GeneBench-Pro Case Study: Parent-Specific Recent Admixture Timing from Local Ancestry

GeneBench-Pro

June 26, 2026

1 Overview

This case study focuses on a practical failure mode in local-ancestry analysis: compact tract summaries can look ready for single-pulse timing even when phase artifacts, chromosome-local label conventions, and incomplete map coverage change the estimand. The released files contain only phased ancestry tracts for one recently admixed individual and a four-row genetic map. The requested output contains four target quantities: the ancestry-A fraction and single-pulse admixture time for the lower-A parental haplotype and for the higher-A parental haplotype. A defensible analysis must first recognize reciprocal low-confidence micro-tracts as phase-related local-ancestry artifacts, repair or equivalently simplify the retained tract path after filtering, detect that the ancestry labels on chr3 are inverted relative to the other chromosomes, and finally use map-declared chromosome lengths rather than the shorter called-span length when converting switch counts to time since admixture.

The scientific core is standard: local ancestry represents an admixed genome as ancestry-labeled chromosomal tracts, and under a single recent pulse the expected rate of ancestry switches is proportional to $2tp(1-p)$ per Morgan, where p is the ancestry fraction and t is the number of generations since admixture [1,2]. Published local-ancestry methods and evaluations show how phasing uncertainty, reference-panel choice, and ancestry-call quality can affect tract summaries [3–5]. This staged data example makes the remaining answer-changing conventions directly visible in the released files: arbitrary haplotype labels, chromosome-local ancestry-label inversion, chromosome-boundary artifacts, and incomplete tract coverage are diagnosed in the walkthrough rather than asserted from the citations alone.

2 Released Prompt and Files

Prompt

```
You are given phased local-ancestry tracts for one admixed individual. Estimate, for each
     transmitted parental haplotype, the fraction of ancestry A across the called tract
    span and the number of generations since a single recent admixture pulse. Label
    parent1 as the haplotype with the smaller ancestry-A fraction and parent2 as the
    haplotype with the larger ancestry-A fraction.
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": {
    "parent1_A_fraction": <float>,
    "parent1_t": <float>,
    "parent2_A_fraction": <float>,
    "parent2_t": <float>
  },
  "reasoning": "<description of method and QC>"
}
```

Released data files

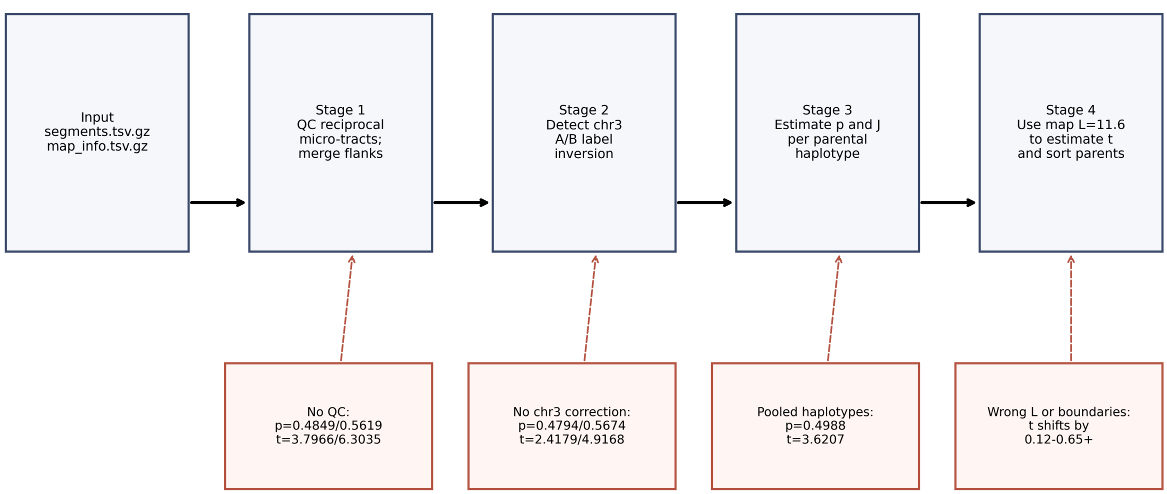
File	Format	Contents
`segments.tsv.gz`	.tsv.gz	Sixty-six phased local-ancestry tract rows across four chromosomes and two haplotypes, with chromosome-local Morgan coordinates, ancestry labels, posterior probabilities, and low-complexity fractions.
`map_info.tsv.gz`	.tsv.gz	Four chromosome map lengths in Morgan units; these lengths define the recombination-map denominator for the single-pulse timing estimator.

3 Answer Fields and Tolerances

Answer field	Ground truth	Tolerance / matching rule	Interpretation
`parent1_A_fraction`	0.273383	Absolute error ≤ 0.010; valid range $[0, 1]$	Ancestry-A fraction for the lower-A parental haplotype after artifact repair and haplotype sorting.
`parent1_t`	3.037828	Absolute error ≤ 0.050; valid range ≥ 0	Single-pulse generation estimate for the lower-A parental haplotype.
`parent2_A_fraction`	0.724795	Absolute error ≤ 0.010; valid range $[0, 1]$	Ancestry-A fraction for the higher-A parental haplotype after artifact repair and haplotype sorting.
`parent2_t`	6.050599	Absolute error ≤ 0.100; valid range ≥ 0	Single-pulse generation estimate for the higher-A parental haplotype.

All four reported values are interpreted together. The ancestry-fraction fields use absolute windows of 0.01, and the timing fields use windows of 0.05 and 0.10. Those timing windows admit the posterior-weighted sensitivity and the chr3-corrected no-merge cleanup sensitivity under the adjacent-different-ancestry switch definition, while still excluding wrong-denominator, no-artifact-QC, no-label-correction, pooled-haplotype, and boundary-counting alternatives.

4 Structure Diagram



Structure diagram. The diagram is read left to right. Blue-gray boxes connected by solid black arrows show the reference workflow; pale red boxes below the workflow show representative wrong analyses; dashed red arrows attach each wrong analysis to the stage it skips or mis-specifies. Here p denotes ancestry-A fraction, J denotes within-chromosome ancestry-switch count, L denotes total map length in Morgans, and t denotes generations since the admixture pulse. Stage 1 reads phased local-ancestry tracts and detects reciprocal low-confidence micro-tracts that appear at the same genomic coordinates on both haplotypes with opposite ancestry labels; these are treated as phase-related local-ancestry artifacts. The reference cleanup removes those artifacts and merges flanking same-ancestry segments, although a chr3-corrected implementation that drops the artifacts without physically merging same-ancestry flanks is answer-equivalent under the released tolerances when J is counted only across adjacent different-ancestry tracts. Stage 2 detects and corrects the ancestry-label inversion on chr3 by examining per-chromosome A fractions and noticing that chr3's pattern is flipped relative to all other chromosomes. Stage 3 computes per-haplotype ancestry-A fractions on the called tract span and within-chromosome switch counts from the cleaned tracts. Stage 4 uses the map-declared chromosome lengths ($L = 11.6$ Morgan from `map_info.tsv.gz`) rather than tract totals or endpoint spans when converting switch counts to single-pulse admixture times. Skipping the answer-affecting stages produces answers that remain numerically separated from the reference analysis; even the closest wrong-denominator shortcut shifts at least one timing estimate by about 0.24 generations.

5 Variables and Assumptions

- $h \in \{h_1, h_2\}$: Haplotype label for the two transmitted parental haplotypes.
- $j_{h_1} = 30$, $j_{h_2} = 15$: Pre-truncation alternating-path switch counts used to construct the clean genome-wide tract paths before chromosome splitting, gaps, telomeric truncation, and filtering.
- $L = 11.6$ Morgan: Total chromosome map length from `map_info.tsv.gz`, split across four chromosomes (2.8, 2.8, 3.1, 2.9 Morgan). The called tract total is about 10.95 Morgan and the endpoint span is about 11.15 Morgan; neither should replace the map length in the time estimator.
- $B = (0, 2.8, 5.6, 8.7, 11.6)$ Morgan: Genome-wide chromosome boundary vector used to convert the constructive tract paths into chromosome-local coordinates.
- $p_h \in (0, 1)$: Called-span ancestry-A fraction for haplotype h, defined as A-tract length divided by total called tract length after cleaning and label correction. After cleaning: $p_{h_1} \approx 0.7248$, $p_{h_2} \approx 0.2734$.
- $J_h \in \mathbb{Z}_{\geq 0}$: Within-chromosome ancestry-switch count for haplotype h. Switches are counted only between adjacent segments on the same chromosome; chromosome boundaries are not recombination events and are excluded. After cleaning and chr3 label correction: $J_{h_1} = 28$, $J_{h_2} = 14$.
- $t_h > 0$: Generations since the single admixture pulse for haplotype h.
- $w_i \sim \text{Gamma}(2, 1)$: Segment-length weights drawn separately for A and B tracts and then rescaled to the target ancestry totals. Seeds are fixed at 3 for h_1 and 4 for h_2.
- $x \in \{0.52, 3.90, 6.00, 10.20\}$ Morgan: Genome-wide centers of reciprocal micro-tracts inserted into both haplotypes.

- Each tract segment has a posterior probability (`posterior`) and a low-complexity fraction (`low_complexity_frac`). Normal tracts have posterior = 0.985 and low-complexity = 0.08. Micro-tracts have posterior = 0.62 and low-complexity = 0.92.
- $g_{ck} \sim \text{Uniform}(0.02, 0.04)$ Morgan: Interior gap lengths inserted on chr1, chr2, and chr4 with fixed seed 77, two gaps per chromosome and haplotype where eligible adjacent different-ancestry tracts are large enough.
- (a_c, z_c): Telomeric start and end trims. Values are $(0.03, 0.08)$ for chr1, $(0.04, 0.07)$ for chr2, $(0.02, 0.10)$ for chr3, and $(0.05, 0.06)$ for chr4.

6 Data-Generating Process

The simulation constructs two haplotypes with pre-specified switch counts and ancestry fractions, then injects reciprocal phasing artifacts.

Step 1: Generate clean tract paths. For haplotype h_1 (high-A), a genome-wide alternating tract path is created with ancestry labels starting from A and target ancestry fraction $p = 0.72$. Segment lengths are drawn from Gamma(2, 1) and rescaled so that A-ancestry segments sum to $p \cdot L$ and B-ancestry segments sum to $(1 - p) \cdot L$:

$$w_A^{(i)} \sim \text{Gamma}(2, 1), \quad \ell_A^{(i)} = w_A^{(i)} \cdot \frac{p \cdot L}{\sum_j w_A^{(j)}} \tag{1}$$

with analogous scaling for B-ancestry segments. Haplotype h_2 (low-A) is generated by the same constructive procedure, starting from B and target ancestry fraction $p = 0.28$. Random seeds are fixed (3 for h_1, 4 for h_2). After chromosome splitting, artifact repair, and boundary exclusion, the recoverable switch counts are $J_{h_1} = 28$ and $J_{h_2} = 14$, and the called-span A fractions are approximately $p_{h_1} = 0.7248$ and $p_{h_2} = 0.2734$.

Step 2: Insert reciprocal artifacts. Four artifact positions are fixed at 0.52, 3.90, 6.00, and 10.20 Morgan (genome-wide coordinates). At each position x, on both haplotypes, the tract containing x is split and a 0.03 Morgan micro-tract of the opposite ancestry is inserted:

$$[s, e, \text{anc}] \to [s, x - \delta, \text{anc}],\ [x - \delta, x + \delta, \overline{\text{anc}}],\ [x + \delta, e, \text{anc}] \tag{2}$$

where $\delta = 0.015$ Morgan and $\overline{\text{anc}}$ denotes the opposite ancestry. Artifact tracts are assigned posterior = 0.62 and low-complexity = 0.92. This produces 8 artifact tracts (4 per haplotype) that are reciprocal: they share coordinates across haplotypes but carry opposite ancestry labels.

Step 3: Split into chromosomes. The genome-wide tract path is split at chromosome boundaries

$$(b_0, b_1, b_2, b_3, b_4) = (0,\ 2.8,\ 5.6,\ 8.7,\ 11.6) \text{ Morgan.} \tag{3}$$

For each segment with genome-wide coordinates $[s_i, e_i]$, the portion overlapping chromosome c is emitted with per-chromosome coordinates

$$s_{ic} = \max(s_i, b_{c-1}) - b_{c-1}, \qquad e_{ic} = \min(e_i, b_c) - b_{c-1}. \tag{4}$$

This creates four chromosome-indexed local-ancestry tables without treating chromosome boundaries as recombination switches.

Step 4: Introduce gaps and telomeric truncation. On chr1, chr2, and chr4 (not chr3), two interior gaps are introduced between selected adjacent different-ancestry tracts. Each gap has length

$$g_{ck} \sim \text{Uniform}(0.02, 0.04) \text{ Morgan}, \qquad c \in \{\text{chr1}, \text{chr2}, \text{chr4}\}, \quad k = 1, 2, \tag{5}$$

with random seed 77. The left tract endpoint is moved left by $g_{ck}/2$ and the right tract start is moved right by $g_{ck}/2$. Telomeric truncation is then applied identically to both haplotypes:

c	chr1	chr2	chr3	chr4
a_c (start trim)	0.03	0.04	0.02	0.05
z_c (end trim)	0.08	0.07	0.10	0.06

(6)

so the first surviving tract begins at a_c and the last surviving tract ends at $L_c - z_c$. This makes total called tract length (10.9537 Morgan for h1 and 10.9863 Morgan for h2) and endpoint span (11.15 Morgan per haplotype) both less than the true total map length (11.6 Morgan from `map_info.tsv.gz`).

Step 5: Invert ancestry labels on chr3. All ancestry labels on chr3 are swapped for both haplotypes:

$$\text{anc}_i^{\text{visible}} = \begin{cases} B, & \text{chrom}_i = \text{chr3 and anc}_i^{\text{clean}} = A, \\ A, & \text{chrom}_i = \text{chr3 and anc}_i^{\text{clean}} = B, \\ \text{anc}_i^{\text{clean}}, & \text{chrom}_i \neq \text{chr3}. \end{cases} \tag{7}$$

Micro-tracts on chr3 are also swapped, preserving the reciprocal structure. The construction-only `artifact`, `inverted`, and `gap` annotations are not included in the released tract file.

7 Analyst Walkthrough

The analysis has four decision points. The analyst must identify reciprocal low-posterior micro-tracts, repair the chr3 ancestry-label inversion, keep the two parental haplotypes separate before sorting them by inferred ancestry fraction, and use the full genetic-map length when translating switch counts into pulse timing. Missing the chr3 inversion is the answer-changing step: it keeps both ancestry fractions near the raw 0.48/0.56 basin and makes the subsequent parent-specific timing calculation look superficially reasonable while targeting the wrong haplotype paths.

1. Initial inspection. Load the tract file and examine the data. The released table contains 66 rows across four chromosomes and two haplotypes, with each row giving an ancestry label, a posterior probability, and a low-complexity fraction. The first useful plot is simply the raw tract path by haplotype and chromosome:

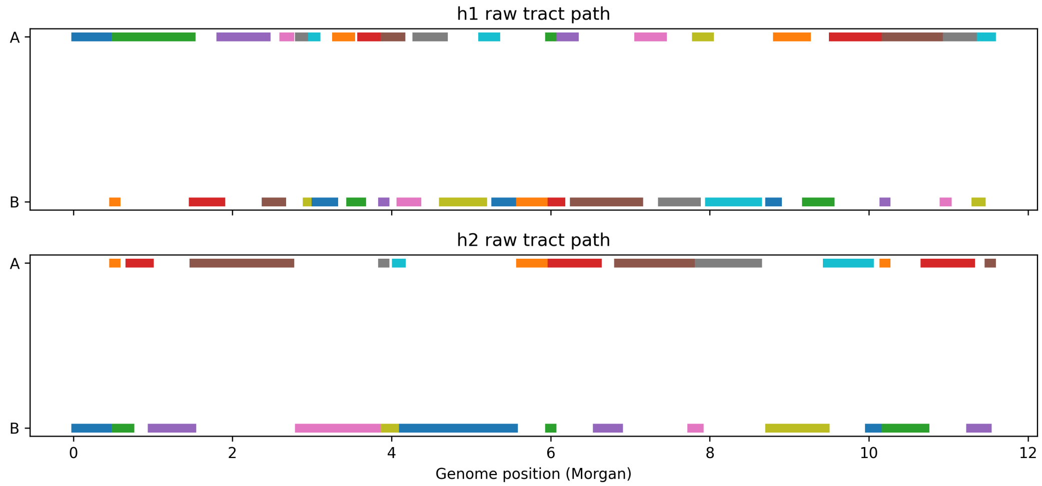


Figure 1. Raw local-ancestry tracts. Each horizontal segment is one called tract, plotted at its genome-wide Morgan position after offsetting chr1–chr4 onto one axis. The y-axis encodes the called ancestry label, with B on the lower row and A on the upper row within each haplotype panel. The colors are arbitrary visual separators for adjacent tract segments and do not encode ancestry, chromosome, posterior probability, or parent identity. Blank intervals between segments are uncalled local-ancestry gaps or telomeric truncation.

A naive first pass would count all adjacent ancestry changes, compute per-haplotype A fractions directly, and plug those summaries into the switch-count estimator. That gives approximately $p = 0.4849/0.5619$ and $t = 3.7966/6.3035$ for the lower-A and higher-A outputs. Those values look superficially plausible because they preserve parent-specific asymmetry and produce recent admixture times. They are wrong because the raw table contains both low-confidence reciprocal micro-tracts and a chromosome-local ancestry-label inversion.

```
seg = read_tsv("segments.tsv.gz")
raw = summarize(seg, map_length=11.6, repair=False, flip_chr3=False)
print(raw[["A_fraction", "switches", "t"]])
```

2. Discover the low-confidence artifact class. Plot posterior probability against low-complexity fraction. This is the first substantive quality-control decision, and it is intentionally data-driven rather than prompt-driven.

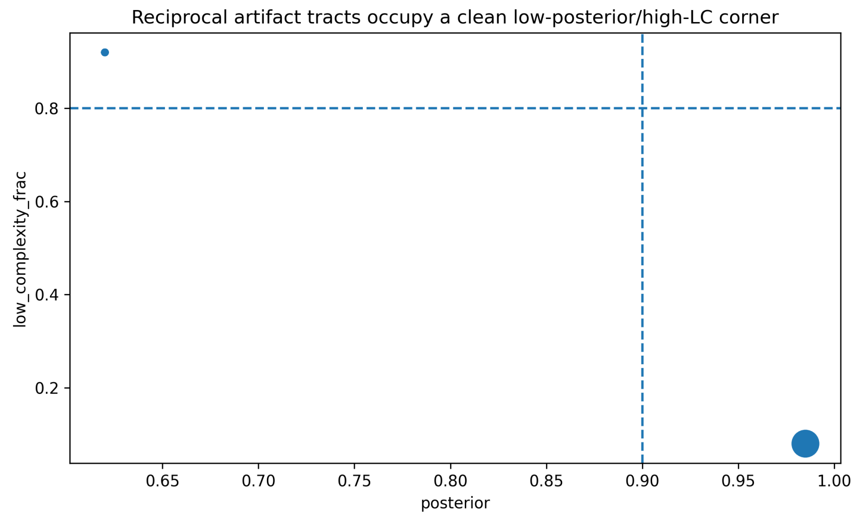


Figure 2. Posterior probability versus low-complexity fraction. Each point is a tract; point area is proportional to tract length, so the large lower-right point represents many normal tracts with the same realized values. The vertical dashed line marks the posterior cutoff 0.9, and the horizontal dashed line marks the low-complexity cutoff 0.8. Normal tracts lie in the high-posterior/low-complexity lower-right corner (posterior 0.985, low-complexity 0.08); suspect reciprocal micro-tracts lie in the low-posterior/high-complexity upper-left corner (posterior 0.62, low-complexity 0.92).

```
suspect = seg[(seg.posterior < 0.9) &
              (seg.low_complexity_frac > 0.8)]
suspect[["chrom", "start_morgan", "end_morgan", "hap", "anc"]]
```

3. Confirm reciprocal structure. The suspect tracts are not random low-quality calls. They form reciprocal pairs: each pair appears at the same chromosome and coordinates on both haplotypes, but with opposite ancestry labels.

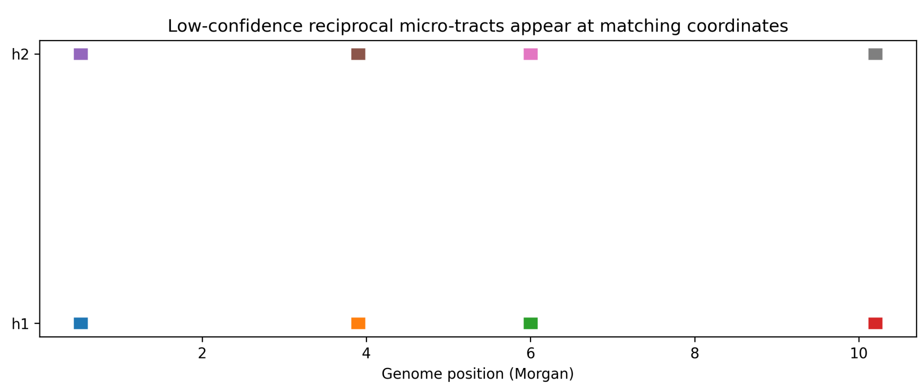


Figure 3. Reciprocal artifact structure. Each thick mark is a suspect micro-tract that passed the Figure 2 low-posterior/high-complexity screen. The x-axis is genome-wide Morgan position, and the y-axis separates the two transmitted haplotypes h1 and h2. Colors are arbitrary visual separators for the four matched coordinate pairs; they do not encode ancestry labels. At each colored coordinate pair, the h1 and

h2 micro-tracts have the same genomic interval but opposite A/B ancestry labels, which is the reciprocal pattern consistent with phase-related local-ancestry error rather than true transmitted ancestry changes.

4. Repair the tract path after filtering. The reference cleanup removes the reciprocal suspect tracts and then merges adjacent same-ancestry segments separated by gaps ≤ 0.05 Morgan. This is the simplest way to represent the intended cleaned tract path. Under the report's switch definition, however, the answer depends on adjacent different-ancestry transitions, not on the physical row count after a same-ancestry micro-tract is removed. Therefore a chr3-corrected implementation that drops the reciprocal suspects but does not merge same-ancestry flanks gives approximately $p = 0.2764/0.7217$ and $t = 3.0172/6.0096$, which remains inside the released answer tolerances. The important answer-affecting decision is to remove the reciprocal low-confidence artifacts and then count only genuine adjacent different-ancestry switches.

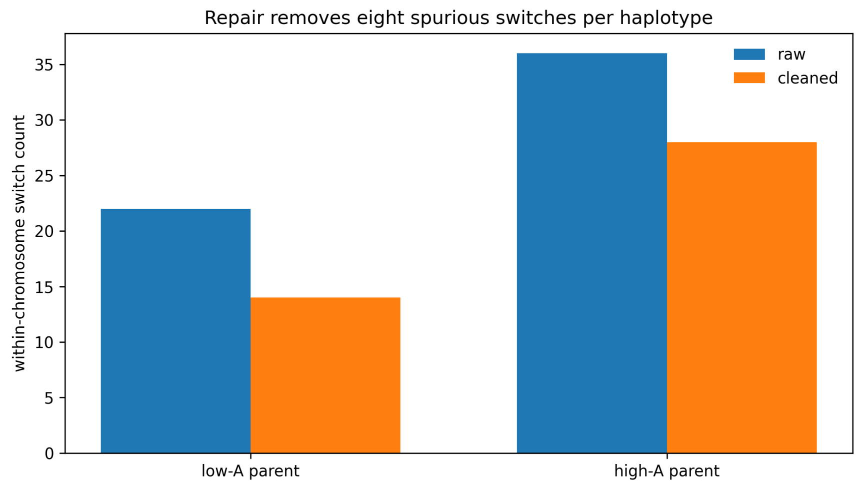


Figure 4. Switch-count impact of repair. Bars show within-chromosome ancestry-switch count J after sorting outputs into parent1 (lower-A parent) and parent2 (higher-A parent). Blue bars use the raw tract table; orange bars use the cleaned table after removing reciprocal micro-tracts and merging same-ancestry flanks separated by small gaps. The drop from blue to orange is the switch-count inflation caused by treating the micro-tract artifacts as real ancestry changes.

5. Correct chr3 labels. After artifact repair, compute per-chromosome ancestry-A fractions for each haplotype. Chr1, chr2, and chr4 agree on the same ordering: h1 is the high-A haplotype and h2 is the low-A haplotype, with h1–h2 A-fraction differences of approximately +0.242, +0.465, and +0.401. Chr3 reverses that ordering with a difference of approximately −0.669. That cross-chromosome inconsistency is the evidence for label harmonization: the A/B labels on chr3 are inverted relative to the other chromosomes, and swapping them restores a positive chr3 difference of approximately +0.669. Without this step, the cleaned genome-wide fractions remain in the wrong basin: $p \approx 0.4794/0.5674$ rather than 0.2734/0.7248.

```
by_chr = repaired.groupby(["chrom", "hap"]).A_fraction()
delta = by_chr["h1"] - by_chr["h2"]
print(delta)   # chr3 has the opposite sign from chr1/chr2/chr4
```

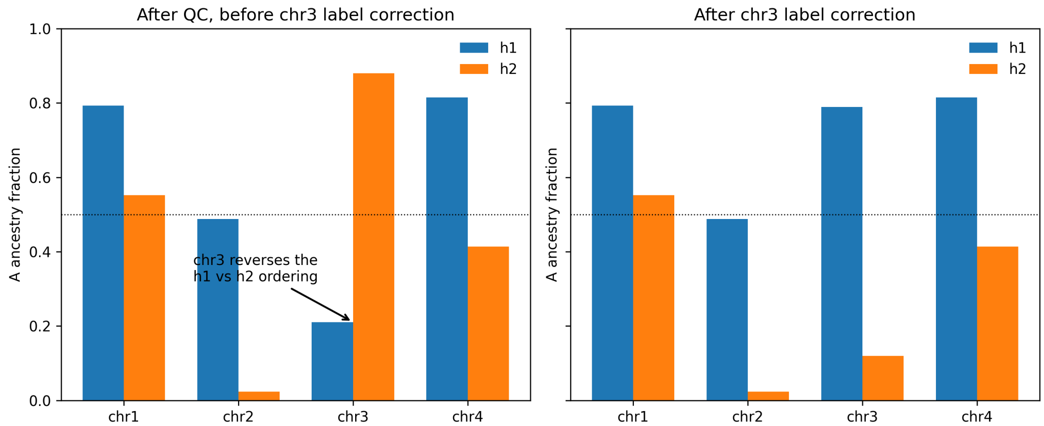


Figure 5. Per-chromosome A fractions before and after label correction. Blue bars are haplotype h1 and orange bars are haplotype h2. The dotted horizontal reference line marks A ancestry fraction 0.5 in both panels. The left panel is after artifact QC but before chr3 label correction; the arrow marks chr3, the only chromosome where h1 has less A ancestry than h2. The right panel swaps A and B labels on chr3, after which all four chromosomes agree that h1 is the higher-A haplotype and h2 is the lower-A haplotype.

6. Compute parent-specific called-span fractions. After artifact repair and chr3 label correction, compute A-tract length divided by total called tract length separately for each haplotype. The called-span convention matters because the prompt asks for the fraction across the called tract span, not across the full map length. The result is:

$$p_{h_2} = 0.2734, \qquad p_{h_1} = 0.7248.$$

Pooling across haplotypes gives $p \approx 0.4988$ even after cleaning. That pooled value is scientifically meaningful only for the individual's aggregate ancestry; it destroys the parent-specific signal that this case study asks the analyst to recover. Parent-asymmetric admixture and recent-ancestor ancestry summaries are exactly the setting where parent-specific estimation is needed [1,6]. Figure 5 shows the same parent-specific separation after label correction.

Hard-call and posterior-weighted fractions. The reference analysis treats each retained, repaired tract as a hard local-ancestry call when computing p_h. A closely related posterior-weighted variant is also defensible: after the same artifact screen, flank merge, chr3 label correction, parent sorting, and map-denominator choice, retained tracts can contribute posterior-weighted ancestry lengths while the switch count J_h remains path-based on the cleaned hard-call tract sequence. That variant gives `parent1_A_fraction`$= 0.2802$, `parent1_t`$= 2.9921$, `parent2_A_fraction`$= 0.7181$, and `parent2_t`$= 5.9613$, which is inside the released answer tolerances. This tolerance does not accept posterior use as a substitute for the earlier decisions: without chr3 label correction, posterior-only filtering remains in the wrong ancestry-fraction basin around 0.48/0.57.

7. Use the map file for total chromosome length. The cleaned tracts still do not tile the chromosomes. Summing cleaned tract lengths gives $L \approx 10.95$ Morgan, and summing per-chromosome endpoint spans gives $L \approx 11.15$ Morgan. Both are smaller than the declared map total, $L = 11.6$ Morgan, because the released local-ancestry calls contain gaps and telomeric truncation. The switch-count estimator is defined over the full genetic map length on which recombination could have generated ancestry switches, so the correct denominator is the map file, not the called span. Using tract length gives $t = 3.2075/6.4076$; using endpoint span gives $t = 3.1604/6.2948$.

```
L = map_info.length_morgan.sum()      # 11.6
t_h = J_h / (2 * L * p_h * (1 - p_h))
```

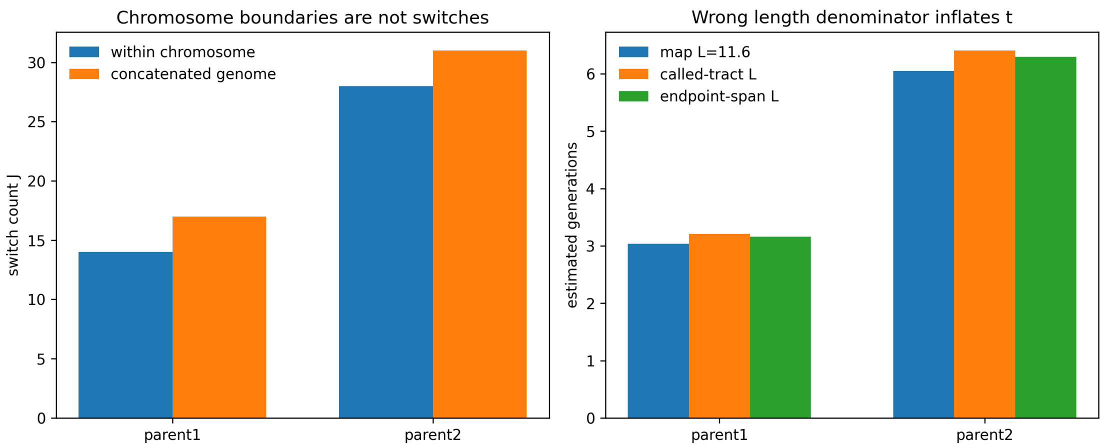


Figure 6. Boundary and denominator diagnostics. In the left panel, blue bars count only within-chromosome ancestry switches J, while orange bars incorrectly concatenate chromosomes and count chromosome-boundary jumps as switches. In the right panel, blue bars use the correct map denominator $L = 11.6$ Morgans from `map_info.tsv.gz`, orange bars use the shorter called-tract length, and green bars use the endpoint-span length. Parent1 is the lower-A parent and parent2 is the higher-A parent. The y-axis is switch count J on the left and estimated generations t on the right.

8. Assemble the requested answer. The final numerical accounting is:

$$t_{h_2} = \frac{14}{2 \cdot 11.6 \cdot 0.2734 \cdot (1 - 0.2734)} = 3.0378,$$
$$t_{h_1} = \frac{28}{2 \cdot 11.6 \cdot 0.7248 \cdot (1 - 0.7248)} = 6.0506.$$

The prompt defines parent1 as the haplotype with the smaller ancestry-A fraction. Therefore h2 maps to parent1 and h1 maps to parent2.

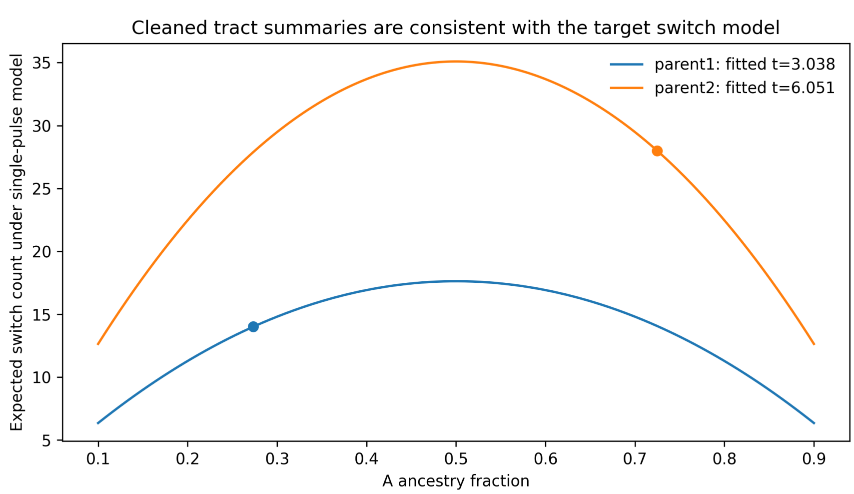


Figure 7. Final switch-count model fit. The x-axis is called-span ancestry-A fraction p, and the y-axis is expected within-chromosome switch count under the single-pulse model $E[J] = 2Ltp(1-p)$. Blue denotes parent1 (the lower-A parent) and orange denotes parent2 (the higher-A parent). Curves show the expected switch count as p varies while holding the fitted t in the legend fixed; circular markers show the observed cleaned summaries (p, J) used for the final answer. The fit uses the full map length $L = 11.6$ Morgans.

8 Estimand

The estimands are the parent-specific ancestry-A fractions and single-pulse admixture times. The ancestry fraction is deliberately defined on the called tract span because that is the quantity requested in the prompt. The time estimator uses the full map length because the Poisson switch process is defined over the chromosome map, including uncalled intervals:

$$p_h = \frac{\sum_i \ell_i^{(h)} \cdot \mathbf{1}[\text{anc}_i = A]}{\sum_i \ell_i^{(h)}}, \quad t_h = \frac{J_h}{2L\, p_h(1-p_h)} \tag{8}$$

where $\ell_i^{(h)}$ are cleaned called tract lengths for haplotype h, J_h is the within-chromosome switch count, and $L = \sum_c L_c$ is the chromosome total from `map_info.tsv.gz`. The naive approach is biased because uncleaned data inflates J_h, the chr3 inversion biases p_h, and using called-tract totals in place of L inflates t_h. The time parameter t_h is the generation-time parameter of the single-pulse ancestry-switch process, not a row index or calendar date. The switch count J_h counts different-ancestry flanks across small repaired gaps within a chromosome after removing reciprocal artifacts, while chromosome boundaries are never counted as recombination switches.

9 Estimator

The correct estimator proceeds in four steps:

Step 1 (quality control and repair): Identify suspect tracts with posterior < 0.9 and low-complexity > 0.8. Check for reciprocal pairs (same chromosome, same coordinates, different haplotype, opposite ancestry). Remove all suspect tracts in reciprocal pairs. Merge flanking segments that share the same ancestry and are separated by a gap ≤ 0.05 Morgan.

Step 2 (Label correction): Examine per-chromosome ancestry-A fractions for each haplotype. Detect that chr3 has an inverted pattern (h1 is low-A on chr3 but high-A on chr1/chr2/chr4). Swap A $\leftrightarrow$ B labels on chr3.

Step 3 (Summary statistics): For each haplotype h, compute:

$$p_h = \frac{\sum_i \ell_i^{(h)} \cdot \mathbf{1}[\text{anc}_i = A]}{\sum_i \ell_i^{(h)}} \tag{9}$$

$$J_h = \left|\left\{ i : \text{anc}_i^{(h)} \neq \text{anc}_{i-1}^{(h)} \text{ and } \text{chrom}_i = \text{chrom}_{i-1} \right\}\right| \tag{10}$$

Step 4 (Single-pulse estimator): Under a single recent admixture pulse, the expected number of within-chromosome switches follows a Poisson process with rate $2tp(1-p)$ per Morgan [1,2]. Here t is the pulse-generation parameter, p_h is the called-span ancestry fraction, J_h counts repaired within-chromosome ancestry switches across adjacent called segments, and L is the full map length supplied by `map_info.tsv.gz`:

$$L = \sum_c L_c \text{ from } \texttt{map\_info.tsv.gz}, \qquad \hat{t}_h = \frac{J_h}{2\,L\,p_h(1-p_h)} \tag{11}$$

Sort results by p_h ascending: parent1 is the haplotype with smaller p_h.

10 Decision-Point and Ablation Walkthrough

The table below combines the ablation and sensitivity results with the stage at which each shortcut fails. The ancestry-fraction fields use absolute windows of 0.01, while the two timing fields use absolute windows of 0.05 and 0.10; the posterior-weighted ancestry sensitivity and chr3-corrected no-merge cleanup sensitivity remain inside those windows and are therefore target-equivalent for the released report. For compactness in the quantitative-output column, $p1$, $t1$, $p2$, and $t2$ abbreviate the released fields `parent1_A_fraction`, `parent1_t`, `parent2_A_fraction`, and `parent2_t`, respectively.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
Reference pipeline	Reference analysis	p1 0.2734, t1 3.0378, p2 0.7248, t2 6.0506	yes	none	Reference artifact repair, chr3 label correction, parent sorting, and full-map denominator.
Acceptable sensitivity	Posterior-weighted ancestry sensitivity	0.2802, 2.9921, 0.7181, 5.9613	yes	none	Posterior-weighted ancestry fractions remain within the public tolerance windows.
Acceptable sensitivity	Drop suspects without merging flanks, with chr3 correction	0.2764, 3.0172, 0.7217, 6.0096	yes	none	Counts only adjacent different-ancestry switches, so physical same-ancestry flank merging is not answer-affecting under released tolerances.
Artifact QC	No artifact QC	0.4849, 3.7966, 0.5619, 6.3035	no	Stage 1	Counts reciprocal low-confidence micro-tracts as real ancestry switches.
Boundary handling	Count chromosome-boundary transitions	0.2734, 3.6888, 0.7248, 6.6989	no	Stage 3	Treats chromosome boundaries as recombination switches.
Compound omission	No QC and boundary transitions counted	0.4849, 4.3143, 0.5619, 6.8288	no	Stages 1 and 3	Combines micro-tract artifacts with chromosome-boundary switch inflation.
Length shortcut	Length filter at 0.10 Morgan	0.4814, 2.5898, 0.5722, 4.4022	no	Stage 1	Uses length filtering as a substitute for reciprocal-artifact identification and repair.
Length shortcut	Length filter at 0.20 Morgan	0.5039, 2.0691, 0.5750, 2.8222	no	Stage 1	Drops real tracts along with artifacts and distorts timing.
Length shortcut	Length filter at 0.40 Morgan	0.4970, 1.3794, 0.5890, 0.8903	no	Stage 1	Over-filters the local-ancestry mosaic.
Output convention	Unsorted h1/h2 output labels	0.7248, 6.0506, 0.2734, 3.0378	no	Stage 3	Reports arbitrary haplotype labels instead of sorting parent1 by lower A fraction.
Parent-specificity	Pool haplotypes before QC	0.5233, 5.0109, 0.5233, 5.0109	no	Stage 3	Averages away the parent-specific admixture histories.
Parent-specificity	Pool haplotypes after cleaning	0.4988, 3.6207, 0.4988, 3.6207	no	Stage 3	Still collapses the lower-A and higher-A haplotypes into one estimand.

Decision point	Analysis / ablation	Quantitative output	Pass?	Failure point	Why the approach is wrong
Label harmonization	Posterior-only filtering without chr3 correction	0.4820, 3.1075, 0.5653, 5.6131	no	Stage 2	Removes the same low-posterior suspects but leaves the chromosome-3 A/B ancestry labels inverted.
Label harmonization	Low-complexity-only filtering without chr3 correction	0.4820, 3.1075, 0.5653, 5.6131	no	Stage 2	Removes the same low-complexity suspects but leaves the chromosome-3 A/B ancestry labels inverted.
Label harmonization	No chr3 label correction	0.4794, 2.4179, 0.5674, 4.9168	no	Stage 2	Leaves the chromosome-3 A/B ancestry labels inverted relative to other chromosomes.
Map denominator	Use cleaned tract length for L	0.2734, 3.2075, 0.7248, 6.4076	no	Stage 4	Uses called span rather than the full recombination-map length.
Map denominator	Use endpoint span for L	0.2734, 3.1604, 0.7248, 6.2948	no	Stage 4	Uses observed endpoint span and misses uncalled/telomeric map length.
Compound omission	No label correction and no map denominator	0.4794, 2.5529, 0.5674, 5.2069	no	Stages 2 and 4	Combines chr3 label inversion with the wrong timing denominator.

Table 2: Unified decision-point and ablation walkthrough for parent-specific recent-pulse ancestry timing.

11 References

1. Avadhanam S, Williams AL. Simultaneous inference of parental admixture proportions and admixture times from unphased local ancestry calls. *American Journal of Human Genetics* 109(8):1405–1420 (2022). doi: 10.1016/j.ajhg.2022.06.016; PMCID: PMC9388397.

2. Gravel S. Population genetics models of local ancestry. *Genetics* 191(2):607–619 (2012). doi: 10.1534/genetics.112.139808.

3. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *American Journal of Human Genetics* 93(2):278–288 (2013). doi: 10.1016/j.ajhg.2013.06.020; PMCID: PMC3738819.

4. Avadhanam S, Williams AL. Phase-free local ancestry inference mitigates the impact of switch errors on phase-based methods. *G3: Genes, Genomes, Genetics* 15(8):jkaf122 (2025). doi: 10.1093/g3journal/jkaf122; PMCID: PMC12341942.

5. Wei Y, Zhi D, Zhang S. Recomb-Mix: fast and accurate local ancestry inference. *Bioinformatics* 41(Supplement_1):i180–i188 (2025). doi: 10.1093/bioinformatics/btaf227; PMCID: PMC12261469.

6. Zhang Y, Zhang H, Wu Y. A general approach for inferring the ancestry of recent ancestors of an admixed individual. *Proceedings of the National Academy of Sciences* 121(2):e2316242120 (2024). doi: 10.1073/pnas.2316242120.