squares and top labels. Cardio = Cardiometabolic; Neuro = Neuropsychiatric; Endo = Endocrine.

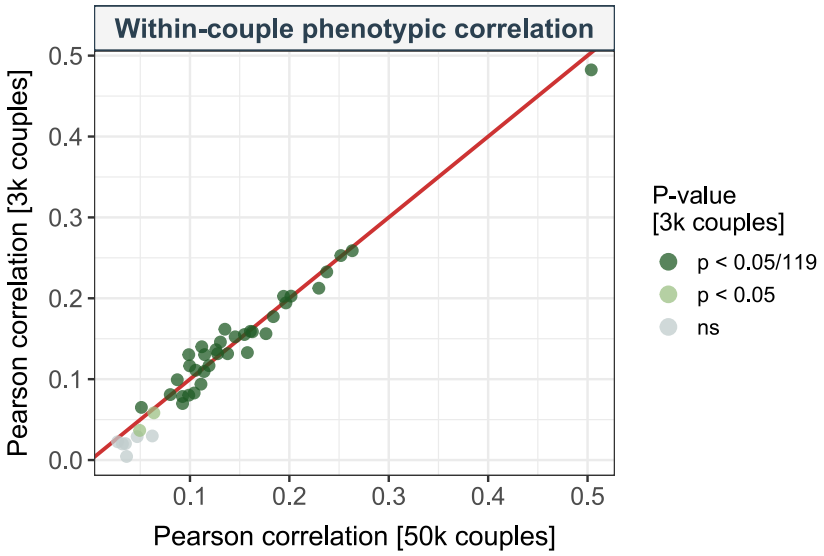

Figure S6. Impact of downsampling on same-trait within-couple phenotypic correlation.
Same-trait, within-couple Pearson phenotypic correlation among the 3,396 couples present in our testing subset (y-axis, "3k") against same-trait, within-couple Pearson phenotypic correlation among the 51,664 couples present in the entire "white British" subset of the UKBB (x-axis, "50k"). All phenotypes show Bonferroni-significant ($p \leq 0.05/119 = 4.2 \times 10^{-4}$) 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.

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):

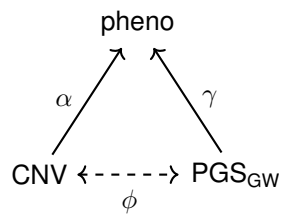

Figure N1.1. Assumed CNV-PGS$_{GW}$ relationship via their influence on the phenotype.
Directed acyclic graph (DAG) in which the solid arrows depict the causal effects of a copy-number variant (CNV) and the genome-wide polygenic score (PGS$_{GW}$) on phenotype (pheno) for a single individual, represented 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 (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 α, γ, and ϕ are bounded between [-1,1]. Since all eigenvalues must be ≥ 0, the determinant is also ≥ 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 ϕ 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 [116], imposes a lower and upper bound for ϕ, 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:

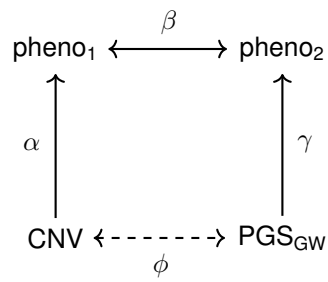

Figure N1.2. Assumed CNV-PGS_{GW} relationship via assortative mating.
DAG in which the solid arrows depict the causal effects of copy-number variants (CNV) and the genome-wide polygenic score (PGS_{GW}) on the phenotypes of the two partners ($pheno_1$ and $pheno_2$), represented by correlations $\alpha = r(CNV, pheno_1)$ and $\gamma = r(PGS_{GW}, pheno_2)$, respectively. The solid bidirectional arrow between the phenotypes indicates the within-couple, same-trait, phenotypic correlation, β. 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 ϕ from the restriction that M_{10}, M_{13}, M_{14} and M_{15} should all be ≥ 0.

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

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

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

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

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

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 4×4 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 116 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 a mere mathematical constraint.

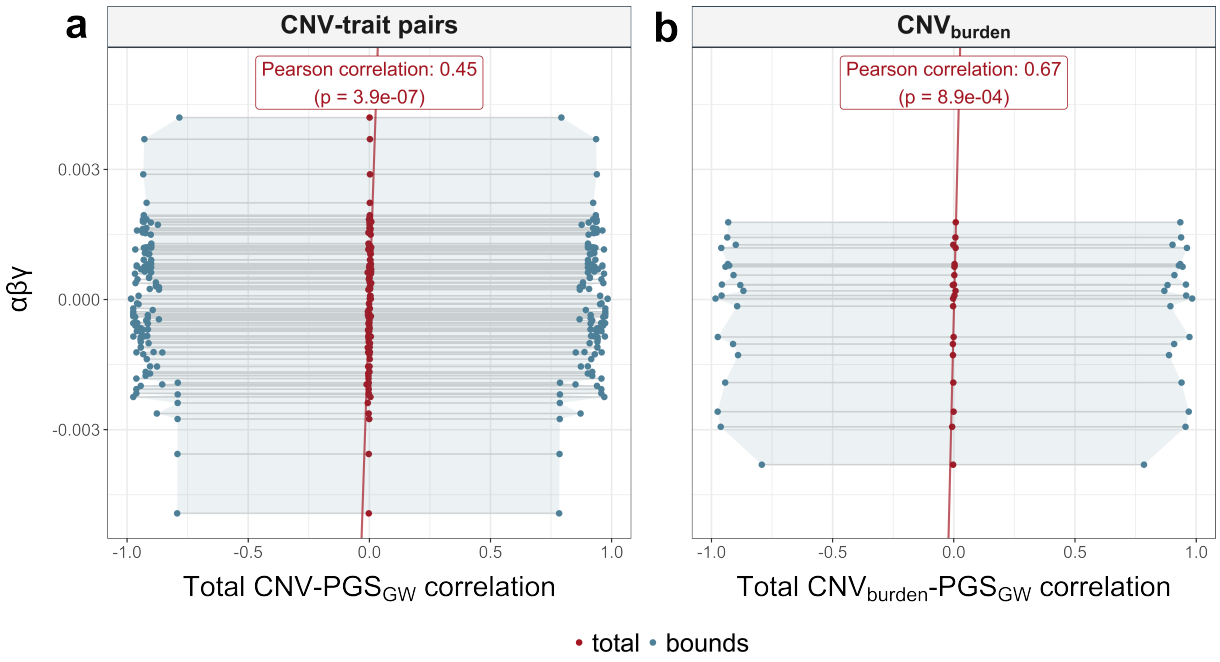

Figure N1.3. Theoretical bounds for the CNV-PGS$_{GW}$ correlation.
a, $\alpha \times \beta \times \gamma$ (y-axis) against the total observed CNV-PGS$_{GW}$ correlation (x-axis) are shown as red dots. Lower ($max(L_{10}, L_{13}, L_{14}, L_{15})$) and upper ($min(U_{10}, U_{13}, U_{14}, U_{15})$) bounds for total CNV-PGS$_{GW}$ correlation are shown in blue. Line segments connect values corresponding to each of the 116 CNV-trait pairs. The Pearson correlation coefficient between $\alpha \times \beta \times \gamma$ and the total CNV-PGS$_{GW}$ correlation is displayed. **b,** same analysis as in **a,** using the CNV burden instead of individual CNV-trait pairs. 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 PGS_{GW} of the second partner (Fig. N2).

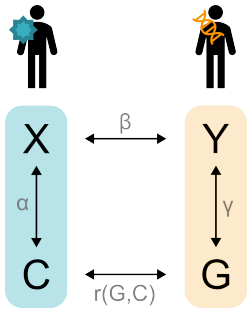

Figure N2. Schematic representation of assortative mating modeling.
X and Y represent the phenotypic values for the same trait among two partners, with X being influenced by the first individual's CNV (C; correlation between the variables is denoted α) and Y being influenced by the second individual's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

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 PGS_{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 ρ 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 β, 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 β, across a grid of α and γ 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 α and γ values. Simulations showed that $\alpha \times \beta \times \gamma$ predicts well the CNV-PGS$_{GW}$ correlation for each value of β. Total and indirect correlations exhibit the largest positive correlation when both α and γ 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.

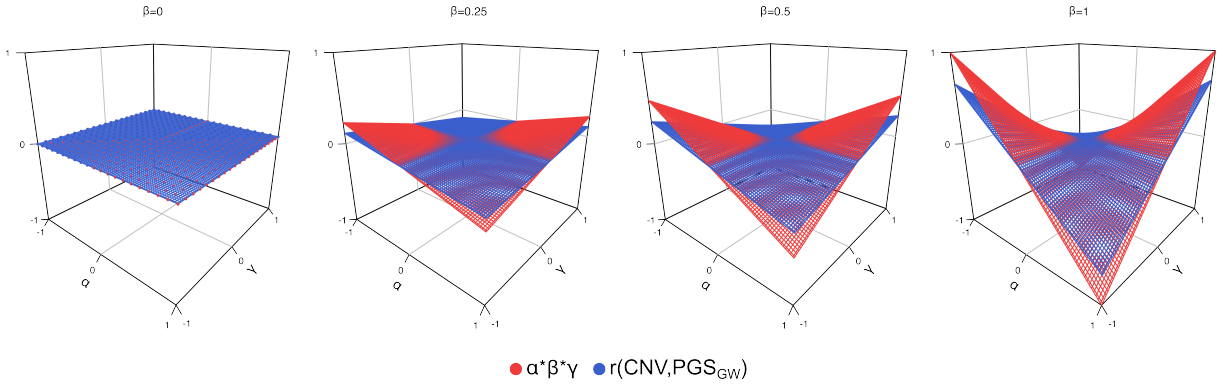

Figure N3. Simulated total versus indirect CNV-PGS$_{GW}$ under various degrees of assortative mating.
Mean of the empirical, total CNV–PGS$_{GW}$ correlations (red) and corresponding $\alpha \times \beta \times \gamma$ predicted indirect correlation (blue) (z-axis) across four levels of assortative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