GeneBench-Pro Case Study: Structural-inversion Subhaplotypes, Expression, and Clinical Risk

GeneBench-Pro

June 26, 2026

1 Overview

This case study asks whether an analyst can analyze an anonymous inversion-like locus whose broad outer orientation is not the same biological object as the nested segment-B sub-haplotype. The released package contains seven gzip-compressed TSV files: a 1,900-person ascertained cohort, 60 anonymous marker dosages, a 780-sample long-read breakpoint calibration panel, 2,240 gene-expression rows, a 108-row sampling-design table, marker annotations, and file-layout notes. The requested answer has four fields: the source-population total clinical log odds ratio per additional calibrated segment-B copy, the expression log fold-change for the credible segment-B-supported gene, a binary sign-concordance support code, and the number of reliable breakpoint-panel carriers.

The correct analysis is a three-stage cascade. Stage 1 distinguishes nested segment-B dosage from global outer orientation and calibrates both continuous dosages from reliable breakpoint labels and marker evidence. Stage 2 selects molecular support only after excluding low-uniqueness expression artifacts and adjusting tissue source, processing batch, baseline covariates, and global orientation. Stage 3 estimates the clinical total effect with inverse sampling-fraction weights and baseline/structural adjustment while leaving downstream expression out of the outcome model. Naive analyses fail for realistic reasons: a top marker or marker PC tracks ancestry and global orientation, raw marker-block means collapse opposite marker phases, apparently high-quality breakpoint rows are biased in repeat-complex neighborhoods, the expression panel contains both a low-uniqueness paralog-like signal and a high-uniqueness global-orientation decoy, and the case-control cohort is not self-weighting.

The biological motivation is the same kind of distinction that arises in complex regulatory loci such as 17q21.31/MAPT: broad structural orientation, nested local haplotype, expression state, and clinical association can point in different directions. The case study is intentionally anonymous, but it preserves that scientific structure. A broad inversion marker can be a useful covariate while still being the wrong exposure; a high-uniqueness expression signal can be real but attributable to global orientation; and an expression-mediated pathway can support biological plausibility while still being downstream of the clinical total-effect estimand. The analysis must therefore keep structural calibration, molecular support, and clinical transport as separate stages until the final answer is assembled.

2 Released Prompt and Files

Prompt

```
Analyze the released files for anonymous Locus Q. Estimate the full-cohort source-population
    clinical association and molecular expression support for the calibrated nested segment-B
    structural copy dosage, separating the nested segment-B dosage from the broader outer-
    orientation dosage. Report subhap_log_or as the natural-log source-population total-effect odds
     ratio for case status per additional calibrated segment-B copy. Report expression_log_fc as
    the natural-log expression fold-change per calibrated segment-B copy for the expression-
    supported gene. Report target_support_code as 1 if the supported gene has a positive
    expression_log_fc and the clinical association is protective (subhap_log_or < 0), otherwise 0.
    Report n_calibrated_carriers as the number of reliable breakpoint-panel samples carrying at
    least one segment-B copy.
These data came from a real experiment; you will be graded not just on numerical correctness but
    the quality of analytical reasoning you exhibit; do not attempt to take any shortcuts.
Return your final answer as exactly one JSON object.
Do not wrap the JSON in markdown.
Do not add prose before or after the JSON.
Do not omit any keys shown in the example.
Return the JSON object in your final answer:
{
  "answer": {
    "n_calibrated_carriers": <int>,
    "target_support_code": <int>,
    "expression_log_fc": <float>,
    "subhap_log_or": <float>
  },
  "reasoning": "<description of method and QC>"
}
```

Released data files

File	Format	Contents
`cohort.tsv.gz`	.tsv.gz	Participant-level clinical status, age, sex, ancestry PCs, ancestry group, clinic stratum, age band, and recruitment stream for the 1,900-person released cohort.
`tag_markers.tsv.gz`	.tsv.gz	Imputed alternate-allele dosages for 60 anonymous markers across the locus for the full released cohort.
`marker_info.tsv.gz`	.tsv.gz	Marker positions, allele labels, anonymous panel-block labels, and imputation-information summaries.
`breakpoint_panel.tsv.gz`	.tsv.gz	Long-read calibration panel for a subset of cohort samples, including outer-orientation fraction, segment-B copy index, segment-B depth, and breakpoint-reliability summaries.
`expression_panel.tsv.gz`	.tsv.gz	Gene-level log-expression summaries with tissue source, processing batch, assay depth, read-uniqueness, multimapper, and GC metrics.

`sampling_design.tsv.gz`	.tsv.gz	Source-population sampling fractions by clinic stratum, recruitment stream, ancestry group, age band, and case status.
`file_notes.tsv.gz`	.tsv.gz	Column-level notes defining the released files and clarifying which fields are sampling-design strata.

3 Answer Fields and Tolerances

The reported fields map the calibrated nested segment-B dosage to the clinical, molecular, and support summaries:

$$\texttt{subhap\_log\_or} = \widehat{\beta}_S, \tag{1}$$

$$\texttt{expression\_log\_fc} = \widehat{\theta}_{GENE2}, \tag{2}$$

$$\texttt{target\_support\_code} = I(\widehat{\beta}_S < 0, \widehat{\theta}_{GENE2} > 0), \tag{3}$$

$$\texttt{n\_calibrated\_carriers} = \sum_{i \in LR} I(R_i^{LR} = 1, S_i^{LR} \geq 1). \tag{4}$$

Answer field	Ground truth	Tolerance / matching rule	Interpretation
`subhap_log_or`	-0.400921	Absolute error ≤ 0.055	Source-population clinical log odds ratio per additional calibrated segment-B copy.
`expression_log_fc`	0.333341	Absolute error ≤ 0.030	Expression log fold-change for the segment-B-supported gene.
`target_support_code`	1	Exact integer match; valid range $\{0, 1\}$	Sign-concordance support code for protective clinical association and positive expression support.
`n_calibrated_carriers`	195	Exact integer match; valid range ≥ 0	Reliable breakpoint-panel samples carrying at least one calibrated segment-B copy.

4 Structure Diagram

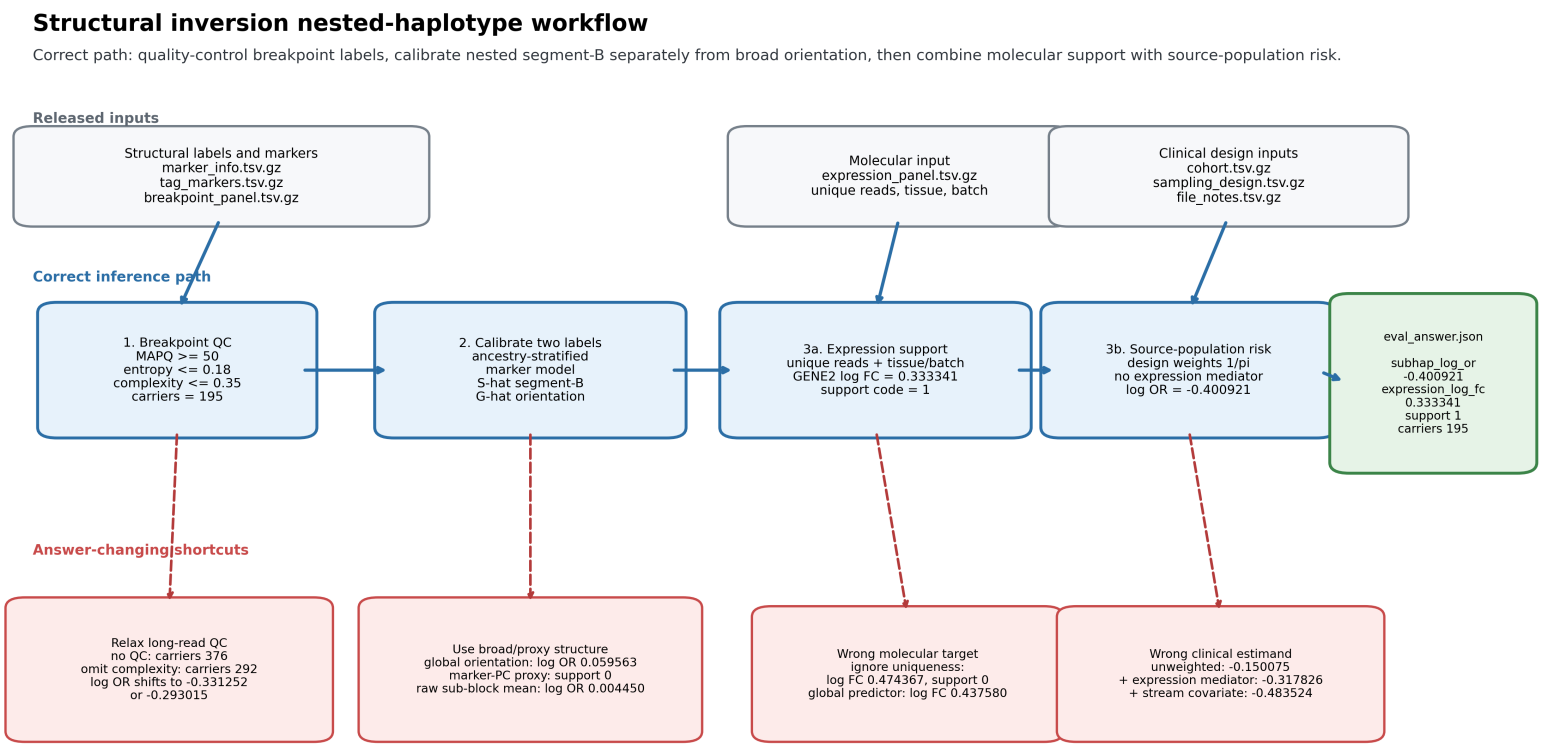

Figure 1: Pipeline-style structure diagram. Gray boxes are released input files, blue boxes and solid blue arrows show the intended workflow, green is the released answer file, and red dashed branches show representative incomplete analyses with their observed wrong outputs. The diagram makes the central contract visible: long-read breakpoint quality control first fixes the reliable segment-B carrier count at 195; ancestry-stratified calibration then separates nested segment-B dosage from broad outer orientation; expression support is estimated as a high-uniqueness tissue/batch-adjusted molecular result; and the clinical answer is a source-population weighted total-effect log odds ratio rather than an unweighted, mediator-adjusted, or recruitment-stream-conditioned contrast.

5 Variables and Assumptions

Variable	Type/domain	Assumption or role
i	person index	Construction source population has 11,500 individuals; the released cohort has 1,900 sampled individuals.
A_i	EUR, AFR, EAS	Broad ancestry group with probabilities 0.55, 0.25, and 0.20.
H_{i1}, H_{i2}	A, B, C, D	Two local haplotypes drawn conditional on ancestry. A and D have global orientation 0; B and C have global orientation 1. B and D carry nested segment B.
S_i	0,1,2	True nested segment-B copy count, $S_i = I(H_{i1} \in \{B, D\}) + I(H_{i2} \in \{B, D\})$.
G_i	0,1,2	True global outer-orientation count, $G_i = I(H_{i1} \in \{B, C\}) + I(H_{i2} \in \{B, C\})$.
PC_{ik}	real	Three ancestry-centered principal components used as baseline covariates and marker-calibration features.
X_i	baseline variables	Age, sex, clinic stratum, ancestry group, and PCs. Clinic is correlated with ancestry and S_i but precedes the clinical outcome.
L_i	real	Latent expression-like mediator downstream of segment-B dosage. It affects disease risk but is not adjusted for in the total-effect clinical estimand.
Y_i	0,1	Clinical case status from the source-population logistic model.
R_i	registry, clinic	Recruitment stream used in the sampling design. It is an ascertainment stratum, not an ordinary outcome covariate in the reference clinical model.
I_i	0,1	Inclusion in the released case-control cohort according to the sampling-design table and deterministic thinning to 1,900 rows.
M_{im}	0–2 dosage	Noisy anonymous marker dosage. Different marker blocks tag global orientation, nonportable segment-B signal, opposite-phase segment-B signal, recombinant state, or ancestry background.
B_i	breakpoint summaries	Long-read molecule count, MAPQ, entropy, pair balance, outer-orientation fraction, segment-B copy index, segment-B depth, and local breakpoint complexity.
E_{ig}	log expression	Gene-level expression with tissue, batch, read-uniqueness, and structural effects.

w_i	positive weight	Normalized inverse sampling fraction, $w_i \propto 1/\pi_i$, used to recover the source-population clinical contrast.

The recoverable estimator assumes that reliable breakpoint calls identify calibration labels for S_i and G_i, that marker dosages can be used to predict continuous structural dosage in the rest of the cohort, that the sampling fractions represent the source-to-release inclusion design, and that the expression and clinical models are finite-sample statistical targets rather than claims about a real named disease locus. The biological analogy is 17q21.31/MAPT-like inversion sub-haplotypes, where broad inversion state, sub-haplotype structure, and tissue-specific expression can diverge [1–3,5].

6 Data-Generating Process

The simulation is deterministic with seed 175431. It creates 11,500 construction-only source-population individuals and releases a compact 1,900-person ascertained case-control cohort with 1,544 cases and 356 controls. The public target is the recoverable released-data estimator, not the hidden DGP coefficient vector.

Source haplotypes and baseline variables. Haplotype states map to (G, S) as A=(0,0), B=(1,1), C=(1,0), and D=(0,1). Conditional haplotype frequencies are ancestry-stratified: EUR draws A/B/C/D with probabilities 0.55/0.25/0.14/0.06, AFR with probabilities 0.68/0.08/0.16/0.08, and EAS with probabilities 0.50/0.18/0.25/0.07. The PCs are independent normals around ancestry-specific means:

$$\begin{aligned} PC_i \mid A_i &\sim N(\mu_{A_i}, \mathrm{diag}(0.18^2, 0.20^2, 0.16^2)), \\ \mu_{EUR} &= (-1.00, 0.00, 0.05), \\ \mu_{AFR} &= (0.95, 0.42, -0.05), \\ \mu_{EAS} &= (0.20, -0.82, 0.08). \end{aligned} \tag{D1}$$

Sex is Bernoulli(0.48). Age is

$$Age_i = \mathrm{clip}\{N(59.5 + 1.5I(A_i = \mathrm{AFR}) + 0.7S_i, 11.5^2), 30, 86\}, \tag{D2}$$

with age bands 30–49, 50–64, and 65+. Clinic is a baseline healthcare-access stratum:

$$p_{T,i} = \mathrm{expit}[-1.35 + 0.50I(A_i = \mathrm{EUR}) + 0.58S_i + 0.18(Age_i - 60)/10], \tag{5}$$

$$p_{R,i} = \mathrm{expit}[-0.15 + 0.15I(A_i = \mathrm{EAS}) - 0.12S_i], \tag{D3}$$

where a uniform draw assigns tertiary clinic with probability $p_{T,i}$ and regional clinic with probability $(1 - p_{T,i})0.48p_{R,i}$; otherwise clinic is community.

Molecular mediator and disease. The latent GENE2-like expression mediator is

$$L_i = 0.305S_i + 0.020G_i + 0.035(Age_i - 60)/10 - 0.020PC_{i2} + \epsilon_i, \quad \epsilon_i \sim N(0, 0.18^2). \tag{D4}$$

The source disease model is

$$\begin{aligned} \mathrm{logit}\, Pr(Y_i = 1) = &-1.00 - 0.32S_i + 0.28G_i - 0.58L_i \\ &+ 0.36(Age_i - 60)/10 + 0.16Sex_i + 0.18PC_{i1} - 0.10PC_{i2} + 0.08PC_{i3} \\ &+ 0.38I(C_i = \mathrm{tertiary}) + 0.16I(C_i = \mathrm{regional}). \end{aligned} \tag{D5}$$

Recruitment stream is generated after disease:

$$Pr(R_i = \text{clinic}) = \text{expit}[-0.75 + 0.85I(C_i = \text{tertiary}) + 0.45Y_i + 0.22S_i + 0.10(Age_i - 60)/10]. \quad \text{(D6)}$$

The case-control sample is drawn from a 108-cell sampling table over clinic, recruitment stream, ancestry, age band, and case status. For cases, the base sampling probabilities by clinic are 0.30, 0.46, and 0.72 for community, regional, and tertiary; stream multipliers are 0.78 for registry and 1.16 for clinic; ancestry multipliers are 1.08, 0.70, and 1.26 for EUR, AFR, and EAS; age-band multipliers are 0.74, 1.00, and 1.22. For controls, the corresponding clinic bases are 0.108, 0.060, and 0.022; stream multipliers are 1.22 and 0.50; ancestry multipliers are 0.84, 1.20, and 0.54; age-band multipliers are 1.34, 0.82, and 0.40. The product is capped at 0.95 and, if the initial draw exceeds 1,900 rows, deterministically thinned to 1,900 with the sampling fractions multiplied by the thinning fraction. In the realized released cohort, sample fractions range from 0.00211 to 0.84267.

Marker panel. Marker dosage is the sum of two haplotype-level alternate-allele draws plus $N(0, 0.040^2)$ noise, clipped to [0,2]:

$$M_{im} = \text{clip}\{Bernoulli(p_{H_{i1},A_i,m}) + Bernoulli(p_{H_{i2},A_i,m}) + \epsilon_{im}, 0, 2\}, \quad \epsilon_{im} \sim N(0, 0.040^2). \quad \text{(D7)}$$

Markers M001–M018 mostly tag global orientation ($p = 0.94$ for global haplotypes and 0.06 otherwise, with weaker AFR/EAS marker subsets and two AFR phase reversals). M019–M024 are proximal segment-B tags that are deliberately ancestry-nonportable: EUR has 0.86/0.14 for segment-B/non-segment-B, AFR has 0.30/0.70, and EAS has 0.60/0.40. M025–M030 are portable positive-phase segment-B tags (0.86/0.14), while M031–M036 are the same biological signal in opposite marker-allele phase (0.14/0.86). M037–M046 tag the recombinant D state (0.87/0.12, weakened to 0.66/0.20 in AFR). M047–M060 are ancestry-differentiated background markers with no direct segment-B role. This design makes raw block averages and marker PCs plausible but wrong.

Long-read breakpoint panel. The breakpoint panel contains 780 rows. Each selected row is drawn from one of three quality regimes. Reliable rows occur with probability 0.62:

$$N_{mol} \sim Poisson(28) + 18, \quad MAPQ \sim \text{clip}\{N(58.3, 1.2^2), 52, 60\}, \quad (6)$$

$$Entropy \sim \text{clip}\{N(0.075, 0.025^2), 0.01, 0.16\}, \quad (7)$$

$$Balance \sim \text{clip}\{N(0.84, 0.06^2), 0.64, 0.98\}, \quad (8)$$

$$AltFrac \sim \text{clip}\{G_i/2 + N(0, 0.035^2), 0, 1\}, \quad (9)$$

$$CopyB \sim \text{clip}\{S_i + N(0, 0.095^2), 0, 2\}, \quad (10)$$

$$Complexity \sim \text{clip}\{N(0.17, 0.035^2), 0.05, 0.30\}. \quad \text{(D8)}$$

Repeat-complex rows occur with probability 0.20. They retain high molecule counts and MAPQ but bias the structural summaries:

$$AltFrac \sim \text{clip}\{0.16 + 0.58G_i/2 + 0.12S_i + N(0, 0.12^2), 0, 1\}, \quad (11)$$

$$CopyB \sim \text{clip}\{0.18 + 0.34S_i + 0.58G_i + N(0, 0.18^2), 0, 2\}, \quad (12)$$

$$Complexity \sim \text{clip}\{N(0.58, 0.06^2), 0.43, 0.76\}. \quad \text{(D9)}$$

The remaining low-quality rows have low molecule count, low MAPQ, high entropy, poor pair balance, and high local complexity. For all regimes, segment-B depth is

$$DepthLog2_i = \log_2\{(2 + 0.46S_i)/2\} + \epsilon_i, \quad \text{(D10)}$$

where the noise SD is 0.035, 0.09, or 0.14 in the reliable, repeat-complex, or low-quality regimes.

Expression panel. The expression panel contains 560 selected samples, enriched to 45% segment-B carriers, with four genes per sample. Tissue source and batch are imbalanced with S_i:

$$Pr(Tissue_i = \text{B}) = \text{expit}[-0.85 + 0.90S_i + 0.45I(C_i = \text{tertiary})], \tag{13}$$

$$Pr(Batch_i = 2) = \text{expit}[-0.25 + 0.50I(Tissue_i = \text{B}) + 0.20S_i]. \tag{D11}$$

For gene g,

$$\begin{aligned} E_{ig} =& a_g + b_{Sg}S_i + b_{Gg}G_i + \tau_{g,T_i} + \kappa_{g,B_i} + 0.035(Age_i - 60)/10 \\ &+ 0.025Sex_i + 0.035PC_{i1} - 0.025PC_{i2} + A_{ig} + \epsilon_{ig}, \end{aligned} \tag{D12}$$

with $\epsilon_{ig} \sim N(0, 0.115^2)$. Baselines are 5.35, 4.75, 5.05, and 4.95 for GENE1–GENE4. Segment-B effects are 0.02, 0.305, 0.00, and 0.00. Global effects are 0.01, 0.02, -0.01, and 0.42. GENE1 has an additional low-uniqueness artifact $A_{i,GENE1} = 0.43S_i + 0.10G_i + N(0, 0.06^2)$ and median unique fraction 0.527. GENE2 has median unique fraction 0.926 and is the credible segment-B expression target; GENE4 has high uniqueness but is a global-orientation decoy.

7 Analyst Walkthrough

Step 1: start with the sampling frame, not a marker scan. The released cohort is visibly case-control ascertained: 1,544 of 1,900 rows are cases, for an unweighted case fraction of 0.8126. Merging `cohort.tsv.gz` to `sampling_design.tsv.gz` reveals why the clinical target cannot be the unweighted assayed-cohort contrast. The normalized design weights range from 0.204 to 33.449, and the weighted source-population case fraction is 0.3181. This is a scientific warning, not a report-only convention: the prompt asks for a source-population association and the sampling table supplies stratum-specific inclusion fractions.

```
cohort = read_tsv("cohort.tsv.gz")
sampling = read_tsv("sampling_design.tsv.gz")
df = cohort.merge(sampling, on=design_keys)
df["w"] = 1.0 / df.sample_fraction
source_case_rate = average(df.case, weights=df.w)
```

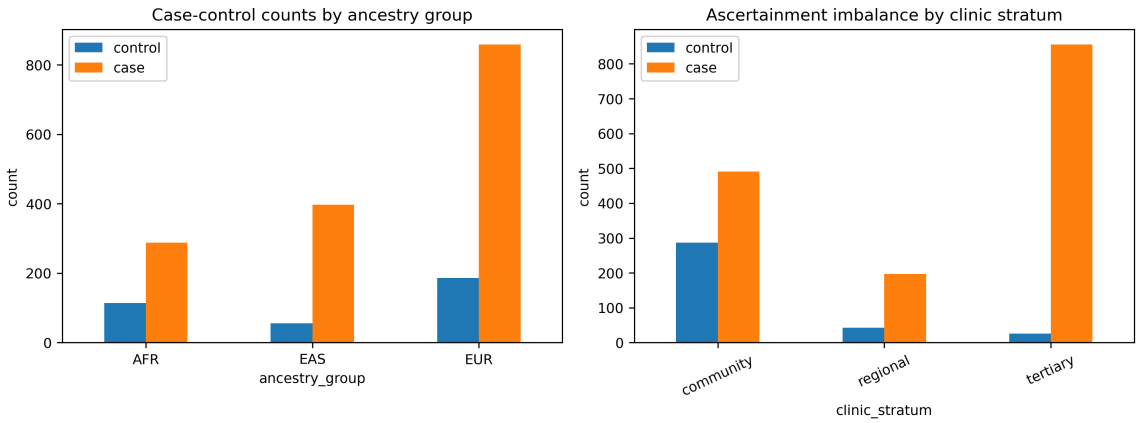

Figure 2: Initial cohort overview. Blue bars are controls and orange bars are cases. Both panels show raw released-cohort counts, first by ancestry group and then by clinic stratum. The released cohort over-represents cases and tertiary-clinic rows relative to the source-population frame, so unweighted clinical regression estimates the wrong population contrast.

This first step prevents a subtle downstream error. If an analyst begins with a marginal marker association scan, the strongest signals are partly driven by how cases and controls were sampled rather than by source-population risk. The sampling table is not optional metadata: it is the bridge from the released case-control cohort back to the requested population. That is why the later clinical regression uses inverse sampling fractions, while recruitment stream and age band remain sampling-design variables rather than extra clinical covariates.

Step 2: read the breakpoint panel as a calibration assay, not a source cohort. A naive analyst might round every `copy_index_segment_B` value and count carriers directly. That gives 376 carriers, which is nearly twice the correct 195. A more careful but still incomplete rule using molecule count, MAPQ, entropy, and pair balance leaves 643 rows and 292 carriers. The reason is visible in the realized data: repeat-complex rows have high molecule count and MAPQ like reliable rows, but their median `local_complexity` is 0.585 rather than 0.170, and their split-molecule fractions and copy indices are biased.

```
reliable = ((molecule_count >= 18) & (mean_mapq >= 50) &
            (alignment_entropy <= 0.18) & (pair_balance >= 0.62) &
            (local_complexity <= 0.35))
sub_lr = round(copy_index_segment_B).clip(0, 2)
global_lr = round(2 * bp_alt_fraction).clip(0, 2)
```

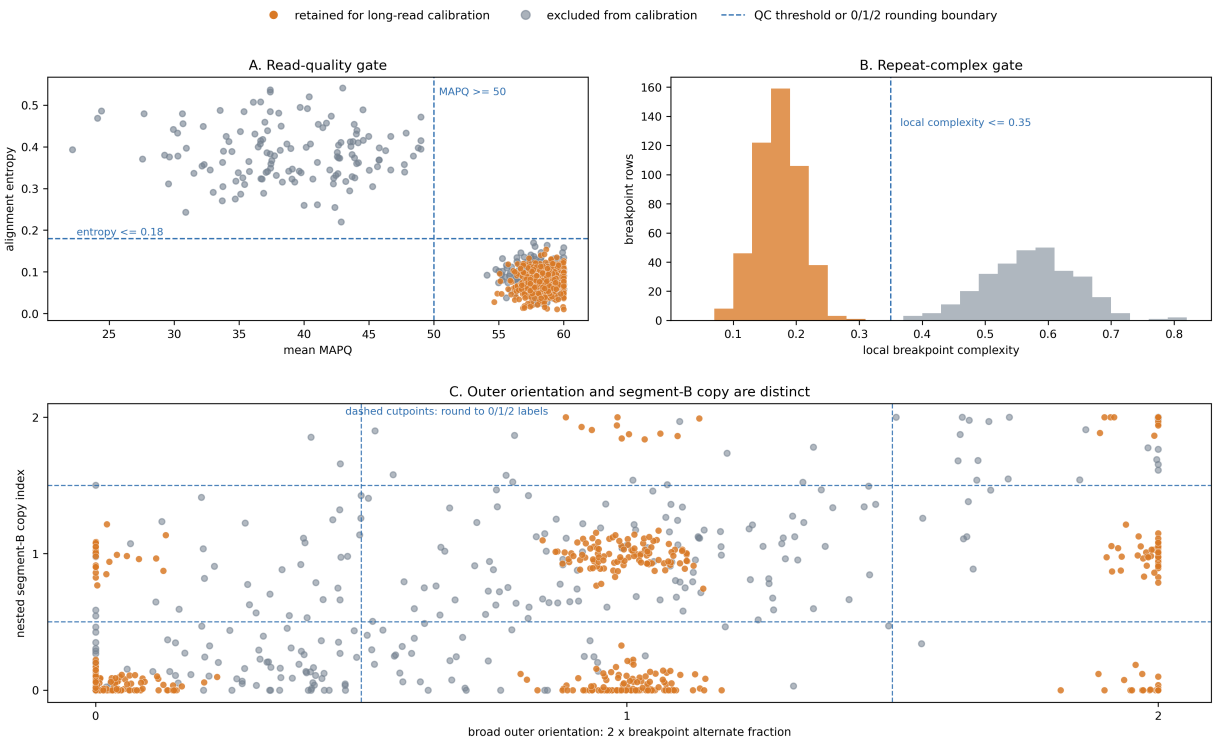

Figure 3: Breakpoint reliability and structural-label derivation diagnostics. Orange marks breakpoint-panel rows retained for long-read calibration and gray marks rows excluded from calibration; color is a calibration-use flag, not a biological class. Panels A and B show the visible QC gates, including the local-complexity threshold that removes repeat-complex rows with otherwise plausible read-quality summaries. The full retained-row rule is molecule count at least 18, mean MAPQ at least 50, alignment entropy at most 0.18, pair balance at least 0.62, and local complexity at most 0.35. Panel C plots the two raw structural summaries separately: broad outer orientation on the x-axis as 2× breakpoint alternate fraction and nested segment-B copy index on the y-axis. Dashed half-integer lines show the rounding boundaries used to form 0/1/2 global-orientation and segment-B calibration labels. The answer-changing counts are 483 retained calibration labels and 195 segment-B carriers with full QC, 292 carriers without local complexity, and 376 carriers if every breakpoint row is used.

The local-complexity threshold is a realized-data diagnostic rather than a hidden constant. The low-complexity reliable regime and high-complexity repeat regime are separated in the released panel, while ordinary MAPQ and molecule-count filters alone do not distinguish them. This is exactly the kind of breakpoint artifact expected in repeat-rich loci: reads can align confidently to the wrong local structure. The diagnostic consequence is visible before fitting any clinical model, because carrier counts jump from 195 to 292 or 376 depending on whether local complexity and reliability are ignored.

Step 3: calibrate nested segment-B and global orientation as separate continuous dosages. The long-read panel gives reliable labels only for a subset, while clinical and expression outputs require full-cohort exposures. The marker panel has enough information, but not as a single portable marker. M001–M018 mostly encode the broad outer orientation; M019–M024 are ancestry-nonportable; M025–M030 and M031–M036 encode the nested segment-B signal in opposite marker-allele phases. Therefore, a raw interval-2 mean collapses signal, a marker PC tracks the broader inversion, and a single top case marker mixes ancestry, global orientation, and disease ascertainment.

```
X = markers[M001:M060].join(cohort[["pc1","pc2","pc3"]])
for ancestry in ["EUR", "AFR", "EAS"]:
    train = reliable_lr_ids in ancestry
    fit RidgeCV(X[train], sub_lr[train])
    predict continuous subhap_dosage for all ancestry rows
```

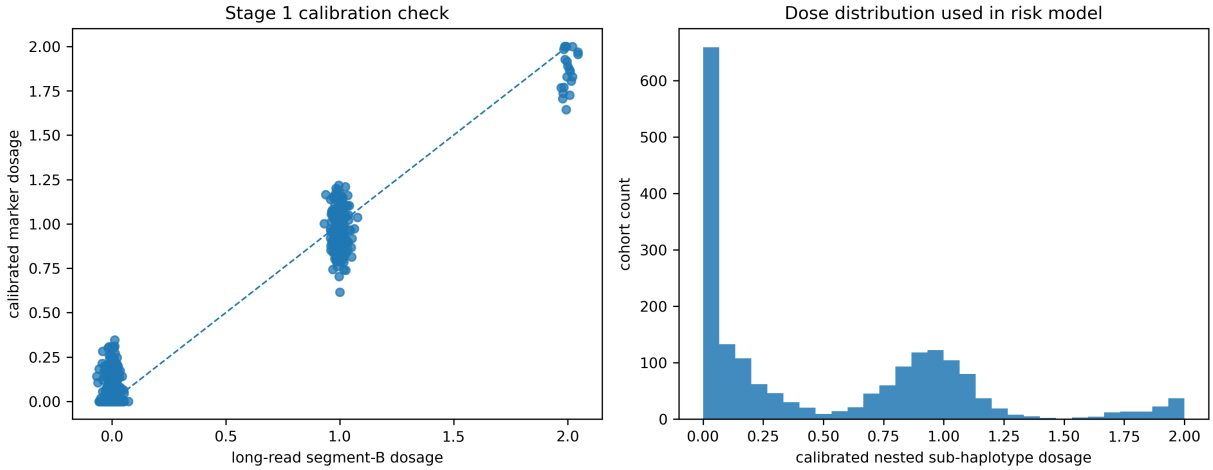

Figure 4: Nested-dosage calibration. The left panel compares reliable long-read segment-B copy labels on the x-axis with calibrated marker-based nested segment-B dosage on the y-axis; points near the dashed 1:1 line indicate agreement between the calibration labels and marker predictions. The right panel is the full-cohort distribution of calibrated nested segment-B dosage used in the clinical risk model. Standard linear and regularized calibration variants agree with reliable long-read labels and recover the target, whereas single-marker, PC, hard-call, and raw-block shortcuts estimate different biological proxies.

The calibration stage is deliberately tolerant of statistically standard implementations. The correct answer does not require one fixed ridge penalty: RidgeCV, ElasticNetCV, OLS with PCs, and OLS without PCs all land within tolerance once they learn the nested dosage from the reliable long-read labels. What fails are different estimands. A single top marker estimates a local marker association, a PC estimates a broad structural axis, a hard call discards continuous calibration, and a distal block mean ignores opposite marker phases. Those are scientifically different analyses, not harmless implementation variants.

The numerical consequences are large enough that this is not cosmetic modeling preference. A top marginal tag gives log OR 0.187080 and expression log FC 0.185875. A global-orientation proxy gives log OR 0.059563. The apparent segment-B block mean gives log OR 0.004450. The distal block mean, even with correct outer-orientation adjustment, overshoots to -0.712076 because opposite marker phases have not been learned from the calibration subset. In contrast, the reference RidgeCV route gives -0.400921, OLS with PCs gives -0.394421, OLS without PCs gives -0.373044, and ElasticNetCV with PCs gives -0.385183, all within the clinical tolerance.

Step 4: make expression support a molecular-read-quality decision. On a first pass, GENE1 is tempting: if read uniqueness is ignored, it has expression log FC 0.474367. But its median unique fraction is only 0.527 and median multimapper fraction is 0.384, consistent with a paralog-like expression artifact in a structurally complex locus. GENE4 is the opposite problem: it has high uniqueness (median 0.911) but is driven by global orientation, not nested segment-B copy. The credible gene is GENE2, with median unique fraction 0.926 and adjusted nested-dosage

coefficient 0.333341.

```
for gene in expression.gene_id.unique():
    keep = median(unique_fraction) >= 0.80
    keep &= median(multimapper_fraction) <= 0.16
    fit log_expr ~ subhap + global + age + sex + PCs + tissue + batch
select high-uniqueness gene with largest |t_subhap|
```

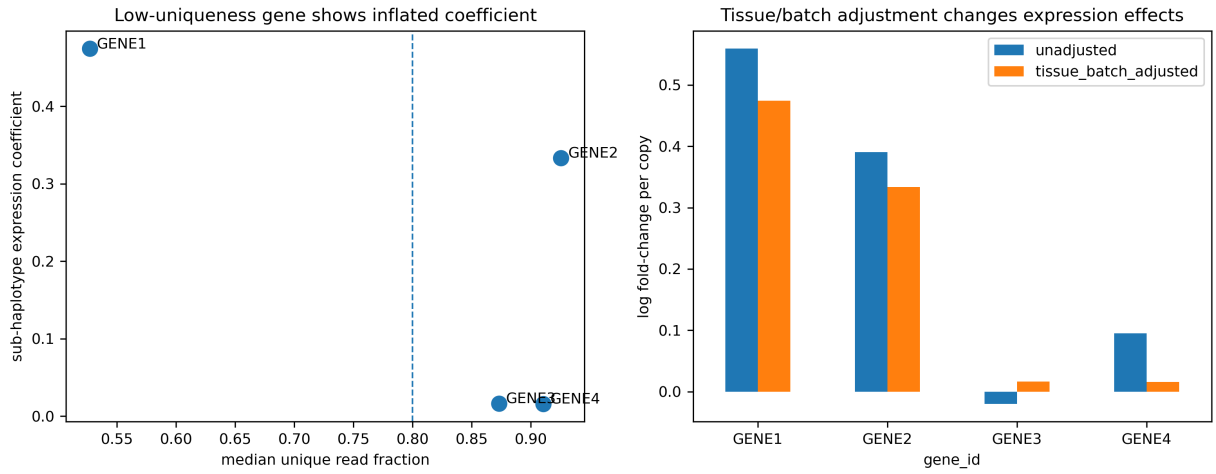

Figure 5: Expression evidence. In the left panel, each labeled point is one candidate gene after covariate adjustment; the x-axis is median unique-read fraction and the y-axis is the nested segment-B expression coefficient. The dashed vertical line is the public estimator's unique-read eligibility threshold of 0.80. GENE1 has a large apparent coefficient but low uniqueness; GENE4 is a high-uniqueness global-orientation decoy. In the right panel, blue bars are expression coefficients before tissue/batch adjustment and orange bars are coefficients after adding tissue source and expression batch covariates. After read-uniqueness filtering and tissue/batch adjustment, GENE2 is the expression-supported segment-B gene.

This stage also explains why molecular support is reported separately from the clinical effect. The expression result is a biological consistency check for the nested dosage, not a covariate to condition on when estimating the total clinical association. If expression is inserted into the outcome model, the clinical coefficient changes from -0.400921 to -0.317826 because the model is now closer to a direct-effect contrast. The prompt asks for the source-population total effect, so the correct report gives expression support alongside the clinical association rather than adjusting the clinical association away.

The tissue and batch adjustments are also answer-affecting. Omitting them gives expression log FC 0.390760, outside the 0.030 tolerance. Using global orientation as the expression predictor gives 0.437580 and selects GENE4. Both wrong answers look biologically plausible if the analyst stops at broad inversion status, which is why the expression model must preserve the nested/global distinction from Stage 1.

Step 5: estimate the clinical total effect on the source-population scale. The final clinical model uses calibrated nested dosage, calibrated global orientation, age, sex, PCs, clinic stratum, ancestry group, and inverse sampling-fraction weights. It does not add recruitment stream or age band as ordinary clinical covariates because those fields define the sampling fractions. It also does not adjust for expression: expression is a downstream molecular support readout, so conditioning on it changes the requested total effect into a direct-effect-like quantity.

```
df = attach_sampling_weights(cohort_with_dosage, sampling_design)
X = [subhap, global, age10, sex, pc1, pc2, pc3,
     clinic indicators, ancestry indicators]
fit GLM(case ~ X, family=Binomial(), weights=1/sample_fraction)
subhap_log_or = coef["subhap"]
```

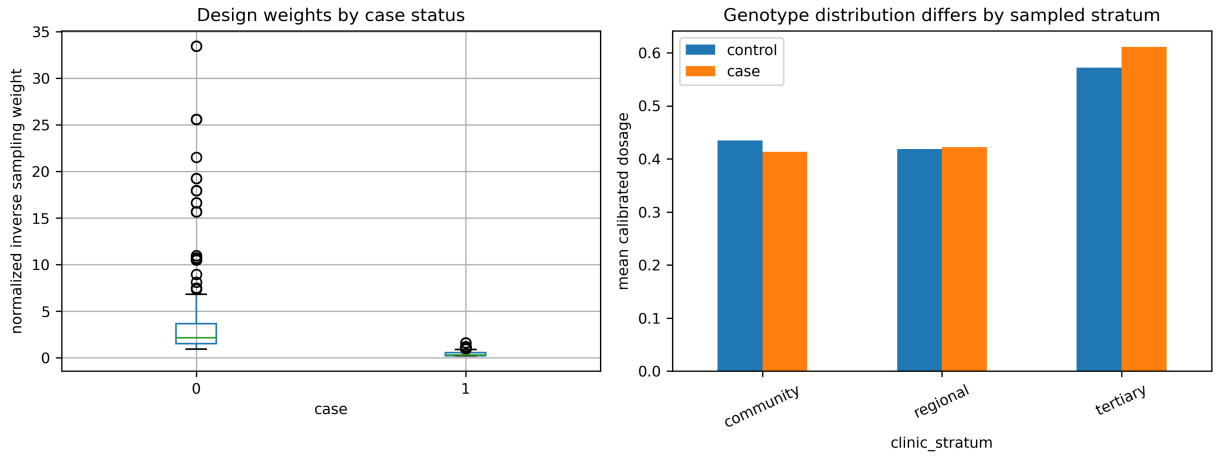

Figure 6: Clinical weighting diagnostics. The left panel is a boxplot of normalized inverse sampling weights by case status, where boxes show the interquartile range, the center line is the median, whiskers show the non-outlier range, and open circles are high-weight outlier rows. The right panel shows mean calibrated nested segment-B dosage by clinic stratum, with blue bars for controls and orange bars for cases. The source-population contrast differs from the assayed-cohort contrast because inclusion depends on case status, clinic, recruitment stream, ancestry, and age band.

The key wrong basins illustrate the estimand distinction. The unweighted outcome model gives -0.150075. Adding expression to the outcome model gives -0.317826. Adding recruitment stream as an extra outcome covariate with otherwise correct dosage gives -0.483524, which is a different conditional contrast rather than the source-population total effect requested by the prompt. The reference design-weighted model gives -0.400921.

The final clinical step is the only stage that uses the entire 1,900-person cohort. The breakpoint panel supplies calibration labels, and the expression subset supplies molecular support, but neither subset is itself the target clinical population. The clinical answer therefore depends on propagating calibrated dosage into the full sampled cohort and then applying the sampling weights. This is the point at which many shortcut analyses failed: they correctly counted 195 reliable breakpoint carriers and often selected GENE2, but they still reported a shallower clinical coefficient because they used hard calls or shallow dosage scores rather than the calibrated continuous exposure.

Step 6: assemble the answer and check tolerance-stable alternatives. The final accounting is:

$$\begin{aligned}
\widehat{\beta}_S &= -0.400921 && \Rightarrow \texttt{subhap_log_or},\\
\widehat{\theta}_{GENE2} &= 0.333341 && \Rightarrow \texttt{expression_log_fc},\\
I(\widehat{\beta}_S < 0, \widehat{\theta}_{GENE2} > 0) &= 1 && \Rightarrow \texttt{target_support_code},\\
\sum_{i \in LR} I(reliable_i = 1, S_i^{LR} \geq 1) &= 195 && \Rightarrow \texttt{n_calibrated_carriers}.
\end{aligned}$$

The sensitivity table is important: the answer is not tied to a fixed RidgeCV alpha. OLS with PCs, OLS without PCs, RidgeCV with PCs, and ElasticNetCV with PCs all recover the same target within tolerance. What fails are shortcut estimands: global proxy, hard calls, raw block means, unreliable breakpoint rows, unweighted clinical regression, mediator adjustment, and expression artifacts.

8 Estimand

Let $\mathcal{D}$ denote the released files. The clinical estimand is the finite-sample recoverable source-population total log odds ratio for one additional calibrated nested segment-B copy:

$$\psi(\mathcal{D}) = \beta_S \quad \text{in} \quad \operatorname{logit} Pr(Y_i = 1 \mid \widehat{S}_i, \widehat{G}_i, W_i) = \alpha + \beta_S \widehat{S}_i + \beta_G \widehat{G}_i + \gamma^T W_i, \tag{E1}$$

fit with normalized design weights proportional to $1/\pi_i$. W_i includes baseline demographic, ancestry, and clinic variables: age, sex, PCs, ancestry group, and clinic stratum. The source-population target uses recruitment stream and age band through π_i; they are not extra baseline covariates in W_i.

The expression estimand for gene g is the nested-dosage coefficient in

$$E_{ig} = a_g + \theta_g \widehat{S}_i + \eta_g \widehat{G}_i + \lambda_g^T W_i + \delta_g^T Tissue_i + \kappa_g^T Batch_i + e_{ig}, \tag{E2}$$

after excluding genes without credible read uniqueness. The reported molecular output is θ_{g^*} for the eligible gene with the strongest nested-dosage evidence. The support code is a sign-concordance summary, not an independent biological estimand. The carrier count is the reliable breakpoint-panel carrier count, not a full-cohort carrier estimate.

9 Estimator

The estimator is a composite of standard components: structural-variant calibration from reliable long-read evidence, marker-based dosage prediction, expression regression with read-mappability and batch/tissue adjustment, inverse-probability weighting for biased sampling, and a total-effect interpretation that avoids conditioning on a downstream mediator [3,4,7–9]. Breakpoint context, alignment ambiguity, and sequence complexity affect SV calls, so the validator first thresholds molecule count, MAPQ, entropy, pair balance, and local complexity before using breakpoint summaries as calibration labels [3,4].

Structural haplotypes often require panels rather than a single portable tag, especially across ancestries [3,6]. The reference analysis therefore fits ancestry-stratified continuous dosage models for nested segment B and global orientation from reliable long-read labels, marker dosages, and PCs. Inversion expression effects can be gene- and tissue-specific, and mapping uniqueness matters in complex regions [1,5], so the molecular stage filters low-uniqueness genes and fits adjusted OLS expression models with nested dosage, global dosage, tissue, batch, and baseline covariates.

The clinical stage targets a source-population contrast rather than the assayed-cohort contrast. Inverse-probability weighting recovers that target under the known sampling and observation mechanisms [8,9], so the final clinical model uses a binomial GLM with weights proportional to inverse sampling fraction and with baseline and structural covariates. Expression is reported as molecular support, but it is excluded from the clinical total-effect model because conditioning on a downstream mediator would change the exposure coefficient's interpretation [7,9].

Stage 1: reliable structural labels and calibrated dosage. The breakpoint reliability indicator is

$$R_i^{LR} = I(N_{mol,i} \geq 18,\ MAPQ_i \geq 50,\ Entropy_i \leq 0.18,\ Balance_i \geq 0.62,\ Complexity_i \leq 0.35). \tag{H1}$$

For reliable rows, labels are

$$S_i^{LR} = \text{round}(CopyB_i), \qquad G_i^{LR} = \text{round}(2AltFrac_i), \tag{H2}$$

both clipped to 0, 1, or 2. Within each ancestry group, standardized marker dosages and PCs are fit to S_i^{LR} and G_i^{LR} among reliable rows using RidgeCV over alphas $\{0.01, 0.05, 0.1, 0.35, 1, 3, 10\}$. Predictions are clipped to [0,2]. The output is continuous calibrated dosage, not a hard copy-state call.

Stage 2: expression support. For each gene, compute median uniqueness and multimapper fraction. Eligible genes satisfy

$$\text{median}(unique\_fraction_g) \geq 0.80, \qquad \text{median}(multimapper\_fraction_g) \leq 0.16. \tag{H3}$$

For eligible genes, fit Eq. E2 by OLS and select the gene with largest absolute nested-dosage t statistic. The support-code positivity requirement additionally requires $\theta_g > 0$ and adequate unique-read support.

Stage 3: clinical model. Merge cohort rows to the sampling table on clinic, stream, ancestry, age band, and case. Use

$$\tilde{w}_i = \frac{1/\pi_i}{n^{-1}\sum_j 1/\pi_j}. \tag{H4}$$

Fit the weighted binomial likelihood

$$\ell(\beta) = \sum_i \tilde{w}_i\{Y_i\eta_i - \log(1+\exp(\eta_i))\}, \quad \eta_i = \alpha + \beta_S\widehat{S}_i + \beta_G\widehat{G}_i + \gamma^T W_i. \tag{H5}$$

The fitted $\widehat{\beta}_S$ is the reported clinical log OR. Expression rows are not joined into this clinical model because the requested output is a total effect.

10 Decision-Point and Ablation Walkthrough

The table below combines the full ablation outputs with the decision point each shortcut fails. It also labels target-equivalent implementation variants from the sensitivity analysis as accepted rows, so they are visibly distinct from incorrect shortcuts. The released answer contract checks the clinical log odds ratio, the expression log fold change, a support code, and the calibrated carrier count; a row fails if any reported field violates its tolerance or exact-match requirement.

Decision point	Analysis variant	Quantitative output	Pass?	Failure point	Why the approach is wrong
Reference pipeline	correct	log OR -0.400921, log FC 0.333341, support 1, carriers 195	yes	none	Reference nested-copy calibration, weighted clinical model, and high-uniqueness expression support.
Accepted variant	OLS + PCs	log OR -0.394421, log FC 0.325151, support 1, carriers 195	yes	tolerance-equivalent	Standard ancestry-stratified linear dosage calibration with PCs recovers all public answer fields within tolerance.
Accepted variant	OLS, no PCs	log OR -0.373044, log FC 0.326148, support 1, carriers 195	yes	tolerance-equivalent	Omitting PCs in the ancestry-stratified calibration is still answer-equivalent for this realized panel.
Accepted variant	RidgeCV + PCs	log OR -0.400921, log FC 0.333341, support 1, carriers 195	yes	tolerance-equivalent	Cross-validated ridge with PCs matches the reference implementation exactly at reported precision.
Accepted variant	ElasticNetCV + PCs	log OR -0.385183, log FC 0.344197, support 1, carriers 195	yes	tolerance-equivalent	Sparse regularized calibration is accepted because all fields remain inside the released tolerances.
Structural proxy	top tag case marker	0.187080, 0.185875, support 0, carriers 195	no	Exposure calibration	A single array tag mixes ancestry, global orientation, and local haplotype background.
Structural proxy	global orientation proxy	0.059563, 0.218790, support 0, carriers 195	no	Exposure calibration	Broad orientation is not the nested segment-B copy exposure.
Structural proxy	marker-PC proxy	0.095368, 0.253726, support 0, carriers 195	no	Exposure calibration	Marker PCs recover broad inversion/ancestry structure rather than phase-aware nested dosage.
Structural proxy	mean internal sub-block	0.004450, 0.270892, support 0, carriers 195	no	Marker phase	Raw block means collapse opposite-phase segment-B tags.
Structural proxy	distal block mean + outer	-0.712076, 0.028038, support 1, carriers 195	no	Marker phase	Distal block averaging is misphased and overcorrected by the outer-orientation covariate.
Structural proxy	proximal block mean	0.185884, 0.124031, support 0, carriers 195	no	Marker phase	Early block tags are not a portable nested-copy dosage.
Structural proxy	proximal block calibration	0.392819, 0.262108, support 0, carriers 195	no	Calibration subset	Calibrates the wrong marker block and therefore preserves a non-target signal.
Long-read shortcut	best single long-read marker	0.094402, -0.087495, support 0, carriers 195	no	Breakpoint calibration	A single long-read marker is not a stable copy-dosage estimator.
Structural shortcut	pooled block classifier	-0.124377, 0.326818, support 1, carriers 195	no	Exposure calibration	Pooled classifier partially captures expression support but misses the clinical nested-copy contrast.

Decision point	Analysis variant	Quantitative output	Pass?	Failure point	Why the approach is wrong
Calibration shortcut	reliable long-read subset	-0.535009, 0.301492, support 1, carriers 195	no	Cohort transport	Uses only reliable long-read carriers rather than calibrated dosage in the full cohort.
Structural proxy	global block hard call	0.320805, 0.256320, support 0, carriers 195	no	Exposure calibration	Hard-calls broad structural state instead of nested segment-B dosage.
Breakpoint QC	omit local-complexity QC	-0.293015, 0.361442, support 1, carriers 292	no	Reliability filtering	Keeps high-MAPQ but locally complex breakpoint rows with biased copy summaries.
Breakpoint QC	no long-read QC	-0.331252, 0.361059, support 1, carriers 376	no	Reliability filtering	Treats all breakpoint rows as reliable and overcounts carriers.
Clinical model	unweighted outcome model	-0.150075, 0.333341, support 1, carriers 195	no	Sampling transport	Ignores ascertainment weights and estimates the sampled cohort contrast.
Clinical estimand	mediator-adjusted outcome	-0.317826, 0.333341, support 1, carriers 195	no	Total-effect model	Conditions on downstream expression and changes the target to a direct-effect-like coefficient.
Clinical estimand	risk + stream covariate	-0.483524, 0.333341, support 1, carriers 195	no	Sampling transport	Adds recruitment stream as an outcome covariate even though stream already defines the sampling fractions, changing the requested source-population contrast.
Expression support	ignore expression uniqueness	-0.400921, 0.474367, support 0, carriers 195	no	Expression QC	Selects a low-uniqueness paralog-like expression artifact.
Expression support	omit tissue/batch adjustment	-0.400921, 0.390760, support 1, carriers 195	no	Expression adjustment	Leaves tissue and processing-batch imbalance in the molecular-support model.
Expression support	global-orientation expression model	-0.400921, 0.437580, support 1, carriers 195	no	Exposure specificity	Uses global orientation as the expression predictor rather than nested segment-B dosage.
Compound shortcut	no-QC unweighted model	-0.205359, 0.390238, support 1, carriers 376	no	Multiple stages	Combines unreliable breakpoint rows, expression imbalance, and unweighted clinical modeling.

Table 3: Unified decision-point and ablation walkthrough for the structural-inversion nested-haplotype problem.

11 References

1. Bowles KR, et al. 17q21.31 sub-haplotypes underlying H1-associated risk for Parkinson's disease are associated with LRRC37A/2 expression in astrocytes. *Molecular Neurodegeneration*. 2022. DOI: 10.1186/s13024-022-00551-x. PMID: 35841044.

2. Espinosa I, et al. Neolithic expansion and the 17q21.31 inversion in Iberia: an evolutionary approach to H2 haplotype distribution in the Near East and Europe. *Molecular Genetics and Genomics*. 2023. DOI: 10.1007/s00438-022-01969-0. PMID: 36355195.

3. Porubsky D, et al. Inversion polymorphism in a complete human genome assembly. *Genome Biology*. 2023. DOI: 10.1186/s13059-023-02919-8. PMID: 37122002.

4. Ahsan MU, et al. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nature Methods*. 2023. DOI: 10.1038/s41592-023-01932-w.

5. de Jong S, et al. Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. *BMC Genomics*. 2012. DOI: 10.1186/1471-2164-13-458.

6. Nguyen DT, et al. A comprehensive evaluation of polygenic score and genotype imputation performances of human SNP arrays in diverse populations. *Scientific Reports*. 2022. DOI: 10.1038/s41598-022-22215-y. PMID: 36266455.

7. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*. 2013. DOI: 10.1093/ije/dyt127.

8. Seaman SR and White IR. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*. 2013. DOI: 10.1177/0962280210395740.

9. Hernan MA and Robins JM. *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC, 2020. Official book page: https://miguelhernan.org/whatifbook.