Signatures of gene conversion in the human two-site frequency spectrum

Matthew D. Schwartz[1] **and Michael M. Desai**[2,†]

[1]Department of Physics, Harvard University, Cambridge, MA 02138, USA, [2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

ABSTRACT Demographic inference from population genomic data typically relies on the Kingman coalescent. However, recent work has argued that departures from Kingman genealogies (e.g. multiple-merger coalescents that arise due to sweepstakes reproduction or pervasive linked selection) are widespread, and can lead to misleading inferences of demography from single-site statistics. Recent work has shown that the joint frequency spectrum of pairs of linked sites (2-SFS) is sensitive to such departures, and used this joint spectrum to reject the Kingman coalescent genome-wide in *Drosophila melanogaster*. Here, we analyze this two-site joint frequency spectrum in humans, using 5.7 billion pairs of nearby sites in 113 Gambian individuals from the 1000 Genomes Project. We find that the 2-SFS of closely linked pairs is inconsistent with the Kingman coalescent under any demographic history. However, multiple-merger (specifically Beta-coalescent) genealogies do not explain the discrepancy. Within bounds set by their measured rates, neither do recombination-map error, sequencing error, population structure, or linked selection. Instead, we find that a Kingman model with gene conversion best explains the data, though some discrepancies remain unresolved. This stands in contrast with data from *Drosophila*, where we found that conversion cannot explain deviations from Kingman expectations, and a model involving multiple merger coalescence is favored instead. Finally, our approach allows us to estimate gene conversion rates from human population genomic data; we find best-fit rates that are of the same order as (but for short tract lengths below) pedigree and sperm-typing estimates.

KEYWORDS two-site frequency spectrum; gene conversion; demographic inference; multiple-merger coalescents; human population genomics

1. Introduction

The Kingman coalescent (Kingman 1982) underlies most methods aiming to infer demographic history from genetic variation. This Kingman process describes the ancestry of a sample by an exchangeable binary branching tree, with branching rate at a given point in the past inversely proportional to the population size at that time. Based on this model, we can infer the demographic history of a population from the distribution of branch lengths, which in turn can be estimated from the patterns of genetic diversity in a set of individuals sampled from the population.

Most approaches to this inference problem focus on the statistics of diversity at individual sites, which can be summarized by the site frequency spectrum (SFS, defined as the fraction of polymorphic sites where the derived allele is observed in i out of n sampled genomes). Numerous widely used computational methods have been introduced to implement this demographic inference, typically under the assumption that the SFS represents the statistics of many independent samples from the Kingman process (as expected if loci are generally unlinked) (Sawyer and Hartl 1992; Nielsen 2000; Marth et al. 2004; Gutenkunst et al. 2009; Excoffier et al. 2013; Liu and Fu 2015; Jouganous et al. 2017; Kamm et al. 2020).

However, recent work has argued that violations of the basic assumptions underlying the Kingman coalescent may be widespread. For example, linked selection or a variety of other processes can lead to multiple-merger processes, so that the statistics of sample genealogies are described instead by Beta or Xi coalescents (Schweinsberg 2003; Durrett and Schweinsberg 2005; Eldon and Wakeley 2006; Neher and Hallatschek 2013; Desai et al. 2013; Birkner et al. 2013; Tellier and Lemaire 2014). When this is the case, inference methods based on the King-

Manuscript compiled: Friday 25th September, 2026
†Corresponding author: [to be set by the authors].

man coalescent are misspecified, and several recent studies have shown that this can lead to misleading results (Eldon et al. 2015; Matuszewski et al. 2018; Sackman et al. 2019; Freund et al. 2023).

Such violations are difficult to diagnose based on the SFS alone: Kingman-based demographic models often provide good fits to the SFS, despite inferring entirely spurious demography, when the underlying process is not Kingman (Eldon et al. 2015; Koskela 2018; Freund et al. 2023). Therefore, analyzing aspects of genetic variation that cannot be computed from the SFS is desirable if we aim to identify violations of the Kingman assumptions. In recent work, Fenton et al. (2025) proposed using the two-site frequency spectrum (2-SFS, defined as the joint distribution of derived allele counts at pairs of linked sites) for this purpose. Their essential idea was that the 2-SFS is sensitive to deviations from the branching structure of the Kingman coalescent: these deviations affect the probability of an ancestral branch subtending j leaves conditional on the existence of an ancestral branch subtending i leaves. They developed a statistical test to detect this signal and applied it to *Drosophila melanogaster*, identifying genome-wide deviations from Kingman assumptions.

Here, we apply this approach to test for deviations from the Kingman coalescent in humans. Specifically, we compute the 2-SFS from the high-coverage 1000 Genomes genotypes of 113 unrelated Gambian individuals (Byrska-Bishop et al. 2022). This involves a total of 5.7 billion pairs of nearby sites across the autosomes, binned by genetic map distance; our primary analysis uses the bin of most closely linked pairs. We chose to focus on these Gambian individuals because they represent the largest set of unrelated individuals from a single population in the high-coverage 1000 Genomes panel. We find that the data are inconsistent with the Kingman coalescent with any single-population demographic history, assuming recombination rates set based on estimates from the deCODE map. However, we also find that deviations in branching structure expected under the multiple-merger (Beta-coalescent) models we tested do not resolve this inconsistency. Instead, introducing gene conversion eliminates much of the observed discrepancy. Thus, in contrast to *Drosophila*, the Kingman assumptions are not necessarily violated in human population history, and multiple-merger coalescents need not be invoked.

Under the Kingman coalescent, recombination or gene conversion does not affect the marginal ancestry of each position, so the SFS by construction is insensitive to these processes. The 2-SFS, by contrast, is sensitive to them even under Kingman assumptions, because they affect the joint ancestry of neighboring sites. We exploit this sensitivity to fit demographic history and gene conversion jointly to the data. We test the fitted model's predictions out of sample on a second pair set binned by physical separation. We find that (with some specific exceptions) our inferred gene conversion rates are consistent with the range estimated from pedigree and sperm studies (Jeffreys and May 2004; Halldorsson et al. 2016; Williams et al. 2015; Palamara et al. 2015; Palsson et al. 2025) (we note that previous estimates of these rates from population polymorphism data have been indirect, based on patterns of linkage disequilibrium or heterozygosity (Frisse et al. 2001; Gay et al. 2007; Setter et al. 2022)). We repeated our analysis on a second population from the 1000 Genomes Project (Yoruba) and in an independently sequenced and processed callset (the Human Genome Diversity Project), and found similar results. We also added gene conversion to the null model of the published *Drosophila* test of Fenton et al. (2025) and found that it does not remove the deviation reported there.

2. Materials and methods

2.1. Definition of the SFS and 2-SFS

Following the notation of Fenton et al. (2025), we define the unfolded site frequency spectrum, ξ_i, as the expected number of segregating sites whose derived allele appears in i out of n sampled genomes. Calculating ξ requires knowing which allele at each site is ancestral. When the ancestral assignment is unknown or uncertain, we can instead define the folded spectrum $\eta_i = (\xi_i + \xi_{n-i})/(1 + \delta_{i,n-i})$, defined for $1 \le i \le n/2$.

We next define the two-site frequency spectrum, $\phi_{ij}(d)$, as the fraction of pairs of polymorphic sites at separation d where the derived alleles are observed in i and j of the n sampled genomes. By definition, the two members of a pair are not ordered, so we have $1 \le i \le j \le n-1$. We write ϕ^{data} for the measured spectrum and ϕ^{null} for the expectation under a model. If ancestral assignment is unknown or uncertain, we can instead define the folded 2-SFS, φ_{ij}, analogously to the folding of the SFS described above. Following Fenton et al. (2025), we compute ϕ^{null} under the Kingman coalescent with a given demographic history, measure ϕ^{data}, and compare the two in a metric calibrated by the measured sampling covariance.

Throughout this paper, we consider spectra involving $n = 8$ haploid genomes, which are projected from the complete data (e.g. the 226 haploid genomes from 113 unrelated Gambian individuals in the high-coverage 1000 Genomes resource (Byrska-Bishop et al. 2022)) by exact hypergeometric projection (see Eq. (1) below). The projection averages each site pair exactly over all subsets of 4 of the 113 individuals (see below for details). We implement this projection for computational efficiency, but we note that it is a projection and not a subsample: every site pair contributes its exact expectation over all possible subsets of 4 of the 113 individuals, so no information is thrown away at random and no resampling noise is added. At $n = 8$ the 2-SFS has 28 elements (27 independent quantities given the normalization constraint).

2.2. Population data, filtering, and polarization

We use the high-coverage whole-genome genotypes of the expanded 1000 Genomes Project cohort (Byrska-Bishop et al. 2022), restricted to the 113 unrelated Gambian (GWD) individuals (226 haploid genomes). We chose GWD because it is the largest single-population subset of the panel's 2,504 unrelated individuals. A single population is desirable because population structure and recent admixture themselves distort the joint distribution of allele frequencies at linked sites, and would be confounded with the departures from the Kingman coalescent that we set out to test. We note that although the 113 GWD individuals carry no close relatives by the panel's pedigree criteria, a haplotype-identity screen flags 22 of them as cryptically related or as offspring of consanguineous unions; repeating the entire analysis on the remaining 91 individuals also does not change our conclusions below. Our analysis uses the autosomes; the numbers reported below are from chromosomes 1–15 and 19–22 (chromosomes 16–18 were initially omitted for computational efficiency, but a later check confirms that including them does not change any of our conclusions below). We label variant alleles as ancestral or derived based on the Ensembl EPO whole-genome alignments (Herrero et al. 2016). We assume that this ancestral-allele identification has some error rate, which we treat as a parameter in the modeling below.

We apply a set of baseline filters for all sites included in our

analysis: sites must be segregating within the cohort, have a FILTER field of PASS in the release VCF, be biallelic strict single-nucleotide variants, and have no missing genotypes across all 113 individuals. Requiring a complete genotype in all 113 keeps the projection of Eq. (1) below exact on a common sample, and in this cohort it excludes only 1.88 million of the 112.3 million VCF records on the autosomes analyzed (1.7%). For larger biobank-scale cohorts, it would be relatively straightforward to relax this requirement (instead requiring only that sites in each projected subsample have no missing genotypes in that subsample), at the cost of some additional bookkeeping overhead. We catalog all sites excluded by these filters in Supplementary Data 1. We also excluded regions containing stretches of sites heterozygous in all 113 of 113 individuals (the largest such window spans 83 kb on chromosome 7). These stretches are probably artifacts of duplications absent from the reference genome: reads from both copies of such a duplication align to the single copy in the reference, so that every individual appears heterozygous. Specifically, we dropped every site at which a one-sided exact test of excess heterozygosity (Wigginton et al. 2005) gives $p < 10^{-6}$ (this removes less than 0.25% of all sites, and no more than 1.2% on single chromosomes).

2.3. Computing the two-site spectrum and its covariance

We computed the 2-SFS among three sets of pairs of sites. First, the *map-distance set* bins pairs by genetic distance on the deCODE genetic map (Palsson et al. 2025), with physical separations of up to 100 kb. The lowest bin in this set contains pairs whose local map distance is consistent with zero, and this lowest-map-distance spectrum ("bin 0"), built from 590.3 million pairs, is our primary focus in this paper. Because the map rather than physical distance defines the bin, bin 0 spans a wide range of separations, with 0.10% of its pairs lying at 3–9 bp, 0.93% at 10–99 bp, 6.4% at 100–999 bp, 34.7% at 1–10 kb, and 57.9% at 10–100 kb. The remaining eight bins are logarithmically spaced in map distance, with edges ranging from 1.9×10^{-7} up to 0.95 cM. Collectively, this map-distance set includes 5.7 billion pairs across its nine bins (the largest single map-distance bin alone holds 2.372 billion). Second, the *physical-separation set* bins pairs by physical separation in base pairs on a uniform map: one window for $d \le 2$; then, up to $d = 24$, alternately one window per multiple of three ($d = 3, 6, 9, \ldots, 24$, each isolated singly) and one window per intervening pair of separations (d = 4–5, 7–8, …, 22–23); then four coarse windows, 25–50, 51–100, 101–300, and 301–1000 bp. The set also includes exact measurements at six named separations (200, 650, 1,500, 3,000, 5,500, and 8,500 bp), which extend it beyond 1 kb and are used for the separation-by-separation validation (Section 3.4). Finally, the *fourfold degenerate set* consists of pairs of fourfold-degenerate sites at each separation up to $d = 300$.

For each pair set and bin b we computed the exact hypergeometric projection of the 113 observed individuals to 4 individuals ($n = 8$ haploid genomes):

$$\phi_{ij}^{\mathrm{data}}(b) \;=\; \frac{1}{|\mathcal{P}_b|} \sum_{p \in \mathcal{P}_b} \Pr\big[\,(i,j) \;\big|\; p;\; 226 \to 8\,\big], \qquad (1)$$

where $\mathcal{P}_b$ is the set of site pairs in bin b and the conditioning is on the observed configuration of pair p among the 226 genomes. The summand is the exact probability that the 8 haploid genomes present in a uniform subsample of 4 of the 113 individuals have derived-allele counts (i, j) at the pair. Note that this estimator depends only on per-site allele counts and site positions, so it is insensitive to phasing errors. We project to $n = 8$ because the null-class spectra of the next subsection can then be computed exactly and all 28 cells remain well populated (as above, we note that this is a projection and not a subsample). We then normalize this spectrum, which removes the overall pair count and leaves 27 independent dimensions.

To verify the accuracy of our computation of the projected spectra, we conducted three checks. Specifically, we first compared the projection of Eq. (1) to explicit enumeration of all subsamples on small synthetic cohorts. Second, we compared two independently written implementations run on the same real chromosomes: the production code (which splits each chromosome into overlapping windows, counts the site pairs of each window in parallel with an optimized routine, and merges the windows with an exact correction for pairs that fall in two windows) against a plain single-pass program that enumerates every pair of the chromosome directly. Finally, we checked the model side of the comparison, by comparing the expected two-site spectrum under the Kingman coalescent with a given piecewise-constant history (which we compute exactly numerically, without simulation) against the same quantity tabulated from large numbers of independent `msprime` coalescent simulations of that history. We find that these checks always produce identical results (within Monte Carlo error in the simulations). In addition, we also validated the complete analysis on synthetic data. Genotypes were simulated with msprime from models whose two-site spectra are known exactly (namely Kingman histories with and without recombination, Beta and Dirac multiple-merger coalescents, and a Kingman model with gene conversion), written in the same input format as the real data and processed identically. We found that for each of these models the recovered spectrum agreed with the known one within its Monte Carlo error in every cell. To measure sensitivity to data errors, we also corrupted such synthetic datasets with one kind of error at a time before processing (genotype errors concentrated at rare variants, errors shared between neighboring sites, added multinucleotide mutations, missing-genotype masks at the observed call rates, ancestral-state mispolarization at the measured rates, and cryptically duplicated individuals), and we recorded how far each kind of error moves the spectrum and whether the filters remove the affected sites.

We next used a delete-one-block jackknife to compute the sampling covariance. We tile the genome into 1-Mb blocks, 2,455 of which enter the bin-0 spectrum, and define

$$\widehat{\Sigma} \;=\; \frac{B-1}{B} \sum_{b=1}^{B} \Big(\phi^{(-b)} - \overline{\phi}\Big)\Big(\phi^{(-b)} - \overline{\phi}\Big)^{\top}, \qquad (2)$$

where $\phi^{(-b)}$ is the spectrum recomputed without block b and $\overline{\phi}$ is the mean of the $\phi^{(-b)}$. Each pair belongs to exactly one block (the block containing its leftmost site) so deleting a block removes a disjoint set of pairs and $\phi^{(-b)}$ is well defined. A chromosome's blocks enter the covariance only when all of them are present, because a partially processed chromosome would bias the covariance. Blocks of 1 Mb are wide compared with every separation we analyze, so the jackknife absorbs the strong correlation of overlapping and nearby pairs without our having to model it. At genome scale the resulting median relative standard error per spectrum cell is about 1.2%.

2.4. A membership test for pooled Kingman histories

We first ask whether *any* single-population demographic history can produce the observed spectrum. To search over all histories

at once rather than over a few-parameter family, we represent a history as follows. We measure depth in the genealogy not in generations but by how far coalescence has progressed: if $g(\tau) = \int_0^\tau \mathrm{d}\tau'/2N(\tau')$ is the cumulative coalescence intensity reached by real time τ in the past, we set $t = e^{-g}$, which is simply the probability that a given pair of lineages has not yet coalesced by τ; $t = 1$ is the present and $t \to 0$ the remote past. In this depth variable every Kingman genealogy unfolds in the same way whatever the demography, and the population-size history enters only through how much real time the population spends at each depth, $\mathrm{d}\tau = 2N\,\mathrm{d}t/t$. We discretize depth into 24 equally spaced levels $t_k = k/25$ ($k = 1, \ldots, 24$) and specify a history by non-negative weights w_k, the real time spent at level t_k.

Converted to a conventional demographic history, this is nothing more than a piecewise-constant $N(t)$ with up to 24 epochs: going back from the present, epoch $k = 24, 23, \ldots, 1$ lasts w_k generations and has diploid size

$$N_k \;=\; \frac{t_k\, w_k}{2\,\Delta t} \;=\; \frac{k\, w_k}{2}, \qquad \Delta t = \tfrac{1}{25}, \tag{3}$$

so that epoch k begins $\tau_k = \sum_{k'>k} w_{k'}$ generations ago. All times and sizes also carry one common scale factor, which a normalized spectrum cannot determine and which we fix from nucleotide diversity whenever a history is plotted in generations (e.g. Figure 1). A large weight at a level is an epoch that is long or has large N, during which lineages accumulate branch length while coalescing little; a small weight is a depth range the history passes through quickly (a bottleneck). The 24-epoch resolution we use is finer than the data require: the fitted weights we find below occupy only a handful of levels, and doubling the number of levels to 48 lowers the best-fit χ^2 of Eq. (5) by only 0.3% (halving it to 12 raises it by 10%).

For such a history the expected two-site spectrum of a pair of fully linked sites is the quadratic form

$$\phi_{ij}^{\mathrm{null}}(\mathbf{w}) \;\propto\; \sum_{a,b=1}^{24} w_a w_b\, C_{ij}^{(ab)}, \qquad w_a \ge 0, \tag{4}$$

normalized to unit sum, which follows in two steps. First, fully linked sites share one genealogy, and a mutation seen in i of the n genomes must have fallen on a branch of that tree subtending exactly i of them. Writing $N_i(t)$ for the number of such branches present at depth t and $L_i = \int N_i\,\mathrm{d}\tau$ for their total length, the probability that two independent mutations on the same tree produce the pair of counts (i, j) is proportional to $\mathbb{E}[L_i L_j]$, i.e. to the double integral of $\mathbb{E}[N_i(\tau)\, N_j(\tau')]$ over the two mutation times. Second, changing the integration variable from real time to depth, with $\mathrm{d}\tau = w_a$ at level t_a, turns that double integral into the double sum of Eq. (4) with

$$C_{ij}^{(ab)} \;=\; \tfrac{1}{2}\Big(\mathbb{E}\big[N_i(t_a)\, N_j(t_b)\big] + \mathbb{E}\big[N_j(t_a)\, N_i(t_b)\big]\Big),$$

the expectation over Kingman genealogies of n genomes viewed at two depths of the same tree, symmetrized because the cell (i, j) does not record which site is which. These coefficients contain no demography at all (the history enters Eq. (4) only through the weights), and we compute them exactly, once, from the jump chain of the Kingman coalescent; they are polynomials in t_a and t_b with rational coefficients. The diagonal blocks $C^{(aa)}$ are, once normalized, the two-site spectra of two-epoch histories. The off-diagonal blocks couple the lineage configuration at one depth to the configuration at another depth of the same genealogy; the two are dependent (singleton lineages that have already merged by depth t_a cannot reappear at a deeper t_b), and it is this coupling that distinguishes a genuine multi-epoch history from a mere superposition of two-epoch spectra.

Equation (4) is thus the zero-recombination limit: the two sites are taken to share a single genealogy. Recombination between the two sites, which would give them distinct but correlated genealogies, is omitted from this null. For the bin-0 pairs analyzed here the map distance is consistent with zero by construction. Crossover at the map rate enters explicitly once separations reach the kilobase scale (Section 3.4) and we treat the effect of any crossover hidden below the map's resolution as one of the alternative candidate processes (Section 3.2; Table 1, crossover-map error).

Statements below that exclude the whole class of possible demographic histories, rather than only its best-fitting member, rest on a guaranteed minimum distance. We compute the smallest χ^2 distance from the data to *any* spectrum the class can produce; because every weighted average of two allowed spectra is itself allowed, this minimization has no local optima and its result is a true floor rather than the best of many tries. When even that floor exceeds the rejection threshold, no member of the class fits, whatever its history. We deliberately take the class to be closed under such averaging: the two-site spectrum of a genome in which different regions have experienced different effective histories is the pair-weighted average of the regional spectra, so we test the data against every mixture of Kingman histories across regions, of which a single genome-wide history is the special case (in mathematical terms this set of mixtures is the convex hull of the single-history spectra, and the floor is found by convex optimization). The mixtures include, in particular, every model in which different regions have experienced different effective histories (e.g. due to background selection, which can rescale the local effective population size differently in different regions of the genome), with the two sites of any one pair (at most 100 kb apart) sharing their region's history. A rejection against the hull therefore holds at once for every single history and for every region-to-region mixture of histories, which is more general than the single genome-wide $N(t)$ assumed by most demographic fits to the site frequency spectrum. We do not, however, estimate region-specific histories (e.g. background selection B-values (McVicker et al. 2009)).

To find the best-fitting member of the pooled-Kingman class, we minimize a chi-squared statistic in the jackknife metric. We also refer to this as the whitened metric: whitening means rotating and rescaling by the measured covariance so that every direction of the 27-dimensional cell space carries unit noise, and distances are meaningful in units of the measurement error. The statistic is

$$\chi^2 \;=\; \min_{\mathbf{w}\ge 0} \Big[\phi^{\mathrm{data}} - \phi^{\mathrm{null}}(\mathbf{w})\Big]^{\top} \widehat{\Sigma}_{0.02}^{-1} \Big[\phi^{\mathrm{data}} - \phi^{\mathrm{null}}(\mathbf{w})\Big], \tag{5}$$

where $\widehat{\Sigma}_{0.02}$ is the jackknife covariance of Eq. (2) restricted to the 27-dimensional sum-zero subspace and regularized by 2% shrinkage, which stabilizes the inversion. We minimize by the Levenberg–Marquardt algorithm (Moré 1978) as implemented in SciPy (Virtanen et al. 2020) from 24 dispersed restarts, which guard against local minima of the constrained fit. Because the demographic weights are fitted under non-negativity constraints, the statistic does not follow a textbook χ^2 distribution, and we calibrate it by parametric bootstrap. We draw replicate spectra from the fitted null with Gaussian noise of the measured

jackknife covariance, refit each replicate under the same non-negativity constraints as the data, and take the 95th percentile of the replicate statistics as the rejection threshold (quoted below as the matched-null 95th percentile beside each χ^2). The construction of the class-wide lower bounds described above and the per-model results are given in Supplementary Data 2.

We also reimplemented the published procedure of Fenton et al. (2025). This procedure first fits a five-epoch demography to the one-site spectrum, then simulates null two-site spectra at separations 3 to 24 bp, with a golden-section search over the scaled recombination rate that keeps the KS-minimizing rate, and compares the multidimensional KS statistic against 1,000 multinomial resamples. On the fourfold degenerate set, which does not depend on the genetic map, this procedure rejects its own best-fit Kingman demography at $p = 0.006$ with the fitted rate interior to the search range. On the physical-separation set restricted to the same separations, which contains all site pairs at those distances, the KS statistic is larger still (13.7 times the 95th percentile of its resampling distribution, $p = 0.001$), but there the fitted rate moves to the upper edge of the search range and remains at the edge when the range is widened. Because a single fitted rate cannot represent a pair set whose per-pair scaled rates span orders of magnitude, we do not interpret this second comparison as a test of the Kingman class. The two designs are complementary: the published procedure fixes the demography from the one-site spectrum and fits the rate; the membership test searches every demography at once so that no choice of history can absorb the deviation.

2.5. Simulated alternative models

To analyze whether alternative models of evolutionary history are consistent with the data, we simulate these models using msprime (Baumdicker et al. 2022). We then compute expected two-site frequency spectra by running these simulations through the same estimator, projection, and whitening as the data. For the case of gene conversion, we use msprime's native model: tracts initiate at rate g per bp per generation and have geometrically distributed lengths of mean L, matching the standard empirical description of conversion tracts. We express conversion strength as the rate $\lambda = gL$, the per-bp, per-generation probability that a site lies inside a transferred tract. Each simulation covers two windows separated by d, with $d = 10$ kb for the primary simulations, and flanking margins of $\max(3000, 10L)$ bp, so that tracts initiated outside the windows still reach them and no edge truncation biases the rate. We consider a range of initiation rates and a series of tract scales, $L \in \{55, 110, 220, 300, 500, 1000\}$ bp, and repeat the simulations at each separation d used in Section 3.4 (588 million simulated replicates in all). For each specific set of parameters, we compute the difference between the simulated spectrum and its own best pooled-Kingman fit under Eq. (5). The subtraction removes whatever a Kingman refit could absorb and leaves only the component that distinguishes the process from a change of demographic history. Alternative-process models are built the same way: crossover recombination at a range of assumed rates representing error in the genetic map, from the best-fitting rate up to and beyond physically realistic map rates; multiple-merger genealogies from Beta coalescents (Schweinsberg 2003) with $\alpha = 19/10, 3/2$, and $11/10$, from nearly Kingman to strongly skewed (other Λ- and Ξ-coalescents were not simulated); a sequencing-error model; population-structure models; linked-selection models; and a mispolarization model. A set of random directions in the whitened space gives the level of alignment expected by chance, and in the Results we compare each of these displacements with the observed deviation.

2.6. Fitting the Kingman model with gene conversion

The model we fit to the data is a Kingman model with gene conversion (hereafter the gene-conversion model): it retains Kingman genealogies under a fitted demographic history and adds gene conversion and ancestral-allele mislabeling (Section 3.2). Its spectrum is

$$\phi^{\text{model}} = F_\varepsilon\left[\phi^{\text{K}}(\mathbf{w}) + \lambda_{\text{short}} D_{L_{\text{short}}} + \lambda_{500} D_{500}\right]. \quad (6)$$

Here $\phi^{\text{K}}(\mathbf{w})$ is the pooled-Kingman spectrum of Eq. (4) with crossover recombination at the deCODE map rate. For the bin-0 pairs that rate is consistent with zero, so in the primary fit to bin 0 the spectrum ϕ^{K} reduces to the zero-recombination form of Eq. (4). In Section 3.4 we also fit the crossover rate on the physical-separation set (with the gene conversion parameters held at these values) and also scan it inside this joint fit. The two D_L terms represent gene conversion at two different scales (short and long); within each scale we assume tract lengths are exponentially distributed with mean L. Two tract-length classes are used, a short class with $L_{\text{short}} = 110$ bp (the pedigree estimate of Palsson et al. (2025)) and a long class with $L = 500$ bp (the scale favored by the tract-length scan of Section 3.2). Their rates λ_{short} and λ_{500} are fitted. Finally, F_ε applies ancestral-allele misidentification: a fraction ε of sites carry their ancestral and derived labels reversed, sending frequency i to $8 - i$. We assume each site is mislabeled independently, so we have

$$\begin{aligned}(F_\varepsilon \phi)_{ij} &= (1-\varepsilon)^2 \phi_{ij} + \varepsilon(1-\varepsilon)\left[\phi_{(8-i)\,j} + \phi_{i\,(8-j)}\right] \\ &\quad + \varepsilon^2 \phi_{(8-i)(8-j)},\end{aligned} \quad (7)$$

where we use the convention $i \le j$. This term describes errors made in the polarization step and is included because assignment errors of 1.5–5.2% are documented for parsimony-style ancestral labeling in the worst-affected frequency bins (Keightley and Jackson 2018), and because the two-site spectrum constrains this rate once it is a free parameter of the fit (Section 3.3).

The fit minimizes the jackknife-metric statistic of Eq. (5) with ϕ^{null} replaced by ϕ^{model}, jointly over the 24 demographic weights and the three non-demographic parameters $(\lambda_{\text{short}}, \lambda_{500}, \varepsilon)$. Because the demography is refit simultaneously with these parameters, no part of the deviation that a change of history could absorb is credited to conversion. The one discrete choice (the long-class tract scale) is made before this fit, using the tract-length scan. Parameter errors are profiled: each parameter is stepped along a scan, and all remaining parameters, including the demographic weights, are re-minimized at every step. The one-sigma interval is taken from a unit rise of the profiled statistic, so the errors account for the uncertainty in demography. Fitted values and the corresponding independent expectations are reported in Table 2, together with the correlation between the two conversion rates.

2.7. Identifiability of the demographic weights

The 24 weights of Eq. (4) parametrize the demographic history and are not 24 separately measured numbers, so we characterized their identifiable content directly. The weights act as a discretized measure over time at the level of pair non-coalescence, so distinct weight vectors that induce the same measure produce

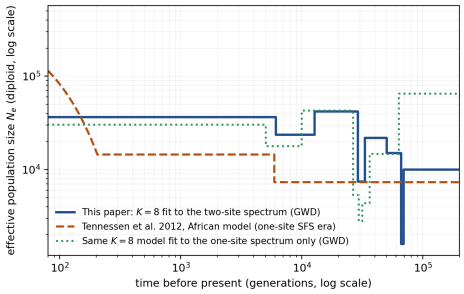

Figure 1 The fitted demographic history against a published African model. The eight-epoch history $N(t)$ built from the identifiable combinations of the fit's 24 demographic weights (solid), with the African model of Tennessen et al. (2012) fitted to one-site spectra (dashed). The two agree on the gross structure: an effective size near 10^4 in the deep past and a several-fold larger size in the recent past. Individual epochs of the fitted curve are not separately measured: the data determine about seven combinations of the 24 weights, more than ten directions are unconstrained, and the overall time scale is set by the assumed mutation rate (so narrow features, such as the deep dip near 6×10^4 generations, lie along weakly constrained directions). The plotted history is the one checked by forward simulation: simulating the gene-conversion model with this history reproduces the fitted spectrum. The dotted curve is the same eight-epoch model fitted instead to the Gambian one-site spectrum alone (the standard demographic-inference exercise, with the same smoothness rule); it agrees with the two-site fit over the last $\sim 2.5 \times 10^4$ generations and departs from it at greater depths, where the one-site spectrum constrains the history only weakly.

the same spectrum. The data determine about seven linear combinations of the 24; more than ten directions are unconstrained, and the overall time scale is set by the assumed mutation rate alone. A smooth eight-epoch history $N(t)$ built from the identifiable combinations fits the spectrum slightly better than the discrete 24-weight history itself ($\chi^2 = 363.5$, marginally below the value for the 24-weight fit reported in Section 3.3); Figure 1 compares this history with a published African model (Tennessen et al. 2012). Simulating the gene-conversion model under that history in msprime (Baumdicker et al. 2022), with the fitted conversion and mislabeling parameters, reproduces the fitted spectrum within simulation noise ($\chi^2 = 0.399$ against a matched-null 95th percentile of 1.252 from 12 million replicates). A concrete single-population history is therefore available for use in standard simulators.

2.8. Replication with Yoruba and HGDP data

We repeat the identical analysis on two other data sets. The first is the Yoruba (YRI) cohort of the 1000 Genomes Project; because 99.6–99.8% of the variant sites in each derived-allele frequency class that enters the projected $n = 8$ spectrum are shared between the two cohorts, the two estimates are correlated by construction. We also analyze the independent Human Genome Diversity Project (HGDP) callset (Bergström et al. 2020) on the same nineteen autosomes. This callset comes from a different cohort, was sequenced at different centers, and was processed with a different variant-calling pipeline; we run it through the same estimator and metric as the data described above.

2.9. Haplotype-identity cluster test

The two-site spectrum summarizes genealogies through allele counts. To examine the *Drosophila* genealogies more directly we also use a statistic of haplotype sharing, which needs resolved haplotypes. For *D. melanogaster* we use the 197 haploid-embryo genomes of the DPGP3 Zambia sample (Lack et al. 2015), in which haplotypes are observed without phasing. We restrict them to fourfold degenerate sites in the central region of each autosome arm (2L 1–17 Mb, 2R 6–19 Mb, 3L 1–17 Mb, 3R 10–26 Mb) and of the X (2–20 Mb; 196 genomes), keep uncalled bases as missing, and obtain 67,000–83,000 variable biallelic sites per arm. The unphased Gambian genotypes used elsewhere in this paper do not resolve haplotypes. For this test we therefore take the 226 haplotypes of the same 113 individuals from the statistically phased 1000 Genomes panel (Byrska-Bishop et al. 2022), in which 111 of the 113 are parents of sequenced trios, so that their phase is largely pedigree-determined. That panel omits variants seen only once among its 3,202 genomes. We therefore keep only sites whose minor allele is carried by at least two of the sampled haplotypes, in the data and in every simulation alike. We never compare the two haplotypes of one individual. We also remove 22 individuals with the excess haplotype sharing expected of the offspring of consanguineous unions or of cryptic relatives. For each pair of haplotypes we sum, over the nineteen autosomes analyzed above, the number of windows (defined below) in which the two are identical; an individual is flagged if this count for its own two haplotypes, or for one of its haplotypes and one of another individual's, exceeds the median over haplotypes of distinct individuals by more than six robust standard deviations (about 14%), and of two individuals related in this way one is removed. This leaves 182 haplotypes. We also analyze the haploid, non-pseudoautosomal X chromosomes of the 55 males, which need no phasing.

On each arm or chromosome we cut the ordered variable sites into consecutive non-overlapping windows of $w = 50$ sites ($w = 25$ and 100 as checks). A window spans a median of 3.8–4.7 kb on the fly arms, 6.9–8.5 kb on the human autosomes and 16 kb on the male X, and identity within it tracks the age of common ancestry more closely than identity over a fixed physical length would. A genome is retained in a window if at least 80% of its sites are called there. Two retained genomes are identical in the window if they are jointly called at $0.8w$ or more sites and differ at none of them. For every window we record the number P of identical pairs, the identity fraction f (P over the number of comparable pairs) and the size $c_{\max}$ of the largest cluster of mutually identical genomes. Recent growth raises P and $c_{\max}$ together and enters the null through the fitted history, whereas a multiple merger in a window's genealogy enlarges the clusters at a given number of identical pairs. We therefore compare data and null at fixed identity. Windows with $P > 0$ are binned by f, with bin edges at 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2 and 0.5, and a bin is used if it holds at least 20 windows in the data and in every null replicate. The score is

$$Z = \frac{1}{\sqrt{B}} \sum_{b=1}^{B} \frac{\bar{c}_b^{\,\mathrm{data}} - \langle \bar{c}_b \rangle}{\sigma(\bar{c}_b)}, \tag{8}$$

where $\bar{c}_b$ is the mean of $c_{\max}$ over the windows in bin b, B is the number of bins used, and $\langle\cdot\rangle$ and σ are the mean and standard deviation of $\bar{c}_b$ over the null replicates. Because σ is the replicate-to-replicate scatter of an average over hundreds of windows, a modest per-window effect gives Z in the tens. Scoring each null replicate against the others gives Z with mean 0.0 and standard

deviation 1.0–1.2 in the fly comparisons (maximum 3.3 among 200 arm replicates) and 1.3–1.4 in the human ones. We always report Z together with the size of the excess in $c_{\max}$. The coarse bins absorb a mismatch between data and null in the overall level of identity only to first order. As a measure insensitive to that level we therefore also match windows on f within twenty quantile cells of the pooled null distribution ($0.002 \le f < 0.5$) and average $(\bar{c}^{\text{data}} - \bar{c}^{\text{null}})/\bar{c}^{\text{null}}$ over cells with data weights. We call this the fine-matched excess; its null scatter comes from scoring each replicate against the rest.

Fly null replicates are samples of 197 genomes (196 on the X) simulated with msprime (Baumdicker et al. 2022) under the Kingman coalescent with haploid samples, in independent 1-Mb segments tiling the analyzed region of each arm. Each uses the arm's five-epoch history fitted to the folded one-site spectrum, at the consistent time unit of Section 3.6, and uniform crossing-over at the arm's fitted scaled rate. Mutations follow a Jukes–Cantor model whose rate is calibrated so that the fraction of fourfold degenerate positions that are variable after masking matches the data within 1%. Genotypes are read at the real fourfold degenerate positions only, each genome's missing bases are imposed on the corresponding simulated genome, and fifty replicates per arm pass through the same computation as the data. For the two-site statistic, each replicate's five-epoch history is refit to its own one-site spectrum and its two-site spectrum is compared with the Kingman expectation under that refit history, as for the data in Section 3.6. For this comparison the folded two-site spectrum of all 197 genomes at separations of 3, 6, …, 24 bp is binned into six minor-allele-count classes (1, 2, 3–4, 5–8, 9–19, ≥ 20), which gives 168 class-pair cells. The statistic χ^2_{K} sums over these cells the squared deviation from the Kingman expectation under the fitted history, in units of the cell's 100-kb block-jackknife standard error, and the per-arm two-site excess is $Q = \chi^2_{\text{K}} - \nu$ with $\nu = 160$ degrees of freedom. Because each cell entering Q is scaled by a standard error that includes the correlation between neighboring site pairs, Q has a common scale across arms and nulls, which the KS distance of Section 3.6 lacks; we report the two side by side. Positive controls are simulated in the same way under Beta$(2-\alpha, \alpha)$ coalescents (Schweinsberg 2003), with $\alpha = 1.9$ and 1.7 on the fitted size schedule and $\alpha = 1.9$ at constant size, three per arm and setting, and with the mutation rate recalibrated to the observed density. They match the data in variant density only, so their scores indicate direction and approximate power. To ask whether one Beta coalescent reproduces both fly statistics at once, we also built Beta-coalescent nulls for $\alpha \in \{1.5, 1.6, 1.7, 1.8, 1.85, 1.9, 1.95\}$. At each α the five-epoch history is refit so that the processed one-site spectrum of the simulations matches the observed one (two fixed-point iterations), and the mutation rate is then recalibrated under that history, giving a simulated variable-site density 0.97–1.07 times the observed. Fifteen replicates per arm and α go through both computations, extended to 30 wherever the first fifteen left the observed Q within 2.5 replicate standard deviations of the simulated mean.

For the human sample, each chromosome's five-epoch history is fitted to the folded spectrum of the 226 haplotypes and simulated as 226 haploid sequences in independent 1-Mb segments at the consistent time unit, with the mutation rate calibrated to the observed variant density. We use three treatments of crossing-over: a uniform scaled rate ($\rho = 4N_e r = 8 \times 10^{-4}$ per bp); the fine-scale GWD recombination map of Spence and Song (2019) rescaled to that mean; and the same map with its overall level lowered until the simulated identity fraction approaches the observed one. The third treatment drives the crossover rate essentially to zero and still leaves the simulated identity fraction 15% below the data in the median (within 3% on 4 of 19 chromosomes); for the male X we lower a uniform rate in the same way. Ten replicates per chromosome and treatment are used, with Beta-coalescent controls ($\alpha = 1.9$ and 1.7) on four chromosomes. We measure the persistence of pair identity across the genome by the rank correlation, over pairs of individuals, between their identity counts on different chromosomes, averaged over chromosome pairs and referred to 2,000 permutations of individual labels; the fly version of this statistic correlates pair identity between arms of chromosomes 2 and 3. To interpret the human results we also simulated two kinds of structure on six chromosomes. The first is an island model with two to eight demes and migration set to give between 0.2 and 6 times the observed persistence; the second is a single admixture pulse of 3% or 8% from a lineage that separated from the ancestors of the sample 0.5–2 coalescent time units ago. Each such dataset was matched to the data in variant density, refit with its own single-population history, and compared with replicates of that history.

The window length, identity rule, bins, null construction, controls and decision criterion were specified before any statistic was computed from the data, separately for the fly arms, for the human sample and for a first Beta-coalescent grid ($\alpha \in \{1.5, 1.7, 1.8, 1.9\}$, with the mutation rate calibrated under the initial fitted history). Two parts of the construction described above were adopted after a first pass, and the statistics reported here were computed only after their adoption. For the human sample, the restriction to sites of minor-allele count ≥ 2, the exclusion of within-individual pairs and the removal of the 22 flagged individuals were adopted after a first comparison on all sites and all 226 haplotypes had exposed the panel's missing singletons and the mismatch in identity level reported in Section 3.7. For the Beta-coalescent nulls, the recalibration of the mutation rate under the refit history, the grid points $\alpha = 1.6$, 1.85 and 1.95 and the 30-replicate extension were added after the first grid had run. We take an arm to show the multiple-merger signal if $Z \ge 3$ with the sign of the Beta controls, and *D. melanogaster* to show it if at least three of the four autosome arms do; the X is analyzed in the same way but is not one of the four. Pairs of genomes identical in more windows than any pair in any null replicate are listed and, if they involve at most ten genomes, removed before rescoring. For the Beta-coalescent nulls, a statistic counts as reproduced at a given α if the data lie within two (Q) or three (Z) replicate standard deviations of the simulations on at least three of the four arms. The spectrum-matched and rate-map nulls, the inversion-karyotype pruning and the crossover-rate stratification were added after the primary fly result was known, and the fine-matched excess, the human recombination-map treatments and the structure simulations after the first human comparison; we present them as robustness analyses.

3. Results

3.1. The genome-wide deviation

We began by analyzing the two-site frequency spectrum ϕ^{data} built from the 590 million pairs in bin 0 of the map-distance set (the pairs whose local distance on the deCODE genetic map is indistinguishable from zero). We find that no pooled-Kingman history fits the genome-wide spectrum from this set. The best fit over the entire class, obtained by optimizing the demographic

weights of Eq. (4), misses the data by more than two hundred times the rejection threshold ($\chi^2 = 7{,}697$ in the jackknife metric on the spectrum's 27 degrees of freedom, against a matched-null 95th percentile of 29.2). This margin changes little across the analysis choices we varied (Table 3; the ratio lies between 204 and 272). Because the fit searches every demographic history $N(t)$ at once, no choice of demography can absorb the excess. The published procedure of Fenton et al. (2025), which instead fixes the demography from the one-site spectrum, also rejects its own best-fit Kingman null when applied to the fourfold degenerate set (Section 2).

Figure 2 shows the deviation as the $\log_2$ ratio of the observed spectrum to the model prediction. In the left panel the observed spectrum lies above the best Kingman fit across most of the frequency range, and the excess grows toward pairs of high-frequency derived alleles. The right panel shows the same data against the gene-conversion model described below, where little of this structure remains.

We write $\Delta = \phi^{\text{data}} - \phi^{\text{null}}$ for the difference between the observed spectrum and the best pooled-Kingman fit. Note that Δ is a vector with one component per spectrum cell. Three features of Δ constrain what could have produced it (Figure 3). First, the deviation rises with derived-allele frequency. Cells involving rare alleles lie close to the fit; because the spectrum is normalized, the large excess at high frequency is balanced by small deficits in these abundant low-frequency cells. Cells pairing two high-frequency alleles lie far above the fit, and the most discrepant cell, $(i, j) = (6, 7)$, exceeds the Kingman prediction by a factor of 4.8. Second, and most distinctively, the deviation alternates: diagonal cells (i, i), in which the two sites carry the same derived-allele count, are close to the Kingman fit, while the flanking off-diagonal cells at high frequency exceed it by factors of two to more than four (Figure 3b). Third, the deviation is coherent across the genome. Recomputed one chromosome at a time, the deviation aligns with the genome-wide direction: the cosine between the two in the jackknife metric has median 0.98 across the 19 autosomes used (minimum 0.93). A handful of unusual loci, or one aberrant chromosome, would not produce this pattern, which suggests that the process generating the deviation operates genome-wide.

3.2. Identifying the source of the deviation

We next tested alternative models that include additional processes, to determine whether any of them can explain the deviation described above. We considered seven candidate processes: (i) gene conversion; (ii) an error in the assumed crossover recombination rate (meaning the genetic map is wrong near the measured pairs); (iii) Beta-coalescent multiple-merger genealogies; (iv) sequencing error; (v) population structure; (vi) linked selection; and (vii) mispolarization due to ancestral-allele misidentification. We describe the construction of each process in Section 2; the genealogical candidates were simulated with msprime (Baumdicker et al. 2022). Variation in the mutation rate among sites is not on this list because, under the infinite-sites assumption, it leaves the normalized two-site spectrum unchanged: the expected joint spectrum of a pair is the product of the two sites' rates and a purely genealogical factor, and the rates cancel when the spectrum is normalized. Recurrent mutation at the same site, however, does not cancel: it is most frequent at hypermutable CpG transitions and perturbs allele counts in a way that partly mimics ancestral mislabeling; we return to it by stratifying on mutation class in Section 3.8, where a CpG-transition recurrence component of the mislabeling rate is measured directly. Recurrent mutation is not, however, what drives our results: repeating the entire analysis with every CpG-associated site removed (45% of the closely linked pairs) leaves the rejection of the Kingman class and the fitted gene-conversion parameters unchanged (Section 3.8).

We judge the candidate processes by fitting their simulated displacements (each relative to its own best pooled-Kingman fit) to the observed deviation, singly and in combination, and measuring how much of the deviation each removes in the jackknife metric. We place bounds on each candidate process based on direct measurements wherever possible; where no direct measurement exists we bound the candidate at the amplitude of a deliberately strong simulated scenario. The candidate processes and corresponding bounds are summarized in Table 1.

Figure 4 compares the direction of the displacement each candidate process produces with the direction of the observed deviation. This comparison uses only the direction of each candidate's displacement relative to the pooled-Kingman model (its alignment, in the jackknife metric, with the observed deviation Δ), rather than its amplitude. We find that gene conversion is the only process whose displacement is close in direction to the deviation observed in the data. The displacement produced by an error in the crossover rate aligns less well: at its best-scoring rate it outperforms a random direction, but at physically plausible map rates its alignment is no better than random. All other candidate processes are roughly consistent with the scatter of random directions.

In particular, we note that in contrast to the findings of Fenton et al. (2025) in *Drosophila*, a multiple-merger Beta coalescent is not consistent with the data we analyze here. Instead, the Beta displacement directions lie within the scatter of random directions relative to the observed deviation. Our analysis is well powered against this family: it separates the Beta-coalescent alternatives from the entire pooled-Kingman class by a wide margin (Table 1). Thus if the deviations from the pooled-Kingman class observed in our data were consistent with this type of multiple-merger coalescent, the test would likely have detected it. Ancestral mispolarization by itself also cannot explain the deviation observed in the data, although correcting for it improves the fit of the gene-conversion model (Section 3.3).

These results suggest that gene conversion provides the best account of the observed deviation, and four further comparisons support this conclusion. When we fit the six non-conversion candidate processes simultaneously, within the bounds of Table 1 and without gene conversion, the best combination removes 74.0% of the deviation ($\chi^2 = 1{,}998.5$), whereas the gene-conversion model (Section 3.3) removes 95.2%. When we add the two gene conversion terms, bounded by their measured ranges, to the full set of alternatives, the fit improves substantially ($\Delta\chi^2 = 1{,}821.5$) and the fitted conversion rates lie within the measured bounds. The data thus favor conversion even after the other processes have been accounted for. The improvement also does not depend on readjusting the demographic history: refitting the gene-conversion model with all 24 demographic weights held fixed at their Kingman-only values retains 98.6% of the improvement in χ^2. Lastly, the ranking of candidate processes is insensitive to the simulated background. In head-to-head comparisons of single simulated processes under each of the demographic histories we tried, the displacement predicted by conversion aligns with the observed deviation at least a factor of 2.1 better than that of each alternative. Together, these results

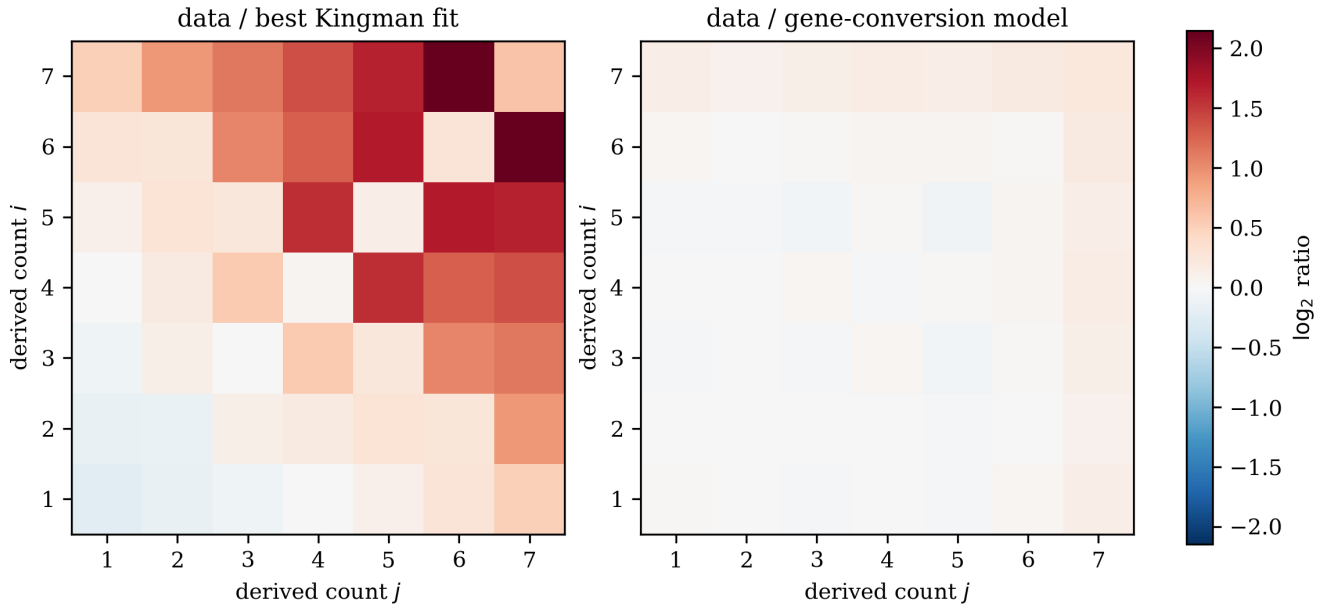

Figure 2 Deviations in the genome-wide two-site frequency spectrum. Log_2 ratio of the observed $n = 8$ two-site spectrum to model predictions for the closely linked genome-wide pair set; i and j are the derived-allele counts at the two sites. Left: the deviation (in the log-ratio form of Fenton et al. (2025)) between the data and the best fit from the entire pooled-Kingman class ($\chi^2 = 7{,}697$ against a matched-null 95th percentile of 29.2). The misfit is structured: it grows toward pairs of high-frequency alleles and peaks in cell $(6, 7)$ at $+380\%$ of the prediction. Right: deviation between the data and the fitted gene-conversion model of Section 3.3, on the same color scale ($\chi^2 = 370.2$ for the analysis shown).

indicate that gene conversion can explain the deviations from the best-fit pooled-Kingman model that we observe in the data, whereas the alternative candidate processes we have considered cannot, at least within the bounds set by their measured strengths.

3.2.1. Tract lengths and the two conversion classes. Our model of gene conversion involves geometrically distributed tract lengths with mean L (the native model in msprime (Baumdicker et al. 2022)). We find that the shape of the displacement this model predicts (across the 28 cells) is sensitive to L, so we can use the fit of our model to the data to estimate the best-fit L. To do so, we simulated a series of single-class conversion processes with mean lengths from 55 to 1,000 bp and, for each L, subtracted the best-scaled displacement from the observed deviation. The residual misfit is smallest at $L \approx 500$ bp (Figure 5). This scale is plausible, given that meiosis lays down a conversion tract at the point where each crossover resolves: if crossover-associated tracts drive the long-range signal, we expect the best single scale to fall between 500 and 1,000 bp.

Direct observations, however, report much shorter tracts (e.g. mean 46 bp from single-sperm long reads (Charmouh et al. 2025), means of 123 and 102 bp (paternal and maternal) from the deCODE trios (Palsson et al. 2025), and 55–290 bp from classic sperm typing (Jeffreys and May 2004)). In the tract-length scan above, mean lengths this short give the poorest fit to the observed deviation, so tracts of the directly observed scale are unlikely to explain the signal by themselves. We therefore fit a model involving two classes of conversion events: a short class at the tract scale observed directly in pedigrees and sperm, and a long class near 500 bp, the scale preferred by the tract-length scan. A two-class tract-length distribution is also consistent with an independent population-scale measurement: from identity-by-descent sharing in the UK Biobank, Masaki and Browning (2025) fit a single geometric tract mean of 459 bp but prefer a two-part mixture with means 16.9 and 724.7 bp. Below, we fit the rates of both of these classes jointly and ask whether the fitted values agree with direct measurements.

3.3. The Kingman model with gene conversion

The comparisons above favor gene conversion as the explanation of the observed deviation from the Kingman class. Here we describe the best-fit model of this kind, the gene-conversion model of Eq. (6). This model retains Kingman ancestry with a fitted demographic history $N(t)$ and crossover at the deCODE map rate, which for the bin-0 pairs of the primary fit is consistent with zero. To this it adds gene conversion in two geometric tract classes (a short-tract class with rate λ_{short} and mean length 110 bp as estimated from pedigree data (Palsson et al. 2025) and a long-tract class with rate λ_{500} and mean length 500 bp), and an ancestral-allele mislabeling rate ε. Beyond the demography, which standard analyses also fit, the model has three fitted parameters (λ_{short}, λ_{500}, and ε) and one discrete tract-scale choice. Fit jointly to the genome-wide bin-0 spectrum, it explains 94.2–96.2% of the deviation, depending on the analysis choices varied below (Table 3); in the reference analysis, with the choices stated in Section 2, χ^2 falls from 7,697 to 370.2. The residual, which is still well above the matched-null threshold, is described in Section 3.8.

3.3.1. Fitted parameters and independent expectations. We summarize the best-fit parameters in Table 2. Our fit constrains the total conversion rate tightly and the split between the classes loosely: the two rate coefficients are anticorrelated at -0.955, so the partition between classes can slide along the constraint while their sum stays fixed. The well-determined quantity is therefore

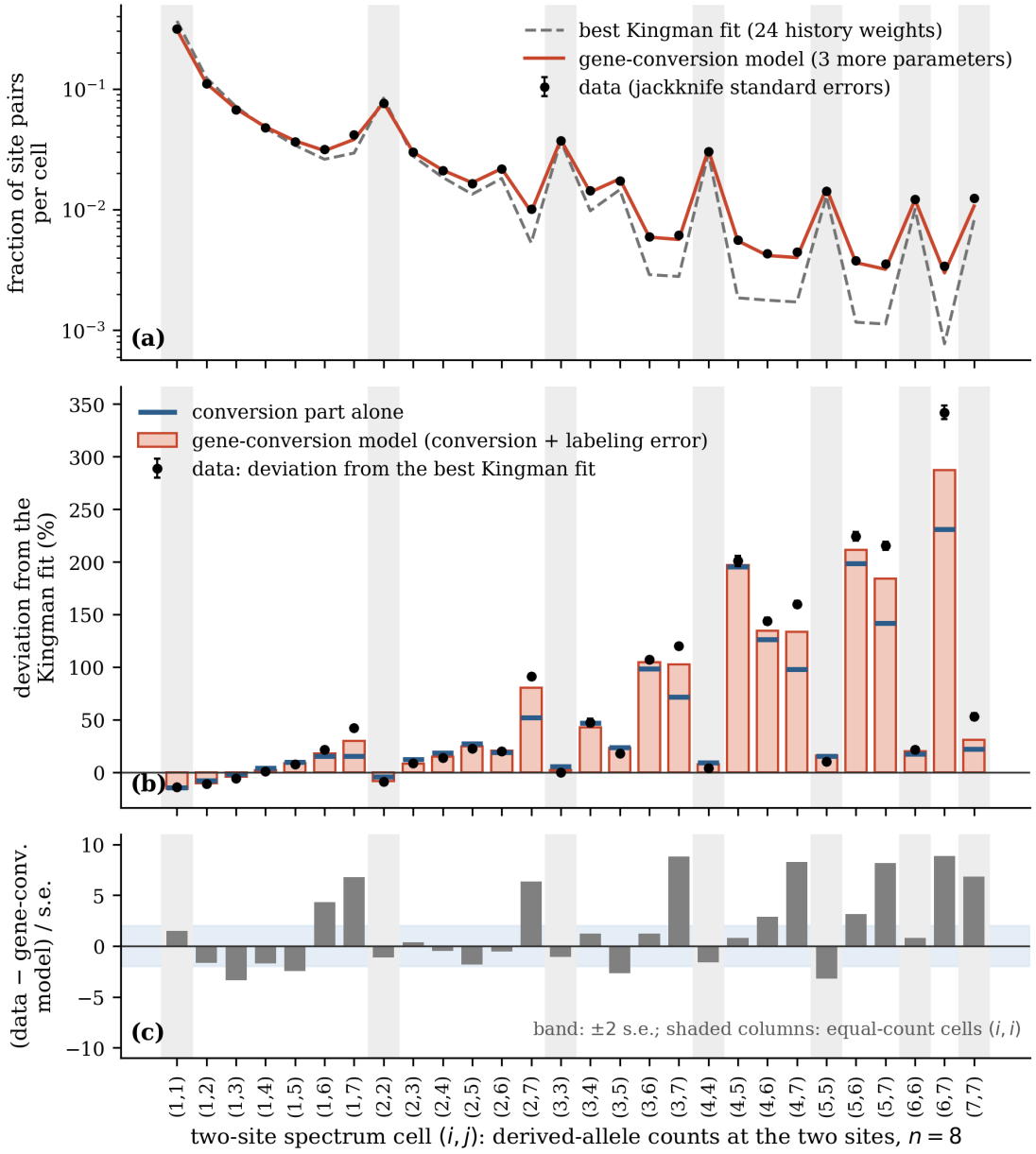

Figure 3 Cell-by-cell deviation of the closely linked two-site spectrum from the pooled-Kingman and gene-conversion models. **(a)** Fraction of site pairs in each of the 28 cells of the $n = 8$ two-site spectrum (bin 0, genome-wide; logarithmic scale). Dashed: the best 24-parameter pooled-Kingman fit; solid: the gene-conversion model of Section 3.3; points: the data with delete-one-block jackknife errors, usually smaller than the markers. **(b)** The same data as the deviation of each cell from the Kingman fit, in percent of the Kingman prediction (points), with the gene-conversion model's prediction (bars) and its gene-conversion component alone (blue dashes); the conversion direction comes from coalescent simulation (Section 2) and is not free to rotate in the fit. Diagonal cells (i, i) (shaded columns) sit near the Kingman fit while off-diagonal high-frequency cells reach two to five times it; the largest, $(6, 7)$, sits at 4.8 times the prediction. **(c)** What is left after the gene-conversion model: we show the data minus the model fit in units of each cell's standard error (band: ± 2). Note that the remaining deviation concentrates in the highest-frequency column; we discuss this in Section 3.8.

the total rate, which we estimate to be $\lambda_{\text{total}} = 8.33 \times 10^{-6}$ per bp per generation (statistical standard error 0.45×10^{-6}; range across reasonable estimator choices 3.8–10.8 $\times 10^{-6}$, Table 3). Neither rate is strongly correlated with the mislabeling term (correlations $+0.12$ and -0.24). The long-class rate is 1.1–1.2 times the rate expected if each crossover is accompanied by one conversion tract of the long-class mean length, the independent expectation listed for λ_{500} in Table 2. The statistical error reflects sampling noise only; the systematic uncertainties, which dominate for the absolute rate, are discussed next.

The systematic uncertainties that accompany the statistical errors of Table 2 are collected in Table 3. The two largest bear directly on λ_{total}, and both follow from how recombination enters the two-site spectrum. The first is the dependence of the

Candidate	Bound	Basis
Gene conversion	per-class rate $\lambda \leq 1.4 \times 10^{-5}$ per bp per generation	twice the deCODE pedigree measurement of 7.0×10^{-6}; the factor of two covers the documented 2.17× maternal–paternal spread (Halldorsson et al. 2016)
Crossover-map error	1.0× the simulated scenarios	the scenarios are neutral two-locus simulations in which the bin-0 pairs, assumed fully linked, in fact recombine at scaled rate $\rho = 4N_e rd = 10^{-4}$, 10^{-3} or 10^{-2} ($N_e = 10^4$); no direct measurement of the map error exists, and the strongest scenario is deliberately large: the bin-0 map edge corresponds to 7.7×10^{-3} of the strongest one
Multiple-merger (Beta)	0.247, 0.048, and 0.033 of the Beta(19/10, 3/2, 11/10) spectra	the amplitude at which the exclusion bound against the pooled-Kingman hull reaches 2σ
Sequencing error	per-call rate ≤ 0.008	the upper end of the measured excess-heterozygosity failure band; generous, since flagged sites are removed from the analysis
Population structure	0.025 to 7.9 in units of the simulated scenario amplitude, per scenario	fourteen neutral structured-coalescent scenarios: a two-deme island model with $Nm = 0.1$–10 sampled from one deme (as GWD is) or from both; a 10% admixture pulse from a deme that diverged 0.5–2 coalescent units ago; and a three-deme stepping-stone with the middle deme sampled. Each bound is the amount of that scenario's displacement actually present in the data (its projection in the jackknife metric) plus the 95% threshold of the same projection under the null
Mispolarization	rate ≤ 0.0183	the fitted mislabeling rate plus two profiled standard errors, the most generous measured rate; an independent ancestral-state error measurement of 0.0145 sits below it
Linked selection	1.0× the simulated scenarios	five simulated scenarios at full strength ($N_e = 10^4$, $\alpha = 2N_e s$): a hard sweep ($\alpha = 1000$) at the locus sampled at fixation and at 0.05–0.4 coalescent units after it; a weaker sweep ($\alpha = 100$) at a random position within the flanking megabase; three recurrent sweeps ($\alpha = 1000$) at random positions and times; a soft sweep from standing variation (initial frequency 0.05); and a background-selection proxy (uniform 30% reduction of N_e). No direct measurement bounds these, so each enters at the full simulated amplitude

Table 1 Bounds on the candidate processes in the combination fits. Each candidate process enters the combination fits as the change ('displacement') it produces in the 28-cell spectrum, obtained by simulating the process at a stated reference strength and subtracting the best pooled-Kingman fit to that simulation; its fitted amplitude is then capped at the physically allowed multiple of the reference strength listed here. We fit the seven listed candidate processes to the data, within bounds on their strength as described here (amplitudes for simulated candidates are in multiples of their simulated scenarios; rates for mispolarization and sequencing error are per site and per call). All fitted amplitudes reported in the text respect these bounds, and we find that in the fit that includes gene conversion all of them lie strictly inside the bounds.

rate estimate on rate heterogeneity across pairs. At fixed demography the spectrum of a pair depends on crossover and conversion only through a single combined quantity, the rate at which recombination events place the two sites on different ancestral lineages (Hudson 1983) (a crossover between them, or a conversion tract covering one site but not the other). In simulations, any mixture of crossover and conversion with the same combined rate produces the same spectrum. A bin, however, pools pairs whose combined rates differ (local rate variation, hotspots, and in bin 0 a wide range of physical separations), and because the spectrum is a nonlinear function of that rate, the pooled spectrum is not the spectrum of the average rate. The data appear to require this heterogeneity to be large, with a best-fitting coefficient of variation of 1.5–1.9 in bin 0 and in the physical-separation set. Our rate estimator matches the observed deviation to templates simulated at a single rate per template, which is exact for homogeneous pairs but not under such heterogeneity. Reasonable alternative estimators applied to the same bin-0 spectrum return a range of total rates (Table 3), with ours near the middle of that range. We therefore regard the absolute value of λ_{total} as determined only to within roughly a factor of two, a much wider range than the statistical error suggests. The rejection of the Kingman class and the identification of conversion do not appear to depend on the choice of estimator. The second effect, unresolved crossover within bin 0, arises because pairs in bin 0 share a deCODE map coordinate, which places only an upper limit on crossover between them at the pedigree map's resolution. Any crossover actually present there is attributed by the fit to conversion; Table 3 gives the resulting downward shift in the inferred conversion rate for a given amount of such crossover. The remaining entries of the table are smaller or affect only the partition between the classes. Among them, the dependence of the short-class coefficient on the strength of background selection (McVicker et al. 2009) bears

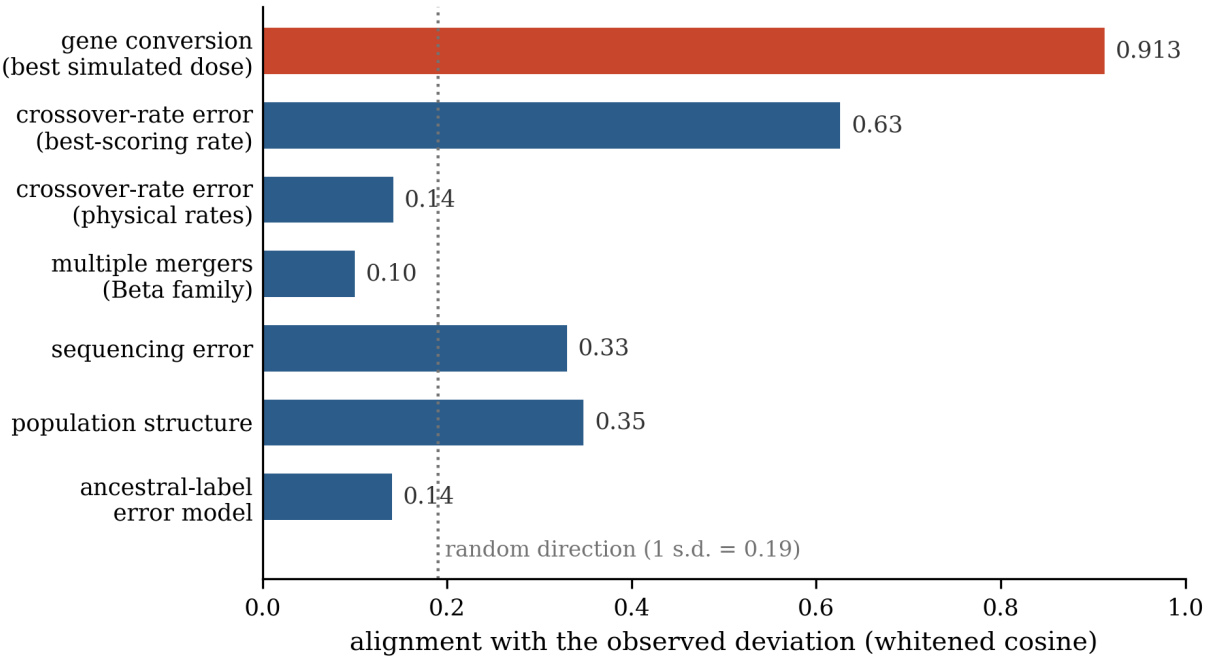

Figure 4 Alignment of each candidate process with the observed deviation. Alignment (cosine in the jackknife metric) between the genome-wide deviation and the displacement each candidate process produces, each with its own best pooled-Kingman fit removed. The dotted line marks the one-standard-deviation spread of a random direction (0.19). Gene conversion reaches 0.913. A crossover-rate error reaches 0.626 only at its best-scoring rate and falls to 0.14 at physical rates; the multiple-merger (Beta) family sits at 0.10, and every remaining alternative scores at or below 0.35.

on the short-tract deficit discussed in Section 3.8.

3.3.2. The ancestral-mislabeling rate. Polarizing the spectrum assigns every site an ancestral allele from the outgroup alignments. When the assignment is wrong, a site with derived count i enters the spectrum at count $8 - i$. The gene-conversion model represents this error by a per-site mislabeling rate ε, applied in the all-orders form of Eq. (7) so that pairs with one or both sites mislabeled are included, and we fit ε jointly with the conversion rates. Although ε describes an error in the ancestral-state assignment and has no genealogical meaning, we treat it as a free parameter. It improves the fit substantially, and mislabeling moves spectrum mass in a direction that the fit can separate from conversion and demography, so the data constrain it. The joint fit gives $\varepsilon = (1.67 \pm 0.09)\%$, inside the 1.5–5.2% band documented for parsimony-style ancestral assignment in the worst-affected frequency bins (Keightley and Jackson 2018). The rate also decomposes: posterior disagreements of the probabilistic est-sfs assignment (Keightley and Jackson 2018), computed with zero free parameters, account for 1.02% of it. Components localized in sequence context supply most of the remainder, and two components remain unexplained (Section 3.8).

We obtain similar mislabeling rates from independent subsets of the data. GC-conservative mutation classes, on which GC-biased gene conversion cannot act, give 1.54%. Transitions and transversions give 1.42% and 0.79% (ratio 1.79, near the substitution-rate prediction of roughly 2; Figure S1), and the 19 chromosomes, fitted separately, give mutually consistent rates. A fit to the folded spectrum, which does not use the ancestral labels and in which single-site flips cancel, gives the same partition of the conversion rate between the two tract classes as the unfolded fit that includes ε. The same subsets also exclude GC-biased conversion as an alternative explanation; the full constructions are given in Appendix A.

3.4. Validation in the physical-separation set

The three parameters of the gene-conversion model (λ_{short}, λ_{500}, and ε) were fixed once, by the fit to bin 0 of the map-distance set (Section 3.3). We simulate the conversion templates in that fit at a reference separation of 10 kb, representative of the pairs in bin 0. With those three parameters held fixed, the model predicts the amplitude of the conversion-attributable deviation at any other pair separation with no adjustable constant. Here, we test those predictions on a *different* pair set, the physical-separation set defined in the Methods. We write $e(d)$ for the ratio of the measured amplitude to the predicted amplitude at separation d, so $e(d) = 1$ means the data match the prediction. We report $e(d)$ first from 200 bp to 8.5 kb and then at separations below 100 bp.

3.4.1. Out-of-sample predictions of the deviation amplitude. At separations where crossover is negligible over the timescales of the sample genealogy, the conversion-only prediction agrees with the data: $e(200) = 1.023 \pm 0.024$ and $e(1500) = 1.026 \pm 0.029$. The result replicates chromosome by chromosome: the conversion-attributable amplitude is positive on 19 of 19 chromosomes at every separation tested, and at $d = 200$ bp the per-chromosome values range from 0.91 to 1.42 with median 1.005. One separation in the crossover-free regime nevertheless fails the comparison: $e(650) = 0.914 \pm 0.030$ after correcting for the mixture of pair separations within the bin, a 9% deficit that none of the corrections we applied removes.

At multi-kilobase separations crossover must enter the pre-

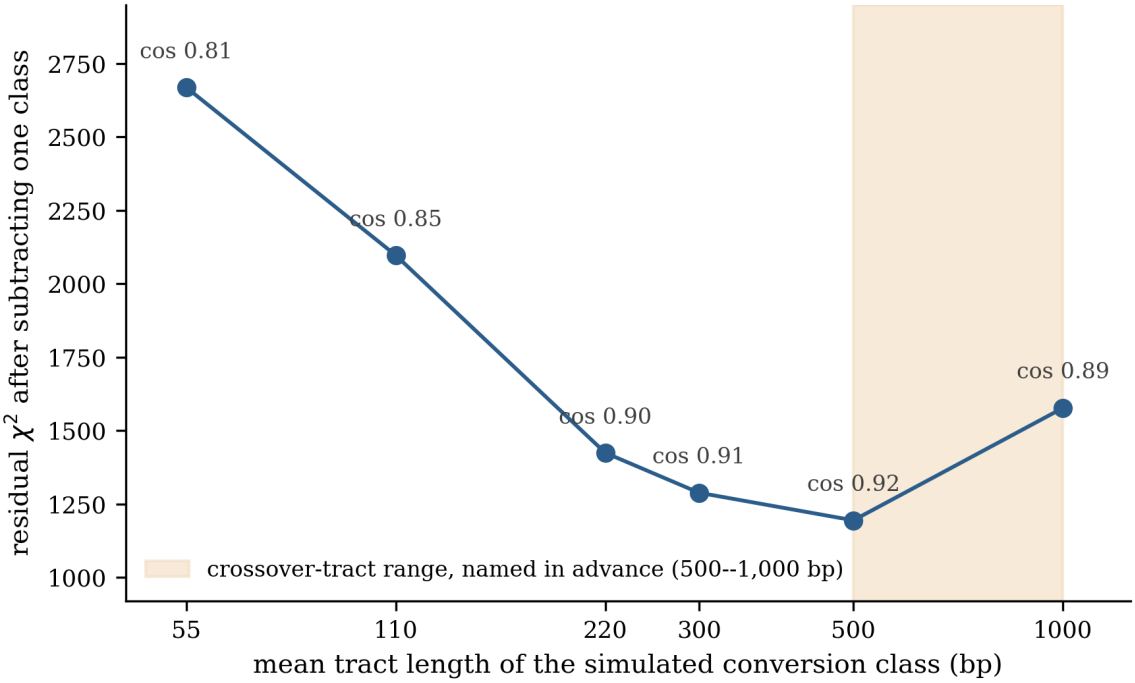

Figure 5 Estimating the mean tract length of gene-conversion events. Residual χ^2 after subtracting the best-scaled displacement of a single simulated conversion class with mean tract length L; each point is labeled with its alignment cosine. The misfit falls with L, reaches its minimum at $L = 500$ bp, and rises again at 1,000 bp. The shaded band is the 500–1,000 bp range expected for crossover-associated tracts (see text). The pedigree-scale tract lengths (55–110 bp) fit worst, which, with the direct observation of short tracts in families, motivates the two-class model of Section 3.3.

diction. With conversion alone the ratio $e(d)$ climbs steadily above one beyond 3 kb, reaching $e = 1.592$ at 8.5 kb; adding crossover at the deCODE map rate, held fixed, brings the ratio at the same separations to between 0.859 and 1.073 (Table 4). The scaled recombination rate $\rho = 4N_e rd$ controls the transition and crosses 1 near $d \approx 2.5$ kb: below that point the data follow the crossover-free prediction, and as ρ grows through 1 they approach the crossover-bearing one (Figure 6). Local variation in the crossover rate does not alter this result: an empirical mixture over pair-weighted local-rate terciles of the deCODE map (0.576, 1.178, and 2.309 cM/Mb) reproduces the mean-rate calculation, as expected when the displacement is near-linear in r.

We can also fit the crossover rate on the physical-separation set itself. With all conversion parameters held at their 10 kb values and the crossover rate as the only free parameter, the fit to the six separations gives a scaled rate of about 0.6 times the deCODE genome-mean value (Table 4, where the χ^2 of each rate is listed). The simulations hold $N_e = 10^4$, so the fitted quantity is the scaled rate $4N_e r$; 0.6 times the map rate and $N_e = 6{,}000$ at the map rate are the same statement in this parametrization. Fitting the crossover rate does not, however, flatten the trend of $e(d)$ with separation. At 0.6 times the map rate the six values of $e(d)$ in Table 4 still depart from unity by between 2.0 and 8.6 standard deviations, below one from 200 bp to 3 kb and above one at 5.5 and 8.5 kb, and the tilt persists at every other crossover rate in the table. In fractional terms the crossover-bearing prediction misses by 10–30% at mid-range separations; we regard this discrepancy as a limitation of the model in the crossover-dominated regime and examine it further in Section 3.8.

We also tested for crossover directly inside the joint fit to bin 0, the set on which the model is fitted. There, a crossover rate scanned from 5 to 520 times the map's own upper bound for those pairs worsens the fit at every step. Over the range in which the fit deteriorates only mildly, the conversion rates are stable: up to 16 times the bound, χ^2 rises by at most 5 units and the fitted rates shift by at most 1.4 profiled standard deviations. At 52 times the bound the long-class rate would shift by 4.2 standard deviations, with χ^2 higher by 7.9 units. We therefore include no crossover parameter in the bin-0 fit. Some crossover may nevertheless occur between bin-0 pairs: a rate of a few percent of the genome mean inside these map-cold intervals would be below the resolution of the pedigree map, and we cannot exclude it with these data. Table 3 gives the size of its effect on the inferred conversion rate.

3.4.2. The short-range regime. Between the validated range and the shortest separations the ratio of measured to predicted amplitude rises steeply, reaching $e = 2.805 \pm 0.072$ at $d = 37$ bp. Below 100 bp, where crossover is negligible and the full model's prediction coincides with the conversion-only one, the measured amplitude exceeds the prediction by a factor of 14.9 at $d = 6$ bp, 8.6 at 12 bp, 4.5 at 24 bp, and 1.7 at 75 bp. The excess has the shape expected when a single tract converts both sites (co-conversion), an amplitude equivalent to a conversion rate of about 4×10^{-6} per bp per generation at $d = 6$ bp, and a decay length near 20 bp.

Roughly half of this excess has a mutational explanation. A family of spectrum cells populated by complex mutations (i.e. single mutational events that affect two or more nearby sites simultaneously) accounts for 40–82% of the excess at $d = 6, 12,$

parameter	fitted value	independent expectation
λ_{500} (long class)	$(7.11 \pm 0.78) \times 10^{-6}$	one tract per crossover
λ_{short} (short class)	$(1.22 \pm 0.34) \times 10^{-6}$	pedigree 4–8 $\times 10^{-6}$ (Section 3.8)
λ_{total}	$(8.33 \pm 0.45) \times 10^{-6}$	pedigree range for total conversion, $\sim$ (4–13) $\times 10^{-6}$
ε	$(1.67 \pm 0.09)\%$	1.02% measured + context-specific components (Section 3.8)

Table 2 Parameters of the gene-conversion model, fit jointly with the demographic history. Point values are given for the reference mutation rate $\mu = 1.25 \times 10^{-8}$ per base pair per generation. The quoted $\pm$ errors are statistical only (the widest of the profiled one-sigma intervals across analysis choices, with demography marginalized); they do not include the systematic uncertainties of Table 3, which for the absolute value of λ_{total} amount to roughly a factor of two. Rates λ are per base pair per generation; the long class has mean tract length 500 bp, the short class 110 bp. The two rate coefficients are anticorrelated at -0.955: the total rate is the well-measured quantity and the partition between classes is soft. Pedigree and sperm-typing rates for non-crossover conversion from Halldorsson et al. (2016); Williams et al. (2015); Palamara et al. (2015); Palsson et al. (2025); Charmouh et al. (2025) are shown in the independent expectation column. The short-class value sits below its pedigree expectation; Section 3.8 analyzes that disagreement.

and 24 bp. Those cells hold only 4.8%, 3.8%, and 2.7% of the pair mass at those separations. These fractions are of the order expected from pedigree studies, in which 1.9% of de novo point mutations arise within 20 bp of another (Besenbacher et al. 2016). Within these cells the ancestral-mislabeling fraction is 12–17%, far above the per-pair rate of 3.3% implied by two independent chances per pair at the fitted per-site ε. Label errors at clustered sites therefore appear to be correlated, and the independent per-site mislabeling term of the model no longer applies in this regime.

The other half of the excess is unlikely to come from conversion: a short-tract conversion class large enough to supply it would raise the predicted amplitude at 10 kb to about 1.7 times the measurement and drive the fitted long-class rate λ_{500} negative, because the total rate is already fixed by the data at long range. We therefore regard the gene-conversion model as a model of linked ancestry whose domain is separations of roughly 100 bp and above. Below that boundary, mutational fine structure and clustered label errors seem to dominate the two-site spectrum, and the model as formulated should not be applied there.

3.5. Replication in a second cohort and an independent callset

The gene conversion parameters reflect molecular processes, so they should be constant across analyses of different data sets. To test whether this is in fact the case, we repeated the analysis twice: once in full on a second cohort, and once on an independently produced callset (on the same nineteen autosomes and in both pair sets analyzed above). The first repetition, with every analysis choice unchanged, was on the 1000 Genomes Yoruba (YRI) cohort (Byrska-Bishop et al. 2022). The genome-wide deviation reappears at a similar multiple of its matched-null threshold, with amplitude 1.010 relative to the Gambian deviation after scaling by pair mass. The joint fit returns the three fitted parameters, and the total rate λ_{total}, within about one joint standard deviation of the Gambian values (Table 5). The ratio of the long-class rate to the rate expected from one conversion tract per crossover is 1.2–1.3 in YRI against 1.1–1.2 in the Gambian cohort. Because these two cohorts share ancestry and most of their variant sites, the two estimators are strongly correlated (0.77–0.96), and the YRI measurement contributes only 0.05 to 0.13 of an independent cohort's worth of information. The agreement is therefore a consistency check and not an independent confirmation. The short-tract deficit discussed in Section 3.8 also carries over: λ_{short} is far below the pedigree range in both cohorts.

The second repetition, on the Human Genome Diversity Project callset (Bergström et al. 2020) described in Section 2, tests the data processing itself. The 22 Yoruba genomes of that callset (taken through the identical extraction, filtering, and estimation), with the conversion templates simulated under the fitted Gambian demography, reproduce the deviation in both pair sets. In the pooled physical-separation spectrum (all pairs within 1 kb) and in bin 0 of the map-distance set (the set to which the model was fitted), the deviation reappears at 341 and 346 times (respectively) the χ^2 expected from that callset's own jackknife sampling noise alone. In the pooled physical-separation spectrum its amplitude is 1.13 relative to the 1000 Genomes Yoruba deviation after scaling by pair mass. Its whitened direction matches the corresponding 1000 Genomes deviation closely (cosines $+0.9993$ with the Yoruba genome-wide deviation and $+0.9988$ with the Gambian bin-0 deviation, for the 22-genome panel). In both sets the three fitted parameters (the two conversion rates and the mislabeling rate) agree with those of the corresponding 1000 Genomes fit to the same pair set within half a joint standard deviation. A second panel adding 22 Mandenka genomes gives the same results. Because both callsets descend from GATK-era calling and share the EPO ancestral assignment, this comparison does not independently test the variant-calling software or the polarization, and the agreement of the mislabeling rate in particular is expected.

3.6. Gene conversion and the Drosophila deviation

In earlier work in *Drosophila melanogaster*, Fenton et al. (2025) found that Kingman coalescence can be rejected from the two-site spectrum of fourfold degenerate sites at separations of 3 to 24 bp on the four autosome arms. Their procedure, which we reimplemented in Section 2, fits a five-epoch population size history to the one-site spectrum and searches the scaled recombination rate for the best Kingman null. Given this history, they found that the observed two-site spectrum lay far outside that null's own resampling distribution on every arm. Their work did not consider gene conversion, and our results here raise the question of whether the deviation these authors observed might also be explained by conversion.

To test whether gene conversion can account for the

Source	Effect on the fitted parameters
Estimator dependence under rate heterogeneity across pairs	alternative estimators (a single fitted rate; the linear two-template estimator used here; a free non-negative mixture of templates; a parametric rate distribution) give λ_{total} from 3.8 to 10.8 $\times 10^{-6}$; roughly a factor of two on the absolute rate
Unresolved crossover within bin 0	each per cent of the genome-mean crossover rate present between bin-0 sites lowers the inferred conversion rate by about 1.2×10^{-6}
Assumed mutation rate μ	all estimated rates scale with μ, which varies substantially across published estimates
Highest-frequency column of the spectrum	refitting without it shifts λ_{total} by 3.9 statistical standard deviations
Covariance shrinkage (0.01 to 0.05)	λ_{short} spans $(1.05\text{–}1.48) \times 10^{-6}$; ε spans 1.56–1.80%
Background selection (McVicker B)	short-class coefficient rises monotonically with B ($z = 3.73$), a systematic of roughly 16% on λ_{short}
Heterozygosity filter (excised mass doubled)	no parameter moves by more than half a statistical standard deviation
Jackknife block size (1 to 5 Mb)	each parameter moves by less than one statistical standard deviation

Table 3 Systematic uncertainties on the parameters of Table 2. The first two rows dominate the uncertainty of the absolute conversion rate; the remaining rows are smaller or affect mainly the partition between the two tract classes and the mislabeling rate.

Drosophila deviation, we reran the released code of Fenton et al. (2025) with their released per-arm configuration for the DPGP3 sample, first unchanged and then with gene conversion enabled in every null simulation. Conversion was added at the two ends of the range of measured fly conversion rates, initiating at κ times the crossover rate, with $\kappa = 0.82$ and mean tract 440 bp (the whole-genome pedigree rate of Miller et al. (2016)) and $\kappa = 4.88$ with mean tract 518 bp (the population estimate of Comeron et al. (2012)), with the crossover rate search re-run in each case. Conversion is specified here by the ratio κ because the single recombination parameter of the Fenton et al. (2025) pipeline is a searched crossover rate, to which conversion must be tied; the human analysis above instead uses absolute per-base-pair rates (λ_{short} and λ_{500} in Eq. (6)), defined without reference to the crossover map. The unchanged rerun reproduces the published rejection on all four arms, with null KS means within 2.5% of the published values and the reproduced KS distances inside the range spanned by ten independent reruns of the pipeline on three arms (on the fourth, 2L, our distance falls below that range by 0.10 pooled standard deviations). Against the same ten reruns, the published KS distance on 3R lies below all of them ($z = -5.03$).

In the course of this rerun we found that the released configuration simulates the Kingman null at an inconsistent time unit. The size history is fitted to the one-site spectrum in units in which a pair of lineages coalesces at rate $1/(2N)$, while the null genealogies are simulated (msprime, `ploidy` $= 1$, as appropriate for haploid embryo genomes) at pair rate $1/N$ on the same epoch times. Every null is therefore run at twice its fitted epoch times in coalescent units, with the consequence that the simulated null's one-site spectrum does not reproduce the one-site spectrum it was fitted to (singleton fraction low by 2%, the $k \geq 20$ class low by 8% on 2L; L_1 distance 0.04 to 0.06 across arms). Because the KS statistic is computed on the joint spectrum without conditioning on its marginal, that mismatch enters the distance. Halving the epoch times passed to the simulator restores the fitted one-site spectrum to within simulation noise ($L_1 \leq 0.009$) and is the only change in the rows labeled "time unit consistent" in Table 6. With this change the Kingman rejection persists on all four arms at reduced size: the KS distances fall relative to the unchanged rerun but remain well above the null means and null maxima (Table 6), with no null resample reaching the data in 1000 draws. The published distances were computed at the inconsistent unit and should be read with this in mind. However, the conclusion of Fenton et al. (2025) that the fly two-site spectrum rejects the Kingman null does not depend on it.

Adding gene conversion, with the time unit consistent, does not remove the rejection (Table 6): on every arm and at either conversion rate, the KS distance between the data and the best null remains about as large as without conversion, and no null resample in 1000 draws reaches the data. Run with $\kappa = 0$, the conversion-enabled code reproduces the time-unit-consistent Kingman rerun within 10% in KS on every arm. The crossover rate selected by the search differs by large factors between these runs, in both directions, while the KS distance varies by a few percent, so this rate should be regarded as a search parameter on a nearly flat surface rather than as an estimate.

The results for the two species therefore differ: in humans the deviation observed in the data lies close to the direction of the displacement that gene conversion produces, is consistent with total conversion rates of the same order as pedigree measurements (though the short-tract rate by itself is far below the pedigree range; see Section 3.8), is predicted out of sample across separations, and reappears with similar fitted parameters in a second cohort and an independent callset. In *Drosophila* the two-site test of Fenton et al. (2025) finds a deviation (even after correcting for the incorrect time unit used in the published work) that remains far outside the null, whether or not gene conversion is included in it. Linked selection, heterogeneity in the recombination rate, undetected population structure and reproductive variance across individuals are all candidate mechanisms for the *Drosophila* deviation. In Section 3.7 we test several of these directly and examine the sampled haplotypes for the signature of multiple-merger genealogies.

3.7. Multiple-merger signatures in haplotype identity

The reanalysis of Section 3.6 shows that the *Drosophila* two-site deviation survives the addition of gene conversion to the null. A distance between spectra does not, however, identify the kind of genealogical departure behind it. Multiple-merger coalescents make a specific prediction about haplotypes. When one ancestor leaves many descendants within a short span of the genealogy, several sampled genomes carry an identical haplo-

crossover rate	χ^2	$e(200)$	$e(650)$	$e(1500)$	$e(3000)$	$e(5500)$	$e(8500)$
none	1,642	1.023	0.884	1.026	1.233	1.442	1.592
0.3× map	319	0.997	0.828	0.920	1.034	1.142	1.222
0.6× map	194	0.954	0.786	0.848	0.939	1.042	1.130
1× map (deCODE)	280	0.911	0.737	0.774	0.859	0.974	1.073
1.7× map	527	0.848	0.668	0.700	0.787	0.909	1.014
3× map	1,002	0.761	0.590	0.619	0.714	0.845	0.962

Table 4 Fit of the crossover rate on the physical-separation set, with the crossover rate as the only free parameter. Each row holds every conversion parameter at its 10 kb value and simulates crossover at the stated multiple of the deCODE genome-mean rate (1.2×10^{-8} per bp per generation; the simulations hold $N_e = 10^4$, so the fitted quantity is the scaled rate). Columns give the ratio of measured to predicted conversion amplitude $e(d)$ at the six separations (bp) and the χ^2 of $e = 1$ over the six points (statistical errors of 0.015–0.029 per point; the 650-bp column is the uncorrected value, 0.914 after the pair-separation mixture correction at zero crossover). The data select about 0.6 times the map rate; the tilt across separations persists at every rate.

parameter	Gambian cohort (primary)	Yoruba cohort (rerun)
λ_{500}	$(7.11 \pm 0.78) \times 10^{-6}$	$(7.97 \pm 0.78) \times 10^{-6}$
λ_{short}	$(1.22 \pm 0.34) \times 10^{-6}$	$(0.87 \pm 0.34) \times 10^{-6}$
λ_{total}	$(8.33 \pm 0.45) \times 10^{-6}$	$(8.84 \pm 0.45) \times 10^{-6}$
ε	$(1.67 \pm 0.09)\%$	$(1.70 \pm 0.09)\%$

Table 5 The four fitted constants in the primary Gambian cohort and in the Yoruba rerun of the identical analysis. The quoted uncertainties are statistical only and are, by construction, the same in both columns: they are the widest profiled one-sigma interval across analysis choices, which was determined on the Gambian data and applied to both fits (the Yoruba fit's own profiled errors, evaluated in the same jackknife metric, are essentially equal to the Gambian ones: 0.60, 0.26 and 0.36 $\times 10^{-6}$ for the three rates and 0.08% for ε, in both cohorts). They exclude the systematic uncertainties of Table 3, which apply equally to both cohorts. Each constant agrees within about one joint standard deviation. The two cohorts share ancestry and most variant sites (estimator correlation 0.77–0.96), so the agreement is a consistency check with limited independent information.

type at that locus, where a Kingman genealogy with the same rate of pairwise coalescence would give identical pairs and an occasional triple (Eldon and Wakeley 2006; Birkner et al. 2013). We therefore asked whether, at a fixed rate of pairwise identity, the Zambian *D. melanogaster* genomes carry larger clusters of identical haplotypes than Kingman genealogies under the fitted history produce. We then put the same question to the Gambian haplotypes (Section 2.9; Table 7).

3.7.1. Identity clusters in the Zambian* D. melanogaster *sample. We first measure the size of the fly two-site deviation against Kingman samples of the full analyzed region of each arm that carry linkage (within 1-Mb segments), the real site positions and the observed pattern of missing bases, processed exactly as the data. These samples give a two-site excess Q of 8–13 per arm, with a replicate standard deviation of 19–22, where the data give 100–120 (Table 7). No replicate of 50 reaches the data on any arm, and χ^2_K per degree of freedom is 1.62–1.75 in the data against 1.05–1.08 in the replicates. Conditioning each replicate's expectation on the sample's own one-site class frequencies removes any effect of misfit in the marginal spectrum; the data then remain above every replicate on 2R, 3L and 3R and above all but one on 2L. On 2L and 3R roughly a third and three fifths of the excess come from the 3-bp separation alone, where multinucleotide mutations, absent from the null, are a non-genealogical alternative. Removing the genomes that carry catalogued cosmopolitan admixture tracts (Lack et al. 2015) changes Q, relative to random subsets of equal size, by no more than about 20%, comparable to its sampling scatter. Distributing each arm's crossing-over along the arm according to the 100-kb map of Comeron et al. (2012) moves the null mean by less than 4.

In windows of 50 variable fourfold degenerate sites, and comparing only windows with the same identity fraction, the largest identity cluster in the fly data exceeds the Kingman expectation in every identity bin used on 2L, 2R and 3L. The excess is 12–32% of the null value; on 2L at identity fractions between 1 and 2%, for example, the largest cluster averages 16.2 genomes against 12.8 ± 0.1 in the replicates. The score of Eq. (8) is $Z = 35.3$, 27.7 and 25.0 on these three arms. Beta-coalescent controls with $\alpha = 1.7$, scored against the same replicates, give Z between 10 and 26 on these arms, and controls with $\alpha = 1.9$ give at most 10, so the data lie above the $\alpha = 1.7$ controls on all three. Since the controls match the data in variant density only, this comparison indicates direction and rough scale. On 3R the score also passes the threshold ($Z = 14.2$), but the fitted history there produces fewer than a third of the observed identical pairs and the score is carried by the windows of lowest identity. At finely matched identity the 3R excess is between +3 and +12% depending on the null, and we regard it as marginal. By the rule recorded in advance, a signal on at least three of the four autosome arms, the Zambian sample shows the multiple-merger signature, whether or not 3R is counted.

The fitted Kingman histories miss the overall level of pair identity on most arms: the ratio of observed to simulated identical pairs ranges from 0.8 to 3.5 on the autosomes and is 0.55 on the X (Table 7). Because the bins of Eq. (8) absorb such a mismatch only approximately, we use the fine-matched excess for comparisons between arms and between species. It is +21.5, +25.4 and +20.7% on 2L, 2R and 3L, with a null scatter of 0.5–0.8%. It moves by a few points at most under each change of null in Table 7: a history refit so that the processed one-site spectrum of the simulations matches the data to $L_1 \leq 0.010$ and a singleton fraction within 0.002 (against $L_1 = 0.020$–0.023

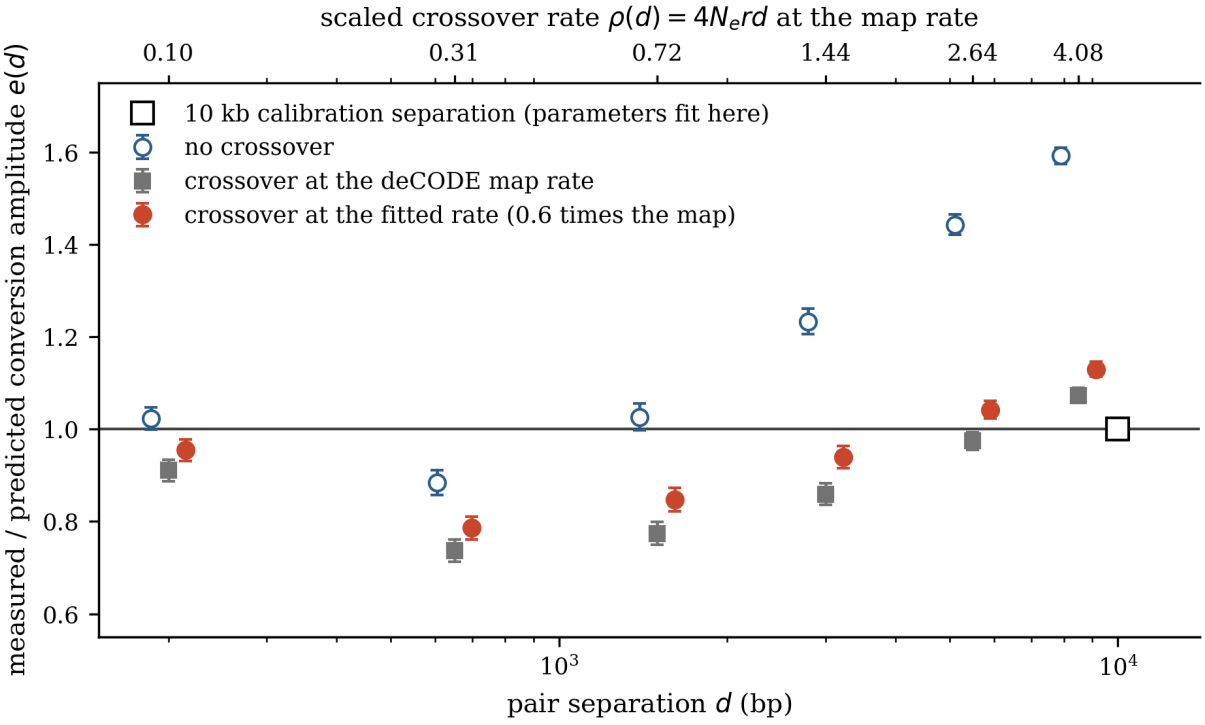

Figure 6 Out-of-sample validation of the gene-conversion model across separations. The measured conversion-attributable deviation amplitude divided by the model prediction, $e(d)$, with all parameters fixed at the 10 kb separation (Section 3.3). One curve shows the prediction without crossover, a second with crossover at the deCODE map rate, and a third at the rate fitted on this pair set, 0.6 times the map (Table 4). The data track the crossover-free prediction where the scaled recombination rate $\rho = 4N_e rd$ is small and move onto the crossover-bearing prediction as ρ grows through 1, near $d \approx 2.5$ kb. The remaining 10–30% mid-range discrepancy is discussed in the text.

for the fitted-history null, whose singleton fraction is 3% low); crossing-over distributed along each arm according to the Comeron et al. (2012) map; and removal of the inversion carriers described below. Windows of 25 or 100 sites, a stricter calling threshold and an identity rule that ignores singleton differences all leave $Z \geq 14$ on these arms. When windows are grouped by the local crossover rate of the Comeron et al. (2012) map, the excess is present in every tercile on every arm and increases from the lowest to the highest tercile on 2L, 3L and 3R ($Z = 24$, 33 and 36 on 2L). Background selection and recurrent sweeps acting mainly where crossing-over is rare would be expected to produce the opposite trend. The null in this comparison was pooled over rate classes, however, and we treat the stratification as descriptive.

The check for persistent co-clustering (Section 2.9) was positive on all four arms. Some pairs of genomes are identical in far more windows than any pair in any null replicate, up to 1,639 of the 1,653 windows on 2L for one nearly clonal pair, against a null maximum of 175. The genomes involved, between 12 and 66 per arm and thus too many to remove under the rule of Section 2.9, are largely that pair and carriers of the common cosmopolitan inversions in the DPGP3 inversion calls (Lack et al. 2015; Corbett-Detig and Hartl 2012). The persistent block on 2L consists exactly of the 44 In(2L)t carriers, the 2R block of 24 genomes contains 23 of the 24 In(2R)NS carriers, and the 3R block comprises every In(3R)K carrier together with 9 of the 21 In(3R)P carriers; the 36 In(3L)Ok carriers form no block. Removing these genomes (49, 32, 7 and 55 on the four arms) from the data, the replicates and the controls alike leaves the cluster excess unchanged ($Z = 36.1$, 28.3, 25.9 and 16.2). It removes the persistence signal on 2L and 2R, while a weaker persistence with carriers we have not identified remains on 3L and 3R. The two-site excess Q is likewise unchanged, relative to random subsets of equal size, when all inversion carriers and close relatives are excluded (ratio 0.99–1.35). Across chromosomes, the identity of a pair of genomes on an arm of chromosome 2 does not predict its identity on an arm of chromosome 3 ($|z| \leq 1.3$ against permutation of genome labels), and simulated sub-populations of five or more of the 197 genomes are always detected by that test ($z \geq 9$). Structure that is not attached to particular genomes, such as an unsampled population that has contributed ancestry to all of them, would give the same window-level signature as multiple mergers, and these checks do not exclude it.

The X chromosome, analyzed identically in 196 genomes but not counted in the decision, shows the excess at least as strongly as the autosomes. Its score is $Z = 38.4$ and its fine-matched excess +26%, slightly above the autosomal range, although its fitted history over-produces identical pairs by nearly a factor of two. In the two-site statistic, its excess over its own Kingman replicates ($Q = 150$ against -1 ± 18) is 1.40–1.65 times the corresponding excess, Q minus the replicate mean, of each autosome. Reproductive skew confined to one sex and linked selection lead to different expectations for the contrast between the X and the autosomes (an X lineage is transmitted through a

arm	null	search r	KS distance	null mean (max)	p
2L	Kingman (published)	0.0107	1046	244 (425)	$< 10^{-3}$
	Kingman (rerun, released configuration)	0.0388	1190	245 (468)	$< 10^{-3}$
	Kingman (rerun, time unit consistent)	0.0388	817	242 (405)	$< 10^{-3}$
	+conversion, $\kappa = 0.82$ (consistent)	0.0044	817	246 (411)	$< 10^{-3}$
	+conversion, $\kappa = 4.88$ (consistent)	0.0021	834	246 (463)	$< 10^{-3}$
2R	Kingman (published)	0.0210	1141	201	$< 10^{-3}$
	Kingman (rerun, released configuration)	0.0333	1199	206 (399)	$< 10^{-3}$
	Kingman (rerun, time unit consistent)	0.0119	609	205 (417)	$< 10^{-3}$
	+conversion, $\kappa = 0.82$ (consistent)	0.0244	715	204 (374)	$< 10^{-3}$
	+conversion, $\kappa = 4.88$ (consistent)	0.0026	696	203 (380)	$< 10^{-3}$
3L	Kingman (published)	0.0158	804	202	$< 10^{-3}$
	Kingman (rerun, released configuration)	0.0331	843	206 (403)	$< 10^{-3}$
	Kingman (rerun, time unit consistent)	0.0244	571	203 (373)	$< 10^{-3}$
	+conversion, $\kappa = 0.82$ (consistent)	0.0244	666	205 (385)	$< 10^{-3}$
	+conversion, $\kappa = 4.88$ (consistent)	0.0039	593	207 (381)	$< 10^{-3}$
3R	Kingman (published)	0.0622	882	208	$< 10^{-3}$
	Kingman (rerun, released configuration)	0.0622	1023	206 (380)	$< 10^{-3}$
	Kingman (rerun, time unit consistent)	0.0674	652	208 (439)	$< 10^{-3}$
	+conversion, $\kappa = 0.82$ (consistent)	0.0236	649	204 (388)	$< 10^{-3}$
	+conversion, $\kappa = 4.88$ (consistent)	0.0044	655	208 (349)	$< 10^{-3}$

Table 6 Reanalysis of the *Drosophila* two-site test of Fenton et al. (2025), with the null's time unit made consistent and with gene conversion added to the null. For each autosome arm: the published run; our unchanged rerun of the released pipeline and configuration; the same rerun with the null's time unit made consistent with the fitted size history (epoch times halved before simulation at `ploidy` $= 1$; see text); and, under the consistent unit, the rerun with gene conversion enabled in every null simulation at the low and high ends of the measured fly bracket (κ = conversion initiations per crossover; tract means 440 and 518 bp). Columns give the crossover-rate search parameter selected by the pipeline (scaled units; a search coordinate on a nearly flat surface, see text), the KS distance between the observed two-site spectrum and the best null, the mean and maximum KS distance among 1000 null resamples, and the resampling p-value (no resample reached the data in any row). The Kingman rejection persists under the time-unit correction and with conversion in the null, on every arm.

male in only one third of generations, while the X recombines only in females and has the smaller effective size), but we have not derived either expectation for these statistics and report the contrast without interpreting it.

3.7.2. A single Beta coalescent as the null. If a Beta coalescent with one value of α described the fly genealogies, a null built on that coalescent should reproduce both the two-site excess and the cluster excess, once its size history is refit to the one-site spectrum and everything else is left unchanged. A spectrum-matched history could be constructed for $1.6 \leq \alpha \leq 1.85$ of the grid $1.5 \leq \alpha \leq 1.95$, and in that range we find no value of α at which the simulations do so. For $1.6 \leq \alpha \leq 1.85$ the refit history stayed off its size bounds, with a one-site L_1 distance from the data of 0.04–0.10 at $\alpha = 1.8$–1.85, 0.06–0.16 at 1.7 and 0.16–0.23 at 1.6 (the Kingman fit gives 0.02). The Beta simulations there fall short of the observed two-site excess on at least two of the four arms at every α, with the data 1.1 to 6.4 replicate standard deviations above them (3.5, 3.5, 3.9 and 6.4 on 2L, 2R, 3L and 3R at $\alpha = 1.85$). At matched identity their largest clusters are smaller than the observed ones in all sixteen combinations of arm and α in that range, by 7–21% in fifteen of them. At $\alpha = 1.5$ the cluster statistic is reproduced on all four arms, but only with a refit history that sits on its upper size bound, whose one-site spectrum misses the data by an L_1 distance near 0.4 (about twenty times that of the Kingman fit); the two-site excess is still not reached on any arm. For $1.85 < \alpha < 2$ the refit develops extreme transient contractions, down to the lower size bound on some arms (L_1 0.11–0.35), which we attribute to the rapidly shrinking time scale of the simulated Beta coalescent as $\alpha \to 2$, and that interval remains untested. Within the family examined, then, a Beta coalescent with a single α and a piecewise-constant history matched to the one-site spectrum produces too little haplotype clustering, and on most arms too small a two-site excess, wherever the comparison could be made. The family is a narrow one, and the result leaves open other sources of multiple mergers: offspring-number distributions with a different tail or with rare large reproduction events, Ξ-coalescents, and recurrent selective sweeps, which produce mergers locally along the genome (Durrett and Schweinsberg 2005; Neher and Hallatschek 2013; Desai et al. 2013).

3.7.3. The Gambian haplotypes. In the Gambian sample we find a different pattern. By the score of Eq. (8) alone, the 182 retained haplotypes pass the per-chromosome threshold ($Z \geq 3$) on all nineteen autosomes under each of the three treatments

of crossing-over in the null, as do the Beta-coalescent controls with $\alpha = 1.9$ and 1.7 ($Z \geq 5$). Every one of these nulls, however, produces fewer identical pairs than the data, by the ratios given below; the coarse identity bins absorb an offset of that size only in part, and when identity is matched within twenty quantile cells the excess of the largest cluster over the Kingman expectation lies between -5.3 and $+3.5$%, depending on the treatment of crossing-over and on the matching construction (medians over the nineteen autosomes; Table 7). We therefore do not read the human score as a multiple-merger signal. The corresponding fine-matched fly figure is $+17$ to $+31$% on four of the five arms under every null we tried. The Beta controls share the level mismatch of the human nulls, and we have not built identity-matched Beta nulls for this sample; the comparison between the species therefore rests on the fine-matched excess, a measure introduced after the coarse-score comparison above had been made (Section 2.9). The Gambian haplotypes instead show two features that the fly genomes lack. First, pairs of haplotypes are identical across a 50-site window (median 6.9–8.5 kb) about 1.5 times as often as in single-history Kingman genealogies fitted to the same frequency spectrum. The median ratio is 1.52 with uniform crossing-over, 1.37 with the fine-scale map, and still 1.15 when the simulated crossover rate is lowered essentially to zero. Second, pair identity persists across chromosomes, in that individuals who share haplotypes on one chromosome tend to share them on the others ($z = 13.1$ against permutation among the 91 unflagged individuals, 14.7 among all 113); the comparable fly statistic gives $|z| \leq 1.3$. Cryptic relatedness and endogamy produce persistence of this kind. Simulated island-model structure strong enough to give up to six times the observed persistence reproduces it, while leaving both the identity level and the cluster statistic at their Kingman values (identity ratio 1.00; fine-matched excess between -0.6 and -0.1%). The level excess, in turn, can be reproduced in large part by a few per cent of ancestry from a lineage that separated from the ancestors of the sample deep in the past, which neither a refit single-population history nor the persistence test detects. Two of the six configurations we tried reach a ratio of 1.34–1.35, about 70% of the observed excess. We offer this only as a demonstration that such structure is sufficient: the configurations are illustrative, recombination hotspots by themselves account for more than a quarter of the excess (the difference between the uniform and map-based nulls), and other contributions may be present. In the 55 male X chromosomes, which involve no statistical phasing, the identity ratio is 2.0 against the uniform-rate null, so the level excess is unlikely to be a phasing artifact. The fine-matched cluster excess there is $+5$ to $+7$%, larger than on the autosomes but measured in a sample too small for that measure to detect the $\alpha = 1.9$ control, and we leave it uninterpreted. The Beta-coalescent displacements of Section 3.2 likewise do not align with the human two-site deviation. Taken together, the Gambian haplotypes show a raised level of pairwise identity and persistence across chromosomes but no cluster excess of the size seen in the flies, whereas the Zambian *D. melanogaster* sample appears to require genealogies with multiple mergers or a process that produces them locally.

3.8. Unresolved discrepancies

The gene-conversion model explains most of the genome-wide deviation (Section 3.3) but leaves two kinds of discrepancy unresolved. One is a quantitative disagreement between the best-fit short-tract gene conversion rate and the rate expected from pedigree data. The other is a set of features of the data that the gene-conversion model does not explain: a tilt at mid-range separations, a localized deficit at one separation, the residual tail of the genome-wide spectrum itself, and two components of the mislabeling rate. We quantify both types of discrepancy below; we have not been able to resolve either.

3.8.1. The short-tract rate discrepancy. The joint fit measures the short-tract conversion rate at $\lambda_{\text{short}} = (1.22 \pm 0.34) \times 10^{-6}$ per base pair per generation (Table 2, where the quoted error is the widest profiled one-sigma interval across analysis choices); in the reference analysis the 2σ interval is $(0.70\text{–}1.75) \times 10^{-6}$. Pedigree and sperm-typing studies of non-crossover gene conversion report rates of $4\text{–}8 \times 10^{-6}$ per base pair per generation (Halldorsson et al. 2016; Williams et al. 2015; Palamara et al. 2015; Palsson et al. 2025; Charmouh et al. 2025). For example, Williams et al. (2015) estimate 5.9×10^{-6} per bp per meiosis (95% CI 4.6–7.4×10^{-6}) from transmitted tracts in pedigrees, a range that lies entirely above our 2σ interval. This discrepancy concerns the short-tract class alone, which is the class these direct counts measure; the total rate in this cohort shows no such deficit (Table 2). The discrepancy is also present in the Yoruba rerun, where the identical fit returns $\lambda_{\text{short}} = (0.87 \pm 0.34) \times 10^{-6}$ (Table 5).

We examined this discrepancy in four ways, none of which resolves it. Changing the tract-length assumption, by moving the mean tract length from 100 bp to 46 or 17 bp, does not close the gap. The missing rate cannot have been absorbed by the mislabeling term either, because in the joint fit the correlation between λ_{short} and ε is only $+0.12$. The softness in the fit lies instead in the partition of the rate between the two tract classes, which are strongly anticorrelated (Table 2). However, the total rate is well measured, and even assigning this entire total to short tracts would result in a rate at the bottom edge of the pedigree range. Adjusting for the local recombination environment of the measured pairs has only a small effect (although the pairs are in colder-than-average map regions). Measured against the deCODE non-crossover map (Palsson et al. 2025), the regions containing the measured pairs have a non-crossover conversion rate $f_{\text{NCO}} = 0.85$ times the genome mean (the corresponding crossover factor is $f_{\text{CO}} = 0.93$), which lowers the local pedigree expectation to $3.8\text{–}6.8 \times 10^{-6}$, still 2–4 times the fitted short-tract rate. Lastly, imposing the pedigree value directly, by fixing $\lambda_{\text{short}} = 5 \times 10^{-6}$, the middle of the pedigree range, and refitting everything else, worsens the fit severely (χ^2 rises from 452.6 for the 46-bp fit to 1,604) and drives λ_{500} negative.

The two kinds of measurement differ in what they count: pedigree sequencing counts conversion events in single parent-to-child transmissions, whereas the two-site spectrum measures the rate at which those events accumulate over the whole sampled genealogy. We found one suggestive correlation, the rise of the short-class coefficient with the McVicker background-selection statistic B (Table 3), but the deficit persists in the weakest-selection third of the genome. Three explanations seem possible to us, and we cannot at present distinguish between them. The gap may arise in the detection modeling behind pedigree counts, since even the two deCODE pedigree estimates differ by nearly a factor of two (4.0 and 7.0×10^{-6}; Palsson et al. 2025; Halldorsson et al. 2016). It may lie in the tract-length law, since a single geometric class at each scale is an assumption, and a two-part mixture was already preferred to any single geometric law in the UK Biobank identity-by-descent analysis (Masaki and Browning 2025). A direct test of that mixture as a remedy for the short-tract deficit, however, worsens the joint fit ($\Delta\chi^2 = +78$)

		two-site excess Q				fine-matched cluster excess (%)			
sample	n	data	null (s.d.)	$P_{\text{data}}/P_{\text{null}}$	Z	(a)	(b)	(c)	(d)
D. melanogaster, DPGP3 Zambia (fourfold degenerate sites; 50-site windows, median 3.8–4.7 kb)									
2L	197	120.1	13.0 (22.2)	1.70	35.3	+21.5	+19.4	+22.4	+23.4
2R	197	117.7	10.2 (20.7)	0.82	27.7	+25.4	+24.9	+26.0	+24.6
3L	197	99.8	8.4 (19.0)	1.31	25.0	+20.7	+19.6	+20.7	+22.3
3R	197	115.0	10.9 (21.1)	3.50	14.2	+10.1	+6.6	+11.1	+11.8
X (not counted)	196	149.9	−0.9 (18.2)	0.55	38.4	+26.1	–	–	–
Gambian (GWD), phased panel (minor-allele count ≥ 2; 50-site windows; autosomal entries are medians over 19 chromosomes)									
autosomes, uniform crossing-over	182	–	–	1.52	‡	−5.3 to +1.2			
autosomes, fine-scale map	182	–	–	1.37	‡	0.0 to +3.0			
autosomes, map at near-zero rate	182	–	–	1.15	13.5	+3.4 to +3.5			
male X, uniform crossing-over	55	–	–	2.04	‡	+5.2			
male X, uniform at near-zero rate	55	–	–	1.16	17.0	+6.8			

Table 7 Haplotype-identity clusters and the two-site excess in *D. melanogaster* and in the Gambian sample. For each fly arm: the number of genomes; the two-site excess $Q = \chi^2_{\text{K}} - \nu$ of Section 2.9 in the data and its mean (standard deviation) over 50 full-region Kingman replicates simulated under the arm's fitted history and processed as the data; the ratio of identical haplotype pairs in 50-site windows in the data to the replicate mean; the cluster score Z of Eq. (8), whose null scatter is 1.0–1.2; and the fine-matched excess of the largest identity cluster over the replicates at equal identity fraction: (a) under the fitted-history null, (b) under a null whose history is refit so that the processed one-site spectrum matches the data, (c) under a null with crossing-over distributed along the arm by the map of Comeron et al. (2012), and (d) with the inversion-karyotype blocks and the near-clonal pair removed from data and replicates alike (null scatter 0.5–0.8% in each case; a second, independently constructed matching agrees within 5 points). Human rows: 182 GWD haplotypes under three treatments of crossing-over in the null, and the haploid X of 55 males, in 50-site windows of median span 6.9–8.5 kb on the autosomes and 16 kb on the male X; fine-matched entries give the range over the two matching constructions. ‡Z withheld: the null's median identity ratio lies outside 0.80–1.25, a condition specified for the human analysis before these nulls were run, and the coarse score is level-confounded there (fine matching moves the excess against the uniform-rate null from +8 to −5%). No such condition was set for the fly arms, where the fine-matched columns serve the same purpose and show, on 3R, that the score overstates the excess when the ratio is far from one. At fixed pair identity the fly genomes carry identity clusters about a fifth to a quarter larger than Kingman genealogies produce on 2L, 2R, 3L and the X under every null tried, whereas the Gambian haplotypes show a raised level of pair identity and no cluster excess of comparable size.

and the fit assigns no weight to its components, so the tract law alone does not appear to resolve the disagreement. Alternatively, it may reflect conversion biology itself: if conversion initiation or resolution targets sequence diversity, the two approaches would measure genuinely different quantities.

3.8.2. Other unexplained features. Three further features of the data remain unexplained by the model. The first is the 10–30% tilt at mid-range separations, where crossover dominates the prediction (Figure 6). The tilt persists under each of the single-population histories we tried (flat histories at two scales, a published African demographic model (Tennessen et al. 2012), and the data's own fitted eight-epoch history), which suggests that no single-population $N(t)$ consistent with these data flattens it. The conclusions drawn in the conversion-dominated regime do not depend on these separations, and the tilt cancels in comparisons between cohorts, because the same simulations enter both sides of any cohort contrast.

The second is the localized deficit at 650 bp in the physical-separation set (Section 3.4), where $e(650) = 0.91 \pm 0.03$ after all corrections, while the flanking separations at 200 bp and 1.5 kb are within a few percent of unity (Figure 6). This is also the separation in that set at which the conversion and mislabeling directions are hardest to tell apart (their cosine in the jackknife metric is 0.43 at 650 bp, rising monotonically from 0.23 at 6 bp), which may be related.

The third is the residual tail of the genome-wide deviation. After the joint fit a residual of $\chi^2 = 370.2$ remains, 4.8% of the original deviation, against a matched-null 95th percentile of 29.2. The residual concentrates in the highest-frequency column of the spectrum, the cells in which one site's derived allele is carried by seven of the eight sampled genomes (Figure 3c). It seems to be distinct from the short-range excess described above: the cosine between the two structures in the jackknife metric is only 0.34, and attributing the tail to that excess would require the excess to be ten times its measured amplitude.

Two components of the mislabeling rate also remain unexplained, while the rest of that rate is accounted for. The value we report throughout is the joint-fit estimate of Table 2. The decomposition below uses a variant in which the regressors are drawn from the fitted demographic family; that variant gives 1.750%, inside the spread across analysis choices listed in Table 3. The est-sfs posterior disagreements (Keightley and Jackson 2018) described in Section 3.3 supply the largest single part of the rate, and most of the remainder lies in components localized in sequence context, which are stable across analysis choices. The largest of these is a CpG-transition recurrence component, a within-class gap of +1.11%. Because CpG transitions are both the most recurrent and the most often mispolarized class, we repeated the complete analysis with every CpG-associated site excluded (any site at which the reference, alternate or inferred ancestral allele forms a CpG dinucleotide with a neighbouring reference base; this removes 30% of sites and 45% of the closely linked pairs). The Kingman class is rejected at least as strongly

per pair ($\chi^2 = 5869$ on the remaining 55% of pairs, against 7697 on all), the gene-conversion parameters are unchanged ($\lambda_{\text{total}} = 8.13$ against 8.33×10^{-6} per bp per generation; tract-class amplitudes 0.240 and 0.272 against 0.222 and 0.285), and the mislabeling rate falls from 1.67% to 1.20%, consistent with the CpG recurrence component identified here. A second is a transversion-specific excess of (0.464 ± 0.059)% over the clean transition class, which is significant under all the metric choices we tried (7.9σ under the fitted-family regressors; 4.2–5.3σ otherwise). None of the mutation-rate, ancestral-context, or artifact explanations we tested accounts for this transversion excess, which is one of the two components we cannot explain; the other is the context-independent part of the rate, about 0.33%. Under the fitted-family regressors the rate is also consistent across pair sets: the physical-separation set gives 1.75–1.76%, whereas the lower values obtained under flat-history regressors appear to be an artifact of those regressors.

4. Discussion

The two-site frequency spectrum was proposed as a diagnostic for genealogies that depart from the Kingman coalescent, for instance through multiple mergers (Fenton et al. 2025). Such departures can leave the single-site spectrum well fit while biasing the inferences built on it, but they should be visible in the joint spectrum of linked sites. When we apply this diagnostic to the Gambian sample analyzed here, the 2-SFS of closely linked sites is inconsistent with the Kingman coalescent under any single-population demographic history, with a misfit more than two hundred times the rejection threshold set by the sampling noise. The direction of the deviation, however, aligns closely with the displacement produced by gene conversion and no better than chance with those produced by the multiple-merger genealogies we tested. Once conversion in two tract classes and an ancestral-mislabeling rate are added to an otherwise standard Kingman model, three fitted parameters account for 94.2–96.2% of the deviation. We therefore suggest that the assumption that fails in these human data is the routine treatment of linkage, in which nearby sites share one genealogy except where crossover separates them (Hudson 1983). We find no evidence for departures from the binary branching structure of the Kingman coalescent itself. Adding conversion to the null of the published *Drosophila* test does not rescue the Kingman coalescent there (Fenton et al. 2025), so the 2-SFS appears to retain its power as a test of genealogical structure once gene conversion, the process that dominates the deviation in humans, is included in the null model. The haplotype-identity analysis of Section 3.7 suggests that the fly deviation reflects multiple-merger genealogies: at a fixed rate of pairwise identity, the Zambian genomes carry clusters of identical 4-kb haplotypes about a fifth to a quarter larger than Kingman genealogies produce, on three autosome arms and the X and under every null we tried, as expected when single ancestors occasionally leave many descendants, while the Gambian haplotypes show no comparable excess. A Beta coalescent with a single α does not, however, reproduce the fly two-site and haplotype statistics together over the range of α we could test, and the two broad sources of multiple mergers, sweepstakes reproduction and recurrent linked selection (Eldon and Wakeley 2006; Durrett and Schweinsberg 2005; Neher and Hallatschek 2013), remain to be distinguished. The persistence of the cluster excess in regions of high crossing-over and its strength on the X bear on that question, but await quantitative predictions under each model.

The single-site spectrum could not have revealed the role of gene conversion. In contrast, the 2-SFS depends on the joint ancestry of the two sites and therefore on any process that places them on different lineages. Methods that use linked variation over hundreds of bases to kilobases, such as linkage-based demographic inference, ancestral-recombination-graph reconstruction, and identity-by-descent methods, typically take recombination from the crossover map and omit conversion. At these separations our fitted rates imply that conversion is likely to be the dominant process placing two nearby sites on different lineages, ahead of crossover, and such methods could now include it at the measured rates. The fitted mislabeling rate describes the EPO ancestral-state assignment and is of the same order in every subset of the data we examined (Table 2; Figure S1); other statistics of derived-allele frequency built on the same EPO assignment will probably inherit an error of this size.

The joint fit to the two-site spectrum also yields a population-genetic measurement of the gene conversion rate. Conversion has been measured directly only where families or single sperm can be sequenced (Halldorsson et al. 2016; Palsson et al. 2025; Charmouh et al. 2025); estimates from population polymorphism data have so far been indirect (Frisse et al. 2001; Gay et al. 2007; Setter et al. 2022). From the two-site frequency spectrum of a population sample and an outgroup alignment we recover a total rate and a two-scale tract structure, which may make the process measurable in species where pedigrees cannot be assembled. In an independent population-scale analysis of identity-by-descent segments, a two-part mixture of tract lengths likewise fits better than a single geometric law, although the fitted scales differ from ours (Masaki and Browning 2025).

The rejection of the Kingman class and the identification of conversion are insensitive to the analysis choices we varied, but the absolute rate obtained this way is less precisely determined. The two-site spectrum of a pair depends on crossover and conversion only through the combined rate at which either process places the two sites on different ancestral lineages, and the data suggest that this rate varies strongly across pairs; under such heterogeneity, reasonable estimators applied to the same spectrum differ by roughly a factor of two (Table 3). Crossover below the resolution of the pedigree map inside bin 0, which the fit would attribute to conversion, is the other main systematic uncertainty on the absolute rate, and all fitted rates scale with the assumed mutation rate. Within these uncertainties the total conversion rate is consistent with pedigree and sperm-typing estimates. The one clear quantitative disagreement is the short-tract rate, which is a factor of 2–8 below direct counts of non-crossover conversion in both West African cohorts and at each tract length we tried (Section 3.8). It remains unclear whether that gap reflects the detection modeling of pedigree counts, the tract-length law, conversion biology that makes the two approaches measure different quantities, or some other effect.

The gene-conversion model does not account for every feature of the data. After the fit a residual well above the matched-null threshold remains, concentrated in the highest-frequency column of the spectrum. With its parameters fixed on the pairs at the smallest genetic map distance (bin 0), the model predicts the deviation amplitude across physical separations. The data follow this prediction at the shortest separations in its domain but depart from it where crossover dominates, with a tilt across separations that no choice of crossover rate in the present model removes. Below about 100 bp the data exceed the model by up to an order of magnitude, and roughly half of that excess

has a mutational explanation in multi-nucleotide events. We interpret each of these three discrepancies as a limitation of the model as fitted here. It is possible that a richer description of rate heterogeneity or tract structure would reduce them; the bounded alternatives of Table 1, multiple mergers included, are less promising candidates, since even against the full genome-wide deviation they account for much less than gene conversion does (Section 3.2).

The conversion parameters describe molecular processes and should not depend on demographic history. The reanalyses of the Yoruba cohort and of the independently produced HGDP callset reproduce the deviation and the fitted parameters, but all of these samples are West African, and the two 1000 Genomes cohorts share ancestry and most of their variant sites. Repeating the analysis in a population with a different demographic history, drawn from the full 3,202-sample panel (Byrska-Bishop et al. 2022), would be the decisive replication. Repolarizing with probabilistic ancestral assignment (Keightley and Jackson 2018) should remove most of the mislabeling term, since posterior disagreements from est-sfs already supply 1.02% of the fitted 1.67% with no free parameter. Larger projected sample sizes, which the exact computation described in the Methods can handle at higher computational cost, may sharpen both the partition of the rate between the two tract-length classes (which is poorly constrained at $n = 8$) and the description of the short-range regime, where the present model does not apply. For the human data analyzed here, our conclusion is that gene conversion has to be part of the null model for the two-site spectrum at these separations. Once it is included, a Kingman coalescent with an ordinary demographic history removes all but a few percent of the deviation we observe in the closely linked pairs, and the multiple-merger genealogies we tested need not be invoked (Figure 4).

5. Use of Claude in population genetics

All of the computational work reported here was carried out by Claude, Anthropic's AI model, running as a small number of long-lived autonomous agent sessions on a few multi-core servers. Working from questions and instructions given in ordinary language, Claude wrote and ran every piece of software used in this study: the site filters and pair-counting pipeline over the 1000 Genomes and HGDP genotype files, the exact computation of the two-site spectrum under the models of Section 2, the fits and matched-null simulations, the alternative-process models of Section 3.2, the certified class-wide bounds of Appendix A, and all figures and tables. It also read the relevant literature, proposed many of the checks, kept a machine-written record in which every number quoted in the paper is tied to the code version and command that produced it, wrote first drafts of most sections, and answered queries left in the manuscript margin, usually with new computations and usually within hours. Neither of us wrote or ran any analysis code.

Our own roles were direction and judgment. M.D.S., a physicist, chose the problem area, set up and operated the agent framework, and made the decisions of scope. The framework's working rules were designed around the ways such a system fails rather than around its strengths: each test was specified in writing before its result was computed; every claimed result had to survive review by a second, independent Claude session instructed to find fault with it before it was shown to us; and results are quoted only from the machine record, never from a session's recollection of it. M.M.D. supplied the population genetics: which departures from the standard model would matter, which confounders and checks a specialist would insist on, how the results should and should not be interpreted, and where successive drafts claimed more than had been shown; he also rewrote the manuscript. Gene conversion entered the project in that spirit, as a confounder written into the pre-specified controls for closely linked pairs before the genome-wide data had been examined, not as a hypothesis; the analysis turned the control into the explanation.

The pace is worth recording. The first genotype files were downloaded on 2 August 2026. The genome-wide deviation from the Kingman class, the exclusion of the alternatives of Table 1, and the gene-conversion model were in hand by the third week of August, and this manuscript was begun on 20 August; the replication cohorts, the reanalyses of Table 3, and the supplementary checks followed over the next five weeks, most of them prompted by questions raised on reading the drafts. Each such check (excluding CpG sites, removing related individuals, extending the analysis to every autosome, repeating it on an independently produced callset) is a complete re-run from genotype files to fitted model, and each was finished within one to three days of being requested. In our estimation a study of this scope, with purpose-built exact theory, a genome-scale data pipeline, and this many rival explanations tested quantitatively rather than discussed, would ordinarily occupy a small research group for a year or more.

We think this way of working can change what is routinely attempted in population genetics. The field is rich in public data and rich in theory, and the scarce resource is usually the skilled labor that connects the two with care: deriving and implementing exact expectations for a statistic that no existing package computes, engineering a pipeline over thousands of genomes, and then pursuing each systematic effect and each alternative explanation to a quantitative conclusion. When that labor becomes cheap, statistics richer than the single-site spectrum, of which the 2-SFS is only the simplest, become practical inference tools; robustness analyses that are customarily left to a sentence in the discussion can simply be run; and existing public datasets, which contain a great deal more than has been extracted from them, can be re-examined question by question at roughly the cost of asking. Depth matters here as much as speed. Most of what separates the present analysis from a first pass (the history-matched nulls, the class-wide certificates, the elimination in Table 1, the five-way measurement of the mislabeling rate) exists because asking for it cost hours rather than months.

The approach has failure modes of its own, and the working rules above exist because we encountered them. Claude makes the mistakes a capable analyst in a hurry makes: a model operator assembled with the diagonal cells of the spectrum counted twice, a covariance routine that silently returned a meaningless matrix when handed more parameters than data points, a summary sentence that claimed more than the test beneath it had established. In this project such errors were caught by built-in consistency checks (an operator must reproduce a known case before it is used), by the prior specification of tests, by the adversarial second session, or by one of us reading critically; they were caught because the process assumed they would occur, and we expect scaffolding of this kind, more than the raw capability of the model, to determine how reliable analyses done this way prove to be. The human contribution does not disappear. It moves from performing analyses to deciding which analyses are worth performing and whether to believe their results,

and the second of those tasks, in this project, still required a population geneticist. The machine record of the project (code, result records, and intermediate outputs) will be deposited as described under Data availability.

Data availability

The 1000 Genomes high-coverage sequencing data analyzed here are publicly available (Byrska-Bishop et al. 2022), as are the HGDP callset (Bergström et al. 2020) and the deCODE genetic maps (Palsson et al. 2025); we took ancestral-allele assignments from the public Ensembl EPO alignments (Herrero et al. 2016). We will deposit the analysis code, together with the intermediate outputs supporting the results reported here, in a public repository upon publication.

Literature Cited

R. B. Corbett-Detig and D. L. Hartl, "Population genomics of inversion polymorphisms in *Drosophila melanogaster*," PLoS Genet. **8**, e1003056 (2012).

J. F. C. Kingman, "The coalescent," Stoch. Proc. Appl. **13**, 235 (1982).

J. B. Lack, C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley, and J. E. Pool, "The *Drosophila* Genome Nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population," Genetics **199**, 1229 (2015).

R. Nielsen, "Estimation of population parameters and recombination rates from single nucleotide polymorphisms," Genetics **154**, 931–942 (2000).

G. T. Marth, E. Czabarka, J. Murvai, and S. T. Sherry, "The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations," Genetics **166**, 351–372 (2004).

R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, "Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data," PLoS Genet. **5**, e1000695 (2009).

L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, "Robust demographic inference from genomic and SNP data," PLoS Genet. **9**, e1003905 (2013).

X. Liu and Y.-X. Fu, "Exploring population size changes using SNP frequency spectra," Nat. Genet. **47**, 555–559 (2015).

J. Jouganous, W. Long, A. P. Ragsdale, and S. Gravel, "Inferring the joint demographic history of multiple populations: beyond the diffusion approximation," Genetics **206**, 1549–1567 (2017).

J. Kamm, J. Terhorst, R. Durbin, and Y. S. Song, "Efficiently inferring the demographic history of many populations with allele count data," J. Am. Stat. Assoc. **115**, 1472–1487 (2020).

J. Schweinsberg, "Coalescent processes obtained from supercritical Galton–Watson processes," Stoch. Proc. Appl. **106**, 107–139 (2003).

R. Durrett and J. Schweinsberg, "A coalescent model for the effect of advantageous mutations on the genealogy of a population," Stoch. Proc. Appl. **115**, 1628–1657 (2005).

B. Eldon and J. Wakeley, "Coalescent processes when the distribution of offspring number among individuals is highly skewed," Genetics **172**, 2621–2633 (2006).

R. A. Neher and O. Hallatschek, "Genealogies of rapidly adapting populations," Proc. Natl. Acad. Sci. USA **110**, 437–442 (2013).

M. M. Desai, A. M. Walczak, and D. S. Fisher, "Genetic diversity and the structure of genealogies in rapidly adapting populations," Genetics **193**, 565–585 (2013).

M. Birkner, J. Blath, and B. Eldon, "An ancestral recombination graph for diploid populations with skewed offspring distribution," Genetics **193**, 255–290 (2013).

J. P. Spence and Y. S. Song, "Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations," Sci. Adv. **5**, eaaw9206 (2019).

A. Tellier and C. Lemaire, "Coalescence 2.0: a multiple branching of recent theoretical developments and their applications," Mol. Ecol. **23**, 2637–2652 (2014).

B. Eldon, M. Birkner, J. Blath, and F. Freund, "Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents?" Genetics **199**, 841–856 (2015).

J. Koskela, "Multi-locus data distinguishes between population growth and multiple merger coalescents," Stat. Appl. Genet. Mol. Biol. **17**, 20170011 (2018).

S. Matuszewski, M. E. Hildebrandt, G. Achaz, and J. D. Jensen, "Coalescent processes with skewed offspring distributions and nonequilibrium demography," Genetics **208**, 323–338 (2018).

A. M. Sackman, R. B. Harris, and J. D. Jensen, "Inferring demography and selection in organisms characterized by skewed offspring distributions," Genetics **211**, 1019–1028 (2019).

F. Freund, E. Kerdoncuff, S. Matuszewski, M. Lapierre, M. Hildebrandt, J. D. Jensen, L. Ferretti, A. Lambert, T. B. Sackton, and G. Achaz, "Interpreting the pervasive observation of U-shaped site frequency spectra," PLoS Genet. **19**, e1010677 (2023).

R. R. Hudson, "Properties of a neutral allele model with intragenic recombination," Theor. Popul. Biol. **23**, 183 (1983).

S. A. Sawyer and D. L. Hartl, "Population genetics of polymorphism and divergence," Genetics **132**, 1161 (1992).

E. F. Fenton, D. P. Rice, J. Novembre, and M. M. Desai, "Detecting deviations from Kingman coalescence using 2-site frequency spectra," Genetics **229**, iyaf023 (2025).

D. E. Miller, C. B. Smith, N. Y. Kazemi, A. J. Cockrell, A. V. Arvanitakis, J. P. Blumenstiel, S. L. Jaspersen, and R. S. Hawley, "Whole-genome analysis of individual meiotic events in *Drosophila melanogaster* reveals that noncrossover gene conversions are insensitive to interference and the centromere effect," Genetics **203**, 159–171 (2016).

J. M. Comeron, R. Ratnappan, and S. Bailin, "The many landscapes of recombination in *Drosophila melanogaster*," PLoS Genet. **8**, e1002905 (2012).

J. Herrero, M. Muffato, K. Beal, S. Fitzgerald, L. Gordon, M. Pignatelli, A. J. Vilella, S. M. J. Searle, R. Amode, S. Brent, W. Spooner, E. Kulesha, A. Yates, and P. Flicek, "Ensembl comparative genomics resources," Database **2016**, bav096 (2016).

J. E. Wigginton, D. J. Cutler, and G. R. Abecasis, "A note on exact tests of Hardy–Weinberg equilibrium," Am. J. Hum. Genet. **76**, 887 (2005).

M. Byrska-Bishop et al., "High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort," Cell **185**, 3426 (2022).

A. J. Jeffreys and C. A. May, Nat. Genet. **36**, 151 (2004).

G. Palsson et al., "Complete human recombination maps," Nature **639**, 700 (2025).

A. P. Charmouh et al., Mol. Biol. Evol. **42**, msaf019 (2025).

B. V. Halldorsson et al., Nat. Genet. **48**, 1377 (2016).

A. L. Williams et al., eLife **4**, e04637 (2015).

P. F. Palamara et al., Am. J. Hum. Genet. **97**, 775 (2015).
L. Frisse et al., Am. J. Hum. Genet. **69**, 831 (2001).
J. Gay, S. Myers, and G. McVean, Genetics **177**, 2417 (2007).
D. Setter, S. Ebdon, B. Jackson, and K. Lohse, Genetics **222**, iyac100 (2022).
S. R. Browning and B. L. Browning, Am. J. Hum. Genet. **112**, 2167 (2025).
N. Masaki and S. R. Browning, PLoS Genet. 10.1371/journal.pgen.1011951 (2025).
P. D. Keightley and B. C. Jackson, Genetics **209**, 897 (2018).
S. Besenbacher et al., "Multi-nucleotide de novo mutations in humans," PLoS Genet. **12**, e1006315 (2016).
F. Baumdicker et al., "Efficient ancestry and mutation simulation with msprime 1.0," Genetics **220**, iyab229 (2022).
A. Bergström et al., "Insights into human genetic variation and population history from 929 diverse genomes," Science **367**, eaay5012 (2020).
G. McVicker, D. Gordon, C. Davis, and P. Green, "Widespread genomic signatures of natural selection in hominid evolution," PLoS Genet. **5**, e1000471 (2009).
J. A. Tennessen et al., "Evolution and functional impact of rare coding variation from deep sequencing of human exomes," Science **337**, 64 (2012).
J. J. Moré, "The Levenberg–Marquardt algorithm: implementation and theory," in *Numerical Analysis*, edited by G. A. Watson, Lecture Notes in Mathematics Vol. 630 (Springer, Berlin, 1978), pp. 105–116.
P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods* **17**, 261–272 (2020).

A. Supplementary methods

Checks of the ancestral-mislabeling rate.

The mislabeling rate ε enters the model only through the flip operator of Eq. (7), so it can be re-estimated on any subset of sites by refitting the same three-parameter model (demographic weights, conversion amplitudes, ε) to the two-site spectrum built from that subset alone, in the same jackknife metric. We use four such subsets, each defined per site from the reference and alternate alleles before any use of the ancestral labels, so that the subset definition cannot itself depend on mislabeling. (i) *GC-conservative classes*: sites whose two alleles are both weak (A/T) or both strong (G/C); GC-biased gene conversion, which favours strong over weak alleles in heteroduplex repair, has no allele to prefer at such sites, so any apparent mislabeling there cannot be a disguised transmission bias ($\varepsilon = 1.54\%$). (ii) *Transitions and transversions*, fitted separately (1.42% and 0.79%); parsimony-style ancestral inference errs roughly in proportion to the substitution rate of the class, so the ratio is expected to lie near the transition/transversion rate ratio of about two (observed 1.79). (iii) *Individual chromosomes*: the nineteen chromosomes fitted one at a time, which agree within their errors. (iv) *The folded spectrum*: the two-site spectrum symmetrized under $i \to n - i$ at each site, in which a single-site label flip cancels identically; fitting the conversion terms to the folded spectrum, with no ε at all, tests whether the conversion signal survives when mislabeling is removed by construction rather than fitted. It does, and the division of the conversion rate between the two tract classes is preserved.

GC-biased gene conversion as an alternative.

Biased conversion would act only at weak/strong-discordant sites and would push derived strong alleles to higher frequency and derived weak alleles to lower frequency; in the two-site spectrum it therefore produces a displacement confined to the W→S and S→W classes and of opposite sign in the two. We test it in two ways. First, the deviation from the Kingman class and the fitted conversion and mislabeling parameters are re-derived on the GC-conservative classes alone, where biased conversion is absent by construction; the deviation persists there with the genome-wide shape and amplitude, and ε is unchanged within error (1.54% against 1.67%). Second, a biased-conversion displacement built from the measured difference between the W→S and S→W class spectra is offered to the combination fit as an additional regressor; it absorbs a negligible share of the deviation and leaves the unbiased-conversion amplitudes unchanged. Biased conversion is thus present in the data at the level expected from the class asymmetry but is not the process that displaces the genome-wide spectrum from the Kingman class.

Class-wide lower bound on the distance to the pooled-Kingman hull.

Let H be the set of all normalized two-site spectra obtainable as mixtures of pooled-Kingman histories: the convex hull of the normalized spectra $q(t_a, t_b) \propto C^{(ab)}$ of Eq. (4) over all pairs of depths (we use the continuum of depths, of which the 24-level grid is a subset, so the bound covers every history and every mixture at any resolution). For a target spectrum ϕ (the data, or a candidate model's spectrum) and any linear functional ψ on the 28 cells, the Cauchy–Schwarz inequality in the jackknife metric gives, for every $v \in H$,

$$\chi^2(v) = (\phi - v)^\top \widehat{\Sigma}^{-1} (\phi - v) \;\geq\; \frac{\left(\psi \cdot \phi - \max_{u \in H} \psi \cdot u\right)^2}{\psi^\top \widehat{\Sigma}\, \psi}$$

whenever $\psi \cdot \phi$ exceeds the maximum of ψ over the hull. The right-hand side is therefore a lower bound on the distance from ϕ to *every* member of the class at once, valid for any choice of ψ; we choose ψ numerically as the direction of the metric projection of ϕ onto H (a quadratic minimization over mixture weights on a fine dictionary of hull generators, which has no local optima because H is convex), and we then *certify* the maximum of ψ over the hull independently of that numerical step: $\psi \cdot q(t_a, t_b)$ is a ratio of polynomials in the two depths with rational coefficients, and its upper bound c is established by verifying $c\,T - \Psi \geq 0$ on the unit square in exact rational arithmetic (Bernstein-basis expansion with de Casteljau subdivision), where Ψ and T are the numerator and normalizing polynomials. The reported exclusion of a family (Table 1) is the amplitude at which this certified bound reaches the rejection threshold; because only the optimality of ψ, not the validity of the bound, depends on floating-point computation, the exclusions hold irrespective of numerical error in the fit.

B. Supplementary data

Supplementary Data 1: Catalogue of excluded sites. One compressed CSV file per autosome (`excluded_sites_chrN.csv.gz`; columns: chromosome, position (GRCh38), reference allele, alternate allele, reason) listing every record of the 1000 Genomes high-coverage genotype VCFs that did not enter the two-site analysis, with the first failing criterion in the order the filters are applied: not PASS; multiallelic; not a strict single-nucleotide variant; missing genotype in at least one of the 113 individuals; monomorphic among the 226 GWD haplotypes; excess heterozygosity (one-sided exact test, $p < 10^{-6}$). A summary table (`SD1_counts_table.csv`) gives the counts by reason for each chromosome and in total (123.6 million records, of which 21.5 million segregating sites are retained), and reproduces the counts recorded by the production pipeline.

Supplementary Data 2: Class-wide bounds and per-model results. (i) The lower-bound construction of Appendix A applied to each excluded family, with its certificate (the separating functional, its certified maximum over the hull, and the resulting bound in standard deviations); (ii) a CSV table of the per-model results quoted in the text: for the pooled-Kingman class and for each candidate process and combination of Section 3.2, the fitted χ^2, its matched-null 95th percentile, the fitted amplitudes with their physical bounds (Table 1), and the fraction of the deviation removed.

C. Supplementary Figures

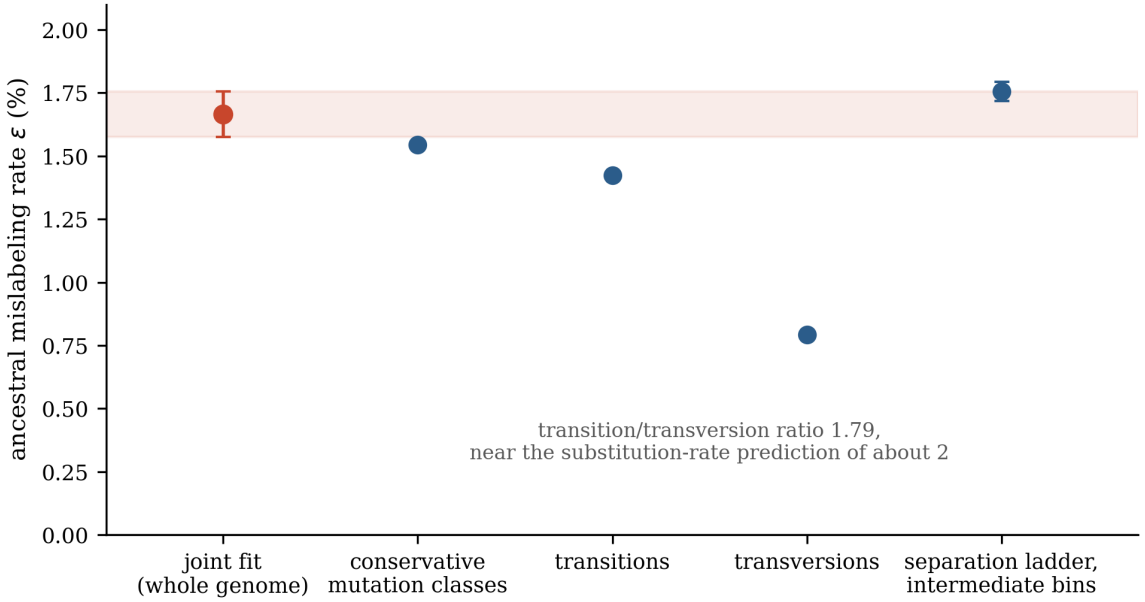

Figure S1 The ancestral-mislabeling rate, measured five ways. The joint-fit value $(1.67 \pm 0.09)\%$ is shown beside four independent slices. Conservative mutation classes where GC-biased conversion cannot act give 1.54%; transitions (1.42%) and transversions (0.79%) have a ratio of 1.79, near the substitution-rate prediction of roughly 2; the intermediate bins of the physical-separation set (Section 2) give 1.75–1.76% once the regressor comes from the fitted demographic family, consistent with the joint fit at one sigma (a flat-history regressor gives 0.96–1.06%, an artifact of that regressor, not shown). Each subset requires a nonzero rate of the same order, as expected for an error of the ancestral-state assignment that is roughly uniform across the genome.