THE UNIVERSITY OF CHICAGO

INFERENCE AND ANALYSIS OF LOW-DIMENSIONAL REPRESENTATIONS OF GENETIC HISTORY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
AND
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY
AHMED SELIM

CHICAGO, ILLINOIS
DECEMBER 2026

© 2026 by Ahmed Selim

This work is openly licensed via CC BY 4.0.

0009-0007-8192-0640

To Mama and Baba,

the reasons for everything good in my life.

بسم الله الرحمن الرحيم

اللهم صل على سيدنا محمد النور وآله

TABLE OF CONTENTS

LIST OF FIGURES viii

LIST OF TABLES xxii

ACKNOWLEDGMENTS xxiii

ABSTRACT xxvi

1 INTRODUCTION 1

2 A LATENT-SPACE MODEL FOR LOW-DIMENSIONAL REPRESENTATIONS OF COALESCENT STRUCTURE THROUGH TIME (COASTT) 12
- 2.1 Introduction 12
- 2.2 Theoretical overview: Latent space models in network analysis 13
- 2.3 A latent position model for coalescent structure through time 16
  - 2.3.1 Method overview: From coalescent rates to latent trajectories through time 16
  - 2.3.2 Likelihood of the coalescent latent position model 17
  - 2.3.3 Computing latent embeddings through time 26
  - 2.3.4 Regularization for smoothing latent space trajectories 29
  - 2.3.5 Characterization of different choices of the link function 33
- 2.4 Properties of CoaSTT under simple demographic scenarios 37
  - 2.4.1 Samples from a panmictic population form circular clusters with cluster size corresponding to population size 37
  - 2.4.2 Samples from isolated populations form distinct circular clusters . . . 42
  - 2.4.3 Gene flow causes population clusters to approach one another at a rate that increases with migration rate 44
- 2.5 CoaSTT applied to ARGs simulated under complex demographic scenarios . 47
  - 2.5.1 Computing coalescent-rate estimates and their variances from ARGs . 48
  - 2.5.2 CoaSTT jointly reveals split times and population size changes without defining population units 56
  - 2.5.3 CoaSTT jointly reveals continuous and discrete coalescent structure in the presence of gene-flow 59
  - 2.5.4 CoaSTT applied to region-specific coalescent rates can resolve the history of admixed individuals 69
- 2.6 Discussion 84

Appendix 2A 86

Appendix 2B 91

Appendix 2C 94

Appendix 2D 96

3 EMPIRICAL ANALYSIS OF THE GENETIC HISTORY OF HUMANS AND CHIMPANZEES USING COASTT . . . 102
3.1 Introduction . . . 102
3.2 Analysis of bias associated with estimating coalescent rates from ARGs . . . 103
3.2.1 ARG inference methods smooth changes in coalescent rates resulting from demographic events . . . 105
3.2.2 CoaSTT applied to ARGs inferred from simulated data . . . 107
3.3 Exploring individual-based coalescent history of human samples using CoaSTT 115
3.3.1 Global continuous structure of human coalescent history . . . 116
3.3.2 Comparing input rate estimates and rates reconstructed from the CoaSTT embeddings of the HGDP ARGs . . . 129
3.3.3 Visualizing uncertainty in the embeddings . . . 135
3.3.4 Regional analysis of human population history with CoaSTT . . . 142
3.4 Application of CoaSTT to chimpanzee data reveals recent and deep structure within subspecies of chimpanzees . . . 152
3.4.1 Description of the chimpanzee dataset and ARG inference using SINGER 154
3.4.2 CoaSTT reveals extensive recent population structure within Central and Eastern chimpanzee subspecies . . . 160
3.4.3 RS-CoaSTT reveals an admixture history for the Nigeria-Cameroon chimpanzees . . . 177
3.5 Discussion . . . 183

4 THEORETICAL LIMITS OF LOW-DIMENSIONAL REPRESENTATIONS OF POPULATION STRUCTURE USING PCA . . . 186
4.1 Introduction . . . 186
4.2 Theoretical Overview . . . 189
4.2.1 Overview of relevant Random Matrix Theory results . . . 189
4.2.2 Writing the theoretical GRM as a function of mean pairwise coalescent times . . . 191
4.3 Methods . . . 196
4.3.1 Spectral decomposition of the theoretical covariance matrix under a generic demography and sampling scheme . . . 197
4.3.2 Critical point of phase transition and cosine similarity expressed in terms of the spectrum of G . . . 206
4.3.3 Empirical identification of critical points from simulated data . . . 212
4.4 Results . . . 215
4.4.1 RMT phenomena reproduced with an effective number of loci L_{eff} . . . 216
4.4.2 Revisiting the two-population detectability conditions . . . 218
4.4.3 Generalizing the detectability conditions for an arbitrary demography with two closely related populations . . . 222
4.4.4 Deriving the conditions for which PCA will reveal spatial structure under homogeneous migration . . . 227

4.4.5 Empirical investigation of the alignment of the top PCs to geography under more realistic sampling and historical scenarios 233
4.5 Discussion 242
Appendix 4A 245
Appendix 4B 254
Appendix 4C 257
Appendix 4D 265

5 CONCLUSION 278

REFERENCES 283

LIST OF FIGURES

1.1 **Figure 1.1: Describing history in terms of coalescent rates through time.** Pairwise coalescent rates computed as described in section 2.5.1 between 25 individuals assigned labels YRI, CEU and CHB, obtained from ARGs inferred by Wohns et al., 2022. **A)** The mean coalescent rate for pairs of samples assigned the same population label provides a summary of the ancestral history of a single population. The inverse of the mean rate is interpreted as the effective population size through time. **B)** The mean coalescent rate between populations provides a proxy for the degree of gene-flow between a pair of populations at a given point in time. **C)** The coalescent rate between each pair of individuals through time provides a comprehensive summary of coalescent history, but is extremely noisy and difficult to interpret without grouping individuals into population units and computing within-population and between-population means. 8

2.1 **Figure 2.1: Illustrative example of CoaSTT:** CoaSTT applied to 20 samples with identical time-varying coalescent rates. **A)** Coalescent rates that are fed to CoaSTT, equivalent to the inverse effective population size, as a function of time. **B)** CoaSTT embeddings shown for two representative time slices, $t = 300$ generations ago (bottom) and $t = 600$ generations ago (top). Dashed lines that connect the same two samples highlight how their distance increases as their coalescent rate decreases. **C)** Latent positions as a function of time, with the lines connecting the two samples at both time points shown. **D)** Pairwise distances between samples' embeddings as a function of time in gray, with the distance between the two focal samples highlighted in black. Average distance over time is plotted with $N_e(t)$ for comparison. 17

2.2 **Figure 2.2: CoaSTT graphical model.** The set of latent positions of samples determines the coalescent rate between each pair of samples at a given point in time. The dashed line indicates the deterministic relationship in equation 2.2. For a given pair of samples, the number of independent trees where a coalescent event occurs is determined by the pairwise coalescent rate and the number of uncoalesced trees. 18

2.3 **Figure 2.3: Gaussian vs. Cauchy link functions.** For a set of pairs with similarities lying within the range $[0.05, 0.1]$, the corresponding distances in the latent space span a much wider range if a Cauchy link function is used compared to a Gaussian link function. 35

2.4 **Figure 2.4: Samples obtained from a single panmictic population form circular clusters.** CoaSTT applied to 20 samples obtained from a panmictic population with constant population size at a single time slice, using the same random initialization and different choices of the link function. Samples uniformly occupy circular clusters except for very slowly or very rapidly decaying link functions. $N_{\min} = 1$ and $N_{\max} = 10^6$ in all examples. 39

2.5 **Figure 2.5: Cluster size as a function of population size under different link functions.** Average pairwise distance between the embeddings of $n = 20$ samples obtained from a single panmictic population increases as population size increases. The relationship between average pairwise distances and N_e depends on the link function, with $f(x) = (1 + x^p)^{-1}$ exhibiting a linear relationship. . . 40
2.6 **Figure 2.6: Cluster size as a function of N_{min} and n. A)** Average pairwise distance as a function of N_{min} with a constant number of samples $n = 20$. **B)** Average pairwise distance as a function of the number of samples, with constant $N_{\text{min}} = 1$. 41
2.7 **Figure 2.7: CoaSTT embeddings for samples obtained from distinct populations.** Embeddings of 20 samples obtained from each of 10 panmictic populations with different population sizes, under a variety of link functions. **A)** $N_{\text{max}} = 10^5$, one order of magnitude larger than the largest population size. Clusters fail to form under several link functions, but when they form, cluster sizes are visually distinguishable. **B)** $N_{\text{max}} = 10^6$, clusters form more robustly for a variety of the link functions; sizes could be more difficult to resolve visually. 43
2.8 **Figure 2.8: Time-varying coalescent rate for coalescence with migration.** Two populations with equal sizes $N = 10^4$ and historical migration with a constant rate m between them. Within-population rate decreases with time while between-population rate increases until the populations become effectively well-mixed with a single coalescent rate. 45
2.9 **Figure 2.9: Limiting and transient behavior of coalescence with migration. A)** Limiting coalescent rate of λ_{within} and λ_{between} as a function of population-scaled migration rate. The inverse of the limiting coalescent rate defines the effective population size of an ancestral well-mixed (panmictic) population. **B)** The exponential rate at which panmixia is reached as a function of population-scaled migration rate. 46
2.10 **Figure 2.10: CoaSTT applied to a population split with migration.** Both populations have a constant size of 10^4 with migration occurring at a constant rate m between them only within a specified period, either in the recent time **(A)** or right after they split **(B)**. 47
2.11 **Figure 2.11: CoaSTT applied to a complex 4-population demography.** CoaSTT embeddings for 20 individuals per population simulated under the demographic history **(A)**, where the width of the rectangle representing each subpopulation reflects population size. **B)** CoaSTT embeddings reveal four clusters in recent time that merge at split times. The size of each cluster varies according to each population's size. **C)** Mismatch between input rate estimates and rates reconstructed from latent positions reveals the loss of information resulting from dimensionality reduction. **D)** Different metrics of reconstruction error as a function of the rate of decay of the inverse-power link function, with higher degree p indicating faster decay. $p = 1$ highlighted with bold points. 58

2.12 **Figure 2.12: Probabilities of tree topologies of the complex 4-population demography.** The probability mass function of gene trees for one haploid sample per population simulated under the demography in figure 2.11A. Obtained numerically by simulating 10^4 single trees using msprime. The true nested topology is shown in red, with four topologies being more probable than it. 59
2.13 **Figure 2.13: Complex 4-population demography with recent migration**. Populations A and B exchange migrants following their split at a rate of $m = 10^{-4}$ **(A)**, $m = 10^{-3}$ **(B)** and $m = 10^{-2}$ **(C)** per generation. **D)** Within-population coalescent rates as a function of time at each migration rate plotted with the trajectory of inverse population size for populations A and B, illustrating a mismatch between inverse population sizes and within-population coalescent rates that gets larger with higher migration rates. 61
2.14 **Figure 2.14: 1D stepping-stone model.** CoaSTT embeddings for individuals sampled from 1D stepping-stone simulations with different effective migration rates ($Nm = 5, 1, 0.1$). **A)** Embeddings computed using the link $f(x) = \frac{1}{1+x}$ form fragmented linear clusters in recent time. **B)** Embeddings computed using the link $f(x) = \frac{1}{1+x^2}$ with a faster-decaying tail are better behaved as they are continuously structured in recent time. 64
2.15 **Figure 2.15: 2D stepping-stone model.** CoaSTT embeddings for individuals sampled from 2D stepping-stone simulations with different effective migration rates ($Nm = 5, 1, 0.1$). **A)** Embeddings computed using the link $f(x) = \frac{1}{1+x}$ form fragmented and uneven clusters of grids in recent time. **B)** Embeddings computed using the link $f(x) = \frac{1}{1+x^2}$ with a faster-decaying tail are more continuously structured in recent time, though the structure is still distorted for $Nm = 0.1$, likely due to the noisy low-rate estimates. Note how a stray sample could escape its natural position on the grid under random initializations, which is the result of the configuration in recent time landing in a local minimum due to the noisy rates there. 65
2.16 **Figure 2.16: CoaSTT jointly reveals discrete and continuous coalescent structure. A)** Schematic of the simulation. A single population split into two groups, each consisting of a 5×5 grid of smaller demes, with migration occurring between neighboring demes at a constant rate, as well as weaker migration between demes of different groups. **B)** 3D CoaSTT embeddings with rates estimated from simulated trees. **C)** 2D Embeddings at four representative time slices. **D)** Plots of input rates (estimated from trees) vs. reconstructed rates (calculated from latent positions) at each of these times. 68

2.17 **Figure 2.17: CoaSTT applied to a simulation with two distinct admixture events. A)** A schematic of a simulated demography with two different admixture events occurring at different times. Each admixture event involves a different pair of source populations. **B)** GW-CoaSTT embeddings. Each admixed cluster lies between the source-population clusters during $t_{\text{admix}} < t < t_{\text{split}}$, pulling both source populations closer together, which is reflected as inflated reconstructed rates in **(C)**. **D)** RS-CoaSTT embeddings resolve the alternative coalescent histories of each set of admixed samples, with the source populations no longer getting pulled towards each other. 70

2.18 **Figure 2.18: RS-CoaSTT embeddings with GNN-based local-ancestry assignments. A)** Embeddings estimated using region-specific input coalescent rates for admixed samples. As a reminder, the population labeled 'Admix_1' has 'A' and 'B' as sources, and 'Admix_2' has 'C' and 'D' as sources, with a more recent admixture time. A genomic region of an admixed sample is assigned one out of four labels based on its genealogical nearest neighbors in this region. **B)** Region-specific estimates of coalescent rates as a function of time for all pairs of samples where one of the samples in the pair is admixed, colored by the population label of the other sample, with the mean rate shown in bold. Rates do not converge to a single value at $t > t_{\text{split}}$ right away due to the biased mis-assignment of local ancestry. **C)** Region-specific rate estimates vs. rates reconstructed from the region-specific embeddings, showing less inflated reconstructed rates compared to GW-CoaSTT. 74

2.19 **Figure 2.19: RS-CoaSTT embeddings for the two-admixture simulation with RFmix local-ancestry assignments. A)** Source populations are still pulled towards each other at $t = t_{\text{admix}}$, but less severely than using genome-wide rates. **B)** Region-specific coalescent rates are biased upward during $t_{\text{admix}} < t < t_{\text{split}}$ between admixed samples and samples from the other source population (e.g. admix-B rates for rates computed using A ancestry segments). **C)** Input vs. reconstructed rates computed from the region-specific embeddings. Reconstructed rates are inflated less severely, especially between the source populations that participated in the more recent admixture event. 76

2.20 **Figure 2.20: GW-CoaSTT applied to three BADMIXTURE simulations that give identical ADMIXTURE results. A)** The demographic history for each example: Three-way admixture, ghost admixture and recent bottleneck. Samples from population A (in red) show identical ADMIXTURE results under the three scenarios. **B)** GW-CoaSTT embeddings computed using genome-wide rates are different for the three scenarios. In the first scenario, the admixed samples lie in between the source-population clusters, pulled closer to the clusters from which they draw higher admixture proportions (B & C). In the second scenario, one of the source populations is not sampled. Therefore, the admixed population is drawn only to population B (the sampled source) around the admixture time. At the time where the ghost population and population C merge, the admixed samples are pulled by the C cluster as they start coalescing with samples from C. The contribution of the ghost population to the admixed samples is detected through coalescence with C lineages. In the last example of recent bottleneck, GW-CoaSTT gives the expected embeddings from a population split with size changes. 78
2.21 **Figure 2.21: RS-CoaSTT embeddings for the BADMIXTURE examples using GNN-based local-ancestry assignment.** Each column represents a demographic scenario, and each row shows the region-specific embeddings for that scenario. 82
2.22 **Figure 2.22: RS-CoaSTT embeddings for the BADMIXTURE examples with local ancestry inferred using RFmix.** Each column represents a demographic scenario, and each row shows the region-specific embeddings for that scenario. 83
2.23 **Figure 2.23: Options for displaying CoaSTT embeddings through time.** Smoothed embeddings of the complex 4-population simulation (figure 2.11) shown as **A)** individual trajectories, **B)** individual trajectories with the median trajectory of each population, and **C)** the support tube of each population (trimmed convex hulls at each time point, smoothed along time), with individual trajectories drawn faintly on top. 100
2.24 **Figure 2.24: Interactive exploration of CoaSTT embeddings.** Snapshot of the Dash application for genome-wide embeddings of the two-admixture simulation (figure 2.17) at $t = 1000$ generations ago, between the admixture and split times. Top left: 3D trajectories with a plane marking the selected time. Top right: latent positions at that time, zoomed on populations A, B and Admix_1, with edges between the five most over-estimated (red, solid) and under-estimated (blue, dashed) pairs. The A–B pairs are drawn too close because the admixed individuals pull their source populations together (section 2.5.4). Bottom: input versus reconstructed rates at that time, colored by population pair, with the median for each population shown as diamonds. 101

3.1 **Figure 3.1: Coalescent rates estimated from ARGs inferred by different methods.** In both panels, black is the analytic truth, gray is the rate estimated from the true simulated trees, and colors correspond to different ARG inference methods. Thin lines are estimates of pairwise coalescent rates between individuals, and thick lines are their mean. **A)** Population split scenario: two populations of size $N_e = 10^4$ diverging from an ancestral population of the same size 2,000 generations ago, with 20 diploid individuals per population. The upper panel shows within-population rates and the lower panel between-population rates. Dashed vertical lines mark the split time estimated for each method as the time at which the ratio of between- to within-population rates reaches 0.5. **B)** Sawtooth demography, with a table of the log-contrast retained defined in equation 3.1 for mean rate trajectories derived from each ARG inference method. The columns indicate the true times of the extrema, but the contrast in estimated rates is evaluated at the corresponding extrema of each estimate. 107

3.2 **Figure 3.2: CoaSTT embeddings with rates estimated from true trees and SINGER-inferred trees. A)** Population split. **B)** Sawtooth demography. **C)** 4-population demography of figure 2.11. **D)** Demography of two spatial populations shown in figure 2.16. 109

3.3 **Figure 3.3: CoaSTT embeddings for a four-population out-of-Africa model, with rates estimated from simulated and SINGER-inferred trees.** Simulations follow the `OutOfAfrica_4J17` model of Jouganous et al. (2017) in `stdpopsim` (Adrion et al., 2020a), with 20 samples per population, and trees are inferred with SINGER + POLEGON as in figure 3.1. **A,B)** 3D embeddings from simulated (A) and SINGER-inferred (B) trees, with years assuming a 29-year generation time. Gray planes mark the time slices shown in **(C,D)**, drawn on a common scale for each slice. 112

3.4 **Figure 3.4: RS-CoaSTT embeddings resulting from SINGER-inferred trees.** Trees were inferred using SINGER + POLEGON from simulations based on the demography shown in figure 2.17. Each column corresponds to one of the four source ancestries. **A)** GNN-based RS-CoaSTT embeddings, where coalescent rates for admixed individuals are computed using only genomic regions assigned to the column's ancestry based on GNNs. **B)** Region-specific coalescent rates between admixed samples and samples from all other populations, with local ancestry assigned based on GNNs: the first two columns show rates for Admix_1 restricted to A- and B-ancestry segments, and the last two show rates for Admix_2 restricted to C- and D-ancestry segments. **C,D)** Same as **(A,B)**, but with local ancestry assigned using RFmix. Incorrect local ancestry assignment in recent time results in non-zero coalescent rates between Admix_1 samples and population B when restricted to A-ancestry segments (first column), and between Admix_1 samples and population A when restricted to B-ancestry segments (second column). A similar effect is visible for Admix_2 in the third and fourth columns. 114

3.5 **Figure 3.5: CoaSTT embeddings for 54 HGDP populations. A)** A map depicting sample locations and population labels with their abbreviations. **B)** CoaSTT embeddings for four individuals from each population, with time (in generations or years ago) stacked vertically and changing from present to past moving downwards. Samples are colored by region, with abbreviations as in A. Six representative time slices, numbered (1)–(6), are highlighted. Within each, sub-insets (i–iii) zoom into sets of samples. 117

3.6 **Figure 3.6: CoaSTT embeddings for 54 HGDP populations 50 generations ago. A)** The full time slice, identical to inset (1) of figure 3.5B. Panels B–G zoom into parts of A to show the arrangement of samples more clearly. **B)** African samples other than the San. **C)** Middle Eastern samples. **D)** European and Central and South Asian samples. **E)** East Asian samples. **F)** American samples. **G)** Oceanian samples. 121

3.7 **Figure 3.7: The Silhouette Coefficient of each population 50 generations ago.** The SC is computed for each sample using distances in the embedding space (small, transparent points) and distances obtained from the rate estimates directly by inverting the link function (small, transparent crosses). The mean SC of samples in each population is shown by a large point and a large cross. The SC computed from the embeddings and from the inverted rates mostly match when they fall above 0.7 (the rough rule of thumb for tight clusters), except for a few exceptions. Below that, when populations do not form tight clusters, the exact set of neighbors in the embedding space tends not to match those obtained from the input rates, so the resulting SCs differ. 123

3.8 **Figure 3.8: The Silhouette Coefficient, computed from embedding distances and inverted rates, as a function of time, for all populations.** Panels are sorted in descending order by the time point at which the mean SC of a population first decreases below 0.5. 128

3.9 **Figure 3.9: Probing the mismatch between input rate estimates and rates reconstructed from human CoaSTT embeddings. A)** Density plot of rates reconstructed from embedding distances and input rate estimates, for all pairs and time points, along with three measures of correlation. The spike of input rates at 10^{-7} is because $N_{\max} = 10^7$, so any rate below 10^{-7} gets pushed to that value. **B)** Decomposition of the source of mismatch between input and reconstructed rates for the top 5% outliers (on the log scale) over all time points. An impossible triangle is a triplet of distances that is not realizable in Euclidean space, likely attributable to admixture, and dominates recent time. **C)** Three measures of correlation between input and reconstructed rates at each time point. The drop in correlation beyond 10^3 generations ago is likely due to crowding. **D)** The standard deviation of the input and reconstructed rates over time, as well as the standard deviation of the reconstructed rates from a configuration fitted to an input of identical pairwise rates, equal to the mean rate at each time point (labeled as 2D floor). 131

3.10 **Figure 3.10: Triplets of human samples most violating the triangle inequality.** Each element is the average violation $\bar{v}(i, j; k)$ over all time points for the triplet formed by the edge samples (i, j) on the row and the bridge sample k on the column, shown for all triplets. Values below the 95th percentile of the non-zero elements are collapsed to a single color. 134
3.11 **Figure 3.11: CoaSTT embeddings from 8 posterior ARG samples for the 54 HGDP populations.** 138
3.12 **Figure 3.12: CoaSTT embeddings from 8 different initializations for the 54 HGDP populations.** 139
3.13 **Figure 3.13: CoaSTT embeddings at two different time slices from 8 posterior ARG samples for the 54 HGDP populations.** 140
3.14 **Figure 3.14: CoaSTT embeddings at two different time slices from 8 different initializations for the 54 HGDP populations.** 141
3.15 **Figure 3.15: CoaSTT embeddings for 104 samples from 7 populations in Africa.** Biaka (n=22), Mandenka (n=22), Yoruba (n=22), Mbuti (n=13), BantuKenya (n=11), BantuSouthAfrica (n=8) and San (n=6). 145
3.16 **Figure 3.16: CoaSTT embeddings for 100 samples from 4 populations in the Middle East.** Bedouin (n=25), Druze (n=25), Mozabite (n=25) and Palestinian (n=25). 146
3.17 **Figure 3.17: CoaSTT embeddings for 155 samples from 8 populations in Europe.** French (n=28), Sardinian (n=28), Russian (n=25), Basque (n=23), Adygei (n=16), Orcadian (n=15), BergamoItalian (n=12) and Tuscan (n=8). . 147
3.18 **Figure 3.18: CoaSTT embeddings for 130 samples from 9 populations in Central and South Asia.** Balochi (n=15), Brahui (n=15), Burusho (n=15), Hazara (n=15), Kalash (n=15), Makrani (n=15), Pathan (n=15), Sindhi (n=15) and Uygur (n=10). 148
3.19 **Figure 3.19: CoaSTT embeddings for 168 samples from 18 populations in East Asia.** Han (n=10), Japanese (n=10), Miao (n=10), NorthernHan (n=10), She (n=10), Tu (n=10), Yakut (n=10), Yi (n=10), Cambodian (n=9), Dai (n=9), Daur (n=9), Hezhen (n=9), Mongolian (n=9), Oroqen (n=9), Tujia (n=9), Xibo (n=9), Lahu (n=8) and Naxi (n=8). 149
3.20 **Figure 3.20: CoaSTT embeddings for 28 samples from 3 populations in Oceania.** Bougainville (n=11), PapuanHighlands (n=9) and PapuanSepik (n=8). 150
3.21 **Figure 3.21: CoaSTT embeddings for 61 samples from 5 populations in the Americas.** Maya (n=21), Pima (n=13), Karitiana (n=12), Surui (n=8) and Colombian (n=7). 151

3.22 **Figure 3.22: An overview of the coalescent history of chimpanzees. A)** An illustration of the range of each of the four subspecies of chimpanzees, as well as bonobos, with locations of samples following table 3.1. PCA is shown to separate samples from the four subspecies into four clusters. **B)** Inverse coalescent rates over time for each pair of haplotypes (thin lines), in addition to the mean of within-individual coalescent rates (thick lines). **C)** Relative cross-coalescence rates over time for each pair of populations, with the time at which each curve crosses 0.5 denoted underneath. 158
3.23 **Figure 3.23: CoaSTT embeddings of chimpanzees between 50 and 4,000 generations ago.** 167
3.24 **Figure 3.24: Recent-time chimpanzee embeddings at 60 (A) and 2,000 (B) generations ago from eight SINGER posterior samples.** 168
3.25 **Figure 3.25: Recent-time chimpanzee embeddings at 60 (A) and 2,000 (B) generations ago from eight different initializations.** 169
3.26 **Figure 3.26: 3D recent-time chimpanzee embeddings from eight SINGER posterior samples.** 170
3.27 **Figure 3.27: 3D recent-time chimpanzee embeddings from eight different initializations.** 171
3.28 **Figure 3.28: CoaSTT embeddings of chimpanzees between 4,000 and 400,000 generations ago.** 173
3.29 **Figure 3.29: Probing the mismatch between input rate estimates and rates reconstructed from chimpanzee CoaSTT embeddings. A)** Density plot of rates reconstructed from embedding distances and input rate estimates, for all pairs and time points, along with three measures of correlation. The block of rates reconstructed at a value much lower than the input of $\approx 2 \times 10^{-6}$ corresponds to the admixed samples, 'Donald' (Dn) and 'Julie' (Jl), which are placed close to the western and NC cluster respectively. Their large distance to the central cluster results in reconstructed rates that are much lower than the input rates. **B)** Decomposition of the source of mismatch between input and reconstructed rates for the top 5% outliers (on log-scale) over all time points. Most of the outliers that are due to an impossible triangle involve rates between 'Donald' and central samples and between 'Julie' and central samples. 175
3.30 **Figure 3.30: Chimpanzee samples contributing to high reconstruction error as a result of admixture.** Triplets of chimpanzee samples most violating the triangle inequality. Each element is the average violation $\bar{v}(i, j; k)$ for edge samples (i, j) on the row and bridge sample k on the column, shown for all triplets; values below the 95th percentile of the non-zero elements are collapsed to a single color (as in figure 3.10). 176

3.31 **Figure 3.31: RS-CoaSTT embeddings with local ancestry assigned for NC samples based on GNNs. A, B)** Embeddings through time computed from the western-assigned **(A)** and central/eastern-assigned **(B)** segments of the NC samples and the western-central hybrid; reference samples contribute all of their segments. **C, D)** Time slices of A and B at 2,000, 8,400, 17,000 and 22,000 generations ago, with the same slice on the same scale in C and D. 179
3.32 **Figure 3.32: Relative cross-coalescence rates based on region-specific rates for NC samples.** RCCR between the NC samples ('Julie' excluded, solid) or 'Julie' (dotted) and each of the western, central and eastern samples, computed genome-wide **(A)**, from western-assigned segments **(B)** and from central/eastern-assigned segments **(C)**. Dashed vertical lines and legend times mark where each curve first crosses 0.5. In B and C only the crossings with the assigned source are dated. Samples with low genotype quality are excluded. 181

4.1 **Figure 4.1: Properties of eigenvalues and eigenvectors of the sample covariance matrix of genotype matrices. A)** The observed distribution of the bulk eigenvalues of a sample covariance matrix computed from a $400 \times 20,000$ genotype matrix with samples simulated under a two-population split scenario (excluding the largest eigenvalue and the zero eigenvalue). **B)** Median (over 8 replicates) of the ratio between the largest sample eigenvalue (the spike) and the second largest one (the edge of the bulk), computed over a range of F_{ST} values. The ratio stays close to 1 before the BBP threshold is crossed, and begins to increase after, tracking the true theoretical eigenvalue (which increases with F_{ST}). **C)** Cosine similarity between PC1 and the first theoretical eigenvector computed for 8 random samples of the genotype matrix (each of size $400 \times 20,000$) at each value of F_{ST}, showing an I-spline fit that closely matches the theoretical prediction from RMT (equation 4.34). The bottom 3 insets show PC1 for three different values of F_{ST}: Below the BBP threshold (where the PC is approximately pure noise), slightly above the threshold (where the samples form two very noisy clusters), and well above the threshold (where the samples form two well-defined clusters, matching the true underlying theoretical eigenvector). 217
4.2 **Figure 4.2: Effective number of loci in a genotype matrix with unnormalized columns.** The genotype matrix is obtained from simulating a panmictic population. **A)** The effective number of columns (L_{eff}) of the unnormalized genotype matrix as a function of the true number of independent columns/SNPs (L). **B)** L_{eff} as a function of N for a fixed number of independent SNPs, showing a decrease that eventually plateaus. Larger samples include more rare polymorphic sites, which skews the heterozygosity spectrum and lowers L_{eff}. 218

4.3 **Figure 4.3: Cosine similarity as a function of F_{ST} for the population split demography.** Genotype matrices were simulated under a population split demography, where the split time is varied, changing the F_{ST} between the populations at each simulation. At each value of F_{ST}, 8 random samples of the genotype matrix are obtained. The cosine similarity is computed between PC1 and the theoretical eigenvector for each of the 8 samples of the genotype matrix at each value of F_{ST} (shown by transparent scatter points). Vertical dotted and dashed lines show the theoretical $(F_{ST})_0$ and $(F_{ST})_{0.9}$ respectively, derived from an approximation that closely matches the exact solution (equations 4.33 and 4.34). The ticks above them show the values derived from a cruder approximation (equation 4.37 which leads to equation 4.43). $(F_{ST})_0$ and $(F_{ST})_{0.9}$ derived from the close approximation match the ones empirically derived from the I-spline fits well. $(F_{ST})_0$ derived from the crude approximation matches the ones derived from the close approximation, while $(F_{ST})_{0.9}$ is biased downwards. **A)** Varying the total number of sampled individuals, holding the sampling proportion fixed at 0.5 per population. **B)** Holding the total number of samples fixed at $N = 200$, and varying the proportion of samples ϕ obtained from population 1. 221

4.4 **Figure 4.4: Detecting recent Yoruban fine structure using PCA under biased sampling:** A recent population split in the YRI population is added to the three-population out-of-Africa demography, where the timing of the split is varied, yielding different values of F_{ST} between the two Yoruban subpopulations. **A)** For a recent population split that occurred 25 generations ago (or $\approx$ 625 years ago), PCA is unable to adequately separate the two YRI subpopulations if each of them represents only 5% of the total sample size. Raising the sampling representation to 20% (such that both populations together contribute 40% of the samples) causes the recent fine structure to be better resolved by PCA. **B)** A deeper split that occurred 255 generations ago ($\approx$ 6375 years ago) is more easily detected by PCA, even under severely biased sampling. 224

4.5 **Figure 4.5: Verifying the sister-population eigenvalue expression under the recent Yoruban split scenario:** Under the demography shown in figure 4.4, the third-largest eigenvalue corresponds to the strength of the axis of variation that separates the two recently split Yoruban populations. We computed this eigenvalue for sample covariance matrices computed from several genotype matrices simulated under a variety of split times and sampled under different sampling schemes. **A)** The empirically observed eigenvalues (median over 8 replicates) closely match the ones obtained from the general theoretically derived expression, with ϵ_0 computed by numerically calculating mean within-population coalescent times using `msprime.DemographyDebugger` helper functions. Note that the empirical eigenvalue itself is slightly biased upward based on RMT, but it serves as a valid reference point when the bias is negligible (strong structure). **B)** Empirically observed eigenvalues plotted against ones obtained from an approximation of the general theoretical expression, showing that the shift ϵ_0 is not negligible, and excluding it underestimates the eigenvalue. 225

4.6 **Figure 4.6: Cosine similarity as a function of F_{ST} for the demography with recent Yoruban structure.** Cosine similarity is computed between PC3 and the theoretical eigenvector in the same manner as figure 4.3. The theoretical thresholds derived from the approximations that closely match the exact solution (equations 4.33 and 4.34) are close to the empirically derived threshold with a small bias. The detectability thresholds derived from the crude approximation (equation 4.45, $\rho = 0$) are close to the empirically derived ones, but the points of crossing squared cosine similarity of 0.8 are severely biased. **A)** Varying the total number of sampled individuals, holding the sampling proportion fixed at 0.2 for the two Yoruban populations and 0.3 for CEU and CHB. **B)** Holding the total number of samples fixed at $N = 1000$, and varying the proportion of samples ϕ obtained from each of the two Yoruban populations by choosing a pair of proportions from the set $\{0.05, 0.1, 0.2\}$ to assign to the two Yoruban populations. CEU and CHB are sampled evenly. 227

4.7 **Figure 4.7: Detecting spatial structure using PCA under stepping-stone models. A)** PC1 shown for genotype matrices simulated under the one-dimensional stepping-stone model with three different migration rates ($N_e = 4000$ for each of the 25 demes). As migration rate decreases, the signal for spatial structure is stronger, and PC1 resolves the linear spatial structure more accurately. **B)** PC1 vs PC2 shown for genotype matrices simulated under the two-dimensional stepping-stone model ($N_e = 1000$ for each of the 81 demes). Similar to the 1D case, spatial structure is more clearly resolved for lower migration rates. $L = 10,000$ independent loci are sampled in all cases. 229

4.8 **Figure 4.8: Verifying the eigenvalue expression for the 1D and 2D stepping-stone models:** The largest eigenvalue of the sample covariance matrix was computed for a genotype matrix simulated under a given migration rate and a given number of samples per deme for the 1D **(A)** and 2D **(B)** stepping-stone models. The median of 8 replicates is plotted against the theoretically derived expression in each case. The slight discrepancy at smaller eigenvalues is because the bias associated with the sample eigenvalues predicted by RMT becomes non-negligible when the magnitude of the eigenvalue (strength of structure) is small. 230

4.9 **Figure 4.9: Cosine similarity as a function of $1/M$ for homogeneous stepping-stone models**. Cosine similarity is computed against the first PC in the 1D case **(A)** and against the first two PCs (Procrustes aligned) in the 2D case **(B)**, where the average is taken. Observed cosine similarity is shown by transparent scatter points, with solid lines showing the best I-spline fit. The close approximation (4.33) and crude approximation (4.48,4.49) for the theoretical phase transition value, in terms of $1/M$, accurately match the empirical transition point in the 1D and 2D cases. For the point at which the squared similarity reaches $\rho^2 = 0.8$, both the close and crude approximations work well for the 1D case, while the crude approximation is biased in the 2D case. The close approximation (4.34) accurately matches the value derived from the I-spline in all cases. 233

4.10 **Figure 4.10: Detecting spatial structure using PCA under stepping-stone models with sparse sampling.** Fixing a total number of samples N, a fraction $\phi \in \{0.2, 0.4, 0.6\}$ of the demes in the habitat is sampled, with an equal number of samples obtained from each deme. Each panel shows a single random draw of demes. The squared cosine similarity is computed between the observed PC ($\hat{u}_\phi$) and the theoretical PC associated with dense sampling (u_1), which aligns with geography. Migration rates ($M = 811$ in 1D, $M = 4$ in 2D) are chosen well above the respective detection thresholds so that differences between panels reflect distortion rather than detectability. **A)** 1D stepping-stone model. For the demes drawn here, sampling the habitat more densely does not improve the alignment of PC1 with geography, because the sparser subsets happen to cover the habitat evenly. **B)** 2D stepping-stone model. Sampling the habitat more densely brings the theoretical PCs under sparse sampling closer to those under dense sampling, so that PC1 vs. PC2 aligns better with geography. 235

4.11 **Figure 4.11: The effect of sparse sampling on distorting the spatial PCs in stepping-stone models.** In both 1D **(A)** and 2D **(B)** stepping-stone models, fixing the total number of samples N and varying the fraction of demes sampled mainly affects the maximum realizable alignment between the top PCs and geography. The squared cosine similarity is measured between the observed PCs and the top theoretical PCs under dense sampling, which themselves align with geography. Dotted horizontal lines show the theoretical maximum realizable similarity $\langle u_\phi, u_1 \rangle^2$, averaged over 16 random subsets of demes for each sampling fraction. Individual realizations that plateau well below their dotted line correspond to subsets of demes with large unsampled gaps. 237

4.12 **Figure 4.12: Detecting recent spatial structure using PCA. A)** Demographic model. An ancestral panmictic population splits $t = 500$ generations ago into a panmictic population A and a population B structured as a 9×9 grid of demes with homogeneous stepping-stone migration of strength M. $N_e = 10^4$ for the ancestral population and population A, while $N_e = 124$ for each deme in the spatially structured population (such that $124 \times 81 = 10,044$ is close to the ancestral N_e). **(B–D)** PC2 vs. PC3 of the sample covariance matrix for $M = 0.25$, 2.5 and 25; PC1 (not shown) separates A from B. Samples from B are colored by the position of their deme on the grid and samples from A are shown in gray. Above each panel is the squared cosine similarity between the observed PC2/PC3 subspace ($\hat{u}_{t=500}$) and the theoretical spatial eigenvectors under equilibrium migration (u_∞). The alignment with geography is highest at the intermediate migration rate, where $M = 2.5$. At $M = 25$ spatial structure revealed by PCA is diffuse, while at $M = 0.25$ the samples separate into discrete clusters corresponding to individual demes but the grid is distorted. 238

4.13 **Figure 4.13: Alignment of PC2/PC3 with geography under recent spatial structure is non-monotonic in the migration rate. (A,B)** Points are simulation replicates, solid lines are smoothing-spline fits, and dots mark their maxima; the vertical dotted line in (C) and (D) marks the maximum in (A). Alignment with geography increases as migration decreases, up to a point beyond which it declines. **(C)** Total projection S_r of PCs 2 to r onto the subspace of the first two eigenvectors of the equilibrium stepping-stone model ($\phi = 0.5$, $n_B = 4$). At weak migration PC2/PC3 capture progressively less of the signal while wider sets of PCs retain it: the signal is redistributed, not lost. The dashed line is the same projection for the noise-free theoretical PC2/PC3, which remain undistorted. **(D)** Relative gap between the largest spatial eigenvalue ($\nu_2 = \nu_3$) and the next one (ν_4) (left axis), and the eigenvalue itself (right axis), from the exact expected coalescent times. In the non-equilibrium model the largest spatial eigenvalue flattens while the next keeps growing, collapsing the gap; in the equilibrium model the gap stays near 0.5. 241

LIST OF TABLES

2.1 **Table 2.1:** Variables and notation 19
2.2 **Table 2.2:** Comparison between different embedding methods 87
2.3 **Table 2.3:** Notation for writing the gradient and the Hessian 90
2.4 **Table 2.4:** Main configuration parameters of the CoaSTT pipeline. Times are in generations. 97

3.1 **Table 3.1:** Chimpanzee samples by subspecies and place of origin. Locations, regional clusters and notes follow de Manuel et al. (2016) (Table S1 and Data S1) and Kuderna et al. (2023). 159

ACKNOWLEDGMENTS

One of the main reasons I have always wanted to write a dissertation is to write an acknowledgments section. It is an excuse to express the degree of love and gratitude I carry for the people who have defined who I am, most notably in the seven years of my PhD (or six, depending on how I choose to count). I started my PhD as a 21-year-old who knew very little about life and science, and here I am, a 28-year-old, knowing a bit more about a very niche branch of science, and growing even more perplexed about life. The immense changes that have shaped me, both personally and scientifically, can be fully attributed to a number of people whom I can never thank enough. I will try to express some of this gratitude, insofar as it is expressible in words.

There are a few people in my life who deserve a dissertation of gratitude of their own, and the most prominent are my parents. Whenever I try to put into words how grateful I am to them, I find myself stepping back, noticing my words undersell my feelings. For the sake of brevity, since this paragraph alone could otherwise run longer than some of the chapters, I will simply say I am absolutely nothing without my parents. As long as I live, I hope to bring into their lives even a tiny fraction of the light they have filled mine with. A similar thing can be said about my sisters, Iman and Aya, two luminous gems who continue to fill my life with vitality.

Moving to my scientific life, I want to first thank my advisors during the first phase of my PhD, Seppe and Arvind, from whom I learned what proper science looks like. John, my advisor for the bulk of my PhD, is the person under whose guidance my scientific mind has matured over the past four years. For that, and for the immense support and guidance he has given me during my time in his group, I cannot thank him enough. Numerous other professors contributed to this growth through invaluable discussions, including Matthew Stephens, Jeremy Berg, and Matthias Steinrücken. I owe Matthias special thanks for introducing me to population genetics through his wonderful class. None of my PhD would have been

possible without the support of the incredible Biophysics program. I am deeply grateful to two people in particular, Adam Hammond and Michele Wittels, who went above and beyond to make sure my PhD ran smoothly. I am also grateful to Adam Fine and Andrew Monlina from my Biophysics cohort, who have been such fun company since our first year.

I additionally owe my brilliant colleagues heartfelt gratitude for teaching me so much about population genetics and high-quality science, and for making my PhD so much fun! I thank the members of the Novembre lab, past and present: Sherif, Marida, Cynthia, Carla, Hao, Nuria, Vivaswat, Maggie, Xinyi, and Luke. I especially thank Luke for our long conversations juggling ideas in my office and his. My gratitude extends to the popgen super-group and to the whole fourth floor of Cummings.

Stepping off the UChicago campus into the broader Hyde Park area, I made friends so dear during my PhD that they have become my second family away from home: Sherif, Alaa, Kiran, Fatima, Xhesika, Abdel-Basset, Wael, Amani, Farah, Radwa, Iqra, and the newest addition to our family, Dima. Some have left and some are still here, but the warmth they bring will keep them in my heart wherever they or I may be. Sherif and Alaa are the two older brothers I was blessed with when I came to Chicago. I think something close to 60% of my PhD work was done during our co-working sessions (maybe 15% of which was actual work). Their wives, Kiran and Fatima, filled our gatherings with so much laughter and made sure we never ran short of material for pointless as well as quite serious debates, to which Wael and Amani add the right amount of fuel. Xhesika and Abdel-Basset were also essential members of these gatherings, and their presence, which we truly miss, always made the family feel complete. Another missed member is the brilliant Iqra, who is currently doing her postdoc with Terence Tao and who will grow to become the most remarkable mathematician Kashmir has given us. These gatherings are never complete without Farah, the wonderfully deep conversations she brings, and the two-word witty comments that instantly pull us out of that depth and into laughter. Nor are they complete without Radwa and the absolute

peace and serenity her presence brings. Finally, I cannot forget the most important person on that list, Dima, who is starting to say her first words and will perhaps use them to give me the dearest "mabrook" I could hear on the day of my defense! These gatherings might be the sole reason I survived my PhD.

Another family (I am so blessed with so many families) that has given me even more warmth in the past year and a half is my in-laws. I hold such deep respect and appreciation especially for my brother- and sister-in-law, Wael and Amani, from whom I learn how to be a *real human* just by observing how they go about their lives. I am also truly grateful to my mother-in-law, whose presence eases the constant, quiet ache of living away from my mother.

Taking a further step back to Egypt, no words can express my appreciation for my childhood friends and the support they have surrounded me with. Ali, Yasser, Omar, Lalosh, Hassan, and Karim are so woven into who I am that I cannot imagine where I would be without them. I am also profoundly indebted to my religious and spiritual mentors, whom I came to know during my PhD and who have truly taught me what it means to purify one's heart: Sheikh Mohamed Nassar, Ahmed Sayed, and Alaa Abdelhamid.

Saving for last those dearest to my heart, who had the greatest impact on my PhD, I want to express my deepest thanks and love to my best friend Yasser and to the light of my eyes, my wife Noor. Yasser is someone without whom I cannot fathom what my life would mean. Every second we spend together is worth a lifetime of happiness. He is the primary reason I am back on my feet today, having pulled me out of one of the deepest pits one could imagine falling into in the middle of my PhD, and he remains a constant pillar supporting the integrity and meaning of my world. Noor is the light that brings absolute serenity and solace into my world. The effect her presence has had on my life over the past year and a half, the most intense part of my PhD, cannot be put into words. Every day I spend with her is a reminder of how beautiful life can be. Without exaggeration, the greatest blessing of this seven-year PhD was Noor.

ABSTRACT

Continued advances in whole-genome sequencing have produced a rapidly expanding wealth of genomic data, fueling sustained interest in methods that make sense of patterns of genetic variation. Because such data is inherently high-dimensional, population structure and history are often studied through low-dimensional representations that are more intuitive. Population history is typically described in terms of discrete, well-mixed populations, a mathematically convenient assumption that is rarely satisfied in nature and can reinforce a typological view of human variation. Principal component analysis (PCA) avoids this assumption by representing each individual with a few coordinates, but summarizes the ancestral history in a single, time-averaged picture. In the first part of this dissertation, I develop CoaSTT (Coalescent Structure Through Time), a latent position model that represents each sample by a smooth trajectory in a low-dimensional space. Distances between samples reflect their pairwise coalescent rates, which capture how closely related their genetic ancestors were at each point in the past. In chapter 2, I show through simulations that CoaSTT reveals both discrete and continuous structure in a time-resolved manner, and introduce RS-CoaSTT, an extension that resolves the coalescent history of specific genomic regions, helping disentangle the different ancestral histories of admixed individuals. In chapter 3, I apply CoaSTT to human and chimpanzee samples, recovering well-established features of their histories while revealing additional nuance through an individual-based picture. In the second part of my dissertation, I turn from developing new representations to understanding the limits of the most widely used one: PCA. In chapter 4, I use random matrix theory to characterize when principal components computed from finite samples faithfully reflect the underlying population structure. The work I present in my dissertation provides a theoretically grounded way of visualizing ancestral history at the individual level, and clarifies when PCA can be trusted to capture population structure.

CHAPTER 1

INTRODUCTION

The genetic material of all individuals within a population is connected by common ancestry. Temporally varying reproductive and ecological interactions between communities in the past affect the degree to which present-day individuals are ancestrally related (Maddison et al., 2023). The lines of genetic ancestry of any individual can be traced retrospectively into the past, through their genetic ancestors, to moments where those lines 'coalesce' with the lines of genetic ancestry of other individuals (Rosenberg & Nordborg, 2002). Mutations accumulating along these lineages result in patterns of genetic variation from which ancestral relationships can be inferred. Understanding the history and structure of coalescence of these genetic lineages has long been a fundamental aspect of phylogenetic and population genetic approaches to understanding the genetic history of living individuals (Mirarab et al., 2021; Nordborg, 2019).

An individual's genetic ancestor is another individual who lived in the past and passed a portion of its genome to that descendant (Mathieson & Scally, 2020). Genetic ancestors of an individual are typically a subset of their genealogical ancestors (parents, grandparents, etc.), since it is possible for an individual to derive no genetic material from one of those genealogical ancestors, especially if that ancestor lived in the deeper past (Donnelly, 1983). It is therefore important to emphasize beforehand that any inference of ancestry or population history that relies on genetic data only pertains to the genetic ancestors of the individuals from which genetic data was obtained, rather than all of their genealogical ancestors. The complete genome-wide representation of genetic ancestry for a set of individuals is contained in their Ancestral Recombination Graph (ARG) (Wong et al., 2024). The ARG is a graph that fully characterizes the set of coalescence (two genomic segments being derived from a single common ancestor) and recombination (a single genomic segment being derived from two ancestors) events that relate a set of genomic samples. In its entirety, the ARG represents

all the genetic ancestors of a set of individuals, when they lived, and how they are related to those individuals and to one another. Due to the comprehensive description an ARG provides regarding the ancestral relationships, it is natural to use it as a starting point to understand population history and structure. However, since the formal introduction of the coalescent with recombination (Hudson, 1983) and the eventual development of the ARG as a graph structure (Griffiths & Marjoram, 1996, 1997), the problem of inferring an ARG for a set of DNA sequences has been formidably difficult.

Before the possibility of ARG inference became a reality, population geneticists developed a wide variety of methods that invoke simpler summary statistics to describe and understand patterns of genetic similarity observed from mutational data. Some of the earliest such statistics were the so-called "fixation indices" developed by Sewall Wright, which quantify the degree to which individuals within a subset of a population are more genetically similar when compared to the whole population (Wright, 1949). This notion of "population structure" has since been a central theme in population genetics. With large-scale genetic data becoming more readily available by the early 2000s, tools have been developed to provide a concise representation of population structure from such extremely high-dimensional data. Of these methods, STRUCTURE and Principal Component Analysis (PCA) stand out as two of the most utilized tools that also inspired the later development of numerous other methods (Patterson et al., 2006; Pritchard et al., 2000). STRUCTURE can be seen as a clustering algorithm, whereby each individual is assigned a fractional membership to each of K clusters rather than being assigned a single cluster. PCA is a prominent statistical method that finds a small set of axes that explain the largest amount of observed genetic variation. Both methods are in fact closely related, and belong to the broader family of low-rank matrix factorization methods (Engelhardt & Stephens, 2010). STRUCTURE, PCA, and related methods have the convenient property of describing patterns of similarity in an individual-based manner, where each individual gets their own set of "admixture proportions"

or "PC coordinates". However, these descriptions lack temporal resolution, as the whole history is summarized by a single picture derived from the mutations accumulated throughout the full ancestral history. For instance, PCA was shown to be equivalent to computing the eigenvectors of a matrix whose elements have an expected value related to the mean coalescent time between each pair of samples (G. McVean, 2009). The exact equivalence is subject to scrutiny though, as the observed PCs will only reflect the true underlying eigenvectors under certain conditions related to the strength of structure and sample size (which is the subject of the second part of the dissertation).

Inference of time-resolved population history is a more ambitious project that gained traction by the late 2000s, building on earlier likelihood and Bayesian coalescent methods for inferring population size change (Drummond et al., 2005; Kuhner et al., 1998; Pybus et al., 2000), migration (Beerli & Felsenstein, 2001), and isolation with migration (Hey & Nielsen, 2004; Nielsen & Wakeley, 2001). In addition to the availability of large-scale genomic data, two simplifications made the theoretical analysis and statistical inference of population history feasible, reducing the otherwise enormous dimensionality of the problem to a tractable size. The first simplification is provided by coalescent theory. A full description of the genetic history of a set of individuals would unrealistically (and unnecessarily) incorporate the complete genetic state of every ancestor that existed at each generation in the past, and how they relate to one another. Most of these genealogical ancestors do not pass any genetic material to their present-day descendants and are therefore invisible in an individual's pool of genetic ancestors; they can be excluded from any description of history that is concerned with genetic ancestry (A description of history in terms of genealogical ancestors would demand different approaches, such as incorporating records of family relationships (Anderson-Trocmé et al., 2023)). The second simplification is to parametrize models of ancestry by population-level parameters, such as allele frequencies or effective population sizes, rather than the prohibitively high-dimensional genetic state of each ancestral individual. This is reminiscent

of the classical paradigm of statistical mechanics, where physical systems with an extremely large number of degrees of freedom are described in terms of a few measurable macro-states (temperature, pressure, volume) rather than the specific micro-state of each particle (position and velocity). In standard coalescent theory, the relevant parameters or "macro-states" that summarize the ancestral history of samples at a given locus are the mutation rate μ and the effective population size N_e (or more accurately, the product of both). Under the assumption that μ is constant over time for a given genetic locus (which might not be true even on a short timescale (Harris, 2015; Harris & Pritchard, 2017; Mathieson & Reich, 2017)), population history can be summarized by a single, time-varying parameter: $N_e(t)$ (figure 1.1A). Therefore, $N_e(t)$ has been a canonical target of inference for a wide range of methods that collectively fall under the problem of "demographic inference".

As a result of coarse-graining groups of individuals into well-mixed "populations", the problem is reduced from inferring the history of ancestral lineages of each individual to inferring the history of population units. In this realm, two primary paradigms exist: "parametric" methods and "flexible" methods. Parametric methods describe history in terms of a set of pre-defined "demographic events", such as population splits, pulse admixture events, periods of continuous migration, and population expansion or decline. Given some summary statistic derived from observed mutational data (e.g. the joint site-frequency spectrum), the parameters of a pre-defined demography are inferred, such as population-split times, admixture times and proportions, migration rates, or rates of population expansion or decline (Bhaskar et al., 2015; Excoffier et al., 2021; R. Gutenkunst et al., 2010; Hu et al., 2023; Jouganous et al., 2017; J. Kamm et al., 2020; J. A. Kamm et al., 2017; Patterson et al., 2012b; Pickrell & Pritchard, 2012). The set of demographic events allowed varies by method, but a common limitation to all such methods is the computational cost associated with more complex demographic scenarios.

On the other hand, flexible methods are not restricted to a pre-defined demography. Such

methods group individuals into population units, and infer time-varying coalescent rates within and between these groups (H. Li & Durbin, 2011; Schiffels & Durbin, 2014; Schweiger & Durbin, 2023; Sheehan et al., 2013; Steinrücken et al., 2019; Terhorst, 2025; Terhorst et al., 2017). In doing so, they avoid the expensive problem of inferring genome-wide genealogies, and instead consider a more tractable state space described in terms of population-level parameters. Within-population coalescent rates are interpreted as the inverse of an effective population size, while between-population coalescent rates are a proxy for the level of gene exchange, and can be used in simple demographic models to infer actual migration rates between populations (Steinrücken et al., 2019; K. Wang et al., 2020).

In both paradigms, grouping samples into populations is essential for deriving well-defined likelihood functions, and for the interpretation of the estimated parameters in terms of intuitive historical processes. However, genetic ancestry is not necessarily discrete (Rosenberg et al., 2005; Serre & Pääbo, 2004). The categorical depiction of groups of individuals in population-based methods more easily permits a typological understanding of population structure and history that can lead to unintended consequences ranging from mild misinterpretations and model misspecification to the misuse and perpetuation of racist ideology (Coop, 2023; Lewis et al., 2022; Panofsky et al., 2021). Matrix factorization methods like PCA have the advantage of avoiding this assumption altogether (Patterson et al., 2006), but they only resolve a time-averaged coalescent history (G. McVean, 2009), whereas our interest is in time-resolved structure.

Flexible methods of demographic inference mostly build on an important approximation to the coalescent with recombination: the Sequentially Markovian Coalescent (SMC) (Marjoram & Wall, 2006; G. A. T. McVean & Cardin, 2005). PSMC (H. Li & Durbin, 2011) was the first method to allow for the flexible inference of the effective population size $N_e(t)$ from a single whole-genome sequence, and was followed by a suite of subsequent methods of similar flavor (reviewed in Spence et al., 2018). From a single whole-genome sequence,

the time-resolved effective population size $N_e(t)$ is defined as the inverse of the estimated coalescent rate: $N_e(t) = \frac{1}{2\lambda(t)}$. The coalescent rate is the rate at which common ancestors are found for a set of samples. Viewing the coalescent as a continuous-time Markov process, the instantaneous coalescent rate between a pair of samples (i, j) is related to the probability that their ancestral lineages coalesce between generations $[t, t + \Delta_t]$ given that they have not coalesced yet.

$$\lambda_{ij}(t) = \lim_{\Delta t \to 0} \frac{P(T_{ij} \in [t, t + \Delta t] | T_{ij} > t)}{\Delta t} \tag{1.1}$$

From a modeling perspective, $\lambda(t)$ is simply the rate or intensity function of a non-homogeneous Poisson process, where the process counts the number of coalescent events that occur over a set of unlinked loci as a function of time.

Kingman's coalescent assumes that the genomes are sampled from a single panmictic population, where any two individuals are equally likely to breed. Therefore, a single function $N_e(t)$ is adequate to describe the history of that population under this paradigm. The structured coalescent was developed later to extend the standard coalescent to model population structure (Notohara, 1990). In that case, history is modeled in terms of a set of panmictic populations, each with its own size $N_e^{(k)}(t)$, allowing for migration to occur between populations, described by a set of between-population migration rates. This process can be equivalently described by a set of K within-population coalescent rates, $\lambda_k(t)$, and $\binom{K}{2}$ between-population coalescent rates, $\lambda_{pq}(t)$ (figure 1.1B) (Rodríguez et al., 2018). Inferring these rates has been the target of MSMC and similar methods (Schiffels & Durbin, 2014). Even though these parameters are not as clearly interpretable as migration rates, their interpretation as the probability of coalescence conditional on not having coalesced yet, or the rate of a non-homogeneous Poisson process still holds.

The natural extension for describing the coalescent process at an individual resolution (or even a haplotype-level one) is to consider the set of pairwise coalescent rates, $\lambda_{ij}(t)$ between each pair of genomic samples, without assigning these samples a priori to any particular

population unit (figure 1.1C). Since the majority of demographic inference methods attempt to infer directly interpretable parameters such as $N_e(t)$ and migration rates, inferring coalescent rates $\lambda_{ij}(t)$ between pairs of samples has not been a standard end-goal. Despite their appeal in providing an individual-level description of history, interpreting history in terms of $\binom{n}{2}$ pairwise rates is not straightforward, especially for a large sample size n. Simple downstream analysis such as PCA on piecewise-constant pairwise coalescent rate matrices computed over large periods of time has proven useful (Speidel et al., 2021). However, to date, this individual-level description of coalescent history remains under-utilized. In the first part of this dissertation, we introduce a novel method, CoaSTT, short for "Coalescent Structure Through Time", which provides intuitive low-dimensional representations of time-resolved pairwise coalescent rates. Such a representation of structure combines the individual-based depiction of structure provided by methods like PCA and STRUCTURE with the flexible, time-resolved depiction of history provided by methods like PSMC.

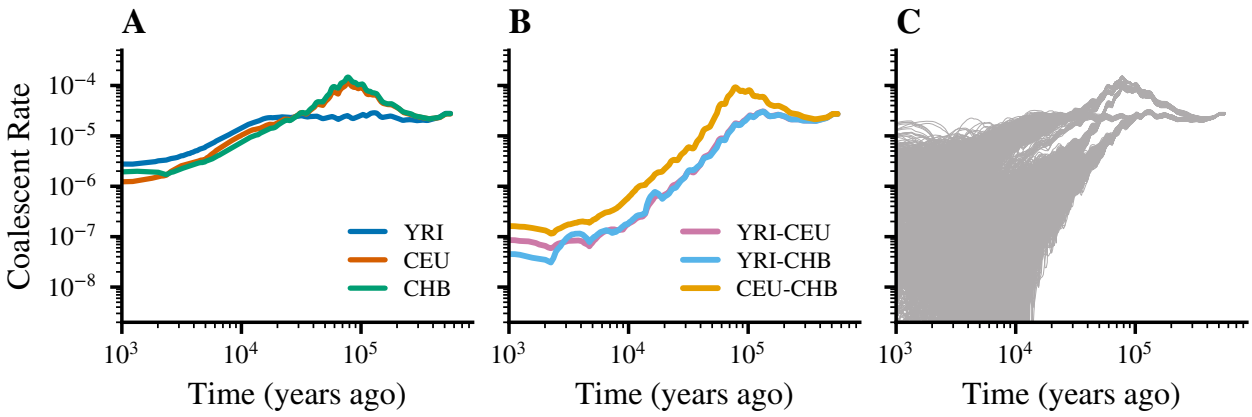

Figure 1.1: Describing history in terms of coalescent rates through time. Pairwise coalescent rates computed as described in section 2.5.1 between 25 individuals assigned labels YRI, CEU and CHB, obtained from ARGs inferred by Wohns et al., 2022. **A)** The mean coalescent rate for pairs of samples assigned the same population label provides a summary of the ancestral history of a single population. The inverse of the mean rate is interpreted as the effective population size through time. **B)** The mean coalescent rate between populations provides a proxy for the degree of gene-flow between a pair of populations at a given point in time. **C)** The coalescent rate between each pair of individuals through time provides a comprehensive summary of coalescent history, but is extremely noisy and difficult to interpret without grouping individuals into population units and computing within-population and between-population means.

Estimates of such coalescent rates between each pair of genomic samples can be obtained from inferred ARGs, which became possible only recently. By the turn of the century, computational tools were mature enough to efficiently simulate ARGs (Hudson, 2002), while accurate, scalable inference of ARGs remained elusive. The primary challenge in ARG inference is that the set of possible ARGs grows extremely rapidly with the number of DNA samples. Additionally, mutational information, which is all that can be observed from sequence data, might be too sparse to allow for reliable inference. Despite the mathematical challenge, the last decade has seen remarkably promising algorithmic innovations in the realm of ARG inference, with a suite of methods being developed, each navigating the tradeoff between accuracy, scalability, and representation of uncertainty (Brandt et al., 2024; Nielsen et al., 2025). While accurate model-based methods were developed over ten years ago, these methods struggled with scalability (Rasmussen et al., 2014). More recent state-of-the-

art methods rely on heuristics and algorithmic innovations that allow inference to scale to thousands of samples (Deng, Nielsen, & Song, 2025; Gunnarsson et al., 2024; Kelleher et al., 2019; Speidel et al., 2019; Wohns et al., 2022; B. C. Zhang et al., 2023).

With ARG inference becoming more readily available in the standard toolkit of a population geneticist, the past few years have seen substantial development of methods that take inferred ARGs as an input to tackle a variety of traditional population genetics questions (Brandt et al., 2024; Nielsen et al., 2025). Summary statistics obtained from ARGs have been utilized to provide a more comprehensive representation of history (C. Fan et al., 2022; Speidel et al., 2019, 2021; Wohns et al., 2022), to resolve complex demographic scenarios involving admixture (Loya et al., 2026; Speidel et al., 2025; Tang & Chiang, 2025; Y. Zhang et al., 2026) and to tackle the ambitious problem of geographically locating genetic ancestors (Grundler et al., 2025; Osmond & Coop, 2024; Wohns et al., 2022). Despite the recent developments, there is still a need for intuitive methods to represent the complex ancestral relationships between samples as they change with time within the ARG, which is what CoaSTT is developed for.

In chapter 2, we introduce CoaSTT as a latent position model (LPM) that provides a low-dimensional representation of coalescent structure through time. In contrast to methods of dimensionality reduction that are based on matrix factorization, time-resolved CoaSTT embeddings are found by minimizing a loss function through gradient descent. We first derive this loss function from the likelihood of observing a number of coalescent events in an ARG given a set of sample latent positions, where distances between samples in the latent space are related to their pairwise coalescent rates through a link function. We describe how smooth latent trajectories for each sample can be obtained by minimizing a regularized version of the loss, then demonstrate how basic demographic processes are revealed intuitively by the embeddings. We then test CoaSTT on more complex demographic scenarios through simulations, highlighting its unique ability to unveil both continuous and discrete structure

in a time-resolved manner. Finally, we provide an extension to CoaSTT that resolves the coalescent history of specific genomic regions rather than providing a genome-wide description. This extension, termed RS-CoaSTT (Region-Specific CoaSTT), has the potential to disentangle alternative coalescent histories of putatively admixed individuals.

We apply CoaSTT to empirical datasets of humans and chimpanzees in chapter 3. The chapter first introduces the biases associated with coalescent rate estimates obtained from inferred ARGs. These biases are invoked throughout the chapter as guidance for how to interpret CoaSTT embeddings appropriately. We infer ARGs for sets of samples from the HGDP dataset (Bergström et al., 2020; Cann et al., 2002), apply CoaSTT to them, and demonstrate its power in reproducing a wide range of known patterns in human population history, while highlighting the additional nuance an individual-based picture provides. We also infer ARGs and apply CoaSTT to a set of chimpanzee samples, again reproducing known signals, while highlighting a few novel signals that the individual-level, coalescent-based description brings to light.

Throughout chapters 2 and 3, we highlight the caveats associated with summarizing the complex relations of pairwise coalescent rates by latent positions in two dimensions. Such low-dimensional summaries are powerful in providing visually intuitive representations of population structure and history, but are unable to faithfully represent all relevant interactions simultaneously. A more fundamental issue regarding the low-dimensional representation of high-dimensional data is that of information limits: how strong must a signal of structure be for it to be recoverable at all?

In chapter 4, we turn back to the more widely used method, PCA, and investigate the conditions under which the observed PCs truly reflect the underlying population structure by employing methods from Random Matrix Theory (RMT). This problem has been touched upon early on for simple demographic scenarios (Bryc et al., 2013; Patterson et al., 2006). We extend the analysis to more general scenarios that result in both discrete and continuous

population structure, with an increased emphasis on the reliability of the observed eigenvectors (or PCs) rather than the eigenvalues, which were the focus of previous studies. We derive and verify several approximate formulas for the sample sizes needed for PCA to faithfully recover underlying population structure for several demographic scenarios, shedding more light on the limits of PCA in resolving population structure.

The work presented in this dissertation provides population geneticists with a novel way of visualizing and describing ancestral history that is rooted in classical population genetic theory. It additionally delves into the nuances associated with low-dimensional representations of genetic data, particularly in the case of PCA, equipping practitioners with a sharper intuition for how to interpret their results.

Declaration of AI usage in writing and research. Drafts of all chapters of this dissertation were written manually, and Claude was used in the final stage of writing of some of the sections to assist in achieving a more clear presentation and for proofreading. Claude code was used in the final stages of result generation to accelerate running multiple workflows in parallel. Assistance was sought from Claude Opus 5.5 and Fable 5.1 to understand theoretical results in Random Matrix Theory specifically pertaining to Appendix 4C, which was written over multiple iterations with assistance from Claude.

CHAPTER 2

A LATENT-SPACE MODEL FOR LOW-DIMENSIONAL REPRESENTATIONS OF COALESCENT STRUCTURE THROUGH TIME (COASTT)

2.1 Introduction

In this chapter, we introduce a method that provides a visually intuitive, individual-based representation of the ancestral history of a set of DNA samples. This is achieved by deriving a low-dimensional representation of the samples' **co**alescent **s**tructure **t**hrough **t**ime, or *CoaSTT* for short. The output of CoaSTT is the position of each sample in some latent space as a function of time. The proximity of any two samples in the latent space at some given time reflects the rate of coalescence between these samples at that time, where closer samples have a higher coalescent rate, and vice versa. The resulting embeddings produce visual summaries of structure through time that are similar to common schematics used by population geneticists to depict population history. Individuals from a single, truly panmictic population will form tube- or branch-like structures whose width represents the size of the population (the coalescent effective size to be precise), while diverged populations that share common ancestry in the past will merge together in this representation. The embeddings thus provide a summary of coalescent structure through time that has intuitive appeal, though, as with other low-dimensional representations of population genetic variation, substantial care and caution are required to avoid misinterpretation (which we address throughout the presentation of our results).

Such a representation draws from the rich literature of latent space models (LSMs) that is pervasive in network science (Hoff et al., 2002; Kaur et al., 2023). Our approach is similar in spirit to methods that model time-varying networks, where latent positions evolve continuously in time (Rastelli & Corneli, 2023a). We first provide an overview of the liter-

ature of latent space models, discussing how this class of models can be invoked to obtain low-dimensional representations of the high-dimensional individual-based coalescent history (section 2.2). We then introduce our method, CoaSTT, deriving the main loss function that is minimized by the method, explaining regularization techniques used for smoothing, and discussing the effect of the choice of a "link function" on the resulting representations (section 2.3). We benchmark CoaSTT against simple demographic scenarios to build a baseline understanding of how simple historical processes are represented by it (section 2.4). We then test it on more complex demographic scenarios through simulations, highlighting its capabilities in providing intuitive representations of complex historical processes in an individual-based manner (section 2.5).

2.2 Theoretical overview: Latent space models in network analysis

In chapter 1, we discussed how coalescent history can be summarized by the coalescent rate between each pair of samples as a function of time. At each generation, the $\binom{n}{2}$ values of instantaneous coalescent rates that summarize the coalescent process can be viewed in terms of parameters that define a network of nodes and edges. Nodes represent genomic samples, and edges represent some random variable parametrized by the pairwise coalescent rate at that generation. This is very similar to the traditional setting of problems in network science, where nodes represent a set of actors, and edges represent some relationship, such as a social relationship (friendship, marriage, etc.) between two actors. Traditional models considered binary relationships, where an edge is either present or absent. Earlier models based on graph theory attempted to model the probability of observing a whole graph as a function of a set of graph statistics (Frank & Strauss, 1986), but these models were highly intractable. Seminal work by Hoff et al., 2002 introduced latent space models (LSMs) or latent position models (LPMs) to model such social networks (reviewed by McLevey et al., 2024). Inspired by the idea that two actors will have a high probability of a relationship if they have similar

characteristics in some latent 'social space', LPMs parametrize the probability of the presence of an edge between two actors as a function of their distance in some Euclidean latent space. Two actors that are close in this space will have a high probability of being related, and two distant actors are most likely not. Specifically, the log-odds of the presence of an edge between actors (i, j) is modeled as

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \alpha + \beta^T x_{ij} - |z_i - z_j| \tag{2.1}$$

where x_{ij} is a vector of covariates (such as age difference or geographic distance) between the actors, and z_i and z_j are the positions of the actors in some d-dimensional latent space. This model naturally incorporates transitivity, the notion that "friends of friends are likely friends". Moreover, it provides a simple, intuitive way to visualize complex networks, and to summarize the $\binom{n}{2}$ parameters of edge presence/absence by nd parameters of positions in latent space. Unlike Multi-Dimensional Scaling (MDS), which aims to find a set of positions that best describe pairwise *dissimilarities*, LPMs provide a model-based representation of observed relationships (i.e. *similarities*) that is parametrized by latent positions and perhaps other covariates. They could be viewed as the continuous counterpart of Stochastic Block Models (SBMs), which model actors in the network as members of "communities", with a single parameter describing within-community or between-community edge probability (Nowicki & Snijders, 2001; Y. J. Wang & Wong, 1987). The so-called Mixed-Membership SBMs are in fact similar in spirit to STRUCTURE-/ADMIXTURE-type models (Airoldi et al., 2008). On the other hand, dynamic SBMs can be compared to PSMC-like methods (Matias & Miele, 2017). LPMs avoid community-assignment (in our case, population assignment) altogether, and project the samples onto a visually intuitive continuum.

LPMs grew in popularity in network science, and were followed by several extensions, including Latent Projection Models that include actor-specific effects and directional interactions (Hoff, 2005; Krivitsky et al., 2009), Latent Position Cluster Models that perform

model-based clustering in addition to inference of positions by placing Gaussian-mixture priors on the positions (Handcock et al., 2007; Krivitsky & Handcock, 2008; Krivitsky et al., 2009), and, most relevant to us, dynamic LPMs, which add a temporal dimension by modeling networks that evolve with time (Rastelli & Corneli, 2023b; Sarkar & Moore, 2005; Sewell & Chen, 2017). It is also worth noting that the inference machinery, either for obtaining point estimates of positions (Hoff et al., 2002; Shortreed et al., 2006) or efficiently sampling from the posterior (Durante & Dunson, 2014, 2016; Ryan et al., 2017; Salter-Townshend & Murphy, 2013; Sewell & Chen, 2015), is well-developed in both the static and dynamic settings, though Bayesian inference in the dynamic case is notorious for being computationally costly, with recent methods being developed to provide solutions for this problem (Loyal, 2025).

Though mostly utilized in social science, LPMs have been used in several biological applications such as ecology (Chiu & Westveld, 2011), protein evolution (Ding et al., 2019), neuroscience (Wilson et al., 2020), and epidemiology (Chu et al., 2021). However, LPMs do not seem to have gained traction in population genetics. Given their potential in providing an individual-specific representation of time-varying structure in a model-based manner, we develop CoaSTT as a latent position model for visualizing coalescent structure. In the next section, we introduce the model, derive the likelihood and the loss function to be minimized, and discuss various aspects of the model such as link function selection and temporal smoothing.

2.3 A latent position model for coalescent structure through time

2.3.1 Method overview: From coalescent rates to latent trajectories through time

The first step in CoaSTT is to derive estimates of time-varying coalescent rates between each pair of samples from inferred ARGs. Details of the estimation of such rates are discussed later in section 2.5.1. Given the input of pairwise coalescent rates through time, embeddings that best reflect these rates are computed as discussed in this section. Figure 2.1 illustrates a simple example of CoaSTT applied to 20 individuals from a single panmictic population that has undergone multiple rounds of historical expansions and bottlenecks according to a 'sawtooth' pattern (Schiffels & Durbin, 2014). Even though noisy estimates of pairwise coalescent rates are used as input in practice, for the sake of illustration here, we use true coalescent rates as input (which are identical for all pairs, and equal to the inverse of the population size over time) (figure 2.1A). In this scenario, the optimal configuration of latent positions at each time point is one where samples are uniformly located within a circle, with the distance between any two given samples increasing/decreasing with the changing coalescent rate. In figure 2.1B, the distance between two representative samples at two different time points is highlighted, showing how the distance increases as the coalescent rate decreases. The latent positions at each time point are stacked over each other, with time as the third dimension, providing a 3-dimensional representation of latent trajectories as a function of time (figure 2.1C). The 3D visualization clearly shows how the radius of the circular cluster changes to reflect $N_e(t)$. Indeed, the average distance over all pairs closely matches $N_e(t)$, even though individual pairs have distances that are scattered around this mean (figure 2.1D). This scatter is an inevitable consequence of trying to fit a set of samples to have identical pairwise distances in two dimensions, and is discussed in more detail in section 2.4.1.

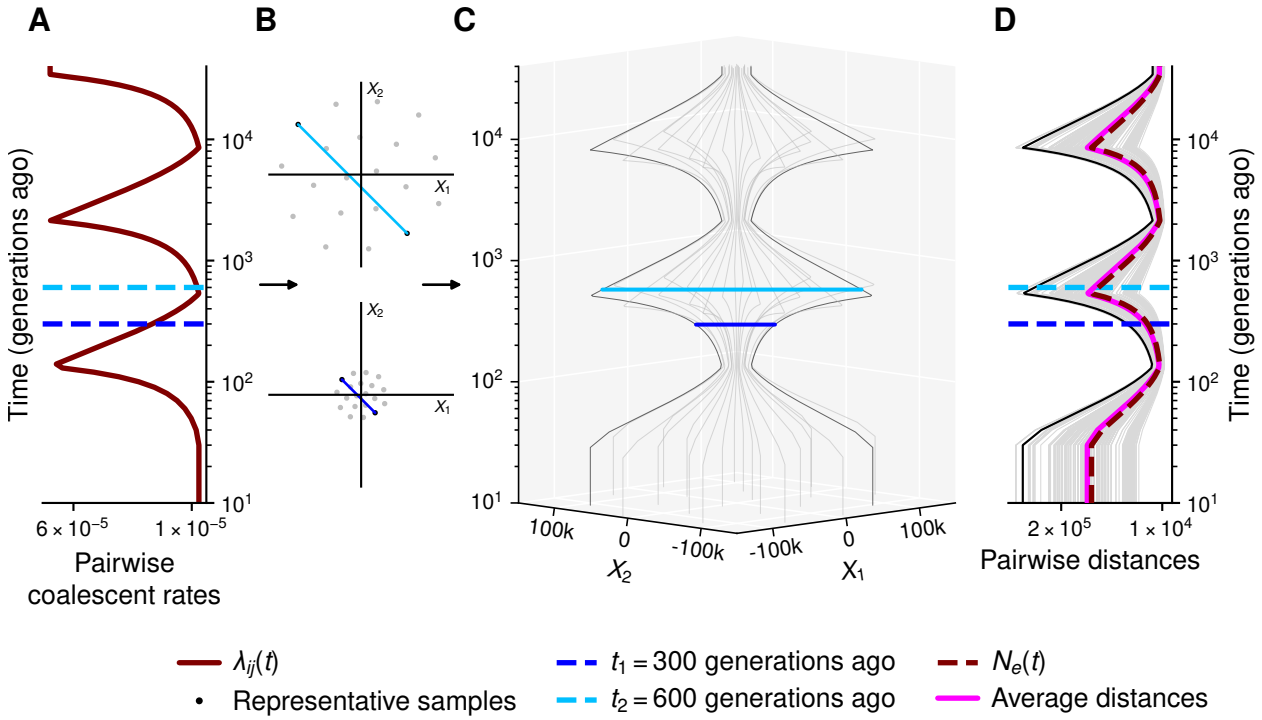

Figure 2.1: Illustrative example of CoaSTT: CoaSTT applied to 20 samples with identical time-varying coalescent rates. **A)** Coalescent rates that are fed to CoaSTT, equivalent to the inverse effective population size, as a function of time. **B)** CoaSTT embeddings shown for two representative time slices, $t = 300$ generations ago (bottom) and $t = 600$ generations ago (top). Dashed lines that connect the same two samples highlight how their distance increases as their coalescent rate decreases. **C)** Latent positions as a function of time, with the lines connecting the two samples at both time points shown. **D)** Pairwise distances between samples' embeddings as a function of time in gray, with the distance between the two focal samples highlighted in black. Average distance over time is plotted with $N_e(t)$ for comparison.

2.3.2 Likelihood of the coalescent latent position model

With an illustrative example that shows how CoaSTT results in latent trajectories that reflect coalescent structure through time, we proceed to derive the likelihood of these latent trajectories given observed counts of coalescent events (figure 2.2). Counts of coalescent events simply indicate the number of local trees where two samples coalesce within a specified time window. These counts could be obtained from an inferred ARG or otherwise. For the derivation, we assume that these counts are obtained from a set of independent local trees

sampled at distant loci from the ARG to derive an exact likelihood. We later describe how we can use the whole ARG as an input without the need to sample independent trees, treating the target function as a pseudo-likelihood rather than an exact one.

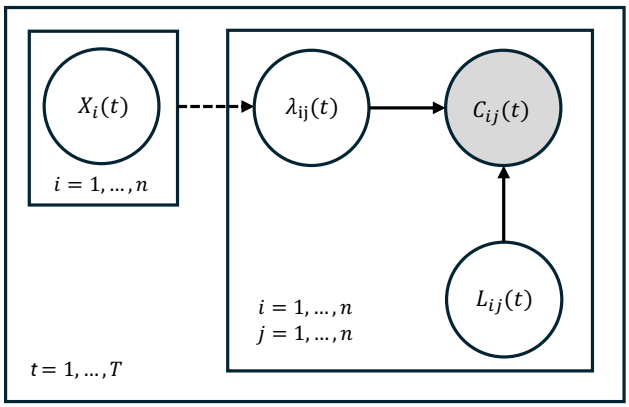

Figure 2.2: CoaSTT graphical model. The set of latent positions of samples determines the coalescent rate between each pair of samples at a given point in time. The dashed line indicates the deterministic relationship in equation 2.2. For a given pair of samples, the number of independent trees where a coalescent event occurs is determined by the pairwise coalescent rate and the number of uncoalesced trees.

We begin by stating the fundamental modeling assumption in the latent position model. We assume that the instantaneous coalescent rate between a pair of haploid samples (i, j) is a function of the distance between them in some latent space

$$\lambda_{ij}(t) = f(d(X_i(t), X_j(t))). \tag{2.2}$$

This is quite a general definition. The distance, which we denote more concisely as $d_{ij}(t) = d(X_i(t), X_j(t))$, can be defined in terms of a general distance metric (Euclidean, hyperbolic, etc.), and the link function $f(.)$ is a general function with no restrictions other than being monotonically decreasing and strictly positive, such that low distances correspond to high rates and vice versa. Another shorthand notation we will be using is $f_{ij}(t) = f(d(X_i(t), X_j(t)))$.

Table 2.1: Variables and notation

Symbol	Description	Shape
n	Number of haploid samples (tips of coalescent trees)	scalar
d	Number of latent space dimensions (default = 2)	scalar
T_{ij}^l	Coalescent time of samples (i, j) at locus l	scalar
$C(t; s)$	Matrix of counts of coalescent events within time interval $[t, t + s]$	(n, n)
$L(t; s)$	Matrix of counts of uncoalesced trees within time interval $[t, t + s]$	(n, n)
$p(t; s)$	Matrix of pairwise coalescence conditional probability within time interval $[t, t + s]$	(n, n)
$\Lambda(t)$	Matrix of pairwise cumulative coalescent rates up until time t	(n, n)
$\lambda(t)$	Matrix of pairwise instantaneous coalescent rates at time t	(n, n)
$X_i(t)$	Latent position of sample i at time t	$(d,)$
$d_{ij}(t)$	distance between samples (i, j) at time t in latent space	scalar
$f_{ij}(t)$	A general monotonically decreasing function of distances between samples (i, j)	scalar

Probabilistic model for independent coalescent trees

Consider a single local coalescent tree l with n haplotype samples at its tips. Define the Bernoulli random variable $Z_{ij}^l(t; s)$ as

$$Z_{ij}^l(t; s) = \begin{cases} 1, & T_{ij}^l \in (t, t + s) \\ 0, & \text{otherwise} \end{cases}$$

where T_{ij}^l is the coalescent time for samples (i, j) in tree l. Let $C_{ij}(t; s)$ be the total count of coalescent events occurring between a pair of samples within the defined time window in the set of independent trees where that pair has not coalesced yet.

$$C_{ij}(t; s) = \sum_{l: T_{ij}^l > t} Z_{ij}^l(t; s). \tag{2.3}$$

Following from the assumption of independence of trees,

$$C_{ij}(t;s) \sim \text{Binomial}(L_{ij}(t), p_{ij}(t;s)), \tag{2.4}$$

where $L_{ij}(t)$ denotes the number of independent trees where the pair has not coalesced yet, and $p_{ij}(t;s)$ denotes the probability of a coalescence event within the window (conditional on not yet coalescing). $C_{ij}(t)$ and $L_{ij}(t)$ are observables, and $p_{ij}(t)$ can be obtained from standard coalescent theory.

For a single pair of samples, the distribution of coalescent times can be generally written as

$$f_{T_{ij}}(t) = \lambda_{ij}(t)e^{-\Lambda_{ij}(t)}, \tag{2.5}$$

where $\Lambda_{ij}(t)$ is the cumulative coalescent rate between samples i and j, defined as

$$\Lambda_{ij}(t) = \int^{t} \lambda_{ij}(s)ds.$$

We defined $p_{ij}(t)$ as the probability of coalescence given that the pair has not coalesced yet. Therefore, it can be written as

$$\begin{aligned} p_{ij}(t;s) &= P(t < T_{ij} < t+s | T_{ij} > t) \\ &= 1 - \frac{P(T_{ij} > t+s)}{P(T_{ij} > t)} \\ &= 1 - \frac{e^{-\Lambda_{ij}(t+s)}}{e^{-\Lambda_{ij}(t)}}. \end{aligned}$$

Assumption 1: s is small compared to the timescale of the change of $\lambda_{ij}(t)$:

$$\begin{aligned}\Lambda_{ij}(t+s) &= \int^{t+s} \lambda_{ij}(s)ds \\ &= \int^{t} \lambda_{ij}(s)ds + \int_{t}^{t+s} \lambda_{ij}(s)ds \\ &\approx \Lambda_{ij}(t) + s\lambda_{ij}(t).\end{aligned}$$

Plugging into the equation for $p_{ij}(t;s)$ we get

$$\begin{aligned}p_{ij}(t;s) &= 1 - \frac{e^{-\Lambda_{ij}(t+s)}}{e^{-\Lambda_{ij}(t)}} \\ &= 1 - \frac{e^{-\Lambda_{ij}(t)}e^{-s\lambda_{ij}(t)}}{e^{-\Lambda_{ij}(t)}} \\ &= 1 - e^{-s\lambda_{ij}(t)}.\end{aligned}$$

We invoke the definition of the link function we introduced in equation 2.2, and rewrite the conditional probability of coalescence as

$$p_{ij}(t;s) = 1 - e^{-sf_{ij}(t)}. \tag{2.6}$$

Assumption 2: Counts of coalescent events between two pairs of samples $(C_{i_1,j_1}(t;s), C_{i_2,j_2}(t;s))$ are independent. This can be achieved by choosing sets of uncoalesced trees with a small number of trees in common between each pair (i_1, j_1) and (i_2, j_2). This can also be phrased in terms of the conditional independence assumption in latent position models defined by Hoff et al., 2002, where the probability of observed counts between a pair of samples is independent of other counts conditional on the positions of that pair in the latent space. Additionally, independence across time points is ensured through the memorylessness of the Markov process. We can therefore write the composite likelihood of observing counts for all pairs over a set of time points by taking the product of the binomial likelihood for each pair

at each time point

$$\mathcal{L}(X) = \prod_{i,j,t} \binom{L_{ij}(t)}{C_{ij}(t;s)} (1 - e^{-sf_{ij}(t)})^{C_{ij}(t;s)} (e^{-sf_{ij}(t)})^{(L_{ij} - C_{ij}(t;s))}. \tag{2.7}$$

In this likelihood, $C_{ij}(t;s)$ are observables that can be obtained from the inferred ARG, and s is a choice of resolution for discretizing time. X is a rank-3 tensor of shape (n, d, T) denoting the positions of each of the n samples in the d-dimensional latent space across T time points. Dependence on X comes from $f_{ij}(t) = f(d(X_i(t), X_j(t)))$. Rearranging and ignoring constants with respect to X, we obtain a negative log-likelihood which will serve as a loss function for minimization

$$l(X) = \sum_{i,j,t} (L_{ij}(t) - C_{ij}(t;s)) s f_{ij}(t) - \sum_{i,j,t} C_{ij}(t;s) \ln(1 - e^{-sf_{ij}(t)}). \tag{2.8}$$

Though we can minimize this loss function directly, we have found in practice that the log term results in a more difficult optimization problem. Our problem can also be rephrased in terms of a Poisson likelihood rather than a binomial, yielding a simpler loss function.

Poisson Approximation for independent trees

Assumption 3: The probability of a coalescent event within the time window s is small, and the number of uncoalesced trees for the times considered is large.

$$\lambda_{ij}(t)s \ll 1$$

$$L_{ij}(t) \gg 1$$

The first condition is guaranteed if a small enough time window is chosen, or if the pairwise coalescent rate is very small, which is usually the case. The second condition holds except

for the very deep past where $L_{ij}(t)$ is expected to become small. With these conditions,

$$C_{ij}(t;s) \sim \text{Poisson}(sL_{ij}(t)f_{ij}(t)), \tag{2.9}$$

where we approximated the probability as

$$p_{ij}(t;s) = 1 - e^{-sf_{ij}(t)} \approx sf_{ij}(t).$$

We rewrite equation (2.7) in terms of the Poisson likelihood rather than the binomial likelihood

$$\begin{aligned}
\mathcal{L}(X) &= \prod_{i,j,t} \frac{(sL_{ij}(t)f(d_{ij}(t)))^{C_{ij}(t;s)} e^{-sL_{ij}(t)f_{ij}(t)}}{C_{ij}(t;s)!} \\
&\propto \prod_{i,j,t} f_{ij}(t)^{C_{ij}(t)} e^{-sL_{ij}(t)f_{ij}(t)} \\
&\propto \prod_{i,j,t} \left[e^{\left(\frac{C_{ij}(t)}{sL_{ij}(t)}\right) \ln f_{ij}(t) - f_{ij}(t)} \right]^{sL_{ij}(t)}
\end{aligned} \tag{2.10}$$

$$l(X) = \sum_{i,j,t} sL_{ij}(t) \left[f_{ij}(t) - \frac{C_{ij}(t)}{sL_{ij}(t)} \ln f_{ij}(t) \right]. \tag{2.11}$$

$l(X)$ is again a loss function based on taking the negative log-likelihood and dropping terms constant with respect to X. In our preliminary implementations, we found that minimizing the Poisson-based loss function is more numerically robust than the binomial-based one. We proceed with it, keeping in mind that the binomial is the more exact analysis.

Pseudo-likelihood-based general loss function

Note that $\hat{\lambda}_{ij}(t) = \left[\frac{C_{ij}(t;s)}{sL_{ij}(t)}\right]$ is the MLE of $\lambda_{ij}(t)$ in the case of independent trees (and small s). We can therefore rewrite equation 2.11 as:

$$l(X) = \sum_{i,j,t} w_{ij}(t) \left[f(d(X_i(t), X_j(t))) - \hat{\lambda}_{ij}(t) \ln f(d(X_i(t), X_j(t))) \right], \tag{2.12}$$

where we have made the dependence on X more explicit and rewritten the factors $sL_{ij}(t)$ as weights $w_{ij}(t)$. The input to the loss function is now the MLEs of instantaneous pairwise coalescent rates for all pairs of samples over all time points, along with a set of weights. These weights specify the information carried by the rate estimate for each pair at each time point.

If the MLE is used for $\hat{\lambda}_{ij}(t)$, then $w_{ij}(t) = sL_{ij}(t)$. However, we can use other estimators for $\hat{\lambda}_{ij}(t)$ as long as we substitute $w_{ij}(t)$ with the appropriate weight. If another estimate of $\hat{\lambda}_{ij}(t)$ is used, we substitute $w_{ij}(t)$ with a factor that results from matching the variance of that estimate with the variance of the MLE. With the MLE defined as $\hat{\lambda}_{ij}(t) = \left[\frac{C_{ij}(t;s)}{sL_{ij}(t)}\right]$, the weight can be rearranged and written as

$$w_{ij}(t) = sL_{ij}(t) = \frac{\lambda_{ij}(t)}{\mathrm{Var}(\hat{\lambda}_{ij}^{\mathrm{MLE}}(t))}. \tag{2.13}$$

For an otherwise general estimate, we define the effective weight as

$$\hat{w}_{ij}(t) = \frac{\hat{\lambda}_{ij}(t)}{\mathrm{Var}(\hat{\lambda}_{ij}(t))}, \tag{2.14}$$

where we substituted the true rate in the numerator with the estimated rate, and used the variance of the general estimate in the denominator. We can therefore define a loss that is based on a pseudo-likelihood, where the MLE of the coalescent rates is substituted with

another estimator of the coalescent rate.

$$\boxed{l(X) = \sum_{i,j,t} \hat{w}_{ij}(t) \left[f_{ij}(t) - \hat{\lambda}_{ij}(t) \ln f_{ij}(t) \right]} \tag{2.15}$$

$\hat{w}_{ij}(t)$ is defined in equation 2.14 as the ratio of the rate estimate to the variance of the estimate, and $\hat{\lambda}_{ij}$ is a general estimate that is not necessarily the MLE.

Minimizing the loss in equation 2.15 is equivalent to minimizing the weighted KL divergence between the distribution of counts of coalescent events induced by the latent positions with mean $f_{ij}(t) = f(d(X_i(t), X_j(t)))$, and the distribution of counts resulting from a rate of $\hat{\lambda}_{ij}(t)$, over all pairs and all time points. To see this more clearly, we write the weighted KL divergence explicitly as

$$\begin{aligned} \hat{w}_{ij} D_{\mathrm{KL}}(\mathrm{Pois}(\hat{\lambda}_{ij}(t)) || \mathrm{Pois}(f_{ij}(t))) &= \hat{w}_{ij} \left[f_{ij}(t) - \hat{\lambda}_{ij}(t) + \hat{\lambda}_{ij}(t) \ln \frac{\hat{\lambda}_{ij}(t)}{f_{ij}(t)} \right] \\ &= \text{constant} + \hat{w}_{ij} \left[f_{ij}(t) - \hat{\lambda}_{ij}(t) \ln f_{ij}(t) \right]. \end{aligned} \tag{2.16}$$

Minimizing the sum of the weighted KL divergences in equation 2.16 is equivalent to minimizing the loss defined in equation 2.15.

The loss function in equation 2.10 was derived under the assumption that we can obtain a set of truly independent trees for each pair of samples and each time point. In reality, what we have is a sequence of trees over the genome. It is possible to sample trees from very distant loci to obtain approximately independent trees, in which case one can directly use equation 2.10. However, the degree of thinning will have to vary for different pairs as it depends on the population-scaled recombination rate $\rho = 4N_e r$ (Terhorst, 2026). Pairs with a high coalescent rate and low N_e will have more frequent recent coalescence events which persist for a longer genomic span, and additionally have a lower ρ. The pseudo-likelihood-based loss in equation 2.15 allows us to utilize more general estimates of coalescent rates

that do not depend on sampling independent trees, as long as the variance of these estimates can be obtained.

The problem of finding latent positions (embeddings) of samples through time has been recast into an optimization problem of minimizing a loss function with respect to the latent positions. The number of parameters is $(n \times d \times T)$ (n positions in d dimensions for T time points), with an input of $\frac{n(n-1)}{2} \times T$ estimates of coalescent rates and their variances. A standard setting would have $n \sim 100$ samples, $d = 2$ dimensions and $T \sim 500$ time points, resulting in $\approx 10^5$ parameters to be estimated given $\approx 2.5 \times 10^6$ coalescent rate estimates. This is a fairly large number of parameters to estimate, but within the reach of modern gradient-based optimization methods. Nevertheless, it is still lower than the full number of input parameters $\left(\frac{n(n-1)}{2} \times T\right)$. In addition to providing an intuitive visualization of coalescent history, CoaSTT can also be seen as a principled method for reducing the number of parameters that represent the coalescent process without assuming population units.

2.3.3 Computing latent embeddings through time

The loss presented in equation 2.15 has no terms that couple the latent positions of a single sample at different time points. Without a term that encourages the latent positions of a sample at nearby time points to be similar, the optimal configuration at each time slice overfits to the noisy rate estimates of that slice, yielding jagged latent trajectories for each sample through time. In the next section, we describe a regularization procedure that addresses this overfitting and smooths the latent trajectories over time. However, minimizing the full regularized loss without a well-calibrated initial solution is computationally expensive and does not reliably converge to a good configuration. We therefore divide the optimization into two phases. In the first phase, we find the optimal configuration one time point at a time, without the regularization term, in a process we call "sequential optimization". The solver at each time point is initialized with the optimal configuration of the previous time

point. We start by finding the optimal configuration for the deepest time point T using a random initialization, and center it at the origin. We use this configuration to initialize the solver for $t = T - 1$, then align the solution for that time point to the previous one by a rigid rotation (Procrustes alignment). We proceed iteratively until the most recent time point is optimized. In the second phase, the full set of aligned, unregularized embeddings over time is used as the initial solution to the regularized optimization problem. In the remainder of this section, we describe how the optimal configuration at a single time point is obtained given the previous time point's configuration as an initialization. The regularization procedure and the solution of the joint problem are described in the next section.

At a given time slice, both the gradient and the Hessian of the loss in equation 2.15 can be computed exactly and efficiently, at least for the sample sizes at which ARG inference is currently feasible. Let $x \in \mathbb{R}^{nd}$ denote the latent positions X at a single time slice flattened into a vector, with $g \in \mathbb{R}^{nd}$ the corresponding gradient and $H \in \mathbb{R}^{nd \times nd}$ the Hessian. In Appendix 2A, we show that these can be written as

$$g = \mathrm{vec}(\nabla_X l) = \mathrm{vec}(2L_g X) \tag{2.17}$$

$$H = \nabla_x^2 l = 2[L_g \otimes I_d + R], \tag{2.18}$$

where L_g is a graph Laplacian and R a block matrix, both of which are cheap to assemble from pairwise distances. A Newton step is defined as $p = -H^{-1}g$, and is the minimizer of the local quadratic model of the loss. However, a Newton step is guaranteed to be a descent direction only if H is positive definite, which is not the case here. Away from the minimum, the repulsive term of the loss contributes negative curvature, and H has many negative eigenvalues. Moreover, the Hessian is badly conditioned under the inverse-power link, with condition numbers of 10^6–10^9 that grow toward the present when populations expand and latent distances are large. Quasi-Newton methods such as L-BFGS, which approximate the inverse Hessian from a small number of recent gradient differences, cannot represent such a

spectrum, and we have found them to occasionally stall well short of convergence.

We therefore use a modified Newton method with line search (Nocedal & Wright, 2006). At iteration k, we replace the exact Hessian H_k by a positive-definite modification $\tilde{H}_k$ (defined below), and take a damped step

$$p_k = -\tilde{H}_k^{-1} g_k, \tag{2.19}$$

$$x_{k+1} = x_k + \alpha_k p_k, \tag{2.20}$$

where the step length α_k satisfies the strong Wolfe conditions (Wolfe, 1969) and is found using the bracketing and zoom line-search algorithm of Nocedal and Wright, 2006 (Algorithms 3.5 and 3.6). Near the minimum, the full step $\alpha_k = 1$ is accepted and the method converges quadratically. Far from it, the line search shortens the step and guarantees a decrease in the loss at every iteration. Because the curvature of the loss depends on X through the link function and changes rapidly as the configuration evolves, the Hessian and its modification are recomputed at every iteration rather than being reused across iterations.

To obtain the positive-definite modified Hessian $\tilde{H}_k$, we compute an eigendecomposition of H_k, and flip the signs of the negative eigenvalues (Greenstadt, 1967). Additionally, we raise all eigenvalues (which are now positive) below a minimum value $\lambda_{\min}$ to that value, such that $\dfrac{\lambda_{\min}}{\lambda_{\max}} = \delta$, clamping the condition number to a maximum of $\dfrac{1}{\delta}$. The Newton direction is then computed as

$$p_k = -V_k \operatorname{diag}\left(\frac{1}{\tilde{\lambda}_{k,j}}\right) V_k^\top g_k, \tag{2.21}$$

with the eigenvectors V_k and modified eigenvalues $\tilde{\lambda}_{k,j}$ given by

$$H_k = V_k \operatorname{diag}\left(\lambda_{k,j}\right) V_k^\top, \tag{2.22}$$

$$\tilde{\lambda}_{k,j} = \max\left(|\lambda_{k,j}|,\ \delta \max_{j'} |\lambda_{k,j'}|\right). \tag{2.23}$$

Convergence is set by the relative change in the loss being lower than some tolerance ε. Default values for δ and ε are 10^{-6}.

2.3.4 Regularization for smoothing latent space trajectories

Even though our sequential optimization procedure yields aligned configurations, the trajectories of individual samples might exhibit erratic fluctuations due to the noise in coalescent rate estimates. In the literature of dynamic LPMs, one solution is to place a Random Walk prior or a Gaussian Process prior on latent trajectories (Durante & Dunson, 2014; Sewell & Chen, 2015). However, such models do not scale well with the number of time points, and stationary GPs additionally risk over-smoothing the trajectories in periods of sharp transitions such as population splits. A more involved method based on nested Gaussian processes has been proposed (Durante & Dunson, 2016), though such a method is based on a latent projection model (where the edge probabilities are a function of the inner product of latent vectors rather than distances in latent space) rather than a latent position model, and the convenient linearity of such a model is what allows for the computational speedup. We develop an alternative regularization approach that is flexible, scales well with the number of samples and time points, and does not over-smooth the trajectories when there is a strong signal.

We draw inspiration from the trend-filtering literature to apply an L_1 regularization term to finite differences of the latent trajectories in the loss function (S.-J. Kim et al., 2009; Rudin et al., 1992; Tibshirani, 2014). In simple terms, the so-called "fused-lasso" adds an L_1 penalty

on absolute differences between adjacent parameters over time or space. Trend-filtering generalizes this to higher-order finite differences (discrete-time higher order derivatives). If the kth-order absolute difference is penalized (corresponding to the kth derivative), solutions are encouraged to be piecewise polynomials of order $k-1$. Because the penalty is linear in the size of each difference, a sharp transition at a single time point is not penalized more than a gradual one of the same total magnitude, so periods of rapid change such as population splits are preserved rather than smeared out, in contrast to quadratic penalties. In the Bayesian sense, this is equivalent to adding a Laplace prior (double-exponential) on the differences rather than a Gaussian prior that would result in an L_2 penalty (Casella et al., 2010). Such a prior has been used to model dynamic networks, but not in an LPM setting (Betancourt et al., 2017).

An exact fused-lasso term for first-order differences has the form

$$\sum_{i,m} \sum_{p=1}^{d} |X_{ip}(t_{m+1}) - X_{ip}(t_m)|.$$

This independently penalizes a changing trajectory in each dimension, and thus breaks the rotational invariance of the likelihood. The order of dimensions in our LPM has no specific meaning. A trajectory that stays constant on the first dimension but varies on the second is not different from one where the opposite is true. What is more appropriate for our case is to penalize change in the whole d-dimensional vector $X_i(t)$. In that case, $X_i(t)$ will remain constant between two time points in *all* dimensions if there is not enough evidence that its coalescent rate changed with respect to all other samples. The appropriate choice in this case is "group fused-lasso" (Bleakley & Vert, 2011; Yuan & Lin, 2006). The corresponding

penalty term for first-order differences has the form

$$\sum_{i,m} \sqrt{\sum_{p=1}^{d} (X_{ip}(t_{m+1}) - X_{ip}(t_m))^2} = \sum_{i,m} \|X_i(t_{m+1}) - X_i(t_m)\|_2 = \sum_{i,m} \|(D_1 X_i)_m\|_2,$$

where D_1 is a $(T-1)\times T$ matrix for the first-order difference operator, X_i is the $T \times d$ matrix of the trajectory of sample i, and $(D_1 X_i)_m$ denotes the mth row of $D_1 X_i$. For k^{th}-order differences, we generally have a $(T-k)\times T$ difference operator D_k, obtained by composing k first-order difference operators and dividing by Δ_t^k, where Δ_t is the spacing of the time grid, so that $D_k X_i$ approximates the kth time derivative of the trajectory and the penalty strength is independent of the chosen time resolution. Finally, since we use L-BFGS for the joint regularized problem, we replace the group norm by its pseudo-Huber approximation $\sqrt{\|(D_k X_i)_m\|_2^2 + \varepsilon_k^2}$, which coincides with the L_1 penalty for $\|(D_k X_i)_m\|_2 \gg \varepsilon_k$ and is quadratic (and hence differentiable) within a small neighborhood of radius ε_k around the non-differentiable point. The full regularized loss is now given by

$$\boxed{l(X) = \sum_{i,j,m} \hat{w}_{ij}(t_m) \left[f_{ij}(t_m) - \hat{\lambda}_{ij}(t_m) \ln f_{ij}(t_m) \right] + \sum_{i,k} \gamma_k \sum_m \sqrt{\|(D_k X_i)_m\|_2^2 + \varepsilon_k^2}} \tag{2.24}$$

Equation 2.24 is written in a quite general sense where multiple values of k penalizing the kth-order difference each with its own penalty strength can be used. By default, we fix the penalty to 2nd-order differences (in addition to a tiny term penalizing 1st-order differences), though the current implementation allows for a more general choice of parameters. First-order differences would yield piecewise constant trajectories, turning the continuous drift of latent positions expected under gradual processes (such as clusters steadily approaching one another during a period of migration, or expanding as population size grows) into a staircase, whereas second-order differences instead yield piecewise linear trajectories. γ_k is

chosen such that the L_1-penalty term (evaluated at an unsmoothed solution $\tilde{X}$) contributes a user-specified fraction α_k of the saturated loss, defined as the data term evaluated at a perfect fit $f_{ij}(t) = \hat{\lambda}_{ij}(t)$ (this is the minimum value the data term can attain and depends only on the input rates and weights). To find an appropriate γ_k and ε_k, the loss is first minimized without the penalty to find unsmoothed trajectories $\tilde{X}_i(t)$ (through sequential optimization). ε_k is set to a small fraction η (default 10^{-3}) of the median of the group norms of the unsmoothed trajectories, so that the penalty is effectively an L_1 penalty on all but a negligible fraction of the differences. For a specified fraction α_k, the penalty strength is then chosen to be

$$\varepsilon_k = \eta \operatorname*{median}_{i,m} \|(D_k \tilde{X}_i)_m\|_2 \tag{2.25}$$

$$\gamma_k = \alpha_k \frac{\sum_{i,j,m} \hat{w}_{ij}(t_m) \hat{\lambda}_{ij}(t_m) \left[1 - \ln \hat{\lambda}_{ij}(t_m)\right]}{\sum_{i,m} \sqrt{\|(D_k \tilde{X}_i)_m\|_2^2 + \varepsilon_k^2}}. \tag{2.26}$$

After obtaining aligned configurations by sequential optimization, we minimize the full loss with the regularization term to obtain smooth latent trajectories. We do so using `L-BFGS` (Byrd et al., 1995; Virtanen et al., 2020; Zhu et al., 1997), with the gradient for the regularization term given by

$$\frac{\partial l_{\text{reg}}}{\partial X_i} = \sum_k \gamma_k D_k^\top \Sigma_i^{(k)} D_k X_i, \qquad \Sigma_i^{(k)} = \operatorname{diag}\left(\frac{1}{\sqrt{\|(D_k X_i)_m\|_2^2 + \varepsilon_k^2}} \right)_{m=1}^{T-k}, \tag{2.27}$$

where $\Sigma_i^{(k)}$ normalizes each kth difference by its own smoothed norm. As in the sequential optimization, convergence is set by the relative change in the loss being lower than a tolerance ε_{tol} (default 10^{-6}).

2.3.5 Characterization of different choices of the link function

So far, we have kept the function that links distances in the latent space to pairwise coalescent rates general. In the traditional LPM literature, a logistic function is used, which is a natural choice given that the parameter it models is the pairwise edge probability. This results in log-odds of edge probabilities that are simply linear in distance (equation 2.1). Further work demonstrated the analytical tractability of studying properties of LPMs when a Gaussian link is used instead (Rastelli et al., 2016). For LPMs where distances correspond to a Poisson rate, similar to our case, a Gaussian link has previously been used (Rastelli & Corneli, 2023b). However, to our knowledge, there is no in-depth analysis of different choices of link functions in LPMs.

In contrast, the literature of non-linear embeddings, especially "Neighbor-Embedding" (NE) methods, has explored the effect of different choices of the link function, or the kernel, in great depth. In Appendix 2A, we show how all NE methods and LPM methods, including CoaSTT, end up minimizing a loss with the same "Attractive-Repulsive" structure (Böhm et al., 2022), given some choice of a "link function" or a kernel that relates distances between latent embeddings to the target similarity. A prominent member of this family, t-SNE (Maaten & Hinton, 2008), was in fact an extension of an earlier method, SNE (Hinton & Roweis, 2002), where one of the main differences was the choice of the link function (or kernel, as it is called in that literature). In classical Stochastic Neighbor Embeddings (SNE), a procedure to construct a similarity matrix from high-dimensional data is described, and the target is to find lower-dimensional embeddings that yield a similarity matrix that best matches the target similarity (based on KL divergence) when passed through a Gaussian link. However, a "crowding problem" was documented and attributed to the choice of the Gaussian link. Samples with high similarity were fitted well (local structure), while clusters with moderate similarity were drawn too close together and to the center, with such clusters occasionally blending into one another. Using an inverse-square link function (t-distribution)

characterized by a heavier tail seemed to alleviate this problem. A further extension, Heavy-Tailed SNE, proposed a more general inverse-power link to control the tail heaviness (Yang et al., 2009). Later, UMAP was developed (McInnes et al., 2020), which proposed a different way of constructing the similarity matrix and a different loss that has similarities to ours (Appendix 2A), and used a general Cauchy distribution as the link. UMAP was characterized by its ability to better differentiate clusters (global structure), exhibiting "less crowding" and more "white space" between them in the embedding space, even though a link similar to t-SNE's was used. Careful analysis showed that the main differences between UMAP and t-SNE embeddings were not attributable to the specifics of how the similarity matrix is defined, but rather to the optimization procedure (which effectively controlled the contribution of the attraction term in the loss), and to the difference in the repulsive term of the loss (Draganov et al., 2023). Appendix 2A describes where CoaSTT falls on that spectrum.

The crowding problem and tail-heaviness

Understanding the crowding problem and the balance between the attractive and repulsive terms of the loss sheds light on the effect of different link functions and motivates our final choice. First, we point out that our loss function is characterized by a repulsive term $\sum_{i,j} \hat{w}_{ij} f_{ij}$ that pushes all pairs of samples apart, and an attractive term $-\sum_{i,j} \hat{w}_{ij} \hat{\lambda}_{ij} \ln f_{ij}$ that pulls samples with higher coalescent rates closer. As a reminder, $f_{ij} = f(d(X_i, X_j))$ is shorthand notation for the link function. The heaviness of the tail of the link function controls two main things. First, it defines the effective range of distances that the loss is sensitive to. For a Gaussian that decays too quickly, the repulsive term becomes negligible once the samples move apart by just a moderate distance, whereas with the Cauchy link, the repulsive term remains active for more substantial distances. Second, the curvature of the link defines the range of distances corresponding to a particular range of similarities. Pairs of samples with moderate similarity (similar coalescent rates in our case) will occupy a small

range of distances under the Gaussian link and a larger range with a Cauchy link. In 2D, it is impossible to fit many points with the same distance or very similar distances, as apparent in the embeddings shown in figure 2.1. A link with a heavier tail alleviates this problem by allowing the moderately close similarities to correspond to a wider range of distances that are more realizable in 2D. It additionally keeps the repulsive force active for a larger range of distances, allowing the "weakly similar" clusters to be better differentiated (figure 2.3).

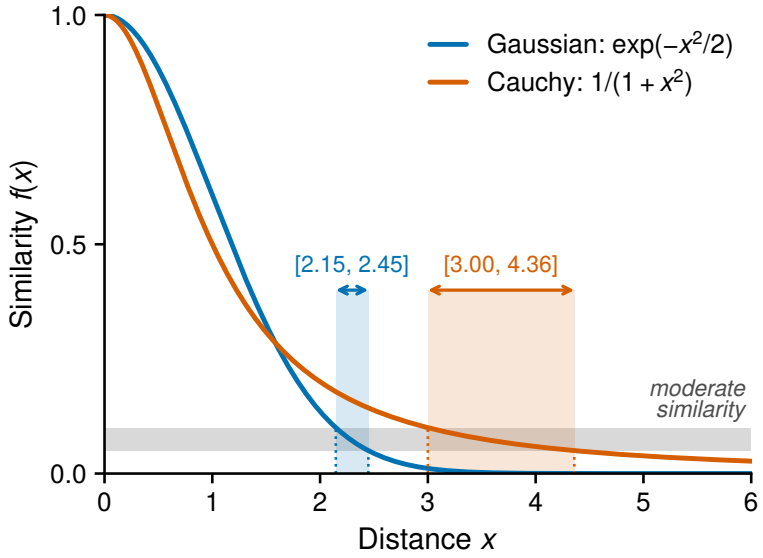

Figure 2.3: Gaussian vs. Cauchy link functions. For a set of pairs with similarities lying within the range $[0.05, 0.1]$, the corresponding distances in the latent space span a much wider range if a Cauchy link function is used compared to a Gaussian link function.

Theoretically, the heaviness of the tail of the link function fully specifies how distinct clusters appear in the latent space, with heavier tails allowing distinct clusters to be better distinguished, and heavier tails are therefore better for resolving fine structure. However, tails that are too heavy were shown to exhibit vanishing gradients, difficulty in convergence, high dependence on initialization and a higher affinity to fragment clusters too much, resulting in spurious discrete clusters with continuous data (Kobak et al., 2020). Therefore, using link functions with heavier and heavier tails in hopes of revealing structure at finer and finer scales has a limit. Moreover, careful analysis in the literature has shown that the

optimization method plays a big role as well in differentiating clusters. For example, with all else held constant, t-SNE and UMAP are shown to reveal global structure equally well given the same initialization (Kobak & Linderman, 2021). Additionally, the optimization algorithm itself affects the attractive and repulsive forces in the loss, with t-SNE's "early exaggeration" increasing the effective attraction (Böhm et al., 2022) and UMAP's negative sampling reducing the effective repulsion (Damrich & Hamprecht, 2021). Both algorithms result in a better resolution of global structure.

Defining CoaSTT's default link function

CoaSTT allows a general user-defined link function. In the next section, we show the effect of a set of link functions on the ability to differentiate distinct population clusters and resolve changes in population size. We argue that an inverse-power link (close to the Cauchy link used by UMAP) with a general power p provides simpler interpretations of cluster sizes and the flexibility of controlling the tail-heaviness through p. CoaSTT uses $p = 1$ by default, since it results in population clusters with a radius directly proportional to N_e. Specifically, we use

$$f_{ij}(t) = \frac{1}{N_{\min} + \|X_i(t) - X_j(t)\|_2^p} - \frac{1}{N_{\max}}, \tag{2.28}$$

with $p = 1$ by default. $N_{\min}$ sets the minimum resolution for differentiating two samples, such that two samples with $\lambda_{ij}(t) > \frac{1}{N_{\min}}$ will optimally be placed with zero distance between them in the latent space. $N_{\max}$ sets a global upper bound on separation, where two samples with $\lambda_{ij}(t) = 0$ will be separated by a maximum distance of $N_{\max}$. Since subtracting $\frac{1}{N_{\max}}$ can result in a negative $f_{ij}(t)$, in practice we add a constant $\frac{1}{N_{\max}}$ to all of the rate estimates $\hat{\lambda}_{ij}(t)$ instead, setting the lower bound for the input rates to $\frac{1}{N_{\max}}$, and therefore resulting in $N_{\max}$ as an upper bound on the distances. Both $N_{\min}$ and $N_{\max}$ are user-specified, and

set to the 0.01 and 0.99 percentiles (1st and 99th) of non-zero values of the complete set of input rates $\frac{1}{\hat{\lambda}_{ij}(t)}$ if not otherwise specified.

2.4 Properties of CoaSTT under simple demographic scenarios

Before diving into examples with complex population histories, we want to demonstrate the basic properties of CoaSTT embeddings under an input of coalescent rates that are obtained from simple demographic scenarios. To that end, we apply CoaSTT to coalescent rates obtained from a single panmictic population as well as multiple structured populations with and without gene flow. This sets a benchmark for how to interpret embeddings inferred for more complex datasets. Since the primary purpose is benchmarking, we use true coalescent rates as inputs for all examples in this section to remove the noise element associated with estimating the rates from ARGs. In doing so, we use weights $\hat{w}_{ij} = 10^6$ for all terms in the loss (equation 2.15).

2.4.1 Samples from a panmictic population form circular clusters with cluster size corresponding to population size

In figure 2.1B, we showed how samples obtained from a single panmictic population form circular configurations with cluster size increasing as pairwise coalescent rates decrease. Figure 2.4 shows that this generally holds under different choices of the link function, except for a few edge cases where it starts to break. We test two representative families of link functions, the inverse-power link with a general degree p

$$f(x) = (N_{\min} + x^p)^{-1} - \frac{1}{N_{\max}} \tag{2.29}$$

and the exponential link with a general degree p

$$f(x) = \frac{1}{N_{\min}} e^{-x^p} - \frac{1}{N_{\max}}. \tag{2.30}$$

To understand why circular configurations form, consider the case where $\lambda_{ij} = \frac{1}{N_e}$ is the same between all pairs of samples. The ideal latent space representation is then a geometric shape where all samples are equidistant from one another, with the characteristic distance given by $x^* = f^{-1}(\lambda_{ij})$. In 2 dimensions, no such configuration exists for $n > 3$ samples. However, if we consider multiple solutions of the optimization problem given random initializations, we can ask what configuration gives, on average, a constant distance between each pair of points. The answer is a circle (in 2 dimensions, and a hyper-sphere in K dimensions) with a filled interior, and indeed, n samples from a panmictic population will correspond to n points that are uniformly distributed within a disk.

Figure 2.4 shows that the circular configuration emerges under a variety of link functions except those that are either very slowly decaying (inverse power with $p = 0.1$) or very quickly decaying (exponential with $p = 4$). We discussed in section 2.3.5 how link functions with very heavy tails might not have enough curvature for robust convergence. For the rapidly decaying link, the sample embeddings are not uniformly distributed within the circle, most likely due to the very short effective range of the attractive and repulsive forces that are not strong enough to rearrange the samples within the circle starting from a random configuration.

We also demonstrated in figure 2.1D how the average pairwise distance corresponds to population size. Figure 2.5 shows this more precisely, illustrating how the relationship between average pairwise distances and population size changes with different choices of the link function. The inverse-power link has the very convenient property that the average pairwise distance is almost identical to population size (up to a small additive constant), since the characteristic distance between each pair is $x^* = N_e - N_{\min}$. The average pairwise distance between two points uniformly sampled from a circle in 2D is $\frac{128}{45\pi} r \approx 0.9054r$, which

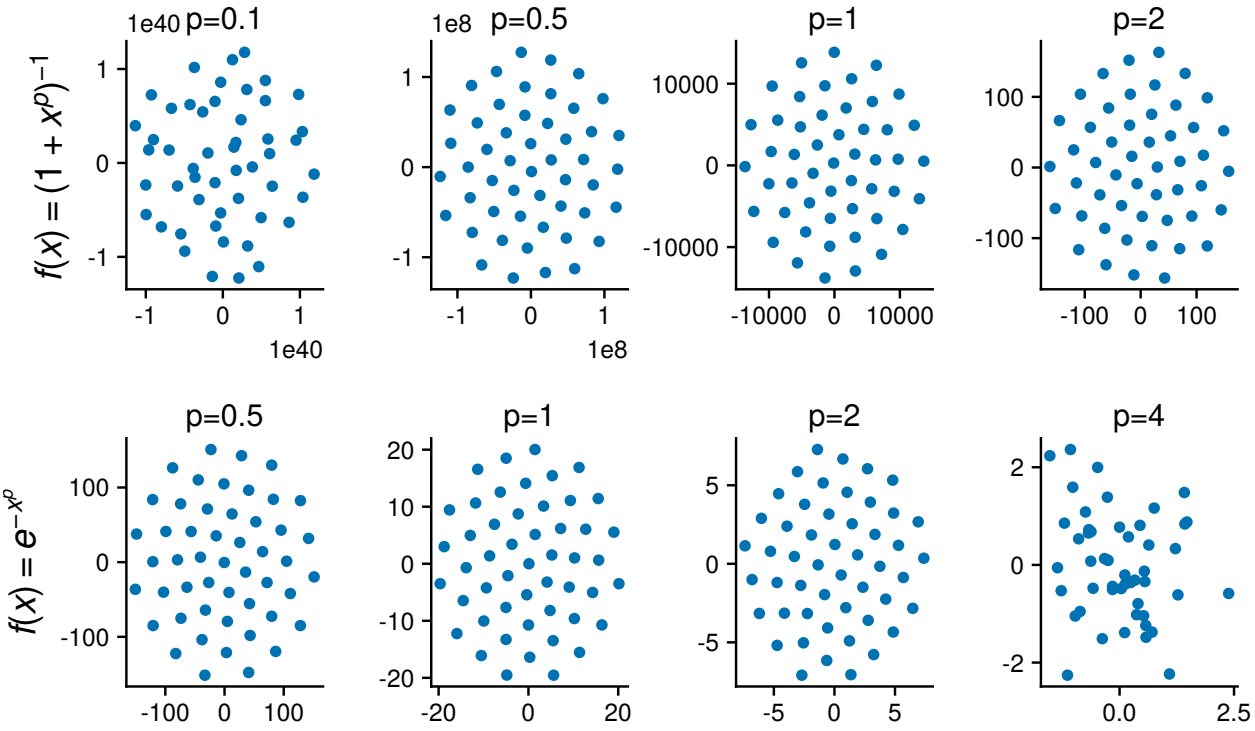

Figure 2.4: Samples obtained from a single panmictic population form circular clusters. CoaSTT applied to 20 samples obtained from a panmictic population with constant population size at a single time slice, using the same random initialization and different choices of the link function. Samples uniformly occupy circular clusters except for very slowly or very rapidly decaying link functions. $N_{\min} = 1$ and $N_{\max} = 10^6$ in all examples.

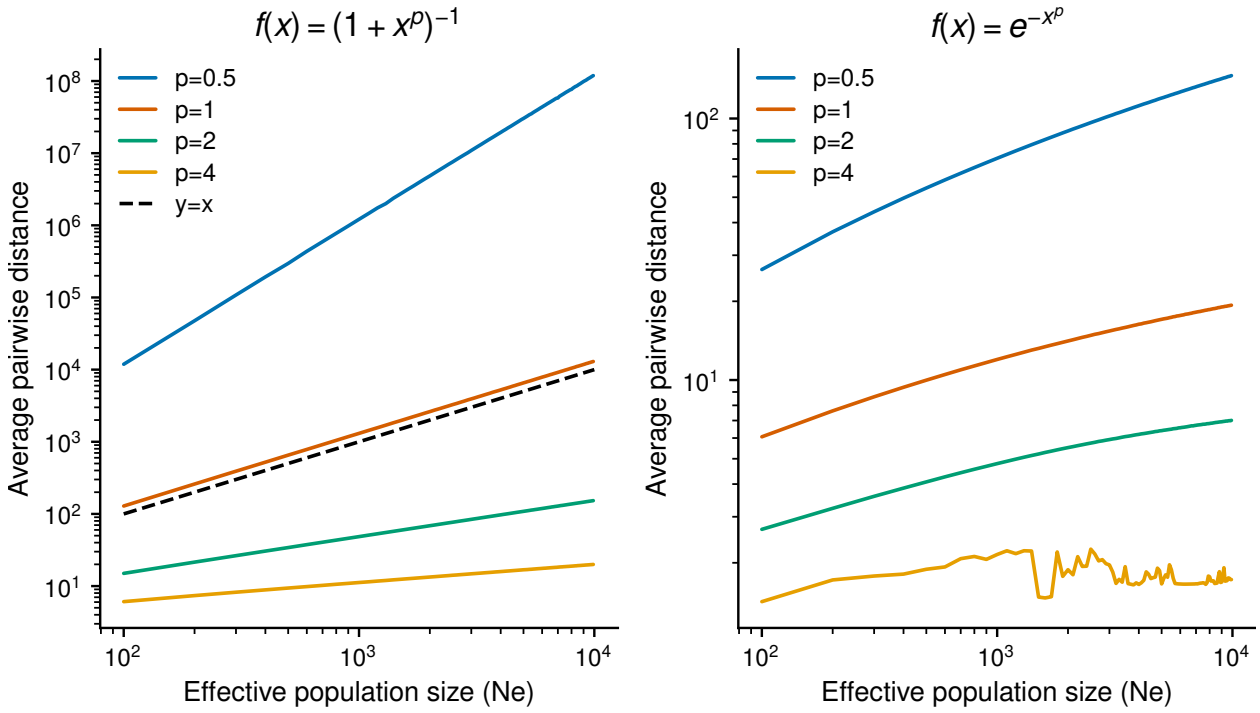

Figure 2.5: Cluster size as a function of population size under different link functions. Average pairwise distance between the embeddings of $n = 20$ samples obtained from a single panmictic population increases as population size increases. The relationship between average pairwise distances and N_e depends on the link function, with $f(x) = (1 + x^p)^{-1}$ exhibiting a linear relationship.

is very close to the radius itself. Therefore, embeddings obtained using the link function $f(x) = (N_{\text{min}} + x)^{-1}$ have the convenient property that the radius of the cluster roughly corresponds to population size if samples are obtained from a panmictic population. A cluster with double the radius corresponds to a population of double the size. Because of this convenient property, we set this as the "default" link function for our analyses, unless otherwise stated.

In the examples shown so far, we have kept N_{min} fixed at 1 and the number of samples n fixed at 20. We investigate how the cluster size varies with N_{min} and n under the inverse-power link. Remember that N_{min} defines a minimum resolution for the embeddings, where two samples with $\lambda_{ij} > \frac{1}{N_{\text{min}}}$ should occupy the same position in the latent space. Therefore, if $N_{\text{min}} > N_e$, we expect the configuration to collapse to a single point. As N_{min} is

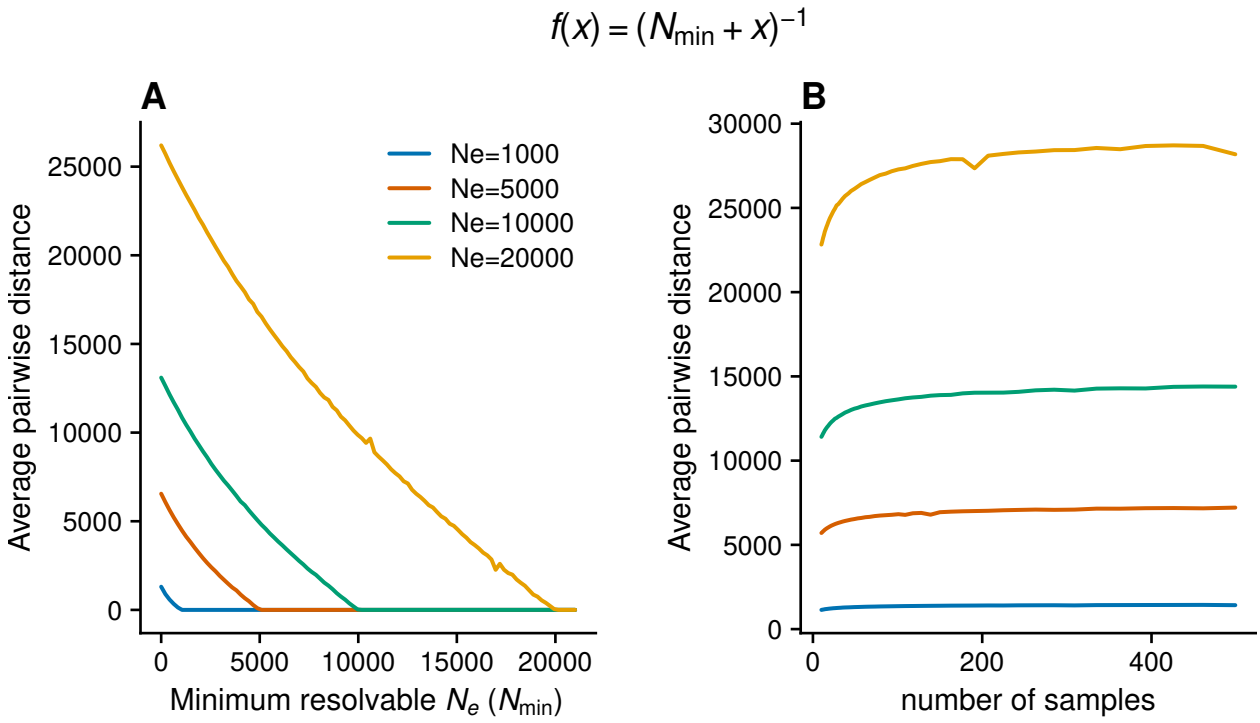

Figure 2.6: Cluster size as a function of N_{min} and n. A) Average pairwise distance as a function of N_{min} with a constant number of samples $n = 20$. **B)** Average pairwise distance as a function of the number of samples, with constant $N_{\text{min}} = 1$.

reduced below N_e, a circular cluster begins to emerge, increasing in size as N_{min} approaches 1 (figure 2.6A). With N_{min} held constant and n increasing, the cluster size increases until a plateau is reached. The number of samples required to reach this plateau increases with population size (figure 2.6B). Therefore, we emphasize the important caveat that average pairwise distances (or the cluster radius) only roughly correspond to population size, as they also depend on the choice of resolution specified by N_{min} and the number of samples n in a population. A single N_{min} is defined for all time points in CoaSTT, so the dependence on N_{min} should not be a problem. However, caution is advised when comparing clusters of different sizes that contain a small but significantly different number of samples, especially for large N_e.

2.4.2 Samples from isolated populations form distinct circular clusters

Samples obtained from isolated populations have a within-population coalescent rate defined by the effective population size ($\lambda_{ij}^{(k)} = 1/N_e^{(k)}$), and between-population coalescent rate of 0 if there is no gene flow. Therefore, it is expected that several isolated populations will form separate clusters, where differences in cluster sizes are informative of differences in population sizes. Figure 2.7 shows that this is true for most choices of link functions, but again begins to break for link functions with rapidly decaying tails or very heavy tails. Additionally, link functions with heavier tails (such as inverse-power of first or second degree) result in clusters that are better separated, or with "less crowding", which is expected following our discussion in section 2.3.5. Highly crowded configurations additionally suffer from a failure to robustly place samples into their correct clusters under random initializations. A likely explanation for this is that the misplaced sample needs to pass through a very crowded area in the latent space to reach its correct cluster. All other samples in this crowded area exert a repulsive force on that sample. If the attractive force coming from the correct cluster is rapidly decaying, it might not be strong enough to pull the sample through the aggregate of many short-range repulsive forces. Such configurations can be avoided by appropriate initialization, as discussed in Appendix 2A.

With many pairwise coalescent rates being identically zero, $N_{\max}$ becomes a relevant parameter (equations 2.29, 2.30). It defines the maximum space available for the embeddings by setting a maximum distance between samples with zero rates. Setting a higher value of $N_{\max}$ can allow samples to be better resolved into distinct clusters (figure 2.7). However, if clusters are already resolved (such as under an inverse-power link with $p = 1$), it pushes the samples even further apart, resulting in visual difficulty in distinguishing cluster sizes. Therefore, we do not recommend choosing an $N_{\max}$ that is orders of magnitude larger than the average expected N_e.

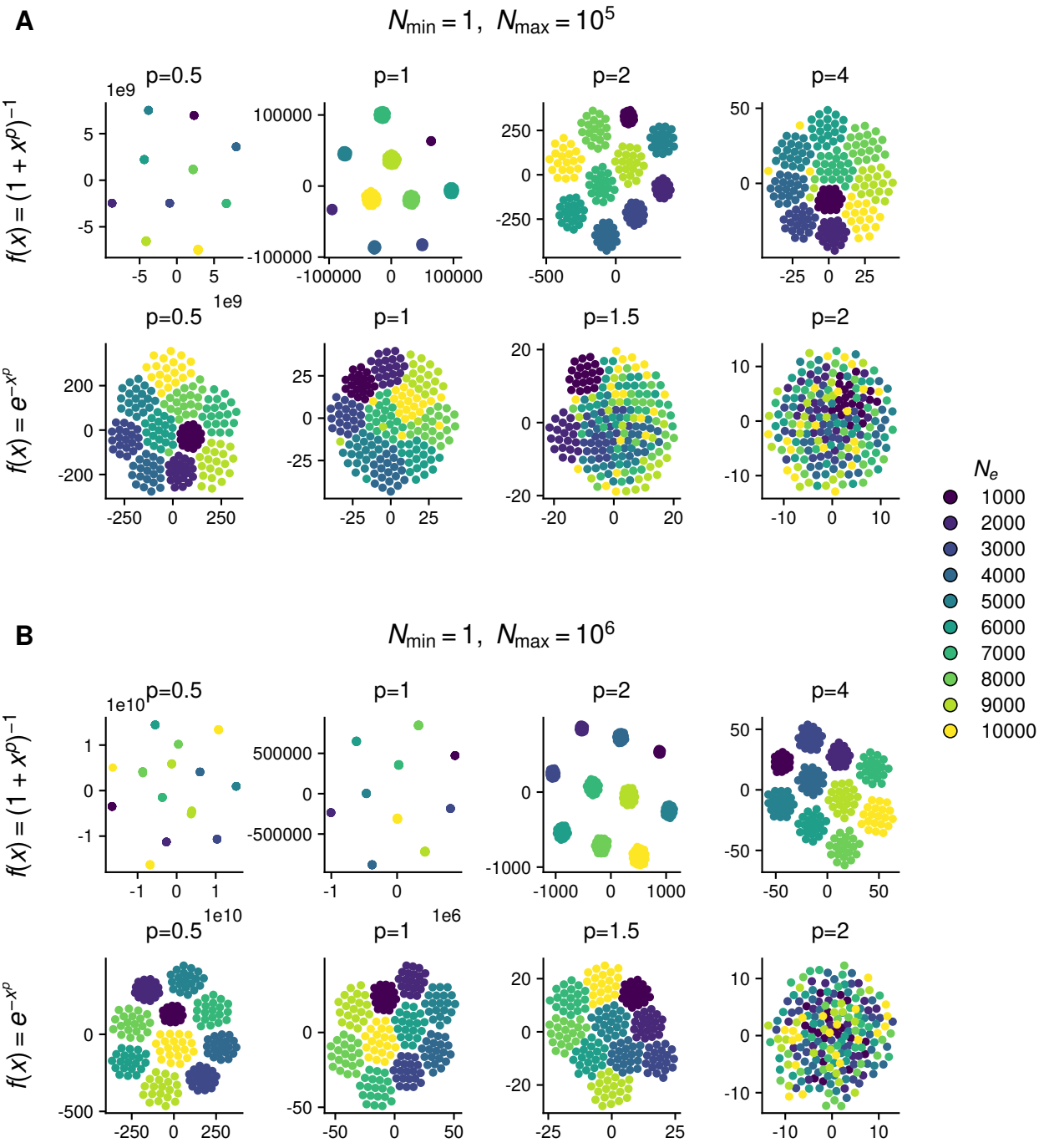

Figure 2.7: CoaSTT embeddings for samples obtained from distinct populations. Embeddings of 20 samples obtained from each of 10 panmictic populations with different population sizes, under a variety of link functions. **A)** $N_{\max} = 10^5$, one order of magnitude larger than the largest population size. Clusters fail to form under several link functions, but when they form, cluster sizes are visually distinguishable. **B)** $N_{\max} = 10^6$, clusters form more robustly for a variety of the link functions; sizes could be more difficult to resolve visually.

2.4.3 Gene flow causes population clusters to approach one another at a rate that increases with migration rate

If subpopulations are not completely isolated, but there is some degree of gene flow between them due to historical migration, we would not expect the clusters shown in figure 2.7 to be constant in time. Coalescent rates between samples from different populations are no longer identically zero, and will rather increase as we go back in time. Specifically, for K subpopulations, pairwise coalescent rates are now given by

$$\lambda_{ij}(t) = \frac{1}{S_{ij}(t)} \sum_k \frac{P_{ij}(k;t)}{2N^{(k)}(t)} \tag{2.31}$$

where $S_{ij}(t)$ is the probability that lineages (i, j) have not coalesced by time t (the survival function), $P_{ij}(k;t)$ is the joint probability that they have not coalesced by time t *and* they are both present in population k at time t, and $N^{(k)}(t)$ is the size of population k at time t. The joint probability cannot be factored into a product of probabilities, and will depend on both the between-population migration rates (that could themselves be time-varying) and the size of each subpopulation.

For a general structured population, solving equation 2.31 analytically to obtain closed-form solutions of the pairwise coalescent rates is non-trivial, but is not difficult to do numerically (especially with the `msprime` debugging tools). Figure 2.8 shows the time-varying within-population and between-population coalescent rates for the simple case of two subpopulations of the same size N that exchange migrants at a constant rate m, under different values of the population-scaled migration rate Nm. Between-population coalescent rates increase as we go back in time since migration events accumulate, allowing the two lineages to coalesce when they are in the same population. This increase eventually plateaus as the populations become well-mixed, and are effectively a single population with a new effective population size given by the inverse of the limiting coalescent rate. On the other hand,

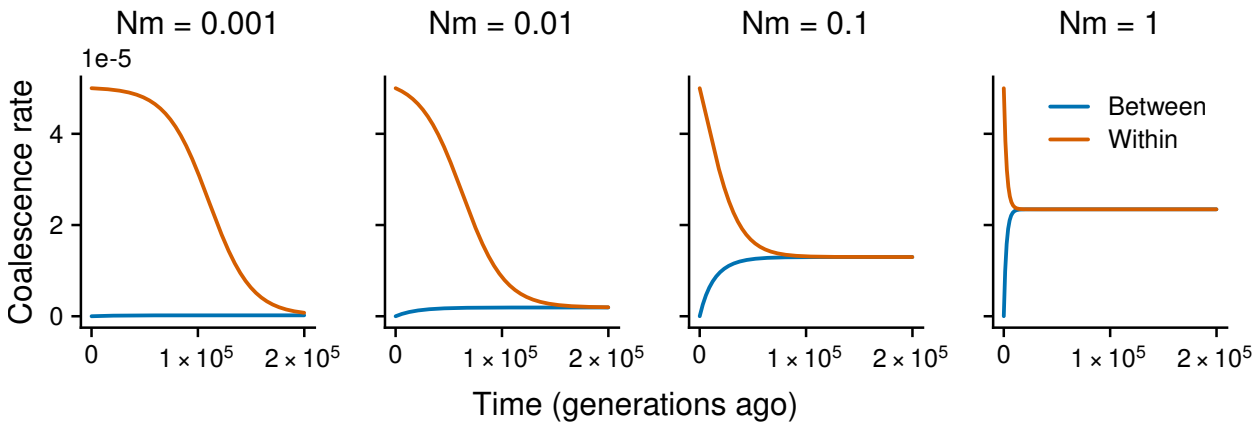

Figure 2.8: Time-varying coalescent rate for coalescence with migration. Two populations with equal sizes $N = 10^4$ and historical migration with a constant rate m between them. Within-population rate decreases with time while between-population rate increases until the populations become effectively well-mixed with a single coalescent rate.

within-population coalescent rates decay until they reach the limiting rate. We can find the limiting coalescent rate and the rate of becoming well-mixed by fitting the between-population coalescent rate to the function

$$\lambda_{\text{between}} = a(1 - e^{-bt}).$$

The factor a gives the limiting coalescent rate, while the rate constant b is a proxy for the rate at which panmixia is reached. Figure 2.9 shows how the limiting coalescent rate and the rate of becoming well-mixed vary with Nm.

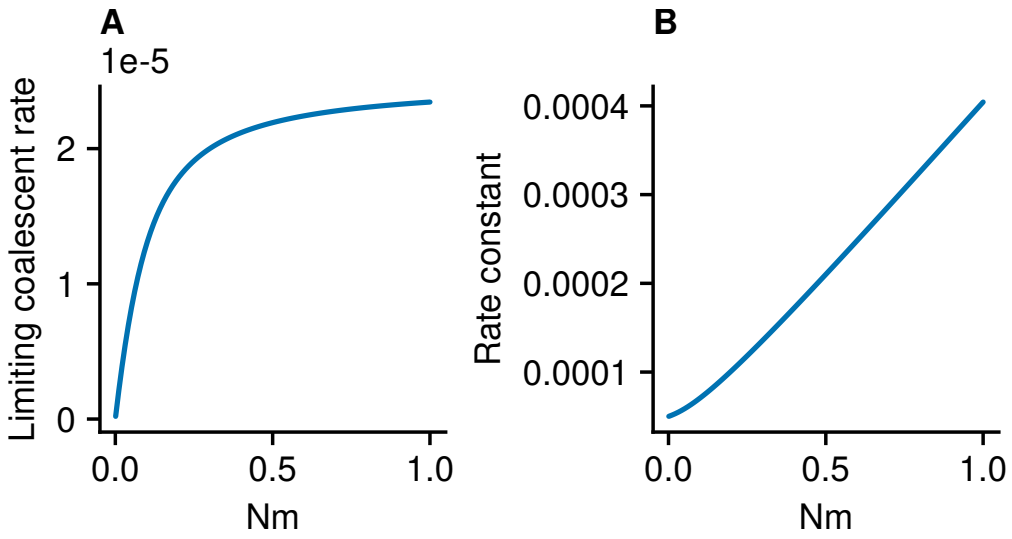

Figure 2.9: Limiting and transient behavior of coalescence with migration. A) Limiting coalescent rate of λ_{within} and λ_{between} as a function of population-scaled migration rate. The inverse of the limiting coalescent rate defines the effective population size of an ancestral well-mixed (panmictic) population. **B)** The exponential rate at which panmixia is reached as a function of population-scaled migration rate.

Since the between-population coalescent rate increases as we go back in time, CoaSTT embeddings will reflect this with clusters corresponding to different populations approaching each other backward in time during the period of migration. Figure 2.10 shows an example of this for two different migration scenarios: A) Recent migration, where populations come back in contact after being separated for a long period, and B) A soft split, where migration continues to occur for some time following the separation of populations. Clusters approach each other during the time when migration is occurring, with the rate of approach reflecting the rate at which the between-population coalescent rate is increasing. Applying CoaSTT to this simple example clearly illustrates a critical property of the method: CoaSTT embeddings reflect coalescent history, not historical population structure. The true historical structure is that of two populations that are constant in size and are exchanging migrants. CoaSTT reflects this history in terms of a visual representation of the corresponding coalescent rates, where within-population rates could themselves vary with time as a result of migration, even though the actual size of each population is constant.

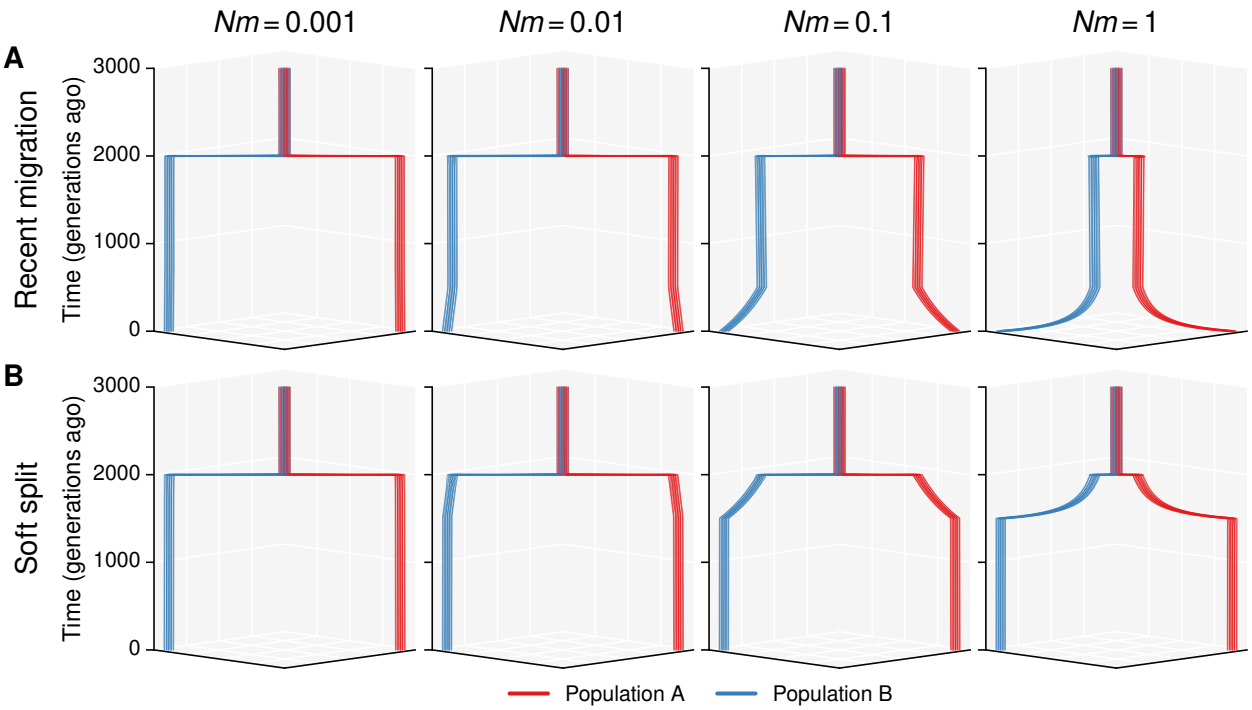

Figure 2.10: CoaSTT applied to a population split with migration. Both populations have a constant size of 10^4 with migration occurring at a constant rate m between them only within a specified period, either in the recent time **(A)** or right after they split **(B)**.

2.5 CoaSTT applied to ARGs simulated under complex demographic scenarios

After describing how time-resolved latent embeddings are inferred given input coalescent rates, we explored the properties of these embeddings under simple demographic scenarios using true coalescent rates as inputs. In practice though, CoaSTT takes *estimates* of coalescent rates as input. Most of the methods that were developed in the past decade to infer coalescent rates from sequence data do so in terms of population units, inferring within-population and between-population coalescent rates, rather than coalescent rates between individuals. Such methods also tend to discretize time in some way, with the choice of discretization being critical to the bias-variance tradeoff of the rate estimates (Spence et al.,

2018). ARGs provide a promising avenue for the inference of coalescent rates between pairs of individuals without assigning them to populations a priori. In this section, we describe how to obtain estimates of pairwise coalescent rates (and the variances) at every generation (without discretizing time) from ARGs. We then demonstrate the ability of CoaSTT to unravel more complex population histories using ARGs simulated under non-trivial demographies as inputs rather than the true rates.

2.5.1 Computing coalescent-rate estimates and their variances from ARGs

If we had access to truly independent coalescent trees, we could compute the MLE of the rates simply as $\frac{C_{ij}(t)}{sL_{ij}(t)}$, where C_{ij} is the number of trees where a coalescent event occurs within $[t, t+s]$, and $L_{ij}(t)$ is the number of trees where (i, j) have not yet coalesced by time t. However, as discussed in section 2.3.2, it is not straightforward to obtain truly independent trees for all pairs of samples in an ARG jointly. Therefore, we have recast our likelihood into a pseudo-likelihood that takes general estimates of the rates and the variances of these estimates as input. Fortunately, we can use information from the full ARG to infer coalescent rates without the need for thinning or down-sampling. We invoke an important result from the Sequentially-Markovian Coalescent (SMC) (Marjoram & Wall, 2006; G. A. T. McVean & Cardin, 2005), which is an approximation to the point-process description of the coalescent with recombination (Wiuf & Hein, 1999). In the SMC, the coalescent with recombination for a pair of genomic samples is described as a Markov chain over the genomic sequence, with the state at each genomic position l being the pairwise coalescent time at that position T_{ij}^{l}. H. Li and Durbin, 2011 showed that the limiting distribution of pairwise coalescent times is given by

$$f_{T_{ij}}(t) = \lambda_{ij}(t) e^{-\int^{t} \lambda_{ij}(s) ds} \tag{2.32}$$

Such a distribution can be estimated for each pair of samples by invoking the ergodicity of the chain using genome-wide estimates of pairwise coalescent times that are readily available in an inferred ARG. If the distribution of pairwise coalescent times is known, the associated survival function can be computed as

$$S_{ij}(t) = P(T_{ij} > t) = e^{-\int^t \lambda_{ij}(s)ds} \tag{2.33}$$

and the instantaneous pairwise coalescent rate can be directly obtained by taking the derivative of the log of the survival function

$$\lambda_{ij}(t) = -\frac{d}{dt} \ln S_{ij}(t) \tag{2.34}$$

In the rest of the section, we describe how we obtain estimates of the distribution of pairwise coalescent times from inferred ARGs that are represented in the succinct tree sequence format (Kelleher et al., 2016, 2018), and how efficient numerical derivatives of the noisy estimates $-\ln \hat{S}_{ij}(t)$ can be obtained, as well as a method for estimating the variances.

Estimating the distribution of pairwise coalescent times from ARGs

The SMC is geometrically ergodic in total variation (Terhorst, 2026). What this means is that wherever the chain starts, it converges to the stationary distribution at an exponential rate. Consequently the Markov Chain CLT holds, and the distribution of coalescent times over the whole chain is an unbiased estimate of the genome-wide stationary distribution. We therefore estimate the binned pdf of pairwise coalescent times by simply counting the positions where coalescence occurs during a time interval. For our purposes, we can narrow the bins to a single generation (or a small number of generations Δ_t), obtaining a very noisy estimate of the distribution of pairwise coalescent times. With L being the total genomic

length, the estimate is given by

$$\hat{f}_{T_{ij}}(t) = P(t < \hat{T}_{ij}^{l} < t + \Delta_t) = \frac{1}{L} \sum_{l} \mathbf{1}\{t < \hat{T}_{ij}^{l} < t + \Delta_t\} \tag{2.35}$$

The estimates of pairwise coalescent times at each genomic position $\hat{T}_{ij}^{l}$ are simply read off from an inferred ARG. Note that this is the main quantity that is obtained from ARGs to be used downstream for our method. Any other method that provides estimates of pairwise coalescent times over the genome from sample data can serve as an equally valid source of input to CoaSTT. In practice, we obtain the above estimate from inferred ARGs stored in the `tskit` format using the function `TreeSequence.pair_coalescence_counts()` from the `tskit` library (Kelleher et al., 2018). An estimate of the binned distribution of coalescent times is obtained for all pairs of samples by appropriately normalizing the computed counts. The normalization takes into account masked regions of the genome or regions where ancestry could not be estimated, for which samples do not have ancestry in the tree sequence and could not coalesce. From the pdf estimate, an estimate of the survival function is computed as

$$\hat{S}_{ij}(t) = 1 - \sum_{r<t} \hat{P}(r < T_{ij}^{l} < r + \Delta_t) \tag{2.36}$$

The function `TreeSequence.pair_coalescence_counts()` scales quadratically with sample size (in memory and time). For $\sim$ 500+ samples, runtime becomes prohibitively long. Nevertheless, for large sample sizes, we can afford to feed in noisy estimates of the rates to the loss function as each sample now draws information from many more samples to inform its position in the latent space. We can obtain such noisy estimates at lower runtime by chopping up the inferred tree sequence into intervals of equal size (e.g. 5Mb), and counting the coalescent segments for a randomly chosen fraction of the samples (specified by the user) in each segment, then normalizing the counts appropriately in the end. We couple this with

performing the computation across segments in parallel to massively reduce compute time.

Estimating $\hat{\lambda}_{ij}(t)$ from survival estimates

With $\hat{S}_{ij}(t)$ estimated, $\hat{\lambda}_{ij}(t)$ can simply be obtained by taking the numerical derivative of $-\ln \hat{S}_{ij}(t)$. Since the coalescence information in the genome is finite, $-\ln \hat{S}_{ij}(t)$ will be quite jagged. Therefore, taking a numerical derivative at each time point by simple finite differences will yield a very noisy estimate for $\hat{\lambda}_{ij}(t)$, especially for time points further in the past. One solution to this is to compute piecewise-constant estimates of the rates. However, it is more desirable in our case to obtain estimates of coalescent rates at almost every generation, avoiding the need to discretize time in a particular manner, and allowing for approximately continuous representations of latent trajectories. Rather than smoothing $\binom{n}{2}$ noisy rate estimates, we use a non-parametric filter that can both smooth the cumulative rate estimates and take the derivatives for all samples and all time points by a single matrix multiplication.

For a given time point t_i (where $t_i = i\Delta_t$, Δ_t being the bin size of the pdf estimate), we estimate the rate as the slope of a local linear fit for time-points in the range $[t_i - \frac{b_i}{2}, t_i + \frac{b_i}{2}]$. The bandwidth b_i increases with time since less signal is present in the past (fewer uncoalesced local trees). For the start and end boundaries ($t_i < \frac{b_0}{2}$ and $t_i > T - \frac{b_T}{2}$), the local linear fit is obtained over $[0, b_0)$ and $[T - b_T, T)$ respectively, and the same rate is assigned to all time points on a boundary. We construct a filter matrix F such that applying F to the estimate of cumulative rates for each pair ($-\ln \hat{S}_{ij}(t)$) will simultaneously fit the polynomial and take the derivative for all points in time. It can therefore be applied to all pairs simultaneously by a matrix multiplication.

If $\hat{\Lambda}$ is a $\binom{n}{2} \times T$ matrix of cumulative rate estimates over time, with each row representing a pair of samples and each column representing a time point, and $\hat{\lambda}$ is a matrix of the same

shape for pairwise rate estimates over time, we construct the $T \times T$ matrix F such that

$$\hat{\lambda} = \hat{\Lambda} F^T \tag{2.37}$$

$$F_{i,j} = \begin{cases} \dfrac{b_0\, x_j^{(0)} - S_1^{(0)}}{b_0 S_2^{(0)} - \left(S_1^{(0)}\right)^2}, & j \in [0,\, b_0),\ i < \dfrac{b_0}{2} \\ \dfrac{(j-i)\Delta_t}{\sum\limits_{k=-\frac{b_i}{2}}^{\frac{b_i}{2}} (k\Delta_t)^2}, & j \in \left[i - \dfrac{b_i}{2},\, i + \dfrac{b_i}{2}\right],\ \dfrac{b_0}{2} \le i < T - \dfrac{b_T}{2} \\ \dfrac{b_T\, x_j^{(T)} - S_1^{(T)}}{b_T S_2^{(T)} - \left(S_1^{(T)}\right)^2}, & j \in [T - b_T,\, T),\ i \ge T - \dfrac{b_T}{2} \\ 0, & \text{otherwise} \end{cases} \tag{2.38}$$

$$\text{Early boundary } (t < \frac{b_0}{2}): \quad x_j^{(0)} = (j-t)\Delta_t,$$
$$S_1^{(0)} = \sum_{j=0}^{b_0-1} x_j^{(0)}, \qquad S_2^{(0)} = \sum_{j=0}^{b_0-1} \left(x_j^{(0)}\right)^2$$

$$\text{Late boundary } (t \ge T - \frac{b_T}{2}): \quad s = T - b_T,$$
$$x_j^{(T)} = \big(j - (t-s)\big)\Delta_t,$$
$$S_1^{(T)} = \sum_{j=s}^{T-1} x_j^{(T)}, \qquad S_2^{(T)} = \sum_{j=s}^{T-1} \left(x_j^{(T)}\right)^2$$

Bandwidth sizes b_i are user-specified, but are rounded to the nearest odd value. For simplicity, we only define minimum and maximum window sizes, and take windows of linearly

increasing sizes in between.

Coalescent rates between individuals

The tips of local trees in a tree sequence typically represent haplotypes rather than individuals. For a valid estimate of coalescent rates between each pair of haplotypes, the whole genome has to be consistently phased across chromosomes. State-of-the-art phasing methods can phase haplotypes within the same chromosome with respectable accuracy, while genome-wide phasing methods are still highly error-prone. Therefore, a genome-wide tree sequence is expected to have phase switches between chromosomes. It is more natural, at least for now, to consider between-individual coalescent rates rather than between-haplotype coalescent rates. We define a pooled survival function estimate between individuals i and j as the average of the survival estimates between their respective haplotypes

$$\hat{S}_{ij}(t) = \frac{\hat{S}_{i_1 j_1}(t) + \hat{S}_{i_1 j_2}(t) + \hat{S}_{i_2 j_1}(t) + \hat{S}_{i_2 j_2}(t)}{4} \tag{2.39}$$

where $\hat{S}_{i_p j_q}(t)$ is the survival estimate between haplotype p of individual i and haplotype q of individual j. The coalescent rate is then computed as the slope of local linear fits of $-\ln \hat{S}_{ij}$ as described.

Estimating the weights associated with rate estimates

For each rate estimate computed for a given pair of samples at a given time point, an associated weight (equation 2.14) has to be supplied to the loss function (2.15). To compute the weight, the variance of each rate estimate has to be evaluated. The rate estimate at time t_a is a linear combination of the estimated cumulative rates $\hat{\Lambda}_{ij} = -\ln \hat{S}_{ij}$. Dropping the

(i, j) indices for conciseness, we have

$$\hat{\lambda}(t_a) = \sum_b F_{ab} \hat{\Lambda}(t_b). \tag{2.40}$$

The variance of this sum is not the weighted sum of the variances of $\hat{\Lambda}(t_b)$, since cumulative estimates at different time points are strongly correlated. However, the sum can be re-expressed in terms of the increments of the cumulative rate, which are uncorrelated, as

$$\hat{\lambda}(t_a) = \sum_b F_{ab} \hat{\Lambda}(t_b) = \sum_c \Phi_{ac} \hat{q}_c, \tag{2.41}$$

with

$$\Phi_{ac} = \sum_{b \geq c} F_{ab}, \tag{2.42}$$

$$\hat{q}_c = \hat{\Lambda}(t_c) - \hat{\Lambda}(t_{c-1}) = -\ln(1 - \hat{p}_c) \approx \hat{p}_c, \tag{2.43}$$

where $\hat{p}_c$ estimates the probability of coalescing in time bin c, i.e. in $(t_{c-1}, t_c]$, conditional on not having coalesced before t_{c-1}. Since $\hat{\Lambda}(t_b)$ accumulates the increments of all bins up to and including b, the coefficient Φ_{ac} of increment c is the sum of the filter coefficients over all $b \geq c$ (equation 2.42).

To characterize the distribution of the increments, we return to the counting-process model underlying equation 2.14 (Andersen et al., 1993). Let $Y(t_c)$ denote the number of independent loci that have not coalesced by the start of bin c (the at-risk set), with $Y(t_0) = L_0$, and let C_c denote the number of these loci that coalesce within bin c. Under Assumption 3 in section 2.3.2, $C_c \mid Y(t_c) \sim \text{Poisson}\big(\lambda(t_c)\, Y(t_c)\, \Delta_t\big)$, and the increment estimate is the Nelson–Aalen increment $\hat{q}_c = C_c / Y(t_c)$. Conditional on the at-risk set, $\hat{q}_c$ is

unbiased for $\lambda(t_c)\Delta_t$ with

$$\mathrm{Var}\big(\hat{q}_c \mid Y(t_c)\big) = \frac{\lambda(t_c)\Delta_t}{Y(t_c)}. \tag{2.44}$$

Because each increment is conditionally unbiased given the history of the process up to the start of its bin, the increments form a martingale-difference sequence and are uncorrelated (Andersen et al., 1993), so the variance of the rate estimate is the sum of their variances weighted by the squared filter coefficients. The at-risk set (the set of *independent* trees where samples i, j have not coalesced yet) is not observed directly, but its expectation is $L_0 S(t_c)$, which we estimate by $L_0 \hat{S}(t_c)$. The variance of the rate estimate can therefore be written in terms of the local linear filter and the survival estimates as

$$\mathrm{Var}\big(\hat{\lambda}(t_a)\big) = \sum_c \Phi_{ac}^2 \frac{\Delta_t \lambda(t_c)}{L_0 \hat{S}(t_c)}. \tag{2.45}$$

The unknown $\lambda(t_c)$ could be replaced by the corresponding rate estimates, but it is more convenient to assume that the rate is approximately constant over the window in which the ath row of Φ is non-zero, which is centered around t_a. It is also more principled, since a weight should not inherit the noise of the very estimate it weights, and the resulting weight would be ill-defined for pairs with no observed coalescence. The rate can then be taken outside the sum as $\lambda(t_a)$, and it cancels in the weight $\hat{w}(t_a) = \lambda(t_a)/\mathrm{Var}(\hat{\lambda}(t_a))$. The constant L_0 can also be dropped, since it scales all weights equally and only relative weights affect the minimizer of the loss. The weight is thus defined as

$$\hat{w}(t_a) = \left(\Delta_t \sum_c \frac{\Phi_{ac}^2}{\hat{S}(t_c)} \right)^{-1}. \tag{2.46}$$

This is the appropriate counterpart, for the filter-based estimator of rates, of the weight $w(t) = sL(t)$ that would result from estimating the rates by counting events in truly inde-

pendent trees (equation 2.14).

2.5.2 CoaSTT jointly reveals split times and population size changes without defining population units

To understand how CoaSTT embeddings over time should be interpreted, we demonstrated a suite of configurations that emerge when CoaSTT is applied to simple demographic scenarios in section 2.4. Panmictic populations form circular clusters whose size increases with increasing N_e (section 2.4.1), a structured population results in separated clusters with the degree of separation increasing with the heaviness of the link function's tail (section 2.4.2), and gene flow results in the distance between clusters decreasing backward in time (section 2.4.3). In reality, the coalescent history of a set of samples is shaped by a potentially complex combination of historical processes, and these simple examples help build intuition on how to interpret embeddings resulting from such complex scenarios. We use msprime (Kelleher et al., 2016) to simulate tree sequences under more complex demographic scenarios that serve as "challenging cases" that mimic realistic complexity, illustrating the ability of CoaSTT to provide intuitive visual representations of non-trivial historical processes without the need to assign samples to populations.

The first example demonstrates CoaSTT's ability to jointly reveal population splits and split times, along with variations in population size that occur at different rates for each subpopulation. We simulate trees under the demography shown in figure 2.11A, where the size of each subpopulation varies at a different rate following its split from its ancestral population. Figure 2.11B shows that the latent embeddings remarkably mimic the simulated demography without the need to assign population labels to samples. Such assignment would be necessary to resolve a history like this one using current methods.

An interesting aspect of this simulation is that the split times are chosen such that the most likely gene tree is discordant with the true demography (figure 2.12), or in the "anomaly

zone" of the phylogeny (Degnan & Rosenberg, 2006). This choice is made to emphasize the utility of coalescent-based methods like CoaSTT in cases where standard phylogenetic methods that rely on the most likely gene tree would fail to accurately represent the history.

We also investigate the effect of reducing the dimensionality of the pairwise coalescent rates. At each time slice, $\binom{n}{2}$ rate estimates are summarized by $2n$ latent positions. In this particular example, with a total of $n = 80$ individuals, 3160 rate estimates are summarized by 160 positions, two for each individual. Rates can be reconstructed from the latent positions by inverting the link function, and such a $\sim 95\%$ reduction in the number of parameters would naturally result in a mismatch between the input rates and the reconstructed rates, the extent of which is shown in figure 2.11C. The geometric reason for the mismatch is that it is impossible to position $n > 3$ samples with very similar (or identical) distances in two dimensions, and such loss of information is an inevitable trade-off associated with visually interpretable low-dimensional representations of this sort. As discussed in section 2.3.5, this effect is mitigated by link functions with heavier tails. The embeddings in figure 2.11B are computed using an inverse-power link of degree $p = 1$. Clusters in this case have a radius that conveniently scales linearly with N_e (section 2.4.1). Lower values of p could potentially result in embeddings with lower reconstruction error (figure 2.11D) at the expense of less intuitive visualizations, since cluster radii will scale differently with N_e. Additionally, with heavy-tailed link functions, distances between clusters can grow to an extent where resolving differences between cluster sizes becomes visually impossible (section 2.4.2). The choice of link function should therefore be guided by whether the ultimate goal is visualization or minimization of reconstruction error.

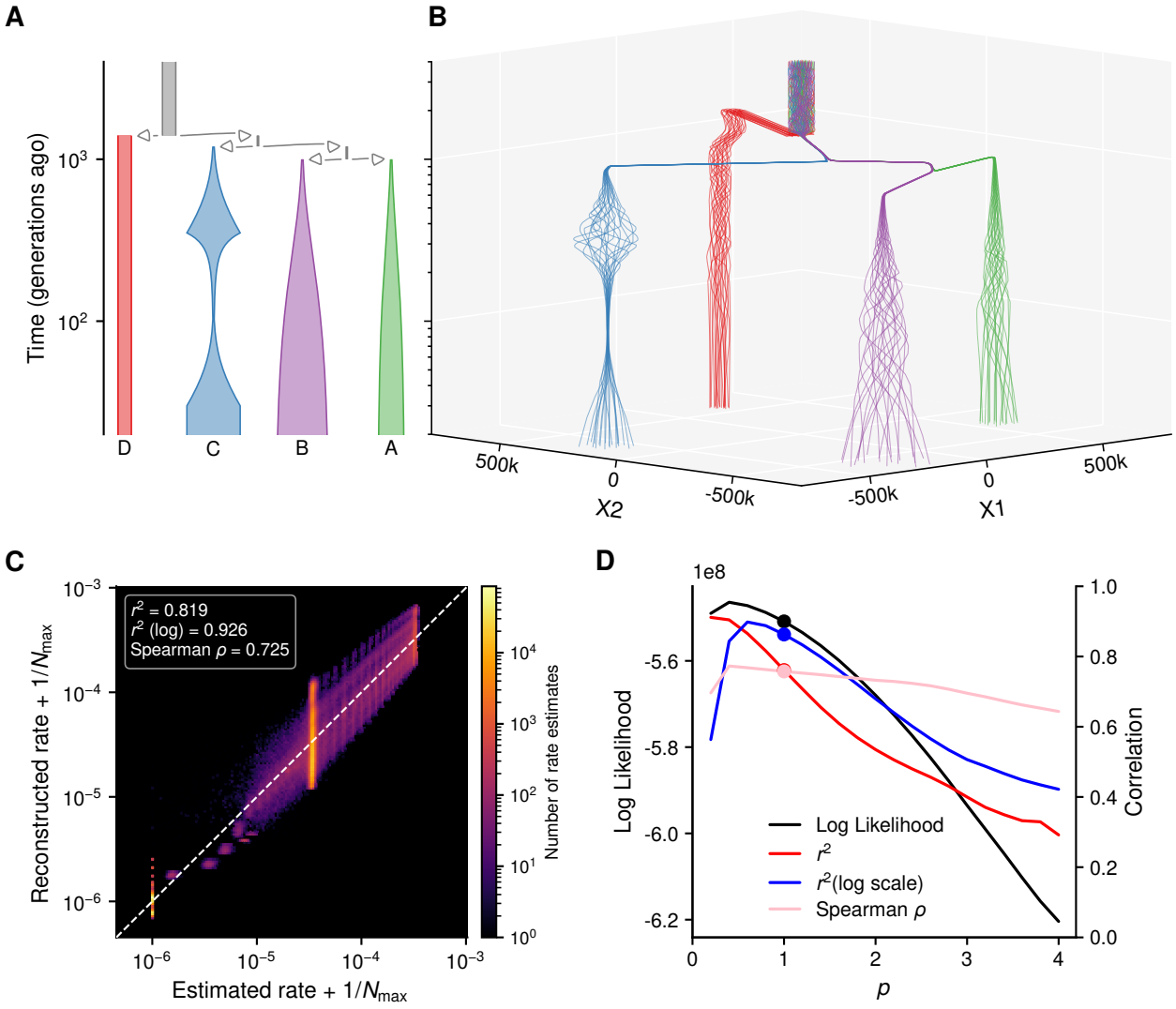

Figure 2.11: CoaSTT applied to a complex 4-population demography. CoaSTT embeddings for 20 individuals per population simulated under the demographic history **(A)**, where the width of the rectangle representing each subpopulation reflects population size. **B)** CoaSTT embeddings reveal four clusters in recent time that merge at split times. The size of each cluster varies according to each population's size. **C)** Mismatch between input rate estimates and rates reconstructed from latent positions reveals the loss of information resulting from dimensionality reduction. **D)** Different metrics of reconstruction error as a function of the rate of decay of the inverse-power link function, with higher degree p indicating faster decay. $p = 1$ highlighted with bold points.

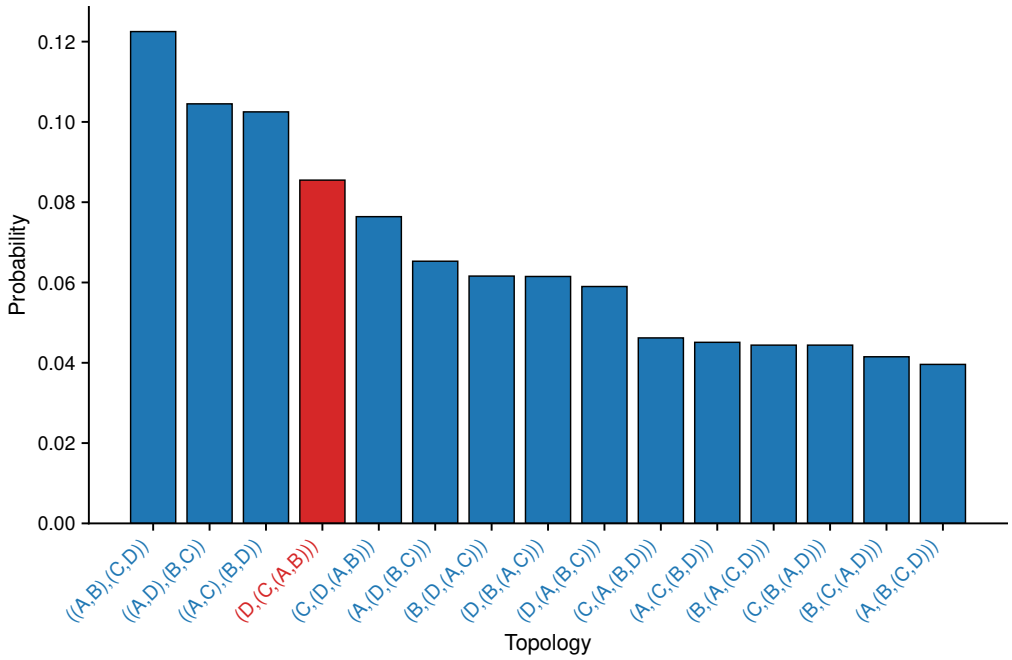

Figure 2.12: Probabilities of tree topologies of the complex 4-population demography. The probability mass function of gene trees for one haploid sample per population simulated under the demography in figure 2.11A. Obtained numerically by simulating 10^4 single trees using msprime. The true nested topology is shown in red, with four topologies being more probable than it.

2.5.3 CoaSTT jointly reveals continuous and discrete coalescent structure in the presence of gene-flow

We have shown the effect of simple patterns of gene flow between discrete populations on CoaSTT embeddings in section 2.4.3. In reality, patterns of gene flow can be much more complex. In this section, we show how CoaSTT reveals several complex patterns of gene flow, emphasizing caveats associated with interpreting cluster sizes in the presence of gene flow, and highlighting the unique potential of CoaSTT to jointly reveal discrete and continuous patterns of historical population structure.

We reiterate a result that was already shown in figure 2.10: clusters corresponding to discrete populations approach one another backward in time in the presence of gene flow. We now tie this result to the more complex demographic history presented in the previous

section (figure 2.11A). We simulate the same demography with migration occurring at a constant rate between the two recently expanding populations A and B (figure 2.13). As expected, the clusters corresponding to both populations approach each other, with the rate of approach increasing with migration rate (figure 2.13A-C). The more interesting observation worth emphasizing is that the sizes of clusters A and B no longer reflect their population size. More specifically, there is a mismatch between the inverse of the true population size and the within-population coalescent rate that gets wider the higher the migration rate. To understand this, consider two genomes sampled at the present day from population A. Going backward in time, the probability of coalescence at a genomic locus is simply given by $1/(2N^{(A)}(t))$ as long as no migration event occurs. Once a migration moves one of the samples to population B, the lineages could no longer coalesce at that locus until another migration event moves either lineage. If the other lineage migrates to population B as well, they will coalesce at the rate $1/(2N^{(B)}(t))$, even though both lineages were sampled from population A. The within-population coalescent rates shown in figure 2.13D represent the genome-wide average probability of coalescence under such a process. At a high migration rate ($m = 0.01$), individuals sampled from populations A and B coalesce as though they belong to the same population around 200 generations ago, with a within-population coalescence rate that is lower than the inverse of either population's size. This hypothetical population has a higher effective population size. At a much lower migration rate ($m = 0.0001$), the within-population coalescent rate is quite close to inverse population size, with the distance between the clusters decreasing more slowly in the embeddings. This result emphasizes more clearly how CoaSTT embeddings reveal coalescent history, or coalescent structure through time (represented in terms of coalescent rates), as opposed to demographic structure through time (represented in terms of population sizes and migration rates).

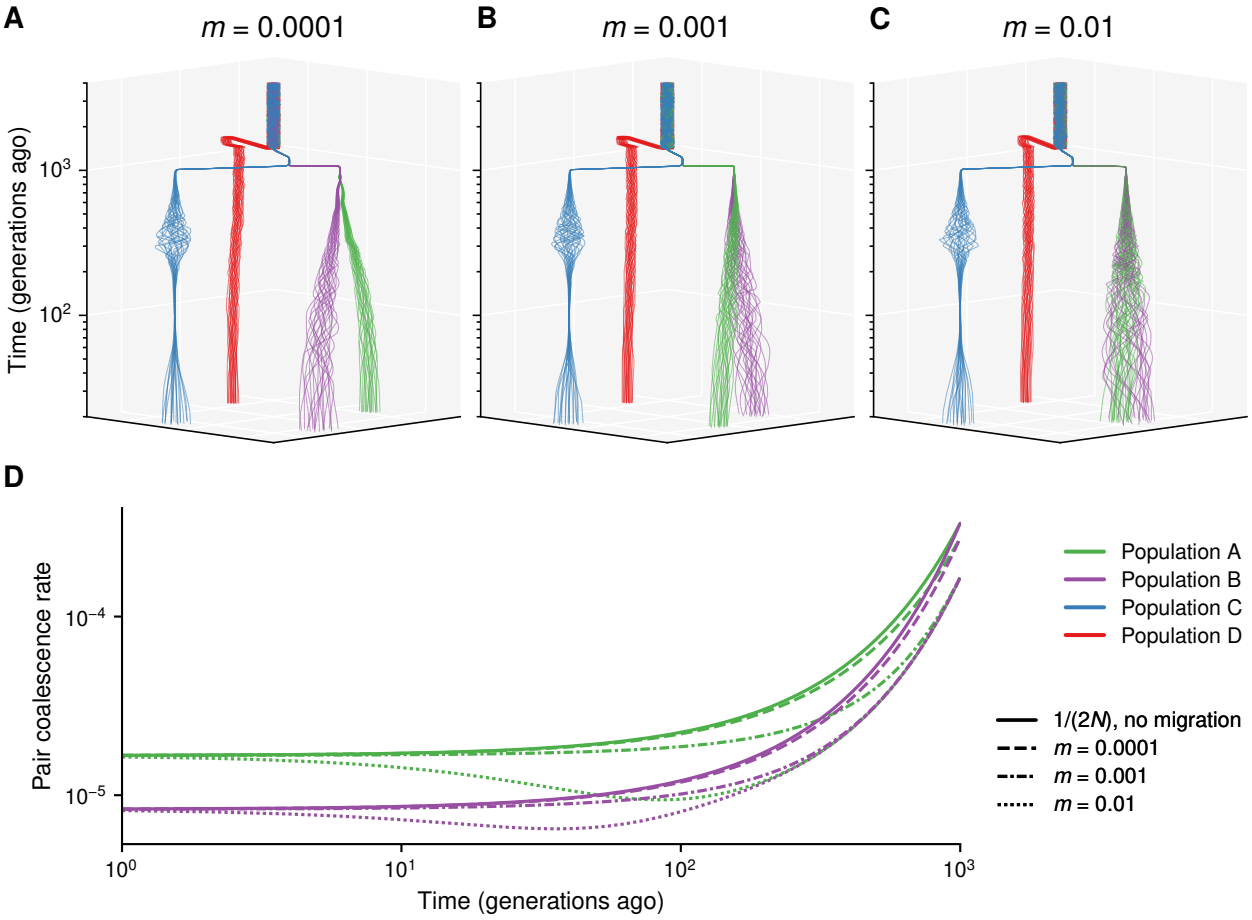

Figure 2.13: Complex 4-population demography with recent migration. Populations A and B exchange migrants following their split at a rate of $m = 10^{-4}$ **(A)**, $m = 10^{-3}$ **(B)** and $m = 10^{-2}$ **(C)** per generation. **D)** Within-population coalescent rates as a function of time at each migration rate plotted with the trajectory of inverse population size for populations A and B, illustrating a mismatch between inverse population sizes and within-population coalescent rates that gets larger with higher migration rates.

All demographic scenarios presented so far are examples where history is presented in terms of discrete population units, perhaps with gene flow. Such discreteness results in embeddings with discrete population clusters. Continuous population structure is prevalent in natural populations, especially for species with continuous spatial distributions and persistent migration patterns such as humans. Therefore, it is crucial to understand the behavior of CoaSTT when applied to samples simulated under historical processes that result in continuous structure. First, we simulate a 1-dimensional linear stepping-stone model

with 100 demes (each with a population size of $N = 100$ individuals), where each deme exchanges migrants with its neighbors with a rate m per generation, sampling one individual per deme. PCA applied to similar simulations produced 'horseshoe' configurations (PC1 vs. PC2) (Novembre & Stephens, 2008). In recent time, CoaSTT embeddings rather lie on a linear manifold, corresponding to the linear spatial structure. The width of this line decreases backward in time, reflecting the increase in pairwise coalescent rates between all pairs, effectively pulling everyone closer together. Eventually, the line transforms into a circle, as the lineages at that period are effectively well-mixed, with all pairs having the same coalescent rate, the inverse of which specifies the effective size of the well-mixed ancestral population (figure 2.14). A key factor affecting the timescale of such loss of structure is the timescale at which the Markov process of migration reaches its limiting distribution, at which point each sample is equally likely to be in any deme regardless of where it started. As expected, the rate of loss of structure is higher for higher migration rates. We also simulated a 2-dimensional stepping-stone model consisting of a 10×10 grid of demes, where each deme also exchanges migrants with its neighbors on the square grid with a constant rate m. The same pattern is observed, where the embeddings are organized on a 2D square grid in recent time, with distances between samples on the grid decreasing backward in time until the embeddings form a well-mixed circular cluster.

An intriguing "fragmentation effect" is observed in recent times for embeddings obtained under the inverse-distance link function ($p = 1$) which is not seen under the inverse-squared distance ($p = 2$) for both the 1D and 2D stepping-stone examples (figures 2.14A and 2.15A). In the 1D case, the continuous linear manifold is fragmented into smaller linear clusters, whereas in the 2D case, groups of samples on the grid cluster together, with uneven distances between these groups. The fragmented embeddings are sustained for a period of time until they merge into a continuous line or grid. This fragmentation effect has been reported in the context of t-SNE when heavy-tailed kernels are used (Kobak et al., 2020). It has

been found that using kernels (link functions) with heavy tails has the potential to reveal discrete structure more clearly and at a finer scale by pushing clusters further apart, but concurrently risks breaking continuously structured clusters into multiple discrete fragments. The exact mechanism for this fragmentation effect is not fully understood in the Neighbor Embedding literature. Our intuition is that in recent times, attractive forces exist only between samples from nearby demes that have a strong enough coalescent rate between them. However, repulsive forces are exerted between every single pair of samples. For a heavy-tailed link function (such as the inverse-distance), the repulsive forces survive for larger distances, whereas the attractive forces are restricted only to nearby samples. The sum of these long-range repulsive forces might dominate the short-range attractive forces, causing fragmented configurations to have a lower total loss than continuous ones. In fact, we discover a bifurcation that occurs when using the true rates of a 1D stepping-stone model a single generation ago, where the fragmented solution is shown to have a lower loss value than the continuous solution for any $p < 2$ (Appendix 2B).

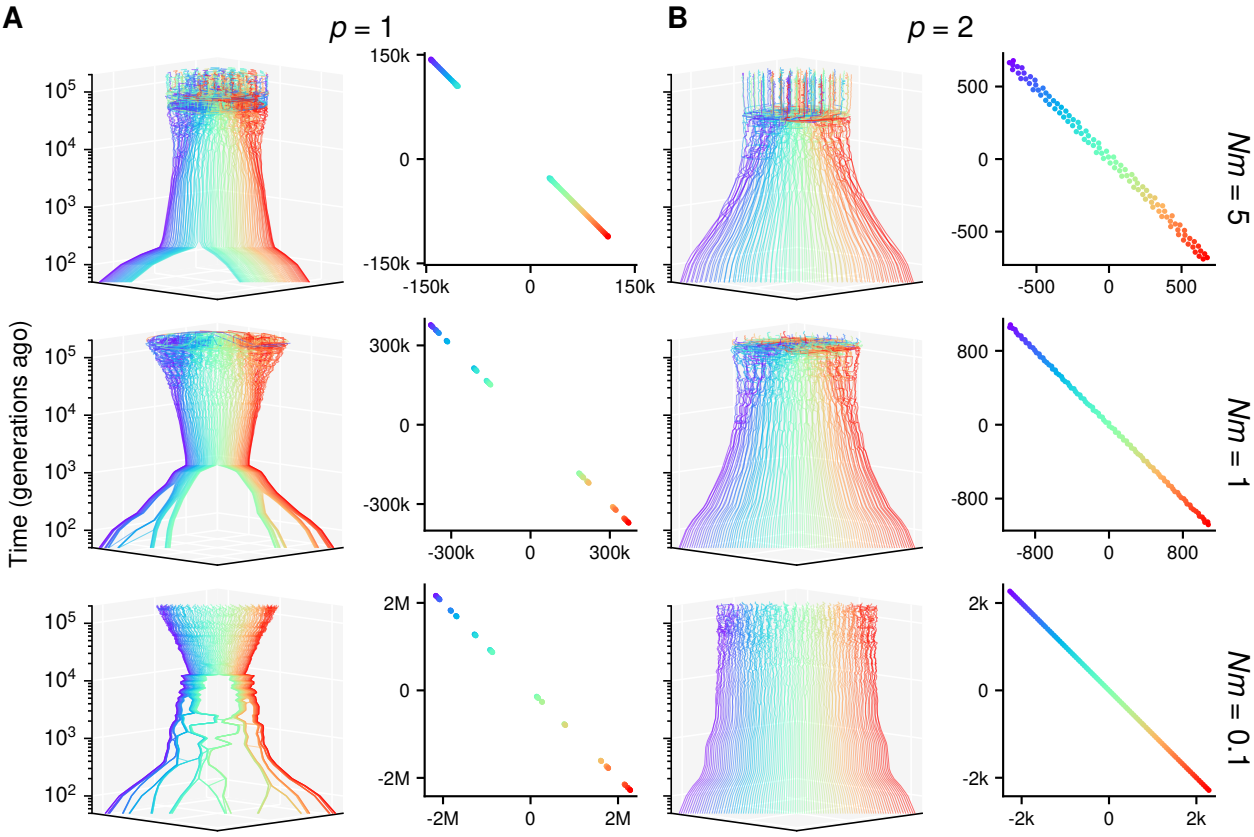

Figure 2.14: 1D stepping-stone model. CoaSTT embeddings for individuals sampled from 1D stepping-stone simulations with different effective migration rates ($Nm = 5, 1, 0.1$). **A)** Embeddings computed using the link $f(x) = \frac{1}{1+x}$ form fragmented linear clusters in recent time. **B)** Embeddings computed using the link $f(x) = \frac{1}{1+x^2}$ with a faster-decaying tail are better behaved as they are continuously structured in recent time.

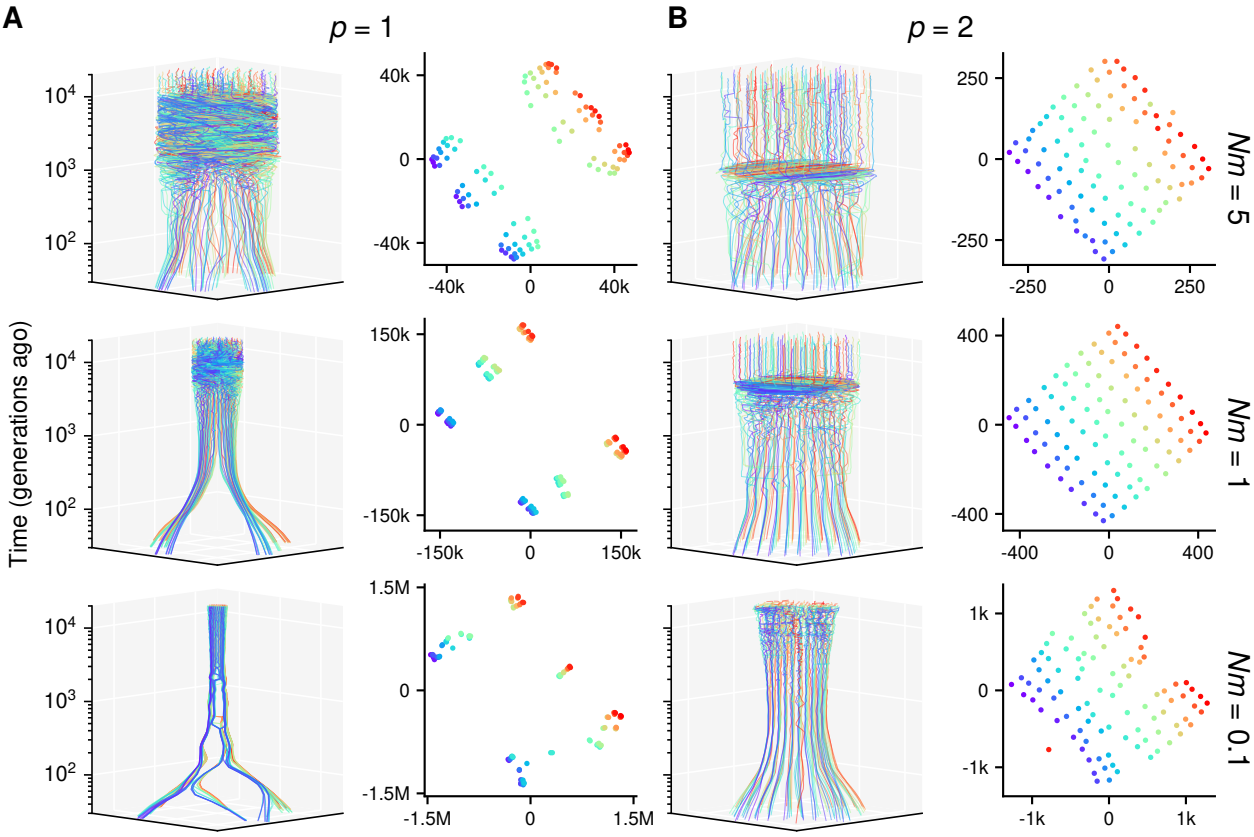

Figure 2.15: 2D stepping-stone model. CoaSTT embeddings for individuals sampled from 2D stepping-stone simulations with different effective migration rates ($Nm = 5, 1, 0.1$). **A)** Embeddings computed using the link $f(x) = \frac{1}{1+x}$ form fragmented and uneven clusters of grids in recent time. **B)** Embeddings computed using the link $f(x) = \frac{1}{1+x^2}$ with a faster-decaying tail are more continuously structured in recent time, though the structure is still distorted for $Nm = 0.1$, likely due to the noisy low-rate estimates. Note how a stray sample could escape its natural position on the grid under random initializations, which is the result of the configuration in recent time landing in a local minimum due to the noisy rates there.

CoaSTT jointly reveals continuous and discrete structure

Realistic populations are expected to exhibit both discrete and continuous modes of structure concurrently. For example, a population consisting of two groups separated by a geographical barrier (such as a river or a desert) will exhibit discrete structure between the subpopulations on opposite sides of the barrier, and continuous spatial structure within each subpopulation on either side. Simultaneously resolving discrete and continuous structure of such a popu-

lation is challenging with current methods. We demonstrate CoaSTT's unique potential to reveal the history of such a population using the simulation depicted in figure 2.16A. An ancestral population (of size $N = 5000$) split 5000 generations ago into two subpopulations, each spatially structured into 5×5 demes ($N = 100$ per deme), with a constant migration rate between neighboring demes. The migration rate between demes in group B is higher than that of group A. Additionally, each pair of demes from the two different groups exchanges migrants at a small constant rate following the split. CoaSTT successfully reveals discrete and continuous spatial structure jointly. Two clusters form at recent time, and they get closer to one another moving backward in time as a result of the weak between-group migration (figure 2.16B). Meanwhile, each cluster is continuously structured as a result of the spatial structure. At a recent time point (150 generations ago), both clusters are internally well-structured into a grid of points. A characteristic of such continuous structure is that pairs of samples from distant demes (with coalescent rates that are still too low) are pulled together as a result of the accumulation of high coalescent rates between the adjacent demes that connect the two distant samples, resulting in reconstructed rates that are much higher than the input rates. By 1500 generations ago, the group with a higher between-deme migration rate has a well-mixed cluster (with similar pairwise rates between all samples within the cluster), while the other group is still structured. Both clusters become well-mixed close to the split time (4500 generations), with a much smaller distance between them compared to recent time. Beyond the split time, the samples are arranged into a single well-mixed cluster corresponding to the ancestral population.

This particular example highlights an issue regarding the smoothness of sample trajectories in latent space. We described in section 2.3.3 how sequential initialization, Procrustes alignment between time slices and a regularization term that penalizes k^{th}-order derivatives are different strategies that ensure the smoothness of latent trajectories despite the noisy rate estimates and the invariance of the loss function with respect to rotation/translation of

sample configurations. This example demonstrates an interesting case where these strategies might not be sufficient. The strategies described seem to preserve the global configuration (group A on one side and group B on the other), as well as prevent erratic changes in the trajectory of a sample. However, the local configuration of each group seems to be continuously rotating through time. Owing to the symmetry of the problem, the grid configuration of the embeddings of each group has rotational freedom. Even with sequential initialization, random noise will cause each group to slightly rotate by a different angle, such that no single Procrustes rotation matrix can align the embeddings of two subsequent time points. Even with a regularization term that penalizes the first- and second-order differences (see Appendix 2C for exact parameters), the rotation of each cluster is not completely removed, but is smoothed out over longer timescales. This highlights the value of inspecting individual time slices (as in figure 2.16C) in addition to the 3D embeddings through time, and illustrates how equivalent solutions (e.g. rotationally equivalent in this case) could appear over different time slices despite regularization. This is similar to the fact that samples within a well-mixed configuration will get shuffled over time as a result of all permutations of samples within the configuration being equivalent (with noise slightly causing one permutation to be preferred). Such shuffling of samples between two time points visually demonstrates that all these solutions are valid, whereas forcibly fixing a specific permutation of samples (or a rotational configuration in this case) will yield clean trajectories, but fail to express the equivalence of the other solutions over time.

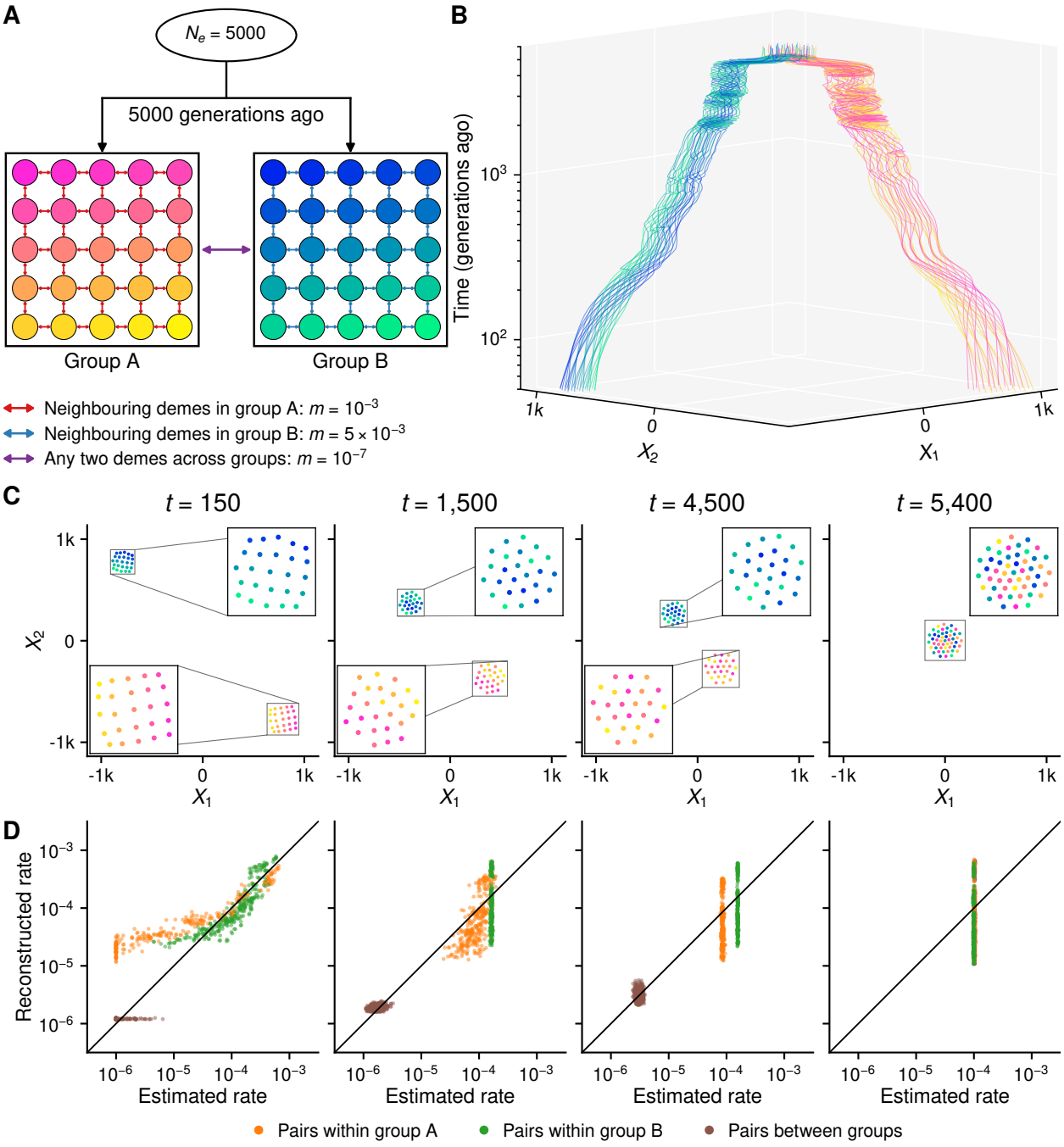

Figure 2.16: CoaSTT jointly reveals discrete and continuous coalescent structure. A) Schematic of the simulation. A single population split into two groups, each consisting of a 5×5 grid of smaller demes, with migration occurring between neighboring demes at a constant rate, as well as weaker migration between demes of different groups. **B)** 3D CoaSTT embeddings with rates estimated from simulated trees. **C)** 2D Embeddings at four representative time slices. **D)** Plots of input rates (estimated from trees) vs. reconstructed rates (calculated from latent positions) at each of these times.

2.5.4 CoaSTT applied to region-specific coalescent rates can resolve the history of admixed individuals

Applying PCA to a dataset that includes samples from an admixed population and from the corresponding source populations is known to result in two separate clusters for the source populations, with the admixed samples projecting between them in PC space (G. McVean, 2009). The proximity of the admixed samples to the source-population clusters is indicative of admixture proportions. If the admixed population itself is structured, PCA (or UMAP) might fail to reveal this fine-scale structure. This has been addressed by developing ancestry-specific variants of PCA (Moreno-Estrada et al., 2013), where only the SNPs lying within genomic regions that have been assigned a particular local ancestry through local-ancestry inference (LAI) methods are included in the analysis. The idea of exploring lower-dimensional representations of genomic data that include admixed individuals using ancestry-specific segments of their genome has been invoked to adapt a few methods to allow them to better resolve population structure for such populations (Brisbin et al., 2012; S. R. Browning et al., 2016; Tang & Chiang, 2025).

The behavior of CoaSTT when applied to ARGs containing admixed samples is similar to that of PCA, but in a time-resolved manner (figure 2.17). To illustrate this, CoaSTT embeddings were computed for samples simulated under a demography with two distinct admixture events, one relatively recent (occurring 100 generations ago), and another occurring at a deeper time (500 generations ago) and involving two distinct source populations (figure 2.17A). At a time more recent than either admixture event, each set of admixed samples forms its own cluster (as they only coalesce within their group). At a time deeper than one of the admixture events, the corresponding admixed cluster is pulled to lie between the clusters of its source populations (figure 2.17B). This reflects the true coalescent history, since, during this period, the admixed samples have a non-zero genome-wide coalescent rate

with samples from both source populations.

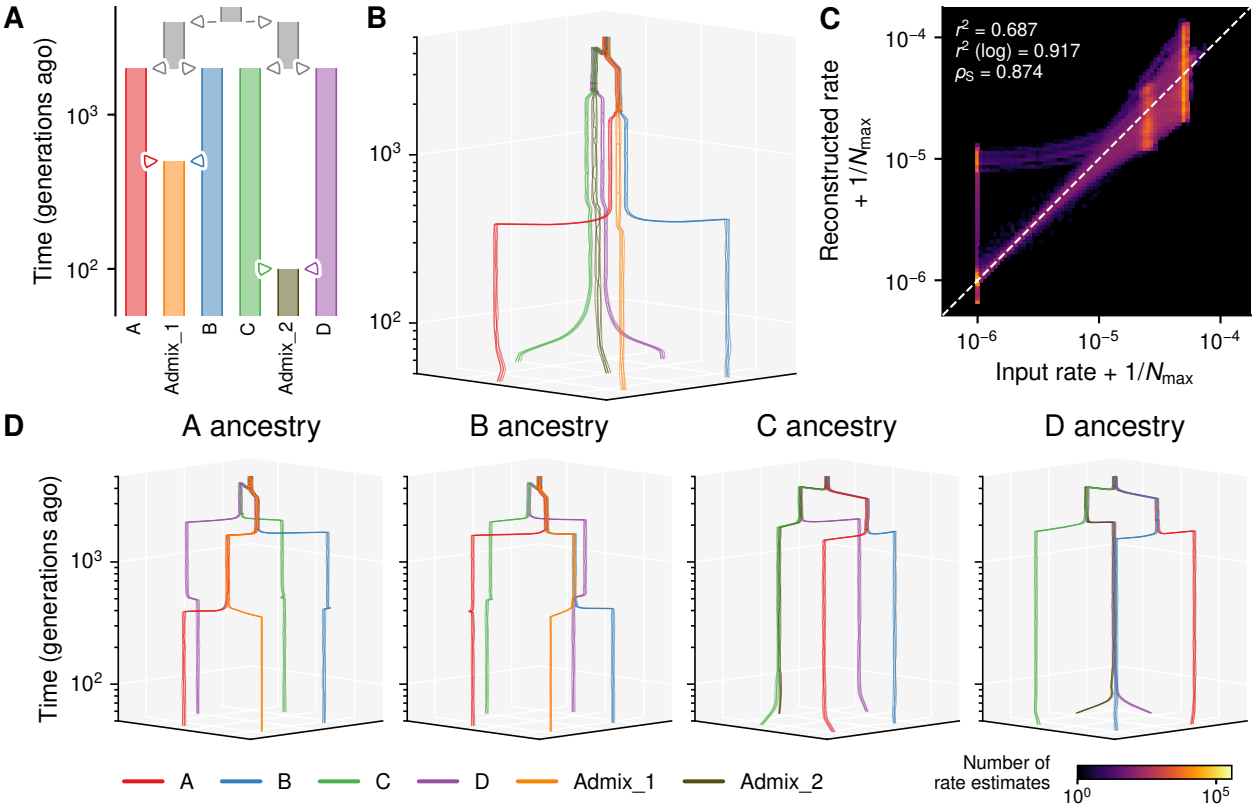

Figure 2.17: CoaSTT applied to a simulation with two distinct admixture events. A) A schematic of a simulated demography with two different admixture events occurring at different times. Each admixture event involves a different pair of source populations. **B)** GW-CoaSTT embeddings. Each admixed cluster lies between the source-population clusters during $t_{\text{admix}} < t < t_{\text{split}}$, pulling both source populations closer together, which is reflected as inflated reconstructed rates in **(C)**. **D)** RS-CoaSTT embeddings resolve the alternative coalescent histories of each set of admixed samples, with the source populations no longer getting pulled towards each other.

This representation of history suffers from two fundamental drawbacks. First, the source populations are themselves drawn to one another as the admixed cluster pulls them both to itself, potentially giving a false indication of gene-flow occurring between them. This is a direct result of the transitivity assumption being violated under this scenario. The transitivity assumption is a foundational assumption in LPMs, where nodes (A, C) are assumed to be connected with high probability if a connection exists between nodes (A, B) and (B, C). In our case, if samples i and j belong to each source population respectively,

and k is from the admixed population, (i, k) and (j, k) will have a substantial coalescent rate (during $t_{\text{admix}} < t < t_{\text{split}}$), whereas (i, j) are not coalescing at all. This can be seen with reconstructed rates that are much higher than their respective input rate estimates (figure 2.17C). Many rate estimates that are identically zero (or very small) have corresponding reconstructed rates that are an order of magnitude higher. These estimates correspond to pairs of samples from different source populations in the period $t_{\text{admix}} < t < t_{\text{split}}$. The second drawback is that such embeddings fail to communicate the alternative coalescent histories of admixed samples. A more accurate depiction would ideally represent each admixed sample by several trajectories rather than one, each reflecting one distinct coalescent history.

We address these issues by adapting CoaSTT in a manner similar to AS-PCA. The normal use case of CoaSTT uses genome-wide estimates of pairwise coalescent rates as input, which results in a genome-wide representation of coalescent structure through time (genome-wide CoaSTT, or GW-CoaSTT). With admixed samples, we could alternatively feed in coalescent rates computed for specific genomic regions (region-specific CoaSTT, or RS-CoaSTT). Specifically, the user specifies a set of samples that are considered to be "admixed". Genomic regions of the admixed samples are then assigned an ancestry label based on some local-ancestry inference method. Rather than estimating a single rate trajectory for a pair of individuals where one of them is admixed, K rate trajectories are computed for this pair, one for each ancestry label. In computing the estimate corresponding to a given ancestry label, coalescent events are only counted within genomic regions where the admixed sample is assigned that ancestry label. Consequently, rather than having a single set of coalescent rate estimates, we end up with K such sets (only for pairs where one of the samples is deemed admixed, whereas all other pairs retain identical coalescent rates in all sets). Each set is used as an input to compute CoaSTT embeddings, resulting in K sets of RS-CoaSTT embeddings. Since only a portion of the genome is used for rate estimation, this portion has to carry sufficient coalescent information or else rate estimates will be excessively noisy. If

an admixed sample has a fraction of its genome assigned to some ancestry that is lower than some threshold (default = 0.1), the rates (and consequently embeddings) specific to that ancestry will not be computed for this admixed sample. This means that not all admixed samples will be present in all RS-CoaSTT embeddings, but only ones that carry a sufficient proportion of that ancestry in their genome.

Figure 2.17D shows the RS-CoaSTT embeddings where the four source populations are regarded as the "reference" populations, and admixed samples are assigned local ancestry given these four references. We therefore end up with four sets of embeddings through time. The true local ancestry can be found in simulations by tracing which population the ancestral lineage of an admixed sample at some locus belonged to prior to the admixture event (using `tskit` methods). With these true local-ancestry assignments, each set of embeddings shows the alternative history for one of the admixed clusters (the other is omitted because it does not have a sufficient proportion of that ancestry), where the admixed cluster merges with one of its source populations at $t = t_{\text{admix}}$. Moreover, the source populations are no longer pulled together at $t = t_{\text{admix}}$. Additionally, true local-ancestry assignment to admixed samples addresses the mismatch between true and reconstructed rates between the source populations.

In practice, we do not have access to true local ancestry assignments, which are rather inferred through some local-ancestry inference method. We compare the ancestry-specific embeddings that arise given two strategies of assigning local ancestry: an ARG-based strategy based on the genealogical nearest-neighbors (GNNs) (Wohns et al., 2022), and a state-of-the-art HMM-based method (RFmix version 2) (Maples et al., 2013). The GNNs of a sample at some locus are defined as the set of samples that descend from the lineage that the focal sample first coalesces with at that locus. Note that a single sample will have multiple GNNs in a single tree if it first coalesces with a lineage that has multiple descendants. Local ancestry can be assigned through the population labels of GNNs. For instance, if

80% of the GNNs of an admixed sample at tree l have label A and 20% have label B, then the admixed sample is assigned label A at that locus. This is a rough method to assign local ancestry straight from the ARG, and might need eventual refinement, for instance, by modeling these proportions as emission probabilities of some HMM. RFmix is a popular method that trains a Random Forest classifier on each genomic segment given a labeled set of reference samples. Ancestry switches are properly modeled through an HMM along the genome, where the RF probabilities are treated as the emissions. One advantage of using GNN-based inference is that it is ARG-based, allowing the local-ancestry assignment step to be seamlessly incorporated into CoaSTT's pipeline without the need for external software and tuning.

Figure 2.18A shows how RS-CoaSTT embeddings with GNN-based local-ancestry assignments closely match the embeddings with true assignments. One issue that is not directly apparent from the 3D plots is the imperfect mixing of the admixed cluster with the rest of the samples at $t > t_{\text{split}}$, when there is a single well-mixed ancestral population. This is a direct result of the manner by which mis-assignment of local ancestry segments using GNNs is biased. In GNN-based local ancestry assignment, a genomic segment will be assigned the wrong source population if it first coalesces with a node whose descendants mostly belong to the wrong source. This occurs occasionally due to incomplete lineage sorting (ILS), where the time when the admixed sample first coalesces with another lineage is deeper than t_{split}. Mis-assigned segments will therefore have a higher probability of having deeper coalescent times. As a result, coalescent rates between the admixed samples and samples of the incorrectly assigned source population are biased upward at $t > t_{\text{split}}$, whereas rates with samples from the population of the correct local ancestry are biased downward. This bias eventually fades away, pushing the time where samples are well-mixed (all having an equal rate) past the true split time (figure 2.18B).

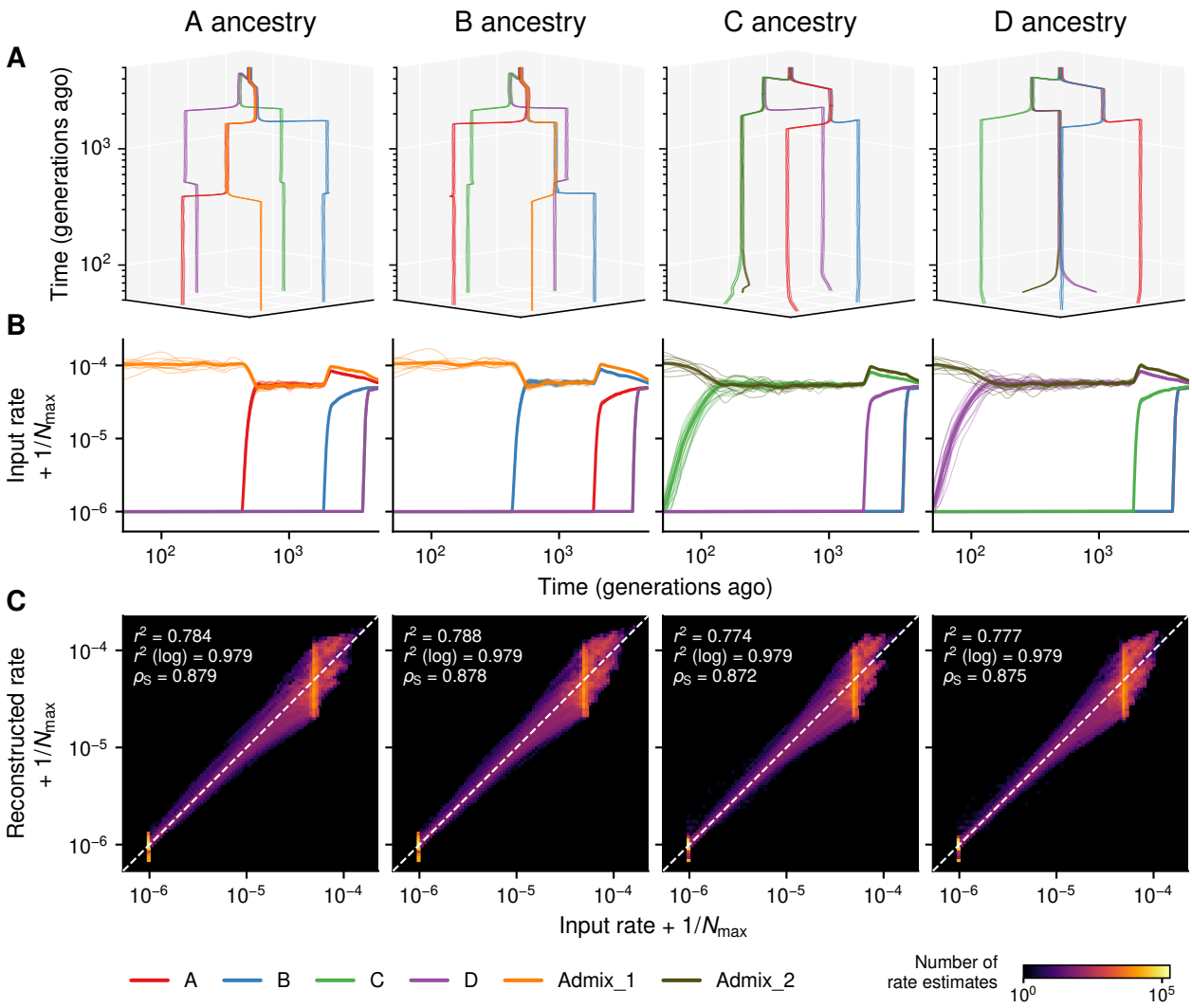

Figure 2.18: RS-CoaSTT embeddings with GNN-based local-ancestry assignments. A) Embeddings estimated using region-specific input coalescent rates for admixed samples. As a reminder, the population labeled 'Admix_1' has 'A' and 'B' as sources, and 'Admix_2' has 'C' and 'D' as sources, with a more recent admixture time. A genomic region of an admixed sample is assigned one out of four labels based on its genealogical nearest neighbors in this region. **B)** Region-specific estimates of coalescent rates as a function of time for all pairs of samples where one of the samples in the pair is admixed, colored by the population label of the other sample, with the mean rate shown in bold. Rates do not converge to a single value at $t > t_{\text{split}}$ right away due to the biased mis-assignment of local ancestry. **C)** Region-specific rate estimates vs. rates reconstructed from the region-specific embeddings, showing less inflated reconstructed rates compared to GW-CoaSTT.

Incorrect local-ancestry assignment arising from RFmix results in a different kind of

bias. Segments that are assigned the incorrect ancestry are usually contiguous extensions of correctly assigned segments, or completely missed short segments, with a higher rate of incorrect assignment the deeper the admixture time (since ancestral genomic segments are broken up into smaller segments due to many rounds of recombination, making the HMM-based assignment problem more challenging). The effect of this on RS-CoaSTT embeddings is that source populations still get slightly pulled together around $t = t_{\text{admix}}$ (figure 2.19A). This is a result of segments assigned an incorrect local ancestry coalescing with samples from the population of the true ancestry (figure 2.19B). Therefore, reconstructed rates are still inflated, albeit less severely for more recent admixture events (figure 2.19C). Interestingly, an additional group of deflated reconstructed rates appears, with samples being further apart than they should. In the region-specific embeddings of a given source population, this corresponds to rates between the admixed samples and samples of the other source population during $t_{\text{admix}} < t < t_{\text{split}}$. These rates are actually incorrectly biased upwards, as they should be exactly zero during that time period. This is an example where the inconsistency between input rates and reconstructed rates is actually good, where the reconstructed rates are closer to the truth than the RFmix-based input rates.

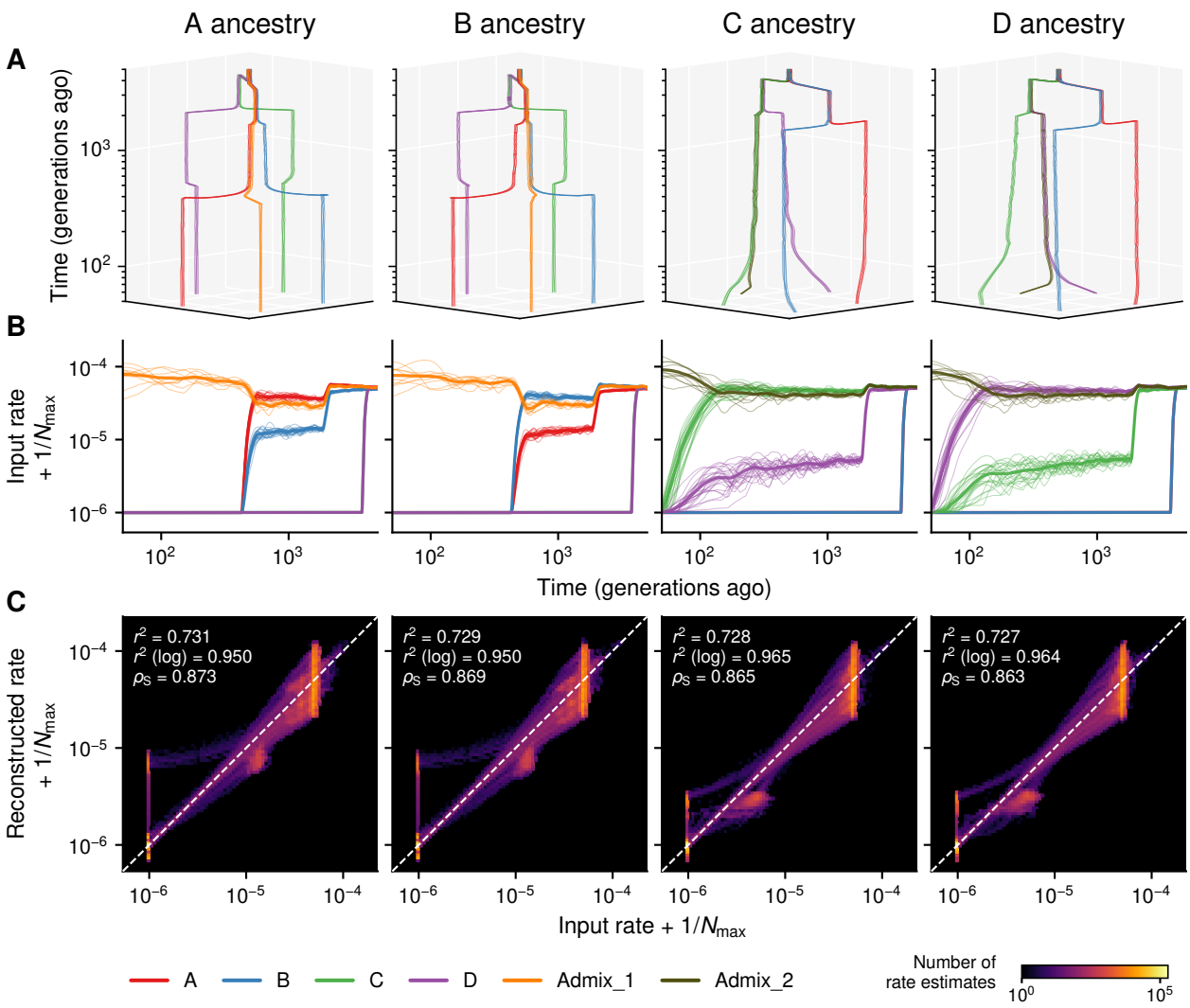

Figure 2.19: RS-CoaSTT embeddings for the two-admixture simulation with RFmix local-ancestry assignments. A) Source populations are still pulled towards each other at $t = t_{\text{admix}}$, but less severely than using genome-wide rates. **B)** Region-specific coalescent rates are biased upward during $t_{\text{admix}} < t < t_{\text{split}}$ between admixed samples and samples from the other source population (e.g. admix-B rates for rates computed using A ancestry segments). **C)** Input vs. reconstructed rates computed from the region-specific embeddings. Reconstructed rates are inflated less severely, especially between the source populations that participated in the more recent admixture event.

RS-CoaSTT applied to BADMIXTURE simulation

We end this section by exploring the results of applying RS-CoaSTT to three challenging case studies proposed by Lawson et al., 2018: Recent three-way admixture, ghost admixture

(where one of the source populations is not sampled), and a recent population split with a sharp bottleneck (figure 2.20A). What makes these scenarios interesting is that they all show almost identical results when ADMIXTURE is applied to them. Each of the three examples represents a relevant realistic admixture scenario. The first represents traditional admixture, and is similar in spirit to the demography shown in figure 2.17A, but with three source populations rather than two, and uneven admixture proportions (50%, 35% and 15% respectively). The results of GW-CoaSTT and RS-CoaSTT applied to samples from this demography are not expected to differ much from the previous example. Ghost admixture is a prevalent case in real genomic datasets, where samples from one of the source populations are not available. Therefore, the genomic segments of admixed samples that originate from the unsampled ghost will be assigned the label of a sampled reference population that is ancestrally closest to the source (population C in our case). The final example is a case where population A is considered by ADMIXTURE to be an admixed population when it is not. It originates from a single source population, that which is ancestral to B. Therefore, any genomic segments of A samples that are assigned local ancestry C or D will be purely a result of ILS. In this case, GW-CoaSTT embeddings represent the correct coalescent history, and RS-CoaSTT risks representing an inaccurate history.

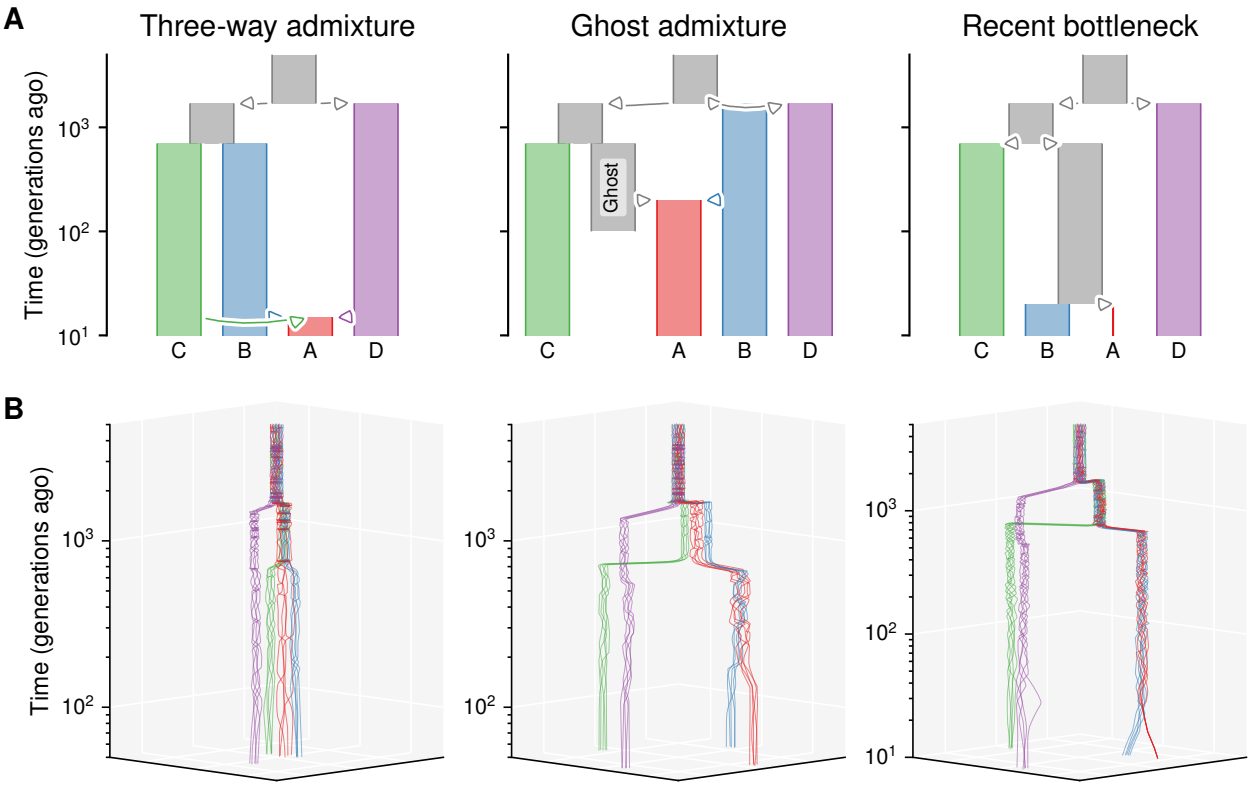

Figure 2.20: GW-CoaSTT applied to three BADMIXTURE simulations that give identical ADMIXTURE results. A) The demographic history for each example: Three-way admixture, ghost admixture and recent bottleneck. Samples from population A (in red) show identical ADMIXTURE results under the three scenarios. **B)** GW-CoaSTT embeddings computed using genome-wide rates are different for the three scenarios. In the first scenario, the admixed samples lie in between the source-population clusters, pulled closer to the clusters from which they draw higher admixture proportions (B & C). In the second scenario, one of the source populations is not sampled. Therefore, the admixed population is drawn only to population B (the sampled source) around the admixture time. At the time where the ghost population and population C merge, the admixed samples are pulled by the C cluster as they start coalescing with samples from C. The contribution of the ghost population to the admixed samples is detected through coalescence with C lineages. In the last example of recent bottleneck, GW-CoaSTT gives the expected embeddings from a population split with size changes.

Figures 2.21 and 2.22 show RS-CoaSTT embeddings when local ancestry is inferred for samples of population A using GNNs and RFmix respectively, with populations (B,C,D) defined as references in the three examples. The three-way admixture case gives results similar to the example with two two-way admixture events (figures 2.18 and 2.19), where

region-specific embeddings for each ancestry show the admixed cluster positioned close to the cluster of the source population of that ancestry. In a similar vein to the previous results, embeddings from GNN-based local ancestry assignment exhibit imperfect mixing in deep time, while those based on RFmix show the source-population clusters getting pulled towards one another through the effect of coalescing with samples from the admixed cluster.

In the ghost admixture scenario, the distinct coalescent histories are revealed despite the lack of samples from one of the source populations in B-specific and C-specific embeddings. The admixed cluster merges with population B at $t = t_{\text{admix}}$ in the B-specific embeddings (in both GNN and RFmix examples), while it merges with population C in the C-specific embeddings at the deeper time when the ghost source and population C share a common ancestral population. This shows that even though the segments derived from the ghost were incorrectly assigned population C as the source, their coalescent history is accurately reflected, showing that the time where they start coalescing with samples from C is different from the time where the B-specific segments start coalescing with the samples from B. The approach of clusters A and B in C-specific embeddings for the RFmix case is due to segments that truly derive from B being assigned ancestry C, a theme that recurs for RFmix-based embeddings. D-specific embeddings are only present for the ghost admixture example in the GNN case, reflecting ILS segments as discussed, and are absent in the RFmix case as their proportion does not pass the threshold.

The recent bottleneck example is an interesting one, where population A is considered to be admixed when it is not, and local ancestry is called on its samples with [B,C,D] as references. In truth, all genomic segments of samples from population A derive from a population ancestral to B. Therefore, any segments assigned ancestry C or D will be due to ILS. Indeed, B-specific embeddings in both GNN and RFmix cases are close to the whole-genome embeddings that represent the correct coalescent history. The C-specific and D-specific embeddings reflect the history of ILS segments of the A samples, which is recognizable

when compared with the whole-genome embeddings. Cluster A merges with clusters C/D in the C-specific and D-specific embeddings, respectively, at the same time as it does in the whole-genome embeddings. Again, cluster A is absent in the D-specific embeddings of the RFmix case due to an insufficient ancestry proportion. More accurately assigned local ancestry (that properly identifies ILS segments) should result in cluster A being only present in the B-specific embeddings (matching the whole-genome embeddings), and absent from both C- and D-specific embeddings.

These results demonstrate the potential of CoaSTT to resolve the history of admixed samples, while highlighting the need to couple it with more refined LAI methods for more reliable results. GNN-based LAI is promising and ARG-native, and already shows reasonable results when applied to simulated data, even with a naive implementation that does not correctly model probabilities of ancestry switches. A couple of supervised ML methods were recently developed for GNN-based LAI (Pearson & Durbin, 2023; Shanks et al., 2026). However, the models are trained on simulations from a single demographic model, as they were intended to be applied to understand particular admixture scenarios in human history. That said, these methods demonstrate the potential of GNN-based LAI, and pave the way for more generalizable methods for LAI using ARGs. Ideally, such methods should model ILS probability and local ancestry switches along the genome in a formal manner, reducing the rate of incorrect local ancestry assignment, and providing an additional measure for the confidence of each assignment. With the naive GNN-based local ancestry assignment, RS-CoaSTT specifically shows the coalescent history of the genomic regions of admixed samples that first coalesced with some other group of samples. These regions either coalesce first with this group because the ancestral population of the group was indeed the source population in an admixture event, or because of ILS. A more refined implementation of GNN-based local ancestry that identifies these ILS regions more accurately, and either removes them from the rate-estimation step or reassigns them the correct local ancestry, would prove useful in

revealing the alternative histories of admixed samples using RS-CoaSTT.

We end this section with an ambitious question: Is it possible to merge all RS-CoaSTT embeddings into a single set of embeddings, where each admixed sample is represented by a single trajectory during $t < t_{\text{admix}}$, that then split into K different trajectories appropriately between $t_{\text{admix}} < t < t_{\text{split}}$? It is perhaps possible for simple admixture events where t_{admix} is known, where this splitting can be implemented manually. However, if t_{admix} is unknown or if the admixture history is more complex, it would be challenging to obtain such a representation using the current CoaSTT framework. If we attempt to represent each admixed sample by K points in the latent space, the admixed cluster will be separated into K clusters during $t < t_{\text{admix}}$. There are essentially no coalescent events to pull these clusters together. The reason is the current implementation of obtaining coalescent rate estimates, where coalescent events are only counted if they occur at the same locus. It is possible for a coalescent event to occur between two samples at two non-overlapping loci, but such events are difficult to pinpoint with current ARG-inference methods, and are excluded from the SMC approximation, which our coalescent rate estimates are based on. A more promising alternative is to develop a variant of CoaSTT that incorporates the alternative histories of admixed samples in a more natural way, perhaps by invoking Latent Position Cluster Models (LPCMs) (Handcock et al., 2007; Sewell & Chen, 2017), where a Gaussian mixture prior is placed on the latent position of each sample, and where inflated reconstructed rates can be used as a hint for identifying sample positions that need to be modeled as a mixture.

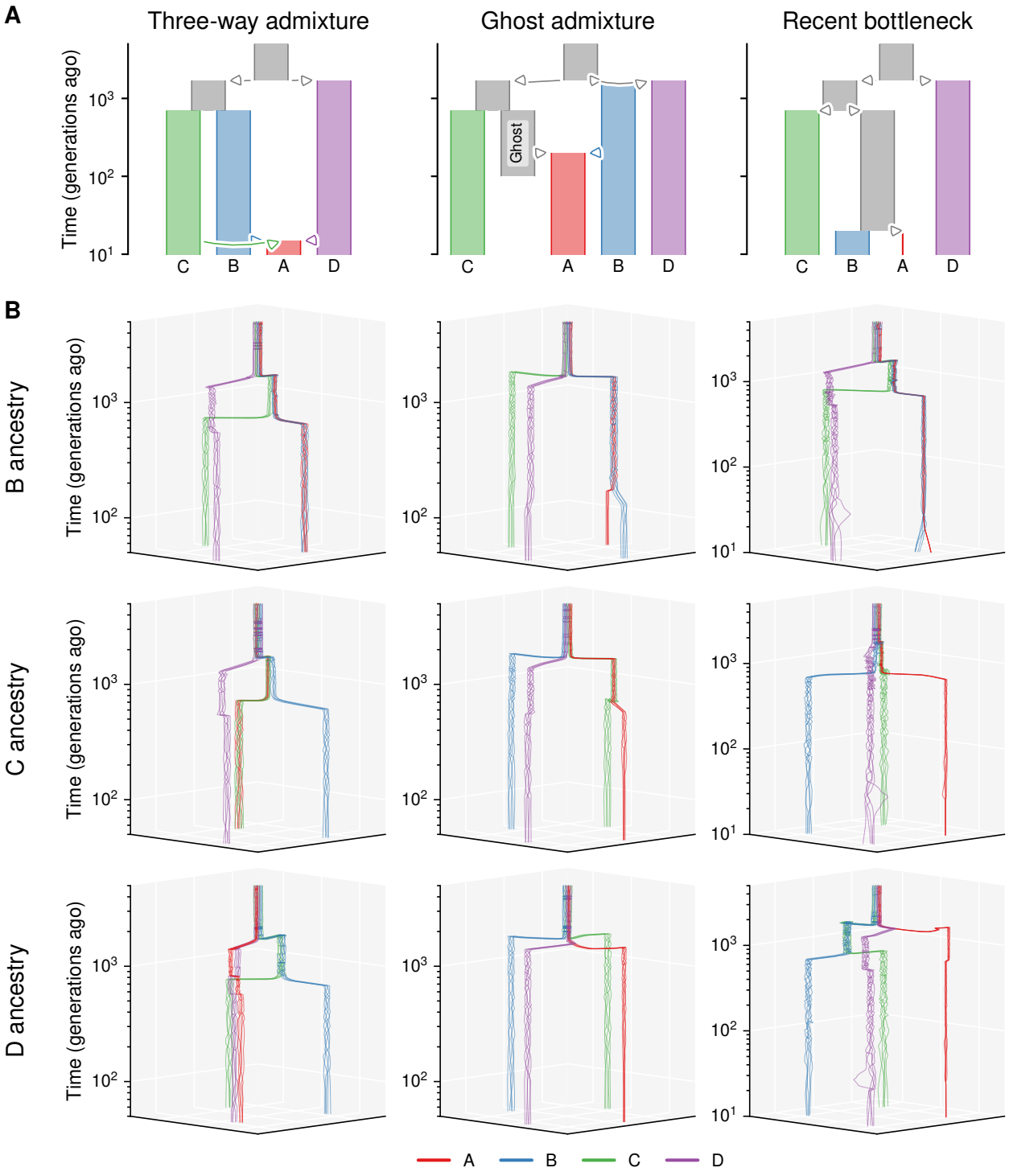

Figure 2.21: RS-CoaSTT embeddings for the BADMIXTURE examples using GNN-based local-ancestry assignment. Each column represents a demographic scenario, and each row shows the region-specific embeddings for that scenario.

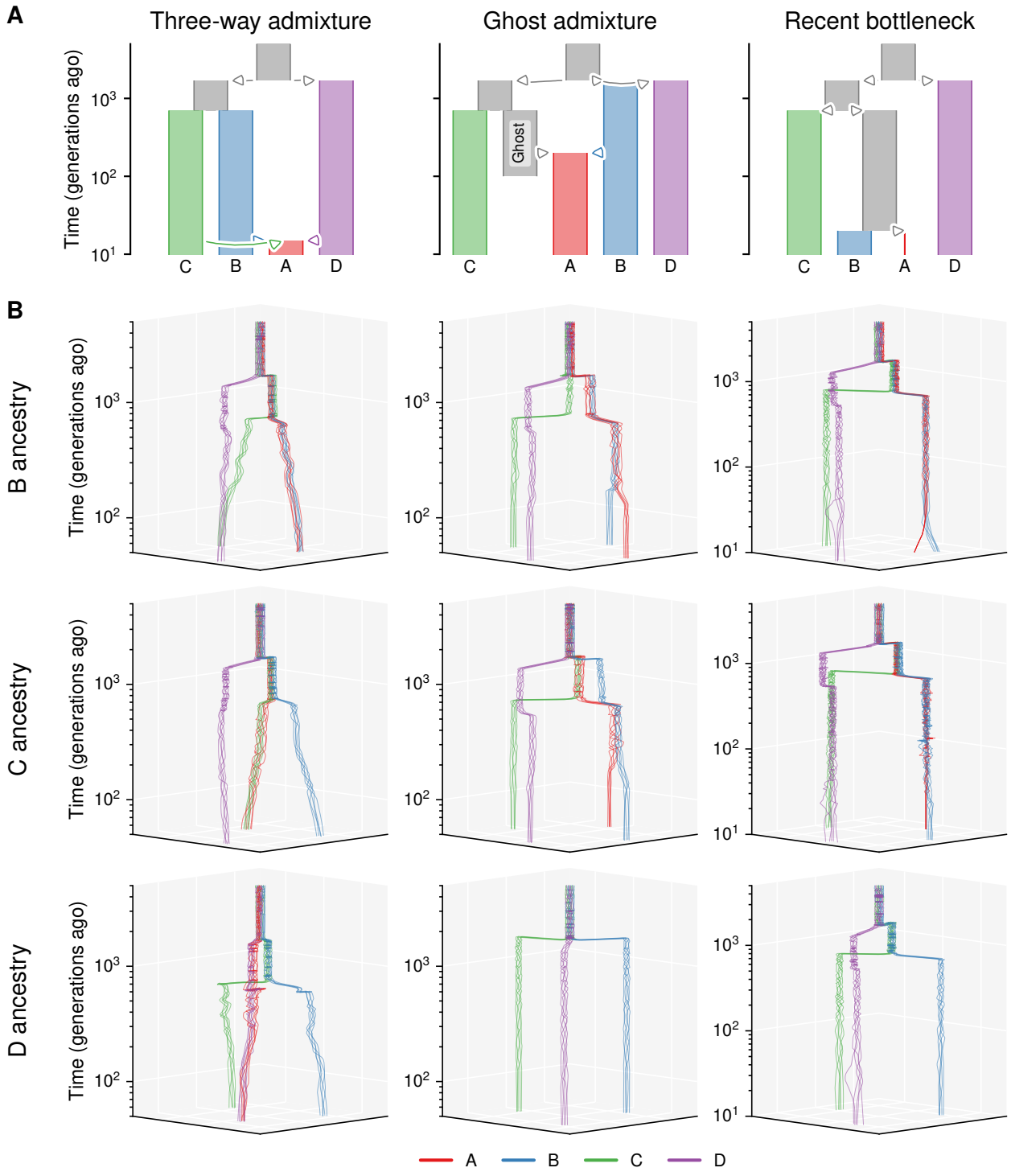

Figure 2.22: RS-CoaSTT embeddings for the BADMIXTURE examples with local ancestry inferred using RFmix. Each column represents a demographic scenario, and each row shows the region-specific embeddings for that scenario.

2.6 Discussion

The ability to express history in terms of coalescent rates has been one of the most prominent outcomes of the sequentially Markovian coalescent (SMC) approximation of the coalescent with recombination (Hudson, 1983; H. Li & Durbin, 2011; Marjoram & Wall, 2006; G. A. T. McVean & Cardin, 2005). Whereas such descriptions are usually given in terms of coalescent rates within and between populations, we develop a method to visualize coalescent structure through time at the individual or haplotype level. The development of efficient ARG inference methods is the main factor that allows coalescent rates to be estimated between each pair of samples without reference to predefined populations (Nielsen et al., 2025), and the encoding of ARGs in the tree-sequence data structure additionally allows these estimates to be obtained efficiently (Kelleher et al., 2018). CoaSTT can therefore be regarded as a natural extension of methods for understanding population history in the era of ARGs.

CoaSTT embeddings are computed by minimizing a loss function that is derived from standard coalescent theory and inspired by latent position models (Hoff et al., 2002). Since it is not immediately obvious what CoaSTT embeddings will look like under different scenarios of population history, we explored the embeddings that emerge across a wide range of simulated scenarios. These examples provide guidance on interpreting embeddings estimated from empirical data, whose analysis is the subject of the next chapter. The results show that CoaSTT reproduces a discrete depiction of coalescent history when that history is indeed described by discrete populations, and a more continuous description when ancestral relationships are better described as such. Importantly, this description pertains to coalescent history as opposed to demographic history. This is most clearly illustrated by the stepping-stone examples, where the demographic history is unchanging (K demes of constant population size exchanging migrants with their nearest neighbors at a constant rate), but the coalescent history gradually changes backward in time until all pairs of samples share the

same coalescent rate. Coalescent history cannot be directly translated into specific historical processes, as different historical processes can result in the same profile of pairwise coalescent rates. It is nevertheless a useful summary of the history of relationships between the genetic ancestors of a set of samples. We additionally provide an extension to CoaSTT that takes as input rates computed from a subset of genomic regions, offering a way to explore the alternative coalescent histories of putatively admixed individuals.

Having tested CoaSTT across a range of simulated scenarios, we turn in the next chapter to ARGs inferred from empirical data. There, we show how CoaSTT embeddings can shed light on different aspects of the history of natural populations.

Appendix 2A: Loss functions of non-linear embedding methods and their relationship to CoaSTT

Where does CoaSTT sit with respect to other non-linear dimensionality reduction methods? Answering this question is not only insightful for situating our method within the broader literature of non-linear embedding methods; it also allows us to draw insight from the extensive literature of these well-studied methods regarding optimization, the effect of link function choice, as well as general intuition around global and local properties of such embeddings.

t-SNE (Maaten & Hinton, 2008) and UMAP (McInnes et al., 2020) are two of the most prominent non-linear embedding methods for visualizing high-dimensional biological data. They each have their distinct way of processing the input data, as well as different modeling assumptions from which they derive their loss functions. They both belong to the class of "Neighbor-Embedding" methods that aim to find lower-dimensional representations of data that preserve neighbor relationships, primarily for visualization purposes. Regardless of their particular details, what concerns us here is that both methods process the data in some way to yield a similarity matrix that is used as an input to some loss function. Similar to our case, the problem reduces to minimizing the loss with respect to the low-dimensional embeddings. A remarkable observation is that both their loss functions, as well as the loss functions of many other related methods, belong to the same class of "Attraction-Repulsion potentials" (Böhm et al., 2022). This fact has actually been realized early on and inspired the development of Elastic-Embeddings (EE), a method related to t-SNE that emerged shortly after (Carreira-Perpiñan, 2010). Loss functions derived for LPMs, including CoaSTT's loss, belong to the same class. For a similarity matrix with elements s_{ij}, and a general link function f_{ij}, such loss functions can be generally written as

$$l(X) = g(s_{ij}, f_{ij}) + r(s_{ij}, f_{ij}) \tag{2.47}$$

The first term is an attractive term that is larger for pairs with larger similarity s_{ij}, and has a gradient that effectively results in an attraction between pairs of particles. The second is a repulsive term that is weaker for stronger similarity (or usually not depending on similarity at all), and results in a repulsive gradient. Table 2.2 shows the link functions as well as the attractive and repulsive terms in the loss functions of t-SNE, UMAP, Hoff's LPM (Hoff et al., 2002) and CoaSTT.

Table 2.2: Comparison between different embedding methods

	Link	Attractive term	Repulsive term
t-SNE	$(1+x^2)^{-1}$	$-\sum_{i,j} s_{ij} \ln f_{ij}$	$\ln\left(\sum_{i,j} f_{ij}\right)$
UMAP	$(1+ax^{2b})^{-1}$	$-\sum_{i,j} s_{ij} \ln f_{ij}$	$-\sum_{i,j}(1-s_{ij}) \ln(1-f_{ij})$
Hoff's LPM	$(1+\alpha e^{x})^{-1}$	$-\sum_{i,j} s_{ij} \ln f_{ij}$	$-\sum_{i,j}(1-s_{ij}) \ln(1-f_{ij})$
CoaSTT	$(N_{\min}+x^{p})^{-1}$	$-\sum_{i,j} w_{ij} s_{ij} \ln f_{ij}$	$\sum_{i,j} w_{ij} f_{ij}$

In the above comparison, we used the same notation for the similarity matrix used by each method (s_{ij}) regardless of how it is obtained. Interestingly, CoaSTT's loss function is very similar to UMAP's for small similarities (which is the case for coalescent rates), with additional weights w_{ij}. Therefore, much of the analysis of UMAP's performance transfers here, such as UMAP's sensitivity to initialization, challenges with optimizing the non-convex loss, and the effect of the heaviness of the link function's tail. The only difference between our loss and t-SNE's is the repulsive term. In t-SNE, the sum over pairs is inside the log, so the repulsive force felt by sample i is a function of all pairwise distances, whereas in our case (and UMAP's), the repulsive force felt by sample i depends only on its distances to all other samples, not on other unrelated pairwise distances. In fact, this has been shown to be the primary difference between t-SNE and UMAP (Draganov et al., 2023).

Loss functions with such attractive and repulsive terms can be highly non-convex in latent positions, inducing a high risk of converging to bad local minima resulting in configurations

that do not reflect real structure. They are also usually applied to very large datasets for which computing exact gradients is computationally expensive. The speed problem is addressed in the NE literature through optimization tricks like the Barnes-Hut approximation (t-SNE) and edge sampling and negative sampling (UMAP), with all tricks designed to avoid iterating over all pairs when computing the gradient. However, the speed issue is not as relevant in our case since we have an earlier bottleneck in the pipeline at the step of estimating coalescent rates from ARGs, limiting our sample sizes to $n < 10^4$. Nevertheless, bad local minima can be an issue, and this has been addressed in the literature through early exaggeration (increasing attractive forces in t-SNE early on), spectral initialization (initializing with Laplacian Eigenmaps in UMAP) and homotopy methods (gradually increasing the strength of the repulsive force in Elastic Embeddings). It has been shown that good initialization can have a substantial effect on retrieving global structure and can be as effective as other more complex optimization tricks (Draganov et al., 2023; Kobak & Linderman, 2021). The standard approach in CoaSTT is sequential initialization, initializing each time slice with the solution of the slice before it. However, we want to extend this default setting to allow for time slices that are too far apart (such that this initialization is no longer reliable), or the case where a solution for an intermediate time slice is needed without access to the one before it.

Initialization

Suppose we want to find the latent embeddings for an intermediate time slice for which we have no previous solution or a solution that is too far off the current slice's minimum. Since our loss's structure is similar to UMAP's, we follow its standard practice and use the Laplacian Eigenmap (LE) as the initial solution (Belkin & Niyogi, 2003). LEs minimize a

loss given by

$$l(X) = \sum_{ij} W_{ij} \|X_i - X_j\|^2 = \mathrm{tr}(X^T L X) \tag{2.48}$$

under the constraint $X^T D X = I$, where $D_{ii} = \sum_j W_{ij}$ is the diagonal degree matrix, and L is the Laplacian $L = D - W$. This is essentially an attraction-only loss with a Gaussian link, with repulsion maintained by the constraint. As mentioned in section 2.3.5, a Gaussian link is good for resolving local structure, ensuring that the optimizer starts with a configuration where the local part of the structure is already resolved. In our case, we construct the weight matrix as $W_{ij} = \hat{w}_{ij}\hat{\lambda}_{ij}$. Solutions to LEs are simply the eigenvectors of L corresponding to the smallest non-zero eigenvalues. Since eigenvectors can have rotational or sign ambiguities, we provide the option of initializing the eigenvector solver with a solution, which could be the solution of a previous yet distant time point. We obtain the eigenvectors through `scipy.sparse.linalg.lobpcg`, which is an iterative solver that allows an initial guess, or `scipy.linalg.eigh` if no initialization is provided (Virtanen et al., 2020).

Gradient & Hessian using Laplacians

Recall that the loss in equation 2.15 for a single time point can be written as

$$l(X) = \underbrace{\sum_{i,j} \hat{w}_{ij} f_{ij}}_{\text{repulsive potential}} \underbrace{- \sum_{i,j} \hat{w}_{ij} \hat{\lambda}_{ij} \ln f_{ij}}_{\text{attractive potential}} \tag{2.49}$$

We introduce the following notation that allows us to write the gradient and the Hessian concisely:

Table 2.3: Notation for writing the gradient and the Hessian

Notation	Definition	Description
d_{ij}	$\|X_i - X_j\|_2$	Euclidean distance between samples i and j
$\hat{u}_{ij}$	$\frac{X_i - X_j}{d_{ij}}$	Unit vector from j to i
P_{ij}	$\hat{u}_{ij}\hat{u}_{ij}^T$	Elements of the projection matrix in $\hat{u}_{ij}$ direction
f_{ij}	$f(d_{ij})$	Link function evaluated at distance d_{ij}
g_{ij}^-	$\frac{\hat{w}_{ij} f'_{ij}}{d_{ij}}$	Elements of the gradient of the repulsive term
κ_{ij}^-	$\hat{w}_{ij} f''_{ij}$	Elements of the curvature of the repulsive term
g_{ij}^+	$-\frac{\hat{\lambda}_{ij}}{f_{ij}} g_{ij}^-$	Elements of the gradient of the attractive term
κ_{ij}^+	$-\hat{w}_{ij}\hat{\lambda}_{ij}\left[\frac{f''_{ij}}{f_{ij}} - \frac{(f'_{ij})^2}{f_{ij}^2}\right]$	Elements of the curvature of the attractive term

We define L^- as the repulsive Laplacian, constructed from weights g_{ij}^-, and L^+ as the attractive Laplacian, constructed from weights g_{ij}^+. With a total Laplacian $L = L^+ + L^-$, the gradient can be written as

$$\nabla_X l(X) = 2LX \tag{2.50}$$

Similarly for the Hessian, we define the block Laplacian with $(d \times d)$ blocks given by

$$R_{ij}^+ = -(\kappa_{ij}^+ - g_{ij}^+)P_{ij}, \qquad R_{ii}^+ = \sum_{j \neq i}(\kappa_{ij}^+ - g_{ij}^+)\, P_{ij} \tag{2.51}$$

$$R_{ij}^- = -(\kappa_{ij}^- - g_{ij}^-)P_{ij}, \qquad R_{ii}^- = \sum_{j \neq i}(\kappa_{ij}^- - g_{ij}^-)P_{ij} \tag{2.52}$$

with $R = R^+ + R^-$, we can write the Hessian succinctly as

$$\nabla_X^2 l(X) = 2\left[L \otimes I_d + R\right] \tag{2.53}$$

$$\nabla_X^2 l(X) = 2\left[(L^+ + L^-) \otimes I_d + R\right] \tag{2.54}$$

Appendix 2B: Fragmentation of a nearest-neighbor chain under heavy-tailed link functions

In section 2.5.3, embeddings of a 1D stepping-stone model fragmented into disconnected segments under the inverse-distance link ($p = 1$) but remained continuous under the inverse-square link ($p = 2$). This appendix shows, in a toy version of that setting, that the ordered (continuous) configuration is unstable and prone to fragmentation for every $p < 2$ and stable for every $p > 2$.

Result. Consider n samples whose coalescent rate is a constant λ between nearest neighbors on a line and zero otherwise, embedded in one dimension under the link $f(x) = (1+x)^{-p}$ with equal weights. Consider two possible configurations for the embeddings: an ordered chain with constant spacing s, and the same chain with a single large gap $L = \alpha n s$ ($\alpha \gg 1$) inserted at its midpoint. As $n \to \infty$, the change in loss produced by the gap is

$$\Delta l(p) \;\approx\; \underbrace{\lambda p \ln n}_{\text{attraction}} \;-\; \underbrace{\frac{n^{2-p}}{s^p}\,\frac{1-2^{p-1}}{(2-p)(1-p)}}_{\text{repulsion}} \qquad (p < 2), \tag{2.55}$$

with the repulsive term equal to $-\frac{n}{s}\ln 2$ at $p=1$, of order $-s^{-2}\ln n$ at $p=2$, and bounded for $p>2$. Hence

$$\boxed{\Delta l(p)\to\begin{cases}-\infty & p<2 \quad \text{(gap lowers the loss: fragmentation favored)}\\ +\infty & p>2 \quad \text{(gap raises the loss: ordered chain favored)}\end{cases}} \tag{2.56}$$

and $p=2$ is the critical exponent, where the two terms are both logarithmic in n and the outcome depends on λs^2. The mechanism is the one proposed in section 2.5.3: the attraction holding the chain together is local and costs only $O(\ln n)$ to break, whereas the repulsion relieved by separating the two halves is a sum over all $\sim n^2/4$ cross pairs and, for $p<2$, grows without bound.

Setup. With $\lambda_{ij}=\lambda$ for $|i-j|=1$ and 0 otherwise, and $f_{ij}=(1+|x_i-x_j|)^{-p}$ ($N_{\min}=1$), the loss in equation 2.15 (with all weights equal) is

$$l(x)=\sum_{i<j}\frac{1}{(1+|x_i-x_j|)^p}+\lambda p\sum_{i=1}^{n-1}\ln\big(1+|x_{i+1}-x_i|\big), \tag{2.57}$$

Place the two halves of the chain symmetrically about a gap of width L,

$$x_i=-\tfrac{L}{2}-(i-1)s, \qquad x'_j=\tfrac{L}{2}+(j-1)s, \qquad i,j=1,\dots,\tfrac{n}{2}, \tag{2.58}$$

so that the ordered chain is the case $L=s$ and the fragmented chain is the case $L=\alpha ns$. Within each half, all distances are independent of L; only the cross-half distances $d_{ij}=L+(i+j-2)s$ and the single bond across the gap change. We write $\Delta l=\Delta l_{\text{att}}+\Delta l_{\text{rep}}$ for the change in going from $L=s$ to $L=\alpha ns$.

Attractive term. Only the bond spanning the gap is affected:

$$\Delta l_{\mathrm{att}} = \lambda p \ln \frac{1+\alpha n s}{1+s} \sim \lambda p \ln n. \tag{2.59}$$

Repulsive term. Writing $k = i+j-2 \in \{0,\dots,n-2\}$ and letting $\gamma(k) = \min(k+1,\, n-1-k)$ be the number of cross pairs with a given k,

$$\Delta l_{\mathrm{rep}} = \sum_{k=0}^{n-2} \gamma(k) \left[\frac{1}{(1+ks+\alpha n s)^p} - \frac{1}{(1+ks+s)^p} \right]. \tag{2.60}$$

The first term is the residual repulsion across the gap; for $\alpha \gg 1$ it is smaller than the second by a factor of order α^{-p} uniformly in k, and we drop it (this only strengthens the inequality we are after, since the dropped term is positive). The second term is the repulsion between the two halves in the ordered chain, which the gap removes.

Case $p < 2$. Rescale $k = n\kappa$. For large n, $\gamma(n\kappa) \approx n \min(\kappa, 1-\kappa)$ and $(1+ks+s)^{-p} \approx (n\kappa s)^{-p}$, so the sum becomes

$$\Delta l_{\mathrm{rep}} \approx -\frac{n^{2-p}}{s^p} \int_0^1 \frac{\min(\kappa, 1-\kappa)}{\kappa^p}\, d\kappa. \tag{2.61}$$

The integrand is integrable at $\kappa = 0$ precisely when $p < 2$, which is why this case is treated separately. Splitting at $\kappa = \tfrac{1}{2}$,

$$\begin{aligned} \int_0^{1/2} \kappa^{1-p}\, d\kappa + \int_{1/2}^1 \left(\kappa^{-p} - \kappa^{1-p}\right) d\kappa &= \frac{2^{p-2}}{2-p} + \frac{1-2^{p-1}}{1-p} - \frac{1-2^{p-2}}{2-p} \\ &= \left(2^{p-1}-1\right)\left[\frac{1}{2-p} - \frac{1}{1-p}\right] = \frac{1-2^{p-1}}{(2-p)(1-p)}, \end{aligned} \tag{2.62}$$

which gives the repulsive term in equation 2.55. The prefactor is positive for all $p < 2$ (both numerator and denominator change sign at $p = 1$), and tends to $\ln 2$ as $p \to 1$, so

$\Delta l_{\text{rep}} \approx -\frac{n}{s}\ln 2$ at the default link. Since $n^{2-p} \to \infty$ while the attractive term grows only as $\ln n$, $\Delta l(p) \to -\infty$.

Case $p > 2$. The sum in equation 2.60 is now dominated by small k, i.e. by the few cross pairs nearest the midpoint, and converges as $n \to \infty$: $\sum_{k\geq 0}(k+1)(1+ks+s)^{-p} < \infty$. Thus Δl_{rep} tends to a finite negative constant, the attractive term $\lambda p \ln n$ wins, and $\Delta l(p) \to +\infty$.

Case $p = 2$. At the boundary, $\sum_k (k+1)(1+ks+s)^{-2} \sim s^{-2}\sum_{k\leq n/2} k^{-1} \sim s^{-2}\ln n$, so $\Delta l(2) \approx \left(2\lambda - s^{-2}\right)\ln n$: the gap is favorable when $2\lambda s^2 < 1$ and unfavorable otherwise. The exponent alone no longer decides, which is the signature of the critical point.

This calculation compares two specific configurations and does not identify the global minimizer. It does show that for any $p < 2$ the ordered chain is not a minimum, since a fragmented configuration with lower loss exists, which is the behavior observed in figures 2.14 and 2.15.

Appendix 2C: Parameters used for simulation examples

All tree sequences were simulated with `msprime` (Kelleher et al., 2016) under the standard coalescent with recombination, with demographies specified as `demes` graphs. Each simulation consists of 500 independent contigs of 5 Mb with a recombination rate of 1.2×10^{-8} per bp per generation, and samples are diploid. Times are in generations ago and population sizes N are diploid.

Complex 4-population demography (figures 2.11, 2.12, 2.13). An ancestral population (N = 15,000) splits at t = 1400 into D (constant N = 15,000) and ABC (N = 1500). ABC splits at t = 1200 into C and AB (N = 1500), and AB splits at t = 1000 into A and B. Starting from N = 1500 at their split, A grows exponentially to N = 30,000 and B to N = 60,000 at present. C grows exponentially from 1500 to 60,000 (t = 1200 $\to$ 351), declines to 1500 (t = 351 $\to$ 103), grows again to 60,000 (t = 103 $\to$ 30), and stays

constant afterwards. 20 individuals are sampled per population. In figure 2.13, A and B also exchange migrants symmetrically from their split to the present at $m \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. Figure 2.12 uses 10^4 single trees with one haploid sample per population.

Two-group 2D stepping-stone (figure 2.16). An ancestral population ($N = 5000$) splits at $t = 5000$ into 50 demes ($N = 100$ each), arranged as two 5×5 grids (groups A and B). Within each group, neighboring demes on the grid (non-diagonal) exchange migrants symmetrically at $m = 10^{-3}$ (A) or $m = 5 \times 10^{-3}$ (B). Every pair of demes from different groups exchanges migrants at $m = 10^{-7}$. All migration runs from the split to the present. One individual is sampled per deme. Embeddings use $p = 2$ and $N_{\min} = 1$, and are smoothed with $\alpha_1 = 10^{-4}$ and $\alpha_2 = 10^{-3}$ (section 2.3.4).

Two admixture events (figures 2.17, 2.18, 2.19). All populations have $N = 10^4$. An ancestral population splits at $t = 4000$ into the ancestors of (A,B) and of (C,D), and each of these splits at $t = 2000$. Admix_1 forms at $t = 500$ from A and B, and Admix_2 at $t = 100$ from C and D, both in equal proportions. 5 individuals are sampled from each of the six populations.

BADMIXTURE scenarios (figures 2.20, 2.21, 2.22). Adapted from Lawson et al., 2018. All populations have $N = 20{,}000$ unless stated, and 5 individuals are sampled per sampled population.

- *Three-way admixture:* the ancestral population splits at $t = 1700$ into D and the ancestor of (B,C), which splits at $t = 700$. A forms at $t = 15$ from B, C and D in proportions 0.5, 0.35 and 0.15.

- *Ghost admixture:* the ancestral population splits at $t = 1700$ into B, D and the ancestor of (C,X), which splits at $t = 700$ into C and an unsampled ghost X. A forms at $t = 200$ from B and X in equal proportions, and X ends at $t = 100$.

- *Recent bottleneck:* the ancestral population splits at $t = 1700$ into D and the ancestor of (B,C), which splits at $t = 700$ into C and the ancestor of B. A splits from the ancestor of B at $t = 20$ with $N = 200$.

Appendix 2D: Snakemake pipeline and visualization options

CoaSTT is implemented in Python as a Snakemake pipeline (Mölder et al., 2025) in the git repository `structure-thru-time` (https://github.com/afselim0/structure-thru-time).

Pipeline overview

A run is fully specified by one YAML configuration file (table 2.4) and launched as

```
snakemake STT --configfile <run>.yaml --cores all
```

The input is either a directory of inferred tree sequences in `tskit` format with sample metadata (target `STT`), or a `demes` file. From a demes file, the pipeline either simulates tree sequences with `msprime` (`STT_sim`), or infers ARGs with `tsinfer`, Relate or SINGER from sequences simulated under it (`STT_infer`). Each run then performs the following steps:

1. **Coalescence-time distributions.** Pairwise coalescence counts are computed per contig in parallel on a grid of bin width Δ_t (optionally for a random fraction of the samples per contig) and normalized into $\hat{f}_{T_{ij}}(t)$ (section 2.5.1).

2. **Rates and weights.** Survival functions are pooled across haplotypes into individual pairs, and $\hat{\lambda}_{ij}(t)$ and $\hat{w}_{ij}(t)$ are computed with the local-linear filter (section 2.5.1).

3. **Embeddings.** Embeddings are computed by sequential optimization (section 2.3.3), followed by the smoothed joint optimization when smoothing penalties are given (section 2.3.4). Rates reconstructed from the embeddings are saved alongside them.

4. **Plots and diagnostics** (listed below).

For simulated demographies, the true pairwise coalescent rates are computed numerically with `msprime` and embedded in the same way (`STT_true`), giving a noise-free reference. When admixed populations are specified, local ancestry is inferred (RFmix, GNN, or the true ancestry in simulations). For each ancestry, steps 1–4 are then repeated on the segments assigned to it, using the admixed individuals whose genome-wide proportion of that ancestry exceeds a threshold. This yields the RS-CoaSTT embeddings (section 2.5.4). Additional targets repeat the optimization under several random seeds (`STT_replicates`) or scan the link exponent p and the dimension d.

Each output file name carries a tag encoding every parameter it depends on, so runs with different settings share an output directory, and changing a parameter recomputes only the affected downstream steps. The standard outputs are:

- `data/`: the coalescence-time distributions, the rates $\hat{\lambda}_{ij}(t)$ and weights $\hat{w}_{ij}(t)$, the unsmoothed and smoothed embeddings $X_i(t)$, the reconstructed rates $f_{ij}(t)$ and, for inferred trees, a genotype PCA.

- `plots/`: pairwise rates through time colored by population pair; a per-population effective-size proxy $1/(2\bar{\lambda}_{\text{within}}(t))$; between- relative to within-population rates; interactive 3D embeddings (HTML) with static snapshots; a 2D slice at the most recent time; input (and, in simulations, true) versus reconstructed rates; and a launcher for the interactive app described below.

Population labels are used only to color and summarize the outputs, and never enter the estimation.

Table 2.4: Main configuration parameters of the CoaSTT pipeline. Times are in generations.

Parameter	**Default**	**Description**
Input		

Table 2.4, continued

Parameter	Default	Description
`trees_dir`, `pop_labels`, `pop_colors`	–	Inferred `.trees` files and sample metadata (empirical runs)
`pops`, `n_samples_per_pop`	all	Populations and individuals per population to include
`demes_file_path`	–	Demography to simulate
`num_contigs`, `contig_length`, `n_sim_per_pop`	500, 5 Mb, 20	Simulated contigs, and diploid individuals per population
`infer_method`	none	`tsinfer`, Relate or SINGER; none uses the simulated trees
Rate estimation		
`pdf_max_time`, `pdf_bin_size`	10^4, 1	Extent and bin width Δ_t of the coalescence-time grid
`fraction_samples`	1	Fraction of samples counted per contig
`rates_min_time`, `rates_max_time`	pdf grid	Time interval on which rates are estimated
`min_window_size`, `max_window_size`	50, 500	Filter bandwidth b at the first and last time point
`per_haplotype`	false	Rates between haplotypes rather than individuals
Embeddings		
`n_components`	2	Latent dimension d
`link`, `p`	inverse power, 1	Link function and its power p
`N_min`, `N_max`	1st, 99th percentiles of $1/\hat{\lambda}_{ij}(t)$	Resolution and maximum separation (equation 2.28)
`l1`	none	Smoothing fractions $[\alpha_1, \alpha_2, \dots]$; none disables smoothing
`unweighted`	true	Weight all pairs equally rather than by $\hat{w}_{ij}(t)$
Admixture		

Table 2.4, continued

Parameter	Default	Description
`admixed_pops`	none	Populations treated as admixed; enables RS-CoaSTT
`reference_pops`	all others	Reference panels for local ancestry
`LAI_method`	GNN	GNN, RFmix, or true ancestry (simulations only)
`significant_ancestry_threshold`	0.1	Minimum ancestry proportion to enter that ancestry's embeddings
Plotting		
`time_scale`, `embeddings_min/max_time`, `generation_time`	–	Display only: time axis scale and range, years axis

Visualizing embeddings through time

With many individuals, raw trajectories could be difficult to read, so the plotting module provides three layers that can be combined (figure 2.23). The first layer is the trajectory of each individual. The second is the median trajectory of each population, which summarizes where its cluster is located. The third is a support tube for each population, which summarizes the cluster's extent. At each time point, the 10% of a population's points farthest from its median are discarded, the convex hull of the rest is computed, and the hulls are smoothed along log-time.

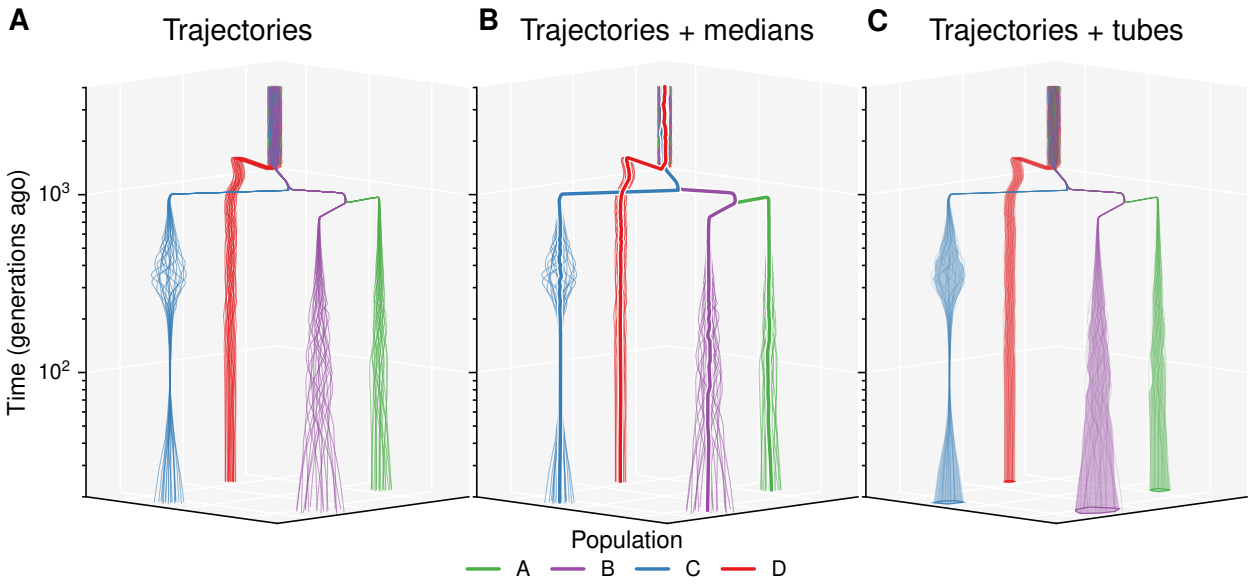

Figure 2.23: Options for displaying CoaSTT embeddings through time. Smoothed embeddings of the complex 4-population simulation (figure 2.11) shown as **A)** individual trajectories, **B)** individual trajectories with the median trajectory of each population, and **C)** the support tube of each population (trimmed convex hulls at each time point, smoothed along time), with individual trajectories drawn faintly on top.

A 3D view alone hides how individuals are arranged within a time slice. Each run therefore also produces an interactive application built with Plotly Dash (figure 2.24), which opens in a web browser through the launcher in `plots/`. A time slider (also driven by the arrow keys, or by a play button at adjustable speed) moves a plane through the 3D trajectories and updates a linked 2D view of the latent positions at that time. The time axis can be switched between linear and log scale, and the 3D view can be restricted to a time window. An optional panel shows the input versus reconstructed rates of all pairs at the current time, together with the median for each population. Hovering over a pair in this panel highlights the two individuals in the 2D view. The k most over- and under-estimated pairs can also be drawn as edges between individuals in the 2D view, which shows where the embedding fails to represent the input rates. For sharing, the app and its data can be packaged into a self-contained archive that runs outside the repository.

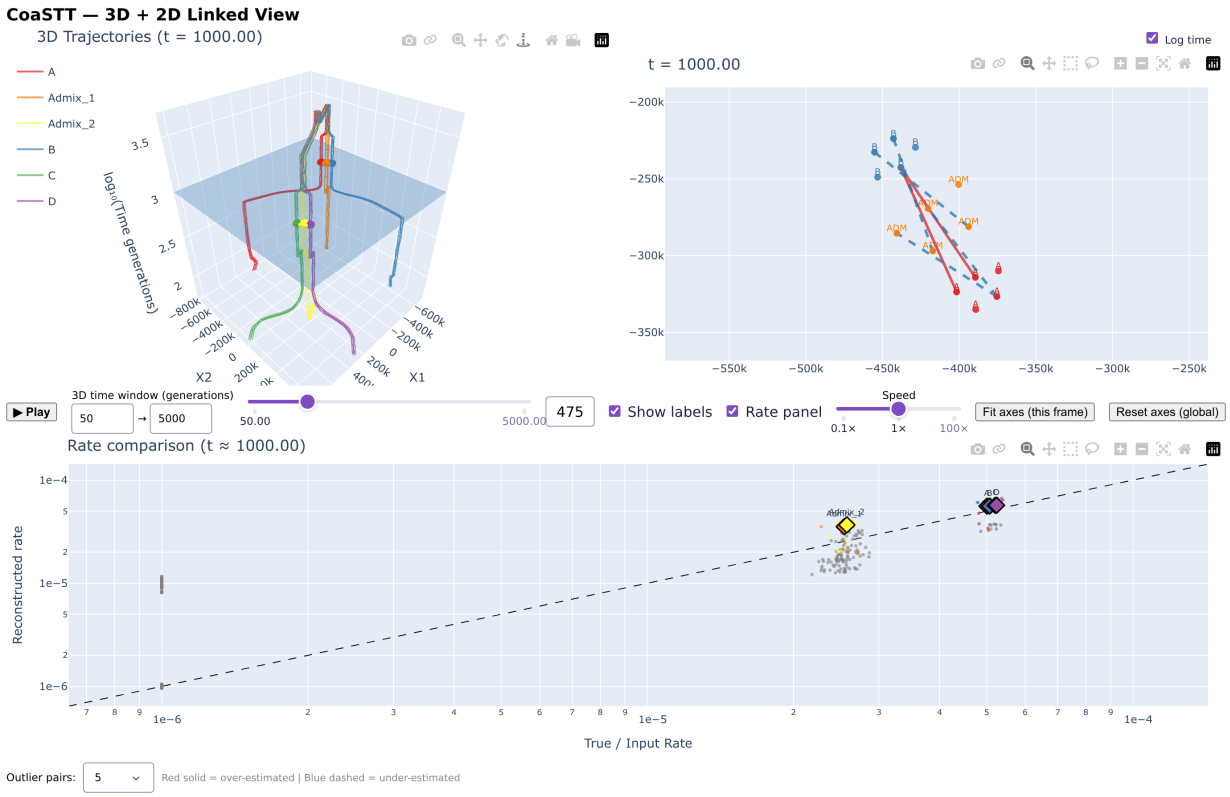

Figure 2.24: Interactive exploration of CoaSTT embeddings. Snapshot of the Dash application for genome-wide embeddings of the two-admixture simulation (figure 2.17) at $t = 1000$ generations ago, between the admixture and split times. Top left: 3D trajectories with a plane marking the selected time. Top right: latent positions at that time, zoomed on populations A, B and Admix_1, with edges between the five most over-estimated (red, solid) and under-estimated (blue, dashed) pairs. The A–B pairs are drawn too close because the admixed individuals pull their source populations together (section 2.5.4). Bottom: input versus reconstructed rates at that time, colored by population pair, with the median for each population shown as diamonds.

CHAPTER 3

EMPIRICAL ANALYSIS OF THE GENETIC HISTORY OF HUMANS AND CHIMPANZEES USING COASTT

3.1 Introduction

Reconstructing population history from genetic data is a daunting task that often requires running the data through a suite of inference methods, each adding a piece to the complex puzzle (Beichman et al., 2018; Schraiber & Akey, 2015). Each method naturally operates under a set of assumptions that allow its inferred parameters to be interpretable at the expense of a simplified description of history. One assumption is nearly ubiquitous in population history inference: that history is describable in terms of discrete population units (Coop, 2023; National Academies of Sciences, Engineering, and Medicine, 2023). Methods that provide time-resolved representations of population history and structure typically require assigning population labels to present-day individuals, after which history is described in terms of these population units. CoaSTT, the method described in chapter 2, was developed in part to circumvent this assumption. It avoids population assignment by inferring a time-varying latent trajectory for each individual, providing a low-dimensional representation of coalescent structure through time.

In this chapter, we revisit the genetic history of two of the best-studied species, humans and chimpanzees, through the lens of the individual-based coalescent history provided by CoaSTT. Humans provide a well-characterized history against which CoaSTT can be validated, while chimpanzees have a well-defined subspecies structure with an active interest in recent history within and between the subspecies. For both species, detailed historical models have been progressively refined over the past two decades as larger genomic datasets and more powerful computational resources have become available (de Manuel et al., 2016; Nielsen et al., 2017). We highlight how CoaSTT can add nuance to these models through a flexible,

individual-based treatment of history, especially in the recent past, where population-based summaries may discard the most information by obscuring fine-scale structure.

All examples shown in chapter 2 demonstrated CoaSTT embeddings computed from either true coalescent rates or rates estimated from true, simulated ARGs. In practice, CoaSTT takes inferred ARGs as input, and ARG inference introduces biases of its own (Y. C. Brandt et al., 2022). We therefore start this chapter by characterizing the biases introduced by ARG inference and how they propagate to CoaSTT embeddings (section 3.2). Throughout, we use SINGER + POLEGON (Deng, DeWitt, et al., 2025; Deng, Nielsen, & Song, 2025) for ARG inference, as we show that it provides reasonable estimates of coalescent rates, and it additionally samples ARGs from an approximate posterior distribution, providing a natural means for quantifying uncertainty. We then apply CoaSTT to ARGs inferred for human (section 3.3) and chimpanzee (section 3.4) datasets, using both genome-wide and genomic-region-specific embeddings, and highlight several cases where CoaSTT provides a more nuanced view of genetic history than that offered by demographic inference methods that group individuals into populations.

3.2 Analysis of bias associated with estimating coalescent rates from ARGs

CoaSTT relies primarily on estimates of coalescent rates between each pair of individuals or haplotypes. These rates are derived from estimates of pairwise coalescent times along the genome, which are in turn informed by the mutational signal and lengths of shared haplotypes between genomic samples. This signal, whose density is determined by the rates of mutation and recombination along the genome, is often too sparse for pairwise coalescent times to be determined precisely, so point estimates of coalescent times carry a substantial degree of uncertainty, often at scales that are a sizable fraction of the estimated time itself (Y. C. Brandt et al., 2022). ARG inference methods that jointly infer coalescent times for all

pairs of samples regularize this uncertainty by placing a prior on coalescent times, typically derived from a panmictic coalescent model. Such priors imply smoothly varying or even uniform coalescent rates through time, and their effect is to smooth sharp changes in the underlying coalescent rate, such as those induced by population splits or bottlenecks.

This limitation is not specific to ARG inference. Coalescent-HMM methods that infer effective population size or relative cross-coalescence rates from the same mutational signal have long been known to have limited temporal resolution, with abrupt changes in coalescent rate being attenuated and spread over neighboring time intervals (J. Kim et al., 2015; H. Li & Durbin, 2011; Sellinger et al., 2021), and with the discretization of time itself introducing a bias-variance tradeoff (Spence et al., 2018). The convention of dating a population split at the time at which the relative cross-coalescence rate inferred by MSMC crosses 0.5 (Schiffels & Durbin, 2014) is an implicit acknowledgment of this. The inferred rate is a smoothed version of a step function, and the halfway point is used as a proxy for the location of the step, even though the smoothed rate is also consistent with a gradual process (K. Wang et al., 2020).

Existing benchmarks of ARG inference methods focus on the accuracy of inferred pairwise coalescent *times* (Y. C. Brandt et al., 2022). CoaSTT, however, takes as input coalescent *rates*, which are obtained from the distribution of coalescent times through the derivative of the log-survival function (section 2.5.1). In this section, we briefly assess the bias in estimates of pairwise coalescent rates derived from ARGs inferred by Relate, tsinfer + tsdate, and SINGER + POLEGON (Deng, DeWitt, et al., 2025; Deng, Nielsen, & Song, 2025; Kelleher et al., 2019; Speidel et al., 2019; Wohns et al., 2022). We choose to proceed with SINGER combined with POLEGON to infer ARGs in the remainder of the chapter, though we note that all three methods result in different levels of bias in rate estimates at different time periods. We additionally show how this bias in rate estimates translates to CoaSTT embeddings as smoothed latent trajectories, a concept that is useful to keep in mind when

interpreting the results of CoaSTT applied to empirical datasets.

3.2.1 ARG inference methods smooth changes in coalescent rates resulting from demographic events

We compare coalescent rates estimated from ARGs inferred by different methods under two simple demographic scenarios. The first is a clean population split, which highlights how ARG inference smooths sharp changes in between-population coalescent rates. The second is the "sawtooth" demography of Schiffels and Durbin (2014), which highlights how it smooths sharp changes in within-population coalescent rates, and hence in inferred effective population sizes. In the split scenario, two populations of constant size $N = 10^4$ diverge from an ancestral population of the same size 2,000 generations ago, and 20 diploid individuals are sampled from each. Under each demography, we simulate 500 independent tree sequences of 5×10^6 bases using msprime (Baumdicker et al., 2022; Kelleher et al., 2016) with equal mutation and recombination rates ($\mu = r = 10^{-8}$ per base pair per generation). For each simulated tree sequence, we infer a tree sequence from the mutational data using Relate, tsinfer + tsdate, and SINGER + POLEGON, supplying the true mutation and recombination rates to all three methods. SINGER is run for 2,500 MCMC iterations, and the last posterior sample is redated using POLEGON. Relate is run once using the default panmictic prior, and then rerun using the re-estimated population size history as a prior. Pairwise coalescent rates are then estimated from each tree sequence, true or inferred, as described in section 2.5.1.

Under a clean population split, the true coalescent rate between samples from the two populations is exactly zero before the split. All three inference methods instead yield non-zero between-population rates well before the split, rising gradually toward the ancestral rate (figure 3.1A, lower panel). After the split, only SINGER + POLEGON reaches the true ancestral rate after a short lag, whereas Relate and tsinfer + tsdate converge to it from below over several thousand generations. The same downward bias is present in their

within-population rates (figure 3.1A, upper panel), so both methods correctly capture that between- and within-population rates coincide after the split, but at an underestimated absolute rate. Since CoaSTT attempts to model absolute rather than relative rates, this bias in the absolute level matters even when the ratio of the two is correct. Estimating the split time as the point where the ratio of between- to within-population rates reaches 0.5 recovers the true split time for SINGER + POLEGON and Relate, but overestimates it by roughly a third for tsinfer + tsdate (dashed lines in figure 3.1A).

Under the sawtooth demography, all three methods attenuate the true amplitude of the oscillations in coalescent rates (figure 3.1B). We measure the log-contrast retained between consecutive peaks/troughs for each rate estimate as

$$\rho(t_1, t_2) = \frac{\ln \hat{\lambda}(\hat{t}_1) - \ln \hat{\lambda}(\hat{t}_2)}{\ln \lambda(t_1) - \ln \lambda(t_2)}, \tag{3.1}$$

where (t_1, t_2) indicate the times of the two consecutive extrema, and $(\hat{t}_1, \hat{t}_2)$ indicate the times of the corresponding extrema in the rate estimates. Inferred ARGs mostly retain the shape of the oscillation, but underestimate its magnitude, each method doing so at a different level for different time periods. SINGER + POLEGON retains the most contrast in the two older cycles and the least in the most recent one, with all methods retaining under half of the true log-contrast. In recent time, even the rates estimated from the true trees deviate from the analytic truth because the window size used (within which rates are assumed to be constant) is larger than the time scale at which the rates vary. Using a smaller window size and a larger number of samples would alleviate this effect.

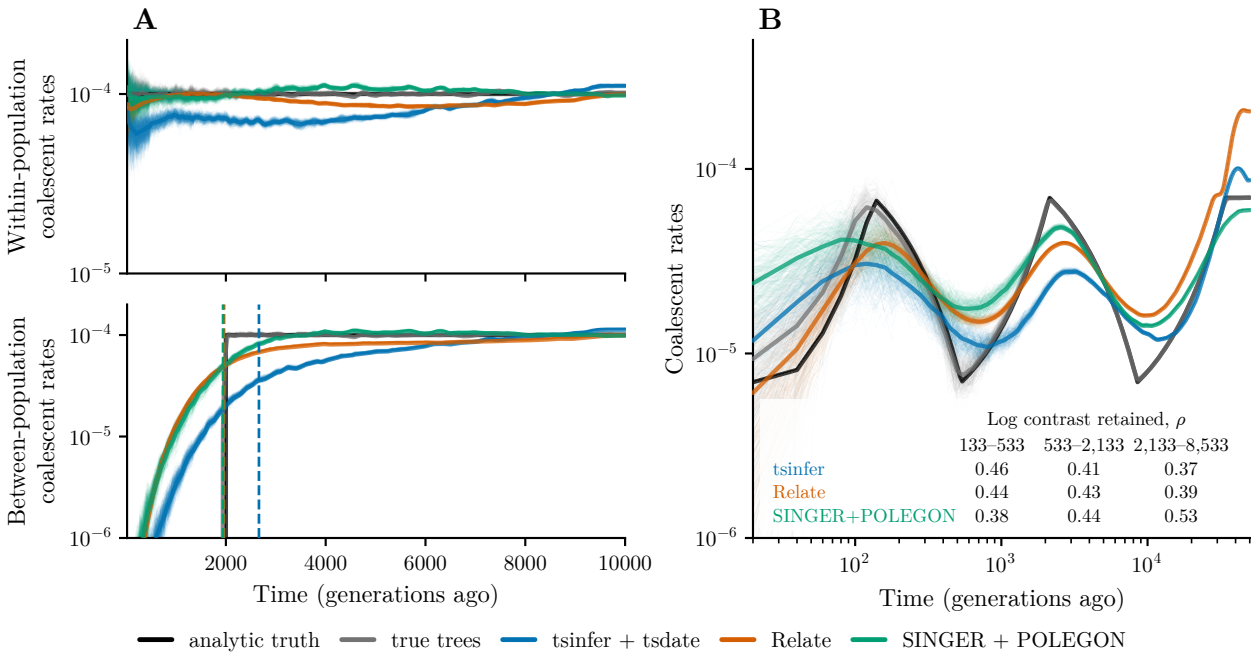

Figure 3.1: Coalescent rates estimated from ARGs inferred by different methods. In both panels, black is the analytic truth, gray is the rate estimated from the true simulated trees, and colors correspond to different ARG inference methods. Thin lines are estimates of pairwise coalescent rates between individuals, and thick lines are their mean. **A)** Population split scenario: two populations of size $N_e = 10^4$ diverging from an ancestral population of the same size 2,000 generations ago, with 20 diploid individuals per population. The upper panel shows within-population rates and the lower panel between-population rates. Dashed vertical lines mark the split time estimated for each method as the time at which the ratio of between- to within-population rates reaches 0.5. **B)** Sawtooth demography, with a table of the log-contrast retained defined in equation 3.1 for mean rate trajectories derived from each ARG inference method. The columns indicate the true times of the extrema, but the contrast in estimated rates is evaluated at the corresponding extrema of each estimate.

3.2.2 CoaSTT applied to ARGs inferred from simulated data

To explore how the bias in rates estimated from inferred ARGs affects CoaSTT embeddings, we compare the embeddings obtained using simulated trees vs. inferred trees as input, using SINGER + POLEGON for inference. We conduct this comparison for the two simple demographic scenarios tested in figure 3.1 (figure 3.2A,B) as well as two more complex ones that were presented in chapter 2 (figure 3.2C,D). The gradual approach of clusters backward

in time in the population split scenario (figure 3.2A) and the suppressed variation in cluster size in the sawtooth demography (figure 3.2B) are both expected effects from the results shown in figure 3.1. The example in figure 3.2C shows the result of the combination of these effects. The periodic size changes of population C are present but suppressed and more difficult to visually discern. Plotting the embeddings of samples from population C on their own would allow such size changes to be better discerned as in figure 3.2B. For the demography involving two spatial populations (simulation shown in figure 2.16), the continuous within-group structure is preserved and the two groups approach each other more rapidly than in the true trees, consistent with panel A (figure 3.2D).

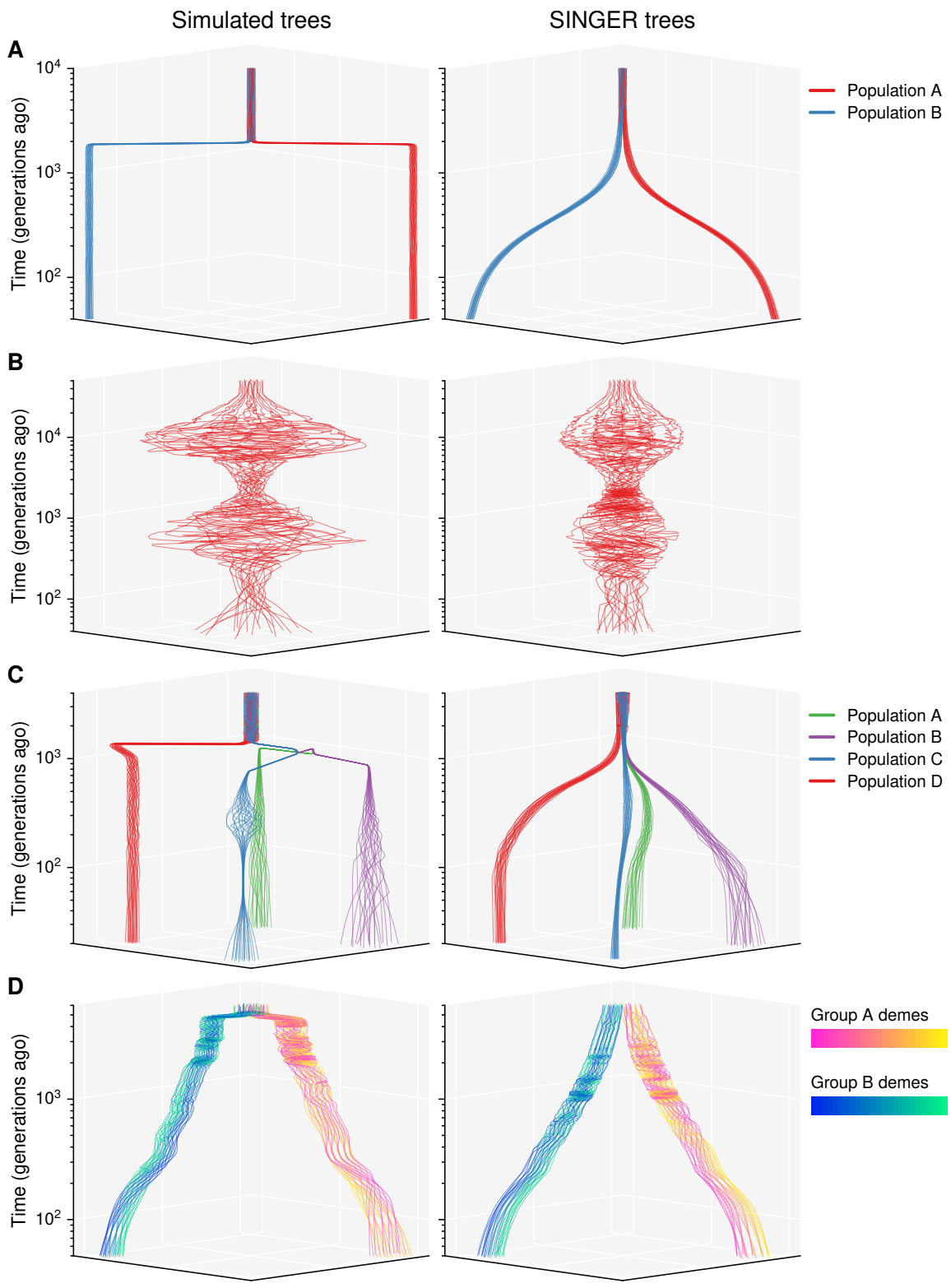

Figure 3.2: CoaSTT embeddings with rates estimated from true trees and SINGER-inferred trees. A) Population split. **B)** Sawtooth demography. **C)** 4-population demography of figure 2.11. **D)** Demography of two spatial populations shown in figure 2.16.

We additionally show a more detailed example based on the four-population out-of-Africa demographic model (Jouganous et al., 2017), which involves simulation parameters relevant for human populations, and will be useful in interpreting CoaSTT embeddings resulting from human data (figure 3.3). The model includes four groups of samples: Japanese in Tokyo (JPT), Han Chinese in Beijing (CHB), Utah residents with Northern and Western European ancestry (CEU) and Yoruba (YRI). The JPT-CHB split is 9 kya, the (JPT,CHB)-CEU split 46 kya, and the (JPT,CHB,CEU)-YRI split, which marks the out-of-Africa event, is 119 kya. Gene flow exists between CHB-JPT, CEU-CHB, YRI-CHB, YRI-CEU and YRI-(JPT,CHB,CEU) at rates defined in the `stdpopsim` catalog (Adrion et al., 2020a). The upward bias in between-population coalescent rates brings all clusters together in recent time (figure 3.3C,D). By 29 kya, the JPT and CHB samples form a single cluster in both simulation-based and SINGER-based embeddings, but the latter cluster is slightly structured, with samples not perfectly mixing as a result of the downward bias in between-population coalescent rates estimated from SINGER ARGs. A similar pattern is seen 72 kya, where a central cluster is formed by all non-African samples, which seems more structured in the SINGER-based embeddings. The CEU samples are pulled slightly to the edge of the cluster in the simulation-based embeddings because they have more segments shared with YRI ancestors in their genome due to the higher rate of gene flow with YRI ancestry, and this signal is accentuated by SINGER, creating a halo of CEU samples around the non-African cluster. The halo formed by the YRI samples around the CEU/JPT/CHB core is an interesting signal, caused by the fact that by 72 kya, the true coalescent rate between YRI individuals is lower than that between YRI and non-YRI individuals, which is a combined result of the large population size of the YRI ancestral population at that time, the low population size of the CEU/JPT/CHB population modeled by the out-of-Africa bottleneck, and the degree of gene flow between the two populations. By 174 kya, a single ancestral population exists, reflected by a circular cluster in the simulation-based embeddings. Due

to SINGER's tendency to push split signals further into the past, the YRI halo persists in the SINGER-based embeddings until this time. Based on these results, we recommend that whenever circular clusters that are slightly structured or that are surrounded by a very close halo appear in SINGER-based embeddings, this should not be taken as evidence of lingering structure on its own, since the same signature arises from the bias in SINGER-based rates.

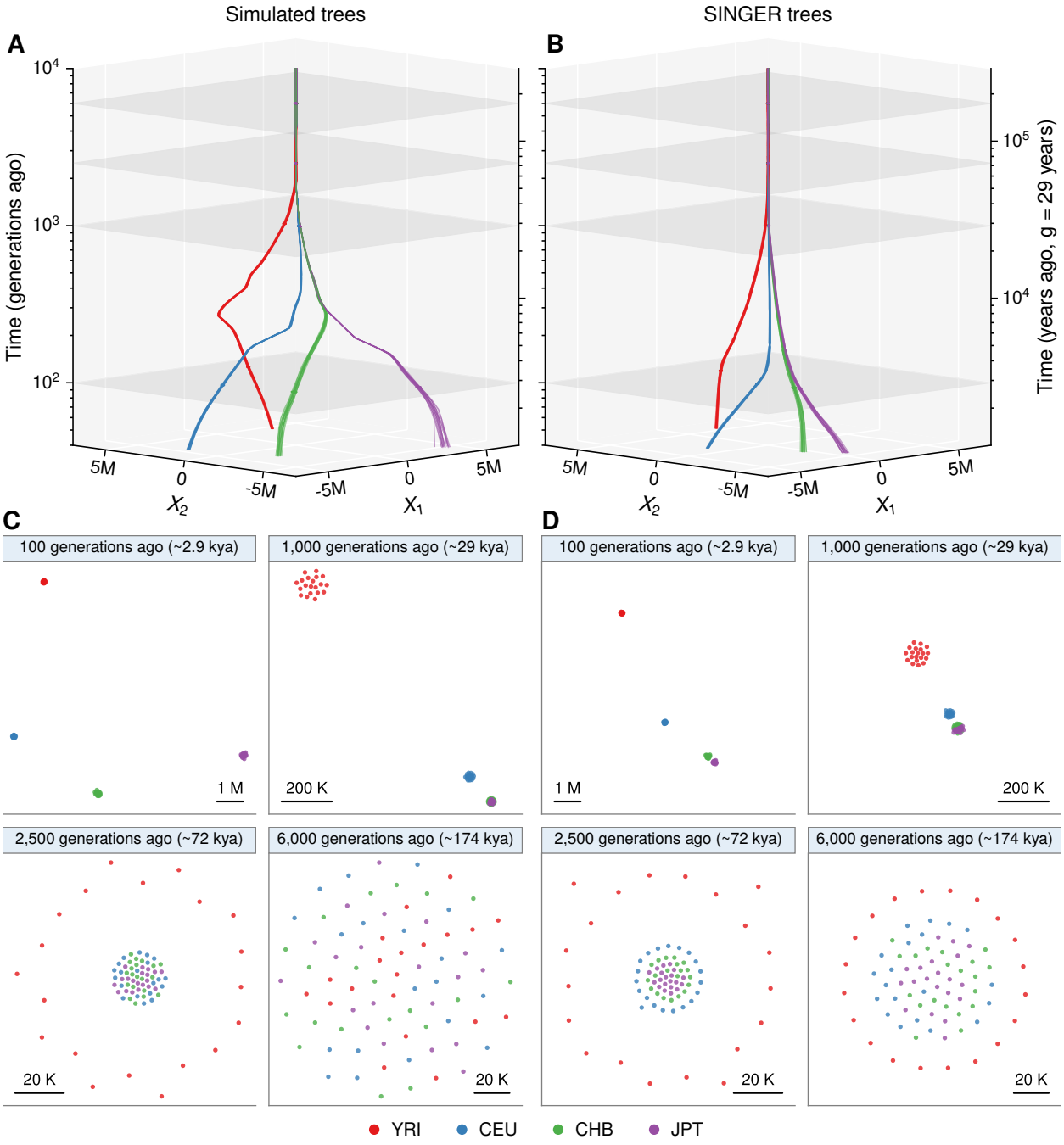

Figure 3.3: CoaSTT embeddings for a four-population out-of-Africa model, with rates estimated from simulated and SINGER-inferred trees. Simulations follow the `OutOfAfrica_4J17` model of Jouganous et al. (2017) in `stdpopsim` (Adrion et al., 2020a), with 20 samples per population, and trees are inferred with SINGER + POLEGON as in figure 3.1. **A,B)** 3D embeddings from simulated (A) and SINGER-inferred (B) trees, with years assuming a 29-year generation time. Gray planes mark the time slices shown in **(C,D)**, drawn on a common scale for each slice.

We also apply RS-CoaSTT to the admixture scenario of figure 2.17, with trees inferred using SINGER + POLEGON and local ancestry assigned using either RFmix or GNNs (Maples et al., 2013; Wohns et al., 2022). As discussed in section 2.5.4, when rates are restricted to genomic regions assigned to one source ancestry, RFmix-based assignment yields non-zero recent coalescent rates between the admixed samples and the *other* source population, because a fraction of segments is misassigned. In the embeddings, the admixed cluster consequently pulls that other source cluster slightly toward itself (figure 3.4C,D). GNN-based assignment produces fewer such misassignments, so the corresponding rates stay near zero and the admixed cluster remains aligned with the source whose ancestry the genomic segments are assigned to, without pulling in the other source cluster (figure 3.4A,B). This matches the behavior observed with true trees in figures 2.19 and 2.18, with sharp changes smoothed out as expected from inferred ARGs.

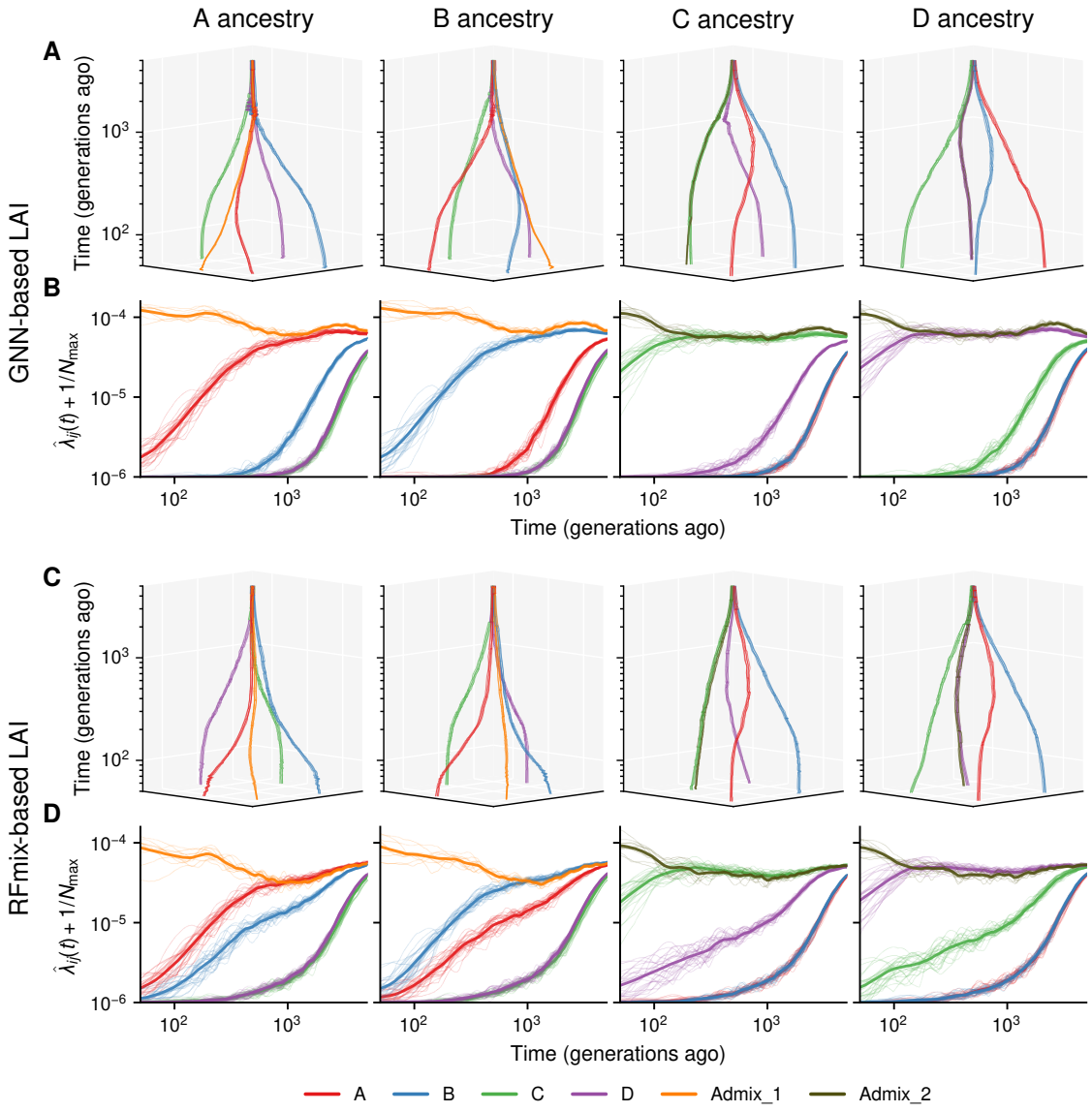

Figure 3.4: RS-CoaSTT embeddings resulting from SINGER-inferred trees. Trees were inferred using SINGER + POLEGON from simulations based on the demography shown in figure 2.17. Each column corresponds to one of the four source ancestries. **A)** GNN-based RS-CoaSTT embeddings, where coalescent rates for admixed individuals are computed using only genomic regions assigned to the column's ancestry based on GNNs. **B)** Region-specific coalescent rates between admixed samples and samples from all other populations, with local ancestry assigned based on GNNs: the first two columns show rates for Admix_1 restricted to A- and B-ancestry segments, and the last two show rates for Admix_2 restricted to C- and D-ancestry segments. **C,D)** Same as **(A,B)**, but with local ancestry assigned using RFmix. Incorrect local ancestry assignment in recent time results in non-zero coalescent rates between Admix_1 samples and population B when restricted to A-ancestry segments (first column), and between Admix_1 samples and population A when restricted to B-ancestry segments (second column). A similar effect is visible for Admix_2 in the third and fourth columns.

3.3 Exploring individual-based coalescent history of human samples using CoaSTT

Over the past three decades, global genomic datasets have helped us form a detailed understanding of human history. As sequencing technology has developed, these datasets have grown in resolution, from a few hundred microsatellite markers in the early 2000s (Rosenberg et al., 2002), to hundreds of thousands of SNPs in the late 2000s (Jakobsson et al., 2008; J. Z. Li et al., 2008), to whole-genome sequences of thousands of individuals worldwide since the 2010s (Auton et al., 2015; Bergström et al., 2020; Mallick et al., 2016). A suite of methods has been developed to analyze this global genetic diversity in a manner that allows meaningful conclusions to be drawn about population structure and history. CoaSTT allows such diversity to be explored in an individual-based manner, similar to PCA or STRUCTURE/ADMIXTURE (Alexander et al., 2009; Patterson et al., 2006; Pritchard et al., 2000), but it is also time-resolved, similar to coalescent-HMM methods such as PSMC and MSMC (H. Li & Durbin, 2011; Schiffels & Durbin, 2014).

We apply CoaSTT to ARGs inferred for one such global dataset, the Human Genome Diversity Project (HGDP) panel (Cann et al., 2002), which has been used extensively to understand human population structure and history. It has been the subject of several seminal studies (J. Z. Li et al., 2008; Rosenberg et al., 2002) and a focal point in debates over how the structure they revealed should be interpreted, in particular whether human diversity is better described by discrete clusters or by continuous clines (Rosenberg et al., 2005; Serre & Pääbo, 2004). We use the recent whole-genome version of the dataset provided by Bergström et al. (2020), which includes 929 statistically phased whole genomes sampled from 54 diverse human populations. Because SINGER scales poorly with sample size, we infer ARGs with SINGER + POLEGON for subsets of samples grouped by region. The 54 HGDP populations are conventionally divided into seven geographic regions (Bergström

et al., 2020; J. Z. Li et al., 2008): Africa, Middle East, Europe, Central and South Asia, East Asia, America and Oceania. We infer ARGs for samples from each region independently, and for an eighth set, formed by choosing the four individuals with the highest sequencing coverage from each population, and apply CoaSTT to each of these ARGs. Throughout, times in generations are converted to years assuming a generation time of 29 years. In doing so, we reproduce multiple well-known aspects of human history and highlight some features of recent history that population-based summaries may obscure, demonstrating CoaSTT's potential for visualizing the coalescent history encoded in ARGs in an interpretable and label-free manner.

3.3.1 Global continuous structure of human coalescent history

Figure 3.5 shows the results of applying CoaSTT to the ARG inferred for all HGDP populations, with four individuals sampled from each of the 54 populations. The 3D embeddings immediately reveal major patterns of human population history in a visually intuitive way by exhibiting a tree-like structure with a narrow trunk that extends from the deep past until approximately 2,000 generations ago (about 58 kya), after which lineages begin to branch out into different areas of the embedding space. This branching results in a clear, almost linear cline in the recent past. Reading it from left to right, the cline tracks African lineages followed by Middle Eastern lineages, West Eurasians, Central and South Asians, Northeast Asians and Southeast Asians. The Oceanian lineages lie on the rightmost side of the cline but markedly shifted on the second dimension, whereas the Americans lie in the middle of the cline, also shifted in the second dimension but to the other side. A notable aspect of this cline is the bridge between East and West Eurasian samples, occupied by Uygur and Hazara lineages, both well known to be admixed populations with East and West Eurasian ancestral components (Hellenthal et al., 2014).

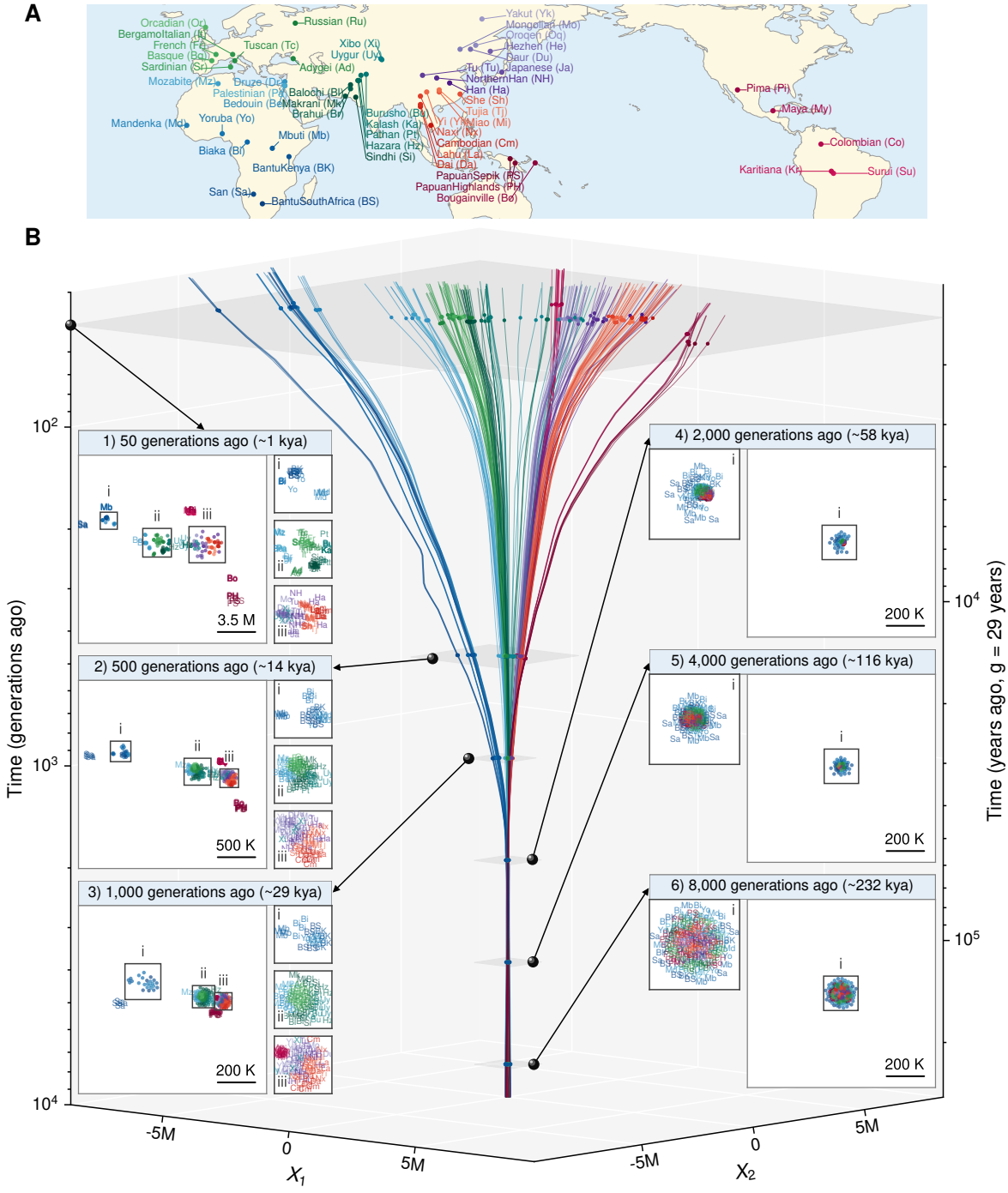

Figure 3.5: CoaSTT embeddings for 54 HGDP populations. A) A map depicting sample locations and population labels with their abbreviations. **B)** CoaSTT embeddings for four individuals from each population, with time (in generations or years ago) stacked vertically and changing from present to past moving downwards. Samples are colored by region, with abbreviations as in A. Six representative time slices, numbered (1)–(6), are highlighted. Within each, sub-insets (i–iii) zoom into sets of samples.

Looking more closely into the six representative time slices shown in figure 3.5, many established patterns for these populations can be observed without the need to assign population labels to samples beforehand. Samples are most differentiated according to their coalescent rates at the most recent time point, with this differentiation decreasing as we move backward in time, as coalescent rates between samples of different groups increase as a result of pervasive gene flow (which is amplified by the bias in SINGER-derived coalescent rates). In the recent past, the width of the linear cline described above rapidly decreases as we go back in time, which is apparent from the scale of insets (1) and (2) in figure 3.5B. By 500 generations ago (about 14 kya), populations from East and West Eurasia still form two distinct clusters (ii and iii), consistent with their divergence around 40 kya (Malaspinas et al., 2016; Nielsen et al., 2017), with the African, American and Oceanian clusters all pulled much closer to the Eurasian clusters. By 1,000 generations ago (about 29 kya), the American cluster appears to be in the process of being absorbed into the East Asian cluster, as expected from the divergence of the ancestors of Native Americans from East Asians around 25 kya (Moreno-Mayar et al., 2018; Raghavan et al., 2015). Next, the East Asian cluster itself loses much of its remaining structure, as does the West Eurasian one. The Oceanian cluster mostly loses its internal structure as well, and appears on the edge of the larger East Asian cluster: closer to East Asians than to West Eurasians, but still distinct from either, consistent with the ancestors of Papuans having diverged from other non-Africans before the East–West Eurasian split, around 50–60 kya (Malaspinas et al., 2016).

By 2,000 generations ago (about 58 kya), all non-African samples form a single cluster, matching the timing of the out-of-Africa dispersal (Nielsen et al., 2017), and the African samples form a crescent-like cluster around them, with San samples visibly on the edge of that cluster, as expected for the most deeply diverged present-day lineage (Schlebusch et al., 2017; Skoglund et al., 2017). By 4,000 generations ago, the crescent-like cluster of African samples fully wraps around the central cluster of non-African samples, with San samples still

slightly separated on the edge. Finally, in the deepest shown time slice of 8,000 generations ago, a single circular cluster appears, with the African samples still forming a ring on the edge of the circle.

Three caveats are worth keeping in mind when interpreting these embeddings. First, the gap between the African cluster and the Eurasian cluster on the recent cline is influenced by the lack of samples from more diverse populations in Africa and the Middle East, which we expect would fill this gap and showcase more clearly the continuity of the cline. Second, it might be appealing to interpret this cline in terms of a particular historical process like a serial founder effect. We caution against these kinds of interpretations, which have been mistakenly applied to similar PCA plots as well. Such a cline is simply a result of the pattern of coalescent similarity in recent time slices (Novembre & Stephens, 2008). It could arise from longstanding isolation by distance or a more complex pattern of migrations (such as a serial founder effect coupled with different levels of sustained gene exchange). It is worthwhile reading plots like these for what they are: low-dimensional representations of time-varying ancestral similarity measured in terms of the rate of coalescent events, regardless of the exact historical processes that led to this pattern of similarity. Third, the African ring that surrounds the non-African cluster is only a reliable signal up to the time of the out-of-Africa split. Figure 3.3 shows that such a ring arises from true coalescent rates before the split, when the large African population coalesces more slowly within itself than with the small non-African population, but that in SINGER-based embeddings it persists well past the split, where the true rates have already equalized. The ring at 2,000 generations ago (58 kya) may therefore be genuine, whereas its persistence at 4,000 and 8,000 generations ago (116 and 232 kya), both at or beyond the split, is most likely an artifact of the same bias from using an inferred ARG. The same reasoning applies to residual structure *within* clusters that have already merged, such as the imperfect mixing of the non-African cluster at 4,000 generations ago or within the East and West Eurasian clusters at 1,000 generations

ago (the two clusters themselves being distinct at that time is likely genuine signal, as discussed above). Such residual structure is consistent with lingering coalescent structure, but it cannot be distinguished from the bias in the rates on its own, since ARGs inferred by Relate and tsinfer + tsdate produce the same persistent ring (not shown); ultimately, it has to be verified with other methods, perhaps through the inclusion of ancient samples.

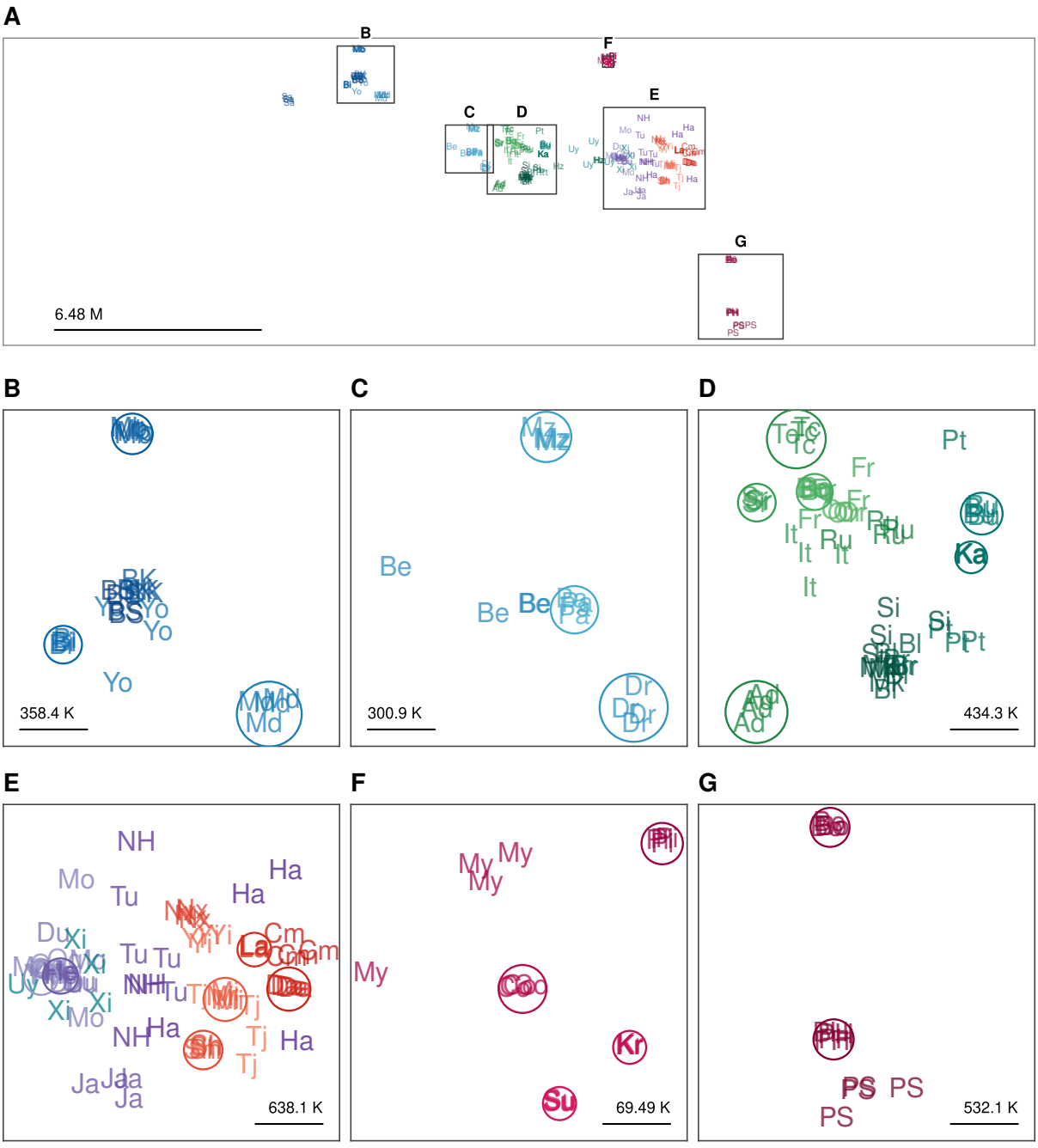

Figure 3.6: CoaSTT embeddings for 54 HGDP populations 50 generations ago. **A)** The full time slice, identical to inset (1) of figure 3.5B. Panels B–G zoom into parts of A to show the arrangement of samples more clearly. **B)** African samples other than the San. **C)** Middle Eastern samples. **D)** European and Central and South Asian samples. **E)** East Asian samples. **F)** American samples. **G)** Oceanian samples.

In addition to the major historical patterns, a closer look at the embedding at a particular time slice reveals distinct ways in which samples sharing a population label are arranged relative to one another. These patterns are clearest in the most recent slice, 50 generations ago (figure 3.5B (1)), which we show enlarged in figure 3.6. Though different clustering patterns can be observed visually, we supplement the visual inspection with a quantitative measure of clustering. To quantify how clustered each population is, we compute for each sample a nearest-neighbor-based variant of the silhouette coefficient (SC) (Rousseeuw, 1987), $s_i = \dfrac{b_i - a_i}{\max(a_i, b_i)}$, where a_i is the mean embedding distance from sample i to the other samples of its population and b_i is its mean distance to the four nearest samples from any other population. We summarize each population by the mean of s_i over its samples, and describe values above 0.7 as indicating a well-defined cluster, following the interpretation proposed by Kaufman and Rousseeuw (1990, Ch.2, Table 4). As those authors note, this interpretation is subjective, and we use it descriptively. They also suggest values of 0.5–0.7 to indicate reasonable evidence for clustering, and 0.25–0.5 to indicate weak evidence. We additionally compute the SC that would result from an embedding that reproduced every input rate exactly, by inverting the link function on the input coalescent rates and averaging the resulting distances over each sample's population-mates and over its four highest-rate samples from other populations to obtain a_i and b_i.

Figure 3.7 shows the SC for each population 50 generations ago, computed from the embedding distances as well as from the distances obtained by inverting the link function. Populations that are strongly clustered in the embedding space are also tightly clustered when distances are computed from inverted input rates. Two exceptions are the Sardinian population, whose embedding-based SC falls below the rate-based one for this time slice only after smoothing (the SC computed from the unsmoothed embeddings matches the SC from the input rates), and the Adygei, who have many weak links with European, Central Asian and Middle Eastern populations that the 2D embeddings cannot honor at once. Generally,

agreement between the embedding-based and rate-based silhouettes is close for well-separated populations (SC > 0.7) but weaker below this value. When a population's within- and between-population distances are similar, the silhouette is most sensitive to small distortions of either. Such populations lie in dense regions of near-equal rates, or are admixed between more distant groups, which are precisely the configurations a two-dimensional embedding cannot represent exactly.

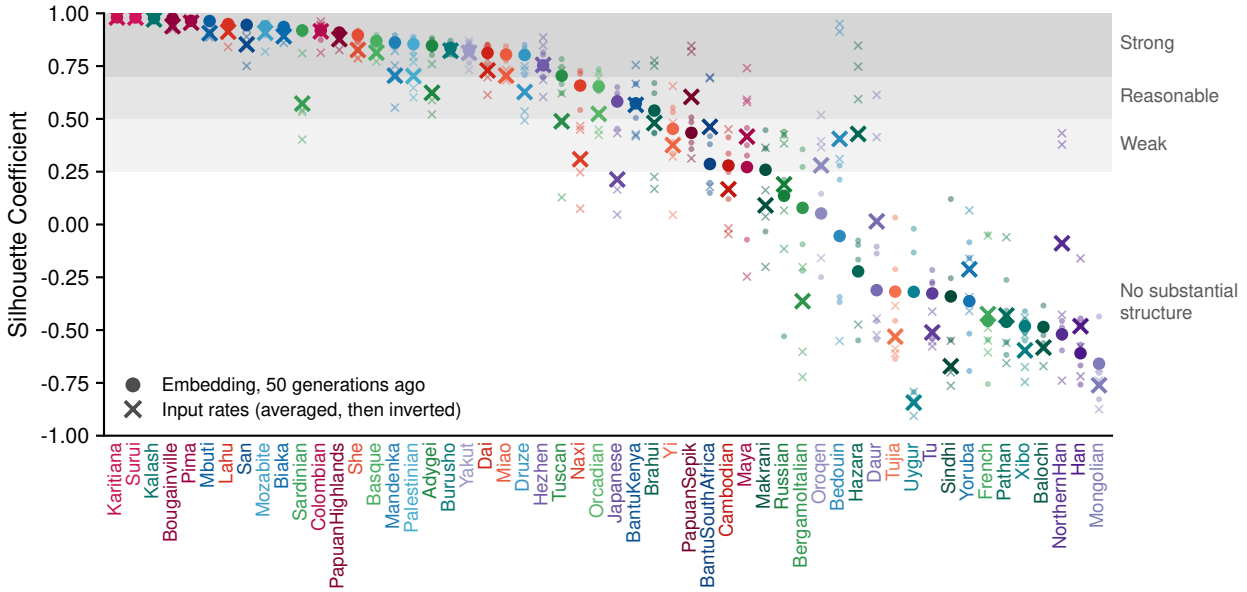

Figure 3.7: The Silhouette Coefficient of each population 50 generations ago. The SC is computed for each sample using distances in the embedding space (small, transparent points) and distances obtained from the rate estimates directly by inverting the link function (small, transparent crosses). The mean SC of samples in each population is shown by a large point and a large cross. The SC computed from the embeddings and from the inverted rates mostly match when they fall above 0.7 (the rough rule of thumb for tight clusters), except for a few exceptions. Below that, when populations do not form tight clusters, the exact set of neighbors in the embedding space tends not to match those obtained from the input rates, so the resulting SCs differ.

From the scatter plot in figure 3.6 and the SCs in figure 3.7, we see that samples assigned the same population label may or may not form clusters in the embedding space. For samples that form clusters, these clusters can either be tight (reflecting a high rate of recent

coalescence due to a small population size, a high rate of endogamy, or a different mechanism) or loose (reflecting a relatively lower rate of recent coalescence that is still higher than with other populations). In figure 3.6(B–G), populations that are tightly clustered, with an SC above 0.7, include Karitiana and Surui in the Americas (F), Kalash in northern Pakistan (D), Lahu in East Asia (E) and Mbuti and Biaka in Africa (B). Karitiana and Surui are Amazonian groups with some of the lowest documented heterozygosity worldwide, possibly resulting from a serial founder effect (Pemberton et al., 2012; S. Wang et al., 2007). The Kalash population is also known to form a distinct cluster in STRUCTURE and PCA (J. Z. Li et al., 2008; Rosenberg et al., 2002), with a maintained low effective population size and weak gene flow with neighboring populations (Ayub et al., 2015). Mbuti and Biaka are African hunter-gatherers that have maintained small long-term effective sizes following the spread of agriculture, and possibly even declined (Bergström et al., 2020; Patin et al., 2009). Samples that form clusters that are not as tight include San and Mandenka in Africa (A,B), Druze in the Middle East (C) and Tuscans and Adygei in Europe (D). The San are a deeply diverged population with a high level of diversity, consistent with their large long-term effective population size (Schlebusch et al., 2012, 2017). Unlike the Mbuti and Biaka hunter-gatherers, the Mandenka are hypothesized to have increased in size as they adopted agriculture (Bergström et al., 2020). The Druze are a religious community closed to converts for about 1,000 years, but founded from a broad Levantine population, resulting in a recently isolated yet genetically diverse group (Shlush et al., 2008). Given these numerous examples, it is clear that many distinct demographic mechanisms can give rise to such clusters in the embedding space in recent time. Regardless of the details of the mechanism, clusters form when members of a group have a higher rate of sharing genetic ancestors among themselves, represented by their coalescent rate, than they do with other groups.

Samples from the same group (assigned the same population label) that do not form clusters in recent time can either stay close to one another in the embedding space without

forming a distinct tight cluster, or be more dispersed as a result of high variability in their coalescent profiles with the rest of the samples. Examples of the first include the samples from the two Bantu-speaking populations of Kenya and South Africa, as well as the Balochi, Brahui and Makrani in Pakistan. The Bantu populations are separated by thousands of kilometers, and are a result of the recent Bantu expansion (Patin et al., 2017). Despite their geographic separation, coalescent rates between them are already quite high by 50 generations ago, presumably as a reflection of the Bantu expansion, but also likely due to the slightly inflated rates preceding a split that result from SINGER. The Balochi and Makrani speak an Iranian language, while the Brahui speak a Dravidian language, yet the three have been shown to genetically cluster together (J. Z. Li et al., 2008; Rosenberg et al., 2002). By 50 generations ago, samples from the three populations are coalescing with one another at very similar rates. Examples of populations with samples that are more dispersed include the Hazara and Uygur in Central Asia, two long-studied examples of admixed populations with components from East and West Eurasia (Hellenthal et al., 2014; J. Z. Li et al., 2008), and the Han and Northern Han in East Asia, which are both part of a larger population with a very sharp recent expansion following agriculture, and a north–south gradient (Chen et al., 2009; Xu et al., 2009). Again, two distinct mechanisms, admixture and rapid recent expansion, lead to similar patterns of spreading in the embedding space. The two can be distinguished by looking at the coalescent rates directly: admixed samples have elevated rates with two distinct groups of samples, whereas samples from a rapidly expanded population have uniformly low rates with all other samples, including one another. In summary, a tight cluster indicates a high rate of recent coalescence within a group, a loose cluster indicates a rate of recent coalescence that is modest but still above that with other groups, and dispersal indicates either admixture or a group whose genetic ancestors barely coalesce with one another at that time relative to their coalescence with genetic ancestors of other groups.

Clustering patterns over time

The clustering patterns described so far arise from a recent time slice of 50 generations ago. The main utility of CoaSTT is the dynamic view it provides of such clustering and similarity patterns in coalescent structure. Figure 3.8 shows how the SC changes over time for each individual, as well as for each population by taking the mean of individual SCs. For comparison, we also show the SC computed from the inverted rates rather than the embeddings as a function of time. As in the recent time slice, the SC computed in both ways mostly matches when it is large and there is a strong clustering signal, and the discrepancy increases as the clustering signal weakens. Clustering indices like this are useful insofar as their exact numerical value is not strongly interpreted but is rather used as a rough measure of relative clustering strength. In fact, the most common pattern in deep time, which is all samples being almost equidistant from one another, is not even realizable in 2D, resulting in the embedding-based SCs decreasing to -1 while the rate-based SCs decrease to 0. However, the general signal of a lack of clustering, given by an SC below 0.7 or 0.5, is still detectable from the embeddings.

The panels in figure 3.8 are sorted in descending order by the time point at which the mean SC of a population first decreases below 0.5. Being the most clustered in recent time does not entail that this clustering is maintained for the longest time. Whereas PCA reflects similarity and clustering patterns that average the whole coalescent history, CoaSTT attaches a time scale to these clustering or similarity patterns (compare Mozabites and Bougainville for example). An example worth focusing on is the San population, which maintains a tight cluster for the longest time, as expected for one of the most diverged lineages. The cluster is maintained for about 1,500 generations, or $\approx 45,000$ years. This is much less than the proposed divergence of the San from other African populations, estimated to be $\geq 100,000$ years ago. The mismatch is a consequence of the suggested gene flow between the deeply diverged southern African populations and other African populations (Ragsdale et al., 2023;

Skoglund et al., 2017), compounded by the tendency of SINGER-derived between-population rates to rise well before the true split (section 3.2). This mismatch directly highlights the effect of gene flow on the timescale at which the coalescent structure collapses for such deeply diverged lineages. The embeddings and the clustering patterns are therefore better understood in terms of the coalescent history of the set of samples under study, rather than a mechanistic demographic history of populations to which these samples are assigned.

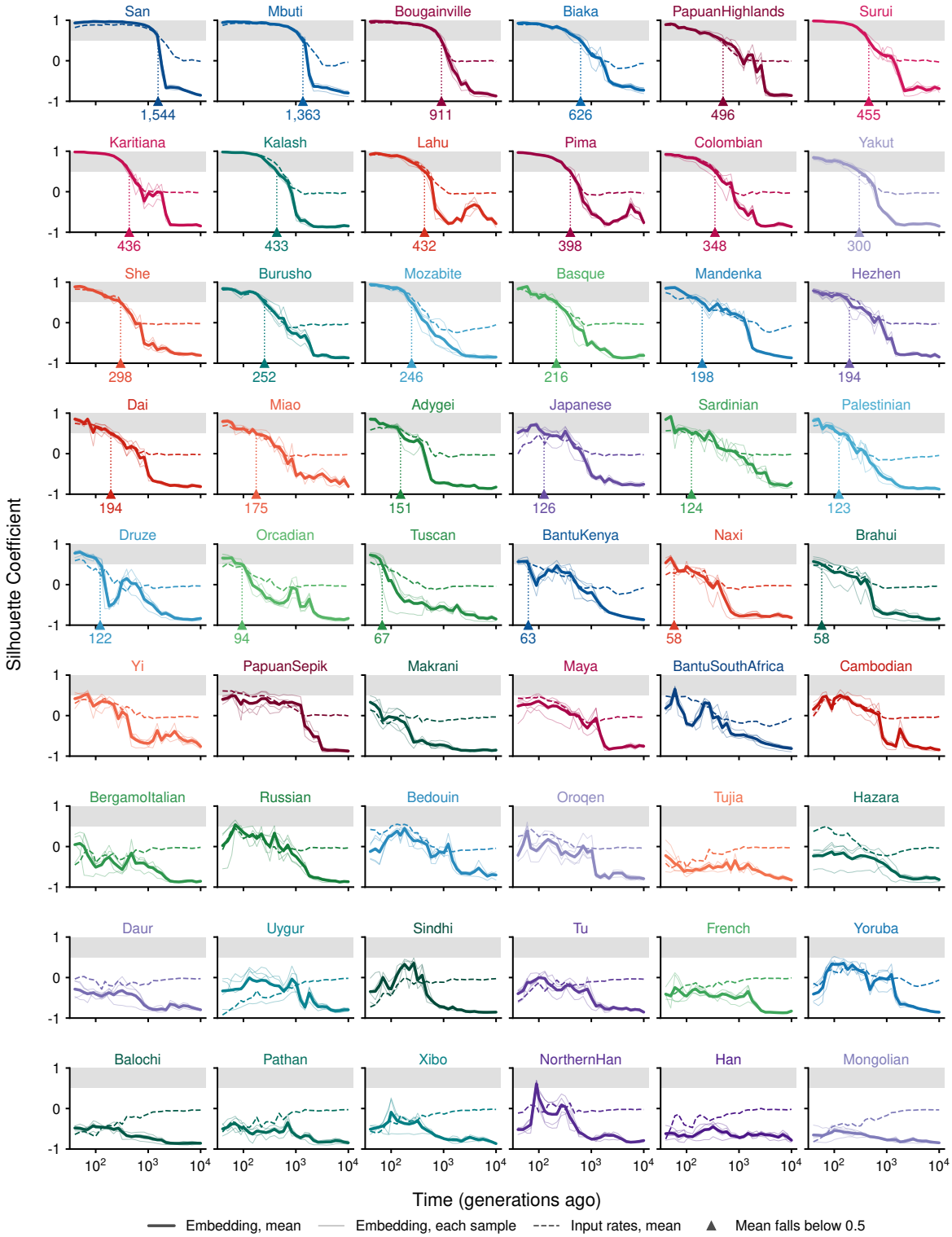

Figure 3.8: The Silhouette Coefficient, computed from embedding distances and inverted rates, as a function of time, for all populations. Panels are sorted in descending order by the time point at which the mean SC of a population first decreases below 0.5.

3.3.2 Comparing input rate estimates and rates reconstructed from the CoaSTT embeddings of the HGDP ARGs

With such low-dimensional representations, it is useful to investigate how much of the observed pattern is an artifact of dimensionality reduction, from $\binom{n}{2} = 23,220$ rate estimates per time slice to $2n = 432$ coordinates. Figure 3.9A shows a density plot of the input rates vs. the rates reconstructed from the embeddings, for all pairs and all time points. On a linear scale, $r^2 = 0.475$ is dominated by rates from deeper time slices, which make up the large majority of points and are of comparable magnitude across all pairs of samples: a geometry that is impossible to realize in 2D. Nevertheless, reconstruction error of this type can be considered mild, as configurations resulting from input rates that are almost identical have a distinct geometry, that of a circular disk. The resulting circular clusters should be interpreted as clusters where samples have almost identical pairwise coalescent rates, without over-interpreting the precise distances between each pair of samples within such clusters. On the log scale, $r^2 = 0.808$, and the Spearman rank correlation is $\rho = 0.842$, both reasonable levels of correlation given the extensive dimensionality reduction.

We can further decompose reconstructed rates that most disagree with the input rates (top 5% outliers on the log-log scale) into three categories (figure 3.9B). Some rate triplets demand distances that are not representable in Euclidean space (of any dimension) as they violate the triangle inequality, which states that for any three samples (i, j, k) in Euclidean space, their distances satisfy $d_{ij} \leq d_{ik} + d_{jk}$. In terms of the rates, this may happen when samples (i, j) are not coalescing at a high rate, whereas (i, k) and (j, k) are, which is also a violation of the transitivity assumption in a latent position model. A common source for such triplets is admixture, where i and j (denoted as "edge samples") belong to source populations A and B that are not coalescing at a high rate, and k (denoted as the "bridge sample") is an admixed individual with genetic ancestors from A and B, such that it has a relatively

high coalescent rate with samples from both A and B at times deeper than the admixture event. Since no configuration in latent space can represent such distances, the loss function compromises in two ways, or a combination of both: the bridge sample either pulls the two edge samples closer together, resulting in an overestimate of their reconstructed rate, or is pushed away from one of the edge samples, resulting in an underestimate of the reconstructed rate between that edge sample and the bridge sample. Which compromise is made depends on the relative strengths of the three coalescent rates involved. The loss function is more sensitive to higher coalescent rates, and is therefore more likely to compromise by misfitting distances that correspond to lower rates. Therefore, we often see the bridge sample pulling the edge samples together, resulting in an overestimate of their rate. Around 20% of the top 5% outliers are of this category, and they are mostly concentrated in recent time.

A second category of outliers is attributed to smoothing. A reconstructed rate is assigned to this category when it is not an outlier in the unsmoothed embeddings but gets pushed into the outlier zone after minimizing the loss with the smoothing term. Around 15% of the top 5% outliers are of this type, and they are dispersed over all time points. Such error is not necessarily undesirable, since smoothing prevents overfitting to highly fluctuating noisy rate estimates, especially in recent time.

The remaining 65% of outliers cannot be assigned a particular category, but are likely due to the crowding effect that is dominant in deep time. In figure 3.9C, the correlation between input and reconstructed rates is computed at each time point, and shows a monotonic decrease beyond $\sim$ 1,000 generations ago. Looking at the standard deviation of the input rates over that period, we see that it also starts to decrease, reflecting that all rates are starting to become more similar (figure 3.9D). The decline in correlation between input and reconstructed rates in deep time is therefore mostly attributable to the inability of the embeddings to fit many samples with almost identical pairwise distances in two dimensions.

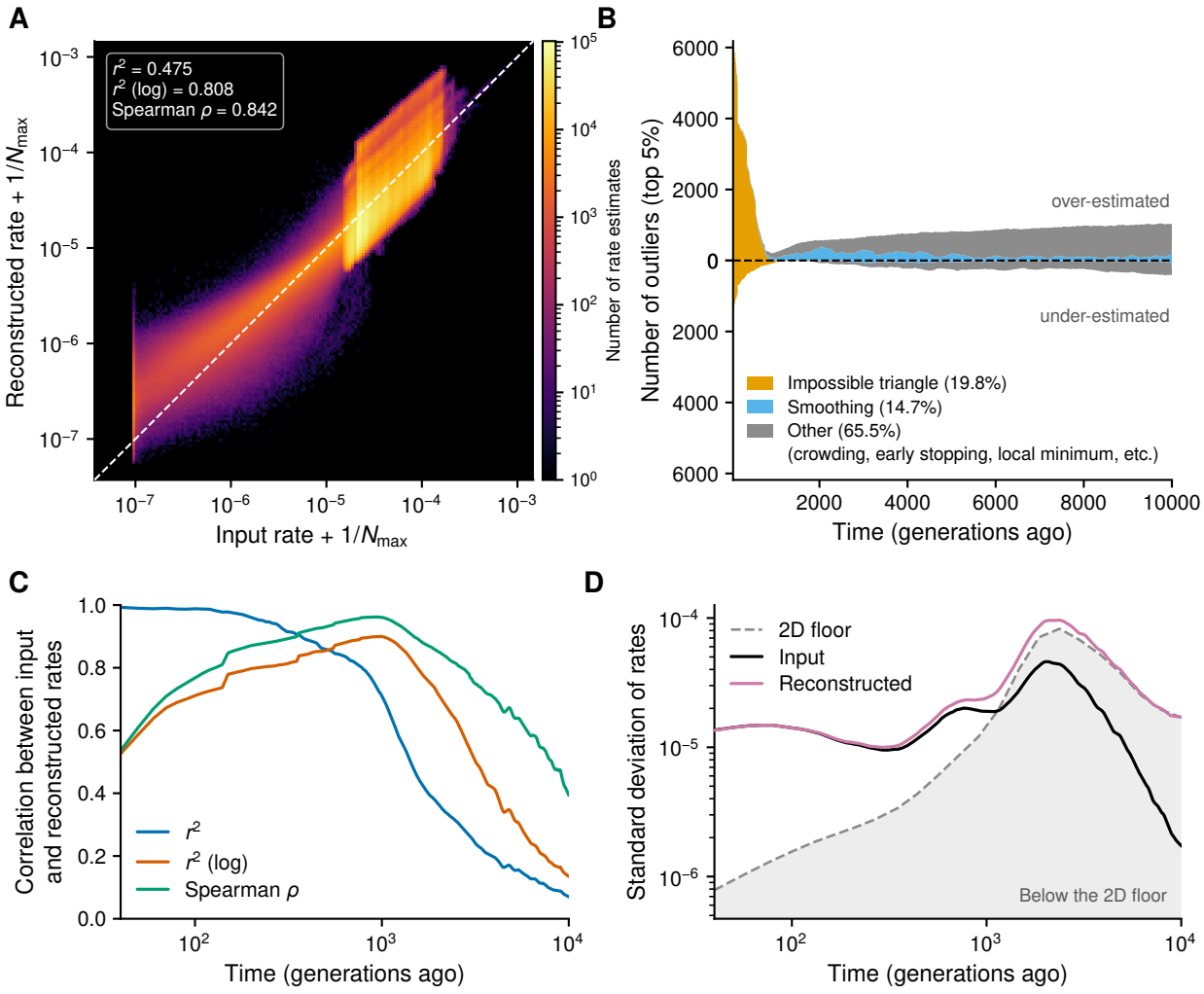

Figure 3.9: Probing the mismatch between input rate estimates and rates reconstructed from human CoaSTT embeddings. A) Density plot of rates reconstructed from embedding distances and input rate estimates, for all pairs and time points, along with three measures of correlation. The spike of input rates at 10^{-7} is because $N_{\max} = 10^7$, so any rate below 10^{-7} gets pushed to that value. **B)** Decomposition of the source of mismatch between input and reconstructed rates for the top 5% outliers (on the log scale) over all time points. An impossible triangle is a triplet of distances that is not realizable in Euclidean space, likely attributable to admixture, and dominates recent time. **C)** Three measures of correlation between input and reconstructed rates at each time point. The drop in correlation beyond 10^3 generations ago is likely due to crowding. **D)** The standard deviation of the input and reconstructed rates over time, as well as the standard deviation of the reconstructed rates from a configuration fitted to an input of identical pairwise rates, equal to the mean rate at each time point (labeled as 2D floor).

Triplets that compose impossible triangles

Triplets of samples with pairwise rates that demand distances violating the triangle inequality are numerous, and deserve a separate investigation. Identifying these triplets can serve as a useful complement to the CoaSTT embeddings, marking more clearly information that could not be adequately conveyed by the embeddings. Denoting the link function by $f(\cdot)$, we mark a triplet of samples (i, j, k) as part of an impossible triangle if

$$f^{-1}(\hat{\lambda}_{ij}) > f^{-1}(\hat{\lambda}_{ik}) + f^{-1}(\hat{\lambda}_{jk}), \tag{3.2}$$

where $\hat{\lambda}$ is the input rate estimate. Samples (i, j) are the edge samples and the intermediate sample k is the bridge sample. Under the inverse-distance link (and dropping the small factors of $N_{\min}$ and $1/N_{\max}$ for simplicity), this approximately entails

$$\hat{\lambda}_{ij} < \frac{\hat{\lambda}_{ik}\hat{\lambda}_{jk}}{\hat{\lambda}_{ik} + \hat{\lambda}_{jk}}. \tag{3.3}$$

That is, the coalescent rate estimate between the edge samples is less than half of the harmonic mean of the rates between each of the edge samples and the bridge.

We can additionally quantify the magnitude by which the triangle inequality is violated. For a given time point, we define the degree of violation as

$$v(i,j;k)(t) = \begin{cases} \ln \dfrac{f^{-1}(\hat{\lambda}_{ij}(t))}{f^{-1}(\hat{\lambda}_{ik}(t)) + f^{-1}(\hat{\lambda}_{jk}(t))}, & \text{if } f^{-1}(\hat{\lambda}_{ij}(t)) > f^{-1}(\hat{\lambda}_{ik}(t)) + f^{-1}(\hat{\lambda}_{jk}(t)), \\ 0, & \text{otherwise,} \end{cases} \tag{3.4}$$

and an average violation over time as

$$\bar{v}(i,j;k) = \frac{1}{T}\sum_{t=1}^{T} v(i,j;k)(t). \tag{3.5}$$

A heatmap of the average violation for every triplet of samples is shown in figure 3.10. Rows of the heatmap represent edge samples (i,j), while columns represent the bridge sample k. A large number of triplets have a non-zero value in this heatmap. To highlight triplets with the most extreme violation of the triangle inequality, any value below the 95th percentile of the non-zero elements is set to the same color.

The most prominent triplets involve bridges that belong to populations that are well known to be admixed. For instance, Bedouin, Mozabite and Palestinian bridges have corresponding edges where one of the samples is African, reflecting the high coalescent rate between those bridges and both African and non-African samples simultaneously, compared to the relatively lower coalescent rate between the African and non-African samples on the edges of the triangle. Another two prominent populations with samples acting as bridges are the Hazara and Uygur, with edges involving samples from (Middle East, East Asia), (Europe, East Asia) and (Central/South Asia, East Asia). There are a few interesting individual samples that act as bridges as well, such as a Mayan individual with edges from a wide range of samples from all regions, consistent with the post-colonial European admixture documented in the HGDP Maya (Rosenberg et al., 2002; S. Wang et al., 2007), Bougainville individuals with edges from East Asian and Oceanian populations, and a few Makrani, Balochi and Brahui individuals with edges involving samples from Africa. African ancestry resulting from recent admixture is documented in Makranis of Pakistan (Laso-Jadart et al., 2017), but the Makrani samples in the HGDP correspond to the Makrani Baluch, a distinct community with substantially lower and more heterogeneous African ancestry ($\sim 6.7\%$, SD 10.7%) than the African-descent Makranis ($\sim 25.5\%$) studied by Laso-Jadart et al. (2017). The same study documents weak African ancestry among the Balochi and Brahui.

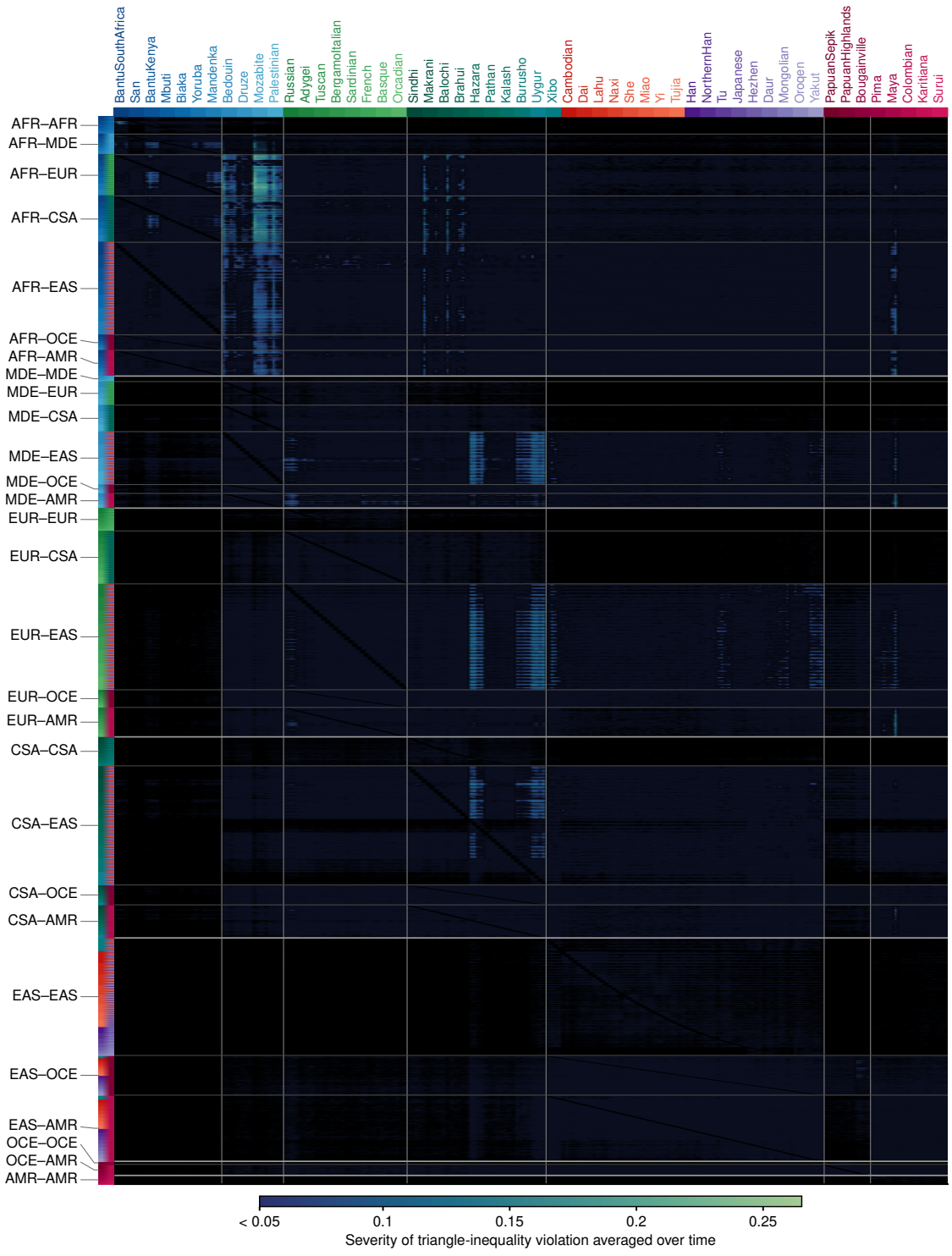

Figure 3.10: Triplets of human samples most violating the triangle inequality. Each element is the average violation $\bar{v}(i, j; k)$ over all time points for the triplet formed by the edge samples (i, j) on the row and the bridge sample k on the column, shown for all triplets. Values below the 95th percentile of the non-zero elements are collapsed to a single color.

3.3.3 Visualizing uncertainty in the embeddings

CoaSTT embeddings have three sources of uncertainty. The first is associated with uncertainty in the rate estimates derived from the inferred ARG. The second is due to the noise associated with rates estimated from a finite number of events in a fixed ARG, and is taken into account by the weight terms in the loss function (equation 2.15). The third arises because the loss function minimized by CoaSTT is highly non-convex, so different initializations can lead to different local minima. We explore the first source of uncertainty by recomputing the embeddings from the same initialization, with rates estimated from different posterior draws of ARGs provided by SINGER. SINGER was run for 3,000 MCMC iterations as a burn-in, then a posterior sample of the tree sequence was taken every 250 MCMC iterations, for a total of 8 posterior samples. The uncertainty associated with the optimization itself is explored by rerunning the optimizer on the same rates from different initializations. Since the optimization proceeds in two phases (section 2.3.3), the random initialization enters only at the very first step, the optimization of the deepest time slice during sequential optimization; every subsequent step, including the final smoothed optimization, is initialized from a previous solution.

Figures 3.11 and 3.12 show the embeddings resulting from different posterior draws of the ARG and from different initializations respectively, each Procrustes-aligned to the solution shown in figure 3.5. The 3D embeddings look qualitatively similar, so it is instructive to look at the variability in the embeddings at two representative time slices (figures 3.13 and 3.14). In recent time (50 generations ago), the different solutions obtained either from different initializations or from different SINGER posterior samples are highly similar, with a few clusters being rotated as a result of the rates not being informative enough to pin them down to a single configuration. For instance, the San samples always cluster further from the other African samples, but sometimes get placed on one side of the Mbuti cluster, and sometimes on the other side. Two other obvious examples of this are the Mozabite

and the Adygei clusters. They both cluster at the edge of their respective regions but are not pinned down to a particular orientation. At the deeper time slice (2,000 generations ago), the embeddings are qualitatively similar over the replicates, with a central non-African cluster surrounded by a crescent of African samples. In all replicates, the crescent is slightly layered, with the inner layer closer to the central cluster composed of Bantu, Mandenka and Yoruba samples, and the outer layer formed by a less structured set of Mbuti, Biaka and San samples. In some replicates, the San samples form a loose cluster, and in others they do not. The inner layer forms because Bantu, Mandenka and Yoruba samples consistently have a higher coalescent rate with non-Africans 2,000 generations ago than the San and the rainforest hunter-gatherers (Mbuti and Biaka) do, consistent with the deeper divergence of the latter lineages from the ancestors of West Africans and non-Africans (S. Fan et al., 2019; Lipson et al., 2020; Ragsdale et al., 2023). The outer layer is less structured because the rates there barely differ, with Mbuti and Biaka having very similar rates at that time in all SINGER samples. The San samples have a consistently elevated coalescent rate among themselves compared to their rates with other African samples (approximately 2× that of the closest African samples), in all SINGER samples. This elevation does not seem strong enough for the loss to pin down the positions of the San samples in the 2D embeddings. As the optimizer attempts to satisfy all the other constraints, it occasionally compromises by pushing the San samples around because their coalescent signal is not strong enough to discourage it.

Ignoring the weight parameters, the data term of the loss function (section 2.3.2) is

$$l(X) = \sum_{i,j,t} f(\|X_i(t) - X_j(t)\|) - \hat{\lambda}_{ijt} \ln f(\|X_i(t) - X_j(t)\|). \tag{3.6}$$

This loss is not equally sensitive to all pairs. Its curvature with respect to f is $\hat{\lambda}_{ijt}/f^2$, so a given relative misfit in f is penalized in proportion to the coalescent rate: pairs with

high rates are fit tightly, while pairs with low rates can be misfit at little cost. This is the opposite of classical MDS applied to inverted rates, where the stress is dominated by the largest distances, i.e. the smallest rates. As a consequence, the positions of samples relative to those they coalesce with rapidly are well constrained by the data, whereas their positions relative to samples they coalesce with slowly are only weakly constrained. These weakly constrained directions act as soft modes of the configuration. The optimizer can move along them at little cost to the data term, and they are correspondingly more influenced by the smoothing term and by the initial solution. The two sets of replicates therefore separate the two sources of uncertainty. A feature that varies across initializations is weakly determined by the rates, so its uncertainty lies in the fit, whereas a feature that is stable across initializations but varies across ARG posterior samples is well determined by any one set of rates but sensitive to which rates are supplied, so its uncertainty lies in the data.

We therefore recommend running CoaSTT on several posterior samples of the ARG and from several initializations whenever possible, and reading a feature as robust only when it is stable across both. When the sample size is too large for sampling from the posterior distribution of ARGs to be feasible, we recommend at least running CoaSTT from several initializations.

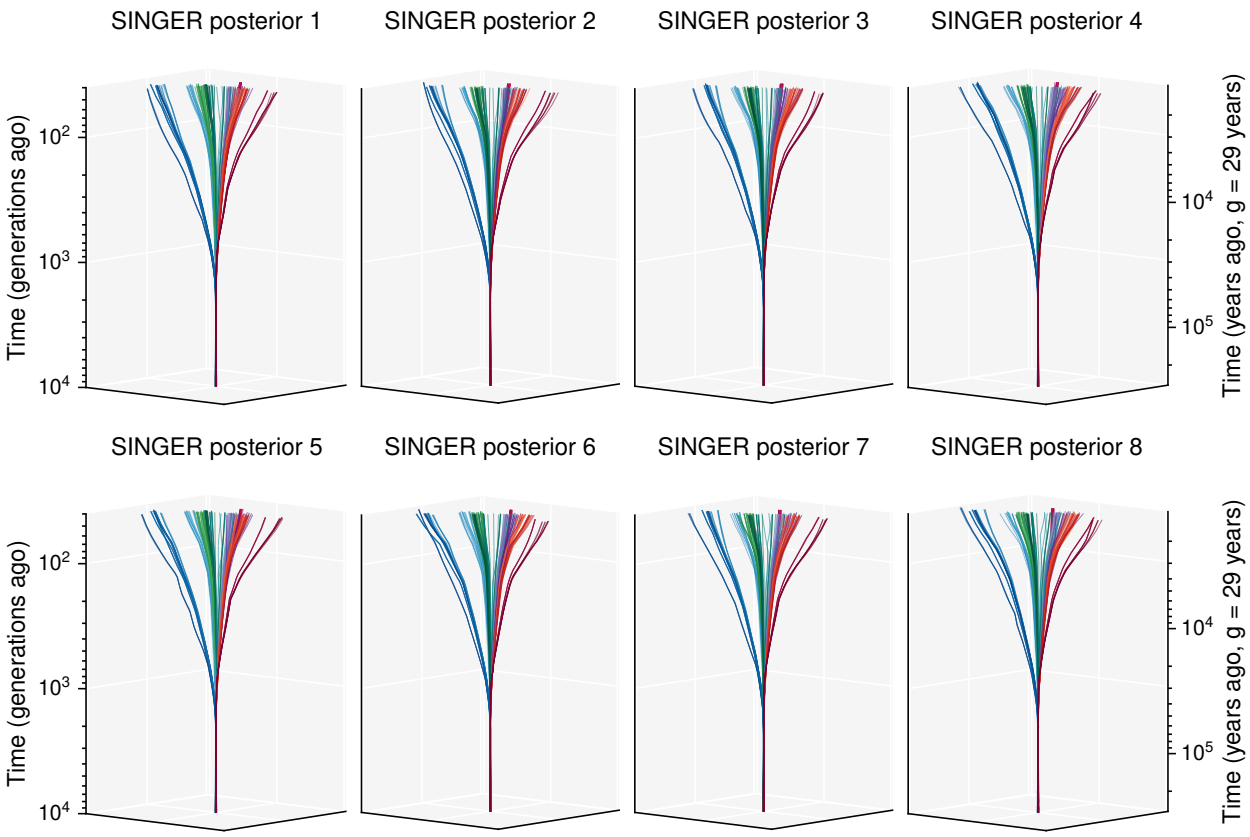

Figure 3.11: CoaSTT embeddings from 8 posterior ARG samples for the 54 HGDP populations.

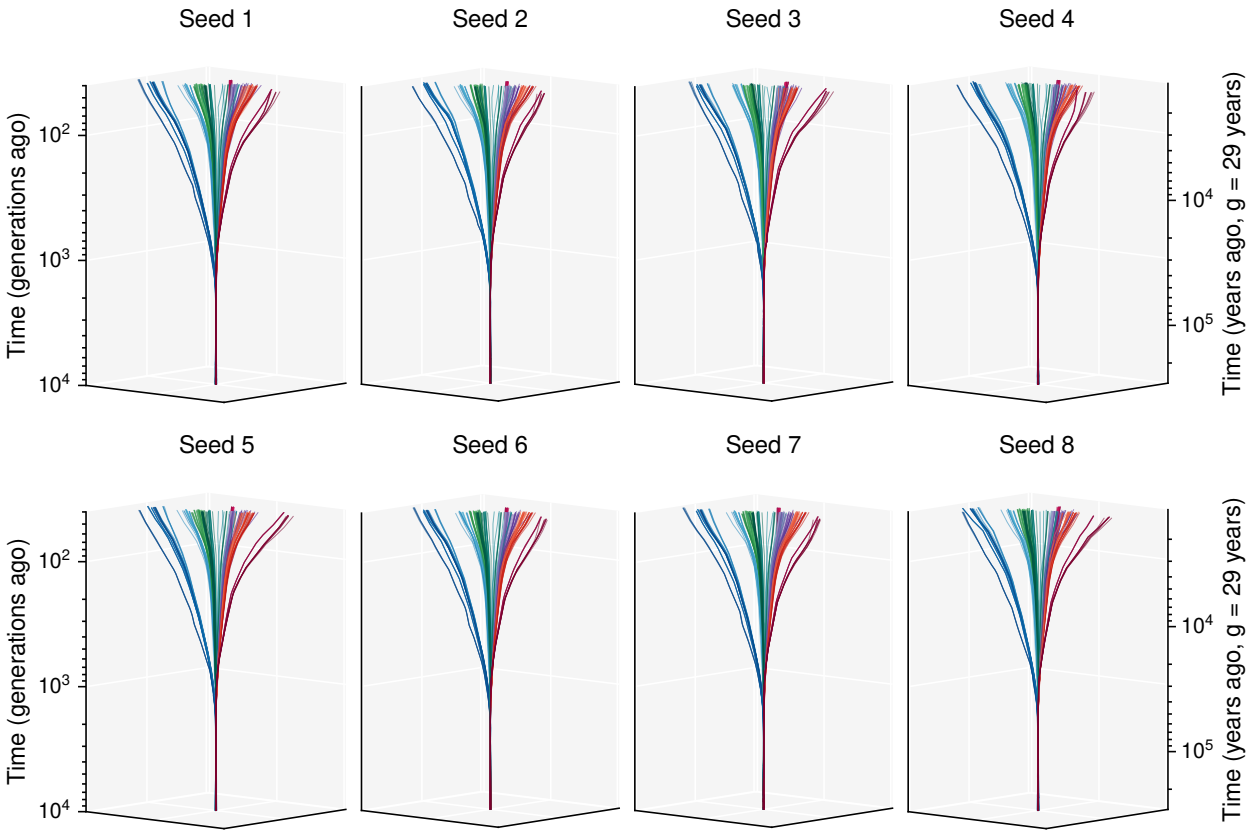

Figure 3.12: CoaSTT embeddings from 8 different initializations for the 54 HGDP populations.

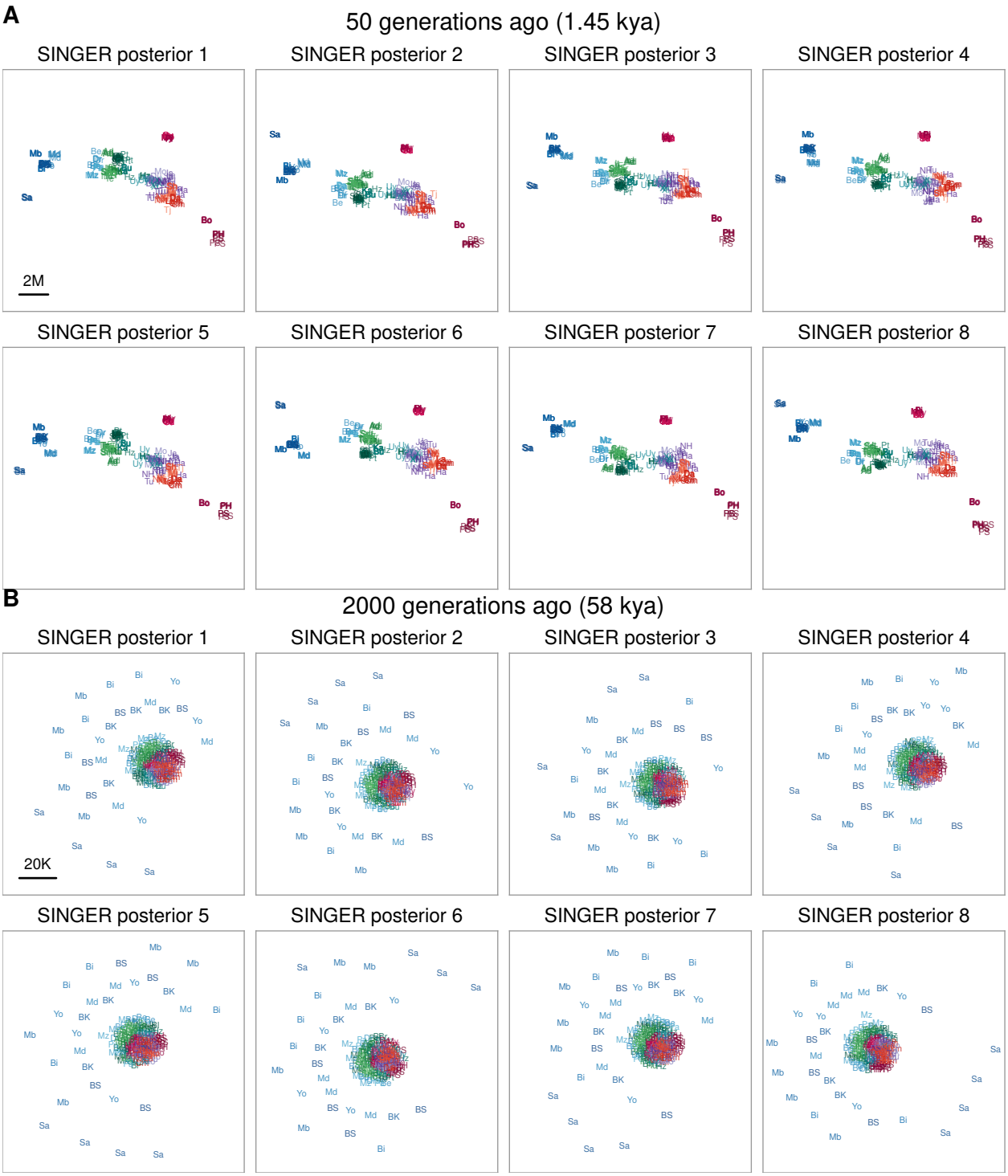

Figure 3.13: CoaSTT embeddings at two different time slices from 8 posterior ARG samples for the 54 HGDP populations.

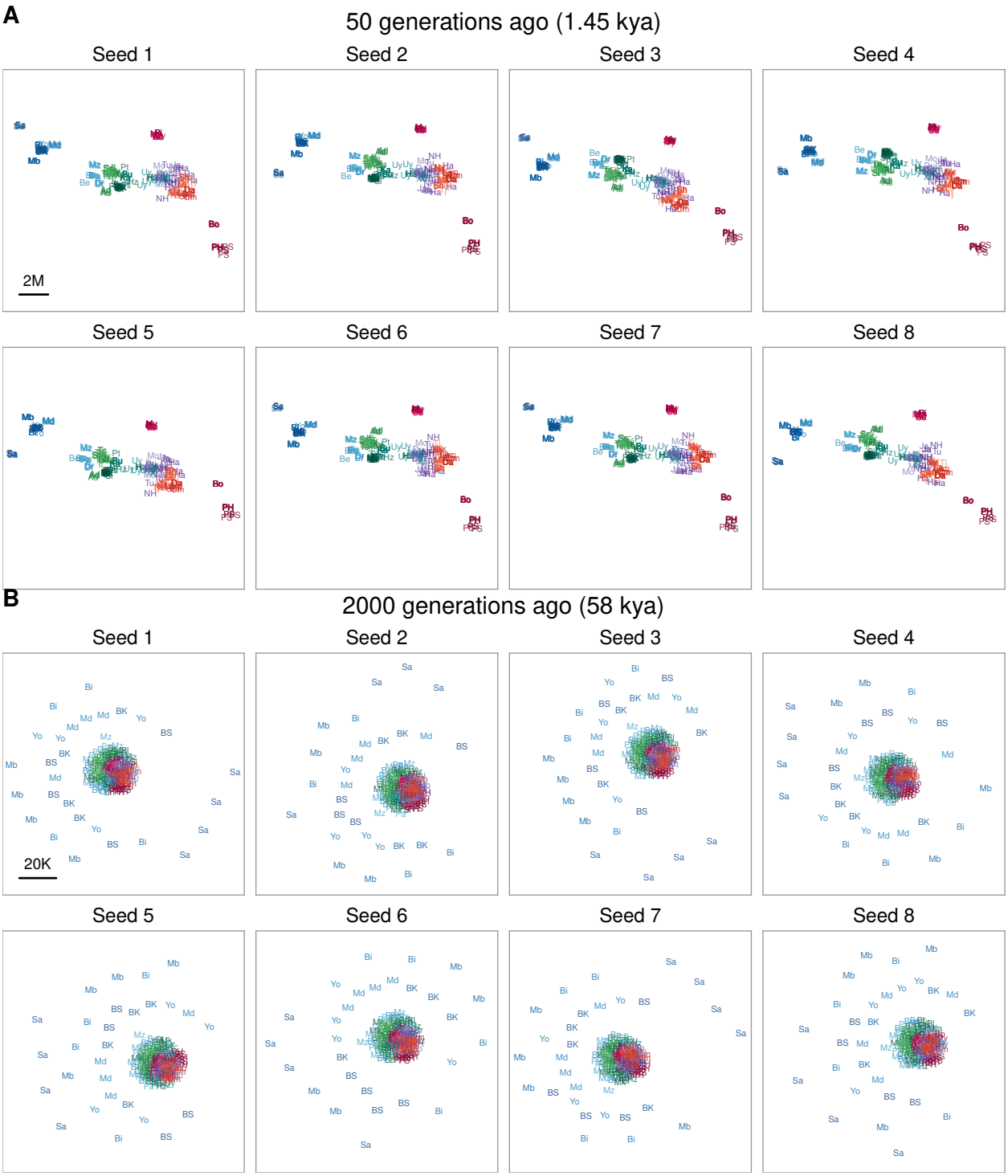

Figure 3.14: CoaSTT embeddings at two different time slices from 8 different initializations for the 54 HGDP populations.

3.3.4 Regional analysis of human population history with CoaSTT

In addition to running CoaSTT on samples from all 54 HGDP populations, we independently inferred ARGs using a subset of samples restricted to each of the seven regional labels in the HGDP dataset and ran CoaSTT on them. For some of the regions, the sample size was still too high for SINGER, so we further subsampled by choosing the set of n samples with the highest coverage in some populations. The number of samples n chosen from each population in each region is provided under the corresponding figure captions. This section is not meant to provide a comprehensive analysis of the CoaSTT embeddings obtained from each region. Rather, its main purpose is to demonstrate how applying CoaSTT to a smaller subset of populations with a higher number of samples per population can reveal more detailed patterns in coalescent history.

Figures 3.15–3.21 show 3D CoaSTT embeddings, as well as four representative time slices, for populations from Africa, the Middle East, Europe, Central and South Asia, East Asia, Oceania and the Americas respectively. An immediately clear pattern from comparing the 3D embeddings is the presence of the sharp out-of-Africa bottleneck around 2,000 generations ago ($\sim$ 58 kya) in all regional embeddings except for the African embeddings. Additionally, for non-African populations, the time of regional diversification can be read off the 3D embeddings. For instance, the American embeddings show that the recent structure of the ancestral lineages of present-day populations started emerging much more recently than in other regions, only about 400 generations ago ($\sim$ 11 kya; figure 3.21), consistent with the diversification of Native American lineages following their entry into the Americas around 15 kya (Moreno-Mayar et al., 2018; Raghavan et al., 2015).

The details of this process of diversification can also be explored by looking at the embeddings at different time slices. For instance, the Oceanian populations had already formed two major clusters (the Papuans and the Bougainville) by 700 generations ago ($\sim$ 20 kya), but the Papuan populations started forming two distinct clusters more recently, around 200

generations ago ($\sim$ 6 kya) (figure 3.20), matching the reported separation of highland and lowland New Guinean groups by 10–20 kya and the diversification of highland groups during the Holocene (Bergström et al., 2017). Embeddings in recent time slices show how different groups might have had strikingly different ancestral histories. Following the same Oceanian example, samples from the Papuan Highlands form a well-defined, tight cluster in recent time, whereas samples from the Papuan Sepik are spread out more in the embedding space, forming three different clusters, one of which has a single individual markedly pulled towards the Bougainville cluster, likely due to admixture (figure 3.20). Another clear example is the Bedouin population in the Middle East. In recent time, two or three distinct clusters can be seen, with five other individuals more scattered, one of them projecting at the very edge of the embedding space and only joining the regional cluster much deeper in time (figure 3.16). The fine structure within the Bedouin samples of HGDP is well-documented, with three clusters also previously reported and attributed to tribal co-ancestry (Lawson et al., 2012; Markus et al., 2014). The outlier at the edge coalesces more slowly with all other Middle Eastern samples, which we interpret as elevated African ancestry, since heterogeneous levels of African ancestry have been reported for this set of Bedouin samples (Moorjani et al., 2011). The same can be said about the Mozabite outlier in figure 3.16.

With more samples per population, the recent regional substructure is revealed in higher detail. For instance, in the European embeddings, clear clusters are formed by Sardinian, Russian and Adygei samples in recent time and are located at the edge, whereas Orcadian and Basque samples form even tighter clusters but fall in the middle (figure 3.17). The tightness of the latter two clusters reflects the documented isolation of both populations (Flores-Bello et al., 2021; Lawson et al., 2012), while their central position indicates that, unlike the populations at the edge, they coalesce at comparable rates with the other European groups. The location of the Basque cluster closer to the side of the Sardinian cluster is an expected signal, as it matches genetic, linguistic and archaeological evidence of the connection

between both populations (Chiang et al., 2018). The French samples on the other hand are quite dispersed throughout the European embeddings, highlighting the high diversity in the coalescent history of present-day HGDP French individuals.

Much more can be said about these regional examples, especially regarding the rich East Asian and Central/South Asian embeddings. We invite the reader to explore the details of these embeddings, with an emphasis on how such an individual-level description of population history provides an avenue for a more nuanced understanding of the history of humans and other populations. Regional and global analyses are complementary, as more samples per population resolve within-region structure and outliers, but an outlier's affinity to populations outside the region can only be read from an ARG that includes them.

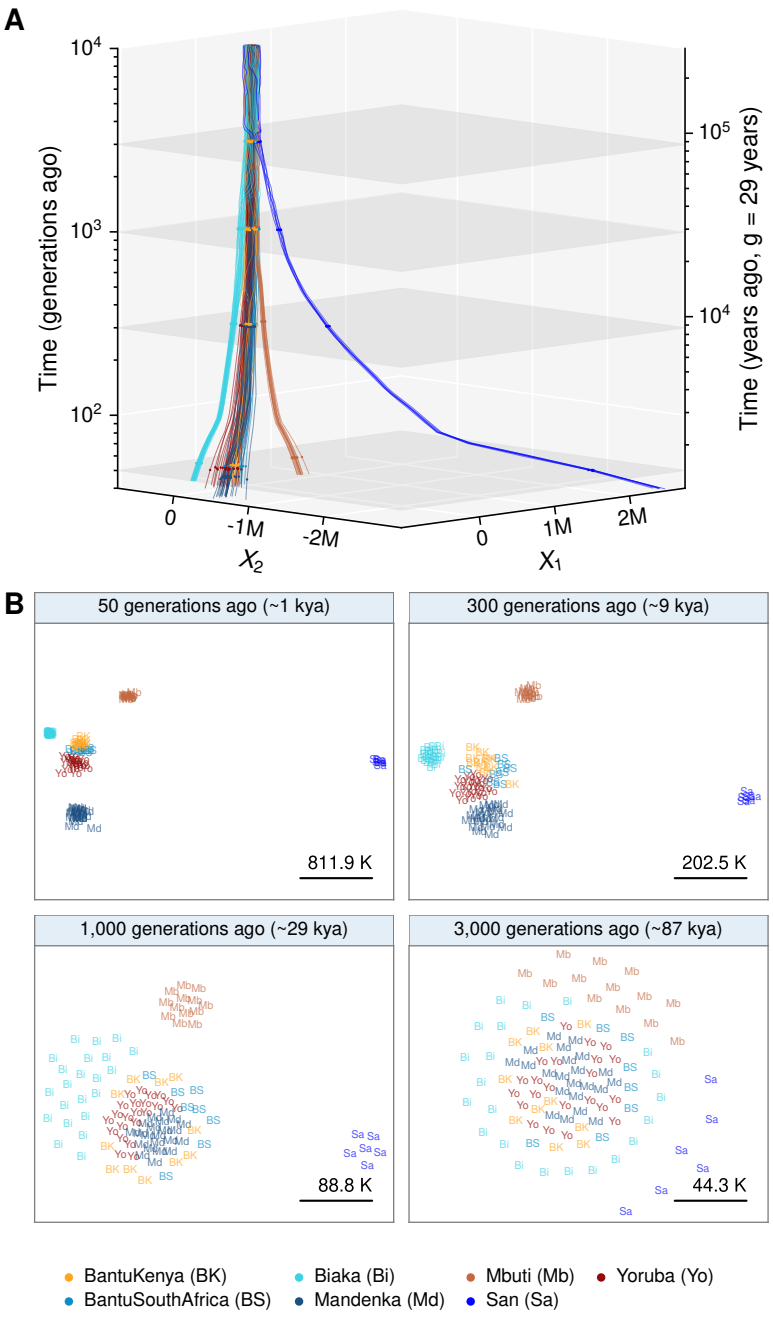

Figure 3.15: CoaSTT embeddings for 104 samples from 7 populations in Africa. Biaka (n=22), Mandenka (n=22), Yoruba (n=22), Mbuti (n=13), BantuKenya (n=11), BantuSouthAfrica (n=8) and San (n=6).

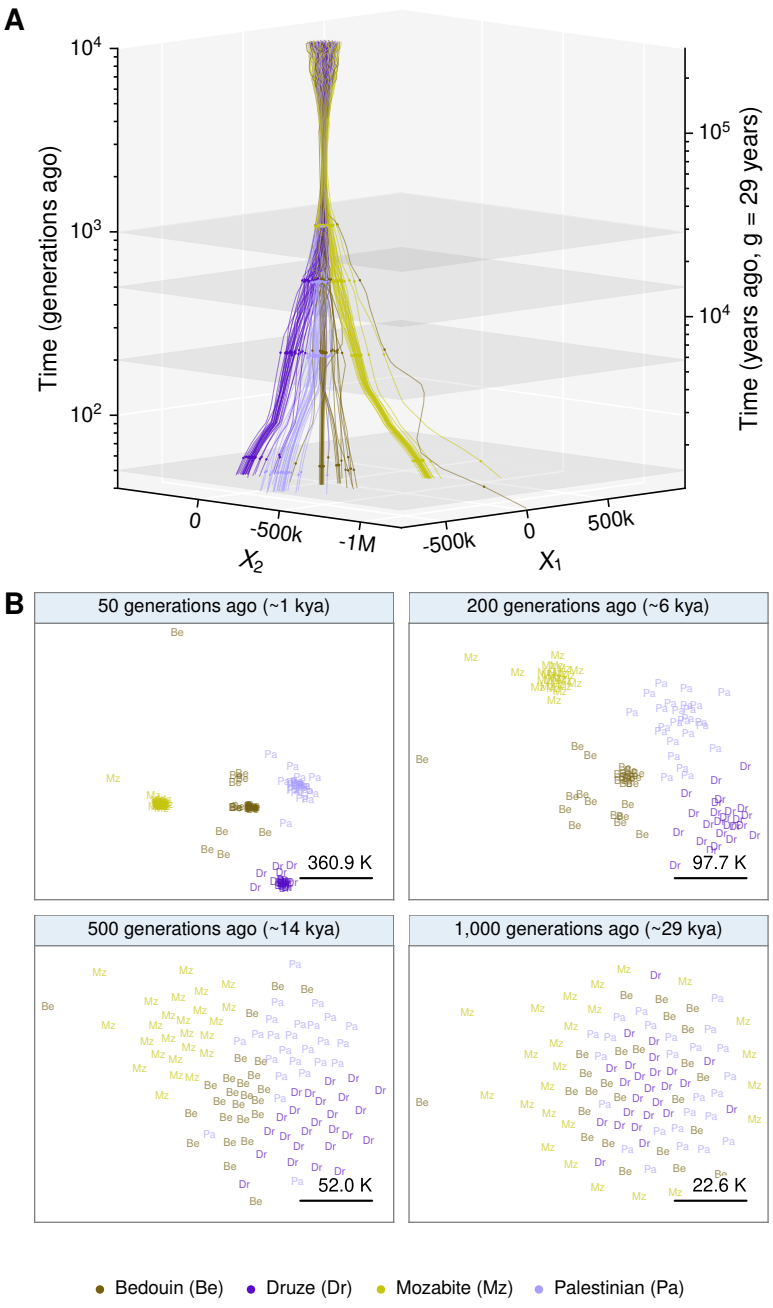

Figure 3.16: CoaSTT embeddings for 100 samples from 4 populations in the Middle East. Bedouin (n=25), Druze (n=25), Mozabite (n=25) and Palestinian (n=25).

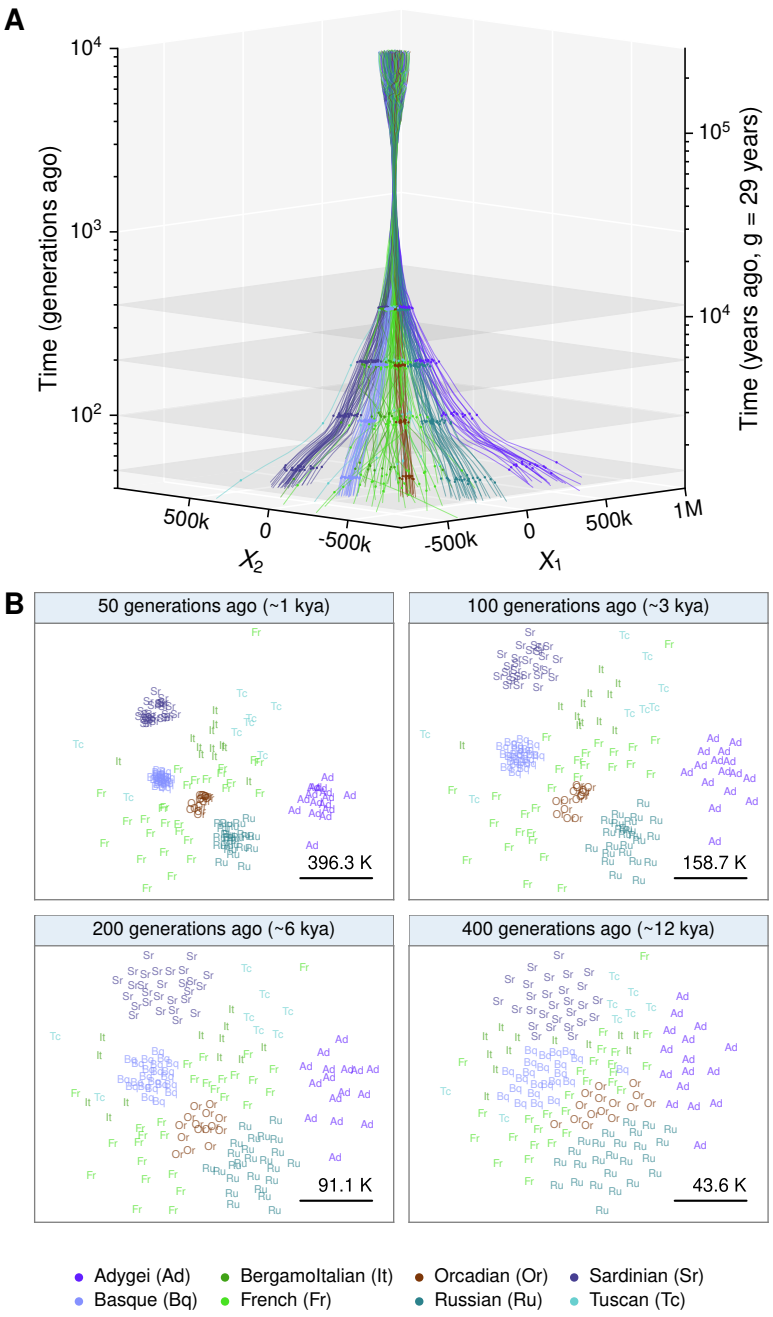

Figure 3.17: CoaSTT embeddings for 155 samples from 8 populations in Europe. French (n=28), Sardinian (n=28), Russian (n=25), Basque (n=23), Adygei (n=16), Orcadian (n=15), BergamoItalian (n=12) and Tuscan (n=8).

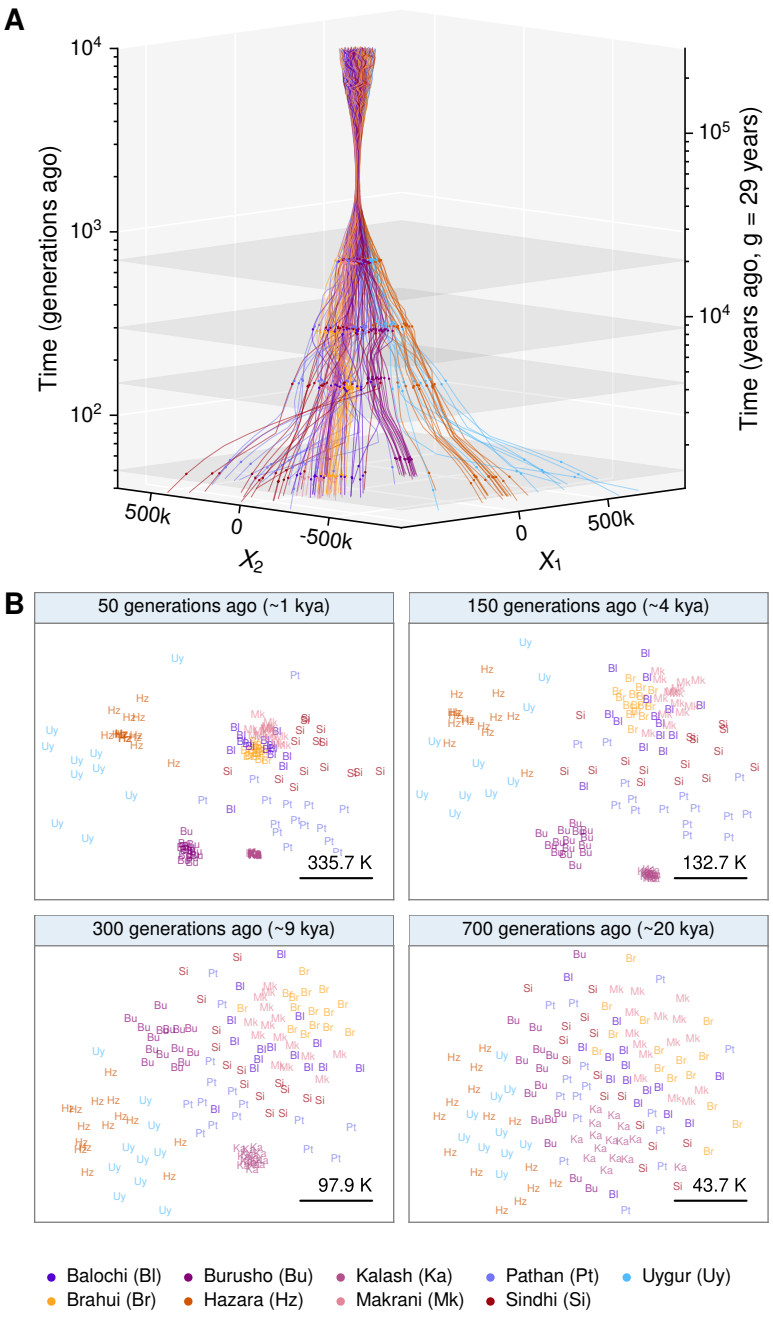

Figure 3.18: CoaSTT embeddings for 130 samples from 9 populations in Central and South Asia. Balochi (n=15), Brahui (n=15), Burusho (n=15), Hazara (n=15), Kalash (n=15), Makrani (n=15), Pathan (n=15), Sindhi (n=15) and Uygur (n=10).

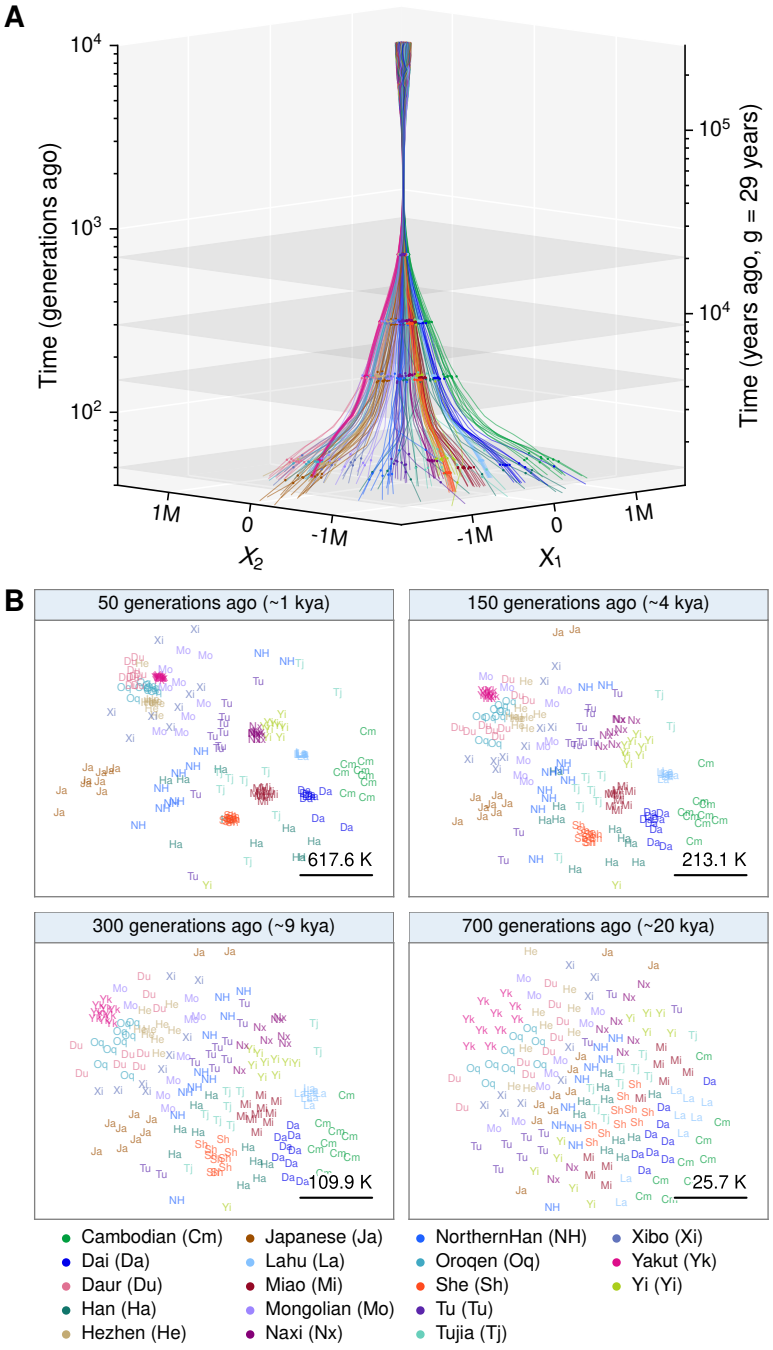

Figure 3.19: CoaSTT embeddings for 168 samples from 18 populations in East Asia. Han (n=10), Japanese (n=10), Miao (n=10), NorthernHan (n=10), She (n=10), Tu (n=10), Yakut (n=10), Yi (n=10), Cambodian (n=9), Dai (n=9), Daur (n=9), Hezhen (n=9), Mongolian (n=9), Oroqen (n=9), Tujia (n=9), Xibo (n=9), Lahu (n=8) and Naxi (n=8).

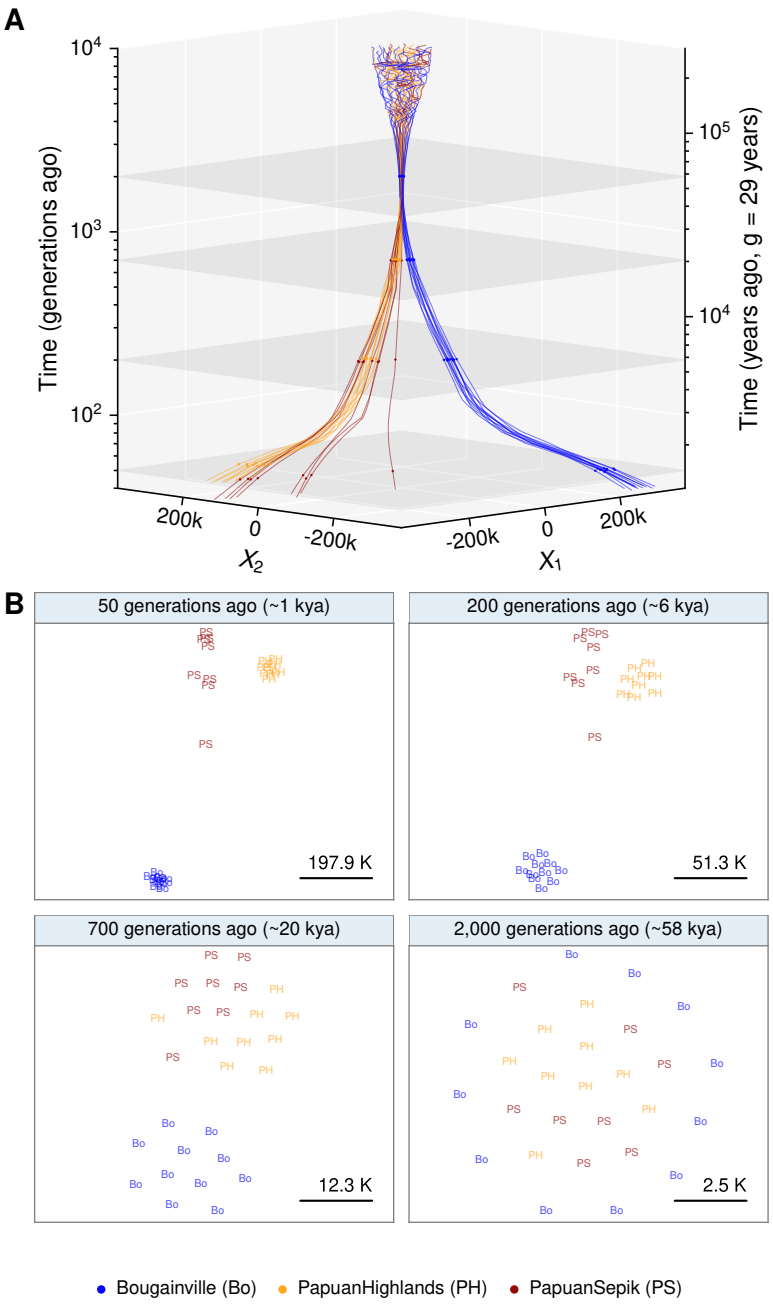

Figure 3.20: CoaSTT embeddings for 28 samples from 3 populations in Oceania. Bougainville (n=11), PapuanHighlands (n=9) and PapuanSepik (n=8).

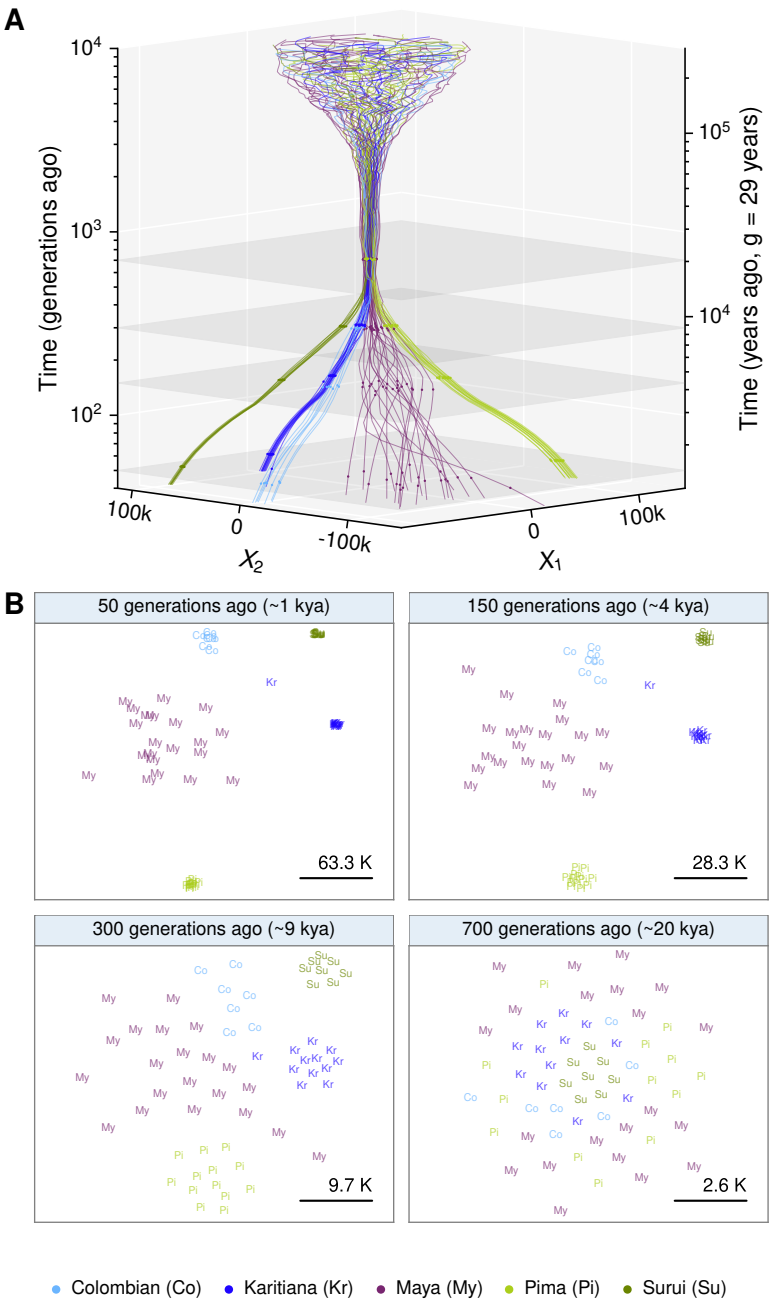

Figure 3.21: CoaSTT embeddings for 61 samples from 5 populations in the Americas. Maya (n=21), Pima (n=13), Karitiana (n=12), Surui (n=8) and Colombian (n=7).

3.4 Application of CoaSTT to chimpanzee data reveals recent and deep structure within subspecies of chimpanzees

We apply CoaSTT to ARGs inferred for a set of chimpanzee samples to explore the utility of an individual-based depiction of coalescent history in understanding the recent and deep history of a non-human species. The genetic history of chimpanzees offers an excellent case study for CoaSTT, as it combines a rich literature against which we can validate our results with a number of unresolved questions. As early as the 1990s, mtDNA data established the existence of three separate subspecies of chimpanzees (Morin et al., 1994) and led to the proposal of a fourth (Gonder et al., 1997). The western (*P. t. verus*) and Nigeria-Cameroon (NC; *P. t. ellioti*) subspecies form one clade, and the central (*P. t. troglodytes*) and eastern (*P. t. schweinfurthii*) subspecies a second, more recently diverged one. The present-day ranges of these subspecies are separated by prominent geographic barriers: the western and NC ranges are separated by the Dahomey Gap, a belt of savanna interrupting the West African forest across Togo, Benin and Ghana; the NC and central ranges by the Sanaga River in central Cameroon; and the central and eastern ranges by the Ubangi River, along the border between the Republic of the Congo and the Democratic Republic of the Congo.

Isolation-with-Migration (IM) models suggested that gene flow had persisted between the ancestors of these four subspecies, with uncertainty regarding the extent and period of this gene flow, and the strength of evidence supporting it (Becquet et al., 2007; Won & Hey, 2005). With the arrival of genome-scale data and larger sample sizes, split times between subspecies were better resolved, the presence of ancestral gene flow among subspecies was established with more certainty, and the population size history of the different subspecies was better understood (Caswell et al., 2008; Hey, 2010; Prado-Martinez et al., 2013). Whole-genome sequencing allowed older gene flow events to be established, for instance from bonobos into the ancestors of eastern and central chimpanzees, as well as between NC chimpanzees and

the central and eastern lineages (de Manuel et al., 2016).

The distinctiveness of the ancestral history of the NC subspecies has been the focus of several studies, which have established its strong differentiation from the other subspecies, the persistence of gene flow with them, and posited the role of local adaptation in shaping the present-day NC population (Bowden et al., 2012; Gonder et al., 2011; Mitchell et al., 2024). The divergence time between NC and western chimpanzees seems uncertain. Autosomal microsatellites give an NC–western TMRCA of 0.46 Mya (Gonder et al., 2011), while MSMC2 applied to the X chromosomes of males places the end of gene flow between the two lineages near 0.25 Mya (de Manuel et al., 2016). These are not strictly commensurable, as they derive from different markers, estimators and chromosome classes, and the microsatellite dates assume a 20-year generation time (vs. 25 years).

Most of these studies explored the ancestral history of chimpanzees by grouping the present-day individuals into four subspecies, estimating parameters such as time-varying population sizes, split times, pulse admixture and continuous gene flow in terms of these population units. More recent studies have begun to appreciate the extent of within-subspecies structure in chimpanzees, arguing that treating the four subspecies as panmictic units is a substantial oversimplification (Fontsere et al., 2022; Steux et al., 2025). We explore the ancestral history of the four subspecies of chimpanzees using CoaSTT, with a focus on recent within-subspecies substructure and a more in-depth investigation of the history of the ancestors of the NC subspecies.

3.4.1 Description of the chimpanzee dataset and ARG inference using SINGER

We infer ARGs using SINGER + POLEGON for 54 publicly available whole-genome sequences: 14 western, 19 eastern, 15 central, 5 Nigeria-Cameroon (NC), and 1 captive western × central hybrid. Fifty samples are obtained from the de Manuel et al. (2016) dataset, and 4 additional ones from Kuderna et al. (2023), with variants jointly called and mapped to the panTro6 assembly. Table 3.1 shows details on the country of origin of samples from each subspecies, and figure 3.22A provides a rough illustration of sample locations. Throughout, times in generations are converted to years assuming a generation time of 25 years (Langergraber et al., 2012).

To prepare the data for ARG inference, we first build an accessibility mask by running `genmap -K 36 -E 0` (Pockrandt et al., 2020) on the panTro6 assembly, marking as unreliable any region in which a 36-mer does not map uniquely. We additionally mask assembly gaps taken from the UCSC gap track (Perez et al., 2025). Together these mark $\sim 12\%$ of the genome as inaccessible, and these regions are excluded from ARG inference. We then apply site-level filters to remove low-quality and non-biallelic variants. Specifically, we apply the GATK best-practices hard filters (Van der Auwera et al., 2013) (`QD<2 | FS>60 | MQ<40 | SOR>3 | ReadPosRankSum<-8 | MQRankSum<-12.5`), we remove sites at which the pooled alternate-allele fraction across heterozygous calls falls outside $[0.25, 0.75]$, and we remove sites at which the total sequencing depth, summed across all samples, falls outside $0.25\times$–$2\times$ the chromosome-wide median depth. Filtered sites are not used in ARG inference, but their fraction is informative for POLEGON as putatively varying sites that can be used to adjust the mutation rate. Applying all filters removes $\sim 13\%$ of the variant sites (some of which fall within inaccessible regions), leaving a total of $42,900,324$ bi-allelic SNPs (spanning 3.1 Gb) for ARG inference. The genomes were phased using SHAPEIT5 (`phase_common -hmm-ne`

1000) with no reference panel and no genetic map (Hofmeister et al., 2023). `-hmm-ne` was swept over 10,000 / 25,000 / 30,000 and made no difference. Beagle (B. L. Browning et al., 2021) was tested and performed slightly worse than SHAPEIT5, with a 4.66% switch error against SHAPEIT5's 4.03%, where accuracy was measured against GATK read-backed phase blocks.

Ancestral states were assigned based on the human hg38 reference bases projected into the panTro6 assembly using a single liftover. This polarization is incorrect at sites where the human lineage carries a derived allele, which should be rare among sites polymorphic in chimpanzees. Positions with no human alignment (2–3%) are left as N and fall through to SINGER's own per-site polarization. Of the variant sites, 85.5% had an ancestral allele that matched the panTro6 reference allele, and 11.5% had one that matched the alternative. REF and ALT labels were flipped for the latter in the VCF as SINGER treats the REF allele as ancestral.

Finally, an LD-based recombination map inferred from the 10 Nigeria-Cameroon chimpanzee genomes of the Great Ape Genome Project (Stevison et al., 2016) was used as an input for SINGER after a double liftover (hg18 $\rightarrow$ hg38 $\rightarrow$ panTro6) and a conversion from ρ/kb to cM/Mb.

Estimates of coalescent rates from inferred ARGs indicate extensive recent subspecies structure

Before running CoaSTT on the inferred ARGs, we show how pooling coalescent rate estimates can obscure true population structure, especially in recent time. In figure 3.22B, we compute an estimate for pairwise coalescent rates between the pair of haplotypes of each individual, and show the average of the inverse within-individual coalescent rate for each subspecies, in a manner similar to how PSMC is applied (plotted in bold lines). We additionally show the inverse coalescent rate between each pair of haplotypes in each subspecies (shown in

thin lines). If the mean inverse coalescent rate between the haplotypes of each individual were interpreted as an effective population size, the result would read as a recent decline in the effective population size of all subspecies. Such a reading is ambiguous, however, since population structure produces the same signal in PSMC-like curves (Steux et al., 2025). Within-individual rates show a recent increase because each individual is highly likely to be a descendant of two parents that belonged to the same subpopulation. Coalescent rates between all pairs of haplotypes in each subspecies support the opposite story, where they rapidly decline in recent time as a result of recent structure within each subspecies (a decline that can equally be mistaken for a recent population expansion). It is difficult, however, to make sense of the trajectories of pairwise coalescent rates as shown in figure 3.22B, and CoaSTT was developed to provide intuitive low-dimensional representations for scenarios exactly like this.

We additionally compute an estimate for the split time between each pair of subspecies as the time at which the relative cross-coalescence rate (RCCR) first crosses 0.5 (figure 3.22C). This gives an estimate for the split time between the eastern and central clades of ~ 210 kya, which is slightly higher than 182 kya, the upper bound of the estimate given by de Manuel et al. (2016). The same is the case for our estimate of the split time between the western clade and the eastern–central clade of ~ 800 kya, which is much higher than their upper bound of 633 kya. The accuracy of the half-life rule applied to SINGER-derived rates depends on the demography: it recovered the true split time under the clean split of section 3.2, but the smoothing of between-population rates across a split can shift the crossing in more complex histories, which may account for the central–eastern disagreement. Our different estimate of the western–(central, eastern) split time is too large to be accounted for only by that, and could be a result of asymmetric bonobo introgression, or perhaps an inflated denominator in the cross-coalescence ratio due to the bottleneck in the western effective size, which renders the half-life rule unreliable.

More interesting are the split times between the NC subspecies and the western, central and eastern subspecies. The three estimated split times are very similar ($\sim$ 530–580 kya), in disagreement with the notion that the western and NC subspecies form a clade, whereas the eastern and central subspecies form a clade. Such a signal is suggestive of the NC subspecies being a result of admixture between a population ancestral to the western population, and one ancestral to the eastern and central populations. The split time between western and NC chimpanzees was estimated to be $\sim$ 250 kya by de Manuel et al. (2016) as the time at which MSMC2 (applied to the X chromosomes of males) shows gene flow between the two lineages ceasing, which corresponds more closely to the point at which the RCCR curve rises above 0.1 than to where it crosses 0.5. The conflicting results point to a more interesting story for the NC subspecies, which we explore further in section 3.4.3.

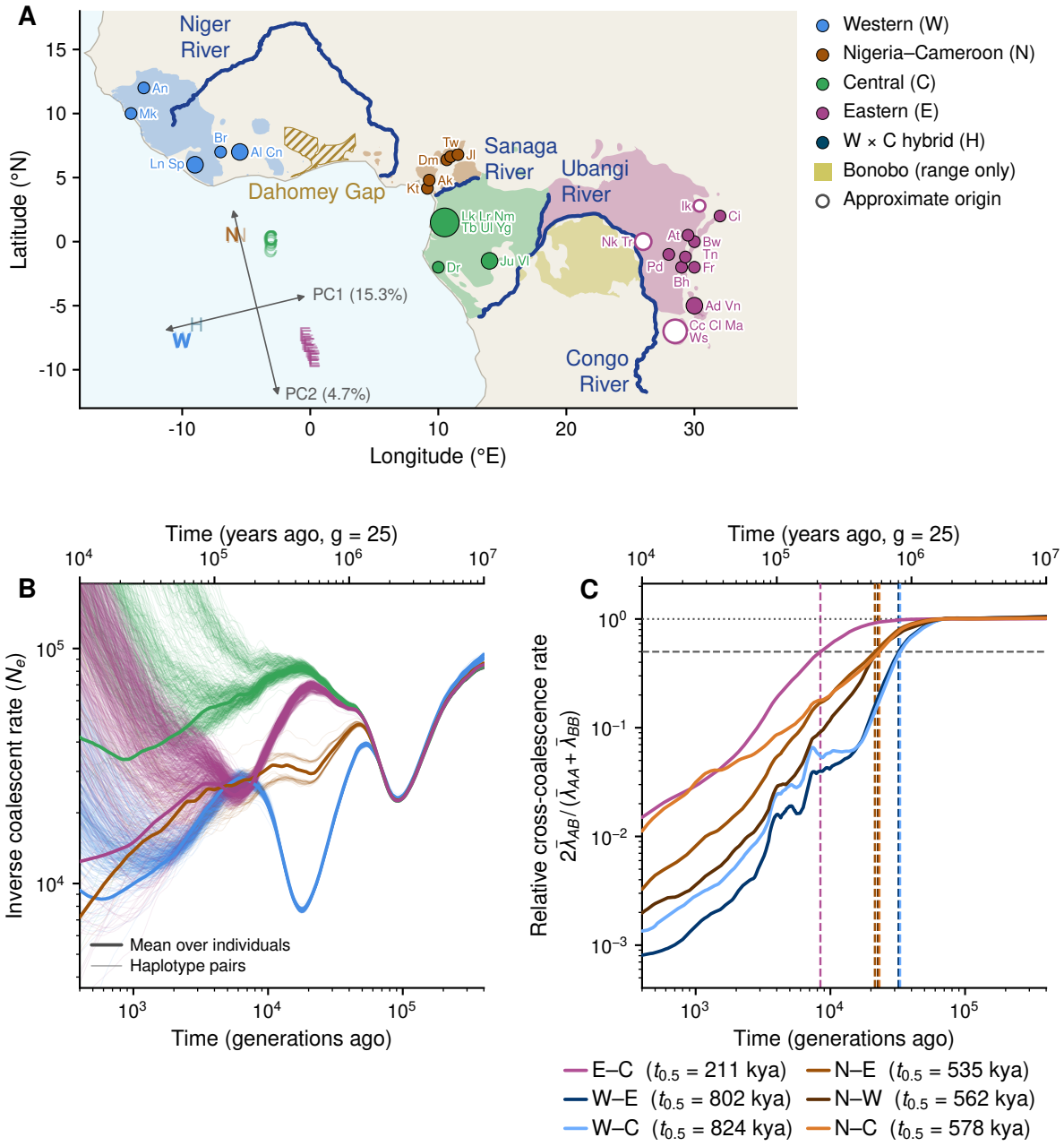

Figure 3.22: An overview of the coalescent history of chimpanzees. A) An illustration of the range of each of the four subspecies of chimpanzees, as well as bonobos, with locations of samples following table 3.1. PCA is shown to separate samples from the four subspecies into four clusters. **B)** Inverse coalescent rates over time for each pair of haplotypes (thin lines), in addition to the mean of within-individual coalescent rates (thick lines). **C)** Relative cross-coalescence rates over time for each pair of populations, with the time at which each curve crosses 0.5 denoted underneath.

Table 3.1: Chimpanzee samples by subspecies and place of origin. Locations, regional clusters and notes follow de Manuel et al. (2016) (Table S1 and Data S1) and Kuderna et al. (2023).

Origin	Individual (code)	Recorded place	Notes
Western (*P. t. verus*)			
Ivory Coast	Alice (Al)	Ivory Coast	Bought from poachers at 7 days old
	Berta (Br)	Man	Captured during forest logging
	Cindy (Cn)	Ivory Coast	Bought at 8 months old
Guinea	Annie (An)	Haute or Moyenne Guinea	Low genotype quality; record also allows south Mali, Senegal, or Koundara (Badiar NP, Guinea)
	Mike (Mk)	Boké–Boffa	Low genotype quality; bought from a local hunter
Liberia	Linda (Ln)	Liberia	Place of embarkation, not of capture; imported to a European zoo
	SeppToni (Sp)	Liberia	Place of embarkation, not of capture; imported to a European zoo
Captive born	Clint (Ct)	–	Captive born
	PD_0257 (P1)	–	Captive born
	PD_0258 (P2)	–	Captive born
	PD_0427 (P4)	Basel Zoo	Captive born
Unknown	Bosco (Bs)	–	No origin recorded
	Jimmie (Jm)	West Africa	Imported
	Koby (Kb)	–	No origin recorded
Nigeria–Cameroon (*P. t. ellioti*)			
Western	Akwaya-Jean (Ak)	Akwaya	Confiscation site
Cameroon	Damian (Dm)	Tinto	Confiscation site
	Julie (Jl)	Douala	Confiscation site; probably smuggled from the north-west
	Koto (Kt)	Edea	Confiscation site; probably smuggled from Mt. Cameroon
	Taweh (Tw)	Numba Nfe, near Mamfe	Confiscation site
Central (*P. t. troglodytes*)			
Equatorial	Lara (Lr)	Eq. Guinea	Taken by customs
Guinea	Luky (Lk)	Eq. Guinea	Taken by customs
	Noemie (Nm)	Eq. Guinea	Low genotype quality; lived as a pet
	Tibe (Tb)	Eq. Guinea	Taken by customs
	Ula (Ul)	Eq. Guinea	Imported
	Yogui (Yg)	Eq. Guinea	–
Gabon, east	Julie (Ju)	Haut-Ogooué	Found on the road; the record also names Gamba, which is in Ogooué-Maritime
	Vaillant (Vl)	Haut-Ogooué	–
Gabon, west	Doris (Dr)	Rabi (oil field)	Found as an infant after her mother was hunted
Unknown	Alfred (Af)	–	Imported to a European zoo
	Brigitta (Bg)	–	Imported to a European zoo
	Cindy (Cd)	–	Imported to a European zoo
	Gamin (Gm)	–	Imported to a European zoo
	Marlin (Mr)	–	Imported to a European zoo; listed by de Manuel as "correct name Marina"
	Mirinda (Mn)	–	Imported to a European zoo
Eastern (*P. t. schweinfurthii*)			
DRC, east / Rwanda	Bihati (Bh)	Goma	Region assigned here from the town
	Frederike (Fr)	Rwanda	Imported from Rwanda
	Nakuu (Nk)	East DRC	Brought to Kampala by a soldier; region assigned here from the record
	Padda (Pd)	Walikale	Low genotype quality
	Tongo (Tn)	Tongo, Virunga NP	Bought in the village
	Trixie (Tr)	DRC, east near Rwanda	Smuggled to Rwanda
DRC, north	Athanga (At)	Beni	–
	Ikuru (Ik)	Arua, Uganda	Taken there by a Ugandan soldier during the war
	Mgbadolite (Mg)	Gbadolite	–
DRC, south	Maya (Ma)	Mbuji-Mayi	Sold in this trading city
	Washu (Ws)	Lubumbashi	Sold in this trading city
DRC, region	Cleo (Cl)	–	Smuggled to Zambia, most likely from DRC
unknown	Coco (Cc)	Kitwe, Zambia	Confiscated there; most likely from DRC
	Kidongo (Kd)	DRC	No record beyond the country
Uganda, west	Bwambale (Bw)	Busumba, Bundibugyo	Probably from Kibale, Queen Elizabeth or Rwenzori NP
	Cindy (Ci)	Masindi	Probably from Budongo Forest
Tanzania,	Andromeda (Ad)	Gombe NP	–
Gombe	Vincent (Vn)	Gombe NP	–
Captive born	PD_0259 (P3)	–	Captive born
Western × Central hybrid			
Captive born	Donald (Dn)	–	Second-generation Western × Central hybrid

3.4.2 CoaSTT reveals extensive recent population structure within Central and Eastern chimpanzee subspecies

We computed two sets of CoaSTT embeddings using the last posterior ARG sample inferred by SINGER, one for a recent time period between 50 and 4,000 generations ago (figure 3.23) and one for a deeper period from 4,000 to 400,000 generations ago (figure 3.28). The embeddings from both periods are plotted separately since their x-y limits (dictated by the lowest coalescent rates) differ by almost two orders of magnitude. We first focus on the recent time embeddings (figure 3.23), shown at four representative time slices in figure 3.23(B–E), with insets that highlight crowded regions. The embeddings show four clusters, corresponding to the four subspecies, that remain distinct for the whole period until 4,000 generations ago ($\sim$ 100 kya). In contrast, by the same number of generations in the past, most human samples have approximately equal pairwise coalescent rates and form a single cluster (figure 3.5B), reflecting the much deeper divergence between chimpanzee subspecies than between human populations. To distinguish features of the embeddings that are robust from those that vary with the ARG sample or the optimization, we additionally show the embeddings obtained from different posterior samples from SINGER and from different initializations, both at two time slices (figures 3.24 and 3.25) and over all time points (figures 3.26 and 3.27).

Around 60 generations ago ($\sim$ 1.5 kya), lineages from all four subspecies show within-subspecies structure in their coalescent rates, which is most pronounced in the central and eastern chimpanzees. The scale bar is a useful guide to the extent of this structure at each time slice, since the coalescent rate between two samples is approximately the inverse of their distance in the embedding. Two samples separated by 2×10^6 have a coalescent rate of roughly 0.5×10^{-6} per generation at that time. Additionally, a sign that samples within a cluster have similar pairwise coalescent rates is that the cluster is approximately circular (or an equilateral triangle for 3 samples). Using these rules, the structure is seen to mostly collapse

for the western chimpanzees by 500 generations ago ($\sim$ 12 kya), and for the eastern and NC chimpanzees by 2,000 generations ago ($\sim$ 50 kya), while the central chimpanzees remain structured throughout the 500–2,000 generations ago period. Two central chimpanzees, 'Marlin' (Mr) and 'Doris' (Dr), stand out by remaining separated from the central cluster for the whole period.

A natural question is whether the observed structure correlates with geography, as previously suggested for central and eastern chimpanzees (de Manuel et al., 2016; Fontsere et al., 2022). At a fine scale, this question cannot be easily answered for our samples, since most of them lack accurate geographic information. Many have no recorded origin or are captive born, and for those with a recorded place, it usually denotes a confiscation site, a location of sale, or a country of export or import. It is still worth examining how the embeddings are structured in recent time and comparing them against the rough locations available. In doing so, it is important to also compare against the embeddings obtained from different SINGER samples and different initializations, to ensure that the placement of a given sample is consistent and is not an artifact of ARG uncertainty or of the low-dimensional embedding. Consistency alone does not rule out data artifacts, however, as samples with a large fraction of low-quality genotypes tend to be pushed away from their cluster, as we show below.

Starting with the NC chimpanzees, four of the five samples form a small cline, with chimpanzee 'Julie' (Jl) slightly distant. The exact arrangement of the four is highly variable between ARG samples and initializations, but 'Akwaya-Jean' (Ak) and 'Koto' (Kt) are consistently nearest neighbors, with 'Damian' (Dm) close to them and 'Taweh' (Tw) slightly more distant. All four recorded places are confiscation sites, so this arrangement cannot be compared against geography. 'Julie' is reported in multiple studies to be distinct from the rest of the NC samples (de Manuel et al., 2016; Prado-Martinez et al., 2013). Her pairwise coalescent rates with the central chimpanzees are elevated relative to the other NC samples, consistent with the central ancestry component found in her genome using sNMF (de Manuel

et al., 2016), and with the hypothesis of de Manuel et al. (2016) that she originates from the forest–savanna ecotone of central Cameroon, where NC chimpanzees exchange gene flow with central chimpanzees across the Sanaga (Mitchell et al., 2015). 'Julie' illustrates how an admixed sample with the majority of ancestry coming from one population appears in CoaSTT embeddings. She is pushed away from her own cluster rather than placed between the two source clusters, since her minority central ancestry contributes little to her coalescent rate with the central samples, while lowering her coalescent rate with the other NC samples. Backward in time, the five NC samples approach one another, and the distances between them shrink by more than an order of magnitude. By 2,000 generations ago ($\sim$ 50 kya), the NC samples other than 'Julie' form an approximately rhomboid configuration, indicating that they are approximately exchangeable with respect to the coalescent process. The local structure among NC samples seems recent, and their lineages trace back to a common ancestral population by that time.

The cluster formed by the western chimpanzees is the tightest over all time slices, but is slightly structured as well. The core of the cluster is formed by the four captive samples that share elevated recent ancestry: 'Clint' (Ct) and the three samples from Kuderna et al. (2023). Around the core, the five samples from Ivory Coast and Liberia ('Alice' (Al), 'Berta' (Br), 'Cindy' (Cn), 'Linda' (Ln) and 'SeppToni' (Sp)) form a small cline. Their exact arrangement differs between initializations and ARG samples, but 'Linda' is consistently closer to the captive group. The three samples with no recorded origin ('Jimmie' (Jm), 'Bosco' (Bs) and 'Koby' (Kb)) are consistently placed outside the core. 'Annie' (An) and 'Mike' (Mk) (both from Guinea) are also consistently placed away from the core, but both have a large fraction of low-quality genotypes: after filtering, 56.6% of the genotypes of 'Mike' and 36% of those of 'Annie' have a sequencing depth below 8 or a GATK genotype quality (GQ) below 20, against a median of 3.4% across all samples ('Koby' is moderately affected, with 16%). When such genotypes are set to missing and the masked stretches are excluded

from each sample's ancestry in the re-inferred ARG, so that coalescence is measured only on high-quality genotypes, 'Annie' and 'Mike' return to typical levels of coalescence with the other western chimpanzees. Notably, the substructure previously reported between Guinea and the rest of the western chimpanzees, which was detected by fineSTRUCTURE but not by PCA or sNMF (de Manuel et al., 2016), rests on these same two samples, which are the only Guinea individuals in that dataset, and may therefore partly reflect genotype quality. 'Donald' (Dn), a second-generation western/central hybrid (de Manuel et al., 2016; Prado-Martinez et al., 2013), sits next to the captive group rather than between the western and central clusters. As with 'Julie', his minority central ancestry barely affects his position. By 200 generations ago ($\sim$ 5 kya), the core tightens further, and the samples that remain furthest from it are 'Mike', 'Koby', 'Jimmie' and 'Bosco'. Overall, the ancestral lineages of the sampled western individuals show the least extensive and most recent structure of the four subspecies, consistent with previous reports of weak population structure and high connectivity across the western chimpanzee range (de Manuel et al., 2016; Fontsere et al., 2022).

The eastern lineages exhibit strong coalescent structure in recent time, shown by a cline in the embeddings 60 generations ago ($\sim$ 1.5 kya), which progressively decays back in time. Additionally, a few pairs of samples consistently form two-member clusters in recent time: 'Cleo' (Cl) and 'Coco' (Cc) (both smuggled to Zambia, most likely from DRC), 'Washu' (Ws) and 'Maya' (Ma) (both from southern DRC), 'Andromeda' (Ad) and 'Vincent' (Vn) (both from Gombe, Tanzania), and 'Nakuu' (Nk) and 'Cindy' (Ci) (one brought by a soldier from eastern DRC, the other from western Uganda). These tight pairs are not close relatives (de Manuel et al., 2016). Their elevated coalescence is concentrated between roughly 100 and 500 generations ago rather than in the last few generations, suggesting that each pair descends from the same local population. The Gombe pair was also previously identified as a distinct genetic component (Prado-Martinez et al., 2013). Comparing against the putatively

assigned locations, the samples are ordered along a north–south cline, starting from 'Washu' and 'Maya' in the south (bottom of the embeddings) together with 'Cleo' and 'Coco', going through 'Nakuu' and 'Cindy', and ending with 'PD_0259' (P3), 'Mgbadolite' (Mg) and 'Bihati' (Bh) (the last two cluster together despite their distant recorded locations). The grouping of 'Cleo' and 'Coco' with the two southern DRC samples supports a southern origin for them, providing an example of how the embeddings can help resolve the uncertain provenance of confiscated individuals. This cline is consistent with the north–south structure and isolation by distance previously reported for eastern chimpanzees (de Manuel et al., 2016; Fontsere et al., 2022). 'Trixie' (Tr) (smuggled to Rwanda), 'Frederike' (Fr) (imported from Rwanda) and 'Tongo' (Tn) (from the Virunga National Park, DRC) are consistently offset from the cline under different initializations and ARG samples, matching the DRC-East/Rwanda regional cluster identified by de Manuel et al. (2016). 'Padda' (Pd), assigned to the same cluster, is consistently offset as well, though 52.9% of his genotypes are of low quality (depth below 8 or GQ below 20), and masking them on one chromosome largely removes his offset. Backward in time, this structure progressively dissolves, though the separation between the southern and northern samples persists until at least 2,000 generations ago ($\sim$ 50 kya). The arrangement of samples along the cline is consistent between initializations, and less consistent between ARG samples, as a result of the variance in the number of coalescent events between pairs of samples on the cline in recent time. We conclude that the eastern chimpanzees descend from a continuously structured population, likely shaped by geography, with uncertainty about the details of this structure in recent time.

Similar to the eastern population, the central population shows extensive coalescent structure in recent time, with a structured core formed by the Equatorial Guinea samples around which the Gabon samples ('Julie' (Ju), 'Vaillant' (Vl) and 'Doris' (Dr)) are scattered, with no single consistent configuration (Unfortunately, there is a central Julie and a Nigeria-Cameroon Julie). This inconsistency is expected in the most recent time slices, where most

pairs of central samples have few or no coalescent events, leaving their relative placement weakly constrained. The core formed by the Equatorial Guinea samples itself forms a cline, with a few samples (notably 'Gamin' (Gm) and 'Cindy' (Cd)) positioned away from it, a pattern similar to the one observed for the eastern population. 'Noemie' (Nm) is also offset from the core, but 54.5% of her genotypes are of low quality, and masking them on one chromosome removes her offset. The Gabon samples as well as 'Gamin' progressively move closer to the core backward in time as their rates of coalescence with the core increase, but remain slightly separated from it even 2,000 generations ago ($\sim$ 50 kya). 'Marlin' (Mr) and 'Doris' (Dr) are two remarkable central samples, whose lineages remain separated from the central cluster until about 5,000 generations ago ($\sim$ 125 kya; figure 3.28). Although they are placed far apart in recent time, from about 2,000 generations ago onward they are closer to each other than to the rest of the central samples, indicating that they descend from a common ancestral population, which itself is structured, that was isolated from the other sampled central chimpanzees for a substantial fraction of the time over which the central and eastern subspecies have been distinct. An ADMIXTURE analysis at $K = 5$ by de Manuel et al. (2016) assigns 3–5% bonobo ancestry to 'Marlin' and 'Doris', and 'Doris' comes from south of the Ogooué river, where central chimpanzees carry more bonobo-derived segments and form a distinct genetic cluster (Fontsere et al., 2022). However, such a small fraction of foreign ancestry cannot explain their separation, as 'Marlin' and 'Doris' coalesce with the other central samples at less than half the typical rate even by 4,000 generations ago. Their coalescent rates with the eastern and NC chimpanzees are also lower than those of the other central samples, while their rates with the western chimpanzees are not, suggesting that this lineage did not take part in the gene flow between central and eastern chimpanzees that followed their split. To our knowledge, a central lineage this deeply separated has not been previously reported. If 'Marlin', whose origin is unrecorded, shares the origin of 'Doris' south of the Ogooué, this implies that the structure of central chimpanzees in that region is

several times older than previously suggested (Fontsere et al., 2022).

Together, the recent-time embeddings illustrate how CoaSTT summarizes population structure at the level of individuals. Within each subspecies, the distance between two samples reads directly as the inverse of their coalescent rate, and the spread of the clusters in recent time reflects how structured the population is rather than how large it is. Backward in time, this structure progressively dissolves until the ancestral lineages of a subspecies become approximately exchangeable with respect to the coalescent process. Admixed samples, such as 'Julie' and 'Donald', appear at the edge of the cluster of their majority ancestry rather than between clusters, so modest admixture can be missed in the embeddings and is better examined through the pairwise coalescent rates themselves. Finally, the placement of individual samples should be checked across ARG samples and initializations, and against data quality, since samples with many low-quality genotypes are consistently, but artificially, pushed away from their clusters. Read this way, the embeddings recover the weak and recent structure of western chimpanzees and the geographic structure of central and eastern chimpanzees reported in previous studies, and reveal two deeply diverged central lineages, 'Marlin' and 'Doris', that have not been previously described.

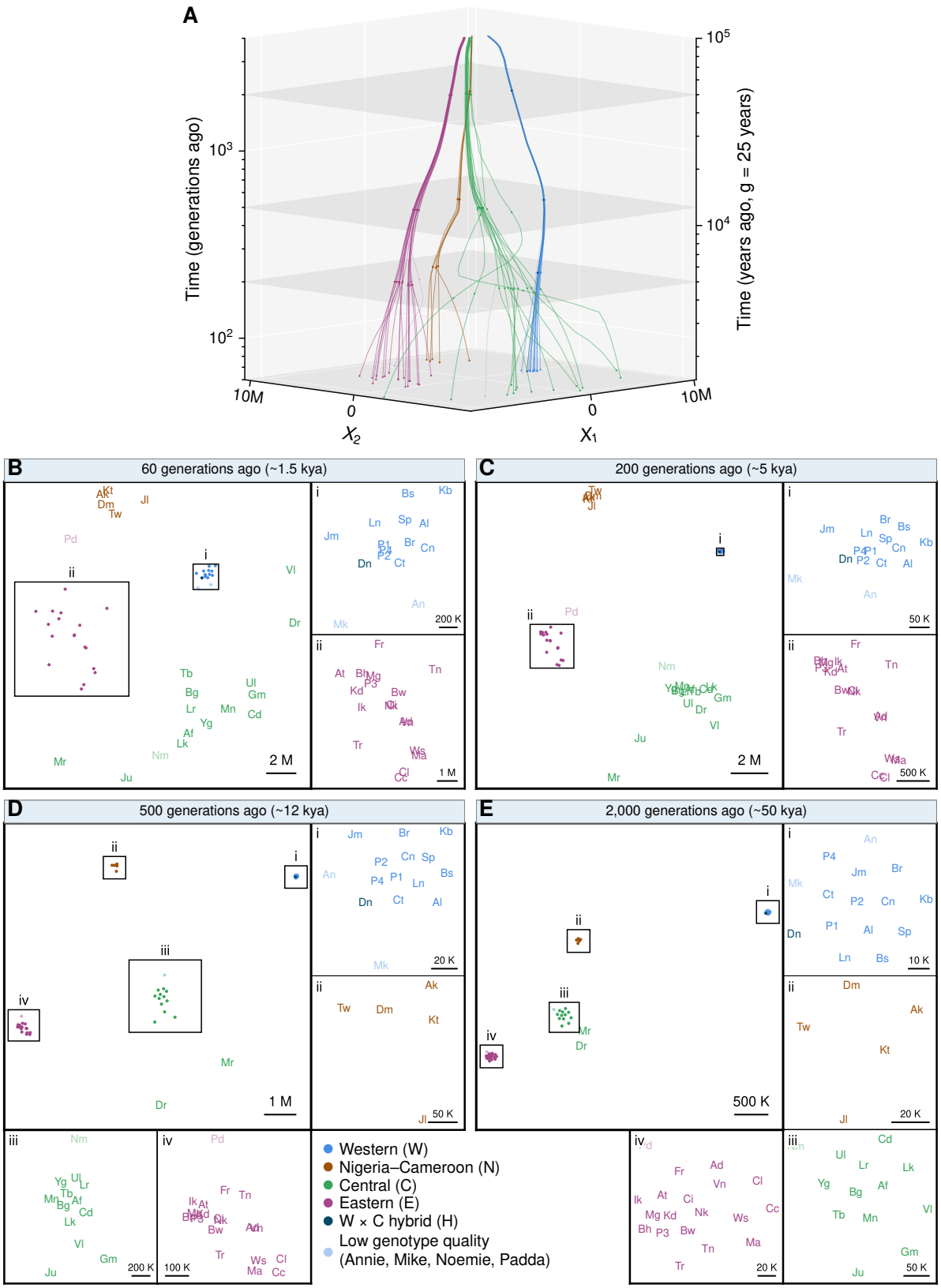

Figure 3.23: CoaSTT embeddings of chimpanzees between 50 and 4,000 generations ago.

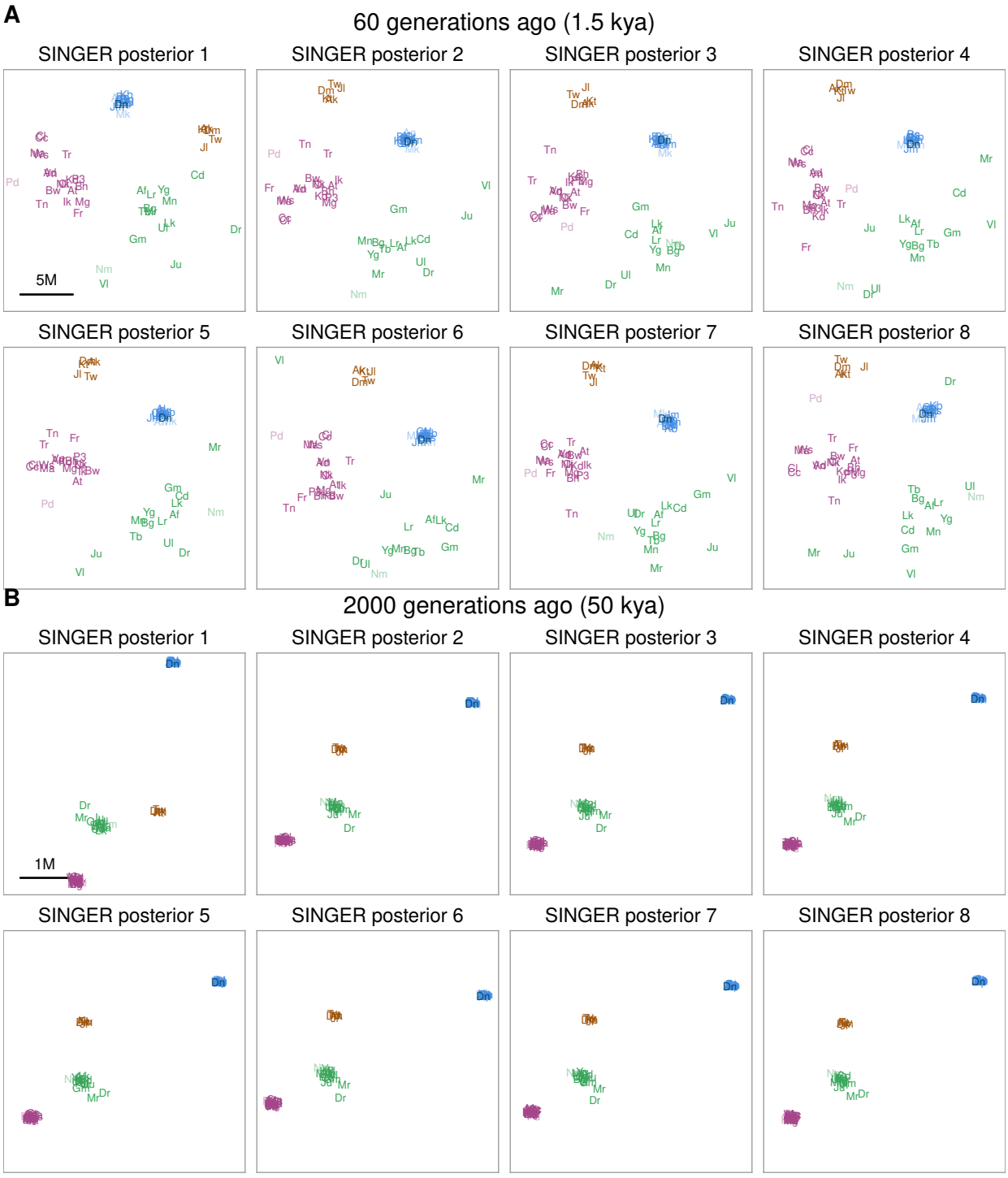

Figure 3.24: Recent-time chimpanzee embeddings at 60 (A) and 2,000 (B) generations ago from eight SINGER posterior samples.

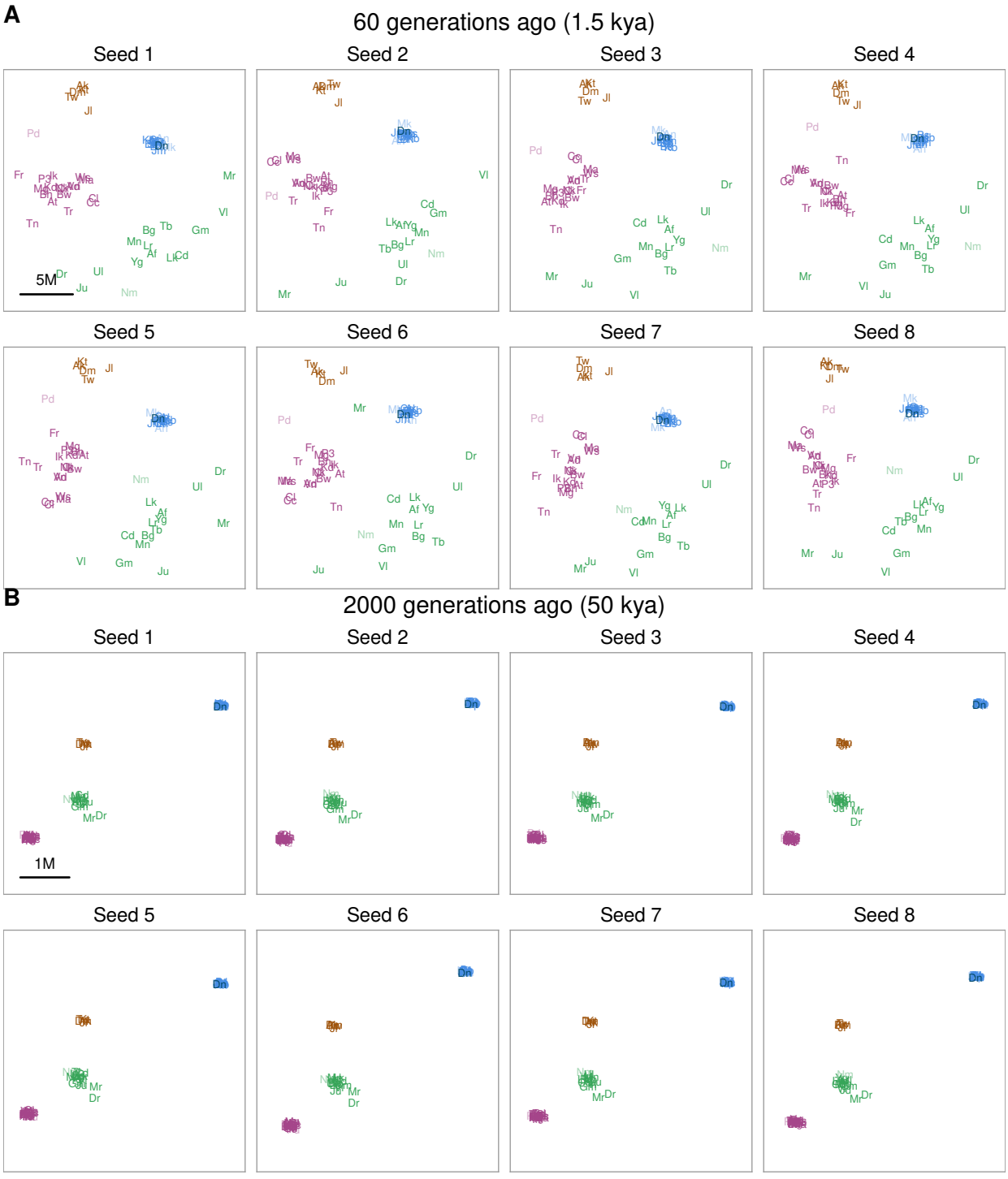

Figure 3.25: Recent-time chimpanzee embeddings at 60 (A) and 2,000 (B) generations ago from eight different initializations.

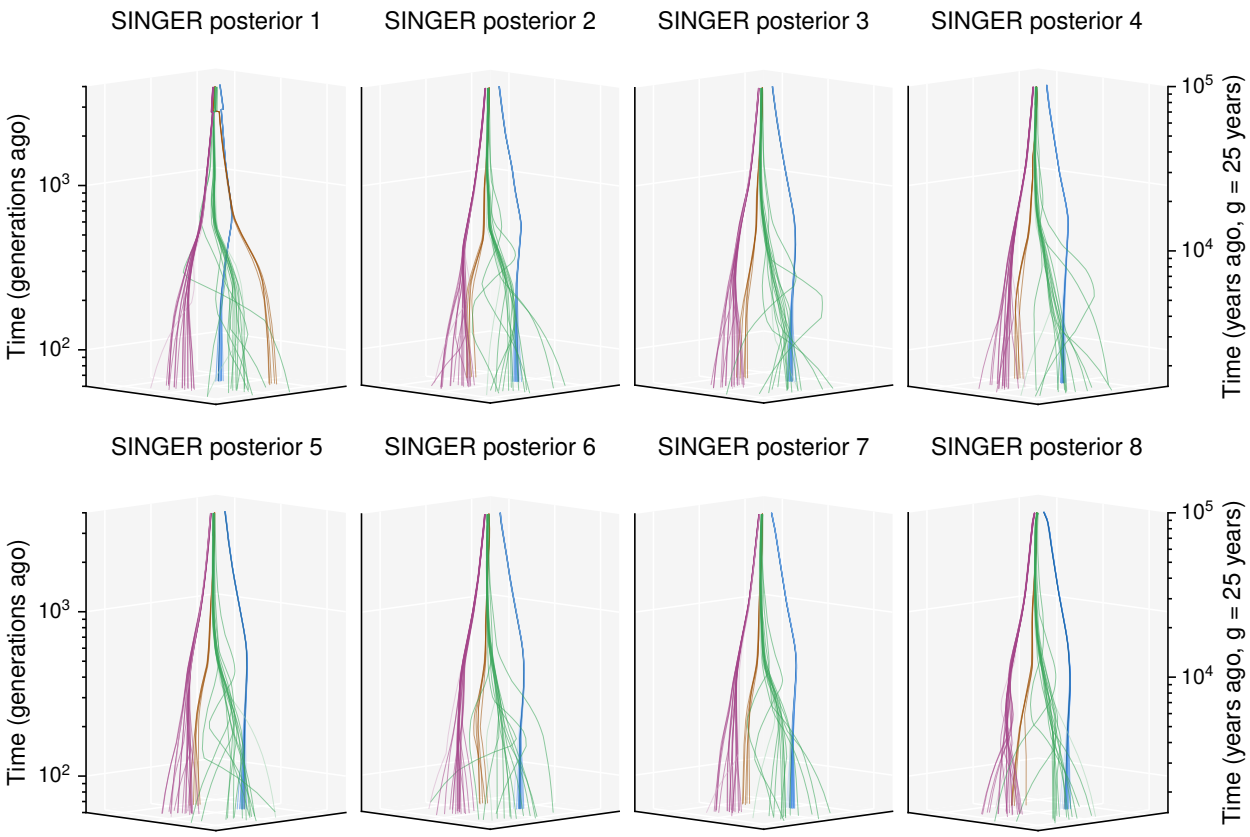

Figure 3.26: 3D recent-time chimpanzee embeddings from eight SINGER posterior samples.

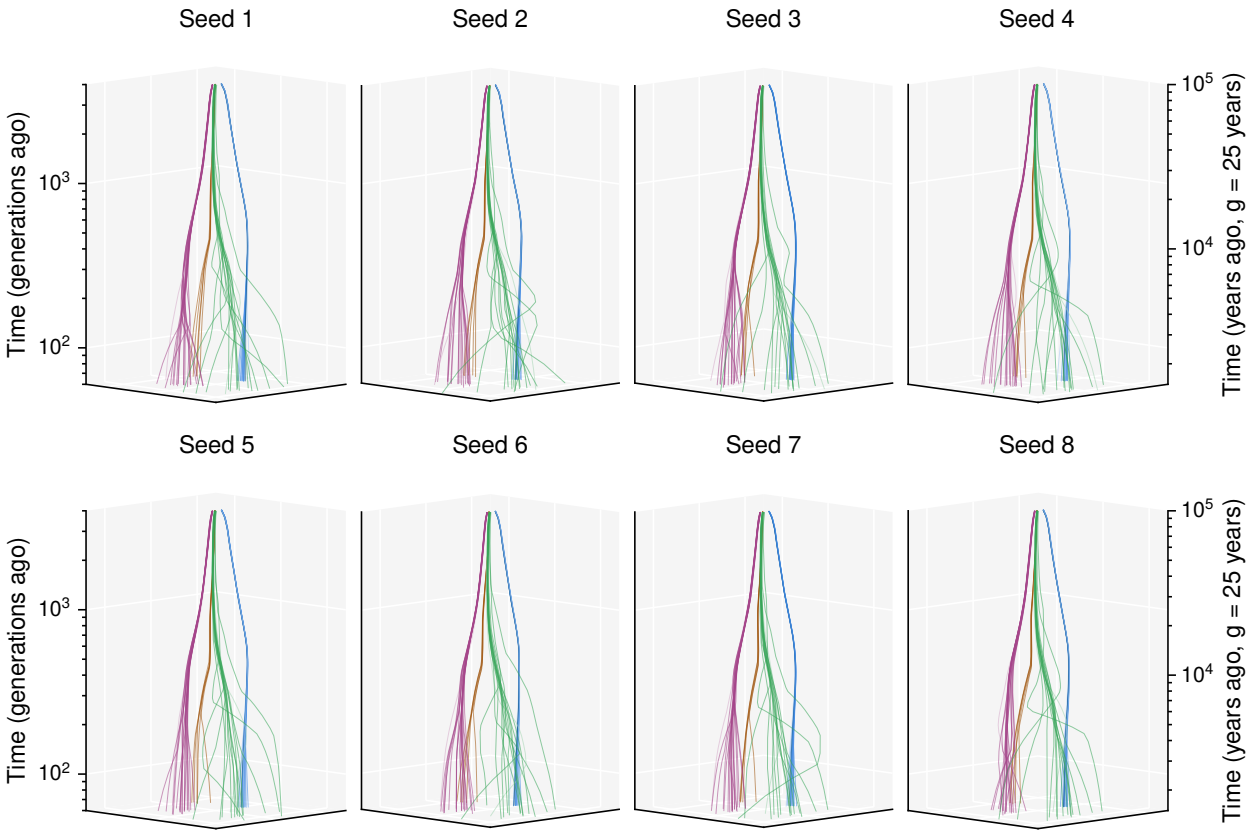

Figure 3.27: 3D recent-time chimpanzee embeddings from eight different initializations.

CoaSTT embeddings going back to the deep past

By 5,000 generations ago ($\sim$ 125 kya), the samples from each subspecies become exchangeable with respect to coalescence, forming circular clusters in the embeddings, except for the two central chimpanzees, 'Marlin' (Mr) and 'Doris' (Dr), which still have a slightly lower coalescent rate with the central cluster. By 10,000 generations ago ($\sim$ 250 kya), the central and eastern chimpanzees approximately form a single circular cluster. Applying the half-life rule to the relative cross-coalescence rate (figure 3.22C) places the central–eastern split around 8,400 generations ago ($\sim$ 210 kya), slightly older than previous estimates of 139–182 kya (de Manuel et al., 2016), a shift that is plausibly due to the smoothing of between-population rates in SINGER-inferred ARGs (section 3.2). For the same reason, the mixing of the two

subspecies remains imperfect well beyond the split, as SINGER-inferred between-population coalescent rates stay under-estimated past the split time. The NC cluster, meanwhile, is still closer to the central–eastern cluster than to the western cluster 5,000 generations ago, and lies right between them by 10,000 generations ago, at which point the NC samples coalesce with the western samples at about the same rate as with the central and eastern samples. Further back in time, the NC samples are pulled toward the western cluster, and by 20,000 generations ago ($\sim$ 500 kya), NC lineages coalesce with western lineages about three times faster than with central or eastern lineages, and by about 22,000 generations ago as fast as among themselves, indicating that by this time the NC and western lineages have entered a common ancestral population. Beyond this point, the clusters collapse toward the center of the embedding as the central and eastern lineages also begin to coalesce with the western and NC lineages.

A relative cross-coalescence rate of 0.5 is reached almost simultaneously between the NC chimpanzees and each of the other three subspecies (535–578 kya). This is much older than the $\sim$ 250 kya western–NC split de Manuel et al. (2016) estimate from the X chromosome, and more recent than our crossing between the western chimpanzees and the central and eastern chimpanzees ($\sim$ 800 kya). The latter crossings are also older than the 544–633 kya estimated for this split (de Manuel et al., 2016), in part because a strong bottleneck in the western lineage around this time raises the western within-population rate, delaying the crossing of the relative rate. At a deep time slice of 36,000 generations ago ($\sim$ 900 kya), a single circular cluster exists, indicating that all samples have approximately the same coalescent rate with one another. The leftover structure is mostly a result of the ARG inference bias, most apparent in the western samples, which remain more tightly clustered as a lingering signature of their bottleneck.

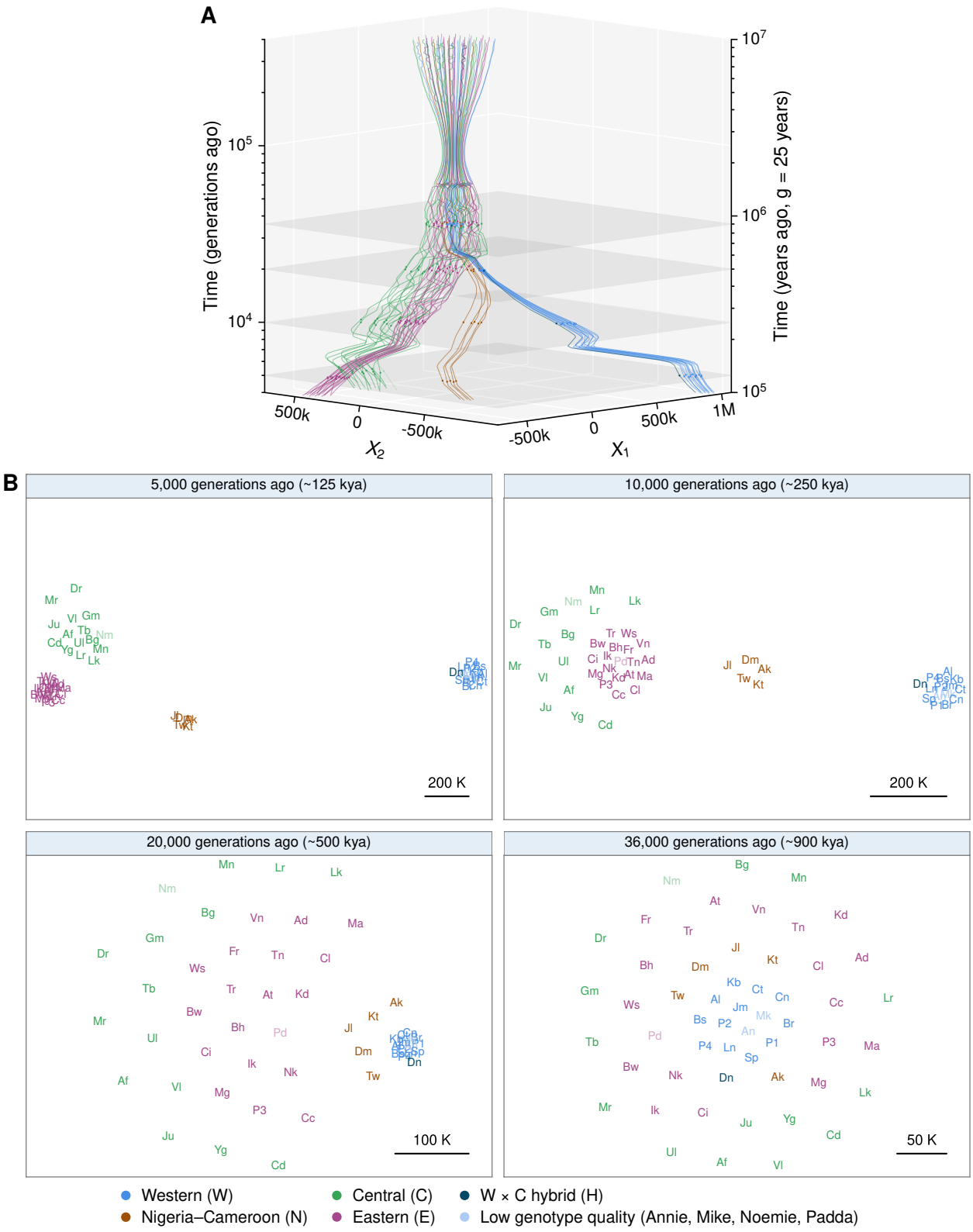

Figure 3.28: CoaSTT embeddings of chimpanzees between 4,000 and 400,000 generations ago.

Comparing input rate estimates with rates reconstructed from the embeddings

We evaluate the ability of CoaSTT to summarize the $\binom{n}{2} = 1{,}431$ rate estimates at each time slice by the $2n = 108$ latent coordinates, in a manner similar to section 3.3.2. Figure 3.29A shows correlation measures that are similar to those of the human embeddings (figure 3.9A), where, again, the reduced correlation on the linear scale is predominantly a result of the circular clusters in deep time, in which most pairwise rates are nearly identical (figure 3.29B). The percentage of outliers that result from an impossible triangle is much lower for the chimpanzees ($\sim 2.5\%$ compared to $\sim 20\%$ in humans), partly because we extend the embeddings to a much deeper time, so the outliers resulting from crowding dominate the top outliers. Figure 3.30 shows that most of the impossible-triangle outliers involve the two admixed samples, 'Julie' (Jl) and 'Donald' (Dn), as bridges. Rather than pulling their two source populations together, which would over-estimate the reconstructed rates between them, each sample is placed near its dominant source (NC for 'Julie' and western for 'Donald'), whose rates with it are reproduced well, while its rates with the other source (central for both samples) are under-estimated. Additionally, all NC samples serve as bridges in highly prominent triangles that involve western and eastern samples on the edges. This reflects time periods where ancestral lineages of NC chimpanzees have relatively high coalescent rates with both western and eastern chimpanzees compared to a very low coalescent rate between western and eastern lineages, which, again, points clearly to a history of admixture for the NC chimpanzees.

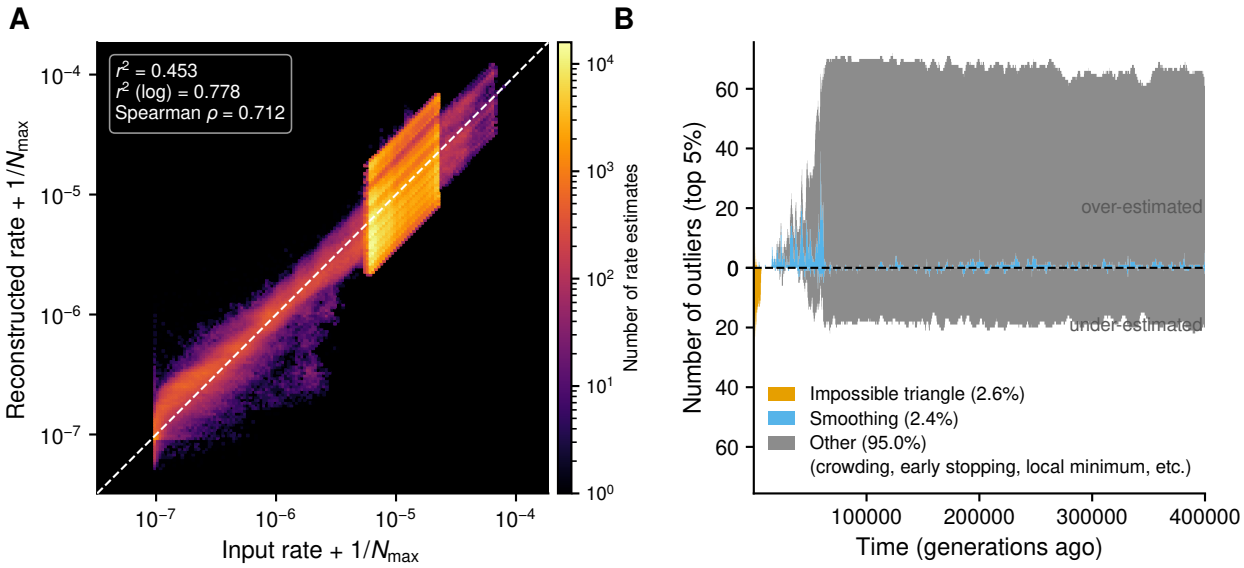

Figure 3.29: Probing the mismatch between input rate estimates and rates reconstructed from chimpanzee CoaSTT embeddings. A) Density plot of rates reconstructed from embedding distances and input rate estimates, for all pairs and time points, along with three measures of correlation. The block of rates reconstructed at a value much lower than the input of $\approx 2 \times 10^{-6}$ corresponds to the admixed samples, 'Donald' (Dn) and 'Julie' (Jl), which are placed close to the western and NC cluster respectively. Their large distance to the central cluster results in reconstructed rates that are much lower than the input rates. **B)** Decomposition of the source of mismatch between input and reconstructed rates for the top 5% outliers (on log-scale) over all time points. Most of the outliers that are due to an impossible triangle involve rates between 'Donald' and central samples and between 'Julie' and central samples.

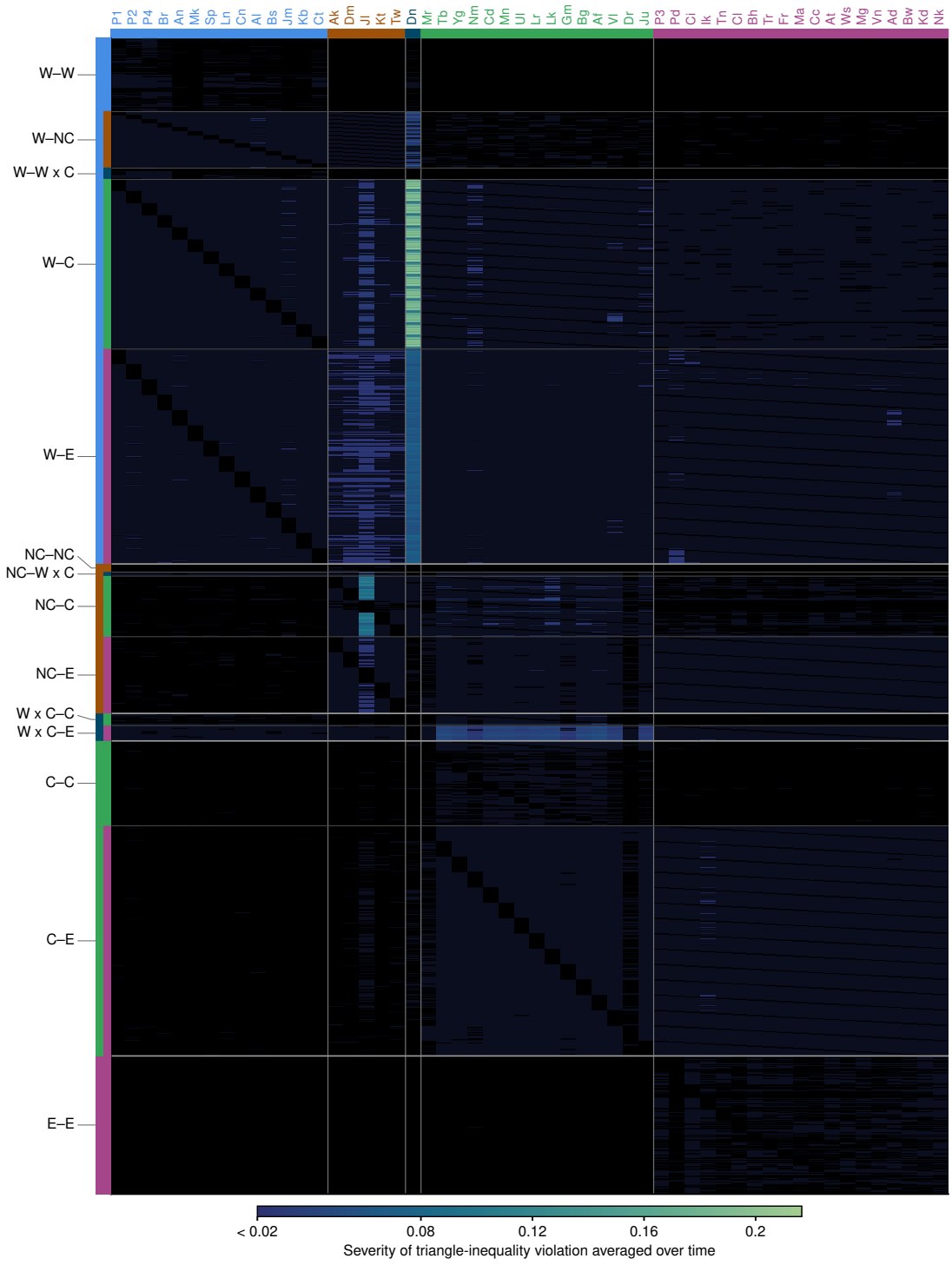

Figure 3.30: Chimpanzee samples contributing to high reconstruction error as a result of admixture. Triplets of chimpanzee samples most violating the triangle inequality. Each element is the average violation $\bar{v}(i, j; k)$ for edge samples (i, j) on the row and bridge sample k on the column, shown for all triplets; values below the 95th percentile of the non-zero elements are collapsed to a single color (as in figure 3.10).

3.4.3 RS-CoaSTT reveals an admixture history for the Nigeria-Cameroon chimpanzees

Between 10,000 and 20,000 generations ago ($\sim$ 250–500 kya), the NC samples are placed between the western and central–eastern clusters (figure 3.28B). This is consistent with an admixed history, but an intermediate position alone does not establish one. Some studies place NC as sister to the western subspecies (Brand et al., 2022; de Manuel et al., 2016; Prado-Martinez et al., 2013), yet NC was shown in other studies to be genetically closer to central and eastern chimpanzees than western chimpanzees are. This affinity was first detected from microsatellite data as a TMRCA estimate between NC and equatorial chimpanzees (central and eastern) of 0.32 Mya that is younger than that between NC and western chimpanzees (0.46 Mya), and was attributed to recent introgression through a hybrid zone in central Cameroon (Gonder et al., 2011; Mitchell et al., 2015). With whole genomes, de Manuel et al. (2016) placed NC and western chimpanzees in the same clade, but recovered a single dominant TreeMix migration edge between NC and eastern chimpanzees (weight 0.41 ± 0.007), and their SFS-based modeling infers that NC exchanged migrants with the extant central and eastern subspecies until $\sim$ 15 kya, in contrast to western chimpanzees, which became isolated $\sim$ 112 kya. In a model without western chimpanzees, the same study places the separation of NC from the central/eastern stem at $\sim$ 429 kya. None of these studies partitioned the NC genome by ancestry and dated each part separately, which is what we explore in this section.

We form two reference panels of 10 samples each: the first composed of 10 western samples, and the second of five central and five eastern samples (the set with the lowest sample index in each subspecies). For each genomic segment in the NC samples and in the western–central hybrid 'Donald' (Dn), ancestry is assigned based on whether the majority of its GNNs belong to the first or the second reference panel, as described in section 2.5.4.

Excluding 'Julie' (Jl), 66.2% of the NC genome is assigned to the central/eastern panel and 33.8% to the western panel; the fractions of 'Julie' are similar (70.5% and 29.5%), and those of 'Donald' are 17.5% and 82.5%, as expected for a second-generation hybrid. These fractions should not be read as admixture proportions, since a segment whose lineage coalesces in the population ancestral to all chimpanzees is assigned to whichever panel it happens to coalesce with first, and this is more often the larger and more diverse central/eastern panel, so the central/eastern fraction over-represents that ancestry. What the assignment provides is a partition of the genome whose two parts can be dated separately. Two sets of coalescent rates are computed between each NC sample and all other samples, one based on the regions assigned western ancestry, and the other based on the regions assigned central/eastern ancestry. From these, two sets of RS-CoaSTT embeddings are computed, shown in figure 3.31. As in the simulations involving admixture (section 2.5.4), an admixed group joins each source population in that source's embedding before the two sources join each other. The western–central hybrid behaves as expected and serves as a positive control for the partitioning, sitting within the western cluster in the western-assigned embedding and within the central cluster in the central/eastern-assigned embedding (2,000 generations ago; figure 3.31C,D). The NC samples show the same pattern at older times. In the western-assigned embedding they begin to merge with the western cluster by 17,000 generations ago. In the central/eastern-assigned embedding, they begin to merge closer to 22,000 generations ago while the western cluster is still apart (figure 3.31C,D).

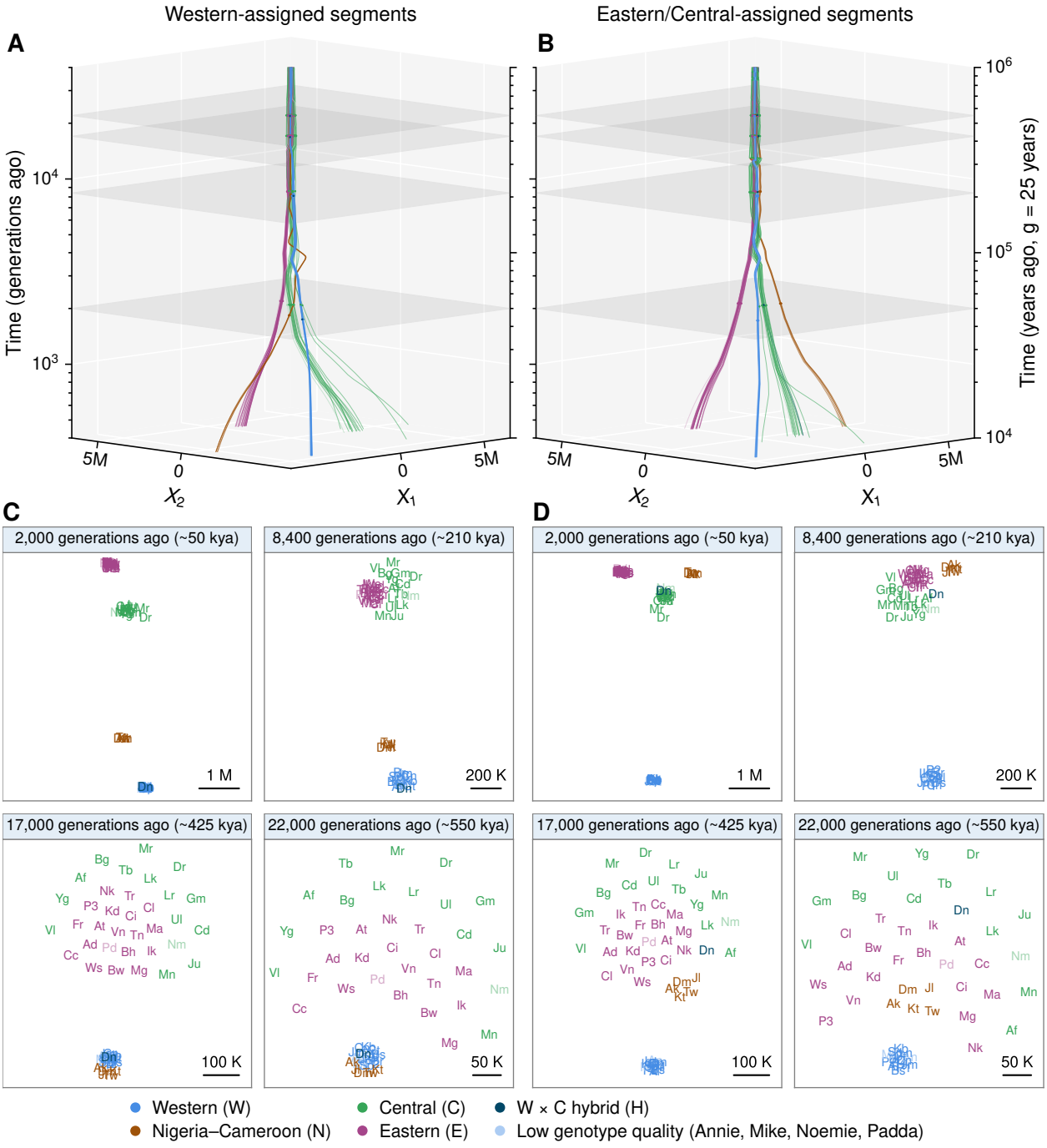

Figure 3.31: RS-CoaSTT embeddings with local ancestry assigned for NC samples based on GNNs. A, B) Embeddings through time computed from the western-assigned **(A)** and central/eastern-assigned **(B)** segments of the NC samples and the western-central hybrid; reference samples contribute all of their segments. **C, D)** Time slices of A and B at 2,000, 8,400, 17,000 and 22,000 generations ago, with the same slice on the same scale in C and D.

In each of the embeddings, the time at which the NC cluster begins to join either the western or the central/eastern cluster is not immediately obvious, so we compute a rough estimate for this time as the time at which the RCCR crosses 0.5, which we denote by $t_{0.5}$ (figure 3.32). Restricted to western-assigned segments, $t_{0.5}$ between NC and western samples is $\sim$ 419 kya. Restricted to central/eastern-assigned segments, $t_{0.5}$ between NC and eastern/central samples is $\sim$ 532 kya and $\sim$ 569 kya respectively. That is, the western-assigned segments reach $t_{0.5}$ with western samples more recently than the central/eastern-assigned segments reach it with eastern/central samples. The restriction barely changes $t_{0.5}$ with eastern/central samples relative to the genome-wide values (555 and 603 kya; these differ from the values in section 3.4 because 'Julie' is excluded here), because the central/eastern-assigned segments make up two thirds of the genome, whereas it lowers $t_{0.5}$ with western samples by $\sim$ 135 kyr, because the western-assigned third is diluted genome-wide by segments that coalesce with western samples only near the root. The near-simultaneous genome-wide crossings with all three subspecies noted in section 3.4 are therefore not a three-way separation but the average of two components with distinct histories.

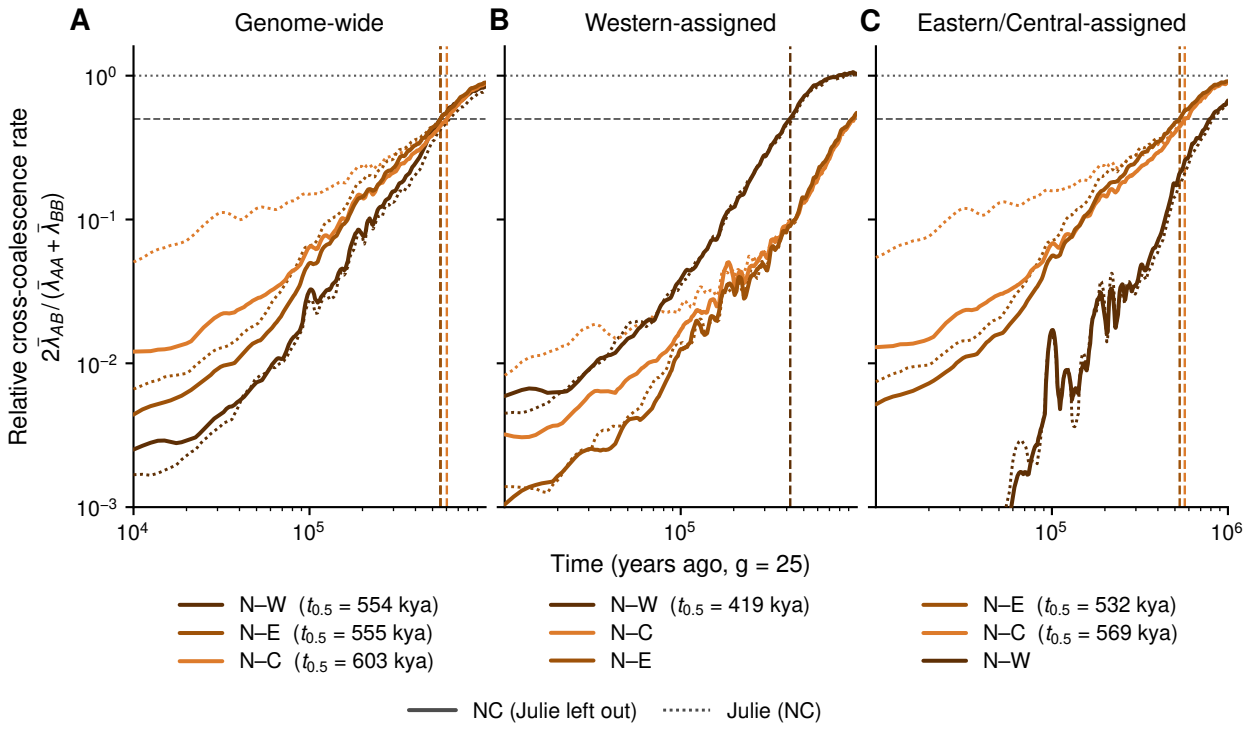

Figure 3.32: Relative cross-coalescence rates based on region-specific rates for NC samples. RCCR between the NC samples ('Julie' excluded, solid) or 'Julie' (dotted) and each of the western, central and eastern samples, computed genome-wide **(A)**, from western-assigned segments **(B)** and from central/eastern-assigned segments **(C)**. Dashed vertical lines and legend times mark where each curve first crosses 0.5. In B and C only the crossings with the assigned source are dated. Samples with low genotype quality are excluded.

These estimates are best read not as split times but as the times by which NC lineages (ancestral to specific genomic segments) and those of each source coalesce at comparable rates (section 2.4.3). The western-assigned segments behave as expected for a sister lineage of the western chimpanzees (de Manuel et al., 2016; Prado-Martinez et al., 2013), reaching $t_{0.5}$ with them ($\sim$ 419 kya) more recently than the central/eastern-assigned segments reach it with central and eastern chimpanzees. The remaining estimates suggest two refinements to existing models of NC history. First, the central/eastern-assigned segments reach $t_{0.5}$ with central and eastern chimpanzees at close times ($\sim$ 530–570 kya), both much deeper than central and eastern lineages reach it with each other ($\sim$ 210 kya). This suggests that

the gene exchange between NC and central/eastern chimpanzees, modeled by de Manuel et al. (2016), extended back to the population ancestral to both subspecies, before their separation. Because this class also contains segments that meet central/eastern lineages first only through incomplete lineage sorting, its $t_{0.5}$ does not date the gene flow directly. An alternative is gene flow from an unsampled (ghost) population related to the central/eastern stem but already distinct from it. Second, recent gene flow from central chimpanzees is visible as the central curve rising above the eastern curve within the last $\sim$ 100 kyr, and is much stronger in 'Julie', whose coalescence with central samples is 3.4–4.7 times that of the other NC samples (figure 3.32A). This refines the hybrid zone across the Sanaga (Gonder et al., 2011; Mitchell et al., 2015) as recent contact concentrated in individuals from near the river, consistent with the hypothesized origin of 'Julie' (de Manuel et al., 2016). The extent of gene flow with the central/eastern stem, and whether an unsampled source is required, can be tested by further fitting parametric demographic models.

The value of the region-specific rates analyzed here lies less in any single estimate than in how they decompose the signal. Partitioning the genome by first coalescence and computing rates conditional on each partition separates the coalescent history of the NC samples into two distinct components, and the western–central hybrid ('Donald') acts as a positive control that confirms the partition recovers known ancestry. The resulting time-resolved, sample-level picture does not select a demographic model by itself, but it generates specific hypotheses and questions that model-based methods can then test directly, such as the extent of gene flow with the central/eastern stem, the possibility of an unsampled source, and how recent central contact is distributed across NC individuals.

3.5 Discussion

In this chapter, CoaSTT is applied to two of the best-studied species in population genetics: humans and chimpanzees. As a novel method for visualizing history through low-dimensional representations of coalescent structure through time, testing it on species whose history is relatively well understood is a natural first step. The results presented here reproduce several established features of the histories of both species and provide general guidelines for how CoaSTT embeddings should be interpreted when applied to a novel dataset. Comparing embeddings obtained from different initializations and from rates derived from different ARG samples helps identify robust signals. The formation of clines in recent time was one such signal that showed up repeatedly, either on a global scale (in human embeddings) or within a population/subspecies (in chimpanzee embeddings). Applying CoaSTT to more samples that are restricted to specific geographic regions was shown to resolve recent coalescent structure at a finer scale in humans. In chimpanzees, we identified two central chimpanzees whose ancestral lineages are deeply diverged from those of the rest of the central samples. This suggests the existence of deep, sustained structure within the central subspecies. We additionally analyzed the coalescent history of NC samples with their genome partitioned according to which group they coalesce with first, resulting in a few hypotheses that could be further tested using model-based methods. This shows the utility of CoaSTT in generating hypotheses, informing users about what demographic models should be explored, and which samples should be used in fitting these models.

The quality of CoaSTT embeddings is bounded by the quality of coalescent rate estimates, which were shown to be biased when derived from ARGs inferred using current methods. The effort to characterize this bias and its effect on the embeddings in section 3.2 can help guard against over-interpretation. For instance, we have repeatedly shown that a sign that a set of samples has become exchangeable with respect to coalescence is the formation of

circular clusters. Genuine fine-scale coalescent structure within these clusters cannot be distinguished from the lingering structure produced by biased coalescent rates derived from inferred ARGs, and should therefore be verified with other methods. Future ARG-inference methods can invoke a structure-aware Empirical-Bayes strategy, using pairwise or group-specific coalescent rates estimated from an initial round of inference to improve estimates of tree topologies and internal branch lengths in subsequent rounds, thereby resulting in better estimates of pairwise coalescent rates.

Even when supplied with unbiased coalescent rates, there is a limit to what can be represented in a two-dimensional space. This issue is a common limitation of any dimensionality reduction method. By analyzing rates reconstructed from the embeddings, we have characterized two primary sources of reconstruction error: crowding, and triplets of samples that violate the triangle inequality. The former leaves a clear signature in the form of circular configurations, a configuration the optimizer places samples into when they have very similar pairwise coalescent rates. The latter can be identified directly from the input rates by constructing heatmaps that highlight such triplets of samples. We have shown that these triplets mostly correspond to admixed samples. Identifying such triplets is interesting in its own right, as it opens an avenue for studying gene flow on an individual level without invoking source populations or admixture proportions.

A common feature of all of the embeddings shown in this chapter is the dissolution of coalescent structure in deep time. Notably, the time at which this structure dissolves is much deeper for chimpanzees than it is for humans. That said, the fact that coalescent structure disappears does not entail that there was no population structure at that time. It rather indicates that, despite any structure that may have existed at that time, the ancestral lineages of present-day samples had mixed sufficiently by that time, resulting in coalescent rates that are similar between every pair of samples. The inclusion of ancient samples can unravel the coalescent structure that existed among past individuals. The current implementation

of CoaSTT supports the processing of ARGs that include both present-day and ancient samples, and exploring the coalescent structure encoded in these ARGs is a natural next step.

The analyses in this chapter illustrate the potential of CoaSTT in providing a time-resolved, individual-level view of the coalescent history encoded in an ARG. We reproduced established features of the population histories of humans and chimpanzees and proposed methods to identify patterns in the coalescent rates that are not fully captured by the embeddings, as well as patterns in the embeddings that are not robust. Applied to species whose histories are far less understood, CoaSTT can serve as a first exploratory step similar to PCA, providing an intuitive visual summary of coalescent structure through time and guiding which further questions are worth pursuing.

CHAPTER 4

THEORETICAL LIMITS OF LOW-DIMENSIONAL REPRESENTATIONS OF POPULATION STRUCTURE USING PCA

4.1 Introduction

Principal Component Analysis (PCA) provides a model-free description of population structure, and is among the most widely used statistical tools in population genetics. Its two most prominent applications are providing low-dimensional visualizations of population structure (Edwards & Cavalli-Sforza, 1965; Menozzi et al., 1978; Patterson et al., 2006) and correcting for the confounding effects of such structure in GWAS (Price et al., 2006). The pervasiveness of PCA in population genetics is likely attributable to its simplicity, as it only requires a genotype matrix as an input, which is a matrix that encodes the number of derived alleles (0, 1 or 2) carried by each individual at each bi-allelic polymorphic site. The principal components (PCs) are essentially the eigenvectors of the genetic relatedness matrix (GRM), which is obtained by computing the sample covariance between each pair of individual genotypes in the genotype matrix, so the GRM is a sample covariance matrix (in practice, the columns of the genotype matrices are mean-centered, scaled to unit variance, and perhaps dropped if the frequency of derived alleles is below some minimum threshold). The sample covariance matrix is an unbiased estimate of a true underlying theoretical covariance matrix, and the eigenvectors of that theoretical covariance matrix are the true axes of variation that the PCs attempt to estimate. Although the sample covariance matrix is an unbiased estimate of the theoretical covariance matrix, this is not necessarily the case for the sample eigenvalues and eigenvectors. For the eigenvalues and eigenvectors of the sample covariance matrix to resemble those of the underlying theoretical covariance matrix, certain conditions regarding the sample sizes and the strength of the underlying signal have to be met. This intriguing

behavior is studied by the field of Random Matrix Theory (RMT) (Johnstone & Paul, 2018).

The consequences of some of the phenomena described by RMT in the context of population genetics were discussed in a seminal paper by Patterson et al. (2006), around the time PCA began to be applied more extensively to large-scale genetic data. Such discussions grew in relevance as large-scale datasets (including a large number of individuals and markers) became more readily available with next-generation sequencing. Before then, PCA was still relevant, but was largely restricted to smaller datasets with far fewer genetic markers (Cavalli-Sforza et al., 1994; Edwards & Cavalli-Sforza, 1965; Menozzi et al., 1978). Moreover, it was still unclear how to reliably relate the results of PCA to the historical processes that gave rise to the observed genetic variation. With increasing availability of large genomic datasets, the statistical properties of PCA in the context of population genetics were more carefully studied (Patterson et al., 2006), and the relationship between the observed PCs and the underlying population history was made more apparent, providing more clarity on how PCA should be interpreted (and more importantly, how it should not) (G. McVean, 2009; Novembre & Stephens, 2008).

One of the relevant statistical phenomena described by RMT is that of "spike eigenvalue separation", and this phenomenon was the focus of the few studies that bridged RMT with population genetics (Bryc et al., 2013; Patterson et al., 2006). In this context, RMT provides the conditions under which the non-trivial eigenvalues of the sample covariance matrix are good estimates of the true eigenvalues (those of the theoretical covariance matrix). In particular, if individual genotypes are sampled from K subpopulations, then the theoretical covariance matrix is expected to have K significant eigenvalues (or $K-1$ if the genotype matrix is mean-centered), with their associated eigenvectors each reflecting a mode of population structure. However, it is not guaranteed that exactly K (or $K-1$) significant sample eigenvalues will be observed. The kth sample eigenvalue will separate from the bulk of the non-significant eigenvalues only if the corresponding k^{th} mode of population structure is

"strong enough". This transition from being absorbed by the bulk of the spectrum, which is populated by non-significant eigenvalues, to separating from the bulk is known in the RMT literature as the Baik–Ben Arous–Péché (BBP) phase transition (Baik et al., 2005). Linking it to population genetics, Patterson et al. (2006) proved the now well-known result that, if N individuals are sampled at L independent SNPs from two subpopulations that are separated by some value of F_{ST} (with $\frac{N}{2}$ samples from each subpopulation), the largest eigenvalue of the sample covariance matrix will separate from the bulk only if

$$F_{ST} > \frac{1}{\sqrt{NL}}. \tag{4.1}$$

This result provided important intuition for practitioners regarding the limits of PCA in revealing fine structure, despite being derived for the specific demographic scenario of a simple population split. Bryc et al. (2013) formalized this behavior further, proving that with sufficient sample sizes, it is possible to estimate the underlying number of subpopulations by counting the number of eigenvalues that separate from the bulk. They briefly allude to the significance of studying the properties of sample eigenvectors (which are the estimated PCs), but state that they are "notoriously difficult to analyze mathematically." Practically though, the PCs are the relevant observables that are further utilized for visualizing structure or correcting for stratification in GWAS, and quantifying the conditions under which the true axes of variation are well-estimated by the sample PCs is crucial.

In this chapter, we invoke the genealogical description of the GRM developed by G. McVean (2009) in addition to recent developments from RMT (Ding & Yang, 2021) to describe the conditions under which PCA reliably reveals the underlying axes of variation for a generic demographic history (beyond the simple population split case) and an arbitrary sampling scheme. We start by briefly reviewing a set of results from RMT that are relevant to our setting, then we present a general method to compute the significant eigenvalues of the theoretical covariance matrix, which correspond to the strength of each mode of population

structure. We combine this method with RMT results that are adapted to our setting to obtain expressions for the conditions under which a PC is a good estimate of the true axis of variation, as a function of sample sizes and relevant demographic parameters. We then apply these methods to demographic scenarios that result in discrete population structure as well as spatial structure, generalizing the condition specified by Patterson et al. (2006) to further relevant demographic scenarios, with a focus on the properties of the sample eigenvectors (the PCs) as opposed to the sample eigenvalues.

4.2 Theoretical Overview

4.2.1 Overview of relevant Random Matrix Theory results

Johnstone and Paul (2018) provide an excellent review of the applications of RMT results to PCA. In this section, we present a summary of the relevant RMT phenomena in the context of genotype data. Consider N haplotypes sampled at L independent, bi-allelic loci, with an associated sample covariance matrix $\hat{G} = \frac{1}{L}\tilde{X}\tilde{X}^T$, where $\tilde{X}$ is the genotype matrix with columns centered to zero mean and normalized to unit variance. $\hat{G}$ (the observed GRM) is an unbiased estimate of the theoretical covariance matrix G (where $G = E[\hat{G}]$). Denote $\{\lambda_i, v_i : i = 1, \cdots, N\}$ as the set of theoretical eigenvalues and eigenvectors of G, and the corresponding set of sample eigenvalues and eigenvectors of $\hat{G}$ by $\{\hat{\lambda}_i, \hat{v}_i : i = 1, \cdots, N\}$. Assume that the haplotypes were sampled from two different populations that have the same diversity, with $N/2$ samples from each population. In this case, the eigenvalues of the theoretical covariance matrix will all be approximately 1 except for a single eigenvalue with a value greater than 1, and the theoretical eigenvector corresponding to this eigenvalue clusters the samples into two groups corresponding to the two populations. We refer to the theoretical eigenvalues of value 1 as the "bulk" of the spectrum, and the single eigenvalue that is larger than 1 as the "spike". In the limit where N and L both go to infinity, with

their ratio $\gamma = \frac{N}{L}$ held constant, the sample eigenvalues and eigenvectors of $\hat{G}$ will exhibit the following phenomena:

1. **Eigenvalue spreading:** Even though the theoretical bulk of the spectrum is composed of identical eigenvalues, the associated sample eigenvalues will not be identical, and will be distributed according to the Marchenko–Pastur (MP) distribution. The edge of the MP distribution is located around $\lambda_+ = (1 + \sqrt{\gamma})^2$. Additionally, the largest sample eigenvalue in the bulk will be distributed according to the so-called Tracy–Widom (TW) distribution.

2. **BBP phase transition:** If the spike of the theoretical spectrum is below the "BBP threshold" ($\lambda_1 < 1 + \sqrt{\gamma}$), then the corresponding largest sample eigenvalue ($\hat{\lambda}_1$) will be absorbed into the bulk (and become TW-distributed). On the other hand, if it is strong enough and exceeds that threshold ($\lambda_1 > 1 + \sqrt{\gamma}$), then it will separate from the bulk, and have a normal distribution with a mean slightly larger than the actual value of the spike. Note that in this simple case, $\lambda_1 \approx 1 + \frac{N}{2} F_{ST}$, reproducing the critical F_{ST} of $F_{ST}^c = \frac{2}{\sqrt{NL}}$. The factor of 2 relative to equation 4.1 arises because N counts haploid rows here rather than diploid individuals: halving the row count to $N/2$ diploids accounts for a factor of $\sqrt{2}$, and the doubled covariance between diploid genotypes from the same population accounts for the other.

3. **Eigenvector inconsistency:** The eigenvector v_1 associated with the spike λ_1 also exhibits a phase transition. If the spike λ_1 is not strong enough, then the sample eigenvector (the first PC) is almost exactly orthogonal to the theoretical eigenvector. In this case, the first PC will capture pure noise rather than structure. If the spike is strong enough, the squared cosine similarity between the sample eigenvector and the theoretical eigenvector will increase, reaching 1 as $\gamma \to 0$. Specifically (Benaych-

Georges & Nadakuditi, 2011; Paul, 2007)

$$\langle \hat{v}_1, v_1 \rangle^2 = \begin{cases} 0, & 1 < \lambda_1 < 1 + \sqrt{\gamma} \\ \dfrac{1 - \gamma/(\lambda_1 - 1)^2}{1 + \gamma/(\lambda_1 - 1)}, & \lambda_1 > 1 + \sqrt{\gamma}. \end{cases} \tag{4.2}$$

This last phenomenon of eigenvector inconsistency is the one that interests us the most, as it exactly quantifies the similarity between the PCs and the true axes of variation in terms of sample sizes (N, L) and the strength of the modes of structure specified by the spike eigenvalues. To obtain a similar expression for the cosine similarity under a generic demography, we need two ingredients. First, we need to derive the $K-1$ "structure eigenvalues" or "spikes" that correspond to the axes of variation as a function of relevant demographic parameters and the sampling scheme. We can do this by invoking G. McVean (2009)'s results, where it was shown that G can be expressed in terms of mean pairwise coalescent times. This expression is only valid if the columns of the genotype matrix are not normalized to unit variance. Additionally, as we will show, the bulk eigenvalues are not generically identical as in the case described above. Therefore, the second thing we need is to find a corresponding expression for the cosine similarity between a given PC and the true axis of variation when the columns of the data matrix have non-identical variance, and the bulk eigenvalues are also non-identical.

4.2.2 Writing the theoretical GRM as a function of mean pairwise coalescent times

To derive the spikes of the theoretical covariance matrix as a function of demographic parameters, we invoke a result by G. McVean (2009) that the theoretical covariance matrix can be written in terms of mean pairwise coalescent times (which themselves depend on demographic parameters). We first reproduce this result using slightly different notation that

will simplify later calculations.

Throughout this chapter, we consider genotype matrices where rows represent haplotypes rather than individuals, in which case the elements of the matrix take values $\{0, 1\}$ instead of $\{0, 1, 2\}$. This simplifies much of the analyses without loss of generality. Let $Z \in \{0, 1\}^{N \times L}$ be a binary genotype matrix of N haplotypes sampled at L independent bi-allelic SNPs, where $Z_{il} = 1$ if sample i has the derived allele at site l. Define the mean-centered version of Z as

$$X_{il} = Z_{il} - p_l,$$

where $p_l = \frac{1}{N} \sum_{j=1}^{N} Z_{jl}$. More conveniently, we can use a centering matrix $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ to define the mean-centered genotype matrix as

$$X = HZ.$$

$\hat{G}$ can be expressed more conveniently in terms of the matrix Δ that counts the fraction of *polymorphic* sites where each pair of samples differs, defined as $\Delta_{ij} = \frac{1}{L} \sum_{l=1}^{L} (Z_{il} - Z_{jl})^2$ (in our later analyses, we only consider polymorphic sites). Expressing it this way highlights how PCA can be thought of as simply applying classical MDS on the matrix of pairwise genetic distance. We can then rewrite $\hat{G}$ as

$$\begin{aligned} \hat{G} &= \frac{1}{L} X X^T \\ &= H \frac{Z Z^T}{L} H \\ &= -\frac{1}{2} H \Delta H, \end{aligned} \tag{4.3}$$

where the covariance of the uncentered binary genotype matrix is rewritten in terms of the

matrix of pairwise differences. This follows from the fact that $H(L\Delta)H = -2HZZ^TH$.

$$\begin{aligned} H(L\Delta)H &= H\left[\mathrm{diag}(ZZ^T)\mathbf{1}^T + \mathbf{1}\mathrm{diag}(ZZ^T)^T - 2ZZ^T\right]H \\ &= H\,\mathrm{diag}(ZZ^T)\left(\mathbf{1}^TH\right) + (H\mathbf{1})\,\mathrm{diag}(ZZ^T)^TH - 2HZZ^TH \\ &= -2HZZ^TH. \end{aligned}$$

The last step follows from $H\mathbf{1} = 0$ and $\mathbf{1}^TH = 0$.

The theoretical covariance matrix $G = E[\hat{G}]$ can be found by computing $E[\Delta_{ij}]$. For a given pair of samples, define $D_l = (Z_{il} - Z_{jl})^2$, a Bernoulli random variable with a value of 1 if samples (i, j) differ at site l. Define another Bernoulli random variable S_l that takes a value of 1 if a mutation occurs on site l (that is, site l is polymorphic). With these definitions, $E[\Delta_{ij}] = E[D_l|S_l = 1]$, since we are only considering polymorphic SNPs. This can be computed using Bayes' rule and using the rule of total expectation, conditioning on the genealogy $\mathcal{T}^l$ at locus l. Under the infinite-sites model, mutations occur on a genealogy at a Poisson rate of μ per unit branch length. Therefore, $P(D_l|\mathcal{T}^l) = 1 - e^{-2\mu T^l_{ij}}$, where T^l_{ij} is the coalescent time between samples (i, j) at locus l. Similarly, $P(S_l|\mathcal{T}^l) = 1 - e^{-\mu T^l_{\mathrm{tot}}}$, where T^l_{tot} is the total branch length of the tree at locus l. Defining $G_X(s) = E[e^{-sX}]$ as the generating function of a random variable X, we can write down an exact expression for $E[\Delta_{ij}]$.

$$
\begin{aligned}
E[\Delta_{ij}] &= E[D_l | S_l = 1] \\
&= P(D_l | S_l = 1) \\
&= \frac{P(D_l = 1, S_l = 1)}{P(S_l = 1)} \\
&= \frac{P(D_l = 1)}{P(S_l = 1)} \\
&= \frac{E[P(D_l = 1 | \mathcal{T}^l)]}{E[P(S_l = 1 | \mathcal{T}^l)]} \\
&= \frac{E[1 - e^{-2\mu T^l_{ij}}]}{E[1 - e^{-\mu T^l_{\text{tot}}}]} \\
&= \frac{1 - G_{T_{ij}}(2\mu)}{1 - G_{T_{\text{tot}}}(\mu)} \\
&\approx 2 \frac{E[T_{ij}]}{E[T_{\text{tot}}]}.
\end{aligned}
$$

The last approximation is valid under small mutation rate μ, which cancels from the numerator and denominator. Therefore,

$$
G = E[\hat{G}] = -\frac{1}{E[T_{\text{tot}}]} HTH. \tag{4.4}
$$

Note that we have slightly abused the notation for convenience, using T to express the matrix of expected values of pairwise coalescent times, where $T \in \mathbb{R}^{N \times N} : T_{ij} \equiv E[T_{ij}]$, dropping the expectation for a more concise notation.

Without normalization, G has a nuisance scaling factor $\frac{1}{E[T_{\text{tot}}]}$ that is not trivial to compute for a general demography. We can alternatively define a normalized version of the

genotype matrix, where it is normalized by the *pooled* average of pairwise differences.

$$\begin{aligned}\bar{\Delta} &= \frac{1}{N(N-1)}\sum_{i\neq j}\Delta_{ij},\\ Y &= \frac{X}{\sqrt{\bar{\Delta}/2}}\\ \hat{G} &= \frac{1}{L}YY^T = -\frac{H\Delta H}{\bar{\Delta}}\end{aligned} \tag{4.5}$$

Under this normalization the theoretical GRM can be conveniently expressed as a function of mean pairwise coalescent times only. Since $E[\bar{\Delta}] = \frac{2\bar{T}}{E[T_{\text{tot}}]}$, where $\bar{T}$ is the sample average of mean coalescent times ($\bar{T} = \frac{1}{N(N-1)}\sum_{i\neq j} T_{ij}$), the theoretical GRM is given by

$$\begin{aligned}G &= E\left[\frac{-H\Delta H}{\bar{\Delta}}\right]\\ &\approx -\frac{2}{E[T_{\text{tot}}]}HTH \cdot \frac{E[T_{\text{tot}}]}{2\overline{T}}\\ &= -\frac{HTH}{\overline{T}},\end{aligned} \tag{4.6}$$

where the expectation of the ratio is approximated as the ratio of the expectations, an approximation that is valid for large L.

Note that this normalization is different from the per-locus normalization that is more standard in PCA, where each column of the genotype matrix is individually normalized by $\sqrt{p_l(1-p_l)}$. Normalizing by the pooled heterozygosity allows the theoretical GRM to be expressed purely in terms of the matrix of pairwise coalescent times, which is more tractable to analytically calculate for a variety of demographic models, and simple to numerically compute for more complex demographies.

If, on the other hand, the sample covariance matrix was computed with the columns of the genotype matrix each normalized to unit variance ($\hat{G} = \frac{1}{L}\tilde{X}\tilde{X}^T$, with $\tilde{X}_{il} = \frac{Z_{il}-p_l}{\sqrt{p_l(1-p_l)}}$),

the corresponding theoretical covariance matrix would be given by

$$G = -\frac{E[V]}{E[T_{\text{tot}}]} HTH + H\Sigma H, \tag{4.7}$$

where $V_l = \frac{1}{p_l(1-p_l)}$ is the inverse of the heterozygosity at locus l, and $\Sigma \in \mathbb{R}^{N\times N} : \Sigma_{ij} \equiv \text{Cov}(X_{il}X_{jl}, V_l \mid S_l = 1)$. The second covariance term is quite non-trivial to compute. The theoretical covariance matrix derived from normalized genotype matrices is therefore not simple to work with, and finding analytical expressions (or even numerical solutions) for the eigenvalues resulting from a generic demography will be intractable. Since our aim is to express the cosine similarity in terms of demographic parameters and sample sizes, we proceed with G resulting from genotype matrices with the pooled normalization rather than the per-locus normalization, while recognizing that the expressions for the thresholds that we derive might be slightly different for the normalized version. We conjecture that the insights gained from analysis under this analytically tractable normalization carry over to the more standard normalization, at least to leading order.

4.3 Methods

The expression for the expected squared cosine similarity between the observed PC and the theoretical eigenvector (equation 4.2) is quite a remarkable result from RMT. It is informative about how well the observed PC estimates the true axis of variation as a function of the theoretical eigenvalue corresponding to that PC and the aspect ratio $\gamma = \frac{N}{L}$. The theoretical eigenvalue is a proxy for the strength of the mode of structure resolved by that PC, and can be obtained by diagonalizing the theoretical covariance matrix G. Conveniently, G can be expressed in terms of pairwise coalescent times, which themselves can be expressed in terms of relevant demographic parameters which correspond to the strength of population structure. Therefore, the eigenvalues that determine the cosine similarity can be expressed

in terms of these demographic parameters, yielding expressions for how well the PCs are expected to reflect true structure as a function of these parameters. The caveat, however, is that we cannot use equation 4.2 for the cosine similarity directly in our case, since it assumes that the columns of the data matrix have unit variance, and that the bulk eigenvalues are all identical. Neither of these assumptions holds in our case.

To answer the question of how well the PCs estimate the true axis of variation, we need the following two ingredients. First, we need to express the eigenvalues of the theoretical covariance matrix in terms of the relevant demographic parameters and the sampling scheme. Second, we need to find the more general expression for cosine similarity that is valid for our setting. The next sections show how the spectrum of G can be decomposed into a set of "structure" eigenvalues and "bulk" eigenvalues that are both directly related to intuitive demographic parameters, and how this decomposition can be coupled with more general RMT results to obtain a general expression for cosine similarity. Finally, we describe a strategy to compute the expected cosine similarity from simulation data, which would allow us to compare the theoretical expectation with simulation experiments.

4.3.1 Spectral decomposition of the theoretical covariance matrix under a generic demography and sampling scheme

Consider a demographic history with K extant subpopulations (demes), where $\vec{n} = (n_1, n_2, ..., n_K)^T$ haplotypes are sampled from each subpopulation. Define $T^{\text{pop}} \in \mathbb{R}^{K \times K}$ as the matrix of mean coalescent times between two samples from populations a and b. Specifically, for haplotypes i, j sampled from populations $a(i), a(j)$, their mean coalescent time is given by $T_{ij} = T^{\text{pop}}_{a(i),a(j)}$. Note that the diagonal of T is zero (since the average coalescent time of a sample with itself is zero by construction), whereas the diagonal of T^{pop} is non-zero and represents mean within-population coalescent times (which will also be denoted by the vector

$\vec{\tau}$). In this setting, T has a block structure given by

$$T_{[a][b]} = \begin{cases} \tau_a \left(\mathbf{1}_{n_a \times n_a} - I_{n_a} \right), & a = b \\ T_{ab}^{\text{pop}} \, \mathbf{1}_{n_a \times n_b}, & a \neq b \end{cases} \tag{4.8}$$

Theorem: If T has a block structure described by equation 4.8, the spectrum of $G = -\frac{HTH}{\bar{T}}$ can be decomposed as

$$\text{spec}(G) = \{0\} \cup \left\{ \frac{1+\nu_k}{1+\alpha} : \nu_k \in \text{spec}(\hat{F}|_{\psi^\perp}) \right\} \cup \left\{ \frac{1+\epsilon_a}{1+\alpha} \text{ with multiplicity } n_a - 1, a = 1, ..., K \right\}$$

with $(\nu_k, \hat{F}, \alpha, \psi, \epsilon_a)$ defined below. This theorem tells us that the N eigenvalues of G can be decomposed into 3 sets. The first is the trivial zero eigenvalue with an associated eigenvector $\mathbf{1}_N$ that emerges because of the mean-centering operation. The second and most important set contains the $K-1$ *structure* eigenvalues, or the spikes, with their associated eigenvectors being the axes of variation estimated by PCA. The structure eigenvalues (and associated eigenvectors) can be found by solving the smaller problem of diagonalizing the $K \times K$ matrix $\hat{F}$, and discarding the trivial eigenvalue associated with an eigenvector ψ. The last set of eigenvalues represents the *floor* or the *bulk* of the spectrum, and are important for specifying the exact detectability threshold. $\vec{\epsilon}$, $\hat{F}$ and α are all quantities with an intuitive meaning and are fully described in terms of population-level mean pairwise coalescent times T^{pop} and sampling scheme $\vec{n}$. Given the following derived quantities:

$$N = \sum_a n_a, \qquad \phi_a = \frac{n_a}{N}, \qquad \psi_a = \sqrt{\phi_a}, \qquad \hat{P} = I_K - \psi\psi^T,$$

$$\tau_a = T_{aa}, \qquad \bar{\tau} = \sum_a \phi_a \tau_a, \qquad \epsilon_a = \frac{\tau_a}{\bar{\tau}} - 1, \qquad (F_2)_{ab} = T_{ab} - \frac{\tau_a + \tau_b}{2}.$$

$\hat{F}$ and α are defined as:

$$\tilde{F}_{ab} = D_{\vec{\epsilon}} - \frac{N}{\bar{\tau}} D_{\vec{\psi}} F_2 D_{\vec{\psi}} = \begin{cases} \epsilon_a, & a = b \\ -N\sqrt{\phi_a \phi_b} \dfrac{(F_2)_{ab}}{\bar{\tau}}, & a \neq b \end{cases} \tag{4.9}$$

$$\hat{F} = \hat{P}\tilde{F}\hat{P}, \tag{4.10}$$

$$\alpha = \frac{\vec{n}^T F_2 \vec{n}}{\bar{\tau} N(N-1)}, \tag{4.11}$$

where $D_{\vec{x}} = \text{diag}(\vec{x})$. In these definitions, ϵ_a represents how much the diversity of population a (represented by its mean within-population coalescent time) differs from the average diversity of the populations ($\bar{\tau}$) weighted by their sample sizes. F_2 is the known F-statistic expressed in terms of mean pairwise coalescent times, and is a measure of how divergent two populations are (Patterson et al., 2012a; Peter, 2016). Matrices $D_{\vec{\psi}}$ and $\hat{P}$ are defined by the sampling scheme. Therefore, $\hat{F}$ can be thought of as the matrix of expected F_2 values between pairs of populations, with a diagonal representing how the diversity of each population differs from the sample mean, transformed in a manner defined by the sampling scheme. The structure eigenvalues that signify the strength of each axis of structure can be simply obtained by finding the $K-1$ non-trivial eigenvalues of $\hat{F}$. The floor eigenvalues, representing the within-population noise, are directly obtained from $\vec{\epsilon}$, and will generically deviate from 1 by the extent to which each population's diversity deviates from the mean diversity.

To prove this theorem, we first show that decomposing $\mathbb{R}^N$ into two orthogonal subspaces $\{\mathcal{B}, \mathcal{B}^\perp\}$ naturally separates the structure spectrum (which lives purely in $\mathcal{B}$) from the floor spectrum (which lives in $\mathcal{B}^\perp$). We then derive the spectrum associated with each subspace, showing how the particular definition of $\hat{F}$ allows the structure eigenvalues to be written as $\lambda_k = \frac{1+\nu_k}{1+\alpha}$ (see Appendix 4A for the complete proof). The advantage of writing it this way is that the eigenvalues ν_k of $\hat{F}$ directly give the excess signal contained in a given axis of

structure, which is the relevant quantity for computing the cosine similarity.

Note that $\hat{F}$ can also be written in terms of the matrix of expected pairwise F_{ST} between populations. Following the genealogical definition of F_{ST} (Slatkin, 1991)

$$(F_{ST})_{ab} = 1 - \frac{\tau_a + \tau_b}{2T_{ab}} \tag{4.12}$$

Elements of $\tilde{F}$ can be written in terms of F_{ST} by substituting

$$\frac{(F_2)_{ab}}{\bar{\tau}} = \frac{(F_{ST})_{ab}}{1 - (F_{ST})_{ab}} \frac{\tau_a + \tau_b}{2\bar{\tau}}. \tag{4.13}$$

The structure eigenvalues can therefore be viewed as the eigenvalues of a $K \times K$ matrix of F_{ST} values, whose elements are scaled in a particular way due to unequal diversity and uneven sampling.

Extending the spectrum calculation for spatially structured populations

One of the main advantages of PCA is its ability to reveal continuous structure as well as discrete structure, and spatial structure is one of the most prevalent sources of continuous structure in natural populations (Slatkin, 1987, 1993). When PCA is applied to a spatially structured population (under certain conditions), the largest PCs tend to align with the geographic location of the samples, with the lower PCs reflecting structure at finer spatial resolution (Novembre & Stephens, 2008; Novembre et al., 2008). We are interested in studying the conditions under which PCA can reveal such spatial structure. To achieve this, we need to find a general way to compute the eigenvalues of the theoretical covariance matrix for samples obtained from a spatially structured population. The eigenvectors of theoretical covariance matrices resulting from a spatially structured population were studied by Novembre and Stephens (2008), and were shown to be the columns of 1D/2D discrete cosine transform (DCT) matrices, which are 1D/2D sinusoidal functions of increasing frequency.

We extend this to more general spatial settings (i.e. inhomogeneous migration and uneven sampling), highlighting how the *eigenvalues* that reflect the strength of spatial structure can be derived as well.

Consider K subpopulations (demes), and define m_{ij} as the probability that a lineage was present in deme j in the previous generation and migrated to deme i. We are interested in deriving a general expression for the structure eigenvalues ν_k as a function of migration rates m_{ij} and a sampling scheme $\vec{n}$. We have already established that finding the matrix of mean coalescent times between all demes and within each deme is sufficient for deriving ν_k. In a general setting, this matrix can be found by solving the equation

$$QT + TQ^T + J = \text{diag}(\vec{c})\text{diag}(\vec{\tau}), \tag{4.14}$$

where Q is the generator of the migration process ($Q_{ij} = m_{ij}$ for $i \neq j$ and $Q_{ii} = -\sum_{j \neq i} m_{ij}$), i.e. $Q = -\Lambda^{(m)}$ with $\Lambda^{(m)}$ the (positive semi-definite, under symmetric migration) graph Laplacian of the migration rate matrix, T is the matrix of pairwise coalescent times, $\vec{\tau}$ is a vector of the diagonal of T, $\vec{c}$ is a vector of within-population coalescent rates ($c_k = \frac{1}{2N_e^{(k)}}$) and $J = \mathbf{1}\mathbf{1}^T$ (Lundgren & Ralph, 2019; Notohara, 1990; Tufto et al., 1998; Whitlock & Barton, 1997). This is a matrix equation that can be numerically solved for T given the migration matrix and per-deme effective population sizes. In the special case of symmetric hitting times (the expected time that a lineage starting from deme i reaches deme j = expected time that a lineage starting from deme j reaches deme i), a neat approximation is available in terms of the resistance distance (McRae, 2006). Using an analogy from electric circuits where the migration rate between two demes can be thought of as the conductance between two nodes in a circuit, with R_{ij} being the "effective resistance" between the nodes, it can be shown that

$$(F_2)_{ij} = T_{ij} - \frac{\tau_i + \tau_j}{2} \approx \frac{K}{4} R_{ij}^{(m)}, \tag{4.15}$$

where the resistance distance between two demes is related to the Moore–Penrose pseudo-inverse of the Laplacian of the migration rate matrix

$$R_{ij}^{(m)} = (\Lambda^{(m)})_{ii}^{+} + (\Lambda^{(m)})_{jj}^{+} - 2(\Lambda^{(m)})_{ij}^{+}. \tag{4.16}$$

McRae (2006) and Petkova et al. (2016) state this relation as one quarter of the commute time, and the reason a factor of K appears is that the commute time is equal to $KR_{ab}^{(m)}$ for a uniform and stationary chain (Chandra et al., 1996; Tetali, 1991).

Using the resistance distance approximation for pairwise coalescent times opens the door to a simple path for deriving the structure eigenvalues ν_k given a set of migration rates, where we compute the pseudo-inverse of the Laplacian of the migration rate matrix, then the resistance distance, which is proportional to F_2, from which we then compute $\hat{F}$. There is a subtlety worth mentioning here. $\hat{F}$ additionally requires knowledge of within-population coalescent times $\vec{\tau}$. When only R is specified, it is possible to find a set of coalescent rates $\vec{c}$ and coalescent times $\vec{\tau}$ that satisfy equation 4.14. However, given a set of specified coalescent rates $\vec{c}$ in addition to R, it is not always possible to find a set of $\vec{\tau}$ that satisfies the equation (Lundgren & Ralph, 2019). In the special case where N_e is identical for all demes, and under conservative and isotropic migration, the within-population coalescent times are constant among all demes, and are given by $\tau = 2N_eK$ (Strobeck, 1987). Using this identical τ for all demes is a useful approximation in many cases, and would allow $\hat{F}$ and therefore ν_k to be computed directly from the migration rate matrix.

We have two paths for computing the structure eigenvalues given a spatially structured population with a migration matrix. The first is to solve equation 4.14 in terms of T, which is possible numerically (Shen & Novembre, 2026), but will not yield closed-form solutions. The other is to assume that within-population diversity is approximately constant and given by $\tau \approx 2N_eK$ for all demes. In this case $\bar{\tau} \approx 2N_eK$ and the diagonal elements of $\tilde{F}$ in

equation 4.9 vanish. Additionally, if we define

$$\bar{m} = \frac{1}{|\{(i,j) : m_{ij} \neq 0\}|} \sum_{i,j:m_{ij}\neq 0} m_{ij} \quad \text{(mean nonzero migration rate)}$$

$$M = 4N_e\bar{m}$$

$$\Lambda_{ij} = \frac{1}{\bar{m}}\Lambda^{(m)}_{ij} = \begin{cases} \frac{1}{\bar{m}}\sum_{j\neq i} m_{ij}, & i = j \\ -\frac{1}{\bar{m}} m_{ij}, & i \neq j \end{cases}$$

$$R_{ij} = \Lambda^+_{ii} + \Lambda^+_{jj} - 2\Lambda^+_{ij} = \bar{m}\, R^{(m)}_{ij}$$

$\hat{F}$ can be expressed directly in terms of the resistance distance as

$$\hat{F} = -\frac{N}{2M}\hat{P}D_{\vec{\psi}}RD_{\vec{\psi}}\hat{P}. \tag{4.17}$$

This expression simplifies even further under even sampling. With $n_a = n$ equal samples drawn evenly from each deme, $D_{\vec{\psi}} = \frac{1}{\sqrt{K}}I$, and $\hat{P} \equiv P = I - \frac{1}{K}\mathbf{1}\mathbf{1}^T$ is the projector onto the subspace orthogonal to $\mathbf{1}$. Note that for any $v \in \mathbf{1}^\perp$

$$v^T R v = -2v^T\Lambda^+ v.$$

This follows directly from the definition of R, which can be written in matrix form as $R = \text{diag}(\Lambda^+)^T J + J\text{diag}(\Lambda^+) - 2\Lambda^+$, with $J = \mathbf{1}\mathbf{1}^T$, as J kills any vector orthogonal to $\mathbf{1}$. This means that, restricted to the subspace orthogonal to the trivial eigenvector $\mathbf{1}$, both R and Λ^+ have the same eigenvectors, with eigenvalues scaled by -2. Additionally, Λ^+ and Λ share the same eigenvectors, with reciprocal eigenvalues (in the case of symmetric Λ). Therefore, computing R or Λ^+ is not needed at all, and the eigenvalues of Λ are sufficient to obtain ν_k. Defining ℓ_k as the non-zero eigenvalues of Λ, the structure eigenvalues are given

by

$$\nu_k = \frac{N}{MK}\frac{1}{\ell_k}. \tag{4.18}$$

Therefore, under even sampling and the validity of the resistance distance approximation, finding the structure eigenvalues that represent the strength of spatial structure amounts to finding the eigenvalues of the migration Laplacian. With uneven sampling, the eigenvalues can be computed from the more general matrix defined in equation 4.17.

Defining F_{ST} for spatially structured populations. If we additionally define an F_{ST} for the spatial population, we can re-express the structure eigenvalues in terms of this pooled F_{ST} by relating it to the scaled migration rate M. We can define a pooled F_{ST} of the whole population as

$$F_{ST} = 1 - \frac{\bar{\tau}}{\bar{T}}, \tag{4.19}$$

where $\bar{\tau}$ is the average mean coalescent time of two individuals sampled from the same deme, and $\bar{T}$ is the average mean coalescent time over all distinct pairs of sampled individuals, including pairs from the same deme (equation 4.53). For any sampling scheme, $\bar{T} = \bar{\tau}(1+\alpha)$ (Appendix 4A, Lemma 0), so the pooled F_{ST} can be re-expressed in terms of α (defined in (4.11)) as

$$F_{ST} = \frac{\alpha}{1+\alpha}. \tag{4.20}$$

In the special case of even sampling and equal diversity, α can be simply expressed in terms of the eigenvalues ℓ_k of the migration Laplacian.

$$\begin{aligned}
\alpha &= \frac{n^T F_2 n}{\tau N(N-1)} \\
&= \frac{n^2 \sum_{a,b}(F_2)_{ab}}{2N_e K \cdot nK(nK-1)} \\
&\approx \frac{n^2 \sum_{a,b}(F_2)_{ab}}{2N_e K \cdot n^2K^2} \text{ (large } nK) \\
&\approx \frac{\sum_{a,b} R^{(m)}_{ab}}{8N_e K^2} \text{ (using (4.15))} \\
&= \frac{\sum_{a,b} R_{ab}}{8N_e \bar{m} K^2} \text{ (using } R^{(m)} = R/\bar{m}) \\
&= \frac{2K\mathrm{tr}(\Lambda^+) - 2\cancelto{0}{\mathbf{1}^T\Lambda^+\mathbf{1}}}{8N_e\bar{m}K^2} \\
&\approx \frac{\mathrm{tr}(\Lambda^+)}{MK} \\
&\approx \frac{1}{MK}\sum_{k=1}^{K-1}\frac{1}{\ell_k}.
\end{aligned}$$

With α expressed in terms of the eigenvalues of the migration Laplacian, the pooled F_{ST} can be expressed as

$$F_{ST} = \frac{\sum_{k=1}^{K-1}\frac{1}{\ell_k}}{MK + \sum_{k=1}^{K-1}\frac{1}{\ell_k}} \tag{4.21}$$

This expression will be useful for relating how well a PC recovers continuous spatial structure to the pooled F_{ST} of the spatial population.

4.3.2 Critical point of phase transition and cosine similarity expressed in terms of the spectrum of G

The classical results of RMT summarized by Johnstone and Paul (2018) depend on two assumptions that are violated in our case. First, the bulk eigenvalues are assumed to be identical. We have shown that this is not generally true, as they differ when within-population diversity is not identical for all populations. This case is treated by invoking generalized MP distributions (Silverstein, 1995), which have an edge that is different from the simple $\lambda_+ = (1 + \sqrt{\gamma})^2$, and generalized expressions for the point of phase transition and the cosine similarity. Second, the columns of the data matrix (the genotype matrix in our case) are assumed to have unit variance. This is also not true in our case, since we have chosen to skip normalization for the sake of expressing the theoretical eigenvalues (and hence the cosine similarity) directly in terms of demographic parameters, which will be difficult if the genotype matrices are normalized (see equation 4.7). We are therefore left with the problem of finding how the RMT results change as a result of non-identical bulk eigenvalues (or heteroskedasticity on the row side) combined with unequal variances of the data matrix (heteroskedasticity on the column side).

The separable spiked covariance model

Fortunately, results for this particular setting have been derived in the RMT literature by Ding and Yang (2021). The treatment of Ding and Yang rests on the primary assumption that the sample covariance matrix is "separable". An $N \times L$ data matrix Y has a separable covariance structure if

$$\mathrm{Cov}(\mathrm{vec}(Y)) = B \otimes A, \tag{4.22}$$

where B is a deterministic $L \times L$ matrix that defines the theoretical covariance between the columns of Y, and A is an $N \times N$ matrix for the theoretical covariance between the rows of Y. In our case, the SNPs are independent, so B is a diagonal matrix. A is the theoretical covariance between samples defined in equation 4.6. For the separable covariance assumption to hold, the covariance between samples at each SNP l should be the same for all l, and only scaled by the entries B_{ll} that are related to the variance of column l. More precisely, we need

$$\mathrm{Cov}(Y_{il}, Y_{jl}) = \eta_l G_{ij}, \tag{4.23}$$

where $\eta_l = B_{ll}$ is the lth diagonal entry of B. Another way of viewing this assumption is that the covariance between two samples at a site segregating at frequency p only depends on p through the multiplicative factor $\eta(p)$, and that $G(p)$ is identical for all frequencies p. Admittedly, this does not hold for a general demography under the coalescent process. We acknowledge that the separability of the covariance is an approximation, and show that employing the Ding & Yang framework that relies on this approximation still yields useful approximations for the cosine similarity between sample PCs and theoretical PCs.

Mapping to the Ding & Yang framework

The separable covariance assumption allows the mean-centered and rescaled genotype matrix Y (in the form $Y/\sqrt{L}$, so that $\hat{G} = (Y/\sqrt{L})(Y/\sqrt{L})^T$) to be factored as

$$\frac{Y}{\sqrt{L}} = A^{1/2} W B^{1/2}, \tag{4.24}$$

matching the Ding & Yang model, where W is a random noise matrix with $E[W_{ij}] = 0$ and $E[W_{ij}^2] = \frac{1}{L}$. We show in Appendix 4B that a consistent choice for A and B is:

$$A = -\frac{HTH}{\bar{T}} = G \tag{4.25}$$

$$B = \text{diag}(\eta_l), \tag{4.26}$$

where $\eta_l = \frac{h_l}{\bar{h}}$, $h_l = p_l(1 - p_l)$ and $\bar{h} = \frac{1}{L}\sum_l h_l$.

All quantities of interest, such as the edge of the bulk of the spectrum, the critical value at which a spike separates from the bulk, and the theoretical cosine similarity between a PC and the true axis of variation can be written in terms of the following functions

$$\Phi_1(\nu) = \sum_a \phi_a \frac{1 + \epsilon_a}{\nu - \epsilon_a}, \qquad \Phi_2(\nu) = \sum_a \phi_a \left(\frac{1 + \epsilon_a}{\nu - \epsilon_a}\right)^2 \tag{4.27}$$

$$\Psi_1(w) = \frac{1}{L}\sum_l \frac{\eta_l}{w - \eta_l}, \qquad \Psi_2(w) = \frac{1}{L}\sum_l \left(\frac{\eta_l}{w - \eta_l}\right)^2. \tag{4.28}$$

These expressions are related to $m_{1c}(z)$ and $m_{2c}(z)$ defined by Ding & Yang, but re-expressed more clearly in terms of the bulk eigenvalues of A and B. Specifically, they are expressed in terms of quantities that specify the two sources of heteroskedasticity: ϵ_a and η_l. The pair of coordinates (ν, w) are coupled through the equation

$$\Psi_1(w) = \gamma\ \Phi_1(\nu). \tag{4.29}$$

In Appendix 4C, we show how the quantities defined by Ding & Yang can be re-expressed in that manner. With these definitions, the critical point ν_c is given by solving the equation

$$\frac{\Phi_2(\nu_c)}{\Phi_1(\nu_c)} \cdot \frac{\Psi_2(w_c)}{\Psi_1(w_c)} = 1, \tag{4.30}$$

which, jointly with (4.29), can be solved to find (ν_c, w_c). A structure eigenvalue ν_k of $\hat{F}$

will separate from the bulk only if $\nu_k > \nu_c$. If $\nu_k < \nu_c$, the corresponding sample eigenvalue $\hat{\nu}_k$ will be absorbed into the bulk, and the sample eigenvector $\hat{u}_k$ will be asymptotically orthogonal to the true eigenvector u_k. If ν_k separates from the bulk, the cosine similarity between u_k and $\hat{u}_k$ will be given by

$$\langle u_k, \hat{u}_k \rangle^2 = 1 - \frac{\big(\Phi_1(\nu_k) + \Phi_2(\nu_k)\big)\Psi_2(w_k)}{\Phi_1(\nu_k)\big(\Psi_1(w_k) + \Psi_2(w_k)\big)}, \tag{4.31}$$

where w_k is found by solving the equation that couples ν and w (4.29).

In Appendix 4C, we also show that expanding the ratio $\dfrac{\Psi_2(w)}{\Psi_1(w)}$ in powers of $\dfrac{1}{w}$ leads to convenient approximations to these expressions. Specifically, these approximations allow us to define an effective aspect ratio $\gamma_{\text{eff}}(\nu)$, where the point of phase transition and the expression for squared cosine similarity reduce to those of the generalized spiked covariance model described by Bai and Yao (2012), but with γ_{eff} in place of γ. Defining $\kappa_2 = \langle \eta_l^2 \rangle$ and $\kappa_3 = \langle \eta_l^3 \rangle$, with $\delta = \kappa_3 - \kappa_2^2$

$$\gamma_{\text{eff}} := \gamma\kappa_2 \left[1 + \frac{2\delta}{\kappa_2}\gamma\Phi_1(\nu)\right]. \tag{4.32}$$

Expressions for ν_c and $\langle u_k, \hat{u}_k \rangle^2$ are expressed more simply as

$$\gamma_{\text{eff}}(\nu_c)\,\Phi_2(\nu_c) = 1 \qquad \text{(threshold)}. \tag{4.33}$$

$$\langle u_k, \hat{u}_k \rangle^2 = \frac{1 - \gamma_{\text{eff}}(\nu_k)\,\Phi_2(\nu_k)}{1 + \gamma_{\text{eff}}(\nu_k)\,\Phi_1(\nu_k)} \qquad \text{(cosine similarity)}, \tag{4.34}$$

the same as the expressions derived for the generalized spiked covariance model (Bai & Yao, 2012). The critical value is found by solving the first equation, which is much simpler to solve than jointly solving (4.29) and (4.30). The cosine similarity is found by plugging the computed ν_k directly into the second equation.

When all populations have the same diversity ($\epsilon_a = 0$ for all a), both equations reduce

to an even simpler form, since $\Phi_1(\nu) = 1/\nu$ and $\Phi_2(\nu) = 1/\nu^2$. In fact, if the second term of γ_{eff} is neglected, which is valid whenever $\gamma \ll 1$, the threshold reduces to the classical threshold $\nu_c = \sqrt{\gamma_{\text{eff}}}$, and the cosine similarity reduces to the classical expression in (4.2). In that case, $\gamma_{\text{eff}} = \dfrac{N}{L_{\text{eff}}}$, where

$$L_{\text{eff}} = \frac{(\sum_l h_l)^2}{\sum_l h_l^2}. \tag{4.35}$$

Despite being an approximation, this definition of L_{eff} is quite instructive. It shows that having SNPs with unequal variances reduces the effective number of SNPs, even when the SNPs are independent. Arriving at this expression for L_{eff} is reassuring, as it matches expressions derived for similar settings. It matches the Kish (1965) effective sample size applied to per-locus variances, and coincides with the effective number of markers of Patterson et al. (2006) when loci are unlinked, and with M_e of Goddard (2009) in the absence of LD. This reduction in the effective number of SNPs pushes the critical value of detectability upward. If, on the other hand, the columns of the genotype matrix were normalized to unit variance, $\gamma = \dfrac{N}{L}$ will hold exactly for the true number of independent SNPs, fixing ν_c at $\sqrt{\gamma}$, but each structure eigenvalue ν_k will be pushed downward in a manner that is difficult to analytically quantify.

The critical point of detectability ν_c is a useful quantity, but the quantity we are more interested in is the cosine similarity, which quantifies how well the observed PC matches the true eigenvector. It would be useful to write down the strength of a mode of structure that is required for the cosine similarity to exceed some value ρ. If we use the approximate expression of the cosine similarity in terms of L_{eff}, we would like to find ν_ρ that solves

$$\rho^2 = \frac{1 - \gamma_{\text{eff}}/\nu_\rho^2}{1 + \gamma_{\text{eff}}/\nu_\rho}, \tag{4.36}$$

which has a closed-form solution in terms of the quadratic formula. However, when $\gamma_{\text{eff}}/\nu_\rho \ll$

1 (i.e. $\nu_\rho \gg \gamma_{\text{eff}}$), the denominator is approximately 1, and ν_ρ can be expressed in a more insightful manner

$$\boxed{\nu_\rho = \sqrt{\frac{N}{L_{\text{eff}}(1-\rho^2)}}} \tag{4.37}$$

With the ingredients developed so far, we can describe a procedure for computing the critical value for detecting structure, as well as the expected cosine similarity as a function of a demographic parameter of interest, for a general demographic scenario. More importantly, we can write down the value that a demographic parameter that controls the strength of population structure has to exceed for the observed PC to align at some degree ρ with the true axis of variation, using equation 4.37. Given N individuals sampled at L independent SNPs from K populations, with a fraction ϕ_a sampled from each population:

1. Construct the matrix $\hat{F}$ (equation 4.10) and compute its $K-1$ eigenvalues that define the strength of each axis of structure. These eigenvalues ν_k are expressed in terms of demographic parameters θ and sampling parameters N and ϕ. In the process, compute the excess diversity ϵ_a of each population.

2. Find ν_c that solves equation 4.33. Express the condition $\nu_k(N, \vec{\theta}, \vec{\phi}) > \nu_c(N, L, \vec{\epsilon}, \vec{\phi})$ as a condition on a single parameter of interest. Alternatively, use the "crude" approximation of $\gamma_{\text{eff}} = \dfrac{N}{L_{\text{eff}}}$, and define $\nu_c = \sqrt{\gamma_{\text{eff}}}$

3. Write the cosine similarity in (4.34) as a function of $(N, L, \vec{\theta}, \vec{\epsilon}, \vec{\phi})$. Invert the expression to find the value of some demographic parameter at which the cosine similarity exceeds ρ. Alternatively, use the "crude" approximate expression (4.37) for a more explicit relationship.

4.3.3 Empirical identification of critical points from simulated data

With a recipe to compute the critical value of a demographic parameter below which the PC will be pure noise, and another point at which the PC will be almost aligned with the true axis of structure (cosine similarity = ρ, with $\rho = 0.9$ or equivalently $\rho^2 \approx 0.8$, for example), we would like to validate these theoretical predictions against simulations. First, we describe how genotype matrices are obtained using coalescent simulations, then we describe a method to estimate empirical values for the two points of interest by running PCA on random subsets of the simulated genotype matrices and fitting a curve through the observed cosine similarities that allows these two points to be identified.

Simulations

For a given demographic history, we simulate independent coalescent trees with a number of samples larger than the maximum tested sample size per population using `msprime` (Baumdicker et al., 2022; Kelleher et al., 2016). A mutation is dropped on a tree with probability $1 - e^{-\mu T_{\text{tot}}}$, where μ is the mutation rate, and T_{tot} is the total branch length of the simulated tree. We choose a default value of $\mu = 5 \times 10^{-7}$, higher than the human rate, since otherwise most of the simulated trees will be dropped, causing the simulations to be prohibitively long. If a tree is selected, a random branch on the tree is mutated with probability proportional to the length of that branch. All samples descending from that branch are assigned the derived allele. This is repeated until a large number of independent SNPs are obtained. Typically, we simulate a genotype matrix with a large number of samples and a large number of SNPs that can be later sub-sampled, for 100 log-spaced values of the demographic parameter of interest.

Given these simulations, a sampling scheme is specified. The sampling scheme dictates the total number of haplotypes N (rows) and independent polymorphic sites L (columns)

to sample, with a fraction ϕ_a of the haplotypes sampled from population a. A subsample of the full genotype matrix is obtained given the sampling scheme. PCA is performed on the subsample, and the cosine similarity between a given PC and the corresponding true axis of structure is computed. For each simulation indexed by a demographic parameter, this is repeated n_{rep} times to obtain n_{rep} observations of the cosine similarity at this value of the parameter (e.g. 100 different migration rates in a spatially structured population).

Fitting empirical cosine similarity curves

For a given demography and sampling scheme, we have $n_{\text{rep}} \times 100$ observations of the cosine similarity. The goal is to obtain an empirical fit of the cosine similarity as a function of the varied demographic parameter. From this fitted curve, we can read off an empirical estimate of the phase transition point (when the curve starts increasing) and the point where the curve crosses a certain level of similarity (default: $\rho^2 = 0.8$).

For an observation i, denote the value of the demographic parameter that is varied as x_i and the observed cosine similarity by y_i, with $t_i = \log x_i$. Given the set of observations $\{t_i, y_i\}$, we fit the following function

$$y(t) = \begin{cases} c, & t < t_0 \\ c \ + \ \sum_j \beta_j I_j(t - t_0), & t > t_0 \end{cases} \tag{4.38}$$

where I_j are I-spline basis functions (Ramsay, 1988). I-splines are monotonically increasing splines defined as the integrals of M-spline (non-negative spline) basis functions, normalized to have a range between zero and one. They provide a convenient basis for a non-parametric fit of the cosine similarity above the transition point, which is expected to increase monotonically after the phase transition point. The parameters to be fitted are the coefficients β_j of the spline basis functions, the phase transition point t_0, and the minimum cosine similarity c,

which is not exactly zero due to finite N (the cosine similarity below the threshold is exactly zero only in the infinite N limit). We observed that the variance of y_i steadily decreases as the demographic parameter x_i increases (for a certain choice of a demographic parameter that correlates with the strength of structure). Therefore, a weight $w_i = \frac{1}{\hat{\sigma}_i^2}$ is assigned to each observation. To estimate the variance of a given y_i, observations are divided into 20 quantile bins, and the sample variance $\hat{\sigma}_i^2$ of observations within each bin is computed. All observations within the same bin have the same variance and are therefore assigned the same weight. Variances are clipped from the bottom by $0.05 \times \text{median}(\hat{\sigma}_i^2)$ to avoid very large weights that dominate the fits.

For a given value of t_0, estimates for $\{\hat{c}(t_0), \hat{\beta}(t_0)\}$ are computed as

$$\hat{c}(t_0) = \frac{\sum_{i:t_i<t_0} w_i y_i}{\sum_{i:t_i<t_0} w_i}, \tag{4.39}$$

$$\begin{aligned}\hat{\beta}(t_0) &= \underset{\beta\geq 0}{\text{argmin}} \sum_{i:t_i>t_0} w_i(y_i - \hat{c} - \sum_{j=1}^{p} \beta_j I_j(t_i - t_0))^2 + \lambda_{\text{pen}} s^2 \sum_{j=1}^{p-2} (\beta_j - 2\beta_{j+1} + \beta_{j+2})^2 \\ &= \underset{\beta\geq 0}{\text{argmin}} \, ||W_R^{1/2} (y_R - (\hat{c} + \mathbf{S}\beta))||^2 + \lambda_{\text{pen}} s^2 ||D_2 \beta||^2. \end{aligned} \tag{4.40}$$

The subscript R in W_R and y_R denotes that only observations where $t_i > t_0$ are used to fit $\hat{\beta}$. The additional penalty term turns out to be essential for the identification of t_0. A fully flexible arm at $t > t_0$ can essentially fit equally well to any given t_0, since it can curve sharply around the point of phase transition to reproduce the elbow that t_0 is meant to locate. A default penalty strength of $\lambda_{\text{pen}} = 16$ is used, and $s^2 = \frac{||W^{1/2}\mathbf{S}||_F^2}{n_R}$ tunes the penalty strength to the magnitude of the weighted basis matrix $\mathbf{S}$ (whose columns are $I_j(t_i - t_0)$).

A grid search is performed over a set of candidate t_0 values to select t_0. A separate set of parameters $\{\hat{c}, \hat{\beta}\}$ is estimated for each candidate t_0, and the t_0 with minimum residual

sum of squares is selected

$$\hat{t}_0 = \underset{t_0}{\operatorname{argmin}} \sum_{i:t_i<t_0} w_i(y_i - \hat{c}(t_0))^2 + \sum_{i:t_i>t_0} w_i(y_i - \hat{c} - \sum_{j=1}^{p} \hat{\beta}_j(t_0) I_j(t_i - t_0))^2. \quad (4.41)$$

Finally, point-wise uncertainty of the spline fit is quantified by a non-parametric case-resampling bootstrap (Efron, 1979), where observations are resampled with replacement for each bootstrap replicate, and the entire model is refit to obtain a set of estimates $\{\hat{c}, \hat{\beta}, \hat{t}_0\}$. The point-wise percentile confidence bands for the fitted curve and confidence intervals for the phase transition point t_0 were obtained from the empirical 16th/84th percentiles of the resulting bootstrap distributions (Efron & Tibshirani, 1994).

4.4 Results

Consider a demographic model parametrized by some demographic parameter θ that specifies the strength of population structure, and a general sampling scheme defined by the number of sampled haplotypes N, the number of independent SNPs L and the proportion ϕ_a of haplotypes sampled from each subpopulation a. We are interested in finding expressions for $\theta_\rho^{(k)}$, which define the value of θ required for the cosine similarity between the kth PC and the corresponding true eigenvector to exceed some value ρ. $\theta_0^{(k)}$ is the detectability threshold for the kth PC, where the cosine similarity between the kth PC and the true eigenvector will be asymptotically zero for any demography with $\theta < \theta_0^{(k)}$. We derive the theoretical detection threshold and the point at which cosine similarity exceeds some value ρ for a variety of relevant demographic scenarios with both discrete structure and continuous (spatial) population structure, and compare the theoretically derived values to empirically obtained values, demonstrating the utility of our theoretical framework in identifying these points of interest.

4.4.1 RMT phenomena reproduced with an effective number of loci L_{eff}

The asymptotic properties of sample eigenvalues and eigenvectors in the classical RMT literature were established under the assumption that the columns of the data matrix are drawn from a multi-variate Gaussian distribution. Later work showed that these properties are "universal" and should extend to random matrices described by more general distributions (Johnstone & Paul, 2018). In the case of genotype matrices with columns representing SNPs, columns are Bernoulli/Binomial draws. Additionally, the classical results were established for the case where the random noise matrix has columns with unit variance. We explore how the three RMT phenomena described in section 4.2.1 (eigenvalue spread, BBP phase transition, and eigenvector inconsistency) extend to the case of genotype matrices whose columns have unequal variances. In figure 4.1, we compute the sample eigenvalues and eigenvectors resulting from a genotype matrix simulated under a simple two-population split demography, with $n = 200$ haplotypes sampled from each subpopulation. This demography results in a theoretical covariance matrix with a single spike, whose bulk eigenvalues are all identical and close to 1. Figure 4.1A demonstrates the "eigenvalue spread" phenomenon, where the bulk eigenvalues of the sample covariance matrix are distributed around their true value of 1 according to a Marchenko–Pastur (MP) distribution. As described in section 4.3.2, the effect of not normalizing the columns of the genotype matrix to unit variance can be approximated by replacing the number of columns L by an effective number of SNPs (equation 4.35). Indeed, an MP distribution parametrized by L_{eff} matches the observed distribution of the bulk eigenvalues much better than one that uses the raw number of columns L. The effective number of SNPs linearly increases with the raw number of SNPs, with a slope that is higher for lower sample size N (figure 4.2). Using an aspect ratio defined by L_{eff} ($\gamma = \frac{N}{L_{\text{eff}}}$) yields a more accurate theoretical prediction for the BBP phase transition point, where the largest sample eigenvalue separates from the bulk (figure 4.1B), and the cosine similarity between the theoretical eigenvector and the PC begins to increase (figure 4.1C). The figure

also shows that the fit of the empirical cosine similarity (obtained by the method described in section 4.3.3) closely matches the theoretical expression for cosine similarity, which was computed as described in section 4.3.2 using L_{eff} rather than the raw number of SNPs L.

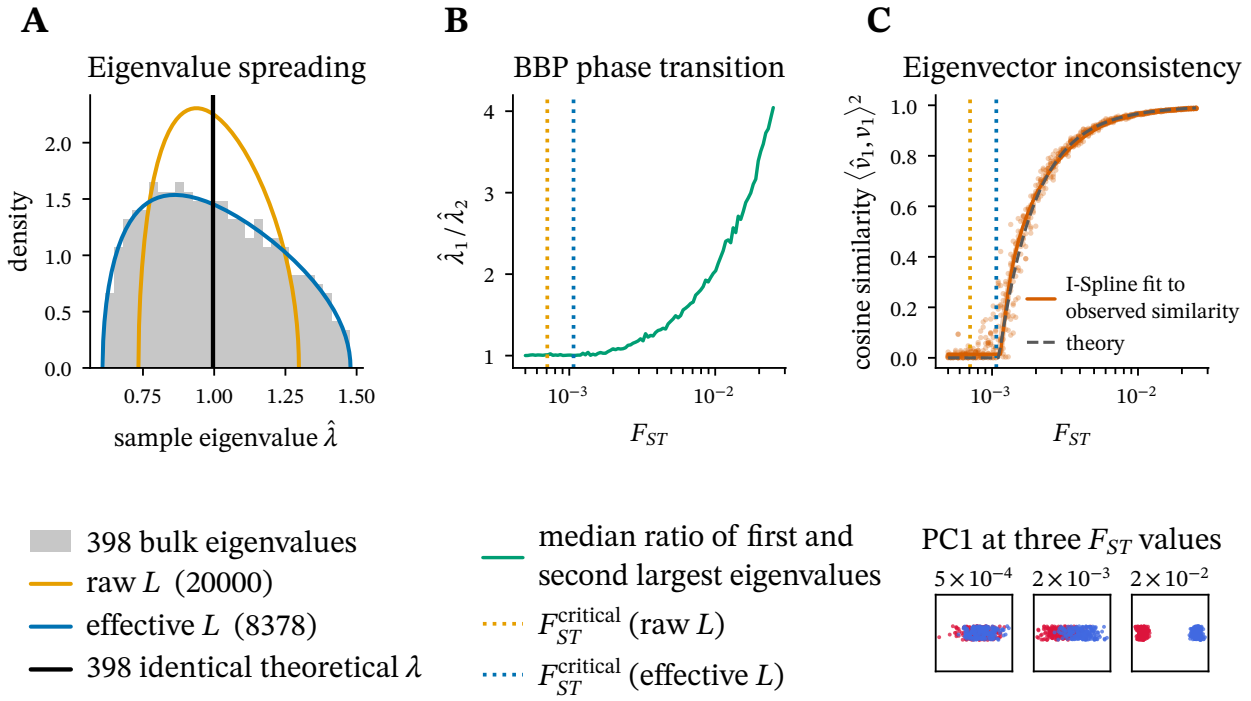

Figure 4.1: Properties of eigenvalues and eigenvectors of the sample covariance matrix of genotype matrices. A) The observed distribution of the bulk eigenvalues of a sample covariance matrix computed from a $400 \times 20,000$ genotype matrix with samples simulated under a two-population split scenario (excluding the largest eigenvalue and the zero eigenvalue). **B)** Median (over 8 replicates) of the ratio between the largest sample eigenvalue (the spike) and the second largest one (the edge of the bulk), computed over a range of F_{ST} values. The ratio stays close to 1 before the BBP threshold is crossed, and begins to increase after, tracking the true theoretical eigenvalue (which increases with F_{ST}). **C)** Cosine similarity between PC1 and the first theoretical eigenvector computed for 8 random samples of the genotype matrix (each of size $400 \times 20,000$) at each value of F_{ST}, showing an I-spline fit that closely matches the theoretical prediction from RMT (equation 4.34). The bottom 3 insets show PC1 for three different values of F_{ST}: Below the BBP threshold (where the PC is approximately pure noise), slightly above the threshold (where the samples form two very noisy clusters), and well above the threshold (where the samples form two well-defined clusters, matching the true underlying theoretical eigenvector).

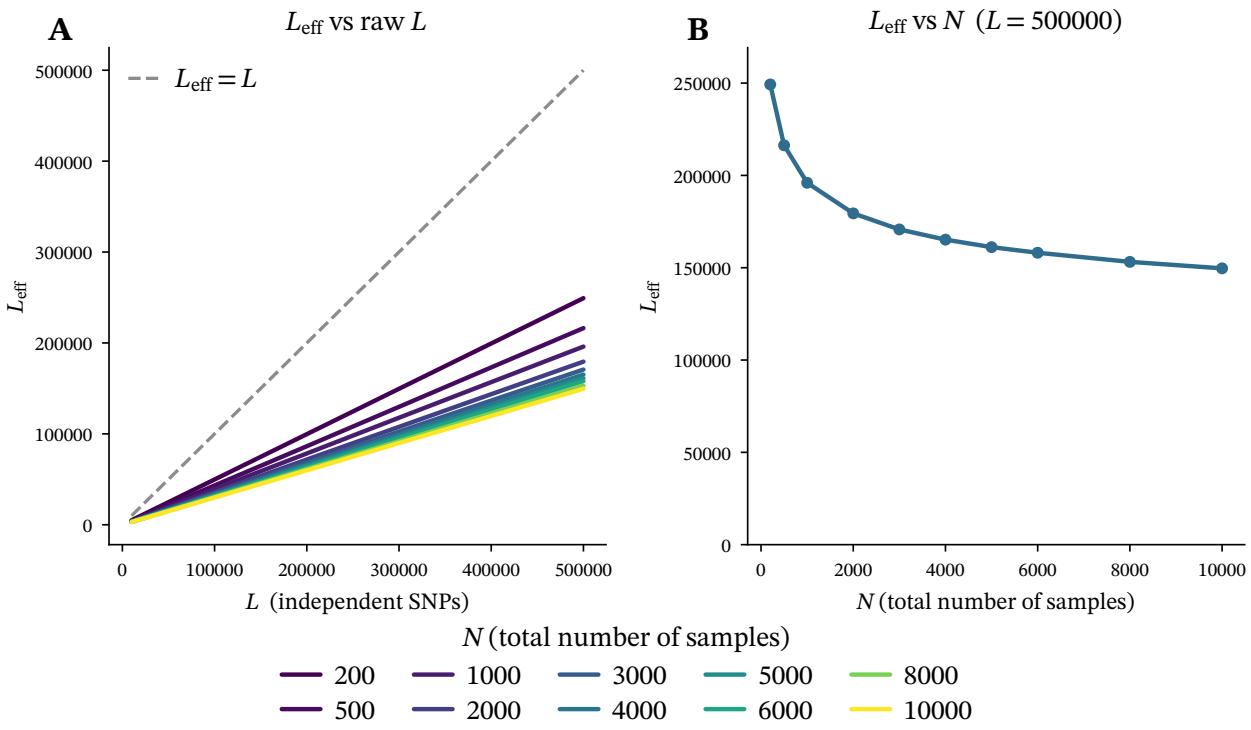

Figure 4.2: Effective number of loci in a genotype matrix with unnormalized columns. The genotype matrix is obtained from simulating a panmictic population. **A)** The effective number of columns (L_{eff}) of the unnormalized genotype matrix as a function of the true number of independent columns/SNPs (L). **B)** L_{eff} as a function of N for a fixed number of independent SNPs, showing a decrease that eventually plateaus. Larger samples include more rare polymorphic sites, which skews the heterozygosity spectrum and lowers L_{eff}.

4.4.2 Revisiting the two-population detectability conditions

Patterson et al. (2006) established the BBP phase transition point for detecting population structure in the case of two populations that diverged from each other recently, with G. McVean (2009) reproducing the result by deriving the eigenvalue explicitly in terms of coalescent times. We revisit this two-population example in a more general setting, where the two populations have unequal diversity (or unequal within-population coalescent times, $\tau_1 \neq \tau_2$), and are separated by some general F_{ST} defined by the mean coalescent time

between them T_{12}:

$$F_{ST} = 1 - \frac{\tau_1 + \tau_2}{2T_{12}}.$$

We additionally assume uneven sampling, with $N\phi_1$ samples obtained from population 1 and $N\phi_2$ samples from population 2. In Appendix 4D, we show that the strength of structure described by the excess in the largest theoretical eigenvalue is

$$\nu \approx N\tilde{\phi}F_{ST}(1 + \langle\epsilon\rangle) - (\phi_1 - \phi_2)(\epsilon_1 - \epsilon_2), \tag{4.42}$$

where $\tilde{\phi} = \dfrac{2\phi_1\phi_2}{\phi_1 + \phi_2}$ is the harmonic mean of the sample proportions, $\epsilon_a = \dfrac{\tau_a}{\bar{\tau}} - 1$ is the excess (or deficit) in the diversity of population a compared to the average diversity ($\bar{\tau} = \phi_1\tau_1 + \phi_2\tau_2$), and $\langle\epsilon\rangle = \dfrac{\epsilon_1 + \epsilon_2}{2}$ is the average of this excess (which is only non-zero under uneven sampling, since $\bar{\epsilon} = \phi_1\epsilon_1 + \phi_2\epsilon_2 = 0$ by definition).

This expression generalizes the result derived by G. McVean (2009) beyond the population-split case to any two-population demography (where $\tau_1 \neq \tau_2$). However, we argue that the simpler expression that ignores uneven diversity is the more relevant expression in the problem of detecting structure using PCA. Note that if τ_1 and τ_2 are vastly different, we would expect the F_{ST} to also be very large, as any demographic scenario that results in very different diversities between the populations can only be realized if the level of differentiation between the populations is also high. In that case, F_{ST} will be well beyond the detection threshold, which is not our regime of interest. In the small-F_{ST} regime where there is a risk that PCA will not reveal the underlying structure, it is safe to neglect the ϵ terms in equation 4.42. In that case, the level of differentiation (F_{ST}) required to yield a cosine similarity ρ between PC1 and the true eigenvectors is given by

$$\boxed{(F_{ST})_\rho \approx \frac{1}{\tilde{\phi}\sqrt{NL_{\text{eff}}(1-\rho^2)}}} \tag{4.43}$$

where $(F_{ST})_0$ is the critical point of detectability.

Two things are immediately obvious from this expression. First, unlike unequal diversity, uneven sampling can have a strong effect on detectability, where a population sampled $10\times$ more than the other reduces the effective F_{ST} to about a third of its even-sampling value, tripling the detectability threshold. Second, the point at which the PC reliably reflects the underlying structure lags behind the detectability threshold by an approximate factor of $\frac{1}{\sqrt{1-\rho^2}}$.

Figure 4.3 shows how the cosine similarity varies as a function of F_{ST} for genotype matrices simulated under a population split demography with the split time varied. We show the effect of varying the total number of samples N while holding the sample proportions ϕ constant, and the effect of varying the sample proportions while holding the total number of samples constant. In each case, our theoretical prediction for the point of phase transition $(F_{ST})_0$ and the point at which the cosine similarity exceeds $(0.9)^2 \approx 0.8$ (computed numerically using (4.33) and (4.34)) both closely match the empirically observed points derived from the I-spline fits. Moreover, the approximate solution for $(F_{ST})_\rho$ (equation 4.43) is quite close to the numerical solution, closely matching the empirical points as well. Therefore, equation 4.43 serves as an extension to the classical Patterson et al. (2006) result of $(F_{ST})_c = \frac{1}{\sqrt{NL}}$ (with N there counting diploid individuals), generalizing it to any two-population demography with low F_{ST}, highlighting more explicitly the effect of uneven sampling, and, most importantly, including the expected degree of alignment between the observed PC and the underlying axis of structure through the parameter ρ.

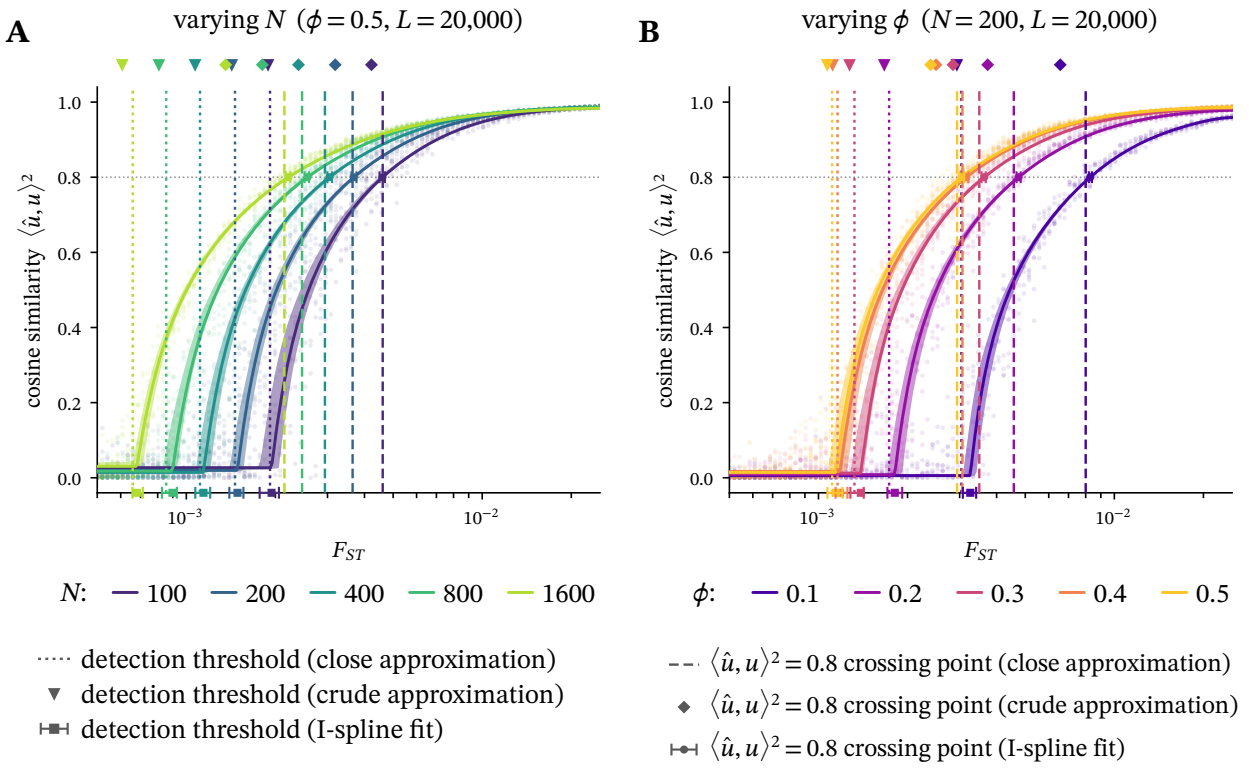

Figure 4.3: Cosine similarity as a function of F_{ST} for the population split demography. Genotype matrices were simulated under a population split demography, where the split time is varied, changing the F_{ST} between the populations at each simulation. At each value of F_{ST}, 8 random samples of the genotype matrix are obtained. The cosine similarity is computed between PC1 and the theoretical eigenvector for each of the 8 samples of the genotype matrix at each value of F_{ST} (shown by transparent scatter points). Vertical dotted and dashed lines show the theoretical $(F_{ST})_0$ and $(F_{ST})_{0.9}$ respectively, derived from an approximation that closely matches the exact solution (equations 4.33 and 4.34). The ticks above them show the values derived from a cruder approximation (equation 4.37 which leads to equation 4.43). $(F_{ST})_0$ and $(F_{ST})_{0.9}$ derived from the close approximation match the ones empirically derived from the I-spline fits well. $(F_{ST})_0$ derived from the crude approximation matches the ones derived from the close approximation, while $(F_{ST})_{0.9}$ is biased downwards. **A)** Varying the total number of sampled individuals, holding the sampling proportion fixed at 0.5 per population. **B)** Holding the total number of samples fixed at $N = 200$, and varying the proportion of samples ϕ obtained from population 1.

4.4.3 Generalizing the detectability conditions for an arbitrary demography with two closely related populations

The general two-population demography is a useful toy example to build a solid understanding of the factors that affect the detectability of structure by PCA, such as the strength of structure represented as F_{ST} and the degree of biased sampling represented as $\tilde{\phi}$. However, PCA is realistically applied to genotype samples obtained from multiple populations with an unknown underlying demographic history. The framework developed in section 4.3.2 would allow the detection thresholds to be numerically computed for any arbitrary demography and sampling scheme. However, we are interested in deriving analytical expressions for the thresholds that would still be relevant for samples obtained from populations with a complex unknown demographic history.

A realistic scenario that is analytically tractable is that of two closely related "sister populations" that exist within a more complex underlying demography. In particular, consider the setting where individuals are sampled from $K > 2$ populations, and two of those populations (for which we use indices 1 and 2) are very closely related, such that they are equally related to all other populations. Specifically, $T_{1c} \approx T_{2c}$ for all $c \notin \{1, 2\}$, and $\tau = T_{11} \approx T_{22}$. We simulate genotype matrices for such a setting by adding a recent population split in the Yoruban population (YRI) in the three-population out-of-Africa demography (Adrion et al., 2020b; R. N. Gutenkunst et al., 2009). Figure 4.4 shows an illustration of this demography, with four examples of PC3 (which should separate the YRI subpopulations) computed under two values of F_{ST} (between the two sister populations) and two levels of biased sampling. In Appendix 4D, we show that the excess eigenvalue associated with the axis of structure that separates the two sister populations is generally given by

$$\nu \approx \epsilon_0 + N\tilde{\phi}F_{ST}(1 + \epsilon_0), \tag{4.44}$$

where $\tilde{\phi}$ is again the harmonic mean of the sample proportions of the two sister populations, and $\epsilon_0 = \frac{\tau}{\bar{\tau}} - 1$ is the excess diversity of each sister population relative to the average diversity in the sample. This expression is exact when both sister populations are evenly sampled ($\phi_1 = \phi_2$), and accurate to second order in the sampling imbalance otherwise. We verify the expression in figure 4.5, showing that approximating it by the two-population version $N\tilde{\phi}F_{ST}$ induces a non-negligible bias. This expression indicates that the strength of structure relevant for detecting this axis of variation is scaled by a value of $1 + \epsilon_0 = \frac{\tau}{\bar{\tau}}$ and shifted by ϵ_0, which will strengthen the apparent structure if the sister populations are more diverse than the sample average, and weaken structure otherwise. Unlike the two-population demography, unequal diversity cannot be ignored in this case as it affects the strength of the spike as well as the bulk distribution. In the two-population case, the excess/deficit in diversity ϵ could be ignored because any non-negligible ϵ will have a correspondingly large F_{ST}, moving the problem to a regime where detectability is almost guaranteed. In the multi-population case, ϵ_0 of the two sister populations is dictated by the diversity of other populations as well, so it can be large even with a small F_{ST} between the sister populations.

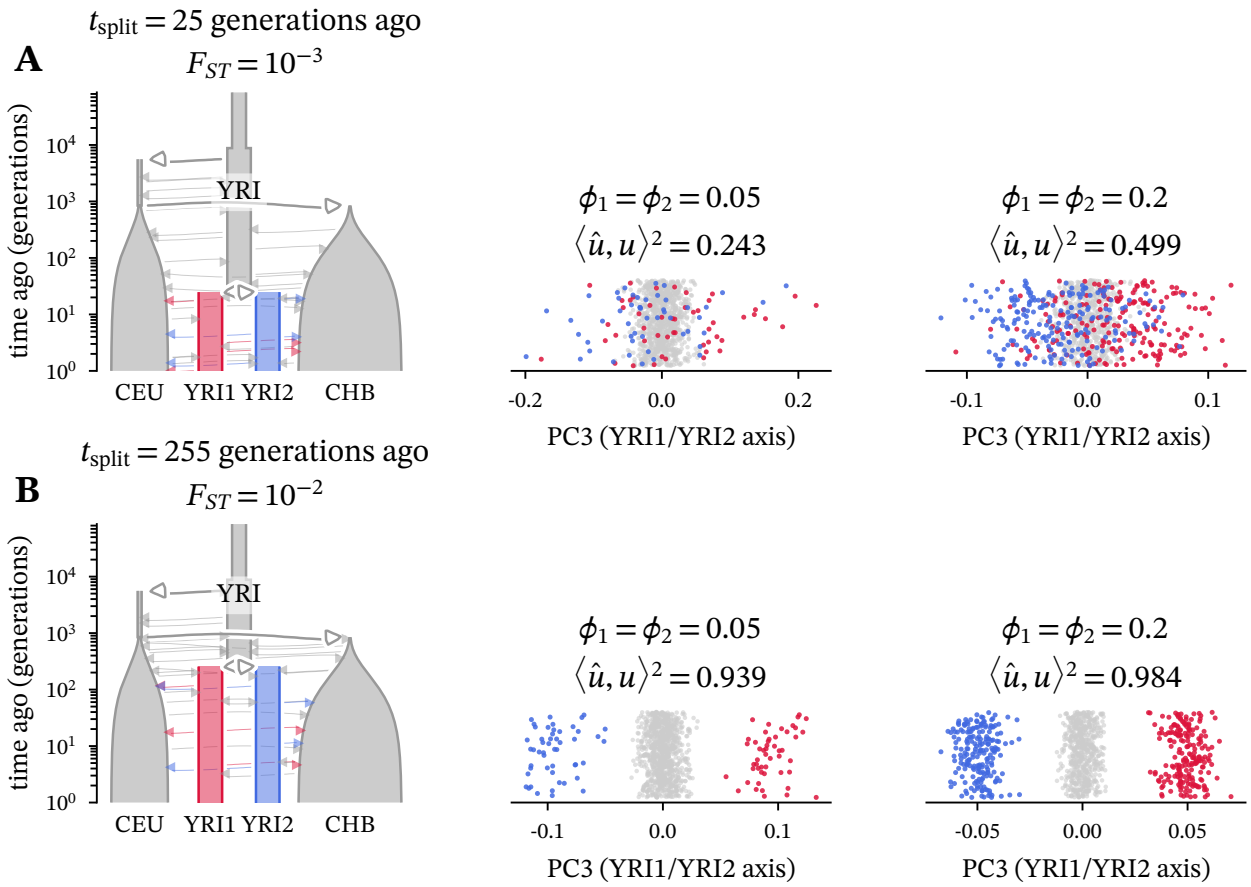

Figure 4.4: Detecting recent Yoruban fine structure using PCA under biased sampling: A recent population split in the YRI population is added to the three-population out-of-Africa demography, where the timing of the split is varied, yielding different values of F_{ST} between the two Yoruban subpopulations. **A)** For a recent population split that occurred 25 generations ago (or ≈ 625 years ago), PCA is unable to adequately separate the two YRI subpopulations if each of them represents only 5% of the total sample size. Raising the sampling representation to 20% (such that both populations together contribute 40% of the samples) causes the recent fine structure to be better resolved by PCA. **B)** A deeper split that occurred 255 generations ago (≈ 6375 years ago) is more easily detected by PCA, even under severely biased sampling.

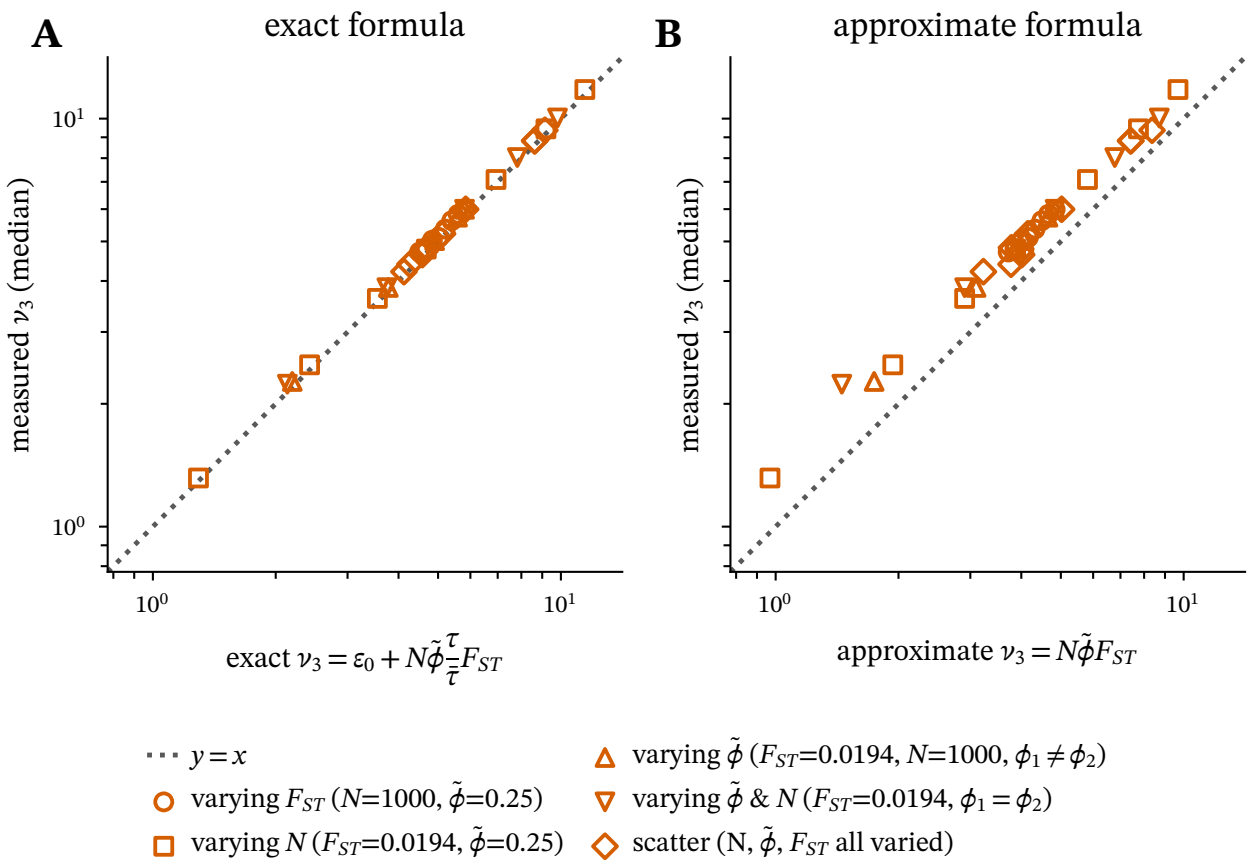

Figure 4.5: Verifying the sister-population eigenvalue expression under the recent Yoruban split scenario: Under the demography shown in figure 4.4, the third-largest eigenvalue corresponds to the strength of the axis of variation that separates the two recently split Yoruban populations. We computed this eigenvalue for sample covariance matrices computed from several genotype matrices simulated under a variety of split times and sampled under different sampling schemes. **A)** The empirically observed eigenvalues (median over 8 replicates) closely match the ones obtained from the general theoretically derived expression, with ϵ_0 computed by numerically calculating mean within-population coalescent times using `msprime.DemographyDebugger` helper functions. Note that the empirical eigenvalue itself is slightly biased upward based on RMT, but it serves as a valid reference point when the bias is negligible (strong structure). **B)** Empirically observed eigenvalues plotted against ones obtained from an approximation of the general theoretical expression, showing that the shift ϵ_0 is not negligible, and excluding it underestimates the eigenvalue.

In the same spirit as equation 4.43, we write an expression for the approximate $(F_{ST})_\rho$

as

$$(F_{ST})_\rho \approx \frac{1}{\tilde{\phi}(1+\epsilon_0)\sqrt{NL_{\text{eff}}(1-\rho^2)}} - \frac{\epsilon_0}{N\tilde{\phi}(1+\epsilon_0)}. \tag{4.45}$$

We test how well this approximation holds up by comparing it to numerically computed $(F_{ST})_\rho$ obtained from solving (4.33) and (4.34), and the values obtained from the I-spline fits to observed cosine similarity values. We conduct the same two experiments as the two-population case, where the total number of samples is varied with the sampling scheme held constant ($\phi_{\text{YRI},1} = \phi_{\text{YRI},2} = 0.2, \phi_{\text{CHB}} = \phi_{\text{CEU}} = 0.3$), and the sampling proportions for the two sister populations $\{\phi_{\text{YRI},1} = \phi_{\text{YRI},2}\}$ are varied with the total number of samples held constant. The numerical solutions to the threshold values of F_{ST} obtained from the close approximations match the empirically derived ones quite well. However, the cruder approximation (equation 4.45) seems to work only for the threshold $(F_{ST})_0$, and fails to accurately predict the point where the squared cosine similarity exceeds 0.8, even more severely than in the two-population case. The more general expression for cosine similarity that uses $\Phi_1(\nu)$ and $\Phi_2(\nu)$ (equation 4.34) seems to be more relevant in this case where diversity is unequal and sampling is uneven.

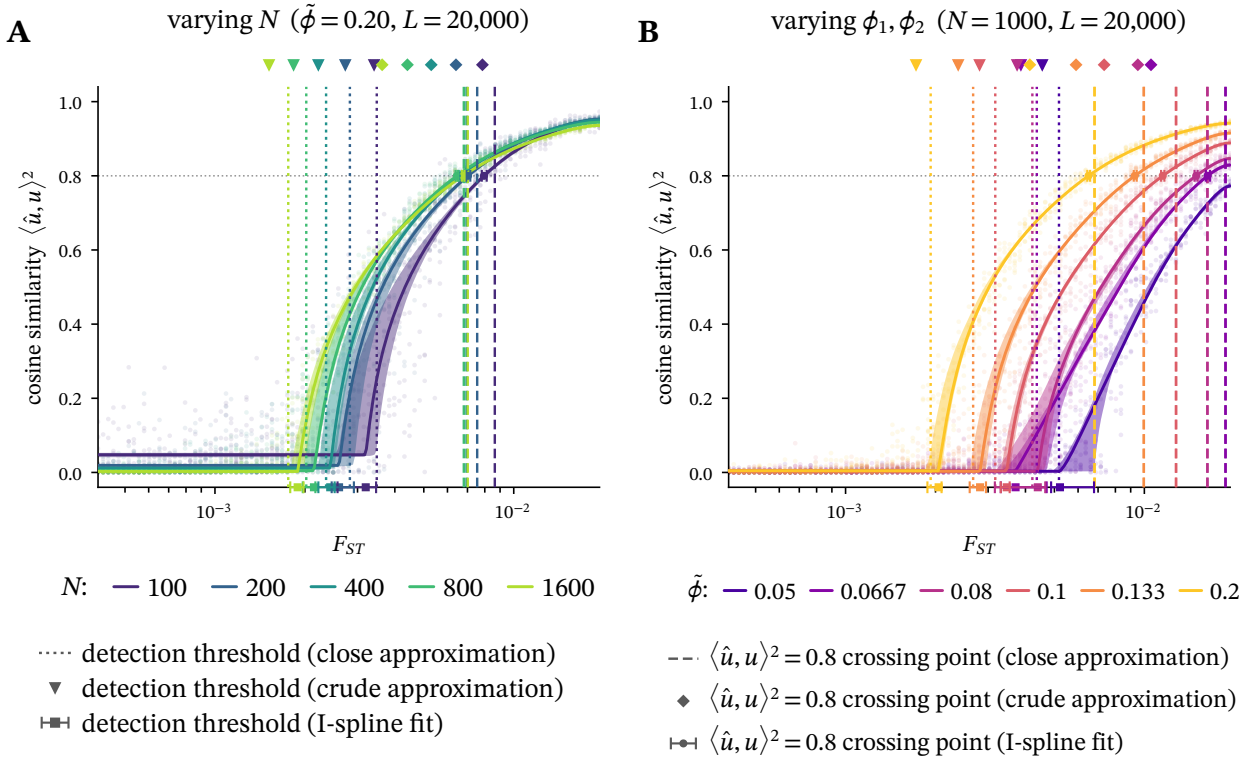

Figure 4.6: Cosine similarity as a function of F_{ST} for the demography with recent Yoruban structure. Cosine similarity is computed between PC3 and the theoretical eigenvector in the same manner as figure 4.3. The theoretical thresholds derived from the approximations that closely match the exact solution (equations 4.33 and 4.34) are close to the empirically derived threshold with a small bias. The detectability thresholds derived from the crude approximation (equation 4.45, $\rho = 0$) are close to the empirically derived ones, but the points of crossing squared cosine similarity of 0.8 are severely biased. **A)** Varying the total number of sampled individuals, holding the sampling proportion fixed at 0.2 for the two Yoruban populations and 0.3 for CEU and CHB. **B)** Holding the total number of samples fixed at $N = 1000$, and varying the proportion of samples ϕ obtained from each of the two Yoruban populations by choosing a pair of proportions from the set $\{0.05, 0.1, 0.2\}$ to assign to the two Yoruban populations. CEU and CHB are sampled evenly.

4.4.4 Deriving the conditions for which PCA will reveal spatial structure under homogeneous migration

One of the main appeals of using PCA in studying population structure is its ability to reveal continuous axes of variation, or clines, that result from continuous population structure, in

addition to clusters that result from discrete population structure. In particular, when applied to spatially structured populations, the first PC (if space is one-dimensional) or the first two PCs (if space is two-dimensional) were shown to correlate with the spatial location of samples under certain settings (Novembre & Stephens, 2008; Novembre et al., 2008). It would be interesting to study the conditions under which PCA is able to reveal such spatial structure as a function of the strength of this structure.

In section 4.3.1, we extended the procedure for computing the structure eigenvalues of the theoretical covariance matrix to a spatial setting, where the eigenvalues and eigenvectors could be derived for a general spatial demography given the Laplacian of the matrix of migration rates between demes, and an arbitrary sampling scheme. To obtain informative analytical expressions like those in equations (4.43) and (4.45), we consider the one-dimensional and two-dimensional stepping-stone models, where subpopulations (demes) are arranged on a chain in 1D or a grid in 2D, with a constant per-generation migration rate m between each pair of neighboring demes. We assume that each deme has the same population size N_e, and define a population-scaled migration rate $M = 4N_e m$. Figure 4.7 shows examples of PCA applied to genotype matrices simulated under 1D and 2D stepping-stone models with different migration rates. When migration rates are very high, PCs are almost pure noise, similar to the very low F_{ST} case, and the sample positions on the top PCs are not informative about spatial location. At intermediate migration rates, PCA can reveal large-scale spatial structure (such as north-south/east-west) but fails to reveal finer-scale spatial patterns. Only when the migration rate is below a particular point do the PCs accurately correlate with geographic location (or, more accurately, correlate with the underlying theoretical eigenvectors which themselves correlate with geographic location).

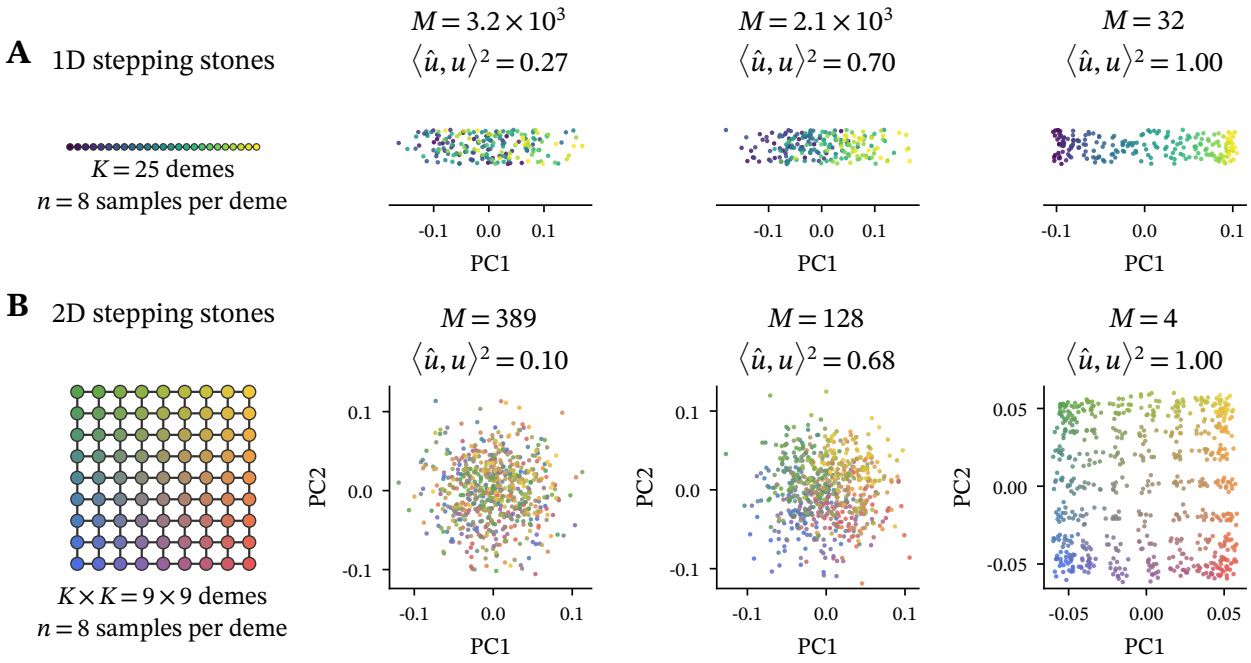

Figure 4.7: Detecting spatial structure using PCA under stepping-stone models. A) PC1 shown for genotype matrices simulated under the one-dimensional stepping-stone model with three different migration rates ($N_e = 4000$ for each of the 25 demes). As migration rate decreases, the signal for spatial structure is stronger, and PC1 resolves the linear spatial structure more accurately. **B)** PC1 vs PC2 shown for genotype matrices simulated under the two-dimensional stepping-stone model ($N_e = 1000$ for each of the 81 demes). Similar to the 1D case, spatial structure is more clearly resolved for lower migration rates. $L = 10,000$ independent loci are sampled in all cases.

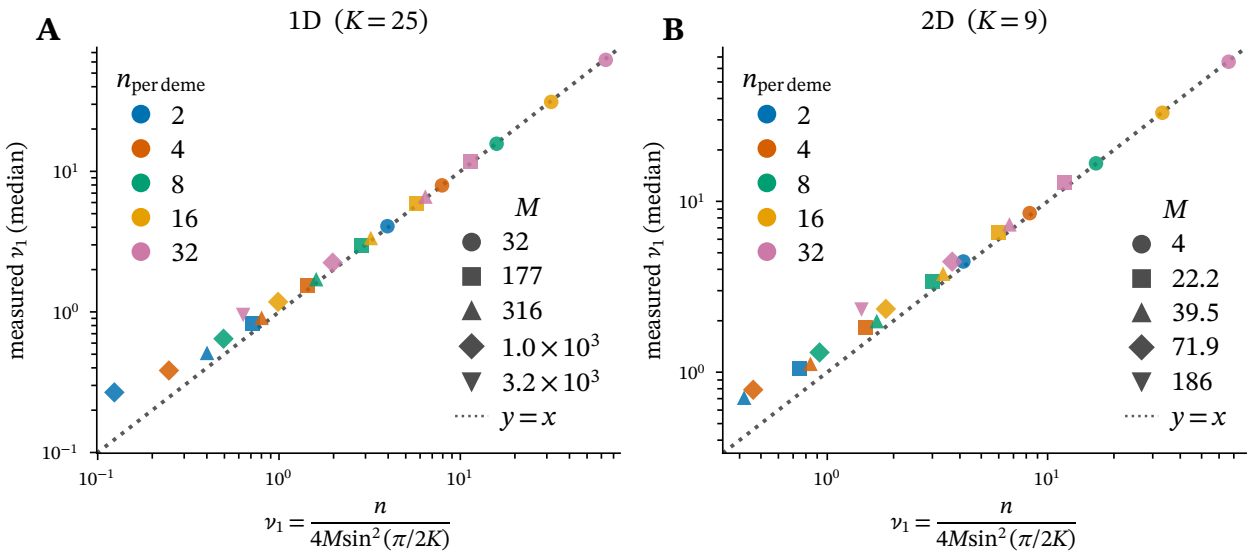

Figure 4.8: Verifying the eigenvalue expression for the 1D and 2D stepping-stone models: The largest eigenvalue of the sample covariance matrix was computed for a genotype matrix simulated under a given migration rate and a given number of samples per deme for the 1D **(A)** and 2D **(B)** stepping-stone models. The median of 8 replicates is plotted against the theoretically derived expression in each case. The slight discrepancy at smaller eigenvalues is because the bias associated with the sample eigenvalues predicted by RMT becomes non-negligible when the magnitude of the eigenvalue (strength of structure) is small.

In Appendix 4D, we derive analytical expressions for the excess in the eigenvalue corresponding to PC1 in 1D and PC1/PC2 in 2D (the first two eigenvalues are degenerate in 2D) under even sampling, where n haploid individuals are sampled from each deme. The expression is identical in the 1D and 2D case, and is given by

$$\nu = \frac{n}{4M \sin^2\left(\frac{\pi}{2K}\right)} \tag{4.46}$$

$$\approx \frac{nK^2}{\pi^2 M} \quad \text{(for large } K\text{)} \tag{4.47}$$

The only difference is that K represents the total number of demes in 1D, while K^2 represents the number of demes in 2D. We verify this expression using genotypes simulated under a handful of migration rates, with different numbers of samples per deme (figure 4.8). An

expression for the point at which the cosine similarity between the spatial PC and the theoretical eigenvector (which correlates with spatial location) exceeds ρ can thus be written in terms of M

$$M_\rho^{1D} \approx \frac{K\sqrt{NL_{\text{eff}}(1-\rho^2)}}{\pi^2} \tag{4.48}$$

$$M_\rho^{2D} \approx \frac{\sqrt{NL_{\text{eff}}(1-\rho^2)}}{\pi^2}, \tag{4.49}$$

where $N = nK$ in 1D and $N = nK^2$ in 2D, hence the extra factor of K in the 1D case. These expressions can also be written in terms of the pooled F_{ST}, whose relationship with M was derived in Appendix 4D for the 1D and 2D cases. Specifically,

$$\begin{aligned}(F_{ST})_\rho^{1D} &\approx \frac{\pi^2}{\pi^2 + 6\sqrt{NL_{\text{eff}}(1-\rho^2)}} \\ &\approx \frac{\pi^2}{6}\frac{1}{\sqrt{NL_{\text{eff}}(1-\rho^2)}}\end{aligned} \tag{4.50}$$

$$\begin{aligned}(F_{ST})_\rho^{2D} &\approx \frac{\ln K}{\ln K + \dfrac{2\sqrt{NL_{\text{eff}}(1-\rho^2)}}{\pi}} \\ &\approx \frac{\pi \ln K}{2}\frac{1}{\sqrt{NL_{\text{eff}}(1-\rho^2)}}.\end{aligned} \tag{4.51}$$

These expressions are quite similar to the expression derived for the two-population case (4.43), with an additional multiplicative factor that depends on the geography. The 2D expression uses the large-K asymptotic of the pooled F_{ST} and underestimates it by about a factor of 2 for the 9×9 grid simulated here (Appendix 4D). For a finite grid, the exact relationship between M and F_{ST} is obtained from equation (4.21).

Figure 4.9 shows that the theoretical point of phase transition and the point of crossing squared cosine similarity of 0.8 (derived from (4.33) and (4.34)) remarkably agree with the empirical points derived from the I-spline fits. In the 1D case, the corresponding values obtained from the crude approximation (4.48) agree quite well with the numerically derived

ones. However, in 2D, the agreement only holds for the detectability point and does not hold well for the point of crossing 0.8 squared similarity, which is a pattern that has been observed in other examples above (e.g. figure 4.6).

A complication is that we must be careful when measuring the cosine similarity for the 2D case, since the first two theoretical eigenvectors are not unique and fall into a two-dimensional subspace. We compute the cosine similarity by first Procrustes aligning the first two PCs with the first two theoretical eigenvectors. Then, we take the average of the squared cosine similarity computed between PC1 and the first theoretical eigenvector, and PC2 and the second theoretical eigenvector, and define this average as the cosine similarity for the 2D case.

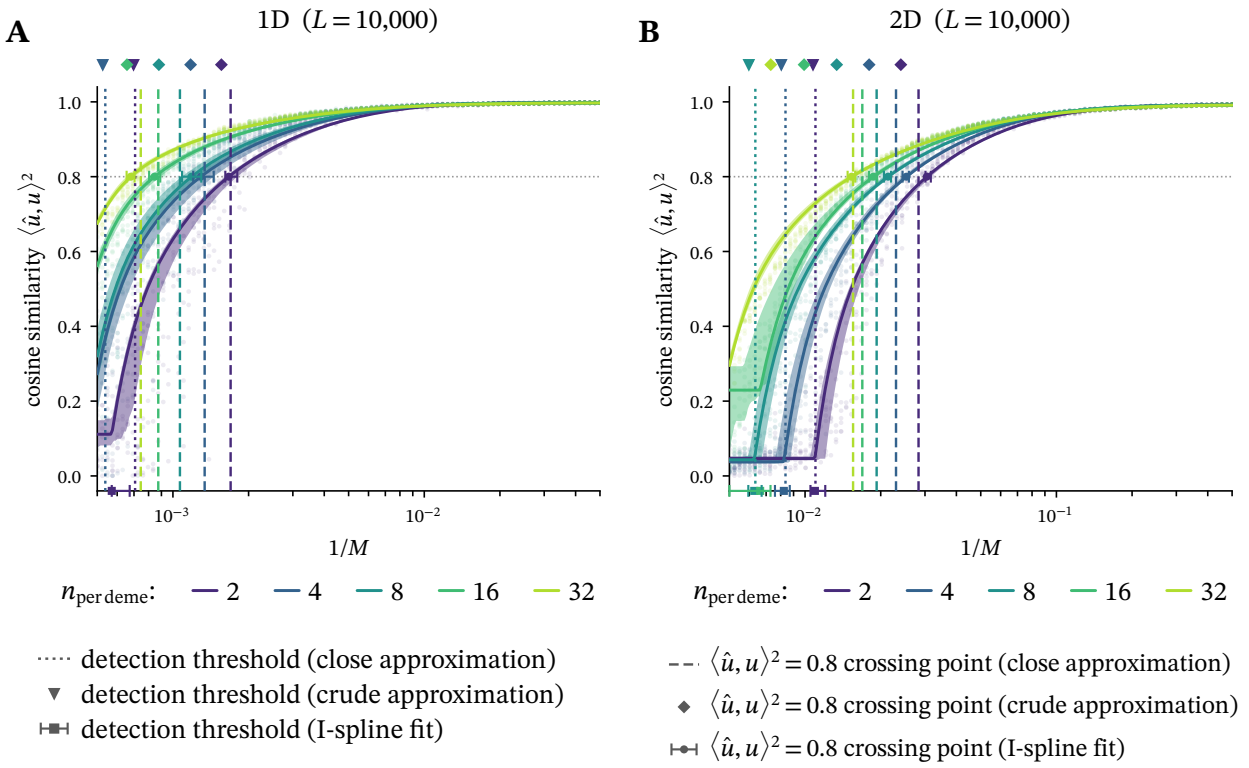

Figure 4.9: Cosine similarity as a function of $1/M$ for homogeneous stepping-stone models. Cosine similarity is computed against the first PC in the 1D case **(A)** and against the first two PCs (Procrustes aligned) in the 2D case **(B)**, where the average is taken. Observed cosine similarity is shown by transparent scatter points, with solid lines showing the best I-spline fit. The close approximation (4.33) and crude approximation (4.48,4.49) for the theoretical phase transition value, in terms of $1/M$, accurately match the empirical transition point in the 1D and 2D cases. For the point at which the squared similarity reaches $\rho^2 = 0.8$, both the close and crude approximations work well for the 1D case, while the crude approximation is biased in the 2D case. The close approximation (4.34) accurately matches the value derived from the I-spline in all cases.

4.4.5 Empirical investigation of the alignment of the top PCs to geography under more realistic sampling and historical scenarios

Deriving the detectability thresholds under the stepping-stone models with even sampling and homogeneous, equilibrium migration is insightful in itself for understanding how these can be expressed in terms of the migration rate or the pooled F_{ST} of the spatial population. However, these simple expressions are expected to change under more realistic migration

patterns and sampling scenarios. The framework of section 4.3.1 offers a promising avenue to study the spectral properties of spatially structured genotype matrices under such general scenarios, though closed-form expressions are not available in general.

Sparse sampling reduces the maximum realizable alignment of the top PCs with geography

Sampling every single deme in a population (or region in a habitat) equally is not a realistically feasible sampling scheme. Practically, a set of regions is chosen to be sampled, and this choice can be made by design or simply enforced by feasibility. Novembre and Stephens (2008) and G. McVean (2009) conducted limited simulations that show that uneven sampling can distort the PCs. We revisit this issue here by conducting further simulation experiments to investigate the effect of sparse sampling on recovering spatial structure using PCA. Rather than sampling n individuals from every deme, a fixed number of N individuals is sampled from a fraction ϕ of the demes in the habitat, where the sampled demes are chosen at random, and an equal number of individuals is sampled from each of them. We denote the theoretical eigenvector associated with a given sparse sampling scheme by u_ϕ, its observed counterpart (PC) by $\hat{u}_\phi$, and the theoretical eigenvector under dense sampling (which is a proxy for geographic location) by u_1 ($\phi = 1$). Figure 4.10 shows an illustration of a few representative sparse sampling schemes, with PCA applied to each scenario. When the total number of samples N is held constant, the main effect of such sparse sampling is distorting the theoretical eigenvector rather than affecting detectability. The illustration also makes a point that will recur below: in the 1D case, the alignment of PC1 with geography barely changes across the three sampling fractions, because the particular subsets of demes drawn for $\phi = 0.2$ and $\phi = 0.4$ happen to cover the habitat almost as evenly as the denser one.

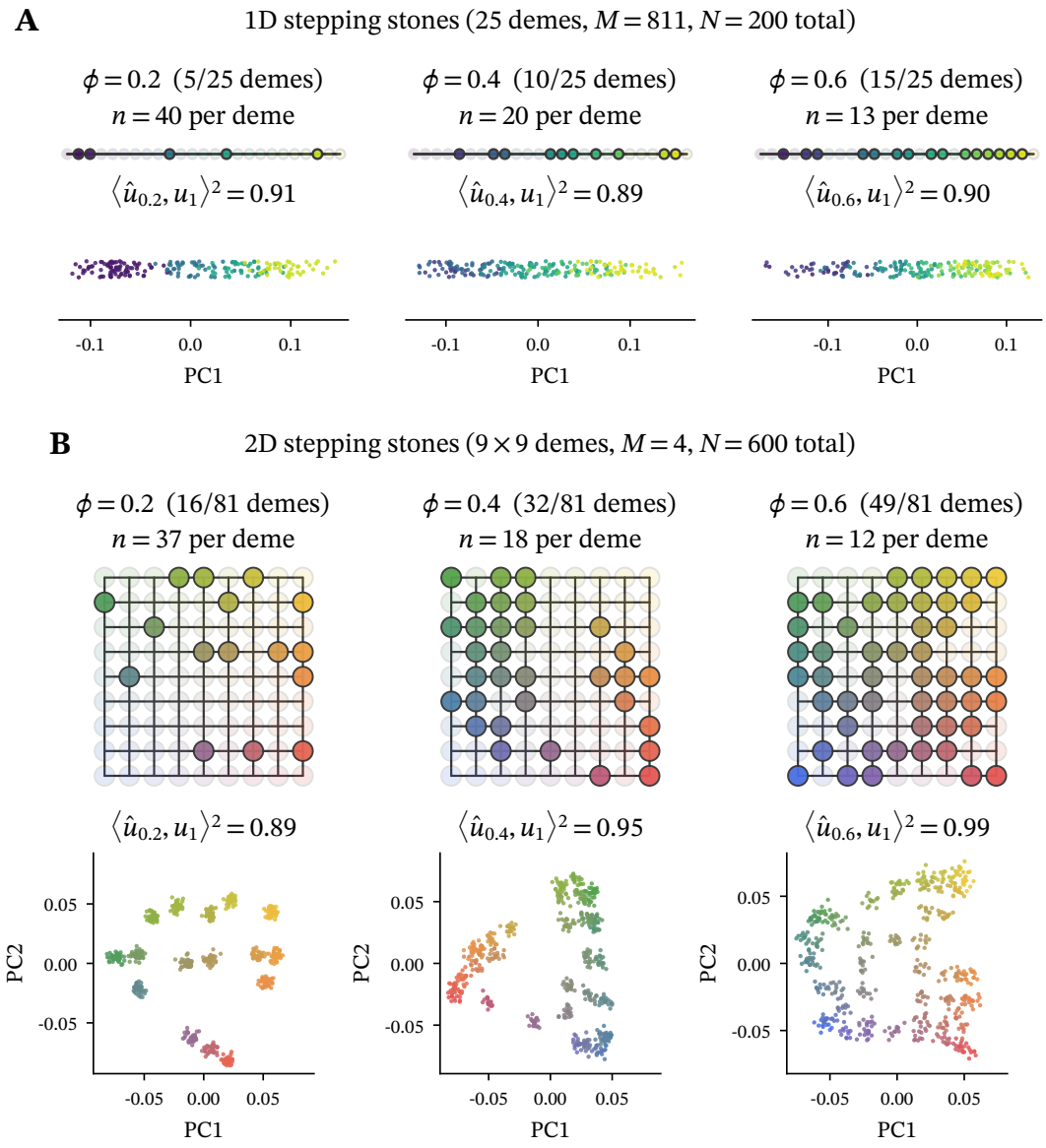

Figure 4.10: Detecting spatial structure using PCA under stepping-stone models with sparse sampling. Fixing a total number of samples N, a fraction $\phi \in \{0.2, 0.4, 0.6\}$ of the demes in the habitat is sampled, with an equal number of samples obtained from each deme. Each panel shows a single random draw of demes. The squared cosine similarity is computed between the observed PC ($\hat{u}_\phi$) and the theoretical PC associated with dense sampling (u_1), which aligns with geography. Migration rates ($M = 811$ in 1D, $M = 4$ in 2D) are chosen well above the respective detection thresholds so that differences between panels reflect distortion rather than detectability. **A)** 1D stepping-stone model. For the demes drawn here, sampling the habitat more densely does not improve the alignment of PC1 with geography, because the sparser subsets happen to cover the habitat evenly. **B)** 2D stepping-stone model. Sampling the habitat more densely brings the theoretical PCs under sparse sampling closer to those under dense sampling, so that PC1 vs. PC2 aligns better with geography.

We compute the squared cosine similarity between the observed PCs and the theoretical eigenvectors under sparse sampling for multiple random realizations of a given sampling fraction. Figure 4.11 shows that varying the sampling fraction mainly affects the maximum realizable similarity, since u_ϕ and u_1 become more aligned as ϕ approaches 1. To compute the theoretical expectation of this maximum realizable similarity, we compute u_ϕ for 16 random subsets of demes at each sampling fraction, Procrustes-align it to u_1 (PC1 in 1D, the PC1/PC2 subspace in 2D), compute the squared cosine similarity between the two, and average over subsets. In both 1D and 2D this ceiling rises monotonically with ϕ, from roughly 0.9 at $\phi = 0.2$ to 1 at $\phi = 1$. The detection thresholds, by contrast, move comparatively little. Averaged over random subsets of demes, the point at which the squared similarity crosses 0.8 shifts toward weaker migration as the fraction of sampled demes decreases, noticeably in 1D and only marginally in 2D. This result highlights that with N fixed, the dominant effect of sparse sampling is not on whether spatial structure is detected but on how faithfully the resulting PCs reflect geography: the PCs are "distorted" more as fewer demes are sampled, especially if the sampled demes are unevenly spaced, and therefore correlate less with spatial structure. The idea that uneven sampling distorts PCs is a well-known result (G. McVean, 2009; Novembre & Stephens, 2008), and our results demonstrate more quantitatively the effect of this distortion. Importantly, the degree of distortion varies with the random subset of demes chosen for sampling, as we observe that the lowest ceilings arise from subsets with large unsampled gaps. This is visible in figure 4.11 as a few individual realizations that plateau well below the average ceiling for their sampling fraction.

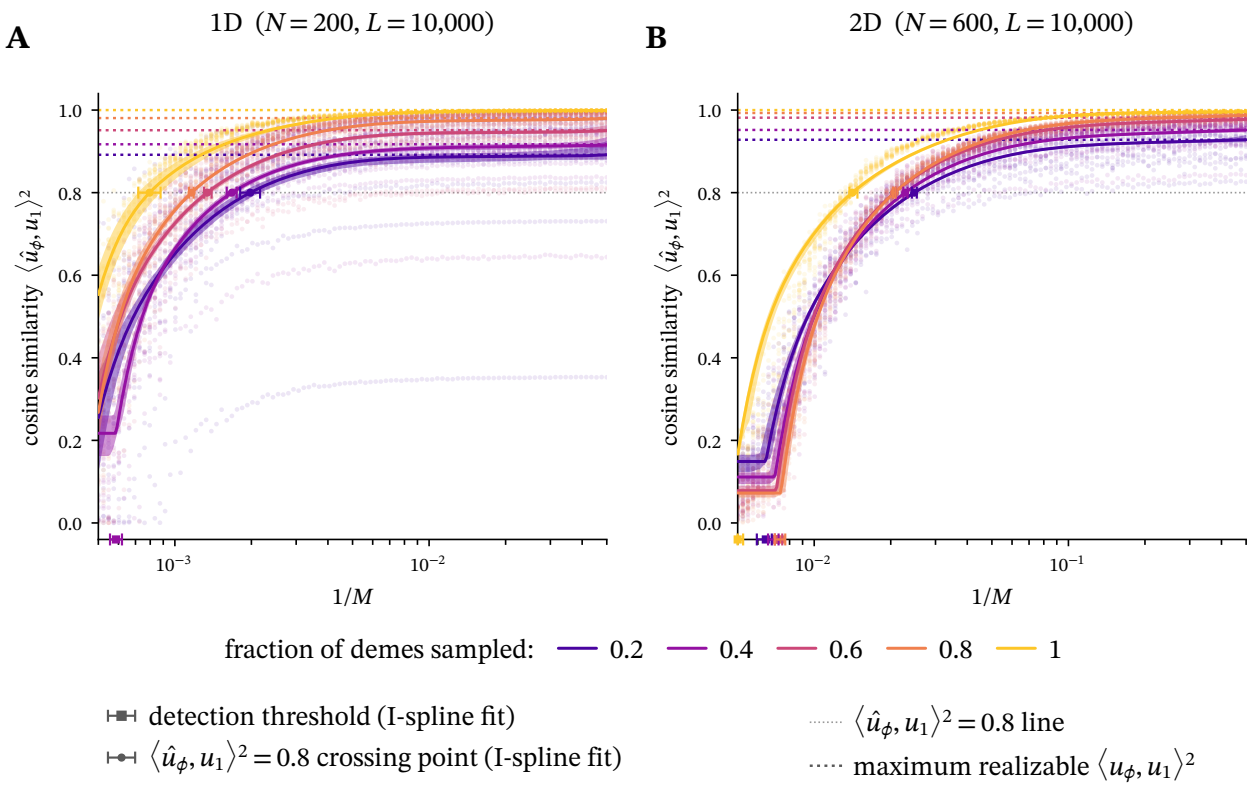

Figure 4.11: The effect of sparse sampling on distorting the spatial PCs in stepping-stone models. In both 1D **(A)** and 2D **(B)** stepping-stone models, fixing the total number of samples N and varying the fraction of demes sampled mainly affects the maximum realizable alignment between the top PCs and geography. The squared cosine similarity is measured between the observed PCs and the top theoretical PCs under dense sampling, which themselves align with geography. Dotted horizontal lines show the theoretical maximum realizable similarity $\langle u_\phi, u_1 \rangle^2$, averaged over 16 random subsets of demes for each sampling fraction. Individual realizations that plateau well below their dotted line correspond to subsets of demes with large unsampled gaps.

Recent spatial structure makes the alignment of the top PCs with geography non-monotonic in the migration rate

The stepping-stone results above assume that migration has been going on long enough for the spatial pattern of coalescent times to have equilibrated. We conduct a simulation experiment in which this is not the case. An ancestral panmictic population splits, $t = 500$ generations ago, into a panmictic population A and a spatially structured population B,

where the latter is arranged as a 9×9 grid of demes with homogeneous migration (figure 4.12A). Of N samples, a fraction ϕ is drawn from the spatially structured population B, evenly across its demes, while the rest is drawn from the panmictic population A. PC1 separates A from B, so we measure the squared cosine similarity between the observed PC2/PC3 subspace ($\hat{u}_{t=500}$) and the top two spatial modes under equilibrium migration ($(u_\infty^{(1)}, u_\infty^{(2)})$), which align with geography.

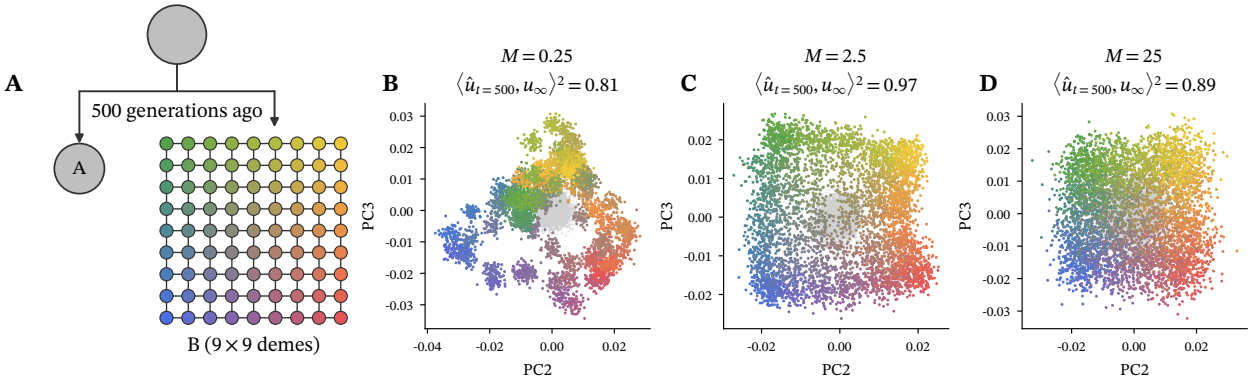

Figure 4.12: Detecting recent spatial structure using PCA. A) Demographic model. An ancestral panmictic population splits $t = 500$ generations ago into a panmictic population A and a population B structured as a 9×9 grid of demes with homogeneous stepping-stone migration of strength M. $N_e = 10^4$ for the ancestral population and population A, while $N_e = 124$ for each deme in the spatially structured population (such that $124 \times 81 = 10,044$ is close to the ancestral N_e). **(B–D)** PC2 vs. PC3 of the sample covariance matrix for $M = 0.25$, 2.5 and 25; PC1 (not shown) separates A from B. Samples from B are colored by the position of their deme on the grid and samples from A are shown in gray. Above each panel is the squared cosine similarity between the observed PC2/PC3 subspace ($\hat{u}_{t=500}$) and the theoretical spatial eigenvectors under equilibrium migration (u_∞). The alignment with geography is highest at the intermediate migration rate, where $M = 2.5$. At $M = 25$ spatial structure revealed by PCA is diffuse, while at $M = 0.25$ the samples separate into discrete clusters corresponding to individual demes but the grid is distorted.

Figure 4.13 shows that, unlike in the equilibrium case, this alignment is not monotonic in $1/M$: it increases with $1/M$ up to a maximum near $1/M \approx 0.4$ and then decreases as migration weakens further. Figure 4.12 illustrates the same behavior in three PCA plots: at $M = 25$ the spatial structure is faintly resolved, at $M = 2.5$ the samples form a smooth grid

aligned with geography, and at $M = 0.25$ they break into discrete clusters corresponding to individual demes, with the grid visibly distorted. At fixed ϕ the curves for n_B, ranging from 2 to 32 samples per deme, collapse onto a single curve (figure 4.13A). This is expected, as all points lie well above the detection threshold, where equation (4.2) gives $\rho^2 \approx 1 - \gamma_{\text{eff}}/\nu$, and since both γ_{eff} and ν scale linearly with N, the alignment depends on L_{eff} but not on N. As ϕ (the proportion of samples drawn from the spatially structured population B) is reduced, the entire curve shifts down without moving its maximum, mirroring the effect of biased sampling that has been described in the case of discrete populations.

The declining ability of the two largest spatial modes (PC2/PC3) to reveal global spatial structure at low migration rates is a manifestation of an additional mechanism, also described by RMT, by which a PC can fail to align with its theoretical eigenvector even when the corresponding mode of structure is strong enough to be detected. The mechanism, which we have not discussed so far, is the mixing of eigenvectors associated with closely spaced spikes. When two or more structure eigenvalues lie within a small gap of one another, the corresponding PCs are no longer individually determined and each becomes a mixture of the theoretical eigenvectors of the group (Benaych-Georges & Nadakuditi, 2011; Bloemendal et al., 2016; Paul, 2007). In our setting, PC2/PC3 no longer fully capture the global spatial signal; part of it leaks into lower PCs. To show this, let $U = [u_\infty^{(1)}, u_\infty^{(2)}]$ be the orthonormal basis of the subspace spanned by the two theoretical spatial eigenvectors of the equilibrium stepping-stone model, which represents global spatial structure. The squared magnitude of the projection of the kth observed PC of the non-equilibrium model onto this subspace is $p_k = \|U^\top \hat{u}_{t=500}^{(k)}\|^2$, which is invariant to rotations within the subspace. If PC2/PC3 capture the global spatial signal, we expect $p_2 + p_3 \approx 2$, and including further terms p_m for $m > 3$ should not increase the total. If the spatial signal leaks into lower PCs, we expect $p_2 + p_3 < 2$, with contributions from $\{p_m : m > 3\}$ bringing the total closer to 2. Define $S_r = \sum_{k=2}^{r} p_k$ as the total spatial signal captured by PCs 2 to r. Figure 4.13C shows

S_r as a function of $1/M$ for several values of r: at low migration rates the global spatial signal is not lost but redistributed over the lower PCs. RMT predicts that such leakage occurs when the theoretical eigenvalues of a set of spikes lie close enough to one another that the corresponding sample eigenvectors are spread over the joint eigenspace, and may even exchange order (Bloemendal et al., 2016). We confirm that the gap between ν_3 and ν_4 indeed decreases in the non-equilibrium stepping-stone model at low migration rates, because ν_3 stops increasing below a certain migration rate while ν_4 continues to increase (figure 4.13D, see Appendix 4D for how the theoretical eigenvalues are computed for the non-equilibrium stepping-stone model). In the equilibrium stepping-stone model, by contrast, the relative gap stays approximately constant, which is why this phenomenon is not observed there. This final experiment thus reveals an additional way in which observed PCs can fail to represent the underlying axes of variation, though a less severe one than those discussed above: the signal is not lost, but spread across neighboring PCs, and can be recovered by considering them jointly.

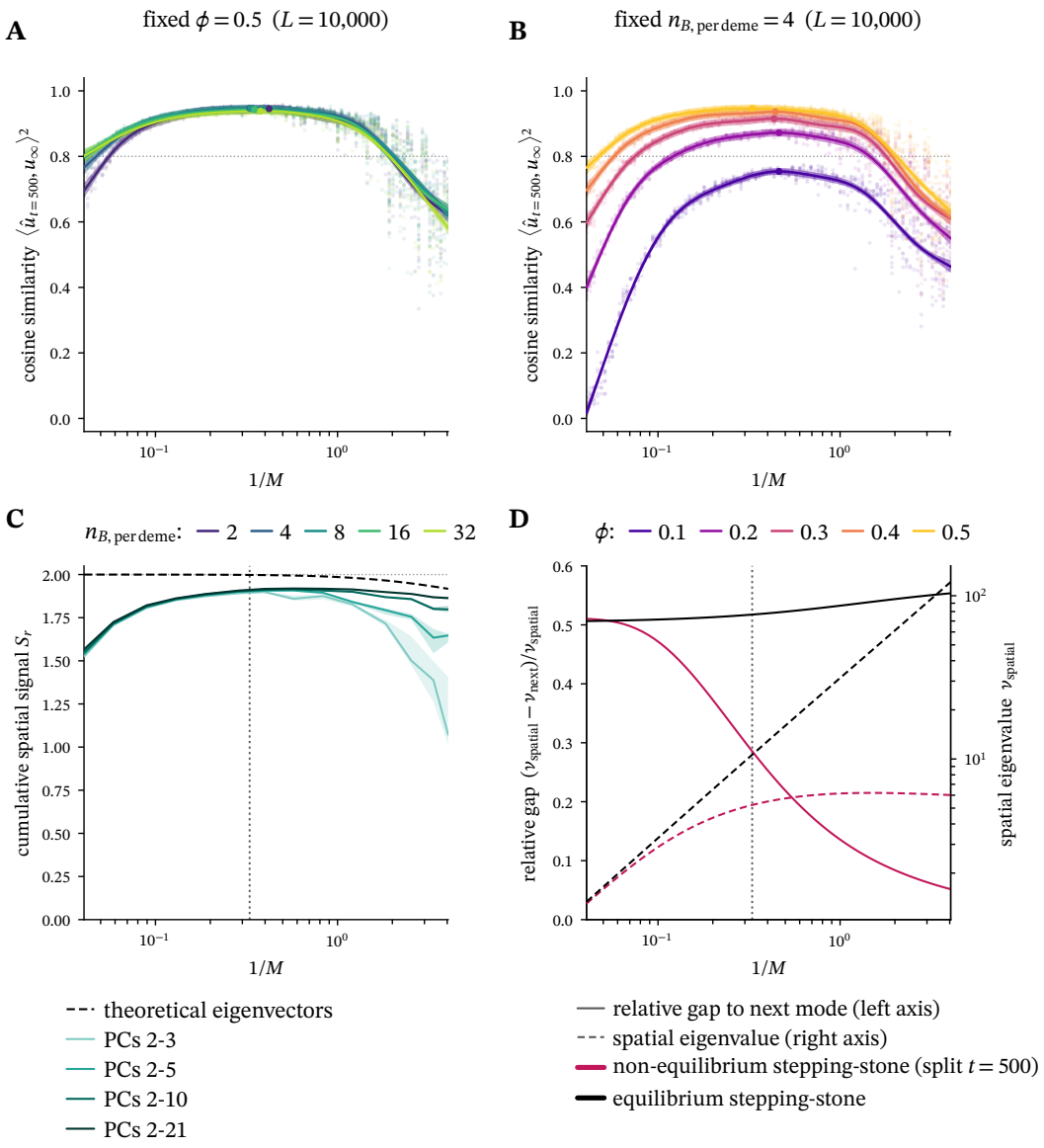

Figure 4.13: Alignment of PC2/PC3 with geography under recent spatial structure is non-monotonic in the migration rate. (A,B) Points are simulation replicates, solid lines are smoothing-spline fits, and dots mark their maxima; the vertical dotted line in (C) and (D) marks the maximum in (A). Alignment with geography increases as migration decreases, up to a point beyond which it declines. **(C)** Total projection S_r of PCs 2 to r onto the subspace of the first two eigenvectors of the equilibrium stepping-stone model ($\phi = 0.5$, $n_B = 4$). At weak migration PC2/PC3 capture progressively less of the signal while wider sets of PCs retain it: the signal is redistributed, not lost. The dashed line is the same projection for the noise-free theoretical PC2/PC3, which remain undistorted. **(D)** Relative gap between the largest spatial eigenvalue ($\nu_2 = \nu_3$) and the next one (ν_4) (left axis), and the eigenvalue itself (right axis), from the exact expected coalescent times. In the non-equilibrium model the largest spatial eigenvalue flattens while the next keeps growing, collapsing the gap; in the equilibrium model the gap stays near 0.5.

4.5 Discussion

The peculiarities associated with a method as pervasive as PCA in the context of population and statistical genetics deserve to be well understood. Previous work tackled two such questions: an axis of variation is detectable only if its associated eigenvalue separates from the bulk (Bryc et al., 2013; Patterson et al., 2006), and how the resulting PCs should and should not be interpreted in the context of population history (G. McVean, 2009; Novembre & Stephens, 2008). In this chapter, we show how detectability does not necessarily imply accuracy, which is a well-known result in RMT. For a given axis of variation, the strength of population structure corresponding to this axis can be quantified in terms of an eigenvalue of a deterministic matrix. If this eigenvalue is below some detectability threshold, the observed PC is asymptotically pure noise. If it exceeds the detectability threshold, the observed PC can remain imperfectly aligned with the true axis of variation. We quantify this exact misalignment and develop a procedure for computing its expected level for general demographic scenarios and sampling schemes. We apply this procedure to derive analytical approximations for the expected level of alignment between the observed PCs and the true axes of variation for several relevant demographic scenarios, with a focus on spatially structured populations, and verify the cases where the analytical approximations agree with simulation experiments. In particular, we identified uneven sampling and unequal population diversity as two factors that affect detectability in addition to the strength of population structure. Unequal diversity enters through the spread of the bulk eigenvalues and through a shift of the structure eigenvalues. It is negligible for a two-population split but not when closely related populations sit inside a diverse sample. Biased sampling acts in two distinct ways. First, it distorts the true axis of variation itself, such that even a perfectly estimated PC no longer reflects geography in a spatially structured population. Second, it dampens the strength of a given mode of structure by reducing its eigenvalue, which raises the detectability threshold as

written in terms of a demographic parameter (such as F_{ST} or M), and lowers the alignment of the observed PC for a given strength of structure specified by such a parameter. In the spatial setting the first effect dominates, with detection thresholds moving comparatively little. Recent spatial structure limits the alignment of the top PCs with geography in a different way. The theoretical eigenvectors are not distorted, but their signal is spread over a larger number of PCs because the corresponding theoretical eigenvalues become close in strength. The example we showed is an instance of a more general phenomenon: the signal for a true axis of variation can be spread over multiple PCs whenever that axis is of similar strength to neighboring modes of structure.

The utility of these results extends beyond quantifying PCA's limitations in visualizing population structure. Another major use of PCA is the correction for population stratification in GWAS by including the top PCs as covariates in the regression (Price et al., 2006). If some of those PCs do not align perfectly with their underlying theoretical eigenvectors, residual stratification will remain even after regressing them out (Mathieson and McVean (2012), Mawass et al., in preparation). Several approaches for regularized PCA have been proposed to compute PCs with better alignment to the true axes of variation for a given strength of structure, and such approaches could potentially reduce such residual stratification in GWAS (Donoho et al., 2018; Lee et al., 2010; Privé et al., 2020). A particularly useful application in the realm of GWAS is to be able to estimate the expected level of alignment just from the observed eigenvalue without any knowledge of the theoretical eigenvalues, which is possible because, above the threshold, the sample eigenvalue is a monotone function of the theoretical eigenvalue (Ding & Yang, 2021; Paul, 2007), so the latter (and hence the expected alignment) can be recovered by inverting this relationship (Dey & Lee, 2019).

Even beyond GWAS, we suspect that methods that rely on the sample covariance of genotype matrices will inherit the limitations of PCA described here. EEMS and FEEMS infer inhomogeneous migration surfaces from the same pairwise dissimilarities that underlie

the sample covariance (Marcus et al., 2021; Petkova et al., 2016). We therefore expect them to inherit the detectability limit we observe for PCs in terms of a lower bound for their ability to resolve migration rates. This remains a hypothesis for now, and perhaps a more general information-theoretic framework can make precise which limits are properties of the data rather than of the method.

In practice, PCA is applied to genotype matrices that are processed differently from the ones we have used in our analyses. SNPs with a minor allele frequency below some threshold are dropped, columns are standardized to unit variance, and they are most likely not independent because of LD. These operations all yield genotype matrices with an underlying theoretical covariance matrix whose eigenvalues and eigenvectors are no longer expressible in closed form in terms of pairwise coalescent times, though they can be computed numerically. Further simulation experiments that extend our results to the class of genotype matrices that are used in practice would be valuable. This has been explored to some extent, but the results were not shown in this chapter as they are still underdeveloped. Moreover, we have invoked the separable covariance model as an approximation to the expected GRM, and it would be worthwhile to study more carefully the conditions under which the separability assumption fails, and perhaps develop more accurate RMT results that are more appropriate for GRM-like covariance matrices.

A particularly interesting phenomenon to characterize within the separable covariance framework is that in which PCs correspond to complex LD structure rather than population structure (Privé et al., 2020). This phenomenon fits naturally within the separable covariance framework, where LD makes the $L \times L$ covariance matrix between SNPs non-diagonal and endows it with spikes of its own. Spikes of the expected GRM and of the expected LD matrix both produce outliers in the spectrum of sample eigenvalues, through different mappings that depend on the aspect ratio (Ding & Yang, 2021), and a PC will reflect LD rather than population structure whenever an outlier induced by a spike of the expected LD

matrix exceeds any of those induced by the spikes of the expected GRM. Our analysis of the simpler, analytically tractable case, where SNPs are independent and not standardized to unit variance, establishes the qualitative behavior that the more realistic settings are likely to inherit, namely a sharp detectability threshold followed by a gradual, quantifiable approach to alignment that is modulated by sampling design and population history, and it opens an avenue for more accurately characterizing the limits of PCA in describing population structure.

Appendix 4A: Deriving the spectrum of the theoretical covariance matrix

Setup and conventions

Index the $N = \sum_{a=1}^{K} n_a$ sampled haplotypes so that the samples from population a occupy the contiguous block $B_a = \{i : a(i) = a\}$, $|B_a| = n_a$. Let $T \in \mathbb{R}^{N \times N}$ be the matrix of mean pairwise coalescent times between sampled haplotypes, $T_{ij} = T^{\text{pop}}_{a(i),a(j)}$ for $i \neq j$ and $T_{ii} = 0$, i.e. in block form

$$T_{[a][b]} = \begin{cases} \tau_a \left(\mathbf{1}_{n_a \times n_a} - I_{n_a} \right), & a = b, \\ T^{\text{pop}}_{ab} \, \mathbf{1}_{n_a \times n_b}, & a \neq b, \end{cases} \tag{4.52}$$

with $\tau_a = T^{\text{pop}}_{aa}$. Write $H = I_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T$ for the centering projector and

$$\bar{T} \;=\; \frac{1}{N(N-1)} \sum_{i \neq j} T_{ij} \tag{4.53}$$

for the mean coalescent time over all distinct pairs of samples. The theoretical covariance matrix is given by

$$G \;=\; -\frac{1}{\bar{\bar{T}}}\, H\, T\, H. \tag{4.54}$$

Throughout we use the derived quantities of the main text,

$$\phi_a = \frac{n_a}{N}, \qquad \psi_a = \sqrt{\phi_a}, \qquad \bar{\tau} = \sum_a \phi_a \tau_a, \qquad \epsilon_a = \frac{\tau_a}{\bar{\tau}} - 1, \qquad (F_2)_{ab} = T^{\mathrm{pop}}_{ab} - \frac{\tau_a + \tau_b}{2},$$

together with the $K \times K$ matrices

$$D = \mathrm{diag}(\vec{n}), \qquad D_{\vec{\psi}} = \mathrm{diag}(\vec{\psi}) = N^{-1/2} D^{1/2}, \qquad D_{\vec{\tau}} = \mathrm{diag}(\vec{\tau}), \qquad \hat{P} = I_K - \psi\psi^T,$$

and $\tilde{F} = D_{\vec{\epsilon}} - \frac{N}{\bar{\tau}} D_{\vec{\psi}} F_2 D_{\vec{\psi}}$, $\hat{F} = \hat{P}\tilde{F}\hat{P}$ as in equations 4.9 and 4.10. Finally, let $U \in \mathbb{R}^{N\times K}$ be the matrix of normalized population indicators,

$$U_{ia} = \frac{1}{\sqrt{n_a}} \mathbf{1}[a(i) = a], \tag{4.55}$$

and set $\Pi = UU^T$, $Q = I_N - \Pi$, $\mathcal{B} = \mathrm{range}(U)$.

Two elementary identities are used repeatedly:

$$\mathbf{1}_N = U\left(\sqrt{N}\,\psi\right), \qquad\qquad D^{1/2}\mathbf{1}_K = \sqrt{N}\,\psi. \tag{4.56}$$

The first holds because $(U\psi)_i = n_{a(i)}^{-1/2}\sqrt{n_{a(i)}/N} = N^{-1/2}$; the second is immediate from $\left(D^{1/2}\mathbf{1}\right)_a = \sqrt{n_a} = \sqrt{N}\psi_a$.

Lemma 0 (normalization identity). $\bar{\bar{T}} = \bar{\tau}\,(1+\alpha)$.

Proof. Using (4.52) and $T^{\text{pop}}_{aa} = \tau_a$,

$$\sum_{i \neq j} T_{ij} = \sum_{a \neq b} n_a n_b T^{\text{pop}}_{ab} + \sum_a n_a(n_a - 1)\tau_a = \vec{n}^T T^{\text{pop}} \vec{n} - \sum_a n_a^2 \tau_a + \sum_a n_a(n_a - 1)\tau_a$$
$$= \vec{n}^T T^{\text{pop}} \vec{n} - N\bar{\tau},$$

since $\sum_a n_a \tau_a = N\bar{\tau}$. On the other hand, because $F_2 = T^{\text{pop}} - \frac{1}{2}\left(\vec{\tau}\mathbf{1}^T + \mathbf{1}\vec{\tau}^T\right)$,

$$\vec{n}^T F_2 \vec{n} = \vec{n}^T T^{\text{pop}} \vec{n} - N \sum_a n_a \tau_a = \vec{n}^T T^{\text{pop}} \vec{n} - N^2 \bar{\tau}.$$

Subtracting, $\sum_{i \neq j} T_{ij} = \vec{n}^T F_2 \vec{n} + N(N-1)\bar{\tau}$, and dividing by $N(N-1)$ gives $\bar{T} = \bar{\tau} + \frac{\vec{n}^T F_2 \vec{n}}{N(N-1)} = \bar{\tau}(1 + \alpha)$. □

Lemma 0 identifies the normalizing constant of G with the quantity $1 + \alpha$ appearing in the theorem: α is simply the relative excess of the mean *between*-sample coalescent time over the sample-size-weighted mean *within*-population coalescent time.

1. Decomposing the space

Lemma 1. U is an isometry, Π and Q are complementary orthogonal projectors, and

$$\mathbb{R}^N = \mathcal{B} \oplus \mathcal{B}^\perp, \qquad \dim \mathcal{B} = K, \qquad \dim \mathcal{B}^\perp = N - K. \tag{4.57}$$

Proof. The columns of U (which is the matrix of normalized population indicators) have disjoint supports B_a, so $\left(U^T U\right)_{ab} = \delta_{ab}$, i.e. $U^T U = I_K$ and $\|Ux\| = \|x\|$ for all $x \in \mathbb{R}^K$, proving that U is an isometry. Consequently $\Pi^2 = UU^T UU^T = UU^T = \Pi$ and $\Pi^T = \Pi$, so Π is the orthogonal projector onto $\mathcal{B} = \text{range}(U)$, and $Q = I_N - \Pi$ is the complementary orthogonal projector onto $\mathcal{B}^\perp$. Since U has full column rank, $\dim \mathcal{B} = \text{rank}\, \Pi = K$ and $\dim \mathcal{B}^\perp = N - K$. □

Concretely, $\mathcal{B}$ is the space of vectors that are constant within each population, and $\mathcal{B}^\perp$ the space of vectors that average to zero within each population. By (4.56), $\mathbf{1}_N \in \mathcal{B}$: the direction removed by centering lives entirely in the "structure" subspace. Lemmas 3 and 6 below show that both subspaces are G-invariant, so that the spectrum of G is the disjoint union of the spectra of G restricted to each of the subspaces.

Deriving the floor spectrum

Lemma 2. Let $V_a = \left\{ v \in \mathbb{R}^N : \ \mathrm{supp}(v) \subseteq B_a, \ \sum_{i \in B_a} v_i = 0 \right\}$ be the subspace of vectors supported on population a and summing to zero. Then

$$\mathcal{B}^\perp = \bigoplus_{a=1}^{K} V_a, \qquad \dim V_a = n_a - 1. \tag{4.58}$$

Proof. V_a consists of vectors with non-zero elements only on B_a, with these non-zero elements summing to zero. Vectors supported only on B_a form a subspace of dimension n_a, and with the single linear constraint of the elements summing to zero, $\dim V_a = n_a - 1$. Additionally, the subspaces V_a are all mutually orthogonal, since any two vectors $v \in V_a$ and $w \in V_b$ with $a \neq b$ have disjoint support and therefore $\langle v, w \rangle = 0$. Therefore, the sum $\bigoplus_{a=1}^{K} V_a$ is a direct sum with dimension $\sum_a (n_a - 1) = N - K$.

We now show that $V_a \subseteq \mathcal{B}^\perp$. Consider a vector $v \in V_a$. $\left(U^T v\right)_b = n_b^{-1/2} \sum_{i \in B_b} v_i$. If $a \neq b$, then the element is zero since v has zero elements in the block B_b. If $a = b$, that element will also be zero by the summing-to-zero constraint of vectors in V_a. Therefore $U^T v = 0$, and $V_a \subseteq \ker U^T = \ker \Pi = \mathcal{B}^\perp$. Thus $\bigoplus_a V_a \subseteq \mathcal{B}^\perp$, and since both spaces have dimension $N - K$ (Lemma 1) they coincide. □

Lemma 3. For any $v \in V_a$,

$$Gv = \frac{1 + \epsilon_a}{1 + \alpha} \, v, \tag{4.59}$$

so V_a is an eigenspace of G with eigenvalue $\frac{1+\epsilon_a}{1+\alpha}$ of multiplicity $n_a - 1$.

Proof. Let $v \in V_a$. Since its values sum to zero, $\mathbf{1}_N^T v = 0$ and $Hv = v$. Therefore, $Gv = -\frac{1}{\bar{T}} HTv$, with T having a block structure defined by (4.52). For $i \notin B_a$,

$$(Tv)_i = \sum_{j \in B_a} T^{\text{pop}}_{a(i),a} v_j = T^{\text{pop}}_{a(i),a} \sum_{j \in B_a} v_j = 0,$$

since the elements of v sum to zero as defined in Lemma 2. For $i \in B_a$,

$$(Tv)_i = \tau_a \sum_{j \in B_a,\, j \neq i} v_j = \tau_a \,(0 - v_i) = -\tau_a v_i,$$

since $T_{ii} = 0$ and $T_{ij} = \tau_a$ in the diagonal block B_a of T. Hence $Tv = -\tau_a v$. Therefore

$$Gv = -\frac{1}{\bar{T}} HTHv = -\frac{1}{\bar{T}} HTv = \frac{\tau_a}{\bar{T}} Hv = \frac{\tau_a}{\bar{T}}\, v.$$

By Lemma 0, $\bar{T} = \bar{\tau}(1+\alpha)$, and by definition $\tau_a/\bar{\tau} = 1 + \epsilon_a$, so $\tau_a/\bar{T} = \frac{1+\epsilon_a}{1+\alpha}$. □

Lemmas 2 and 3 account for $N - K$ eigenvalues of G, the *floor* of the spectrum. It is composed of eigenvalues $\frac{1+\epsilon_a}{1+\alpha}$ each with multiplicity $n_a - 1$ for each $a = 1, \ldots, K$.

Deriving the structure spectrum

Lemma 4. $HU = U\hat{P}$.

Proof. By (4.56), $\mathbf{1}_N^T U = \sqrt{N}\psi^T$ and $\mathbf{1}_N = \sqrt{N}\, U\psi$. Hence

$$HU = U - \frac{1}{N}\mathbf{1}_N \left(\mathbf{1}_N^T U\right) = U - \frac{1}{N}\left(\sqrt{N}U\psi\right)\left(\sqrt{N}\psi^T\right) = U\left(I_K - \psi\psi^T\right) = U\hat{P}. \qquad \square$$

Thus centering in $\mathbb{R}^N$ acts on the coordinates of $\mathcal{B}$ exactly as the sampling-weighted projector $\hat{P}$ acts on $\mathbb{R}^K$.

Lemma 5. $TU = UA$, where

$$A \;=\; D^{1/2} T^{\mathrm{pop}} D^{1/2} - \mathrm{diag}(\vec{\tau}). \tag{4.60}$$

Proof. For $i \in B_a$ and any b,

$$(TU)_{ib} = \frac{1}{\sqrt{n_b}} \sum_{j \in B_b} T_{ij} = \begin{cases} \dfrac{1}{\sqrt{n_b}}\, n_b\, T^{\mathrm{pop}}_{ab} = \sqrt{n_a n_b}\, \dfrac{T^{\mathrm{pop}}_{ab}}{\sqrt{n_a}}, & b \neq a, \\ \dfrac{1}{\sqrt{n_a}}\, (n_a - 1)\, \tau_a = (n_a \tau_a - \tau_a)\, \dfrac{1}{\sqrt{n_a}}, & b = a, \end{cases}$$

where the $b = a$ case uses $T_{ii} = 0$, so that only $n_a - 1$ terms contribute. Since $(UA)_{ib} = n_a^{-1/2} A_{ab}$, the two sides agree precisely when $A_{ab} = \sqrt{n_a n_b}\, T^{\mathrm{pop}}_{ab}$ for $a \neq b$ and $A_{aa} = n_a \tau_a - \tau_a$, which is (4.60). □

The subtraction of $\mathrm{diag}(\vec{\tau})$ is the algebraic trace of the zero diagonal of T: it is the "self-pair" correction that later produces the $+1$ in $\lambda_k = (1 + \nu_k)/(1 + \alpha)$.

Lemma 6. Define $M = -\hat{P} A \hat{P}$. Then $GU = \frac{1}{T} UM$. If w is an eigenvector of M, then Uw is an eigenvector of G. In particular, if $Mw = \beta_k w$ then

$$G(Uw) = \frac{\beta_k}{T} (Uw). \tag{4.61}$$

Proof. Applying Lemmas 4 and 5 in turn,

$$GU = -\frac{1}{T} HTHU = -\frac{1}{T} HTU\hat{P} = -\frac{1}{T} HUA\hat{P} = -\frac{1}{T} U\hat{P}A\hat{P} = \frac{1}{T} UM.$$

If $Mw = \beta_k w$ then $G(Uw) = \frac{1}{T} UMw = \frac{\beta_k}{T} (Uw)$, and $Uw \neq 0$ because U is an isometry (Lemma 1). Since M is symmetric it has an orthonormal eigenbasis $w_1, \ldots, w_K$ of $\mathbb{R}^K$, whose images $Uw_1, \ldots, Uw_K$ form an orthonormal eigenbasis of $\mathcal{B}$; hence $\mathrm{spec}(G|_{\mathcal{B}}) = \frac{1}{T} \mathrm{spec}(M)$. □

Two consequences are immediate. First, $M\psi = -\hat{P}A\hat{P}\psi = 0$ because $\hat{P}\psi = 0$; by (4.56) the corresponding eigenvector of G is $U\psi = N^{-1/2}\mathbf{1}_N$. This is the trivial eigenvalue 0 of the theorem, produced by the centering operation. Second, since M is symmetric and $\psi \in \ker(M)$, the orthogonal complement $\psi^{\perp} \subset \mathbb{R}^K$ is M-invariant and carries the remaining $K-1$ eigenvalues.

Lemma 7. $\hat{P}D^{1/2}T^{\mathrm{pop}}D^{1/2}\hat{P} = \hat{P}D^{1/2}F_2D^{1/2}\hat{P}$: the additive part of T^{pop} that distinguishes it from F_2 is annihilated by centering.

Proof. By definition of F_2,

$$T^{\mathrm{pop}} = F_2 + \tfrac{1}{2}\left(\vec{\tau}\mathbf{1}_K^T + \mathbf{1}_K\vec{\tau}^T\right),$$

so, writing $u = D^{1/2}\vec{\tau}$ and using $D^{1/2}\mathbf{1}_K = \sqrt{N}\psi$ from (4.56),

$$D^{1/2}T^{\mathrm{pop}}D^{1/2} = D^{1/2}F_2D^{1/2} + \frac{\sqrt{N}}{2}\left(u\psi^T + \psi u^T\right).$$

Conjugating by $\hat{P}$ and using $\hat{P}\psi = 0$,

$$\hat{P}\left(u\psi^T + \psi u^T\right)\hat{P} = (\hat{P}u)(\hat{P}\psi)^T + (\hat{P}\psi)(\hat{P}u)^T = 0. \qquad \square$$

This is the algebraic reason why the spectrum of G depends on the demography only through F_2 and the within-population diversities: any additive "branch-length" component $\frac{1}{2}(\tau_a+\tau_b)$ of the pairwise coalescent times is invisible after centering.

Lemma 8. $M|_{\psi^{\perp}} = \bar{\tau}\left(I + \hat{F}\right)\big|_{\psi^{\perp}}$, and consequently the $K-1$ non-trivial eigenvalues of $G|_{\mathcal{B}}$ are $\beta_k/\bar{T} = \frac{1+\nu_k}{1+\alpha}$, where $\nu_k \in \mathrm{spec}\left(\hat{F}|_{\psi^{\perp}}\right)$.

Proof. By Lemma 7 and (4.60),

$$M = -\hat{P}\left(D^{1/2}T^{\mathrm{pop}}D^{1/2} - D_{\vec{\tau}}\right)\hat{P} = \hat{P}\left(D_{\vec{\tau}} - D^{1/2}F_2D^{1/2}\right)\hat{P}. \tag{4.62}$$

Since $D_{\vec{\psi}} = N^{-1/2} D^{1/2}$ we have $\frac{N}{\bar{\tau}} D_{\vec{\psi}} F_2 D_{\vec{\psi}} = \frac{1}{\bar{\tau}} D^{1/2} F_2 D^{1/2}$, and since $\epsilon_a = \tau_a/\bar{\tau} - 1$ we have $D_{\vec{\epsilon}} = \bar{\tau}^{-1} D_{\vec{\tau}} - I_K$. Therefore

$$\bar{\tau}\left(I_K + \tilde{F}\right) = \bar{\tau} I_K + (D_{\vec{\tau}} - \bar{\tau} I_K) - D^{1/2} F_2 D^{1/2} = D_{\vec{\tau}} - D^{1/2} F_2 D^{1/2}.$$

Substituting into (4.62),

$$M = \bar{\tau}\, \hat{P}\left(I_K + \tilde{F}\right)\hat{P} = \bar{\tau}\left(\hat{P} + \hat{F}\right),$$

using $\hat{P}^2 = \hat{P}$ and $\hat{F} = \hat{P}\tilde{F}\hat{P}$. Restricted to $\psi^\perp$, $\hat{P}$ acts as the identity: if $w \in \psi^\perp$ then $\psi^T w = 0$, so $\hat{P}w = w - \psi(\psi^T w) = w$. Consequently, $M|_{\psi^\perp} = \bar{\tau}\left(I + \hat{F}\right)|_{\psi^\perp}$.

Hence if $w \in \psi^\perp$ satisfies $\hat{F}w = \nu_k w$, then $Mw = \bar{\tau}(1+\nu_k)w$, i.e. $\beta_k = \bar{\tau}(1+\nu_k)$, and by Lemma 6 together with Lemma 0,

$$G(Uw) = \frac{\beta_k}{\bar{T}}(Uw) = \frac{\bar{\tau}\,(1+\nu_k)}{\bar{\tau}\,(1+\alpha)}(Uw) = \frac{1+\nu_k}{1+\alpha}\,(Uw). \qquad \square$$

Proof of the theorem

By Lemma 1, $\mathbb{R}^N = \mathcal{B} \oplus \mathcal{B}^\perp$ with $\dim \mathcal{B} = K$; by Lemmas 3 and 6 both subspaces are G-invariant, so $\mathrm{spec}(G)$ is the disjoint union of $\mathrm{spec}\big(G|_{\mathcal{B}^\perp}\big)$ and $\mathrm{spec}(G|_{\mathcal{B}})$.

On $\mathcal{B}^\perp = \bigoplus_a V_a$ (Lemma 2), Lemma 3 gives an orthogonal eigenbasis with eigenvalues $\frac{1+\epsilon_a}{1+\alpha}$ repeated $n_a - 1$ times for each $a = 1, \ldots, K$; these are the $N - K$ floor eigenvalues.

On $\mathcal{B}$, Lemma 6 gives $\mathrm{spec}\,(G|_{\mathcal{B}}) = \frac{1}{\bar{T}}\,\mathrm{spec}(M)$ with $M = \bar{\tau}(\hat{P} + \hat{F})$. Since $\hat{P}\psi = 0$ and $\hat{F}\psi = 0$, the vector ψ contributes the eigenvalue 0, with associated eigenvector $U\psi \propto \mathbf{1}_N$; this is the trivial eigenvalue induced by centering. On the M-invariant complement $\psi^\perp$, Lemma 8 gives the remaining $K-1$ eigenvalues $\frac{1+\nu_k}{1+\alpha}$, with ν_k the eigenvalues of $\hat{F}$ restricted to $\psi^\perp$;

these are the structure eigenvalues. Collecting the three groups,

$$\operatorname{spec}(G) = \{0\} \cup \left\{ \frac{1+\nu_k}{1+\alpha} : \nu_k \in \operatorname{spec}\left(\hat{F}|_{\psi^\perp}\right) \right\} \cup \left\{ \frac{1+\epsilon_a}{1+\alpha} \text{ with multiplicity } n_a - 1,\ a = 1, \ldots, K \right\},$$

which contains $1 + (K-1) + (N-K) = N$ eigenvalues in total, as required. ■

Remarks

Eigenvectors. The structure eigenvectors are Uw_k, i.e. vectors that are constant within each population, with the value $w_{k,a}/\sqrt{n_a}$ for all samples from population a. These are the axes of variation that PCA attempts to estimate; the sampling scheme enters both through the $\sqrt{n_a}$ rescaling and through $\hat{P}$.

Interpretation of the shift. Lemma 8 shows that $G|_{\mathcal{B}}$ is, up to the overall factor $\bar{\tau}/\bar{T} = (1+\alpha)^{-1}$, the identity plus $\hat{F}$. The identity contributes the same 1 that appears in the floor eigenvalues (through $1 + \epsilon_a$), so ν_k measures exactly the *excess* eigenvalue of axis k above the bulk, which is what makes it the natural quantity to compare against a detectability threshold.

Degenerate cases. If two populations have equal within-population coalescent times, $\tau_a = \tau_b$, the corresponding floor eigenvalues coincide and the associated eigenspaces merge. If some $n_a = 1$, that population contributes no floor eigenvalue, consistent with $\dim V_a = 0$.

Appendix 4B: Identification of A and B in the separable factorization

Let Z be the $N \times L$ genotype matrix with binary entries (N samples and L independent SNPs), $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ the centering matrix, and

$$Y = \sqrt{\frac{2}{\bar{\Delta} L}}\, HZ, \qquad \Delta_{ij} = \frac{1}{L}\sum_l \left(Z_{il} - Z_{jl}\right)^2, \qquad \bar{\Delta} = \frac{1}{N(N-1)}\sum_{i\neq j}\Delta_{ij}. \tag{4.63}$$

We write p_l for the population allele frequency at site l and $\hat{p}_l = \frac{1}{N}\sum_i Z_{il}$ for the sample frequency, with $h_l = p_l(1-p_l)$ and $\hat{h}_l = \hat{p}_l(1-\hat{p}_l)$ the corresponding heterozygosities, $\bar{h}$ and $\bar{\hat{h}}$ their averages over the L sites, and

$$\eta_l = \frac{h_l}{\bar{h}}, \qquad \hat{\eta}_l = \frac{\hat{h}_l}{\bar{\hat{h}}}, \qquad \text{so that} \qquad \sum_l \eta_l = \sum_l \hat{\eta}_l = L. \tag{4.64}$$

The separable covariance assumption (4.22) allows Y to be factorized as

$$Y = A^{1/2} W B^{1/2}, \qquad \text{where: } E[W_{ij}] = 0 \ , \ E\big[W_{ip}W_{jq}\big] = \frac{1}{L}\delta_{ij}\delta_{pq}. \tag{4.65}$$

In this appendix we identify the deterministic matrices A and B as $G = -\dfrac{HTH}{\bar{T}}$ and $\mathrm{diag}(\eta_l)$ respectively.

Using $E[WCW^T] = \frac{\mathrm{tr}(C)}{L} I$ and its transpose analogue, (4.65) implies the two moment conditions

$$E\Big[YY^T\Big] = \frac{\mathrm{tr}(B)}{L}\, A, \qquad E\Big[Y^TY\Big] = \frac{\mathrm{tr}(A)}{L}\, B. \tag{4.66}$$

Thus A is proportional to the row moment of Y and B to its column moment, and the identification reduces to computing those two moments. Note that (4.65) cannot fix the

overall scale of either matrix, since $(A, B) \mapsto (cA, c^{-1}B)$ leaves Y and both conditions in (4.66) unchanged. One convention must therefore be imposed; we adopt, without loss of generality,

$$\mathrm{tr}(B) = L. \tag{4.67}$$

The row moment has already been derived in equation 4.6,

$$E\Big[YY^T\Big] = -\frac{HTH}{\bar{T}} = G, \qquad \mathrm{tr}(G) = N - 1. \tag{4.68}$$

For the column moment, note that the diagonal entries of $Y^TY = \frac{2}{\bar{\Delta}L} Z^T H Z$ are

$$\Big(Y^TY\Big)_{ll} = \frac{2}{\bar{\Delta}L} \sum_i (Z_{il} - \hat{p}_l)^2 = \frac{2N\hat{h}_l}{\bar{\Delta}L}, \tag{4.69}$$

where we used $Z_{il}^2 = Z_{il}$. Since Z_{il} is binary, $(Z_{il} - Z_{jl})^2 = |Z_{il} - Z_{jl}|$, so the number of ordered pairs $i \neq j$ differing at site l is $2(N\hat{p}_l)(N(1-\hat{p}_l)) = 2N^2\hat{h}_l$, and

$$\bar{\Delta} = \frac{1}{N(N-1)} \cdot \frac{1}{L} \sum_l 2N^2\hat{h}_l = \frac{2N}{N-1}\,\bar{\hat{h}}. \tag{4.70}$$

Substituting (4.70) into (4.69), we obtain the exact identity

$$\Big(Y^TY\Big)_{ll} = \frac{N-1}{L} \cdot \frac{\hat{h}_l}{\bar{\hat{h}}} = \frac{N-1}{L}\,\hat{\eta}_l. \tag{4.71}$$

No expectation has been taken and no property of the genealogy has been used: because Y is normalized by $\bar{\Delta}$, which is itself the mean sample heterozygosity up to a constant, the diagonal of Y^TY is the scaled site heterozygosity spectrum. In particular $\mathrm{tr}(Y^TY) = \frac{N-1}{L}\sum_l \hat{\eta}_l = N - 1$ by (4.64), in agreement with $\mathrm{tr}(YY^T) = N - 1$.

For $l \neq l'$ the sites are unlinked, so $E\left[(Z^THZ)_{ll'}\right] = 0$ and the column moment is

diagonal. Taking the expectation in (4.71) and replacing the sample heterozygosity ratio by its population counterpart,

$$E\Big[Y^T Y\Big] = \frac{N-1}{L}\,\mathrm{diag}(\eta_l)\,. \tag{4.72}$$

Combining (4.66) with (4.68) and (4.72), the matrices A and B must satisfy

$$\frac{\mathrm{tr}(B)}{L}\,A = G, \qquad \frac{\mathrm{tr}(A)}{L}\,B = \frac{N-1}{L}\,\mathrm{diag}(\eta_l)\,. \tag{4.73}$$

Taking the trace of each and using $\mathrm{tr}(G) = N-1$ and $\sum_l \eta_l = L$, both lines of (4.73) yield the same constraint,

$$\mathrm{tr}(A)\,\mathrm{tr}(B) = L(N-1). \tag{4.74}$$

The fact that the two agree is automatic, since $\mathrm{tr}(E[YY^T]) = \mathrm{tr}(E[Y^T Y])$ for any Y; the content of (4.74) is rather that the common value is $L(N-1)$, which is what the normalization by $\bar{\Delta}$ in (4.63) leads to. Consequently (4.73) determines A and B only up to the one-parameter family $(A, B) \mapsto (cA, c^{-1}B)$ as anticipated above, and a convention is needed to pin down both matrices. The choice $\mathrm{tr}(B) = L$ in (4.67) does not constrain the data: it fixes only how the invariant product (4.74) is apportioned between the two factors.

Finally, imposing the convention (4.67), the first equation of (4.73) gives

$$A = G, \qquad \mathrm{tr}(A) = N-1, \tag{4.75}$$

and substituting $\mathrm{tr}(A) = N-1$ into the second gives

$$B = \mathrm{diag}(\eta_l)\,, \qquad \eta_l = \frac{h_l}{\bar{h}} = \frac{p_l(1-p_l)}{\bar{h}}, \tag{4.76}$$

which indeed satisfies $\mathrm{tr}(B) = \sum_l \eta_l = L$ as required.

Appendix 4C: Concrete mapping to the Ding & Yang model

We show that the self-consistent system of Ding and Yang (2021), equation (2.17), reduces to the pair of equations (4.29) and (4.30) governing (ν, w) stated in the main text.

Setup. Let π_A, π_B denote the empirical spectral distributions of A, B as in Ding and Yang's equation (2.5), and let $m_{1c}(z), m_{2c}(z)$ solve their self-consistent system, equation (2.17):

$$m_{1c}(z) = d_n \int \frac{x}{-z[1 + x m_{2c}(z)]}\, \pi_A(dx), \tag{4.77}$$

$$m_{2c}(z) = \int \frac{y}{-z[1 + y m_{1c}(z)]}\, \pi_B(dy), \tag{4.78}$$

with $d_n = N/L = \gamma$ in our notation. By their Lemma 2.5, for z real and greater than the right spectral edge λ_+, both $m_{1c}(z)$ and $m_{2c}(z)$ are real, negative, and satisfy $m_{2c}(z) \in (-1/\max_i \sigma_i^a,\, 0)$, $m_{1c}(z) \in (-1/\max_l \sigma_l^b,\, 0)$. This justifies defining the positive quantities

$$t := -\frac{1}{m_{2c}(z)}, \qquad w := -\frac{1}{m_{1c}(z)}, \tag{4.79}$$

each strictly exceeding the corresponding population spectral radius. In our model, $t = (1 + \nu)/(1 + \alpha)$ is the eigenvalue of A associated with a mode of structure with strength ν.

Reduction of the second equation. Substituting $m_{1c}(z) = -1/w$ into the bracket of the second equation gives $1 + y m_{1c}(z) = (w - y)/w$, so

$$m_{2c}(z) = \int \frac{y}{-z(w - y)/w}\, \pi_B(dy) = -\frac{w}{z} \int \frac{y}{w - y}\, \pi_B(dy) = -\frac{w}{z} \Psi_1(w), \tag{4.80}$$

using the definition of Ψ_1, where the integral reduces to a sum over the discrete values η_l. Substituting $m_{2c}(z) = -1/t$ on the left and solving for z:

$$z = t\, w\, \Psi_1(w). \tag{4.81}$$

Reduction of the first equation. By the identical argument with $m_{2c}(z) = -1/t$ substituted into the first bracket, $1 + x m_{2c}(z) = (t - x)/t$, and

$$m_{1c}(z) = -\frac{d_n\, t}{z} \int \frac{x}{t - x}\, \pi_A(dx) = -\frac{\gamma\, t}{z} \Phi_1(\nu). \tag{4.82}$$

Substituting $m_{1c}(z) = -1/w$ and solving for z:

$$z = \gamma\, t\, w\, \Phi_1(\nu). \tag{4.83}$$

The coupling equation. Equating the right-hand sides of (4.81) and (4.83) eliminates z, t, and w as common factors, giving

$$\Psi_1(w) = \gamma\, \Phi_1(\nu) \tag{4.84}$$

which is equation (4.29) of the main text. This is the statement that (2.17) is a single equation in (t, w) once z is eliminated; equation (4.81) then recovers the outlier location itself, consistent with Ding and Yang's equation (3.2), $\theta_1(\sigma^a) = g_{2c}(-(\sigma^a)^{-1})$.

The critical-point equation. Eliminating z leaves the coupling equation $\Psi_1(w) = \gamma \Phi_1(\nu)$, which defines w as a function of t (equivalently of ν, since $t = (1+\nu)/(1+\alpha)$). Substituting back into (4.81) gives the outlier location as an explicit function of the population spike,

$$z(t) = t\, w(t)\, \Psi_1\big(w(t)\big), \tag{4.85}$$

which is Ding and Yang's $g_{2c}(-1/t)$, the inverse of m_{2c}. A population spike t produces an outlier only if $z(t)$ lies outside the bulk, and by their Lemma 2.5 the spectral edge λ_+ is the critical value of this inverse map: $\lambda_+ = g_{2c}(m_c)$ with $g'_{2c}(m_c) = 0$ (this is the content of the pair of conditions $f = 0$, $\partial_m f = 0$ in their equation (2.20)). Since $g'_{2c}(-1/t) = t^2\, dz/dt$, the threshold spike t_c is characterized by

$$\left.\frac{d\log z}{d\log t}\right|_{t=t_c} = 0. \tag{4.86}$$

To evaluate this derivative, differentiate (4.85) logarithmically, $\log z = \log t + \log w + \log \Psi_1(w)$, and use the two identities

$$\frac{d\Phi_1}{d\nu} = -\frac{\Phi_1 + \Phi_2}{1+\nu}, \qquad \frac{d\Psi_1}{dw} = -\frac{\Psi_1 + \Psi_2}{w}, \tag{4.87}$$

which follow from differentiating the definitions term by term. The second identity gives $\frac{d\log\Psi_1}{d\log w} = -\frac{\Psi_1+\Psi_2}{\Psi_1}$, so

$$\frac{d\log z}{d\log t} = 1 - \frac{\Psi_2}{\Psi_1}\frac{d\log w}{d\log t}. \tag{4.88}$$

Differentiating the coupling equation with respect to $\log t$ (note $d\log t = d\nu/(1+\nu)$) and using both identities together with $\Psi_1 = \gamma\Phi_1$,

$$\frac{d\log w}{d\log t} = \frac{\gamma(\Phi_1+\Phi_2)}{\Psi_1+\Psi_2} = \frac{\Psi_1\,(1+\Phi_2/\Phi_1)}{\Psi_1+\Psi_2}. \tag{4.89}$$

Substituting (4.89) into (4.88),

$$\frac{d\log z}{d\log t} = 1 - \frac{(\Phi_1+\Phi_2)\,\Psi_2}{\Phi_1\,(\Psi_1+\Psi_2)} = \frac{\Phi_1\Psi_1 - \Phi_2\Psi_2}{\Phi_1(\Psi_1+\Psi_2)}. \tag{4.90}$$

Setting this to zero yields the critical-point equation

$$\Phi_2(\nu_c)\,\Psi_2(w_c) = \Phi_1(\nu_c)\,\Psi_1(w_c) \quad \Longleftrightarrow \quad \frac{\Phi_2(\nu_c)}{\Phi_1(\nu_c)} \cdot \frac{\Psi_2(w_c)}{\Psi_1(w_c)} = 1, \tag{4.91}$$

solved jointly with the coupling equation for (ν_c, w_c). Detectability of a structure eigenvalue ν_k, i.e. $\nu_k > \nu_c$, corresponds exactly to Ding and Yang's supercritical condition, equation (3.1): $\sigma^a > -m_{2c}^{-1}(\lambda_+)$.

The cosine similarity. For a well-separated, non-degenerate spike, Ding and Yang's Theorem 3.10 (see also Example 3.12) gives the leading-order generalized component

$$|\langle v_i^a, \xi_{\alpha(i)}\rangle|^2 = \frac{1}{\sigma_i^a} \cdot \frac{g'_{2c}(-(\sigma_i^a)^{-1})}{g_{2c}(-(\sigma_i^a)^{-1})} + O_{\prec}(\cdot). \tag{4.92}$$

With $\sigma_i^a = t$, $g_{2c}(-1/t) = z$ and $g'_{2c}(-1/t) = t^2\,dz/dt$, the leading term is exactly $\frac{d\log z}{d\log t}$, already computed in (4.90):

$$\langle u_k, \hat{u}_k\rangle^2 = 1 - \frac{\big(\Phi_1(\nu_k) + \Phi_2(\nu_k)\big)\Psi_2(w_k)}{\Phi_1(\nu_k)\big(\Psi_1(w_k) + \Psi_2(w_k)\big)}, \tag{4.93}$$

as stated in the main text. The threshold and the overlap are therefore the same function evaluated at its zero and above it: the overlap vanishes exactly at $\nu_k = \nu_c$ and degrades continuously to zero as ν_k approaches the detection threshold from above.

Small-γ approximation via an effective aspect ratio

When $N \ll L$, so that $\gamma = N/L \to 0$, the exact joint solution described above can be replaced by a closed-form or near-closed-form procedure obtained by expanding the ratio $\Psi_2(w)/\Psi_1(w)$ in powers of $1/w$. Recall $\eta_l = h_l/\bar{h}$ has unit mean by construction. Define the

two locus-side moments and their combination

$$\kappa_2 = \frac{L \sum_l h_l^2}{\left(\sum_l h_l\right)^2}, \qquad \kappa_3 = \frac{L^2 \sum_l h_l^3}{\left(\sum_l h_l\right)^3}, \qquad \delta = \kappa_3 - \kappa_2^2 \ \ (\geq 0 \text{ always}), \tag{4.94}$$

equivalently $\kappa_2 = \frac{1}{L}\sum_l \eta_l^2$, $\kappa_3 = \frac{1}{L}\sum_l \eta_l^3$ in terms of the normalized η_l. Writing $\kappa_2 = L/L_{\text{eff}}$ defines an effective number of loci $L_{\text{eff}} := (\sum_l \eta_l)^2 / \sum_l \eta_l^2$, the inverse participation ratio of the B-spectrum. Since $\kappa_2 \geq 1$ with equality iff $B \propto I$, we have $L_{\text{eff}} \leq L$, with equality only when every locus contributes equal variance to the genotype matrix.

Effective aspect ratio. We derive the expansion of $\Psi_2(w)/\Psi_1(w)$ in three steps: expand the kernels for large w, invert the coupling equation to express w in terms of ν to the order needed, and substitute. We then use the result to define an effective aspect ratio γ_{eff}.

Step 1: large-w expansion of Ψ_1, Ψ_2. Since η_l has unit mean by construction, expanding each kernel $\eta_l/(w - \eta_l) = (\eta_l/w)/(1 - \eta_l/w)$ in powers of η_l/w and averaging over l gives

$$\Psi_1(w) = \frac{1}{w} + \frac{\kappa_2}{w^2} + \frac{\kappa_3}{w^3} + O(w^{-4}), \qquad \Psi_2(w) = \frac{\kappa_2}{w^2} + \frac{2\kappa_3}{w^3} + O(w^{-4}). \tag{4.95}$$

Step 2: inverting the coupling equation. Write $T := \gamma\Phi_1(\nu)$, so the coupling equation (4.29) reads $\Psi_1(w) = T$. As $\gamma \to 0$ we have $T \to 0$ and correspondingly $w \to \infty$, so we solve for $w(T)$ perturbatively using (4.95). Posit $w = 1/T + c + O(T)$ for a constant c to be determined, and substitute into $\Psi_1(w) = T$:

$$\frac{1}{1/T + c} + \frac{\kappa_2}{(1/T + c)^2} = T \implies T - cT^2 + \kappa_2 T^2 + O(T^3) = T. \tag{4.96}$$

Matching the $O(T^2)$ terms forces $c = \kappa_2$, so

$$w = \frac{1}{T} + \kappa_2 + O(T), \qquad T = \gamma\Phi_1(\nu). \tag{4.97}$$

This $O(1)$ shift is the key correction relative to the naive leading order $w \approx 1/T = 1/(\gamma\Phi_1(\nu))$: it is not negligible, because it enters the ratio Ψ_2/Ψ_1 at the same order as the next term of (4.95) itself.

Step 3: substitution. Insert (4.97) into (4.95) and expand $\Psi_2(w)/\Psi_1(w)$ as a series in T:

$$\Psi_1(w) = T + \delta T^3 + O(T^4),$$
$$\Psi_2(w) = \kappa_2 T^2 + 2\delta T^3 + O(T^4),$$

so that to the order retained,

$$\frac{\Psi_2(w)}{\Psi_1(w)} = \kappa_2\, T + 2(\kappa_3 - \kappa_2^2)\, T^2 + O(T^3) = \gamma\kappa_2\Phi_1(\nu) + 2\gamma^2(\kappa_3 - \kappa_2^2)\Phi_1(\nu)^2 + O(\gamma^3). \tag{4.98}$$

Factoring out $\gamma\kappa_2\Phi_1(\nu)$ and writing $\delta = \kappa_3 - \kappa_2^2$ gives

$$\frac{\Psi_2(w)}{\Psi_1(w)} = \gamma\,\kappa_2\,\Phi_1(\nu)\Big[1 + \frac{2\delta}{\kappa_2}\gamma\Phi_1(\nu)\Big] + O(\gamma^3), \tag{4.99}$$

valid for both $\epsilon \equiv 0$ and $\epsilon \neq 0$, since the derivation concerns only the B-side quantities Ψ_1, Ψ_2 and never uses the specific form of Φ.

Definition of γ_{eff}. Equation (4.99) motivates defining the *effective aspect ratio*

$$\gamma_{\text{eff}}(\nu) := \gamma\,\kappa_2\Big[1 + \frac{2\delta}{\kappa_2}\gamma\,\Phi_1(\nu)\Big]. \tag{4.100}$$

Two important points follow directly from this definition. First, γ_{eff} is a function of ν, not a single number: it must be evaluated at ν_c for the threshold equation and separately at each ν_k when computing the cosine similarity. Second, whenever the bracketed correction term is negligible (as it is for small enough γ, or when δ is small relative to κ_2) γ_{eff} reduces to its

leading order,

$$\gamma_{\text{eff}} \approx \gamma\kappa_2 = \frac{N}{L_{\text{eff}}}, \tag{4.101}$$

recovering the effective aspect ratio in terms of the effective locus count L_{eff} defined above.

With $\gamma_{\text{eff}}(\nu)$ defined, both cases below reduce to the classical generalized spiked covariance threshold and overlap formulas of Bai and Yao (2012) under the substitution $\gamma \to \gamma_{\text{eff}}$:

$$\begin{aligned} &\gamma_{\text{eff}}(\nu_c)\,\Phi_2(\nu_c) = 1 && \text{(threshold)}, \\ &\langle u_k, \hat{u}_k\rangle^2 = \frac{1 - \gamma_{\text{eff}}(\nu_k)\,\Phi_2(\nu_k)}{1 + \gamma_{\text{eff}}(\nu_k)\,\Phi_1(\nu_k)} && \text{(cosine similarity)}. \end{aligned}$$

Case A: $\epsilon_a \equiv 0$. In the case where diversity is identical for all populations, such that $\epsilon_a = 0$ for all a, $\Phi_1(\nu) = 1/\nu$ and $\Phi_2(\nu) = 1/\nu^2$ exactly, so (4.100) becomes fully explicit:

$$\gamma_{\text{eff}}(\nu) = \gamma\kappa_2 + \frac{2\gamma^2\delta}{\nu}. \tag{4.102}$$

The threshold condition $\gamma_{\text{eff}}(\nu_c)/\nu_c^2 = 1$ reduces to the cubic equation

$$\nu_c^3 - \gamma\kappa_2\,\nu_c - 2\gamma^2\delta = 0, \tag{4.103}$$

which, by Descartes' rule of signs $(+,-,-)$, has exactly one positive real root. Writing

$$A = 2\sqrt{\gamma\kappa_2/3}, \qquad \Theta = \frac{3\sqrt{3}\,\delta\sqrt{\gamma}}{\kappa_2^{3/2}}, \tag{4.104}$$

the closed-form root is

$$\nu_c = \begin{cases} A\cos\big(\frac{1}{3}\arccos\Theta\big), & \Theta \le 1, \\ A\cosh\big(\frac{1}{3}\operatorname{arccosh}\Theta\big), & \Theta > 1. \end{cases} \tag{4.105}$$

The trigonometric expression is the exact root of the cubic, benchmarked against the exact joint solution of (4.29) and (4.30). The small-γ expansion underlying γ_{eff} achieves relative error in ν_c below 0.15% at $\gamma = 0.05$, rising to about 1% at $\gamma = 0.5$, accurate well beyond the strictly asymptotic regime $\gamma \to 0$ for which it was derived. The overlap in this case is

$$\langle u_k, \hat{u}_k\rangle^2 = \frac{1 - \gamma_{\text{eff}}(\nu_k)/\nu_k^2}{1 + \gamma_{\text{eff}}(\nu_k)/\nu_k}, \qquad \gamma_{\text{eff}}(\nu_k) = \gamma\kappa_2 + \frac{2\gamma^2\delta}{\nu_k}. \tag{4.106}$$

Case B: $\epsilon_a \neq 0$. No closed form is available; ν_c is obtained by fixed-point iteration on (4.100)–(4.33), which converges in about three passes:

$\gamma_{\text{eff}} \leftarrow \gamma\kappa_2$ [initialize]

repeat 3 times:

solve $\gamma_{\text{eff}}\,\Phi_2(\nu) = 1$ for ν [1D bisection, $\nu > \max_a \epsilon_a$]

$\gamma_{\text{eff}} \leftarrow \gamma\kappa_2\Big(1 + \dfrac{2\delta}{\kappa_2}\gamma\,\Phi_1(\nu)\Big)$

$\nu_c \leftarrow \nu$ (final value).

Across different values of ϵ_a ranging from 0 to $O(1)$, the three-iteration scheme tracks the exact critical point to under 1% relative error for $\gamma \in \{0.05, \ldots, 0.5\}$, with no visible degradation as ϵ_a grows.

The cosine similarity for a candidate structure eigenvalue ν_k requires its own $\gamma_{\text{eff}}(\nu_k)$, which is simply computed by plugging ν_k into equation 4.100. This is a single evaluation, not an iteration. The result is then substituted into the general overlap formula (4.34). Reusing

$\gamma_{\text{eff}}(\nu_c)$ from the threshold computation at this step is incorrect: the two quantities depend on ν through $\Phi_1(\nu)$ and are equal only when $\nu_k = \nu_c$, so substituting one for the other introduces a systematic error in the overlap that does not vanish as ν_k moves away from threshold.

Note: This appendix section was written with assistance from Claude Opus 5.5 and Claude Fable 5.1. The models were used to assist in understanding the theoretical framework and in presenting the mathematical results in a seamless manner. This was done through multiple rounds of manual writing coupled with AI-assisted writing and revision.

Appendix 4D: Deriving the structure eigenvalues for the demographic models under study

In the main text, we made use of the analytical structure eigenvalues ν_k for the following demographic models

1. $K = 2$ populations separated by a generic F_{ST}, with different diversities, and a general sampling scheme.

2. A demography with an arbitrary number of populations K, where two subpopulations are closely related with a small F_{ST}, under a general sampling scheme.

3. 1-dimensional stepping-stone model with homogeneous migration and even sampling.

4. 2-dimensional stepping-stone model with homogeneous migration and even sampling.

We derive the analytical expressions of the structure eigenvalues for these demographic models in this section. In each case, we need to write down the matrix $\hat{F}$ defined in equation 4.10 and find its eigenvalues as a function of demographic parameters.

K=2 populations

For $K = 2$ populations, there is a single structure eigenvalue ν, with the corresponding eigenvector u clustering the samples into two clusters. The structure eigenvalue ν can be written in terms of the set of parameters $(F_{ST}, \tau_1, \tau_2, N, \phi)$, which generically describe two populations that are separated by some value of F_{ST}, with within-population diversities given by τ_1 and τ_2. $(N\phi_1, N\phi_2)$ haplotypes are sampled from populations A and B respectively (with $\phi = \phi_1 = 1 - \phi_2$). In this case,

$$\psi = (\sqrt{\phi_1}, \sqrt{\phi_2})^T$$
$$\hat{P} = I - \psi\psi^T = \begin{pmatrix} \phi_2 & -\sqrt{\phi_1\phi_2} \\ -\sqrt{\phi_1\phi_2} & \phi_1 \end{pmatrix}$$
$$\tilde{F} = \begin{pmatrix} \epsilon_1 & -N\sqrt{\phi_1\phi_2}\dfrac{F_2}{\bar{\tau}} \\ -N\sqrt{\phi_1\phi_2}\dfrac{F_2}{\bar{\tau}} & \epsilon_2 \end{pmatrix}$$

Note that the structure eigenvector must be the vector orthogonal to ψ, since it is the only other eigenvector. The vector orthogonal to ψ is $u = (\sqrt{\phi_2}, -\sqrt{\phi_1})^T$. Therefore ν can be found by

$$\nu = u^T\hat{F}u = u^T(\hat{P}\tilde{F}\hat{P})u = u^T(uu^T)\tilde{F}(uu^T)u = u^T\tilde{F}u,$$

where we have used $\hat{P} = I - \psi\psi^T = uu^T$ and $u^Tu = 1$. Carrying out the matrix multiplication, the structure eigenvalue can be generically written as

$$\nu = 2N\phi_1\phi_2\frac{F_{ST}}{1 - F_{ST}}\frac{\tau_1 + \tau_2}{2(\phi_1\tau_1 + \phi_2\tau_2)} - \frac{(\phi_1 - \phi_2)(\tau_1 - \tau_2)}{\phi_1\tau_1 + \phi_2\tau_2} \tag{4.107}$$
$$= 2N\phi_1\phi_2\frac{F_{ST}}{1 - F_{ST}}(1 + \langle\epsilon\rangle) - (\phi_1 - \phi_2)(\epsilon_1 - \epsilon_2), \tag{4.108}$$

where we have used $\langle\epsilon\rangle = \dfrac{\epsilon_1+\epsilon_2}{2}$ (note that the sample-weighted average $\bar{\epsilon}$ is zero by construction). For $F_{ST} \ll 1$

$$\boxed{\nu(F_{ST},\epsilon_1,\epsilon_2,\phi_1,\phi_2) = N\tilde{\phi}F_{ST}(1+\langle\epsilon\rangle) - (\phi_1-\phi_2)(\epsilon_1-\epsilon_2)} \tag{4.109}$$

$\tilde{\phi} = \dfrac{2\phi_1\phi_2}{\phi_1+\phi_2}$ is the harmonic mean of the sample proportions (note that $\phi_1+\phi_2 = 1$ in this case).

Arbitrary K with two closely related subpopulations

Consider the case where samples are obtained from K subpopulations, two of which are separated by a small F_{ST}. Specifically, assume that the matrix of pairwise coalescent times can be written as

$$T^{\text{pop}} = \begin{pmatrix} \tau & \tau(1+\delta) & T_{13} & T_{14} & \cdots & T_{1K} \\ \tau(1+\delta) & \tau & T_{13} & T_{14} & \cdots & T_{1K} \\ T_{13} & T_{13} & \tau_3 & T_{34} & \cdots & T_{3K} \\ T_{14} & T_{14} & T_{34} & \tau_4 & \cdots & T_{4K} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ T_{1K} & T_{1K} & T_{3K} & T_{4K} & \cdots & \tau_K \end{pmatrix}. \tag{4.110}$$

This is a case where the two closely related populations have equal mean within-population coalescent times ($\tau_1 = \tau_2 = \tau$), with the mean coalescent time between them slightly larger ($T_{12} = \tau(1+\delta)$). Since they are closely related, we assume that $T_{1c} = T_{2c}$ for all $c \geq 3$. The rest of the demographic structure is generic. We propose the following ansatz for the eigenvector that separates these populations

$$u_1 = \frac{1}{\sqrt{s}}\left(\sqrt{\phi_2}, -\sqrt{\phi_1}, 0, \ldots, 0\right)^\top, \qquad s \equiv \phi_1+\phi_2, \tag{4.111}$$

which is orthogonal to ψ by construction, so that $\hat{P}u_1 = u_1$ and $u_1^T \hat{F} u_1 = u_1^T \tilde{F} u_1$. We first compute $\tilde{F}u_1$. Writing $g_{ab} = -N\sqrt{\phi_a \phi_b}\,(F_2)_{ab}/\bar{\tau}$ and $\epsilon_1 = \epsilon_2 \equiv \epsilon_0$, its entries are

$$
\begin{aligned}
(\tilde{F}u_1)_1 &= \epsilon_0 \sqrt{\phi_2/s} - g_{12}\sqrt{\phi_1/s}, \\
(\tilde{F}u_1)_2 &= g_{12}\sqrt{\phi_2/s} - \epsilon_0 \sqrt{\phi_1/s}, \\
(\tilde{F}u_1)_c &= g_{1c}\sqrt{\phi_2/s} - g_{2c}\sqrt{\phi_1/s}, \qquad c \geq 3.
\end{aligned}
$$

Since $\tau_1 = \tau_2 = \tau$ and $T_{1c} = T_{2c}$ for all $c \geq 3$,

$$
(F_2)_{1c} = T_{1c} - \frac{\tau + \tau_c}{2} = (F_2)_{2c} \quad \Rightarrow \quad g_{1c} = \sqrt{\phi_1/\phi_2}\, g_{2c},
$$

so that $(\tilde{F}u_1)_c = g_{2c}\sqrt{\phi_1/s} - g_{2c}\sqrt{\phi_1/s} = 0$ for all $c \geq 3$: $\tilde{F}u_1$ is supported on the first two coordinates, exactly like u_1. On those two coordinates, writing $f_{12} \equiv (F_2)_{12}/\bar{\tau}$ so that $g_{12} = -N\sqrt{\phi_1\phi_2}\, f_{12}$, and decomposing $\tilde{F}u_1$ into its components along u_1 and orthogonal to it,

$$
\tilde{F}u_1 = \nu_1\, u_1 + r, \qquad \nu_1 = u_1^T \tilde{F} u_1 = \epsilon_0 + \frac{2N\phi_1\phi_2}{s}\, f_{12}, \qquad r = N\sqrt{\phi_1\phi_2}\, f_{12}\, \frac{\phi_1 - \phi_2}{s^{3/2}} \left(\sqrt{\phi_1}, \sqrt{\phi_2}, 0, \ldots, 0\right. \tag{4.112}
$$

The residual r is proportional to $\phi_1 - \phi_2$. When the two sister populations are sampled in equal proportions, $r = 0$, u_1 is an exact eigenvector of $\tilde{F}$ and hence of $\hat{F}$ (since $\hat{F}u_1 = \hat{P}\tilde{F}\hat{P}u_1 = \hat{P}\tilde{F}u_1 = \nu_1 \hat{P}u_1 = \nu_1 u_1$), and ν_1 is exactly the corresponding structure eigenvalue. This is also the reason the $K = 2$ case above is exact for any ϕ: there $s = 1$ and $\psi^{\perp}$ is one-dimensional, so the residual has nowhere to go.

When $\phi_1 \neq \phi_2$, u_1 is no longer an exact eigenvector: $\hat{F}u_1 = \nu_1 u_1 + \hat{P}r$, with $\hat{P}r \neq 0$ since the vector $(\sqrt{\phi_1}, \sqrt{\phi_2}, 0, \ldots, 0)^{\top}$ is not orthogonal to ψ. To quantify the error in ν_1,

write $\hat{F} = \hat{F}_0 + E$ with

$$E = u_1(\hat{P}r)^\top + (\hat{P}r)\,u_1^\top, \qquad \hat{F}_0 = \hat{F} - E. \tag{4.113}$$

By construction $\hat{F}_0 u_1 = \nu_1 u_1$ exactly (since $u_1 \perp \hat{P}r$), so u_1 is an eigenvector of $\hat{F}_0$ with eigenvalue ν_1, and E is a symmetric perturbation with $\|E\| = \|\hat{P}r\|$. Let (ν_k, v_k), $k \neq 1$, be the remaining eigenpairs of $\hat{F}_0$, so that $\{u_1\} \cup \{v_k\}_{k\neq 1}$ is an orthonormal basis of $\psi^\perp$. Provided ν_1 is non-degenerate (does not coincide with any ν_k), standard second-order perturbation theory for symmetric matrices (Kato, 1995, Chapter II) gives

$$\nu_1^{\text{exact}} = \nu_1 + u_1^\top E u_1 + \sum_{k\neq 1} \frac{(v_k^\top E u_1)^2}{\nu_1 - \nu_k} + O(\|E\|^3). \tag{4.114}$$

The first-order term vanishes, $u_1^\top E u_1 = 2\,(u_1^\top \hat{P}r) = 0$, and in the second-order term $v_k^\top E u_1 = v_k^\top \hat{P}r$ because $v_k \perp u_1$. Bounding every denominator by the smallest gap and using Parseval's identity, $\sum_{k\neq 1}(v_k^\top \hat{P}r)^2 = \|\hat{P}r\|^2$ (as $u_1^\top \hat{P}r = 0$),

$$\left|\nu_1^{\text{exact}} - \nu_1\right| \leq \sum_{k\neq 1} \frac{(v_k^\top \hat{P}r)^2}{|\nu_1 - \nu_k|} + O(\|E\|^3) \leq \frac{\|\hat{P}r\|^2}{\min_{k\neq 1}|\nu_1 - \nu_k|} + O(\|E\|^3), \tag{4.115}$$

with

$$\|\hat{P}r\| = N\sqrt{\phi_1\phi_2}\, f_{12}\, |\phi_1 - \phi_2|\, \frac{\sqrt{1-s}}{s}, \tag{4.116}$$

where we used $\|\hat{P}(\sqrt{\phi_1}, \sqrt{\phi_2}, 0, \ldots)^\top\|^2 = s(1-s)$. When $\phi_1 = \phi_2$, $\hat{P}r = 0$ and the result of the previous paragraph is recovered exactly.

Relative to ν_1 itself, the residual has size $\|\hat{P}r\|/\nu_1 \approx |\phi_1 - \phi_2|\sqrt{1-s}/(2\sqrt{\phi_1\phi_2})$, independent of N and of F_{ST}, so the correction to the eigenvalue is controlled by the gap to the neighboring eigenvalues. In the setting of interest, the sister populations are separated by a

small F_{ST} while the remaining populations are separated by much larger ones, so $\nu_1 \ll \nu_k$ for the other structure eigenvalues and the correction is negligible. This is the regime in which the expression is used in the main text, and the agreement in figure 4.5, which includes unequal sampling proportions for the two sister populations, confirms it empirically.

Since F_{ST} between the two sister populations is small, $f_{12} = \dfrac{F_{ST}}{1 - F_{ST}} \cdot \dfrac{\tau}{\bar{\tau}} \approx \dfrac{\tau}{\bar{\tau}} F_{ST} = (1 + \epsilon_0) F_{ST}$, resulting in the expression

$$\boxed{\nu_1(F_{ST}, \epsilon_0, N, \tilde{\phi}) \approx \epsilon_0 + N\tilde{\phi} F_{ST}(1 + \epsilon_0)} \tag{4.117}$$

where $\tilde{\phi} = \dfrac{2\phi_1\phi_2}{\phi_1 + \phi_2}$ is again the harmonic mean of the two sample proportions.

1D stepping-stone model

To derive a closed-form expression of the eigenvalues of a spatially structured population, we consider the simplified stepping-stone model with homogeneous nearest-neighbor migration and even sampling.

Note that for the linear chain, within-population coalescent times are not identical: the edge demes have longer within-deme coalescent times than interior demes, and the deviation grows with K and with decreasing M (at $K = 25$, $\max_a |\epsilon_a| \approx 0.1$ for $M = 32$ and ≈ 0.4 for $M = 4$). What does hold exactly is that the unweighted mean of τ_a over demes equals $2N_eK$ (Strobeck, 1987), so under even sampling $\bar{\tau} = 2N_eK$ exactly, and the only approximation in setting $\tau_a \equiv \tau$ is dropping the diagonal $D_{\tilde{\epsilon}}$ of $\tilde{F}$. This introduces an error in ν_1 below 4% over the range of M used in figure 4.8, rising to about 13% at $K = 25$, $M = 4$. We therefore proceed with the formulation in section 4.3.1 and derive ν_k directly through the eigenvalues of the Laplacian. Defining $\bar{m} = m$, the scaled Laplacian of the migration matrix will have a

tri-diagonal Toeplitz structure

$$\Lambda = \begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{pmatrix}_{K\times K}$$

This is a second-difference matrix with Neumann/reflective boundary conditions. The eigenvectors of such a matrix are the columns of a DCT-II matrix (discrete cosine transform of second type) (Strang, 1999). Specifically, the kth eigenvector u_k and eigenvalue ℓ_k are given by

$$u_k(j) = \cos\left(\frac{k\pi}{K}(j+\frac{1}{2})\right), \quad j = 0, \ldots, K-1 \tag{4.118}$$

$$\ell_k = 4\sin^2\left(\frac{\pi k}{2K}\right), \quad k = 1, \ldots, K-1 \tag{4.119}$$

$$\ell_1 \approx \frac{\pi^2}{K^2} \quad \text{(large K limit)} \tag{4.120}$$

The eigenvectors being the columns of the discrete cosine transform agrees with the Novembre and Stephens (2008) findings. The first eigenvector exposes global spatial structure, so PCA will successfully reveal the continuous spatial structure if PC1 is properly estimated. What is highlighted more explicitly here are the eigenvalues. Invoking equation 4.18 the largest eigenvalue corresponding to the global spatial structure is given by

$$\boxed{\nu_1(n, M, K) = \frac{n}{4M\sin^2\left(\frac{\pi}{2K}\right)}} \tag{4.121}$$

It is also useful to find an expression for the pooled F_{ST} of the 1D stepping-stone model

in terms of M using (4.21). To do that, we need to compute the sum of the eigenvalues of the pseudo-inverse of the Laplacian.

$$\begin{aligned}
\sum_{k=1}^{K-1} \frac{1}{\ell_k} = \mathrm{tr}(\Lambda^+) &= \frac{1}{2K} \sum_{a,b=1}^{K} R_{ab} \\
&= \frac{1}{K} \sum_{1 \le a < b \le K} R_{ab} \\
&= \frac{1}{K} \sum_{1 \le a < b \le K} (b-a) \\
&= \frac{1}{K} \sum_{d=1}^{K-1} d\,(K-d) \\
&= \frac{1}{K} \left[K \sum_{d=1}^{K-1} d \; - \; \sum_{d=1}^{K-1} d^2 \right] \\
&= \frac{1}{K} \left[K \cdot \frac{(K-1)K}{2} \; - \; \frac{(K-1)K(2K-1)}{6} \right] \\
&= \frac{(K-1)K}{2} - \frac{(K-1)(2K-1)}{6} \\
&= (K-1) \left[\frac{3K - (2K-1)}{6} \right] \\
&= \frac{(K-1)(K+1)}{6} \\
&= \frac{K^2-1}{6}
\end{aligned}$$

Plugging into (4.21),

$$(F_{ST})^{1D} = \frac{K^2-1}{K^2-1+6MK} \tag{4.122}$$

$$\approx \frac{K}{K+6M} \quad (K \gg 1), \tag{4.123}$$

closely matching the expressions derived by Slatkin (1991) expressed in terms of our parameters.

2D stepping-stone model

Extending the 1D result to the 2D setting turns out to be simple. Consider the 2D stepping-stone model, where the habitat is divided into a $K \times K$ grid of demes, with each pair of neighboring demes exchanging migrants at a rate m. Note that demes in the center have 4 neighbors, while demes at the edges have 3 neighbors, and demes on the corners have 2 neighbors. The Laplacian of the 2D grid is now a $K^2 \times K^2$ matrix with diagonal terms of $(4, 3, 2)$ for interior, edge and corner nodes respectively, and off-diagonal terms of -1 for all pairs of nodes that are neighbors on the grid. Such a Laplacian can be conveniently written in terms of the Laplacian of the 1D chain. To see this, write the 2D Laplacian as a $K \times K$ block matrix, with block (a, b) representing the scaled migration rates between rows a and b.

$$
\Lambda_{2D} = \left(\begin{array}{c|ccccc}
 & \text{row}_1 & \text{row}_2 & \text{row}_3 & \cdots & \text{row}_K \\ \hline
\text{row}_1 & \Lambda_{1D}+I & -I & 0 & \cdots & 0 \\
\text{row}_2 & -I & \Lambda_{1D}+2I & -I & \cdots & 0 \\
\text{row}_3 & 0 & -I & \Lambda_{1D}+2I & \cdots & 0 \\
\vdots & \vdots & & & \ddots & -I \\
\text{row}_K & 0 & 0 & 0 & -I & \Lambda_{1D}+I
\end{array}\right)
$$

$$
= \begin{pmatrix}
\Lambda_{1D} & & & & \\
 & \Lambda_{1D} & & & \\
 & & \Lambda_{1D} & & \\
 & & & \ddots & \\
 & & & & \Lambda_{1D}
\end{pmatrix} + \begin{pmatrix}
I & -I & & & \\
-I & 2I & -I & & \\
 & -I & 2I & \ddots & \\
 & & \ddots & \ddots & -I \\
 & & & -I & I
\end{pmatrix}
$$

$$
= I_K \otimes \Lambda_{1D} + \Lambda_{1D} \otimes I_K
$$

Decomposing the 2D Laplacian into the Kronecker sum of two 1D Laplacians allows the eigenvalues and eigenvectors to be written directly in terms of those of the 1D Laplacian.

$$
u_{ks}(p,q) = \cos\left(\frac{k\pi}{K}(p+\frac{1}{2})\right)\cos\left(\frac{s\pi}{K}(q+\frac{1}{2})\right) \tag{4.124}
$$

$$
\ell_{ks} = 4\sin^2\left(\frac{k\pi}{2K}\right) + 4\sin^2\left(\frac{s\pi}{2K}\right). \tag{4.125}
$$

The first eigenvalue has a multiplicity of 2 ($\ell_{10} = \ell_{01}$), with the associated eigenvectors reflecting the 2D spatial arrangement of samples. The associated ν is identical to the 1D

case

$$\boxed{\nu_{10}(n, M, K) = \frac{n}{4M \sin^2\left(\frac{\pi}{2K}\right)}} \tag{4.126}$$

Defining a closed-form expression for the pooled F_{ST} in the 2D case is less straightforward since the double sum

$$\sum_{k=1}^{K^2-1} \frac{1}{\ell_k} = \sum_{\substack{j,l=0 \\ (j,l)\neq(0,0)}}^{K-1} \frac{1}{4\sin^2\left(\frac{\pi j}{2K}\right) + 4\sin^2\left(\frac{\pi l}{2K}\right)} \tag{4.127}$$

has no elementary closed form, unlike its 1D counterpart. However, for large K, this sum is known to grow logarithmically rather than polynomially in K (remember that K^2 is the total number of demes). Building on the identity relating the Kirchhoff index to the trace of the Laplacian pseudo-inverse (Gutman & Mohar, 1996), Ecevit and Boysal (2025) rigorously established the leading-order asymptotic behavior of the Kirchhoff index of the square lattice under free boundary conditions, correcting several earlier erroneous claims of a finite limit for this quantity (Liu & Pan, 2015; Liu et al., 2014; Ye, 2011). Translated to our notation, their result gives

$$\sum_{k=1}^{K^2-1} \frac{1}{\ell_k} = \frac{K^2 \ln K}{2\pi}\left(1 + O\left(\frac{1}{\ln K}\right)\right) \qquad (K \gg 1), \tag{4.128}$$

in which case the F_{ST} has the form

$$(F_{ST})^{2D} \approx \frac{\ln K}{\ln K + 2\pi M} \qquad (K \gg 1), \tag{4.129}$$

matching the insight that population differentiation grows logarithmically with habitat size in 2D, whereas growth is linear in 1D (Cox & Durrett, 2002; Maruyama, 1970). Note that the

$O(1/\ln K)$ correction is not negligible for the grid sizes used in our simulations: numerically, $\sum_k 1/\ell_k \approx K^2\left(\frac{\ln K}{2\pi} + 0.325\right)$, and at $K = 9$ the constant is as large as the leading term, so the leading-order expression underestimates the pooled F_{ST} of a 9×9 grid by roughly a factor of 2. The expression should be read as a statement of the asymptotic scaling with K; for a specific grid, F_{ST} is computed from the exact eigenvalue sum through equation (4.21).

Expected coalescent times under the non-equilibrium stepping-stone model

For the demography of figure 4.12A, the matrix T^{pop} of expected pairwise coalescent times between the K^2 demes of population B and the panmictic population A cannot be obtained from the resistance-distance approximation, which assumes that migration has been ongoing indefinitely. We compute it instead directly from the coalescent process of two lineages, which is a continuous-time Markov chain whose state is the unordered pair of populations the two lineages occupy. During the structured phase, a lineage in deme i of the grid moves to a neighboring deme j at rate m_{ij}; two lineages in the same deme coalesce at rate $1/(2N_e^{\text{deme}})$, two lineages in A coalesce at rate $1/(2N_A)$, and a lineage in A cannot coalesce with a lineage in B. Let Q be the sub-generator of this chain on the non-absorbing states, with Q_{ij} the rate of transition from state i to state j for $i \neq j$ and Q_{ii} minus the total rate out of state i, including coalescence, so that coalescence is the absorbing event. Let τ denote the coalescence time of a pair. Since the chain is time-homogeneous during the structured phase, the survival function of τ for every starting state is obtained at once as

$$\Pr(\tau > s) = e^{Qs}\,\mathbf{1}, \qquad 0 \le s \le t, \tag{4.130}$$

where the entries of the resulting vector are indexed by the starting pair of populations. A pair that has not coalesced by the time of the split finds itself in the ancestral panmictic population of size N_{anc}, where it coalesces after a further $2N_{\text{anc}}$ generations on average.

Since $\int_0^t \Pr(\tau > s)\, ds = \mathbb{E}[\min(\tau, t)]$ already accounts for the t generations elapsed by pairs that survive to the split, the expected coalescent time for a pair starting in state $\{a, b\}$ is

$$T_{ab}^{\mathrm{pop}} = \int_0^t \Pr(\tau > s)\, ds \;+\; 2N_{\mathrm{anc}}\, \Pr(\tau > t), \tag{4.131}$$

which we evaluate numerically by computing $e^{Qs}\mathbf{1}$ on a fine grid of $s \in [0, t]$ with a sparse matrix exponential and integrating by quadrature. The equilibrium reference used in figure 4.13D is obtained from the same chain restricted to population B with no split, for which $T^{\mathrm{pop}} = -Q^{-1}\mathbf{1}$. We verified the implementation against `msprime`'s exact `mean_coalescence_time` for representative pairs (agreement to within 0.2%, the expected discrete- versus continuous-generation offset). The resulting T^{pop} is then converted to $\hat{F}$ through equations (4.9) and (4.10), with the sampling proportions of the experiment, and its eigenvalues and eigenvectors are computed numerically.

CHAPTER 5

CONCLUSION

A major theme that motivated the work presented in this dissertation is the complex nature of the historical processes that give rise to observed genetic variation, and how much this observed variation can tell us about such processes. One common simplification that we particularly address is that of describing history in terms of population units. In an effort to offer alternative representations of history that are agnostic to such population labels, we develop CoaSTT, which provides a low-dimensional representation of coalescent structure through time in the form of three-dimensional plots, where the third dimension depicts time. A key computational innovation that allowed for the development of CoaSTT is ARG inference, which provides an avenue for estimating the coalescent rates between each pair of haplotypes or individuals without the need to partition them into groups.

CoaSTT is presented within the broader context of latent position models (Hoff et al., 2002), which is a class of methods in network science used to derive low-dimensional representations of complex network interactions. Additionally, we have shown parallels between the loss minimized by CoaSTT and that of UMAP (McInnes et al., 2020). The class of methods to which CoaSTT belongs is different from the class of matrix factorization methods to which PCA and STRUCTURE belong (Engelhardt & Stephens, 2010), which are more popular in population genetics. In chapter 2, we derive a loss function motivated by the likelihood of observing a given number of coalescent events between a pair of samples within some time slice. Minimizing the loss function with respect to the latent position of samples over time yields smooth latent trajectories that can be presented as intuitive CoaSTT plots. At a given time slice, the distance between any two samples in the latent space is related to their coalescent rate through a "link function", where the closer the samples are, the higher their coalescent rate is. We benchmark CoaSTT against a variety of simple and complex demographic scenarios, providing more guidance on how CoaSTT plots should be

interpreted. Importantly, we emphasize how CoaSTT provides a representation of coalescent history rather than demographic history. Finally, we present an extension to CoaSTT which takes as an input coalescent rates estimated from particular genomic regions, such as those assigned a specific ancestry, rather than the whole genome, which is useful in visualizing alternative coalescent histories of admixed individuals.

At the beginning of chapter 3, we grapple with the complexity of real data, opening the chapter with an analysis of the bias associated with coalescent rate estimates obtained from state-of-the-art ARG inference methods. This analysis provides us with guidelines to avoid misinterpretation of CoaSTT embeddings obtained from ARGs inferred using current methods, especially in deep time. We additionally demonstrate the utility of comparing CoaSTT embeddings obtained under different initializations, and using different posterior ARG samples. We apply CoaSTT to ARGs inferred for human and chimpanzee samples, reproducing numerous known signals, and identifying a few previously unreported signals for chimpanzees. The coalescent-based approach to population history allows us to propose more nuanced hypotheses about the ancestral history of the Nigeria-Cameroon chimpanzees, which can be further refined using model-based approaches.

CoaSTT can be regarded as a step towards a more general direction of individual-based depiction of genetic history. A method like CoaSTT does not strictly require inference of the full ARG, as estimates of distributions of pairwise coalescent times would suffice. A recent study presented a transformer-based deep learning method for estimating pairwise coalescent times without reconstructing ARGs (Korfmann et al., 2026), suggesting that developments in this direction are promising. An essential limitation of ARG-based estimates of coalescent rates is the over-smoothing of signals, resulting in a bias that can be mistaken for early gene flow or sustained deep structure. This bias is a consequence of the panmictic prior used in the topology reconstruction step of ARG inference, such that genomic segments that do not carry sufficient mutations contribute a panmictic signal to the coalescent time distribution.

Methods of dating the ARG (assigning branch lengths to topologies that have been inferred) have been developed in a manner that avoids such a panmictic prior (Deng, DeWitt, et al., 2025; Pope et al., 2026). Further work is needed to mitigate the effect of the panmictic prior on the topology inference step. A potential Empirical-Bayes approach could make use of estimates of pairwise coalescent rates obtained after an initial round of ARG inference to improve the estimates obtained from subsequent rounds, which is an extended version of what Relate already does (Speidel et al., 2019).

In the current version of CoaSTT, a single latent position is computed for each sample. Samples that are regarded as admixed could be handled through RS-CoaSTT. However, a more faithful representation of the coalescent history of such samples would estimate multiple latent trajectories for them in the same embedding. One way to do that is to compute latent positions for every single ancestral lineage that is extant at a given time slice. Since the number of such extant lineages is expected to explode for intermediate time points, the problem can be simplified by assigning each lineage ancestral to a sample one of K positions (where each ancestral lineage can be associated with multiple samples). The result would be each sample represented by K latent positions at each time point rather than a single latent position. In the context of latent space models, this is similar to latent position cluster models (Handcock et al., 2007). Alternatively, matrix factorization methods can be used to factorize the matrix of pairwise coalescent rates into K components, which can be further projected into a two-dimensional space. An approach in the spirit of the latter suggestion is currently being explored by Deng and Pritchard (unpublished).

Whereas the main aim of chapters 2 and 3 is the development of a method to understand population history, chapter 4 looks more closely into the limitations of a classical method, PCA, in revealing population structure. Using methods from random matrix theory, we derive approximate expressions for how strong population structure must be (as measured by a demographic parameter) for PCA to accurately recover the underlying axis of variation in a

genotype matrix of a given size, and we verify these expressions through simulations. These limits of PCA in revealing population structure have long been documented (Patterson et al., 2006), but only for simple demographic scenarios and with an emphasis on the separation of eigenvalues rather than the reliability of estimated eigenvectors (the observable PCs). This chapter extends the analysis to more general patterns of population structure, and shifts the emphasis towards assessing the reliability of observed PCs, emphasizing that detectability does not necessarily entail reliability. Low-dimensional summaries of population structure are known to have their own caveats related to interpretation, which were explored in the context of PCA (G. McVean, 2009; Novembre & Stephens, 2008), and explored in chapters 2 and 3 in the context of CoaSTT. Chapter 4 addresses a more fundamental question: when is a signal of population structure recoverable by PCA at all?

The extent to which genetic data can be informative about population history is interesting in its own right. Every method that infers some aspect of population structure or demographic history uses some summary statistic derived from observed genetic variation (e.g., estimated pairwise coalescent rates in the case of CoaSTT). Different summary statistics set different bounds on the aspects of history that are recoverable (R. N. Gutenkunst, 2026). A more foundational issue that motivated my projects is the prevalent description of history in terms of population units. The problem is not simply that this depiction is inaccurate; all mathematical and statistical models have a level of inaccuracy that is acceptable in exchange for the utility or explanatory power they provide. This inaccuracy becomes less of a nuisance when caveats associated with these models are clearly acknowledged and emphasized. That said, the dominant paradigm of "population units" seems so integral to depictions of population structure and history that it is difficult to conceive of alternative descriptions that do not rely on it, either methodologically or interpretively (Coop, 2023). Even common interpretations of methods that do not enforce this depiction a priori, such as PCA and STRUCTURE, are prone to drift towards it. One example that resonates with

me personally is a recent analysis of 1024 Egyptian genomes using PCA and STRUCTURE, where the authors argue for the existence of an Egyptian-enriched "genetic component" without an adequate discussion of what this specifically means. This result was released to the public as an official statement by the Ministry of Higher Education and Scientific Research, celebrating the discovery of "a *distinct* genetic component among Egyptians amounting to 18.5%, which helps specialists understand the *unique* hereditary makeup of the Egyptian people" (Ministry of Higher Education and Scientific Research, Egypt, 2026). Notably, this "distinct component" is quite variable within Egyptian samples, and substantially present in Middle Eastern, African, and even admixed American samples (likely pointing to a component that is generally enriched in North Africans and commonly found in such analyses). This statement fueled a wave of hyper-nationalist news articles echoing the "uniqueness" of the genetic composition of the Egyptian population and the continuity of this uniqueness all the way back to ancient Egyptians, a claim the study itself did not make. This anecdote is one instance of a broader societal tendency for such results to be misinterpreted in different ways (Lewis et al., 2022). Such forms of misinterpretation are quite insidious, risking the perpetuation of ideologies that inflict tangible, material harm. In recent years, serious discussions of these questions among geneticists, sociologists, epidemiologists, anthropologists, and historians have gained prominence (National Academies of Sciences, Engineering, and Medicine, 2023). As a population geneticist, I am eager to be part of these discussions and to clarify how the methods I develop or use should and should not be interpreted. I also hope to contribute to a new generation of methods for understanding population structure and history that do not oversimplify the inherent complexity underlying the patterns of genetic similarity observed in humans and other species.

REFERENCES

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-Del Vecchyo, D., … Kern, A. D. (2020a). A community-maintained standard library of population genetic models (G. Coop, P. J. Wittkopp, J. Novembre, A. Sethuraman, & S. Mathieson, Eds.). *eLife*, *9*, e54967. https://doi.org/10.7554/eLife.54967

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-Del Vecchyo, D., … Kern, A. D. (2020b). A community-maintained standard library of population genetic models (G. Coop, P. J. Wittkopp, J. Novembre, A. Sethuraman, & S. Mathieson, Eds.). *eLife*, *9*, e54967. https://doi.org/10.7554/eLife.54967

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, *9*, 1981–2014.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Springer.

Anderson-Trocmé, L., Nelson, D., Zabad, S., Diaz-Papkovich, A., Kryukov, I., Baya, N., Touvier, M., Jeffery, B., Dina, C., Vézina, H., Kelleher, J., & Gravel, S. (2023). On the genes, genealogies, and geographies of Quebec. *Science*, *380*(6647), 849–855. https://doi.org/10.1126/science.add5300

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Ayub, Q., Mezzavilla, M., Pagani, L., Haber, M., Mohyuddin, A., Khaliq, S., Mehdi, S. Q., & Tyler-Smith, C. (2015). The Kalash Genetic Isolate: Ancient Divergence, Drift, and Selection. *The American Journal of Human Genetics*, *96*(5), 775–783. https://doi.org/10.1016/j.ajhg.2015.03.012

Bai, Z., & Yao, J. (2012). On sample eigenvalues in a generalized spiked population model. *Journal of Multivariate Analysis*, *106*, 167–177. https://doi.org/10.1016/j.jmva.2011.10.009

Baik, J., Arous, G. B., & Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, *33*(5), 1643–1697. https://doi.org/10.1214/009117905000000233

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B.,

Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., ... Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229. https://doi.org/10.1093/genetics/iyab229

Becquet, C., Patterson, N., Stone, A. C., Przeworski, M., & Reich, D. (2007). Genetic Structure of Chimpanzee Populations. *PLOS Genetics*, *3*(4), e66. https://doi.org/10.1371/journal.pgen.0030066

Beerli, P., & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, *98*(8), 4563–4568. https://doi.org/10.1073/pnas.081068098

Beichman, A. C., Huerta-Sanchez, E., & Lohmueller, K. E. (2018). Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. *Annual Review of Ecology, Evolution, and Systematics*, *49*, 433–456. https://doi.org/10.1146/annurev-ecolsys-110617-062431

Belkin, M., & Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, *15*(6), 1373–1396. https://doi.org/10.1162/089976603321780317

Benaych-Georges, F., & Nadakuditi, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, *227*(1), 494–521. https://doi.org/10.1016/j.aim.2011.02.007

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., ... Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484), eaay5012. https://doi.org/10.1126/science.aay5012

Bergström, A., Oppenheimer, S. J., Mentzer, A. J., Auckland, K., Robson, K., Attenborough, R., Alpers, M. P., Koki, G., Pomat, W., Siba, P., Xue, Y., Sandhu, M. S., & Tyler-Smith, C. (2017). A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. *Science*, *357*(6356), 1160–1163. https://doi.org/10.1126/science.aan3842

Betancourt, B., Rodríguez, A., & Boyd, N. (2017). Bayesian Fused Lasso regression for dynamic binary networks. *Journal of Computational and Graphical Statistics*, *26*(4), 840–850. https://doi.org/10.1080/10618600.2017.1341323

Bhaskar, A., Wang, Y. R., & Song, Y. S. (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, *25*(2), 268–279. https://doi.org/10.1101/gr.178756.114

Bleakley, K., & Vert, J.-P. (2011, June 21). *The group fused Lasso for multiple change-point detection*. https://doi.org/10.48550/arXiv.1106.4199

Bloemendal, A., Knowles, A., Yau, H.-T., & Yin, J. (2016). On the principal components of sample covariance matrices. *Probability Theory and Related Fields*, *164*(1–2), 459–552. https://doi.org/10.1007/s00440-015-0616-x

Böhm, J. N., Berens, P., & Kobak, D. (2022). Attraction-repulsion spectrum in neighbor embeddings. *The Journal of Machine Learning Research*, *23*(1), 95:4118–95:4149. Retrieved July 2, 2026, from https://dl.acm.org/doi/10.5555/3586589.3586684

Bowden, R., MacFie, T. S., Myers, S., Hellenthal, G., Nerrienet, E., Bontrop, R. E., Freeman, C., Donnelly, P., & Mundy, N. I. (2012). Genomic Tools for Evolution and Conservation in the Chimpanzee: Pan troglodytes ellioti Is a Genetically Distinct Population. *PLOS Genetics*, *8*(3), e1002504. https://doi.org/10.1371/journal.pgen.1002504

Brand, C. M., White, F. J., Rogers, A. R., & Webster, T. H. (2022). Estimating bonobo (Pan paniscus) and chimpanzee (Pan troglodytes) evolutionary history from nucleotide site patterns. *Proceedings of the National Academy of Sciences*, *119*(17), e2200858119. https://doi.org/10.1073/pnas.2200858119

Brandt, D. Y. C., Huber, C. D., Chiang, C. W. K., & Ortega-Del Vecchyo, D. (2024). The Promise of Inferring the Past Using the Ancestral Recombination Graph. *Genome Biology and Evolution*, *16*(2), evae005. https://doi.org/10.1093/gbe/evae005

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G., & Bustamante, C. D. (2012). PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Human Biology*, *84*(4), 343–364. https://doi.org/10.3378/027.084.0401

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, *108*(10), 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Browning, S. R., Grinde, K., Plantinga, A., Gogarten, S. M., Stilp, A. M., Kaplan, R. C., Avilés-Santa, M. L., Browning, B. L., & Laurie, C. C. (2016). Local Ancestry Inference in a Large US-Based Hispanic/Latino Study: Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *G3 Genes|Genomes|Genetics*, *6*(6), 1525–1534. https://doi.org/10.1534/g3.116.028779

Bryc, K., Bryc, W., & Silverstein, J. W. (2013). Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. *Theoretical Population Biology*, *89*, 34–43. https://doi.org/10.1016/j.tpb.2013.08.004

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. https://doi.org/10.1137/0916069

Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., … Cavalli-Sforza, L. L. (2002). A Human Genome Diversity Cell Line Panel. *Science*, *296*(5566), 261–262. https://doi.org/10.1126/science.296.5566.261b

Carreira-Perpiñan, M. Á. (2010). The elastic embedding algorithm for dimensionality reduction. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 167–174.

Casella, G., Ghosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411. https://doi.org/10.1214/10-BA607

Caswell, J. L., Mallick, S., Richter, D. J., Neubauer, J., Schirmer, C., Gnerre, S., & Reich, D. (2008). Analysis of Chimpanzee History Based on Genome Sequence Alignments. *PLOS Genetics*, *4*(4), e1000057. https://doi.org/10.1371/journal.pgen.1000057

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1994). *The History and Geography of Human Genes: Abridged paperback Edition*. https://doi.org/10.2307/j.ctv301gjp

Chandra, A. K., Raghavan, P., Ruzzo, W. L., Smolensky, R., & Tiwari, P. (1996). The electrical resistance of a graph captures its commute and cover times. *computational complexity*, *6*(4), 312–340. https://doi.org/10.1007/BF01270385

Chen, J., Zheng, H., Bei, J.-X., Sun, L., Jia, W.-h., Li, T., Zhang, F., Seielstad, M., Zeng, Y.-X., Zhang, X., & Liu, J. (2009). Genetic Structure of the Han Chinese Population Revealed by Genome-wide SNP Variation. *The American Journal of Human Genetics*, *85*(6), 775–785. https://doi.org/10.1016/j.ajhg.2009.10.016

Chiang, C. W. K., Marcus, J. H., Sidore, C., Biddanda, A., Al-Asadi, H., Zoledziewska, M., Pitzalis, M., Busonero, F., Maschio, A., Pistis, G., Steri, M., Angius, A., Lohmueller, K. E., Abecasis, G. R., Schlessinger, D., Cucca, F., & Novembre, J. (2018). Genomic history of the Sardinian population. *Nature Genetics*, *50*(10), 1426–1434. https://doi.org/10.1038/s41588-018-0215-8

Chiu, G. S., & Westveld, A. H. (2011). A unifying approach for food webs, phylogeny, social networks, and statistics. *Proceedings of the National Academy of Sciences*, *108*(38), 15881–15886. https://doi.org/10.1073/pnas.1015359108

Chu, A. M. Y., Chan, T. W. C., So, M. K. P., & Wong, W.-K. (2021). Dynamic Network Analysis of COVID-19 with a Latent Pandemic Space Model. *International Journal of Environmental Research and Public Health*, *18*(6), 3195. https://doi.org/10.3390/ijerph18063195

Coop, G. (2023, January 7). *Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics*. https://doi.org/10.48550/arXiv.2207.11595

Cox, J. T., & Durrett, R. (2002). The stepping stone model: New formulas expose old myths. *Annals of Applied Probability*, *12*(4), 1348–1377. https://doi.org/10.1214/aoap/1037125866

Damrich, S., & Hamprecht, F. (2021). On UMAP' s True Loss Function. *Advances in Neural Information Processing Systems*, *34*, 5798–5809. Retrieved July 4, 2026, from https://proceedings.neurips.cc/paper/2021/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html

de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., Schmidt, J. M., Heredia-Genestar, J. M., Benazzo, A., Barbujani, G., Peter, B. M., Kuderna, L. F. K., Casals, F., Angedakin, S., Arandjelovic, M., … Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, *354*(6311), 477–481. https://doi.org/10.1126/science.aag2602

Degnan, J. H., & Rosenberg, N. A. (2006). Discordance of Species Trees with Their Most Likely Gene Trees. *PLOS Genetics*, *2*(5), e68. https://doi.org/10.1371/journal.pgen.0020068

Deng, Y., DeWitt, W. S., Song, Y. S., & Nielsen, R. (2025). A general framework for branch length estimation in Ancestral Recombination Graphs. *Proceedings of the National Academy of Sciences*, *122*(48), e2504461122. https://doi.org/10.1073/pnas.2504461122

Deng, Y., Nielsen, R., & Song, Y. S. (2025). Robust and accurate Bayesian inference of genome-wide genealogies for hundreds of genomes. *Nature Genetics*, *57*(9), 2124–2135. https://doi.org/10.1038/s41588-025-02317-9

Dey, R., & Lee, S. (2019). Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model. *Journal of Multivariate Analysis*, *173*, 145–164. https://doi.org/10.1016/j.jmva.2019.02.007

Ding, X., Zou, Z., & Brooks III, C. L. (2019). Deciphering protein evolution and fitness landscapes with latent space models. *Nature Communications*, *10*(1), 5644. https://doi.org/10.1038/s41467-019-13633-0

Ding, X., & Yang, F. (2021). Spiked separable covariance matrices and principal components. *The Annals of Statistics*, *49*(2), 1113–1138. https://doi.org/10.1214/20-AOS1995

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, *23*(1), 34–63. https://doi.org/10.1016/0040-5809(83)90004-7

Donoho, D., Gavish, M., & Johnstone, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, *46*(4), 1742–1778. https://doi.org/10.1214/17-AOS1601

Draganov, A., Jørgensen, J., Scheel, K., Mottin, D., Assent, I., Berry, T., & Aslay, C. (2023). ActUp: Analyzing and Consolidating tSNE and UMAP. *4*, 3651–3658. https://doi.org/10.24963/ijcai.2023/406

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, *22*(5), 1185–1192. https://doi.org/10.1093/molbev/msi103

Durante, D., & Dunson, D. B. (2014). Nonparametric Bayes dynamic modelling of relational data. *Biometrika*, *101*(4), 883–898. Retrieved June 25, 2026, from https://www.jstor.org/stable/43304694

Durante, D., & Dunson, D. B. (2016). Locally adaptive dynamic networks. *The Annals of Applied Statistics*, *10*(4), 2203–2232. https://doi.org/10.1214/16-AOAS971

Ecevit, F., & Boysal, A. (2025). Asymptotic behavior of the Kirchhoff index for square lattice with boundary conditions. *Turkish Journal of Mathematics*, *49*(1), 18–28. https://doi.org/10.55730/1300-0098.3571

Edwards, A. W. F., & Cavalli-Sforza, L. L. (1965). A Method for Cluster Analysis. *Biometrics*, *21*(2), 362–375. https://doi.org/10.2307/2528096

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Efron, B., & Tibshirani, R. J. (1994, May 14). *An Introduction to the Bootstrap*. https://doi.org/10.1201/9780429246593

Engelhardt, B. E., & Stephens, M. (2010). Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. *PLOS Genetics*, *6*(9), e1001117. https://doi.org/10.1371/journal.pgen.1001117

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). Fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics*, *37*(24), 4882–4885. https://doi.org/10.1093/bioinformatics/btab468

Fan, C., Mancuso, N., & Chiang, C. W. K. (2022). A genealogical estimate of genetic relationships. *The American Journal of Human Genetics*, *109*(5), 812–824. https://doi.org/10.1016/j.ajhg.2022.03.016

Fan, S., Kelly, D. E., Beltrame, M. H., Hansen, M. E. B., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T., Omar, S. A., Meskel, D. W., Belay, G., Froment, A., Patterson, N., Reich, D., & Tishkoff, S. A. (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biology*, *20*(1), 82. https://doi.org/10.1186/s13059-019-1679-2

Flores-Bello, A., Bauduer, F., Salaberria, J., Oyharçabal, B., Calafell, F., Bertranpetit, J., Quintana-Murci, L., & Comas, D. (2021). Genetic origins, singularity, and heterogeneity of Basques. *Current Biology*, *31*(10), 2167–2177. https://doi.org/10.1016/j.cub.2021.03.010

Fontsere, C., Kuhlwilm, M., Morcillo-Suarez, C., Alvarez-Estape, M., Lester, J. D., Gratton, P., Schmidt, J. M., Dieguez, P., Aebischer, T., Álvarez-Varona, P., Agbor, A., Angedakin, S., Assumang, A. K., Ayimisin, E. A., Bailey, E., Barubiyo, D., Bessone, M., Carretero-Alonso, A., Chancellor, R., … Marques-Bonet, T. (2022). Population dynamics and genetic connectivity in recent chimpanzee history. *Cell Genomics*, *2*(6), 100133. https://doi.org/10.1016/j.xgen.2022.100133

Frank, O., & Strauss, D. (1986). Markov Graphs. *Journal of the American Statistical Association*, *81*(395), 832–842. https://doi.org/10.1080/01621459.1986.10478342

Goddard, M. E. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*, *136*, 245–257. https://doi.org/10.1007/s10709-008-9308-0

Gonder, M. K., Oates, J. F., Disotell, T. R., Forstner, M. R. J., Morales, J. C., & Melnick, D. J. (1997). A new west African chimpanzee subspecies? *Nature*, *388*(6640), 337–337. https://doi.org/10.1038/41005

Gonder, M. K., Locatelli, S., Ghobrial, L., Mitchell, M. W., Kujawski, J. T., Lankester, F. J., Stewart, C.-B., & Tishkoff, S. A. (2011). Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *Proceedings of the National Academy of Sciences*, *108*(12), 4766–4771. https://doi.org/10.1073/pnas.1015422108

Greenstadt, J. (1967). On the relative efficiencies of gradient methods. *Mathematics of Computation*, *21*(99), 360–367. https://doi.org/10.1090/S0025-5718-1967-0223073-7

Griffiths, R. C., & Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, *3*(4), 479–502. https://doi.org/10.1089/cmb.1996.3.479

Griffiths, R. C., & Marjoram, P. (1997). An ancestral recombination graph. In P. Donnelly & S. Tavaré (Eds.), *Progress in population genetics and human evolution* (pp. 257–270, Vol. 87). Springer. https://doi.org/10.1007/978-1-4757-2609-1_16

Grundler, M. C., Terhorst, J., & Bradburd, G. S. (2025). A geographic history of human genetic ancestry. *Science*, *387*(6741), 1391–1397. https://doi.org/10.1126/science.adp4642

Gunnarsson, Á. F., Zhu, J., Zhang, B. C., Tsangalidou, Z., Allmont, A., & Palamara, P. F. (2024, September 2). *A scalable approach for genome-wide inference of ancestral recombination graphs*. https://doi.org/10.1101/2024.08.31.610248

Gutenkunst, R., Hernandez, R., Williamson, S., & Bustamante, C. (2010). Diffusion Approximations for Demographic Inference: DaDi. *Nature Precedings*, 1–1. https://doi.org/10.1038/npre.2010.4594.1

Gutenkunst, R. N. (2026, August 27). *Two blind spots in the demographic inference of human origins from genomic data*. https://doi.org/10.48550/arXiv.2608.27591

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Gutman, I., & Mohar, B. (1996). The quasi-Wiener and the Kirchhoff indices coincide. *Journal of Chemical Information and Computer Sciences*, *36*(5), 982–985. https://doi.org/10.1021/ci960007t

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(2), 301–354. https://doi.org/10.1111/j.1467-985X.2007.00471.x

Harris, K. (2015). Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences*, *112*(11), 3439–3444. https://doi.org/10.1073/pnas.1418652112

Harris, K., & Pritchard, J. K. (2017). Rapid evolution of the human mutation spectrum. *eLife*, *6*, e24284. https://doi.org/10.7554/eLife.24284

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A Genetic Atlas of Human Admixture History. *Science*, *343*(6172), 747–751. https://doi.org/10.1126/science.1243518

Hey, J. (2010). The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. *Molecular Biology and Evolution*, *27*(4), 921–933. https://doi.org/10.1093/molbev/msp298

Hey, J., & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics*, *167*(2), 747–760. https://doi.org/10.1534/genetics.103.024182

Hinton, G. E., & Roweis, S. (2002). Stochastic Neighbor Embedding. *Advances in Neural Information Processing Systems*, *15*. Retrieved July 2, 2026, from https://papers.nips.cc/paper_files/paper/2002/hash/6150ccc6069bea6b5716254057a194ef-Abstract.html

Hoff, P. D. (2005). Bilinear Mixed-Effects Models for Dyadic Data. *Journal of the American Statistical Association*, *100*, 286–295. Retrieved June 25, 2026, from https://ideas.repec.org//a/bes/jnlasa/v100y2005p286-295.html

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S., & Delaneau, O. (2023). Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nature Genetics*, *55*(7), 1243–1249. https://doi.org/10.1038/s41588-023-01415-w

Hu, W., Hao, Z., Du, P., Di Vincenzo, F., Manzi, G., Cui, J., Fu, Y.-X., Pan, Y.-H., & Li, H. (2023). Genomic inference of a severe human bottleneck during the Early to Middle Pleistocene transition. *Science*, *381*(6661), 979–984. https://doi.org/10.1126/science.abq7487

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, *23*(2), 183–201. https://doi.org/10.1016/0040-5809(83)90013-8

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., … Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, *451*(7181), 998–1003. https://doi.org/10.1038/nature06742

Johnstone, I. M., & Paul, D. (2018). PCA in High Dimensions: An Orientation. *Proceedings of the IEEE*, *106*(8), 1277–1292. https://doi.org/10.1109/JPROC.2018.2846730

Jouganous, J., Long, W., Ragsdale, A. P., & Gravel, S. (2017). Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. *Genetics*, *206*(3), 1549–1567. https://doi.org/10.1534/genetics.117.200493

Kamm, J., Terhorst, J., Durbin, R., & Song, Y. S. (2020). Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. *Journal of the American Statistical Association*, *115*(531), 1472–1487. https://doi.org/10.1080/01621459.2019.1635482

Kamm, J. A., Terhorst, J., & Song, Y. S. (2017). Efficient Computation of the Joint Sample Frequency Spectra for Multiple Populations. *Journal of Computational and Graphical Statistics*, *26*(1), 182–194. https://doi.org/10.1080/10618600.2016.1159212

Kato, T. (1995). *Perturbation theory for linear operators* [Reprint of the 1980 2nd edition]. Springer.

Kaufman, L., & Rousseeuw, P. (1990). Partitioning Around Medoids (Program PAM). In *Finding Groups in Data* (pp. 68–125). https://doi.org/10.1002/9780470316801.ch2

Kaur, H., Rastelli, R., Friel, N., & Raftery, A. E. (2023, April 6). *Latent Position Network Models*. https://doi.org/10.48550/arXiv.2304.02979

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, *12*(5), e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Kelleher, J., Thornton, K. R., Ashander, J., & Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, *14*(11), e1006581. https://doi.org/10.1371/journal.pcbi.1006581

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, *51*(9), 1330–1338. https://doi.org/10.1038/s41588-019-0483-y

Kim, J., Mossel, E., Rácz, M. Z., & Ross, N. (2015). Can one hear the shape of a population history? *Theoretical Population Biology*, *100*, 26–38. https://doi.org/10.1016/j.tpb.2014.12.002

Kim, S.-J., Koh, K., Boyd, S., & Gorinevsky, D. (2009). $\Ell_1$ Trend Filtering. *SIAM Review*, *51*(2), 339–360. https://doi.org/10.1137/070690274

Kish, L. (1965). *Survey sampling*. John Wiley & Sons.

Kobak, D., Linderman, G., Steinerberger, S., Kluger, Y., & Berens, P. (2020). Heavy-Tailed Kernels Reveal a Finer Cluster Structure in t-SNE Visualisations. *Machine Learning and Knowledge Discovery in Databases*, 124–139. https://doi.org/10.1007/978-3-030-46150-8_8

Kobak, D., & Linderman, G. C. (2021). Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, *39*(2), 156–157. https://doi.org/10.1038/s41587-020-00809-z

Korfmann, K., Pope, N. S., Meleghy, M., Tellier, A., & Kern, A. D. (2026). Coalescence and translation: A language model for population genetics. *Proceedings of the National Academy of Sciences*, *123*(15), e2518956123. https://doi.org/10.1073/pnas.2518956123

Krivitsky, P. N., & Handcock, M. S. (2008). Fitting Latent Cluster Models for Networks with latentnet. *Journal of Statistical Software*, *24*, 1–23. https://doi.org/10.18637/jss.v024.i05

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., & Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, *31*(3), 204–213. https://doi.org/10.1016/j.socnet.2009.04.001

Kuderna, L. F. K., Gao, H., Janiak, M. C., Kuhlwilm, M., Orkin, J. D., Bataillon, T., Manu, S., Valenzuela, A., Bergman, J., Rousselle, M., Silva, F. E., Agueda, L., Blanc, J., Gut, M., de Vries, D., Goodhead, I., Harris, R. A., Raveendran, M., Jensen, A., … Marques Bonet, T. (2023). A global catalog of whole-genome diversity from 233 primate species. *Science*, *380*(6648), 906–913. https://doi.org/10.1126/science.abn7829

Kuhner, M. K., Yamato, J., & Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, *149*(1), 429–434. https://doi.org/10.1093/genetics/149.1.429

Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockford, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J. C., Nishida, T., Pusey, A., Reynolds, V., Schu-

bert, G., Wrangham, R. W., Wroblewski, E., & Vigilant, L. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences*, *109*(39), 15716–15721. https://doi.org/10.1073/pnas.1211740109

Laso-Jadart, R., Harmant, C., Quach, H., Zidane, N., Tyler-Smith, C., Mehdi, Q., Ayub, Q., Quintana-Murci, L., & Patin, E. (2017). The Genetic Legacy of the Indian Ocean Slave Trade: Recent Admixture and Post-admixture Selection in the Makranis of Pakistan. *The American Journal of Human Genetics*, *101*(6), 977–984. https://doi.org/10.1016/j.ajhg.2017.09.025

Lawson, D. J., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1), 3258. https://doi.org/10.1038/s41467-018-05257-7

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. *PLOS Genetics*, *8*(1), e1002453. https://doi.org/10.1371/journal.pgen.1002453

Lee, S., Zou, F., & Wright, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *The Annals of Statistics*, *38*(6), 3605–3629. https://doi.org/10.1214/10-AOS821

Lewis, A. C. F., Molina, S. J., Appelbaum, P. S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S. M., Garrison, N. A., Ghosh, N., Hammonds, E. M., Jones, D. S., Kenny, E. E., Kraft, P., Lee, S. S.-J., Mauro, M., Novembre, J., Panofsky, A., Sohail, M., Neale, B. M., & Allen, D. S. (2022). Getting genetic ancestry right for science and society. *Science*, *376*(6590), 250–252. https://doi.org/10.1126/science.abm7530

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, *319*(5866), 1100–1104. https://doi.org/10.1126/science.1153717

Lipson, M., Ribot, I., Mallick, S., Rohland, N., Olalde, I., Adamski, N., Broomandkhoshbacht, N., Lawson, A. M., López, S., Oppenheimer, J., Stewardson, K., Asombang, R. N., Bocherens, H., Bradman, N., Culleton, B. J., Cornelissen, E., Crevecoeur, I., de Maret, P., Fomine, F. L. M., … Reich, D. (2020). Ancient West African foragers in the context of African population history. *Nature*, *577*(7792), 665–670. https://doi.org/10.1038/s41586-020-1929-1

Liu, J.-B., & Pan, X.-F. (2015). A unified approach to the asymptotic topological indices of various lattices. *Applied Mathematics and Computation*, *270*, 62–73. https://doi.org/10.1016/j.amc.2015.08.008

Liu, J.-B., Pan, X.-F., Cao, J., & Huang, X. (2014). The Kirchhoff index of toroidal meshes and variant networks. *Mathematical Problems in Engineering*, *2014*, 1–8. https://doi.org/10.1155/2014/286876

Loya, H., Hinch, A. G., Palamara, P. F., Speidel, L., & Myers, S. R. (2026, April 17). *Genome-wide genealogies reveal deep admixtures forming modern humans*. https://doi.org/10.64898/2026.04.17.719197

Loyal, J. D. (2025). Fast Variational Inference of Latent Space Models for Dynamic Networks Using Bayesian P-Splines. *Bayesian Analysis*, *-1*, 1–29. https://doi.org/10.1214/25-BA1545

Lundgren, E., & Ralph, P. L. (2019). Are populations like a circuit? Comparing isolation by resistance to a new coalescent-based method. *Molecular Ecology Resources*, *19*(6), 1388–1406. https://doi.org/10.1111/1755-0998.13035

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. Retrieved July 2, 2026, from http://jmlr.org/papers/v9/vandermaaten08a.html

Maddison, W. P., Whitton, J., Maddison, W. P., & Whitton, J. (2023). The species as a reproductive community emerging from the past. *Bulletin of the Society of Systematic Biologists*, *2*(1). https://doi.org/10.18061/bssb.v2i1.9358

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., et al. (2016). A genomic history of Aboriginal Australia. *Nature*, *538*(7624), 207–214. https://doi.org/10.1038/nature18299

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Marcus, J., Ha, W., Barber, R. F., & Novembre, J. (2021). Fast and flexible estimation of effective migration surfaces. *eLife*, *10*, e61927. https://doi.org/10.7554/eLife.61927

Marjoram, P., & Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genetics*, *7*(1), 16. https://doi.org/10.1186/1471-2156-7-16

Markus, B., Alshafee, I., & Birk, O. S. (2014). Deciphering the fine-structure of tribal admixture in the Bedouin population using genomic data. *Heredity*, *112*(2), 182–189. https://doi.org/10.1038/hdy.2013.90

Maruyama, T. (1970). Stepping stone models of finite length. *Advances in Applied Probability*, *2*, 229–258.

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, *44*(3), 243–246. https://doi.org/10.1038/ng.1074

Mathieson, I., & Reich, D. (2017). Differences in the rare variant spectrum among human populations. *PLOS Genetics*, *13*(2), e1006581. https://doi.org/10.1371/journal.pgen.1006581

Mathieson, I., & Scally, A. (2020). What is ancestry? *PLoS Genetics*, *16*(3), e1008624. https://doi.org/10.1371/journal.pgen.1008624

Matias, C., & Miele, V. (2017). Statistical Clustering of Temporal Networks Through a Dynamic Stochastic Block Model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(4), 1119–1141. https://doi.org/10.1111/rssb.12200

McInnes, L., Healy, J., & Melville, J. (2020, September 18). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://doi.org/10.48550/arXiv.1802.03426

McLevey, J., Scott, J., J.Carrington, P., Kaur, H., Rastelli, R., Friel, N., & Raftery, A. E. (2024). Latent Position Network Models. In *The Sage Handbook of Social Network Analysis*. https://doi.org/doi.org/10.4135/9781529682618.n36

McRae, B. H. (2006). Isolation by Resistance. *Evolution*, *60*(8), 1551–1561. https://doi.org/10.1111/j.0014-3820.2006.tb00500.x

McVean, G. (2009). A Genealogical Interpretation of Principal Components Analysis. *PLOS Genetics*, *5*(10), e1000686. https://doi.org/10.1371/journal.pgen.1000686

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2005.1673

Menozzi, P., Piazza, A., & Cavalli-Sforza, L. (1978). Synthetic Maps of Human Gene Frequencies in Europeans. *Science*, *201*(4358), 786–792. https://doi.org/10.1126/science.356262

Ministry of Higher Education and Scientific Research, Egypt. (2026, April 7). Wazīr al-ta‘līm al-‘ālī yu‘lin natā’ij aḍkham dirāsa li-l-tasalsul al-jīnī al-kāmil ḍimn mashrū‘ al-jīnūm al-marji‘ī li-l-Miṣriyyīn wa-qudamā’ al-Miṣriyyīn [Minister of Higher Education announces results of the largest whole-genome sequencing study under the Reference Genome Project for Egyptians and Ancient Egyptians] [In Arabic]. Retrieved October 12, 2026, from https://mohesr.gov.eg/allnews/

Mirarab, S., Nakhleh, L., & Warnow, T. (2021). Multispecies Coalescent: Theory and Applications in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *52*, 247–268. https://doi.org/10.1146/annurev-ecolsys-012121-095340

Mitchell, M. W., Alexander, W., Mitchell, D. V., Freedman, A. H., Dordel, J., Harrigan, R. J., Sacan, A., Kentatchime, F., Featherstone, B. S., Abwe, E. E., Clee, P. R. S., Abwe, A. E., Locatelli, S., Morgan, B. J., Fosso, B., Fotso, R., Tishkoff, S. A., Eichler, E. E., Anthony, N. M., … Gonder, M. K. (2024, July 26). *Environmentally-mediated selection parallels population divergence across a chimpanzee subspecies contact zone*. https://doi.org/10.1101/2024.07.26.605171

Mitchell, M. W., Locatelli, S., Ghobrial, L., Pokempner, A. A., Sesink Clee, P. R., Abwe, E. E., Nicholas, A., Nkembi, L., Anthony, N. M., Morgan, B. J., Fotso, R., Peeters, M., Hahn, B. H., & Gonder, M. K. (2015). The population genetics of wild chimpanzees in Cameroon and Nigeria suggests a positive role for selection in the evolution of chimpanzee subspecies. *BMC Evolutionary Biology*, *15*(1), 3. https://doi.org/10.1186/s12862-014-0276-y

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Dyken, P. C. van, Tomkins-Tinch, C. H., Sochat, V., Forster, J., Vieira, F. G., Meesters, C., Lee, S., Twardziok, S. O., Kanitz, A., VanCampen, J., Malladi, V., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2025, September 23). *Sustainable data analysis with Snakemake*. https://doi.org/10.12688/f1000research.29032.3

Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A. L., & Reich, D. (2011). The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLOS Genetics*, *7*(4), e1001373. https://doi.org/10.1371/journal.pgen.1001373

Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J. L., Byrnes, J. K., Gignoux, C. R., Ortiz-Tello, P. A., Martínez, R. J., Hedges, D. J., Morris, R. W., Eng, C., Sandoval, K., Acevedo-Acevedo, S., Norman, P. J., Layrisse, Z., Parham, P., Martínez-Cruzado, J. C., Burchard, E. G., Cuccaro, M. L., … Bustamante, C. D. (2013). Reconstructing the Population Genetic History of the Caribbean. *PLOS Genetics*, *9*(11), e1003925. https://doi.org/10.1371/journal.pgen.1003925

Moreno-Mayar, J. V., Potter, B. A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., Kamm, J. A., Albrechtsen, A., Malaspinas, A.-S., Sikora, M., et al. (2018). Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*, *553*(7687), 203–207. https://doi.org/10.1038/nature25173

Morin, P. A., Moore, J. J., Chakraborty, R., Jin, L., Goodall, J., & Woodruff, D. S. (1994). Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science*, *265*(5176), 1193–1201. https://doi.org/10.1126/science.7915048

National Academies of Sciences, Engineering, and Medicine. (2023). *Using population descriptors in genetics and genomics research: A new framework for an evolving field*. The National Academies Press. Washington, DC. https://doi.org/10.17226/26902

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*(7637), 302–310. https://doi.org/10.1038/nature21347

Nielsen, R., Vaughn, A. H., & Deng, Y. (2025). Inference and applications of ancestral recombination graphs. *Nature Reviews Genetics*, *26*(1), 47–58. https://doi.org/10.1038/s41576-024-00772-4

Nielsen, R., & Wakeley, J. (2001). Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics*, *158*(2), 885–896. https://doi.org/10.1093/genetics/158.2.885

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd). Springer. https://doi.org/10.1007/978-0-387-40065-5

Nordborg, M. (2019). Coalescent Theory. In *Handbook of Statistical Genomics* (pp. 145–30). https://doi.org/10.1002/9781119487845.ch5

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, *29*(1), 59–75. https://doi.org/10.1007/BF00173909

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., & Bustamante, C. D. (2008).

Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101. https://doi.org/10.1038/nature07331

Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, *40*(5), 646–649. https://doi.org/10.1038/ng.139

Nowicki, K., & Snijders, T. A. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*(455), 1077–1087. https://doi.org/10.1198/016214501753208735

Osmond, M., & Coop, G. (2024). Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. *eLife*, *13*, e72177. https://doi.org/10.7554/eLife.72177

Panofsky, A., Dasgupta, K., & Iturriaga, N. (2021). How White nationalists mobilize genetics: From genetic ancestry and human biodiversity to counterscience and metapolitics. *American Journal of Physical Anthropology*, *175*(2), 387–398. https://doi.org/10.1002/ajpa.24150

Patin, E., Laval, G., Barreiro, L. B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K. K., Kidd, J. R., Veen, L. V. der, Hombert, J.-M., Gessain, A., Froment, A., Bahuchet, S., Heyer, E., & Quintana-Murci, L. (2009). Inferring the Demographic History of African Farmers and Pygmy Hunter–Gatherers Using a Multilocus Resequencing Data Set. *PLOS Genetics*, *5*(4), e1000448. https://doi.org/10.1371/journal.pgen.1000448

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J. B., Fernandes, V., Pereira, L., … Quintana-Murci, L. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, *356*(6337), 543–546. https://doi.org/10.1126/science.aal1988

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012a). Ancient Admixture in Human History. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012b). Ancient Admixture in Human History. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLOS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, *17*(4), 1617–1642.

Pearson, A., & Durbin, R. (2023, March 7). *Local Ancestry Inference for Complex Population Histories*. https://doi.org/10.1101/2023.03.06.529121

Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., & Li, J. Z. (2012). Genomic Patterns of Homozygosity in Worldwide Human Populations. *The American Journal of Human Genetics*, *91*(2), 275–292. https://doi.org/10.1016/j.ajhg.2012.06.014

Perez, G., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, C. M., Nassar, L. R., Raney, B. J., Speir, M. L., van Baren, M. J., Vaske, C. J., Haussler, D., Kent, W. J., & Haeussler, M. (2025). The UCSC Genome Browser database: 2025 update. *Nucleic Acids Research*, *53*(D1), D1243–D1249. https://doi.org/10.1093/nar/gkae974

Peter, B. M. (2016). Admixture, Population Structure, and F-Statistics. *Genetics*, *202*(4), 1485–1501. https://doi.org/10.1534/genetics.115.183913

Petkova, D., Novembre, J., & Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, *48*(1), 94–100. https://doi.org/10.1038/ng.3464

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967

Pockrandt, C., Alzamel, M., Iliopoulos, C. S., & Reinert, K. (2020). GenMap: Ultra-fast computation of genome mappability. *Bioinformatics*, *36*(12), 3687–3692. https://doi.org/10.1093/bioinformatics/btaa222

Pope, N. S., Tallman, S., Jeffery, B., Robertson, D., Wong, Y., Karthikeyan, S., Ralph, P. L., & Kelleher, J. (2026, January 12). *Tracing the evolutionary histories of ultra-rare variants using variational dating of large ancestral recombination graphs*. https://doi.org/10.64898/2026.01.07.698223

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., ... Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–475. https://doi.org/10.1038/nature12228

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J., & Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics*, *36*(16), 4449–4457. https://doi.org/10.1093/bioinformatics/btaa520

Pybus, O. G., Rambaut, A., & Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, *155*(3), 1429–1437. https://doi.org/10.1093/genetics/155.3.1429

Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M. C., Malaspinas, A.-S., et al. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science*, *349*(6250), aab3884. https://doi.org/10.1126/science.aab3884

Ragsdale, A. P., Weaver, T. D., Atkinson, E. G., Hoal, E. G., Möller, M., Henn, B. M., & Gravel, S. (2023). A weakly structured stem for human origins in Africa. *Nature*, *617*(7962), 755–763. https://doi.org/10.1038/s41586-023-06055-y

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science*, *3*(4), 425–441. https://doi.org/10.1214/ss/1177012761

Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-Wide Inference of Ancestral Recombination Graphs. *PLOS Genetics*, *10*(5), e1004342. https://doi.org/10.1371/journal.pgen.1004342

Rastelli, R., & Corneli, M. (2023a). Continuous latent position models for instantaneous interactions. *Network Science*, *11*(4), 560–588. https://doi.org/10.1017/nws.2023.14

Rastelli, R., & Corneli, M. (2023b). Continuous latent position models for instantaneous interactions. *Network Science*, *11*(4), 560–588. https://doi.org/10.1017/nws.2023.14

Rastelli, R., Friel, N., & Raftery, A. E. (2016). Properties of latent variable network models. *Network Science*, *4*(4), 407–432. https://doi.org/10.1017/nws.2016.23

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., & Chikhi, L. (2018). The IICR and the non-stationary structured coalescent: Towards demographic inference with arbitrary changes in population structure. *Heredity*, *121*(6), 663–678. https://doi.org/10.1038/s41437-018-0148-0

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., & Feldman, M. W. (2005). Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLOS Genetics*, *1*(6), e70. https://doi.org/10.1371/journal.pgen.0010070

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, *3*(5), 380–390. https://doi.org/10.1038/nrg795

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic Structure of Human Populations. *Science*, *298*(5602), 2381–2385. https://doi.org/10.1126/science.1078311

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, *60*(1), 259–268. https://doi.org/10.1016/0167-2789(92)90242-F

Ryan, C., Wyse, J., & Friel, N. (2017). Bayesian model selection for the latent position cluster model for social networks. *Network Science*, *5*(1), 70–91. https://doi.org/10.1017/nws.2017.6

Salter-Townshend, M., & Murphy, T. B. (2013). Variational Bayesian inference for the Latent Position Cluster Model for network data. *Computational Statistics & Data Analysis*, *57*(1), 661–671. https://doi.org/10.1016/j.csda.2012.08.004

Sarkar, P., & Moore, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, *7*(2), 31–40. https://doi.org/10.1145/1117454.1117459

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H., Lombard, M., & Jakobsson, M. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, *358*(6363), 652–655. https://doi.org/10.1126/science.aao6266

Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G. B., Soodyall, H., & Jakobsson, M. (2012). Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science*, *338*(6105), 374–379. https://doi.org/10.1126/science.1227721

Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, *16*(12), 727–740. https://doi.org/10.1038/nrg4005

Schweiger, R., & Durbin, R. (2023). Ultrafast genome-wide inference of pairwise coalescence times. *Genome Research*, *33*(7), 1023–1031. https://doi.org/10.1101/gr.277665.123

Sellinger, T. P. P., Abu Awad, D., & Tellier, A. (2021). Limits and convergence properties of the sequentially Markovian coalescent. *Molecular Ecology Resources*, *21*(7), 2231–2248. https://doi.org/10.1111/1755-0998.13416

Serre, D., & Pääbo, S. (2004). Evidence for Gradients of Human Genetic Diversity Within and Among Continents. *Genome Research*, *14*(9), 1679–1685. https://doi.org/10.1101/gr.2529604

Sewell, D. K., & Chen, Y. (2015). Latent Space Models for Dynamic Networks. *Journal of the American Statistical Association*, *110*(512), 1646–1657. Retrieved June 22, 2026, from https://www.jstor.org/stable/24740175

Sewell, D. K., & Chen, Y. (2017). Latent Space Approaches to Community Detection in Dynamic Networks. *Bayesian Analysis*, *12*(2), 351–377. https://doi.org/10.1214/16-BA1000

Shanks, C., Bonet, D., Cara, M. C., & Ioannidis, A. (2026, April 7). *Graph transformer for ancient ancestry inference*. https://doi.org/10.64898/2026.04.05.714076

Sheehan, S., Harris, K., & Song, Y. S. (2013). Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics*, *194*(3), 647–662. https://doi.org/10.1534/genetics.112.149096

Shen, H., & Novembre, J. (2026). FRAME: Fine-Resolution Asymmetric Migration Estimation. *Nature Communications*, *17*(1), 7891. https://doi.org/10.1038/s41467-026-74129-2

Shlush, L. I., Behar, D. M., Yudkovsky, G., Templeton, A., Hadid, Y., Basis, F., Hammer, M., Itzkovitz, S., & Skorecki, K. (2008). The Druze: A Population Genetic Refugium of the Near East. *PLOS ONE*, *3*(5), e2105. https://doi.org/10.1371/journal.pone.0002105

Shortreed, S., Handcock, M. S., & Hoff, P. (2006). Positional estimation within a latent space model for networks. *Methodology: European Journal of Research Methods for*

the Behavioral and Social Sciences, *2*(1), 24–33. https://doi.org/10.1027/1614-2241.2.1.24

Silverstein, J. W. (1995). Strong Convergence of the Empirical Distribution of Eigenvalues of Large Dimensional Random Matrices. *Journal of Multivariate Analysis*, *55*(2), 331–339. Retrieved October 15, 2026, from https://ideas.repec.org//a/eee/jmvana/v55y1995i2p331-339.html

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., Heinze, A., Olalde, I., Ferry, M., Harney, E., Michel, M., Stewardson, K., Cerezo-Román, J. I., Chiumia, C., Crowther, A., ... Reich, D. (2017). Reconstructing Prehistoric African Population Structure. *Cell*, *171*(1), 59–71.e21. https://doi.org/10.1016/j.cell.2017.08.049

Slatkin, M. (1987). Gene Flow and the Geographic Structure of Natural Populations. *Science*, *236*(4803), 787–792. https://doi.org/10.1126/science.3576198

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research*, *58*(2), 167–175. https://doi.org/10.1017/S0016672300029827

Slatkin, M. (1993). ISOLATION BY DISTANCE IN EQUILIBRIUM AND NON-EQUILIBRIUM POPULATIONS. *Evolution*, *47*(1), 264–279. https://doi.org/10.1111/j.1558-5646.1993.tb01215.x

Speidel, L., Cassidy, L., Davies, R. W., Hellenthal, G., Skoglund, P., & Myers, S. R. (2021). Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. *Molecular Biology and Evolution*, *38*(9), 3497–3511. https://doi.org/10.1093/molbev/msab174

Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, *51*(9), 1321–1329. https://doi.org/10.1038/s41588-019-0484-x

Speidel, L., Silva, M., Booth, T., Raffield, B., Anastasiadou, K., Barrington, C., Götherström, A., Heather, P., & Skoglund, P. (2025). High-resolution genomic history of early medieval Europe. *Nature*, *637*(8044), 118–126. https://doi.org/10.1038/s41586-024-08275-2

Spence, J. P., Steinrücken, M., Terhorst, J., & Song, Y. S. (2018). Inference of population history using coalescent HMMs: Review and outlook. *Current Opinion in Genetics & Development*, *53*, 70–76. https://doi.org/10.1016/j.gde.2018.07.002

Steinrücken, M., Kamm, J., Spence, J. P., & Song, Y. S. (2019). Inference of complex population histories using whole-genome sequences from multiple populations. *Proceedings of the National Academy of Sciences*, *116*(34), 17115–17120. https://doi.org/10.1073/pnas.1905060116

Steux, C., Couloigner, C., Arredondo, A., Rodríguez, W., Mazet, O., Tournebize, R., & Chikhi, L. (2025). On the Demographic History of Chimpanzees and Some Consequences of Integrating Population Structure in Chimpanzees and Other Great Apes. *Human Population Genetics and Genomics*, *5*(2). https://doi.org/10.47248/hpgg2505020004

Stevison, L. S., Woerner, A. E., Kidd, J. M., Kelley, J. L., Veeramah, K. R., McManus, K. F., Great Ape Genome Project, Bustamante, C. D., Hammer, M. F., & Wall, J. D. (2016).

The Time Scale of Recombination Rate Evolution in Great Apes. *Molecular Biology and Evolution*, *33*(4), 928–945. https://doi.org/10.1093/molbev/msv331

Strang, G. (1999). The Discrete Cosine Transform. *SIAM Review*, *41*(1), 135–147. https://doi.org/10.1137/S0036144598336745

Strobeck, C. (1987). Average Number of Nucleotide Differences in a Sample from a Single Subpopulation: A Test for Population Subdivision. *Genetics*, *117*(1), 149–153. https://doi.org/10.1093/genetics/117.1.149

Tang, J., & Chiang, C. W. K. (2025). A genealogy-based approach for revealing ancestry-specific structures in admixed populations. *The American Journal of Human Genetics*, *112*(8), 1906–1922. https://doi.org/10.1016/j.ajhg.2025.06.016

Terhorst, J. (2025). Accelerated Bayesian inference of population size history from recombining sequence data. *Nature Genetics*, *57*(10), 2570–2577. https://doi.org/10.1038/s41588-025-02323-x

Terhorst, J. (2026, April 22). *Rates of forgetting for the sequentially Markov coalescent*. https://doi.org/10.48550/arXiv.2604.20629

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*(2), 303–309. https://doi.org/10.1038/ng.3748

Tetali, P. (1991). Random walks and the effective resistance of networks. *Journal of Theoretical Probability*, *4*(1), 101–109. https://doi.org/10.1007/BF01046996

Tibshirani, R. J. (2014). Adaptive Piecewise Polynomial Estimation Via Trend Filtering. *The Annals of Statistics*, *42*(1), 285–323. Retrieved June 29, 2026, from https://www.jstor.org/stable/43556281

Tufto, J., Raybould, A. F., Hindar, K., & Engen, S. (1998). Analysis of genetic structure and dispersal patterns in a population of sea beet. *Genetics*, *149*(4), 1975–1985. https://doi.org/10.1093/genetics/149.4.1975

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, K., Mathieson, I., O'Connell, J., & Schiffels, S. (2020). Tracking human population structure through time from whole genome sequences. *PLOS Genetics*, *16*(3), e1008552. https://doi.org/10.1371/journal.pgen.1008552

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., et al. (2007). Genetic variation and population structure in native americans. *PLoS Genetics*, *3*(11), e185. https://doi.org/10.1371/journal.pgen.0030185

Wang, Y. J., & Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, *82*(397), 8–19. https://doi.org/10.1080/01621459.1987.10478385

Whitlock, M. C., & Barton, N. H. (1997). The Effective Size of a Subdivided Population. *Genetics*, *146*(1), 427–441. https://doi.org/10.1093/genetics/146.1.427

Wilson, J. D., Cranmer, S., & Lu, Z.-L. (2020). A Hierarchical Latent Space Network Model for Population Studies of Functional Connectivity. *Computational Brain & Behavior*, *3*(4), 384–399. https://doi.org/10.1007/s42113-020-00080-0

Wiuf, C., & Hein, J. (1999). The Ancestry of a Sample of Sequences Subject to Recombination. *Genetics*, *151*(3), 1217–1228. https://doi.org/10.1093/genetics/151.3.1217

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., & McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, *375*(6583), eabi8264. https://doi.org/10.1126/science.abi8264

Wolfe, P. (1969). Convergence conditions for ascent methods. *SIAM Review*, *11*(2), 226–235. https://doi.org/10.1137/1011036

Won, Y.-J., & Hey, J. (2005). Divergence Population Genetics of Chimpanzees. *Molecular Biology and Evolution*, *22*(2), 297–307. https://doi.org/10.1093/molbev/msi017

Wong, Y., Ignatieva, A., Koskela, J., Gorjanc, G., Wohns, A. W., & Kelleher, J. (2024). A general and efficient representation of ancestral recombination graphs. *Genetics*, *228*(1), iyae100. https://doi.org/10.1093/genetics/iyae100

Wright, S. (1949). THE GENETICAL STRUCTURE OF POPULATIONS. *Annals of Eugenics*, *15*(1), 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., He, Y., Yang, Y., Wang, Y., Fu, W., An, Y., Wang, J., Tan, J., Qian, J., Chen, X., … Jin, L. (2009). Genomic Dissection of Population Substructure of Han Chinese and Its Implication in Association Studies. *The American Journal of Human Genetics*, *85*(6), 762–774. https://doi.org/10.1016/j.ajhg.2009.10.015

Y. C. Brandt, D., Wei, X., Deng, Y., Vaughn, A. H., & Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, *221*(1), iyac044. https://doi.org/10.1093/genetics/iyac044

Yang, Z., King, I., Xu, Z., & Oja, E. (2009). Heavy-Tailed Symmetric Stochastic Neighbor Embedding. *Advances in Neural Information Processing Systems*, *22*. Retrieved July 5, 2026, from https://proceedings.neurips.cc/paper/2009/hash/2291d2ec3b3048d1a6f86c2c4591b7e0Abstract.html

Ye, L. (2011). On the Kirchhoff index of some toroidal lattices. *Linear and Multilinear Algebra*, *59*(6), 645–650. https://doi.org/10.1080/03081081003794233

Yuan, M., & Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F., & Palamara, P. F. (2023). Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics*, *55*(5), 768–776. https://doi.org/10.1038/s41588-023-01379-x

Zhang, Y., Biddanda, A., Johnson, S. A., O'Dushlaine, C., & Moorjani, P. (2026). Recovering signatures of archaic hominin introgression using ancestral recombination graphs. *Science*, *0*(0), eaef8874. https://doi.org/10.1126/science.aef8874

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, *23*(4), 550–560. https://doi.org/10.1145/279232.279236