GeneBench-Pro Case Study: Multi-parent QTL Mapping with Founder Reconstruction

GeneBench-Pro

June 26, 2026

1 Overview

Multi-parent recombinant populations localize quantitative trait loci (QTLs) by observing mosaics of founder haplotypes, but routine analysis files often expose only biallelic marker calls rather than the founder states needed for a founder-specific association scan. Heterogeneous-stock, MAGIC, and Diversity Outbred designs are concrete examples where founder mosaics support QTL fine-mapping [1, 5, 4]. This case study uses 120 recombinant inbred lines, an eight-founder marker panel, 220 markers across two chromosomes, and a quantitative phenotype with a batch covariate to illustrate the practical bottleneck. The required output is the chromosome-1 genetic-map position, in centimorgans, and the founder carrying the high-effect allele.

Three data-supported decisions determine the recoverable target. First, marker IDs must be aligned across the marker table, founder panel, and genotype matrix, then sorted by chromosome and map position. Second, the biallelic observations must be converted into founder-state probabilities using a multipoint hidden Markov model (HMM), a standard component of multi-parent QTL analysis with a generic formulation described by Broman and implemented in R/qtl2 [2, 3]. That same HMM pass reveals two local allele-orientation outliers near the QTL region. Third, the association scan must include the provided batch covariate, because batch is strongly aligned with local ancestry near 20 cM and otherwise produces a plausible nuisance peak. Applying these decisions reports marker `m1_053`, position 48.635551 cM, and founder `F5`, matching the two public fields `qtl_pos_cM` and `high_founder`.

2 Released Prompt and Files

Prompt

```
Map the chromosome 1 QTL in an 8-founder multi-parent population. Report the position (cM) and which founder
    carries the high-effect allele.
Report high_founder as "F1".."F8".
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": {
"high_founder": "<string>",
"qtl_pos_cM": <float>
},
"reasoning": "<description of method and QC>"
}
```

Released data files

File	Format	Contents
`markers.tsv.gz`	TSV.gz	Marker ID, chromosome, and centimorgan position for 220 markers. Rows are deliberately not in map order.
`founders.tsv.gz`	TSV.gz	Founder alleles F1–F8 at the same marker IDs.
`ril_genotypes.npz`	NPZ	A 120 by 220 matrix of line genotypes, with 0/1 biallelic calls and -1 for missing calls, plus sample and marker IDs.
`phenotypes.tsv.gz`	TSV.gz	Sample ID, quantitative phenotype, and binary batch covariate.

3 Answer Fields and Tolerances

The public answer contains two fields. The numeric field `qtl_pos_cM` is the centimorgan coordinate of the selected chromosome-1 marker $k^\star$ after founder reconstruction, orientation repair, and batch adjustment. The categorical field `high_founder` is the founder label $f^\star$ whose fitted founder-dosage coefficient is positive at that selected marker.

Answer field	Ground truth	Tolerance / matching rule	Interpretation
`qtl_pos_cM`	48.635551	Absolute error ≤ 3.0 cM; valid range $[0, 100]$	Chromosome-1 QTL position after founder reconstruction, marker-orientation repair, and batch adjustment.
`high_founder`	`F5`	Exact case-sensitive string match	Founder with positive fitted dosage effect at the selected marker.

4 Structure Diagram

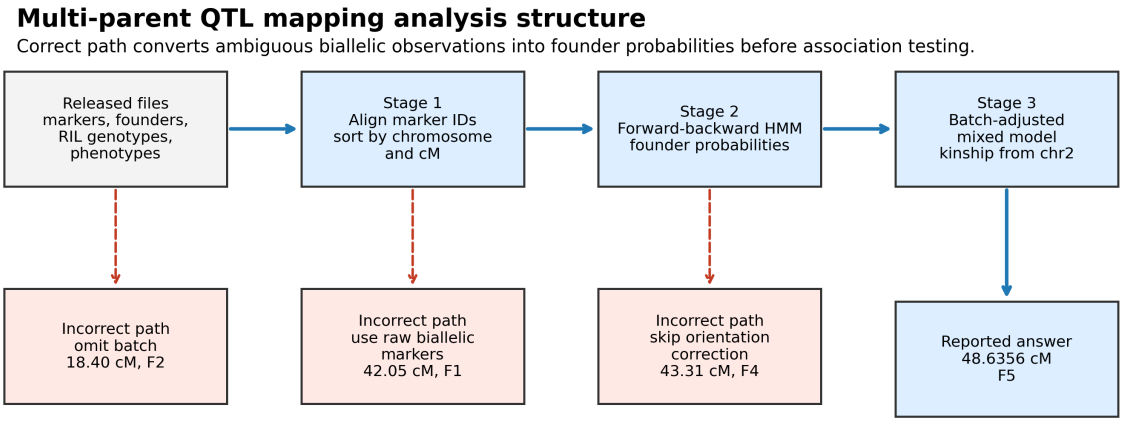

Figure 1: Analysis structure. The gray input box shows the released files for recombinant inbred lines (RILs), eight founders (F1–F8), marker positions in centimorgans (cM), phenotypes, and batch labels. Light-blue boxes and solid blue arrows trace the required workflow: align and sort the marker map, use a hidden Markov model (HMM) to convert biallelic calls into founder probabilities, repair the two local orientation outliers, and scan with batch included. Pink boxes and dashed red arrows mark incomplete workflows and their wrong chromosome-1 outputs; the solid blue path is the workflow that recovers the founder-specific quantitative trait locus (QTL).

5 Variables and Assumptions

Lines are indexed by i, ordered markers by k, and founders by $f \in \{\mathrm{F1}, \ldots, \mathrm{F8}\}$. The released genotype matrix contains biallelic calls $G_{ik} \in \{0, 1\}$ and missing values coded as -1; founder alleles A_{fk} define how each founder would appear at each marker. The inferential state is the hidden founder ancestry S_{ik}, summarized by posterior founder dosages $\gamma_{ikf} = P(S_{ik} = f \mid G_i, A, M)$. The problem asks for the realized-data founder-state association peak, so the answer is determined by the released marker map, founder panel, line genotypes, phenotypes, and batch covariate after the documented orientation repair.

6 Data-Generating Process

The synthetic data contain 120 recombinant inbred lines, 8 founders, and 220 markers split evenly across two chromosomes. On each chromosome, marker positions are generated near an evenly spaced centimorgan grid and then shuffled before release. Founder alleles are locally block-correlated, so nearby markers carry shared haplotype information rather than independent binary labels; this is what makes multipoint founder reconstruction informative.

The latent founder state S_{ik} for line i at marker k follows a chromosome-wise Markov chain. If adjacent markers are separated by ΔM_k Morgans, then

$$P(S_{ik} \neq S_{i,k-1}) = 1 - \exp(-4.0\,\Delta M_k), \tag{1}$$

with a switch drawing uniformly from the seven non-current founders. The observed genotype is

the founder allele with 1 percent genotyping error and 5 percent missingness:

$$\widetilde{G}_{ik} = \begin{cases} 1 - A_{S_{ik},k}, & E_{ik} = 1, \\ A_{S_{ik},k}, & E_{ik} = 0, \end{cases} \qquad E_{ik} \sim \text{Bernoulli}(0.01), \tag{2}$$

$$G_{ik} = \begin{cases} -1, & R_{ik} = 1, \\ \widetilde{G}_{ik}, & R_{ik} = 0, \end{cases} \qquad R_{ik} \sim \text{Bernoulli}(0.05). \tag{3}$$

Two markers near the QTL region, `m1_056` and `m1_058`, are complemented in the line genotype matrix but not in the founder panel. This models a small local orientation error, a familiar genotype-integration problem in which allele or strand conventions are not harmonized across files [6, 7]. The diagnostic is empirical: these two nearby markers are isolated posterior-mismatch outliers after the first HMM pass.

Batch is constructed from local ancestry near chromosome 1 at 20 cM:

$$B_i = \mathbf{1}\{S_{i,k_{20}} \in \{\text{F1}, \text{F2}, \text{F3}, \text{F4}\}\}, \qquad \text{batch effect}_i = 3.0 B_i - 1.5. \tag{4}$$

The major QTL is at the chromosome-1 marker nearest 48.7 cM and gives founder F5 an additive effect:

$$\text{QTL}_i = 2.0\, \mathbf{1}\{S_{i,k_\star} = \text{F5}\}. \tag{5}$$

Chromosome 2 contributes a polygenic background through a state-sharing covariance matrix. The phenotype is

$$Y_i = 3.0 + (3.0 B_i - 1.5) + 2.0\, \mathbf{1}\{S_{i,k_\star} = \text{F5}\} + u_i + \epsilon_i, \qquad \epsilon_i \sim N(0, 1). \tag{6}$$

The scientifically relevant decisions are map alignment, founder reconstruction, orientation correction, and batch adjustment; sample retention does not determine the reported answer in this realized dataset.

7 Analyst Walkthrough

7.1 Establish the map

The first decision is file alignment. The raw marker table has 111 adjacent chromosome-position order violations, so file order cannot be treated as map order. Because an HMM transition model interprets adjacent rows as adjacent recombination intervals, the correct preprocessing step is to align all files by `marker_id`, then sort by chromosome and `pos_cM`. This map-order violation is the relevant initial data check; all 120 released lines are retained, and no sample-retention step drives the answer.

7.2 Reject the direct biallelic scan

A marker-wise ordinary least-squares regression on the raw observed 0/1 line genotypes gives a concrete but misleading position. The all-marker direct biallelic shortcut selects 42.05 cM and maps the binary allele to F1, a label not supported by founder-state inference; a chromosome-1-only biallelic profile peaks at 13.80 cM and cannot justify a founder label at all. This is the analysis one would get by treating `ril_genotypes.npz` as an ordinary marker matrix. Its failure is conceptual rather than numerical. At a multi-parent marker, the observed allele is a many-to-one

measurement: several founders can carry allele 0, and several can carry allele 1. Founder identity is latent and must be inferred from surrounding markers.

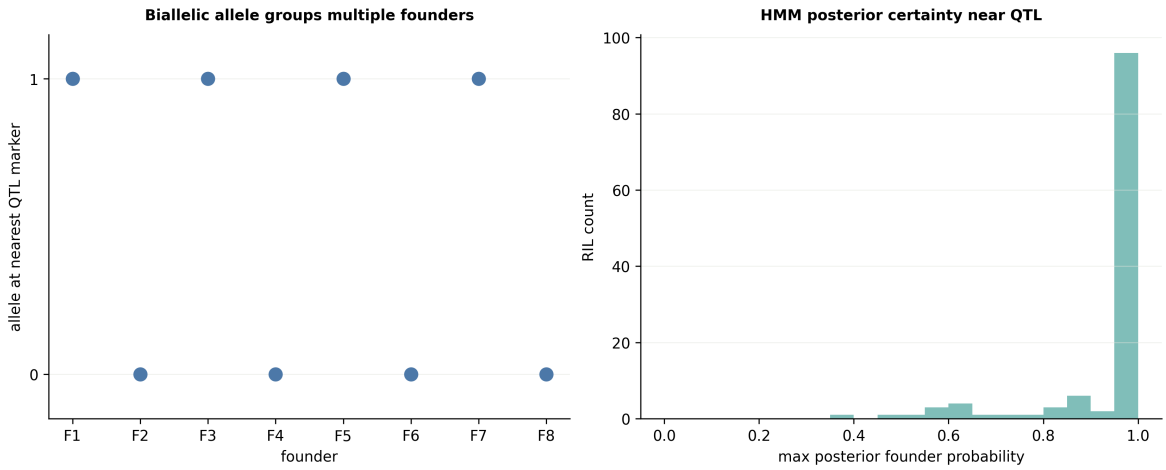

Figure 2: Founder reconstruction evidence. At the nearest quantitative trait locus (QTL) marker, the left panel shows that the observed biallelic allele value, 0 or 1, groups multiple founders (F1–F8) and therefore cannot identify a founder by itself. The right-panel teal histogram counts recombinant inbred lines (RILs) by their maximum hidden Markov model (HMM) posterior founder probability near the QTL after ordering markers by chromosome position, showing that multipoint founder reconstruction supplies the founder-state evidence missing from a direct biallelic scan.

7.3 Use HMM posteriors to detect orientation outliers

The HMM uses founder alleles, line genotypes, and map distance to compute $\gamma_{ikf} = P(S_{ik} = f \mid G_i, A, M)$, the posterior probability that line i inherited founder f at marker k. A first pass also gives a marker-level posterior mismatch diagnostic:

$$\mathrm{mm}_k = \frac{1}{n_k} \sum_{i:G_{ik}\neq -1} \sum_{f=1}^{8} \gamma_{ikf} \mathbf{1}\{A_{fk} \neq G_{ik}\}. \tag{7}$$

Most markers have low expected mismatch. Two nearby chromosome-1 markers are clear outliers: `m1_056` at 51.3985 cM has mismatch 0.4311, and `m1_058` at 53.2289 cM has mismatch 0.4123. The largest non-flipped mismatch is 0.0483. Complementing those two non-missing genotype columns and rerunning the HMM changes the downstream scan from the no-correction result, 43.31 cM/F4, to the target region. The absolute mismatch values depend on the HMM's assumed genotyping error and switching rate, but across plausible HMM settings the same two markers remain well separated from the largest non-outlier marker. The stable observation is therefore the outlier gap itself, not a fragile numerical cutoff.

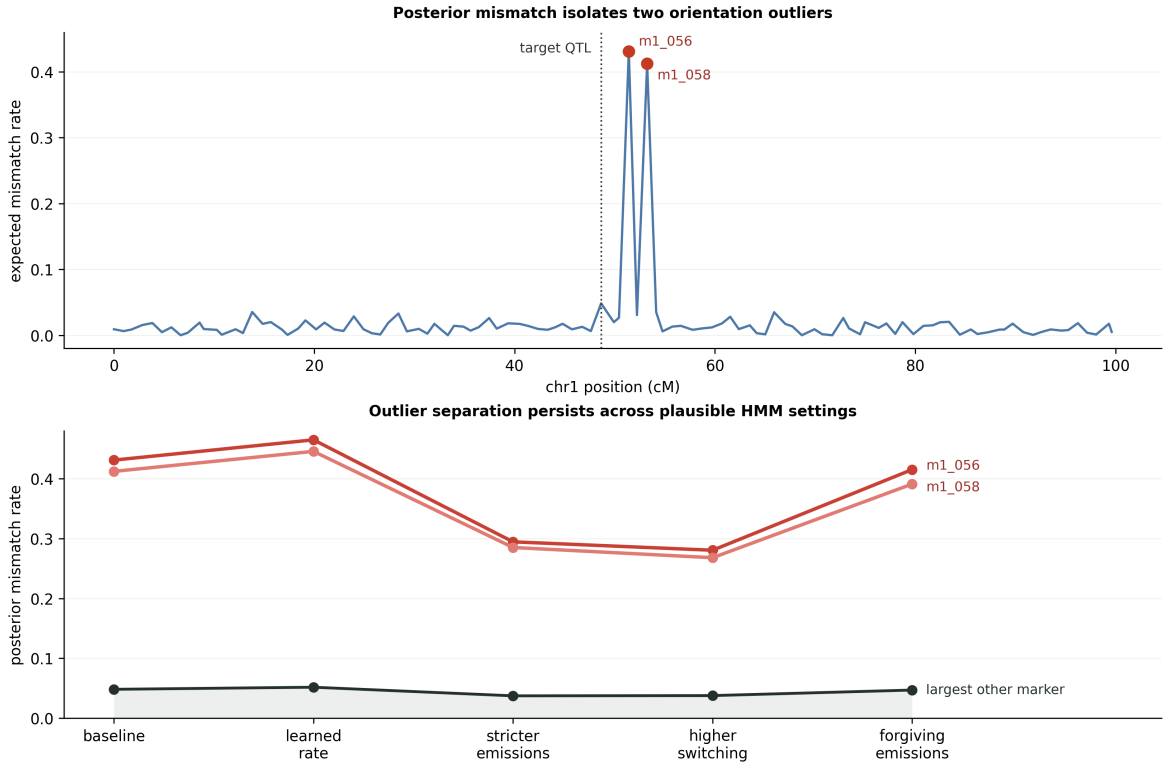

Figure 3: Allele-orientation diagnostic. The upper blue line is the expected posterior mismatch rate for each chromosome-1 marker after the first hidden Markov model (HMM) pass; the dotted vertical reference line marks the target quantitative trait locus (QTL) position, and the red points label the two complemented orientation-outlier markers, `m1_056` and `m1_058`. In the lower sensitivity panel, the red and salmon lines are the same two outlier markers across plausible HMM settings, while the dark line and gray band show the largest non-outlier marker for comparison. The correction is supported by the persistent outlier gap, not by a finely tuned mismatch cutoff.

7.4 Explain the 20 cM peak before accepting it

After founder reconstruction and orientation correction, an association scan that omits batch selects `m1_020` at 18.40 cM with founder F2. This signal is superficially plausible because it is a local ancestry association on the target chromosome, not an obviously nonsensical result. The released data explain why it is not the target QTL. At `m1_022`, the posterior probability of belonging to founder group F1–F4 averages 0.0657 in batch 0 and 0.9720 in batch 1, with Pearson correlation 0.950 against batch. Phenotype means also differ by batch, 1.921 in batch 0 and 4.207 in batch 1. The 20 cM signal is therefore a batch-aligned local-ancestry peak.

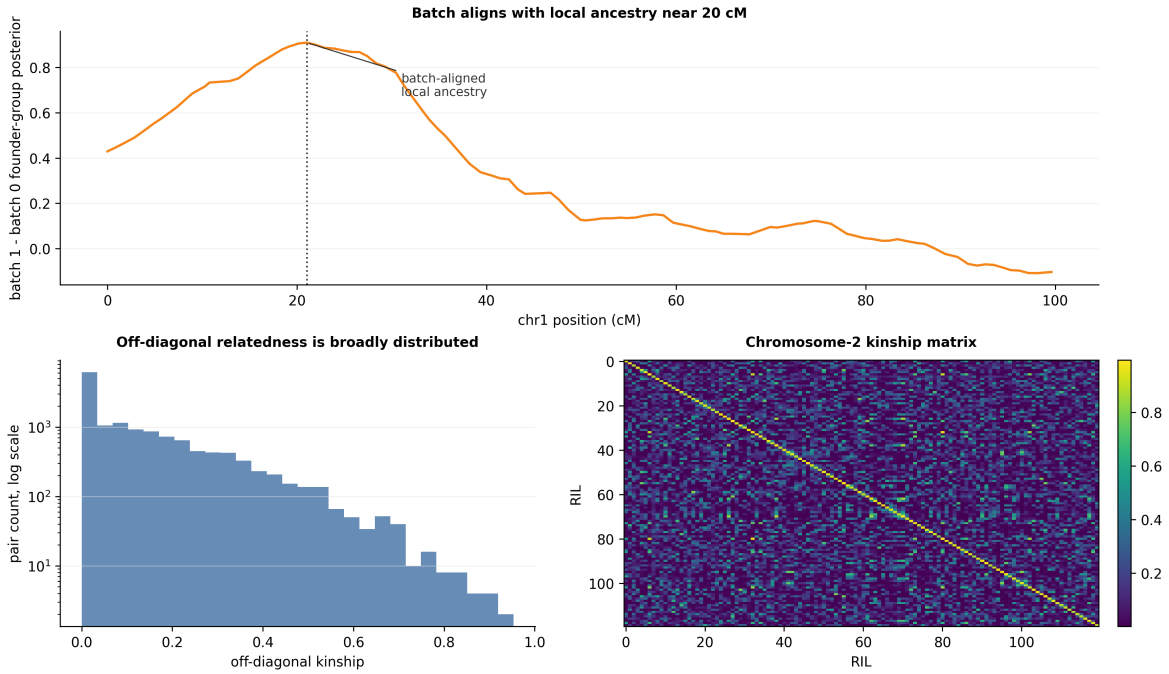

Figure 4: Batch and relatedness diagnostics. The upper orange line plots the batch-1 minus batch-0 founder-group posterior along chromosome 1; the dotted vertical line and arrow identify the batch-aligned local-ancestry peak near 20 centimorgans (cM). The lower-left blue histogram shows the distribution of off-diagonal pairwise kinship values on a log-count scale. The lower-right heatmap is the chromosome-2 recombinant-inbred-line (RIL) by RIL kinship matrix; the color bar runs from low sharing in purple to high sharing in yellow. Batch adjustment determines the recovered target in this realization, while the relatedness term remains a defensible mixed-model component.

Assemble the answer. The final analysis uses corrected HMM founder posteriors, includes batch as a fixed effect, and estimates background relatedness from chromosome 2. The chromosome-1 scan selects `m1_053` with LRT 12.2358. Importantly, this does not make ordinary least squares on the raw biallelic marker matrix a valid analysis. Raw-input OLS fails because it never estimates founder ancestry. In contrast, batch-adjusted OLS on the corrected HMM founder posterior probabilities recovers the same marker and founder in this realization, indicating that the relatedness random effect is not target-determining after founder reconstruction, orientation correction, and batch adjustment have already been performed. The chromosome-2 relatedness term remains a defensible mixed-model component, but here it does not alter the chromosome-1 peak once those three decisions are correct. A chromosome-2-only scan with the same batch-adjusted mixed model has a lower leading statistic, LRT 7.1268 at 12.07 cM/F7, and answers a different question. The reported answer is therefore

$$\texttt{qtl\_pos\_cM} = 48.635551353449436, \qquad \texttt{high\_founder} = \texttt{F5}. \tag{8}$$

8 Estimand

The target is a realized-data association peak, not a hidden simulator constant: the strongest chromosome-1 founder-state association after founder reconstruction and batch adjustment. For-

mally,

$$(k^\star, f^\star) = \arg \max_{k:C_k=1,\ f\in\{1,\ldots,8\}} \Lambda_{kf}, \tag{9}$$

where Λ_{kf} compares a null model with intercept, batch, and a relatedness random effect to an alternative model adding the founder posterior probability $\gamma_{:kf}$. The public answer fields are

$$\texttt{qtl\_pos\_cM} = \texttt{pos\_cM}(k^\star), \qquad \texttt{high\_founder} = \text{label}(f^\star). \tag{10}$$

The high-effect founder is the selected founder whose fitted founder-dosage coefficient is positive in the alternative model. For the released data, this target is marker $k^\star = \texttt{m1\_053}$ with the realized answer values reported in the answer-field table above.

9 Estimator

The estimator mirrors the walkthrough. First, align by marker ID and sort by chromosome and genetic-map position. Second, run a founder-state hidden Markov model following the multi-parent HMM formulation of Broman [2], where the hidden state $S_{ik} \in \{1, \ldots, 8\}$ is the founder ancestry of line i at ordered marker k. The initial state distribution is uniform at the first marker of each chromosome,

$$P(S_{i1} = f) = 1/8. \tag{11}$$

For adjacent markers on the same chromosome, let ΔM_k be the genetic distance in Morgans and $r_k = 1 - \exp(-4.0\,\Delta M_k)$. The transition model is

$$P(S_{ik} = f' \mid S_{i,k-1} = f) = \begin{cases} 1 - r_k, & f' = f, \\ r_k/7, & f' \neq f. \end{cases} \tag{12}$$

At chromosome boundaries the chain is reinitialized to the uniform distribution rather than transitioned across chromosomes. For non-missing genotype G_{ik}, the emission model is

$$P(G_{ik} \mid S_{ik} = f) = \begin{cases} 0.99, & A_{fk} = G_{ik}, \\ 0.01, & A_{fk} \neq G_{ik}, \end{cases} \tag{13}$$

and missing genotypes use emission probability 1 for every founder. Implementation therefore requires the forward-backward algorithm, preferably in scaled or log-space form: the forward pass propagates $\alpha_{ikf} \propto P(G_{i,1:k}, S_{ik} = f)$ through the transition and emission matrices, the backward pass propagates $\beta_{ikf} \propto P(G_{i,k+1:K} \mid S_{ik} = f)$, and the founder dosage used by the association scan is

$$\gamma_{ikf} = P(S_{ik} = f \mid G_i, A, M) = \frac{\alpha_{ikf}\beta_{ikf}}{\sum_{g=1}^{8} \alpha_{ikg}\beta_{ikg}}. \tag{14}$$

The algorithm is run twice. The first pass supplies the marker-level posterior-mismatch diagnostic used to identify the two isolated orientation outliers; those genotype columns are complemented, and the HMM is rerun so that the final γ_{ikf} values are based on the repaired marker matrix. A single-marker regression, a Viterbi-only hard-call path, or an OLS scan on the raw biallelic calls is not equivalent, because those analyses never integrate uncertainty across flanking founder-informative markers.

Third, construct a chromosome-2 relatedness matrix from founder posteriors. Relatedness-adjusted mixed models are a standard way to account for sample structure in genetic association scans [8, 9, 10]:

$$K_{ij} = \frac{1}{m_2} \sum_{k:C_k=2} \sum_{f=1}^{8} \gamma_{ikf}\gamma_{jkf}. \tag{15}$$

For each chromosome-1 marker and founder, fit

$$Y = X\beta + \alpha_{kf}\gamma_{:kf} + u + \epsilon, \qquad X = [\mathbf{1}, B], \quad u \sim N(0, \sigma_g^2 K), \quad \epsilon \sim N(0, \sigma_e^2 I). \tag{16}$$

The implementation diagonalizes K, searches a grid over $\delta = \sigma_e^2/\sigma_g^2$, and evaluates generalized least-squares residual sums of squares. The scan statistic is

$$\Lambda_{kf} = n \log\left(\frac{\mathrm{RSS}_0 + 10^{-12}}{\mathrm{RSS}_{1,kf} + 10^{-12}}\right). \tag{17}$$

10 Decision-Point and Ablation Walkthrough

The table combines ablation values, pass/fail status, and diagnostic interpretation for the released files. The direct-biallelic rows fail at founder reconstruction; the final rows are target-equivalent sensitivities that preserve founder reconstruction and batch adjustment while varying the mixed-model term, HMM transition rate, or handling of the two orientation-outlier markers.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
Reference pipeline	HMM, orientation correction, and batch-adjusted LMM	48.64 cM, error 0.00, F5	yes	none	Reference realized-data target.
Founder reconstruction	Raw all-marker biallelic OLS	42.05 cM, error 6.59, F1	no	Stage 2	Treats a many-founder mosaic as a binary marker scan and never estimates founder ancestry.
Founder reconstruction	Raw chromosome-1-only biallelic OLS	13.80 cM, error 34.83, no founder label	no	Stage 2	Uses the raw biallelic matrix, so the founder identity requested by the prompt is not identifiable.
Orientation repair	HMM without orientation correction	43.31 cM, error 5.33, F4	no	Stage 2 QC	Keeps the two posterior-mismatch outlier markers in the wrong allele orientation.
Batch adjustment	Corrected HMM LMM, batch omitted	18.40 cM, error 30.24, F2	no	Stage 3	Confuses the batch-aligned 20 cM ancestry peak with the target QTL.
Founder posteriors without adjustment	Corrected HMM OLS, batch and kinship omitted	18.40 cM, error 30.24, F2	no	Stage 3	Uses correct founder posteriors but still omits the batch covariate that explains the nuisance peak.
Compound omission	HMM without orientation correction or batch adjustment	18.40 cM, error 30.24, F2	no	Stages 2–3	Combines the local orientation error with the batch-confounded scan.
Compound omission	HMM without orientation correction, batch adjustment, or kinship	38.22 cM, error 10.42, F2	no	Stages 2–3	Leaves both the orientation problem and the model-adjustment problem unresolved.
Target chromosome	Chromosome-2-only batch-adjusted LMM	12.07 cM on chromosome 2, error 36.57, F7	no	Target definition	Scans the relatedness chromosome rather than the requested chromosome-1 QTL.
Target chromosome	Chromosome-2-only batch-adjusted OLS on HMM posteriors	87.25 cM on chromosome 2, error 38.62, F3	no	Target definition	Answers a different chromosome-specific association question.
Acceptable sensitivity	Corrected HMM batch-adjusted OLS, kinship omitted	48.64 cM, error 0.00, F5	yes	none	Target-equivalent here because founder reconstruction, orientation repair, and batch adjustment already determine the chromosome-1 peak.
Acceptable sensitivity	Corrected HMM with transition rate 3.41 per Morgan	48.64 cM, error 0.00, F5	yes	none	Target-equivalent transition-rate sensitivity; the learned-rate-style value leaves the same marker and founder peak.
Acceptable sensitivity	Drop `m1_056`/`m1_058` instead of flipping	48.64 cM, error 0.00, F5	yes	none	Target-equivalent marker-handling sensitivity; discarding the two orientation outliers weakens the LRT but preserves the graded answer.

Table 2: Unified decision-point and ablation walkthrough. Founder reconstruction, orientation correction, and batch adjustment are the decisions that determine the recovered chromosome-1 QTL.

References

[1] R. Mott, C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint. "A method for fine mapping quantitative trait loci in outbred animal stocks." *Proceedings of the National Academy of Sciences*, 97(23):12649–12654, 2000. DOI: 10.1073/pnas.230304397.

[2] K. W. Broman. "A generic hidden Markov model for multiparent populations." *G3: Genes—Genomes—Genetics*, 12(2):jkab396, 2022. DOI: 10.1093/g3journal/jkab396.

[3] K. W. Broman, D. M. Gatti, P. Simecek, N. A. Furlotte, P. Prins, S. Sen, B. S. Yandell, and G. A. Churchill. "R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multiparent populations." *Genetics*, 211(2):495–502, 2019. DOI: 10.1534/genetics.118.301595.

[4] D. M. Gatti, K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar, P. Simecek, N. Goodwin, R. Cheng, D. Pomp, A. Palmer, E. J. Chesler, K. W. Broman, and G. A. Churchill. "Quantitative trait locus mapping methods for Diversity Outbred mice." *G3: Genes—Genomes—Genetics*, 4(9):1623–1633, 2014. DOI: 10.1534/g3.114.013748.

[5] P. X. Kover, W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, M. D. Purugganan, C. Durrant, and R. Mott. "A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana." *PLoS Genetics*, 5(7):e1000551, 2009. DOI: 10.1371/journal.pgen.1000551.

[6] P. Deelen, M. J. Bonder, K. J. van der Velde, H.-J. Westra, E. Winder, D. Hendriksen, L. Franke, and M. A. Swertz. "Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration." *BMC Research Notes*, 7:901, 2014. DOI: 10.1186/1756-0500-7-901.

[7] K. W. Broman, D. M. Gatti, K. L. Svenson, S. Sen, and G. A. Churchill. "Cleaning genotype data from Diversity Outbred mice." *G3: Genes—Genomes—Genetics*, 9(5):1571–1579, 2019. DOI: 10.1534/g3.119.400165.

[8] J. Yu, G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." *Nature Genetics*, 38:203–208, 2006. DOI: 10.1038/ng1702.

[9] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. "Efficient control of population structure in model organism association mapping." *Genetics*, 178(3):1709–1723, 2008. DOI: 10.1534/genetics.107.080101.

[10] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. "Variance component model to account for sample structure in genome-wide association studies." *Nature Genetics*, 42:348–354, 2010. DOI: 10.1038/ng.548.