

When do protein-truncating variants and copy-number loss agree? A pre-specified, power-aware genome-wide audit of cross-perturbation effect equivalence in 885,362 gene-trait pairs

Beiji Lu

DrLubeiji@outlook.com

Xuzhou Central Hospital, Affiliated Hospital of Southeast University; Xuzhou Medical University Clinical College <https://orcid.org/0009-0004-9876-1883>

Jian Cheng

Xuzhou Central Hospital

Meng Han

Xuzhou Central Hospital

Cao Ma

Xuzhou Central Hospital

Yunlong Zhao

Xuzhou Central Hospital

Research Article

Keywords: copy number variation, protein-truncating variants, equivalence testing, UK Biobank, negative result, drug target genetics

Posted Date: October 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11302876/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

When do protein-truncating variants and copy-number loss agree? A pre-specified, power-aware genome-wide audit of cross-perturbation effect equivalence in 885,362 gene-trait pairs

Beiji Lu, PhD^{1,2*}, Jian Cheng^{1,2}, Meng Han^{1,2}, Cao Ma^{1,2}, Yunlong Zhao^{1,2}

¹ Department of Orthopedics, Xuzhou Central Hospital, Affiliated Hospital of Southeast University, No. 99 Jiefang Road, Xuzhou, Jiangsu 221000, China; ² Xuzhou Medical University Clinical College, Xuzhou, Jiangsu 221000, China

* Correspondence: Beiji Lu, PhD, Department of Orthopedics, Xuzhou Central Hospital, No. 99 Jiefang Road, Xuzhou, Jiangsu Province 221000, China. Email: DrLubeiji@outlook.com. ORCID: 0009-0004-9876-1883.

Keywords: copy number variation; protein-truncating variants; equivalence testing; UK Biobank; negative result; drug target genetics

Abstract

Rare-variant association studies now estimate the phenotypic effects of the same gene from two largely independent perturbation types: protein-truncating SNV/indel burden (PTV) and copy-number loss (CNV-LoF). The portability assumption - that these effects are exchangeable - underlies dosage-sensitivity maps and the use of human genetics to prioritize drug targets, yet it is rarely audited with uncertainty-aware equivalence methods at genome-wide scale. We conducted a three-phase, gate-based audit using public UK Biobank summary statistics (Hujouel et al., 2024; CC BY 4.0) and an external AstraZeneca 500k whole-genome-sequencing PTV probe. Across 57 quantitative traits, 885,362 valid gene-trait pairs with separately estimated PTV and CNV-LoF effects were classified by two one-sided tests (margin 0.20 trait SD) into equivalent, heterogeneous, or information-insufficient, with global false-discovery control. After strict correction, 4,252 pairs were equivalent, 216 heterogeneous, and 880,894 information-insufficient. Only nine strict-equivalent pairs were active in the same direction in both modalities, decisively failing the pre-specified minimum of 200 required for any drug-portability label. In a gene-grouped out-of-fold diagnostic, precision features alone separated strict-heterogeneous from strict-equivalent rows almost perfectly (AUROC 1.000), adding CNV architecture features changed average precision by -0.002, and power-matched overlap contained far fewer than the required 50 rows per class. Heterogeneous-effect genes were nevertheless strongly enriched for multi-gene deletions (median 10 vs 2.5 protein-coding genes covered per loss event, $q = 5.2 \times 10^{-12}$), consistent with regional passenger or oligogenic effects. Cross-perturbation equivalence labels derived from public summary statistics are therefore essentially deterministic functions of statistical precision and cannot train a portability predictor under the current design; heterogeneity

signal is real but concentrated in complex CNV regions. We publish the full row-level classification as a genome-wide evidence inventory together with the failed gates, and specify the label designs required before any drug-portability claim can be tested.

Introduction

Gene dosage perturbations arrive in two chemically and structurally distinct forms in human populations. Protein-truncating SNV and indel burden tests aggregate rare nonsense and frameshift alleles into a per-gene loss-of-function estimate, whereas copy-number association tests estimate the phenotypic effect of DNA segments that delete or duplicate one or more genes. Large exome- and genome-wide efforts in the UK Biobank have now produced both kinds of estimate for the same genes on overlapping analytical frameworks, and both are used - often interchangeably - as evidence for drug target direction of effect, dosage sensitivity, and safety liabilities. Dosage-sensitivity maps such as pHaplo and pTriplo integrate multiple perturbation readouts into a single genome-wide prior, and the portability of genetic effect estimates across perturbation types is an implicit load-bearing assumption of this inference chain.

How exchangeable the two perturbation types actually are remains incompletely quantified. Hujoel et al. reported broad direction concordance between pLoF SNV/indel and pLoF CNV associations, while the genome-sequencing CNV PheWAS of Zou et al. reached the opposite corner of the same design space: large and multi-gene CNVs generate passenger and oligogenic associations that a focal gene-level interpretation would misattribute. These two observations are not contradictory, but public summary statistics have not, to our knowledge, been used to build a calibrated, uncertainty-aware genome-wide inventory of where PTV and CNV-LoF effects are statistically supported as

equivalent, where they are supported as genuinely different, and where the data are simply uninformative. Treating non-significance as equivalence is the central methodological trap in such an inventory; equivalence-testing logic, in which compatibility must be affirmatively demonstrated inside a pre-specified margin, avoids it. We therefore conducted a three-phase audit with pre-registered continuation gates. Phase 0 measured cross-perturbation coverage and concordance across 12 quantitative traits and froze the analysis direction. Phase 1 probed an external AstraZeneca UK Biobank 500k whole-genome PTV resource for the same traits under frozen gates. Phase 2 scaled the frozen equivalence classification to all 57 complete non-truncated quantitative-trait files and tested, under pre-specified grouped cross-validation gates, whether CNV architecture features can predict cross-perturbation non-portability beyond statistical precision. Here we report the audit in full, including the gates that failed. The intended drug-portability predictor direction did not pass its frozen continuation gates and we present this as a negative result, together with the row-level classification dataset, the structural associations that did replicate, and the label designs required for any defensible successor study.

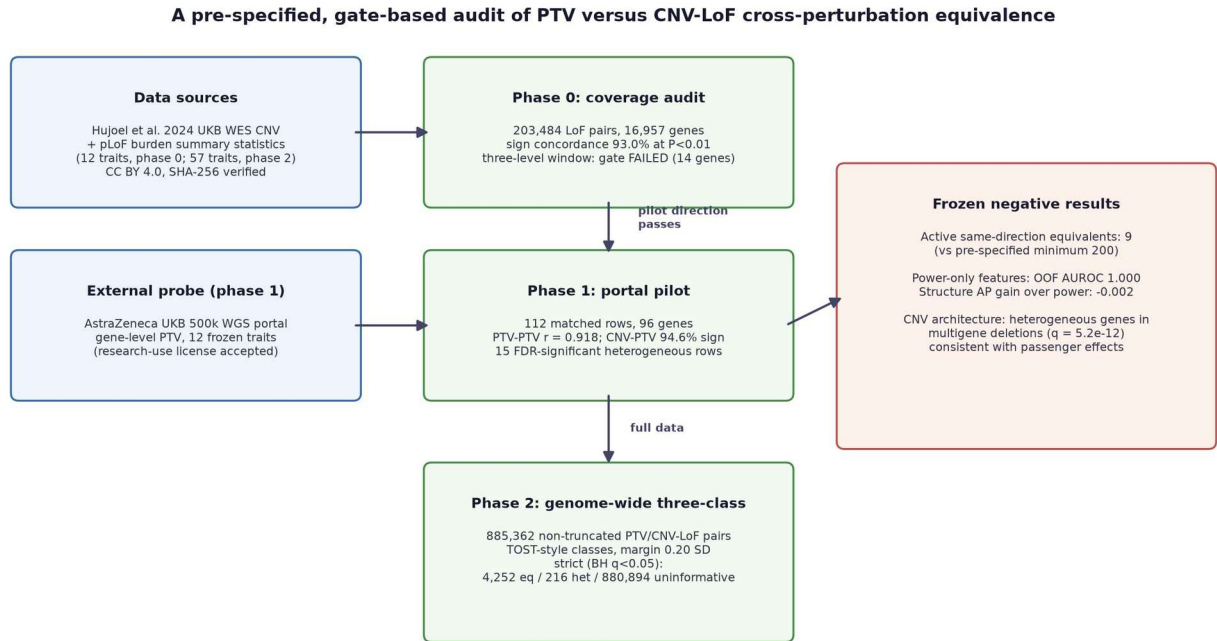


Figure 1. Study design and frozen gate outcomes. Phase 0 audited cross-perturbation coverage on 12 quantitative traits and failed the pre-specified three-level evidence gate; the analysis direction was frozen to PTV-versus-CNV-LoF equivalence. Phase 1 probed the external AstraZeneca 500k WGS PTV resource under frozen gates, which passed. Phase 2 classified 885,362 genome-wide pairs into equivalence, heterogeneity, or uninformative classes and failed every frozen prediction gate, invoking the stop line. Heterogeneous-effect genes were nonetheless strongly enriched for multi-gene deletions (right panel).

Methods

Study design and frozen gates

The audit was organized as three sequential phases, each with continuation and stop criteria fixed in writing before the next data stage began (frozen decision documents dated 2026-07-15 and 2026-07-16). Gates were quantitative: phase 0 required at least 100 independent genes with three-level (PTV, CNV-LoF, duplication) evidence at a per-modality threshold of $P \leq 0.01$ to justify the three-level therapeutic-window direction; phase 1 required at least 50 matched rows, 40 genes, and 75% sign agreement with the external PTV probe; phase 2 required, for any prediction model, at least 200 active same-

direction strict-equivalent rows, at least 50 power-matched rows per strict class, and an out-of-fold average-precision gain of at least +0.03 from CNV architecture features over a precision-only baseline. All gate outcomes below are reported as pass or fail exactly as pre-specified.

Data sources and integrity verification

Phase 0 used the supplementary tables and 12 quantitative-trait summary-statistic files of Hujoel et al. 2024 (Zenodo record 10.5281/zenodo.10529671; Broad Institute UKB WES CNV summary-statistic directory; CC BY 4.0). Phase 2 extended this to all 57 complete quantitative-trait files in the official Broad directory; every file passed remote-size checks, complete gzip decompression, and SHA-256 hashing, and approximately 1.8 GB of raw data were retained. The phase-1 external probe used the AstraZeneca Centre for Genomics Research public UK Biobank 500k whole-genome-sequencing resource (v2), gene-level PTV model, European ancestry, accessed through the public portal after explicit acceptance of the resource license (research use only, with required attribution to Wang et al. 2021 and Zou et al. 2026); portal responses are truncated at approximately $P < 0.1$, and missing rows were never treated as null effects or as negative labels. The structural annotation used the AZ UKB WGS CNV gene-annotation table (390,873 rows; 144,562 CNV events).

Effect-compatibility classification

For each gene-trait pair with valid, non-truncated PTV and CNV-LoF estimates, the primary contrast was the effect difference $\text{beta_CNV-LoF} - \text{beta_PTV}$, with standard error $\sqrt{\text{SE_PTV}^2 + \text{SE_CNV-LoF}^2}$ under a conservative zero-covariance approximation (shared-control covariance is unavailable in public summary statistics,

making this a conservative screen rather than a formal equivalence test). With a pre-specified equivalence margin of 0.20 trait SD, a pair was classified as equivalent when its 90% confidence interval lay fully inside $[-0.20, 0.20]$, as heterogeneous when its 95% confidence interval lay fully outside that interval, and otherwise as information-insufficient. A strict label set applied global Benjamini-Hochberg false-discovery control at $q < 0.05$ separately to the equivalence and heterogeneity tests; strict labels were frozen as the only admissible targets for the prediction audit. Class counts were stable (rank correlation 1.0) when the assumed between-estimate correlation was varied from 0 to 0.25. Phase-0 direction analyses used per-modality nominal P thresholds (0.05, 0.01, 0.001, 0.0001) as pre-specified sensitivity settings, not as multiple-testing-corrected discovery claims.

Predictability diagnostic

The strict strict-heterogeneous ($n = 216$) and strict-equivalent ($n = 4,252$) rows formed a binary diagnostic target (prevalence 0.048). Three logistic-regression feature sets were compared out-of-fold: precision-only (standard errors, allele frequencies and derived ratios), CNV architecture only (event counts, spans, covered-gene counts and focal versus multi-gene event fractions from the AZ annotation), and the combination. Genes never crossed folds (gene-grouped GroupKFold), and phenotype-family holdouts (biochemistry, hematology and other quantitative traits) were used as a secondary stress test. Because separability by precision features could reflect label determinism rather than biology, we additionally measured power-matched overlap: rows were binned by out-of-fold predicted power propensity, and the 0.05-0.95 interval was required to contain at

least 50 rows of each class. Metrics were average precision (AP), area under the ROC curve (AUROC) and balanced accuracy at the 0.5 threshold.

CNV architecture association

Genes carrying at least one strict-heterogeneous row were compared against genes with strict-equivalent rows only, using per-gene CNV architecture features from the AZ annotation: number of available LoF deletion events, median and maximum deletion span, median protein-coding genes covered per event, multi-gene event fraction and focal event fraction. Group differences were tested with Mann-Whitney U tests and Benjamini-Hochberg correction across the nine features.

Software and ethics

Analyses used Python 3.10.12 with numpy 2.2.6, pandas 2.3.3 and scikit-learn 1.7.2 (logistic regression pipelines with median imputation and standardization). Only public, de-identified summary statistics were used; no individual-level data were accessed. AZ resource terms (research use only, no re-identification, required attribution) were followed throughout.

Results

Phase 0: cross-perturbation coverage and the failure of the three-level direction

Parsing the Hujoel et al. supplementary material yielded 100 significant pLoF gene-trait combinations involving 77 genes and 41 traits. Among them, PTV and CNV-LoF effect directions agreed in 93.0% of pairs, with Pearson $r = 0.634$ and Spearman $\rho = 0.764$ across effect sizes; only 10 pairs additionally had significant duplication evidence, and 60% of those duplication effects went in the opposite direction to the loss effects, arguing

against naive reciprocal-dosage symmetry. In the full 203,484 LoF-pair universe (16,957 genes), raw sign concordance was 46.9%, confirming that most numerically estimable pairs are noise. Tightening the per-modality threshold to $P \leq 0.01$ isolated 139 LoF pairs (120 genes) with 91.4% concordance (binomial $P = 1.1e-25$); at the same threshold the three-level evidence shrank to 16 combinations covering only 14 genes, far below the frozen minimum of 100. The three-level therapeutic-window direction was therefore stopped and the analysis froze on the narrower PTV-versus-CNV-LoF equivalence question.

A bone-mineral-density case study illustrates both the recoverability of classical positive controls and the sparsity of strong three-level evidence. At an exploratory per-modality threshold of $P \leq 0.05$, BMD had twelve three-level candidates: SOST, NPIPA5, WNT16 and ZSCAN2 were monotonic across the dose ladder, ARVCF and PI4KA were concordant for loss but not reciprocal for duplication, and CERS2, CLDN4, LZTR1, RHCE, TXNRD2 and YPEL3 disagreed between PTV and CNV-LoF. At $P \leq 0.01$ only SOST (monotonic) and CERS2 (fully discordant) remained.

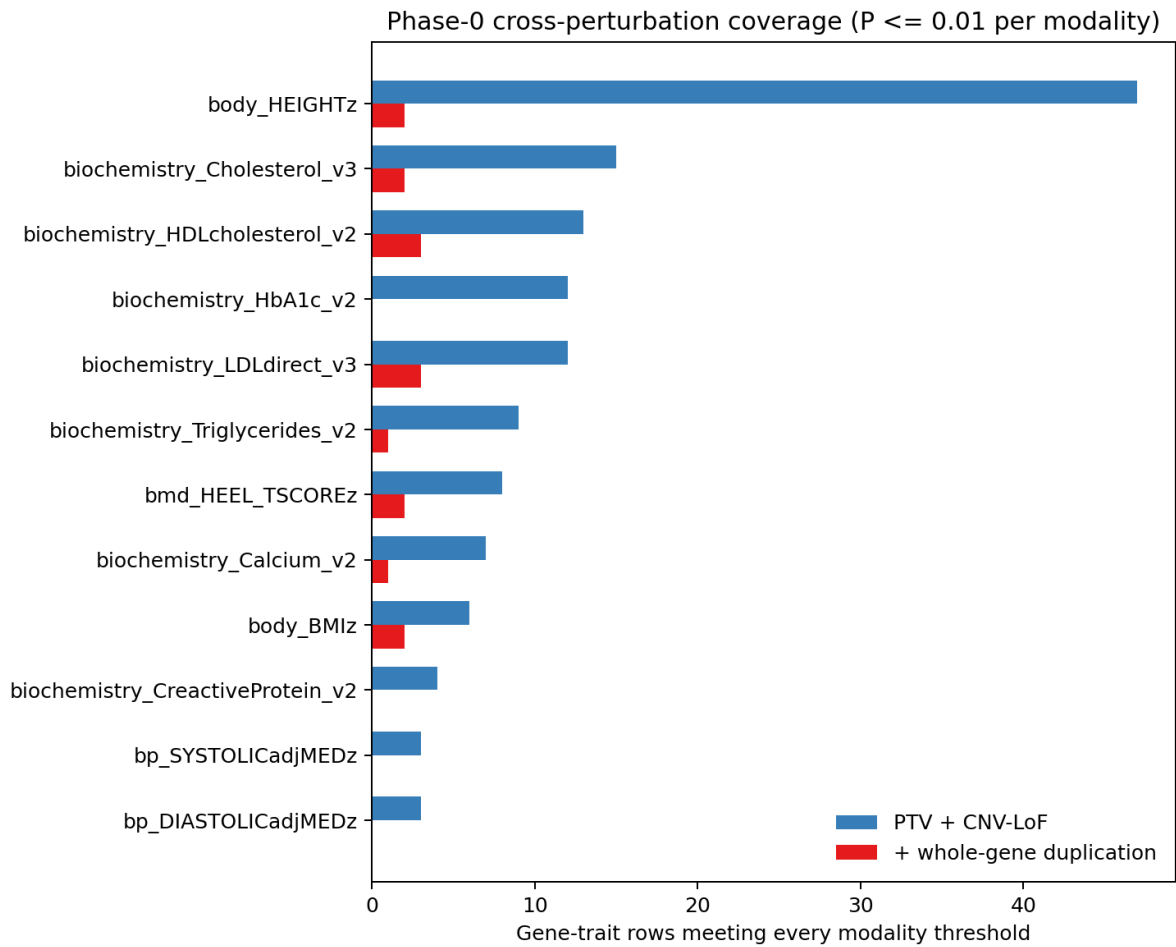


Figure 2A. Phase-0 trait coverage: number of extractable gene-trait estimates per quantitative trait and perturbation class across the 12 phase-0 traits.

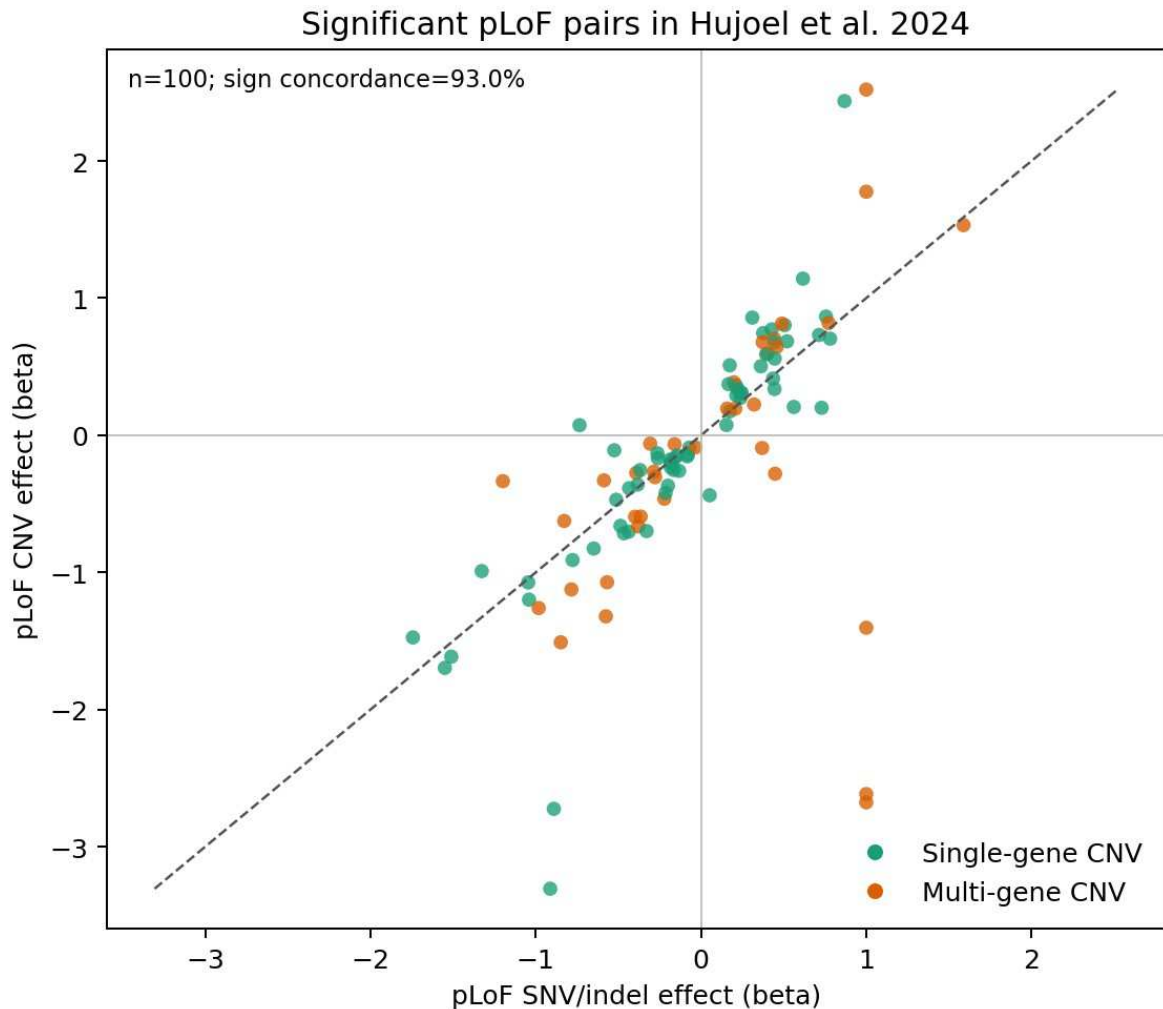


Figure 2B. Phase-0 effect concordance between pLoF SNV/indel and pLoF CNV estimates for significant LoF pairs; identity line shown for reference.

Per-modality threshold	LoF pairs (genes)	LoF sign concordance	Three-level pairs (genes)	Duplication-reciprocal fraction
P ≤ 0.05	952 (842)	67.3%	98 (87)	67.3%
P ≤ 0.01	139 (120)	91.4%	16 (14)	81.2%
P ≤ 0.001	54 (47)	96.3%	5 (4)	100.0%
P ≤ 0.0001	31 (27)	100.0%	2 (2)	100.0%

Table 1. Phase-0 threshold sensitivity of cross-perturbation concordance. All thresholds are nominal per-modality settings, not multiple-testing-corrected discovery claims.

Phase 1: an external PTV probe confirms direction replication but exposes a discordant tail

The frozen phase-1 gate required at least 50 rows and 75% sign agreement with the external AZ PTV probe; it was passed with 112 matched rows covering 96 genes. Phase-

0 PTV and AZ PTV effects agreed in sign in all 112 rows (Pearson $r = 0.918$; Spearman $\rho = 0.972$; fitted slope 0.813). This is replication across analysis pipelines within the strongly overlapping UK Biobank participant base, not an independent-cohort replication. The cross-perturbation contrast was weaker but still strongly concordant: phase-0 CNV-LoF versus AZ PTV sign agreement was 106/112 (94.6%; Pearson $r = 0.825$; Spearman $\rho = 0.888$). Under the conservative independence-variance screen, 15 of 112 rows showed FDR-significant heterogeneity, of which 6 went in opposite directions. Discordant rows had fewer PTV carriers than concordant rows (median 90.5 vs 146.5), consistent with noise or with regional CNV effects rather than systematic biology at this sample size.

Gene	Trait	CNV-LoF beta	AZ PTV beta	Heterogeneity q (BH)
RHCE	Glycated haemoglobin (HbA1c)	+0.009	-0.527	1.4e-05
OR3A1	Calcium	+0.595	-0.226	0.0056
HP	Cholesterol	-0.034	+0.282	0.014
RASGEF1B	Systolic blood pressure automated reading	+1.786	-0.181	0.032
HP	LDL direct	-0.038	+0.240	0.042
FAT4	C-reactive protein	+0.404	-0.123	0.091

Table 2. The six opposite-direction CNV-LoF versus AZ PTV rows from the phase-1 power-aware screen (conservative independence-variance approximation; effect units are trait SD).

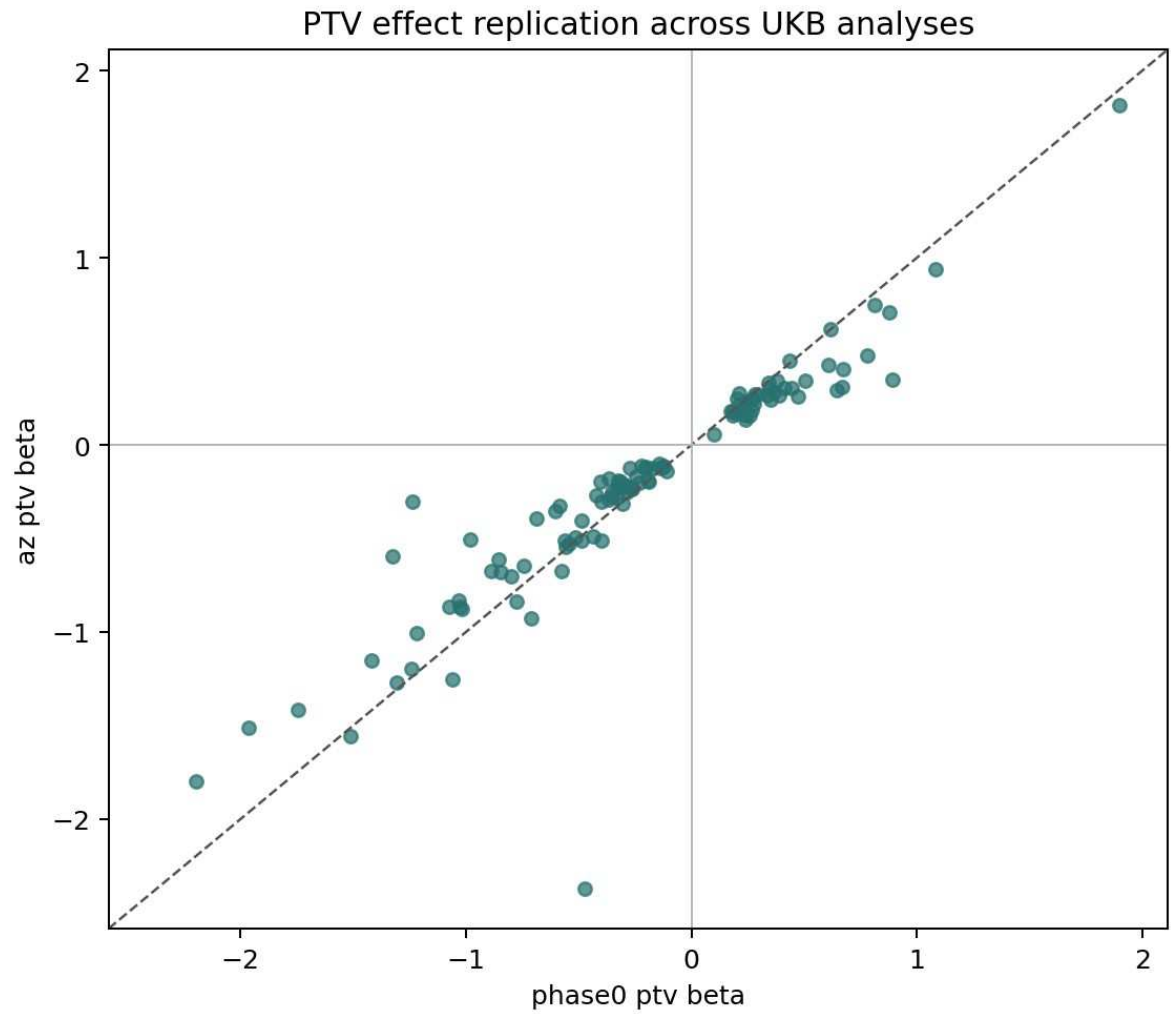


Figure 3A. Phase-0 PTV versus AstraZeneca 500k WGS PTV effect estimates for the 112 matched gene-trait rows; identity line and fitted slope (0.813) shown.

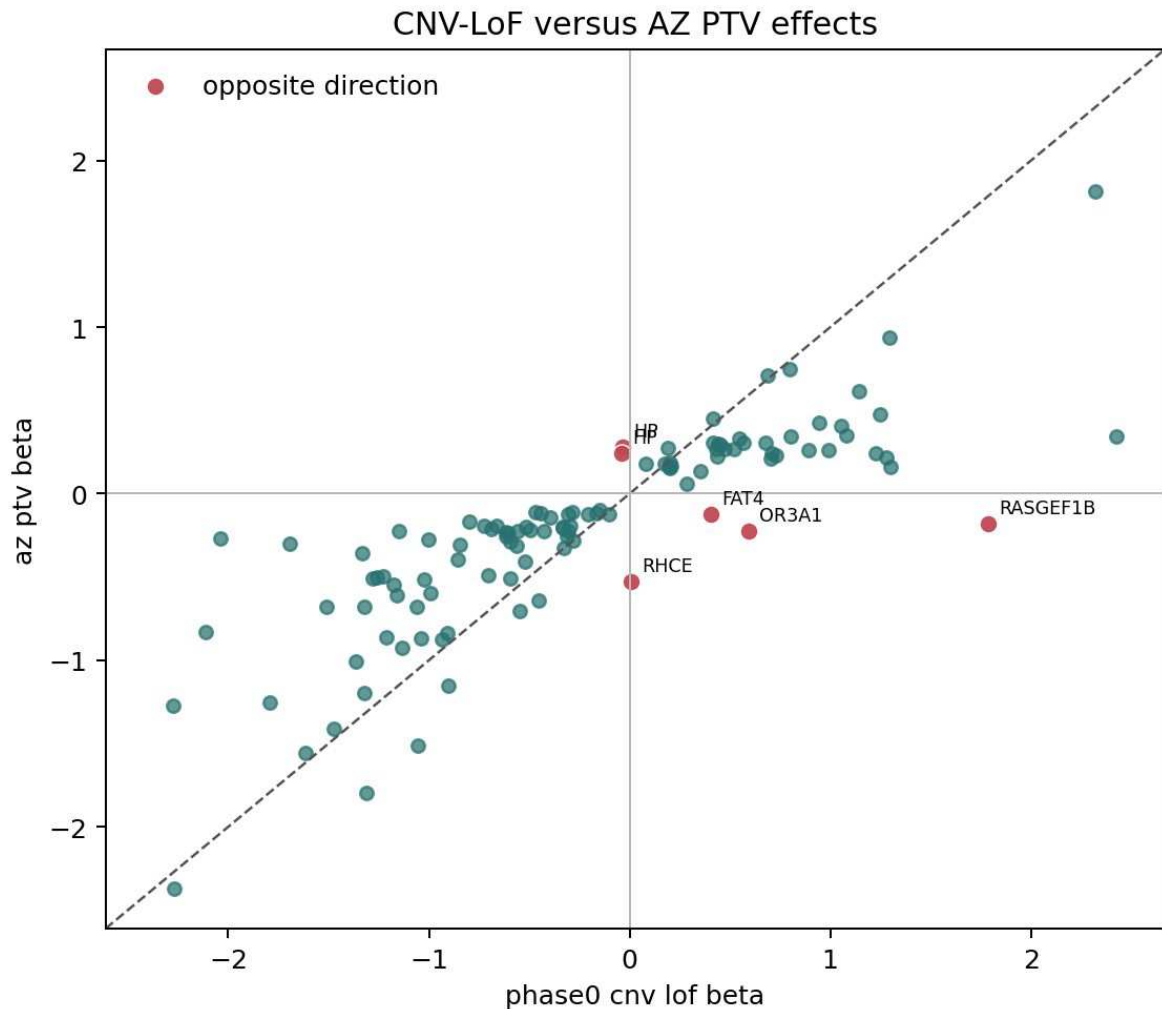


Figure 3B. Phase-0 CNV-LoF versus AstraZeneca PTV effect estimates for the same rows; six opposite-direction points annotated.

Phase 2: a genome-wide three-class inventory in which almost everything is uninformative

All 57 complete quantitative-trait files contributed 1,063,107 outer-ladder gene-trait rows, of which 885,362 pairs (15,693 genes) had valid, non-truncated PTV and CNV-LoF estimates; 177,745 rows were excluded for missing or invalid estimates. No P-value filter was applied to construct this universe. Under the descriptive (unadjusted) confidence-interval classes, 28,761 pairs were equivalent, 14,784 heterogeneous, and

841,817 information-insufficient. After global Benjamini-Hochberg control at $q < 0.05$, the strict classes were 4,252 equivalent, 216 heterogeneous, and 880,894 information-insufficient - the dominance of the uninformative class is the correct outcome, since neither a non-significant contrast nor a single active modality was ever used as evidence. Class assignments were perfectly stable (rank correlation 1.0) when the assumed between-estimate correlation was raised from 0 to 0.25.

Composition analysis explains what the strict classes mean biologically. Of the 216 strict-heterogeneous rows, 196 were active only in the CNV-LoF modality, 13 were active in both modalities in opposite directions, and only 7 were active in the same direction but materially different in magnitude, spanning 98 unique genes and 31 traits. Of the 4,252 strict-equivalent rows, 3,728 were active in neither modality, 425 were CNV-only active, 87 PTV-only active, and only 9 were active in the same direction in both modalities (with 3 opposite-direction pairs equivalent near zero). The activity audit gate - at least 200 active same-direction strict-equivalent rows to support any drug-portability label - therefore failed decisively, with only 9 observed. Strict equivalence under this design is predominantly precise null-equivalence, not evidence that perturbing the target changes the phenotype.

Class	Descriptive CI classes	Strict (global BH $q < 0.05$)
Equivalent	28,761	4,252
Heterogeneous	14,784	216
Information-insufficient	841,817	880,894
Total valid pairs	885,362	885,362

Table 3. Phase-2 three-class counts over 885,362 valid PTV/CNV-LoF gene-trait pairs from 57 quantitative traits and 15,693 genes.

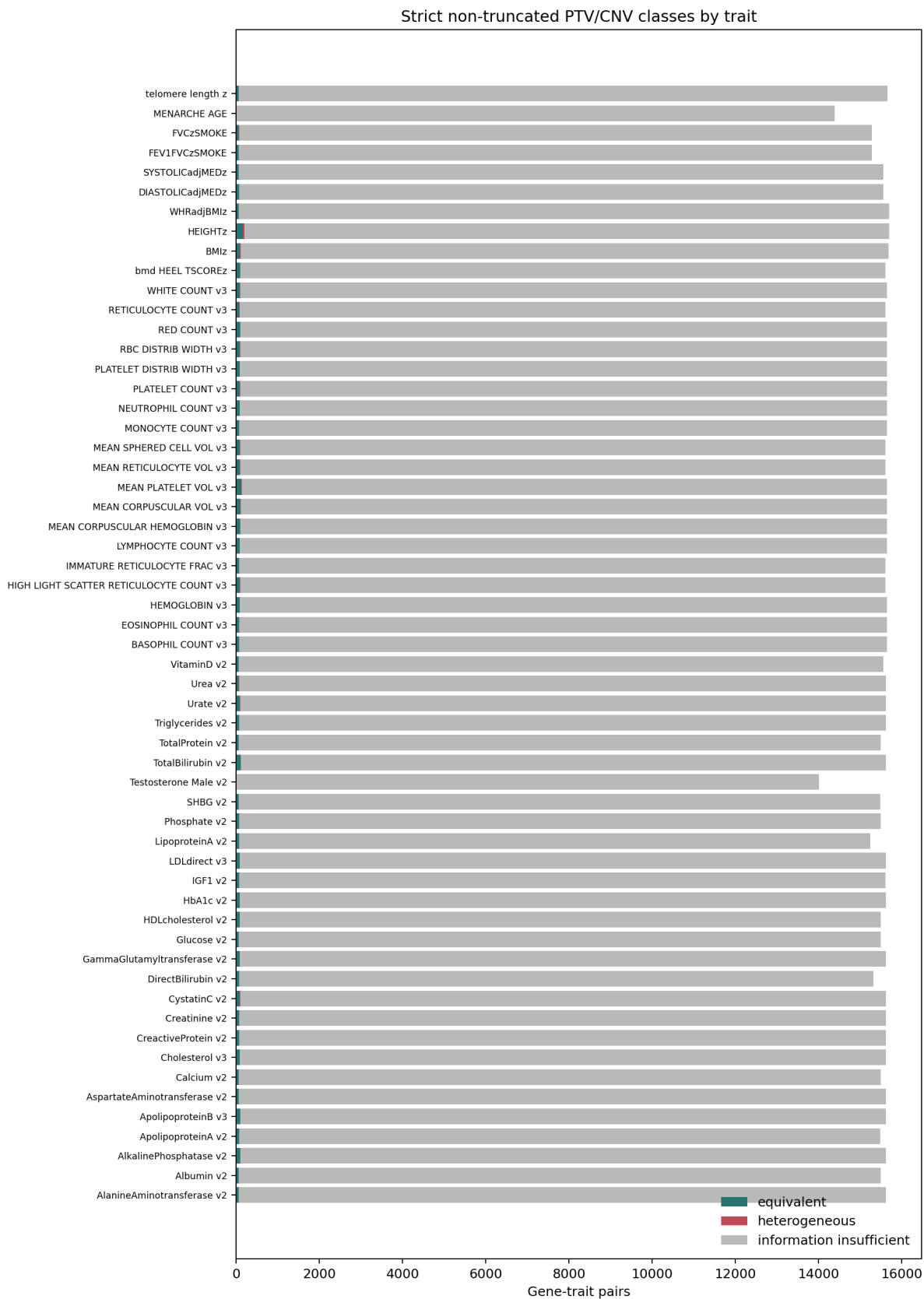


Figure 4A. Distribution of strict three-class assignments across the 57 quantitative traits.

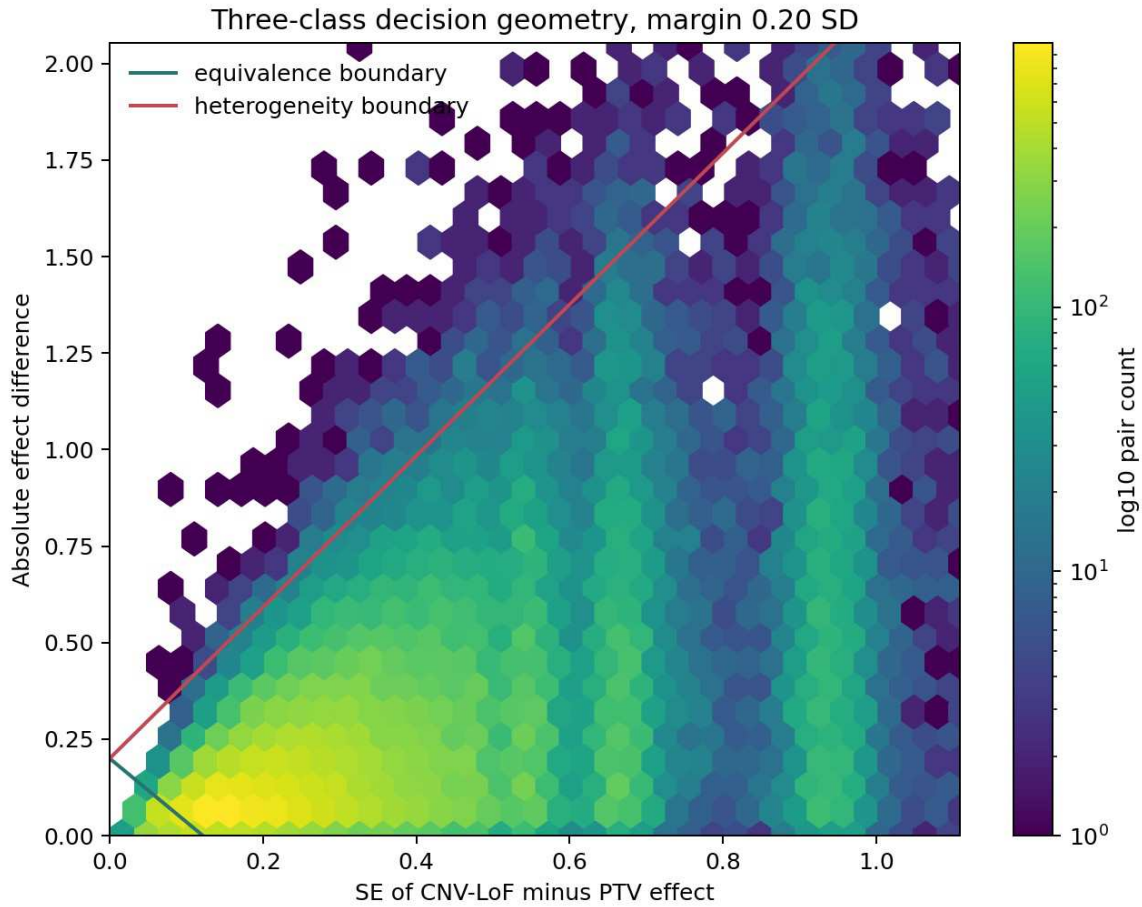


Figure 4B. Decision geometry of the three-class rule: effect difference versus its standard error with the frozen 0.20 SD equivalence margin; strict heterogeneous and strict equivalent rows highlighted.

Strict heterogeneity is predictable from precision alone, and architecture adds nothing out-of-fold

In the gene-grouped out-of-fold diagnostic, the precision-only model separated strict-heterogeneous from strict-equivalent rows almost perfectly (average precision 0.9999; AUROC 1.0000; balanced accuracy 0.9951). The architecture-only model contained real but weak signal (AP 0.1821 vs prevalence 0.0483; AUROC 0.8105), yet adding architecture to the precision baseline changed AP by -0.0024, failing the frozen +0.03 incremental gate. Power-matched overlap was equally decisive: within the 0.05-0.95 out-

of-fold power-propensity interval there were only 15 heterogeneous versus 370 equivalent rows, versus the required minimum of 50 per class. Broad phenotype-family holdouts did not change this: the precision-only model remained near-perfect within biochemistry, hematology and other trait families. All diagnostic gates failed, and the frozen stop line was invoked: no production classifier was trained and no portability predictor is claimed.

Feature set (gene-grouped OOF)	Average precision	AUROC	Balanced accuracy
Precision only	0.9999	1.0000	0.9951
CNV architecture only	0.1821	0.8105	0.7501
Combined	0.9975	0.9999	0.9928

Table 4. Gene-grouped out-of-fold diagnostic performance separating 216 strict-heterogeneous from 4,252 strict-equivalent rows (positive-class prevalence 0.0483). The frozen incremental gate required the combined model to exceed the precision-only AP by at least +0.03; the observed gain was -0.0024.

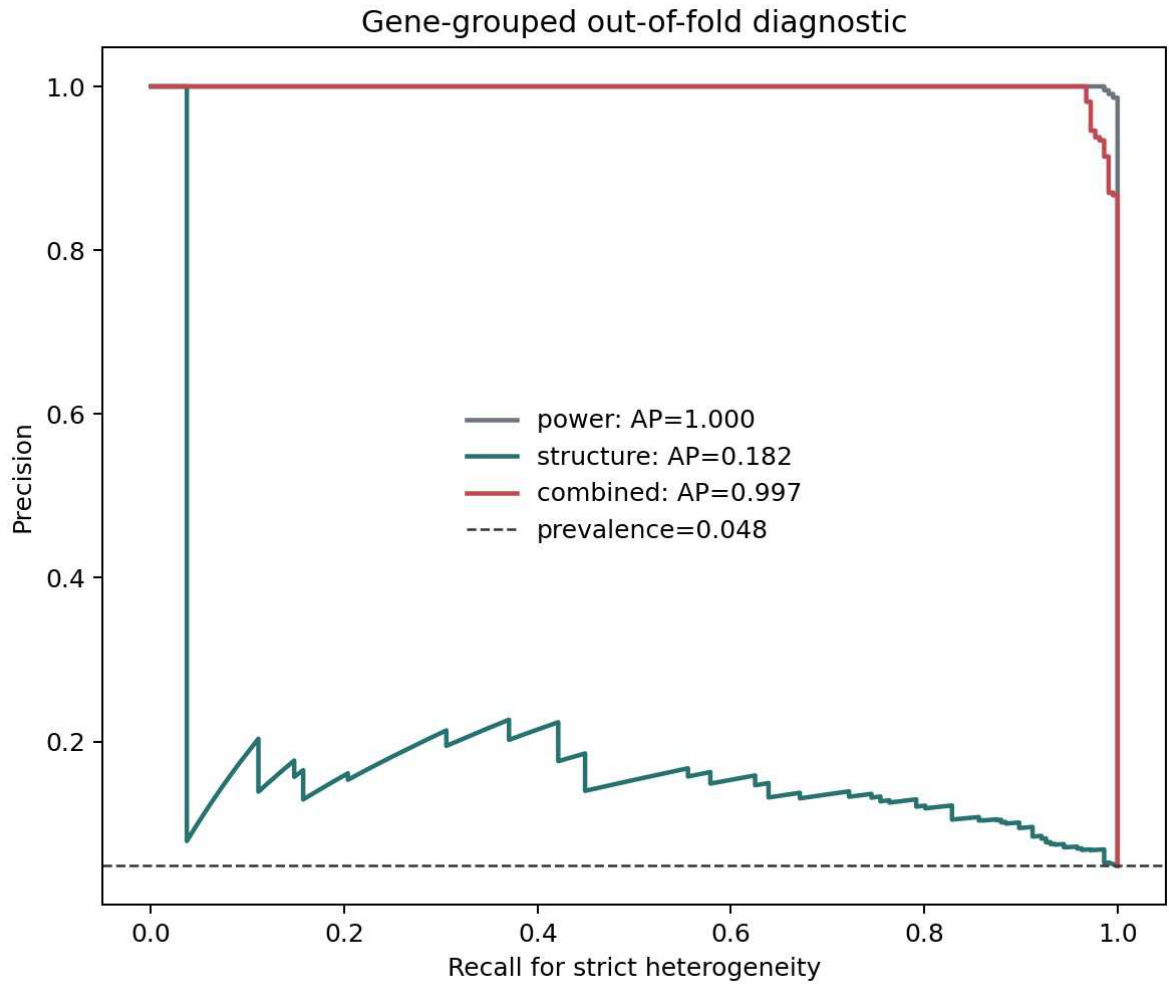


Figure 5. Out-of-fold precision-recall curves for the three feature sets under gene-grouped cross-validation, with the positive-class prevalence baseline shown.

Heterogeneous-effect genes sit in multi-gene deletions

CNV architecture feature	Het genes (n)	Eq-only genes (n)	Heterogeneous genes (median)	Equivalent-only genes (median)	Mann-Whitney P	BH q
LoF deletion events (gene)	97	164	3	4	0.0005	0.00075
Median deletion span (bp)	87	143	191,443	97,259	0.029	0.037
Maximum deletion span (bp)	87	143	509,266	435,563	0.92	1
Median protein-coding genes covered per LoF deletion	97	164	10	2.5	5.8e-13	5.2e-12

Maximum protein-coding genes covered per LoF deletion	97	164	22	10	1e-09	4.7e-09
Multi-gene event fraction	97	164	1	0.75	2.5e-07	4.6e-07
Focal event fraction	97	164	0	0.25	2.5e-07	4.6e-07
Half-gene deletion fraction	97	164	1	0.690476	1.4e-07	4.3e-07
Whole-gene deletion fraction	97	164	0	0	1	1

Table 5. Per-gene CNV architecture comparison between genes carrying at least one strict-heterogeneous row and genes with strict-equivalent rows only, from the AZ UKB WGS CNV annotation. Nine features were tested; q values are Benjamini-Hochberg adjusted across features.

Although the prediction gate failed, the structural association itself was strong: genes carrying strict-heterogeneous rows were far more likely to be covered by multi-gene deletions than equivalent-only genes (median 10 versus 2.5 protein-coding genes per LoF deletion event, BH $q = 5.2 \times 10^{-12}$; median maximum covered genes 22 versus 10, $q = 4.7 \times 10^{-9}$; multi-gene event fraction 1.00 versus 0.75, $q = 4.6 \times 10^{-7}$; focal event fraction 0 versus 0.25, $q = 4.6 \times 10^{-7}$; median span 191 kb versus 97 kb, $q = 0.037$). This pattern is exactly what regional passenger or oligogenic CNV effects predict, and it mirrors the central conclusion of Zou et al. (2026). We emphasize, per the frozen analysis plan, that these correlations are not independent predictive evidence: the strict classes are almost completely separated by precision features, and the annotation describes available events rather than the carrier events driving each association.

Discussion

This audit set out to test whether public human CNV and protein-truncating-variant summary statistics contain enough cross-perturbation signal to support a calibrated predictor of when a focal gene-level loss interpretation is portable between SNV and

CNV perturbations. The answer, after three pre-gated phases and 885,362 genome-wide pairs, is no under the current label design - and the reasons are themselves informative. Three findings stand. First, once equivalence must be affirmatively demonstrated rather than inferred from non-significance, the informative fraction of the genome is tiny: 4,252 supported-equivalent and 216 supported-heterogeneous pairs against 880,894 uninformative ones. Second, drug-portability-style labels are unattainable because only nine strict-equivalent pairs are active in the same direction in both modalities; most strict equivalence is precise null-equivalence. Third, the strict classes are almost deterministically separable by standard errors and frequencies, so any classifier trained on them would learn the sampling design, not biology; CNV architecture added no out-of-fold value over this baseline.

These negative results delimit, rather than merely fail to support, a research program. The observed heterogeneity signal is real and biologically coherent - it concentrates in genes embedded in multi-gene deletions, extending the passenger and oligogenic mechanism that Zou et al. established as a central property of genome-wide CNV associations. What failed was the attempt to predict non-portability from effect-derived labels: labels constructed from the same precision features used for prediction cannot certify those features as biological predictors. A defensible successor study needs labels that are independent of both the predictors and statistical power, such as contrasts between phenotype-matched focal single-gene deletions and multi-gene deletions, carrier-level conditional estimates that isolate the focal gene from its neighbours, and replication of the same focal-versus-multi-gene contrast in an independent biobank. Only after such labels survive grouped holdouts would it be legitimate to test incremental prediction of

external drug direction, clinical progression or safety outcomes against baselines such as carrier count, current Z statistics, LOEUF, and the pHaplo/pTriplo dosage-sensitivity map.

Several limitations bound every number reported here. All estimates derive from the same UK Biobank analytical framework, so the phase-1 'replication' is cross-pipeline, not cross-cohort. The zero-covariance standard error for effect differences is conservative but the true shared-control covariance is unknown, and the AZ portal data are truncated at approximately $P < 0.1$. pLoF_high-confidence CNV calls are not equivalent to focal whole-gene deletions. The 57 traits are quantitative phenotypes only, and gene-level tests mask allelic heterogeneity. Most importantly, genetic effect compatibility is not a pharmacological therapeutic window: nothing in this dataset distinguishes passenger from oligogenic effects at the individual level, and no drug-outcome endpoint was analyzed. We have deliberately published the failed gates alongside the passing ones so that the stop line is as citable as any positive finding.

In summary, public summary statistics are sufficient to build a rigorous genome-wide inventory of cross-perturbation equivalence - which we release at row level - but not to train a portability predictor from effect-derived labels. The inventory, the frozen gates, and the explicit redesign requirements are the durable contributions of this work.

Data availability

The primary summary statistics are public: Hujoel et al. 2024 supplementary tables and UKB WES CNV summary statistics (Zenodo 10.5281/zenodo.10529671; https://data.broadinstitute.org/lohlab/UKB_WES_CNV_sumstats/; CC BY 4.0), and the AstraZeneca CGR public UK Biobank 500k WGS resource

(<https://public.cgr.astrazeneca.com/BULK-DOWNLOAD/V01/>), used under its license (research use only, required attribution to Wang et al. 2021 and Zou et al. 2026). Derived row-level phase-0, phase-1 and phase-2 datasets, including the 885,362-row three-class classification, will be deposited at Zenodo upon posting, and the dataset DOI will be added to this preprint in a subsequent revision.

Code availability

All analysis and download scripts, frozen decision documents and verification manifests (including SHA-256 checksums for all 57 primary files) are in the project repository and will be deposited at a public archive upon posting.

Acknowledgements, funding and competing interests

Computations were run on a private Linux workstation. This work received no specific funding. The authors declare no competing interests.

References

1. Hujuel MLA, Sherman MA, Arvanitis M, et al. Protein-altering variants at copy number-variable regions influence diverse human phenotypes. *Nature Genetics*. 2024;56:569-578.
2. Zou XZ, Hu F, Lou H, et al. Phenome-wide analysis of copy number variants in 470,727 UK Biobank genomes. *Nature*. 2026;652:675-685. doi:10.1038/s41586-025-10087-x. (Author correction: *Nature*. 2026;652:E7. doi:10.1038/s41586-026-10389-8.)
3. Wang Q, Dhindsa RS, Carss K, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*. 2021;597:527-532.
4. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581:434-443.
5. Collins RL, Glessner JT, Porcu E, et al. A cross-disorder dosage sensitivity map of the human genome. *Cell*. 2022;185:3041-3055.
6. Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*. 2017;8:355-362.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*. 1995;57:289-300.
8. Ochoa D, Hercules A, Carmona M, et al. Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Research*. 2021;49(D1):D1302-D1310.
9. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-2830.