A geographic history of human genetic ancestry

Michael C. Grundler[1], **Jonathan Terhorst**[2], **and Gideon S. Bradburd**[1,✉]

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA; [2]Department of Statistics, University of Michigan, Ann Arbor, MI, USA

Describing the distribution of genetic variation across individuals is a fundamental goal of population genetics. In humans, traditional approaches for describing population genetic variation often rely on discrete genetic ancestry labels, which, despite their utility, can obscure the complex, multifaceted nature of human genetic history. These labels risk oversimplifying ancestry by ignoring its temporal depth and geographic continuity, and may therefore conflate notions of race, ethnicity, geography, and genetic ancestry. Here, we present a method that capitalizes on the rich genealogical information encoded in genomic tree sequences to infer the geographic locations of the shared ancestors of a sample of sequenced individuals. We use this method to infer the geographic history of genetic ancestry of a set of human genomes sampled from Europe, Asia, and Africa, accurately recovering major population movements on those continents. Our findings demonstrate the importance of defining the spatial-temporal context of genetic ancestry to describing human genetic variation and caution against the oversimplified interpretations of genetic data prevalent in contemporary discussions of race and ancestry.

Correspondence: ***bradburd@umich.edu***

Introduction

The genomes of present-day individuals existed at every point in the past, scattered across geographic space and contained within the ancestors from whom they will eventually inherit their genetic material. The way those ancestors moved across space through history determines spatial patterns of genetic relatedness in the present (Bradburd and Ralph, 2019). Understanding these spatial patterns is vital both for the identification of the genomic basis of phenotypic variation (Battey et al., 2020) and for knowledge of the demographic history of a species (e.g. Ralph and Coop, 2013). Conversely, ignoring spatial demographic history can have serious implications for genome-wide association studies or the identification of loci involved in local adaptation (Zaidi and Mathieson, 2020).

Genetic variation in humans is often summarized with discrete geographic labels, but these can be inaccurate and misleading (Coop, 2022; Lewis et al., 2023). Even when loosely based on geographic history, genetic ancestry labels oversimplify a complex picture because they implicitly focus on only a single point in time. For example, based on our current best understanding of human origins, all living individuals are “African" (regardless of the geography of their recent ancestors) on considering their ancestry ∼200ky before present. Advances in the study of ancient DNA have revealed a lack of genetic continuity within geographic regions (e.g. Haak et al., 2015; Allentoft et al., 2015; Racimo et al., 2020; Mattila et al., 2023), further highlighting the shortcomings of genetic ancestry labels. The fact that these labels are generated using statistical genetics approaches gives them the veneer of authenticity, further reifying problematic notions of race and ancestry in society (Lewis et al., 2023; Carlson et al., 2022).

At a technical level, many of the existing statistical methods for quantifying ancestry average over the ages of shared ancestors in the sample, effectively “flattening" the temporal component of the genealogy that connects all individuals within a species (Lawson et al., 2012; Pritchard et al., 2000). In reality, any pair of individuals is connected by many shared ancestors from whom each have inherited some portion of their genome (Mathieson and Scally, 2020). This flattening has the effect of painting a static notion of ancestry, rather than one that changes as it proceeds backwards in time.

If we knew the identities, locations, and ages of the ancestors of a sample, we could much more precisely and accurately report the geographic ancestry of a set of modern-day individuals through time. Moreover, we could learn about the history of dispersal in a species, identifying major population movements, demographic events, and barriers to migration. Although such detailed pedigree information is rarely available (but see Aguillon et al. (2017) and Anderson-Trocmé et al. (2023)), it nevertheless *is* possible to learn about the pedigree ancestors that are *shared* among individuals in a sample. This is because genetic relationships between samples, as well as the identities of the shared ancestors via whom they are related, are encoded in an interwoven collection of gene genealogies called an ancestral recombination graph (ARG) (Griffiths and Marjoram, 1997; Lewanski et al., 2024; Wong et al., 2023).

Recent advances in statistical and computational population genetics (Speidel et al., 2019; Kelleher et al., 2019; Wohns et al., 2022; Deng et al., 2024) have facilitated the inference of an ARG from large numbers of whole genomes. The ARG is a record of all coalescence and recombination events since the divergence of the sequences under study, and therefore specifies a complete genealogy of the sample at each genomic

position. This record can be represented as a tree sequence (Kelleher et al., 2016, 2018) – an ordered set of trees, localized to adjacent regions of the genome, describing the gene genealogies of a set of samples at every genomic position. Each internal node in these local genealogies represents a haplotype within an ancestor from whom two or more sampled individuals have co-inherited a portion of their genome. By determining where and when each of these ancestors lived, we can, in principle, reconstruct the geographic history of a set of modern day individuals, documenting the path through space and time by which their genomes came to them. For example, Wohns et al. (2022) introduced a nonparametric approach that estimates ancestor locations by successively averaging the coordinates of sample locations in a postorder traversal of the ARG to their most recent common ancestor. In addition, Osmond and Coop (2021) and Deraje et al. (2024) describe likelihood methods for locating genetic ancestors and estimating migration rates based on a model of branching Brownian motion utilizing either samples of local gene trees or the full ARG.

Here, we present and validate an additional method for achieving this goal. Our method, called `gaia` (geographic ancestor inference algorithm), efficiently infers the geographic locations of the shared ancestors of a modern, georeferenced sample. We apply `gaia` to a tree sequence of humans sampled in Europe, Asia, the Middle East, and Africa (Wong et al., 2023), and are able to reconstruct the geographic history of human ancestry over the last two million years.

Results

Inferring the locations of shared ancestors

`Gaia` works by fitting a minimum migration cost function to each genomic position in an ancestral haplotype using the generalized parsimony algorithm and the local gene genealogy relating the sampled genomes at the given position (Fig. 1). Because the neighboring gene genealogies in a tree sequence are highly correlated, we are able to efficiently maintain the state of parsimony calculations as we iterate over the local genealogies in a tree sequence that contain the ancestral haplotype. Once these cost functions are computed for all genomic positions, we average them and assign the ancestral haplotype to the geographic location that minimizes its average cost function. These assignments then have a straightforward interpretation: they correspond to the geographic location that minimizes the overall migration cost of an average ancestral base pair. In our implementation, migration cost is a function of geographic distance, and the overall migration cost in a local genealogy is simply the sum of all ancestor-descendant migration costs.

To validate the performance of `gaia` for inferring the geographic locations of nodes in a tree sequence, we simulated genetic data under different spatial models using SLiM (Haller and Messer, 2023). `Gaia` performs well under both expansionary and stationary demographic histories and over a range of dispersal kernels (both magnitude and shape) (Figs. S3,S4). We also demonstrate that we can use the reconstructed geographic distances between nodes in the tree sequence to estimate the parent-offspring dispersal distance for both Gaussian and long-tailed dispersal kernels (Figs. S1,S2).

Tracking human ancestors through space and time

We inferred the geographic locations of ancestors of a contemporary sample of 2140 georeferenced human genomes from the Human Genome Diversity Project (Cann et al., 2002; Li et al., 2008) using a dated tree sequence inferred for chromosome 18 by Wohns et al. (2022) (Fig. 3). We focus on ancestry of the subset of individuals sampled from the continents of Europe, Asia, and Africa, consisting of 1070 contemporary individuals. The tree sequence for these individuals consists of 28,154 local genealogies, containing 114,606 ancestral nodes and spanning approximately 80,000 generations of human history. An equal area discrete global grid (Barnes and Sahr, 2023) (cell spacing approximately 800 km) intersected with Earth's landmass provided a set of habitable areas, and individual sample locations were assigned to the nearest grid cell. Although a variety of complex migration cost matrices can be envisioned, including ones that incorporate long distance dispersal as well as temporal variation in the set of habitable areas, we opted for the simplest model that only allows migration between neighboring grid cells and assigns a unit cost to each migration event. With these assumptions, we used `gaia` to locate ancestral nodes to the grid cells with the lowest migration cost. For many ancestral nodes, especially older nodes, multiple grid cells may be optimal or near-optimal (Fig. 2). Because our summaries ignore near-optimal solutions, they do not explore the full range of uncertainty in ancestral locations and should be viewed not as precise statements on where ancestors lived and how they moved, but rather as summaries of major trends.

Our inferred geographic chronology of the ancestors of the sample largely reconstructs major population movements in human prehistory, including the out-of-Africa expansion and the peopling of Eurasia (Fig. 4). Our estimates place some genetic ancestors of the sample in Asia and the Middle East well before the earliest fossil evidence of human dispersal out of Africa. The oldest nodes in the tree sequence are estimated to have occurred roughly 2 million years before present, and a small minority of these are inferred to be located in Asia and the Middle East. A similar pattern was observed by Wohns et al. (2022) in their analysis of chromosome 20. Nonetheless, from ca. 2 million to 200ky before present, the average positions of genetic ancestors to the geographic subsets of samples from Europe,

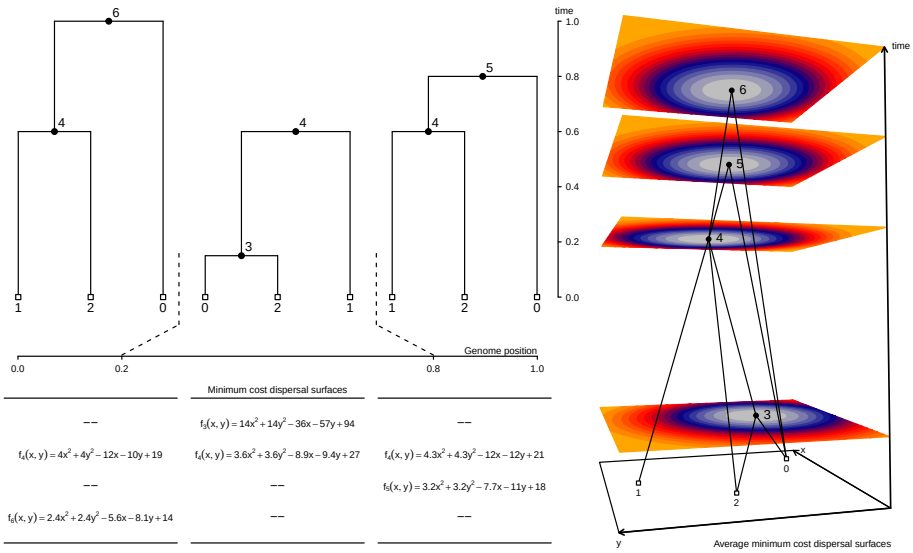

Figure 1. Conceptual overview of method. For each local tree we use the dynamic programming method of Sankoff and Rousseau (1975) to fit a minimum cost dispersal surface to the genealogical relationships of the georeferenced sample nodes. In this example, we use squared Euclidean distance as the cost function and $f_u(x, y)$ returns the smallest sum of squared dispersal distances between all ancestor-descendant node pairs that can be obtained when node u is at location (x, y). Using the genomic spans of local trees as weights, we then take a weighted average of local surfaces to assign each node a single average minimum cost dispersal surface. Here, node 4 appears in all three local trees and its final fitted surface is the weighted average of the three local surfaces. By contrast, nodes 3, 5, and 6 appear in a single local tree and their final fitted surfaces are identical to the surface in the local tree in which they appear. The perspective plot in the rightmost panel displays the ancestral recombination graph encoding the local trees along with the final fitted surface for each node. Non-sample nodes are positioned at the minimum cost point on the surface (warmer colors denote higher costs).

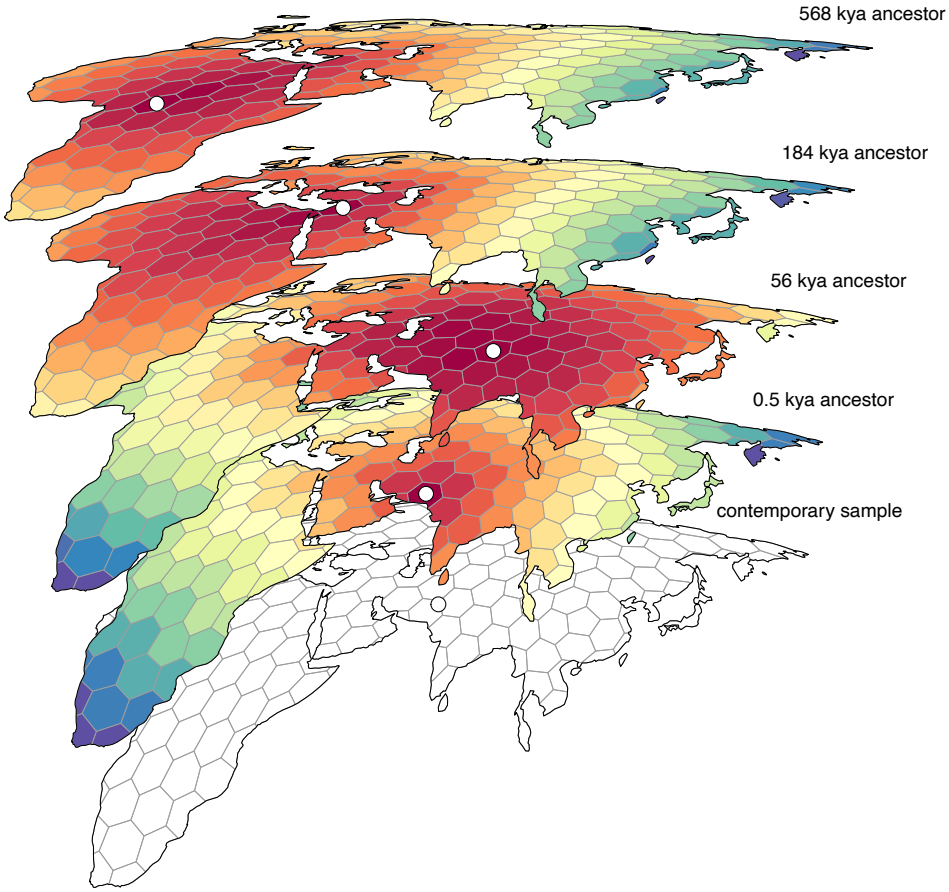

Figure 2. An example set of geographically referenced genetic ancestors of a contemporary sample from Asia. Colored grid cells denote the migration costs calculated by `gaia` (warmer colors denote lower costs) and each ancestor (denoted by a white point) is positioned within the minimum cost grid cell. For most ancestors there are many near-optimal locations that are ignored by our summary methods, which should therefore be viewed as summarizing major trends rather than as precise statements on where ancestors lived and how they moved.

Asia, the Middle East, and Africa are all inferred to be in Africa. This coincides with the time period when the majority of genome positions in these geographic subsets trace their descent to an ancestor in Africa. Between 200kya and the present, the average location of ancestors to the geographic subsets of samples from Europe, Asia, and the Middle East diverge from one another and begin to move toward the average of position of samples from those regions, while the average position of ancestors to the African subset of samples remains in Africa (Fig. 5).

Geographic history of human ancestry

We use the georeferenced tree sequence to define a spatiotemporally explicit ancestry coefficient, which we then track across space and time to understand and quantify the genetic and geographic history of our sample. Only a subset of any given contemporary individual's genome, $A_i(t)$ in individual i, is found in the ancestral nodes in the tree sequence at a given point in time t (once the local genealogy in a particular genomic region has coalesced, that portion of the genome is no longer represented in deeper sections of the ARG; therefore, the farther back in time we look, the less of any modern-day individual's genome can be found). We define this spatiotemporally explicit ancestry coefficient, which we call $z_{ik}(t)$, as $A_{ik}(t)/A_i(t)$: the proportion of individual i's genome that exists in its ancestors in the ARG at time t and that is inherited from ancestors living within some prescribed geographic region k. This ancestry coefficient can be broadly understood as the proportion of a sample's genome inherited from individuals living within a specified geographic region at a specific time.

Unlike ancestry labels, $z_{ik}(t)$ is *explicitly* associated with both a point in time and a region of space; we can therefore use it to understand how a sample's ancestry has changed across space and time. At the present moment, the distribution of z_{ik} simply reflects the geography of modern-day sampling. By 100,000ya, we find that only 2.5% of modern-day samples are inheriting their genomes from ancestors in Europe (Fig. 5); instead, almost all ancestors contributing genomic material to the modern-day sample are inferred to have lived in Africa, Asia, and the Middle East. The spatiotemporal ancestry coefficients show similar trajectories for modern-day samples from Asia. During this same time interval, the proportion of sample ancestors found in Africa increases almost monotonically backward in time, from 16% at the present (again, reflecting the geography of modern-day sampling) to $\sim$90% by 1Mya. Interestingly, the proportion of the genomes of the sample inherited from ancestors inferred to be in Africa plateaus at $\sim$90%, with $\sim$10% of genome continuing to be inherited from ancestors in the Middle East along the eastern Mediterranean. (Note that this region contributed the largest number of samples to the dataset.)

Large scale movement in human ancestry

To study large-scale movements of human ancestry, we introduce a new statistical summary of the georeferenced tree sequence: ancestry flux. Formally, if $A_i(t_l, t_r)$ is the amount of individual i's genome that inherits from ancestors alive between $[t_l, t_r)$ and $A_{ijk}(t_l, t_r)$ is the same amount that inherits from ancestors who moved from j to k, we define ancestry flux as $\phi_{ijk}(t_l, t_r) = \frac{A_{ijk}(t_l,t_r)}{A_i(t_l,t_r)}$: the proportion of i's genome that exists in its ancestors in the ARG during the time period $[t_l, t_r)$ and that is inherited from ancestors who moved from j to k during that same time period. This coefficient can be broadly understood as the proportion of a sample's genome inherited from ancestors who moved between specified geographic regions during a particular time period in the past.

We discretized Europe, Asia, and Africa into equal-area hexagons (each approximately 800 sq km) and quantified ancestry flux between them in 2,500 year intervals between the present and 0.5Mya. We find consistent ancestry flux out of Africa into the Middle East during this time period, with a peak occurring between 100,000ya and 150,000ya (Fig. 6). Interestingly, nearly all ancestry flux out of Africa is estimated to occur through a northern route across the Sinai Peninsula rather than a southern route across the strait at Bab-el-Mandeb. Approximately 30% of genomic positions in the modern-day sample trace ancestry to a northerly migration out of Africa but only 0.1% trace ancestry to a southerly migration route out of Africa.

Discussion

We have described a simple heuristic approach for reconstructing the geographic genetic history of a sample, which can be used to explore the timing and geographic location of demographic events. Patterns of genetic diversity and relatedness between individuals are created by a complex interplay between evolutionary forces acting over the history a population. By uncovering these patterns, our method enables researchers to go beyond the standard practice of grouping individuals into a small number of static and ill-defined "populations." Although this practice may be useful in some limited scenarios (e.g., for conservation and management), our results illustrate that it discards an immense amount of additional information that can be extracted from the data. Furthermore, when applied to humans, population assignment has the potential to lead to greater harm and confusion by conflating race, ethnicity, geography, and ancestry (Coop, 2022). A spatiotemporal summary of ancestors offers an intuitive and informative path toward understanding ancestry, demography, and major population movements through time.

Of course, the geographic histories inferred by `gaia` are not perfectly accurate; individual dispersal decisions are the result of countless external factors that we cannot

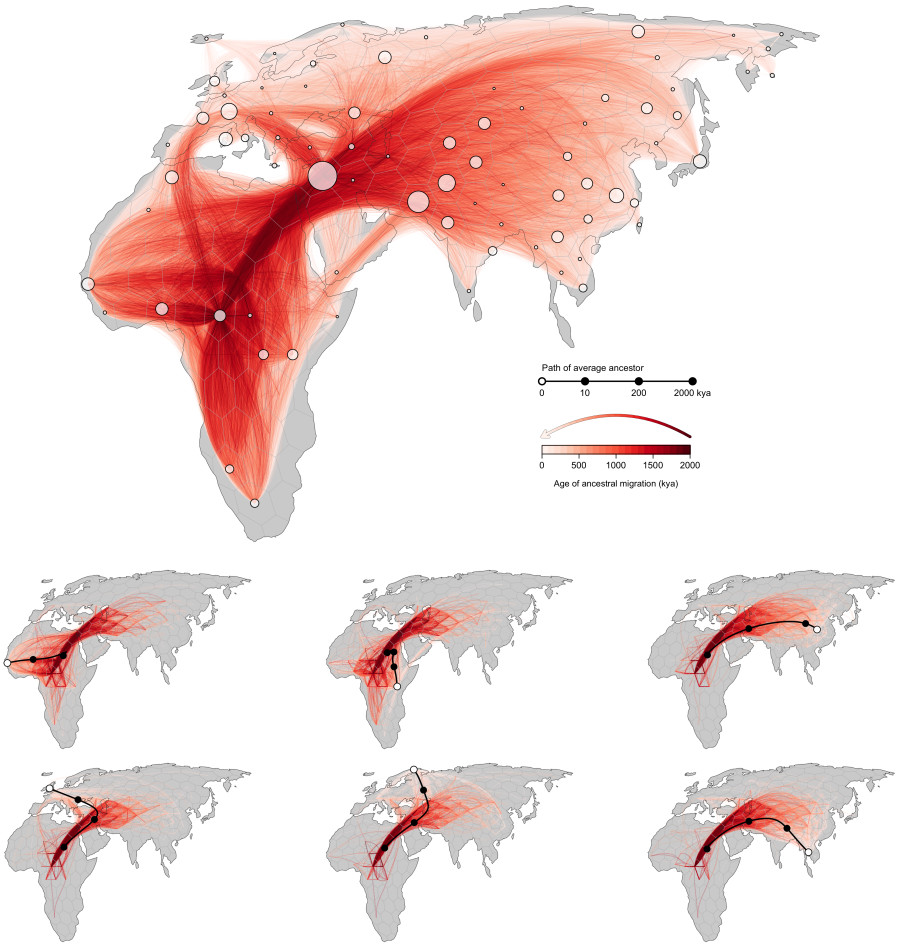

Figure 3. A georeferenced ancestral recombination graph (ARG). Red lines trace the inferred historical migrations of genetic ancestors of the sample (white points), with darker shading used to indicate movements that took place in the more distant past. Six contemporary samples are highlighted in the lower panels; black lines trace the average position of their genetic ancestors back through time, while red lines denote the subset of edges in the ARG that are ancestral to the samples.

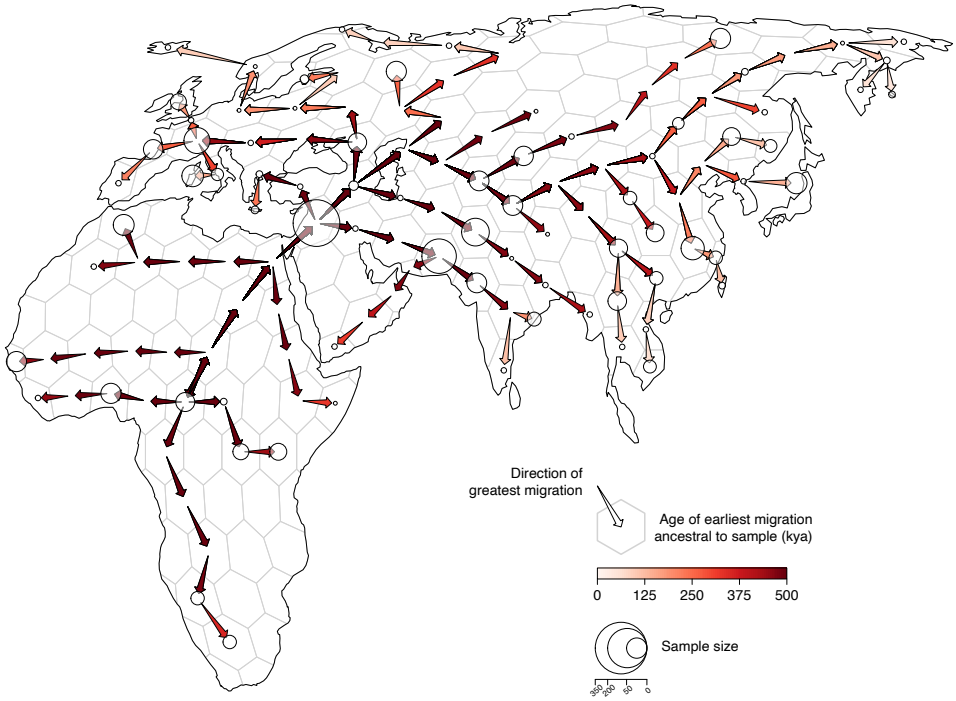

Figure 4. Geographic chronology of human genetic ancestry. Arrows point in the direction of greatest migration of the shared genetic ancestors of the sample and are colored according to the age of the earliest migration. Points show the distribution of sampled modern-day genomes. Point size is proportional to the number of samples from each locality.

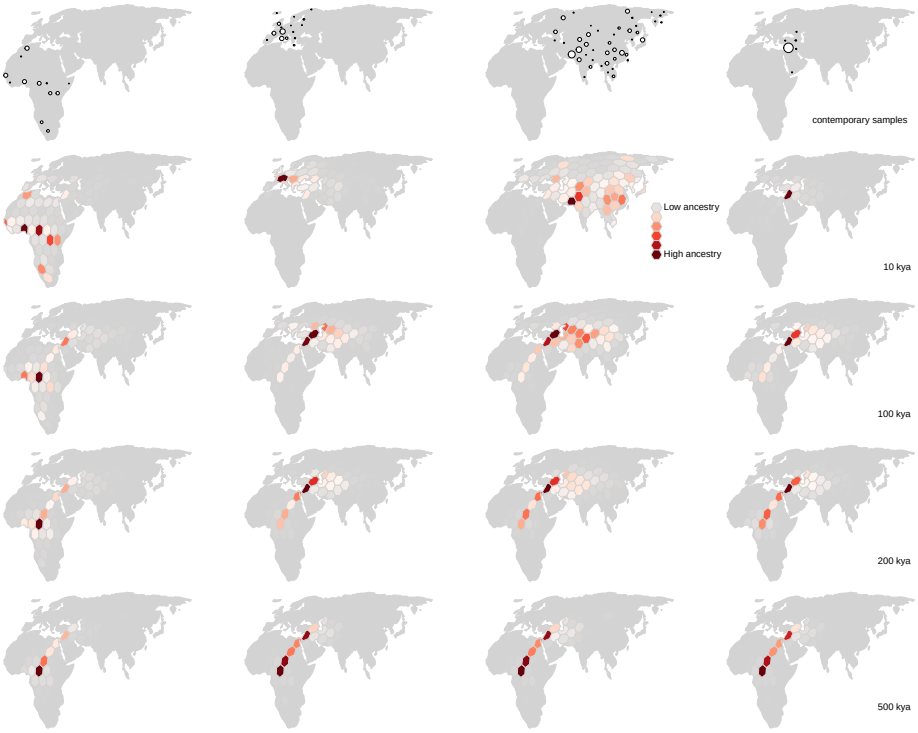

Figure 5. Inferred spatiotemporal ancestry coefficients through time. The fraction of genomic positions in each geographic subset of samples (top row) that trace ancestry to different geographic regions at different times in the past is represented by shades of red, with darker shading indicating greater ancestry (=a larger fraction). Point size is proportional to the number of sampled genomes from each locality.

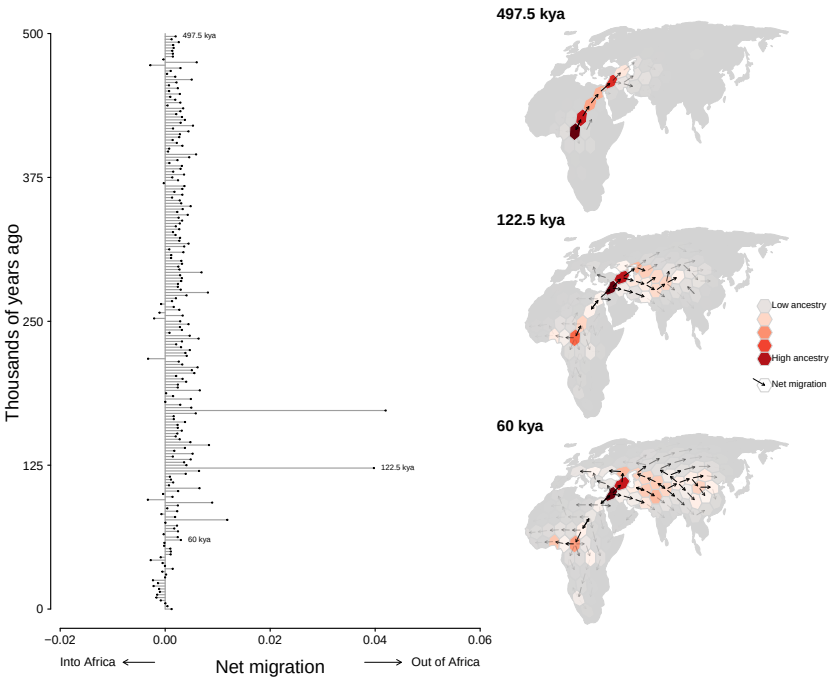

Figure 6. Inferred flux in spatiotemporal ancestry coefficients through time. Each bar depicts the net direction of historical migration of shared genetic ancestors of the modern-day sample. Inset maps highlight three different time intervals. Cells are colored by the fraction of genomic positions that find ancestry in them at that time and arrows depict the direction of greatest migration with opacity scaled by the magnitude migration.

hope to capture using a simple heuristic model. For example, gaia assumes that lineages disperse independently of one another and that the dispersal process is independent of the coalescent process in a reconstructed gene tree; both assumptions are likely to be violated in empirical datasets. Moreover, errors in the tree sequences inferred by upstream tools such as relate (Speidel et al., 2019) and tsinfer (Kelleher et al., 2019) have the potential to propagate into the output of gaia. Additionally, the sampling process itself may impact our results, insofar as the accuracy of geographic inference of ancestor locations depends on the distribution of sampled individuals relative to that of their ancestors. Finally, we caution that gaia sheds light on the geographic and temporal distribution of *shared genetic ancestors*; even if we had perfect information on the geographic locations of all ancestors, it would not necessarily inform our understanding of the geographic distribution of the general census population alive at any point in the past, though the two are likely correlated. As with all population genetic approaches, historical individuals that contributed no ancestry to the modern-day sample remain inscrutable.

Despite these limitations, it is clear that the ability to infer the locations of ancestors in the tree sequence opens exciting avenues of future research. Simply summarizing the georeferenced tree sequence has the potential to yield valuable insights into the evolutionary and natural history of a population, including identifying barriers to dispersal, shifts in dispersal regimes (magnitude and direction) through time, and the geography of ecological dynamics (at least with respect to genetic ancestry). A straightforward direction for extending gaia would be to explicitly incorporate some of these elements – e.g., heterogeneous dispersal rates between different regions. As the transitions between different geographic states occur on an arbitrary network (generated, in the applications presented here, via geographic proximity), we can easily incorporate different geographies; for example, dispersal routes that connect far-flung regions, which may be of interest when species occasionally experience human-mediated dispersal (e.g., (Zhan et al., 2023)). More broadly, the ability to study the geography of genealogies heralds an exciting growth in the ability of the field of population genetics to shed light on population ecological processes governing the movement, distribution, and density of individuals across space and through time.

Methods

Main approach: computing minimum migration fits to local trees

Given a set of sampled genomes with known locations and a tree sequence that relates the samples, gaia efficiently reconstructs the locations of genetic ancestors using a minimum migration heuristic. Conceptually, gaia may be thought of as proceeding in three steps (Fig. 1).

In the first step, a minimum migration cost function f_{uk} is fitted to each node u in each local tree k in which u appears using the generalized parsimony algorithm (Sankoff, 1975; Sankoff and Rousseau, 1975). For each spatial location, $f_{uk}(x)$ then returns the smallest sum of migration costs between all ancestor-descendant pairs in local tree k that can explain the sampled locations when u is at location x. We assume migration cost is a function of migration distance and provide three cost function implementations. In each of these three cases we allow migration distances to optionally be weighted by inverse branch length. For a continuous state space, we use either the squared Euclidean distance or the Manhattan distance. Squared Euclidean distance leads to cost functions that are convex quadratic functions of the spatial locations (Maddison, 1991); Manhattan distance, to convex piecewise linear functions (Csűrös, 2008). For a finite set of geographic locations, a square matrix with arbitrary transition costs between all pairs of geographic states leads to minimum migration costs that are discrete functions defined on the set of geographic locations (Clemente et al., 2009).

In the second step, we form a new function F_u by taking a weighted average of f_{uk} over the local trees in which u appears, using the genomic spans of each local tree as weights. We interpret F_u as the minimum migration cost of an average ancestral base pair. In other words, if one were to take each sampled base pair that inherits from u and fit a minimum migration cost function to its coalescent history, the average of those would look like F_u. In the Appendix, we present a detailed description of the algorithm we use for computing F_u from the local tree fits f_{uk}. Considerable computational savings can be achieved by using the succinct tree sequence encoding (Kelleher et al., 2016), which allows us to efficiently maintain the state of parsimony calculations as we iterate over local trees. As detailed in the Appendix, the local tree fits also allow us to compute an estimate of effective migration rate, which we define as the mean lineage migration cost of an average base pair. That is, if one were to take the coalescent history of each sampled base pair and compute the mean the per-branch migration cost in a most parsimonious migration history, the average of those – taken over all sampled base pairs – would equal the effective migration rate.

Finally, we locate each ancestor by finding the spatial location x where $F_u(x)$ is smallest. A constrained (rather than global) minimization of F_u could also be used to locate ancestors if some spatial locations are known to be uninhabitable. Although we do not pursue this possibility in the current study, uncertainty in ancestor locations could be explored by sampling from $\exp\big(-\lambda[F_u(x) - \min_z F_u(z)]\big)$, where λ is a tuning parameter that controls how strongly deviations from the spatial location that minimizes the overall migration cost

are penalized.

Spatiotemporal ancestry coefficients

We use our ancestor location estimates to define two spatially and temporally explicit ancestry coefficients as follows. In the first case, consider a subset i of the sampled genomes. Define $A_i(t)$ to be the total amount of sample material in the subset that is descended from ancestral material at a point in time t in the past. Define $A_{ik}(t)$ to be the total amount of sample material descended from ancestral material in geographic location k at time t in the past. We then define the spatiotemporal ancestry coefficient for subset i as $z_{ik}(t) = \frac{A_{ik}(t)}{A_i(t)}$, the fraction of sample material that is descended from ancestral material in geographic location k at time t in the past. Thus, if we were to pool all sample material in subset i whose coalescent history extended at least t time units into the past, $z_{ik}(t)$ is the probability that we can trace the coalescent history of a randomly selected base pair from that pool to an ancestral base pair in geographic location k at time t in the past.

Next, define $A_i(t_l, t_r)$ to be the total amount of sample material in the subset that is descended from ancestral material present during an interval in time $[t_l, t_r)$ in the past. Define $A_{ijk}(t_l, t_r)$ to be the total amount of sample material descended from ancestral material that migrated from geographic location j to location k during an interval in time $[t_l, t_r)$ in the past (where j is not equal to k). We then define the spatiotemporal ancestry flux coefficient for subset i as $\phi_{ijk}(t_l, t_r) = \frac{A_{ijk}(t_l,t_r)}{A_i(t_l,t_r)}$. Thus, if we were to pool all sample material in subset i whose coalescent history extended at least $[t_l, t_r)$ time units into the past, $\phi_{ijk}(t_l, t_r)$ is the probabily that we can trace the coalescent history of a randomly selected base pair from that pool to an ancestral base pair that migrated from geographic location j to location k during the interval $[t_l, t_r)$ in the past.

As defined, our spatiotemporal ancestry coefficients require knowing the geographic locations of ancestral lineages at arbitrary points in time in the past but our minimum migration location estimates (i.e., $\arg\min_x F_u(x)$) are available only for the two endpoints of an ancestral lineage. By using the migration cost function and conditioning on these endpoint states, however, we can sample a minimum cost migration history for each lineage and interpolate the history between the endpoint times of the ancestral lineage. In this way we can locate ancestors at arbitrary points in time. Further details about this procedure are presented in the Appendix.

Simulation study: testing ancestor location accuracy

We tested performance of `gaia` using spatially explicit forward-time simulations in SLiM v4.0.1 (Haller and Messer, 2023). Simulated individuals were semelparous and hermaphroditic and diploid for a single chromosome with 10^8 basepairs and a recombination rate of 10^{-8} per basepair per generation. Individuals coexisted on a two-dimensional square plane with reflecting boundaries and a side length of 1000σ units, where σ was the standard deviation of the dispersal kernel.

Population regulation occurred via density dependent effects on fecundity and survival. The number of offspring produced by a focal individual in generation t was Poisson-distributed with a mean equal to $\frac{R}{1+\frac{N_t \times (R-1)}{K}}$, where N_t is the local population density in a circle of radius 3σ around the focal individual at generation t, R is the population growth rate at low density, and K is the local carrying capacity density. During reproduction, each individual (the “mother") chose a mate (the “father") uniformly at random from the set of individuals living within a radius of 3σ from itself. A random location centered on the mother's location was then chosen for each offspring by drawing from a dispersal kernel with a standard deviation of σ. Each offspring produced in generation t survived to reproduce in generation $t+1$ with probability $\min(1.0, \frac{K}{O_t})$, where O_t is the local population density in a circle of radius 3σ around the focal offspring in generation t.

We conducted simulations using either a Gaussian dispersal kernel or a double exponential dispersal kernel and 10 different σ levels that were equally spaced from 0.2 to 2.0. The Gaussian kernel emulates a Brownian motion while the double exponential kernel emulates a Laplace motion, which generates more extreme displacements of offspring relative to Brownian motion.

We conducted 10 replicate simulations for each σ level under each dispersal kernel. For all simulations, we set $R = 2$ and $K = \frac{30}{(3\sigma)^2\pi}$. Each simulation was initiated from a randomly located founder population of 30 individuals that was allowed to grow and disperse for 10,000 generations, after which the simulation was terminated and the recorded (simplified) tree sequence of extant individuals was saved for analysis.

Simulations conducted under the Gaussian dispersal kernel were analyzed using squared Euclidean distance weighted by inverse branch length as the transition cost function. Simulations conducted under the double exponential dispersal kernel were analyzed using Manhattan distance weighted by inverse branch length as the transition cost function. For all simulations we evaluated our ability to accurately estimate effective migration rates and the locations of genetic ancestors. We defined the true effective migration rate (σ_e) to be the maximum likelihood estimate of σ under each dispersal kernel assuming we had complete knowledge of the locations of all ancestral individuals in the tree sequence.

Empirical example: reconstructing human migration out of Africa

We applied `gaia` to a contemporary sample of georeferenced human genomes from the Human Genome Diversity Project using the dated tree sequence inferred

for chromosome 18 by Wohns et al. (2022). Prior to analysis, we first simplified the tree sequence to include only individuals from Africa, Asia, and Europe. Certain non-sample nodes in the tree sequence are clear outliers with respect to the number of edges in which they participate. For example, one node has nearly 3,000 immediate descendants, well above the median number of 6. As these genealogical outliers have the potential to strongly bias geographic reconstructions, we further simplified the tree sequence by removing edges to or from non-sample nodes in the upper two percent of in-degree or out-degree counts. The resulting simplified tree sequence consisted of 28,154 local genealogies, containing 114,606 ancestral nodes and spanning approximately 80,000 generations of human history.

To explore historical geographic patterns in human genetic ancestry over the last one-half million years we created a set of equal-interval time bins spanning 100 generations (or 2,500 years using a 25 year generation time) that extended from the present to 20,000 generations in the past. An equal area discrete global grid (Barnes and Sahr, 2023) (cell spacing approximately 800 km) intersected with Earth's landmass provided a set of habitable locations. Individual samples were then mapped to the nearest grid cell and we used `gaia` to compute F_u for each node in the empirical tree sequence using a simple cost matrix that assigned a unit cost to each migration event and only allowed migration between neighboring grid cells.

Each internal node in the tree sequence was then geographically referenced by sampling uniformly at random one of the grid cells that minimized its migration cost function F_u (for some nodes there may be a unique grid cell that achieves this minimum but for others there may be multiple). This procedure was repeated 100 times and spatiotemporal ancestry coefficients were computed for each realization and then averaged.

Acknowledgements

We thank Yaniv Brandvain, Jed Carlson, Graham Coop, Puneeth Deraje, Doc Edge, James Kitchens, Marjorie Weber, and members of the Bradburd lab for comments and discussion that helped improve the research presented here. This research was supported by National Science Foundation (award number DMS-2052653), the National Institute of General Medical Sciences of the National Institutes of Health under award numbers R35GM151145 (JT) and R35GM137919 (GB). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Bibliography

Aguillon, S. M., Fitzpatrick, J. W., Bowman, R., Schoech, S. J., Clark, A. G., Coop, G., and Chen, N. (2017). Deconstructing isolation-by-distance: The genomic consequences of limited dispersal. *PLOS Genetics*, 13(8):1–27.

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M. L., Stenderup, J. T., de Barros Damgaard, P., Schroeder, H., Ahlström, T., Vinner, L., Malaspinas, A.-S., Margaryan, A., Higham, T., Chivall, D., Lynnerup, N., Harvig, L., Baron, J., Casa, P. D., Dabrowski, P., Duffy, P. R., Ebel, A., Epimakhov, A. V., Frei, K. M., Furmanek, M., Gralak, T., Gromov, A., Gronkiewicz, S., Grupe, G., Hajdu, T., Jarysz, R., Khartanovich, V., Khokhlov, A., Kiss, V., Kolár, J., Kriiska, A., Lasak, I., Longhi, C., McGlynn, G., Merkevicius, A., Merkyte, I., Metspalu, M., Mkrtchyan, R., Moiseyev, V., Paja, L., Pálfi, G., Pokutta, D., Łukasz Pospieszny, Price, T. D., Saag, L., Sablin, M. V., Shishlina, N., Smrčka, V., Soenov, V. I., Szeverényi, V., Tóth, G., Trifanova, S. V., Varul, L., Vicze, M., Yepiskoposyan, L., Zhitenev, V., Orlando, L., Sicheritz-Pontén, T., Brunak, S., Nielsen, R., Kristiansen, K. H., and Willerslev, E. (2015). Population genomics of bronze age eurasia. *Nature*, 522:167–172.

Anderson-Trocmé, L., Nelson, D., Zabad, S., Diaz-Papkovich, A., Kryukov, I., Baya, N., Touvier, M., Jeffery, B., Dina, C., Vézina, H., et al. (2023). On the genes, genealogies, and geographies of quebec. *Science*, 380(6647):849–855.

Barnes, R. and Sahr, K. *dggridR: Discrete Global Grids*, (2023). URL https://CRAN.R-project.org/package=dggridR. R package version 3.0.0.

Battey, C. J., Ralph, P. L., and Kern, A. D. (2020). Space is the place: effects of continuous spatial structure on analysis of population genetic data. *Genetics*, 215(1):193–214.

Bradburd, G. S. and Ralph, P. L. (2019). Spatial population genetics: It's about time. *Annual Review of Ecology, Evolution, and Systematics*, 50(1):427–449. doi: 10.1146/annurev-ecolsys-110316-022659.

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., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J., and Cavalli-Sforza, L. L. (2002). A human genome diversity cell line panel. *Science*, 296(5566):261–262. doi: 10.1126/science.296.5566.261b.

Carlson, J., Henn, B. M., Al-Hindi, D. R., and Ramachandran, S. (2022). Counter the weaponization of genetics research by extremists. *Nature*, 610(7932):444–447.

Clemente, J., Ikeo, K., Valiente, G., and Gojobori, T. (2009). Optimized ancestral state reconstruction using sankoff parsimony. *BMC Bioinformatics*, 10(51). doi: 10.1186/1471-2105-10-51.

Coop, G. (2022). Genetic similarity and genetic ancestry groups. *arXiv*.

Csűrös, M. Ancestral reconstruction by asymmetric wagner parsimony over continuous characters and squared parsimony over distributions. In *RECOMB International Workshop on Comparative Genomics*, pages 72–86. Springer, (2008).

Deng, Y., Nielsen, R., and Song, Y. S. (2024). Robust and accurate bayesian inference of genome-wide genealogies for large samples. *bioRxiv*.

Deraje, P., Kitchens, J., Coop, G., and Osmond, M. M. (2024). Inferring the geographic history of recombinant lineages using the full ancestral recombination graph. *bioRxiv*. doi: 10.1101/2024.04.10.588900.

Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Banffy, E., Economou, C., Francken, M., Friederich, S., Pena, R. G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S. L., Risch, R., Rojo Guerra, M. A., Roth, C., Szecsenyi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K. W., and Reich, D. (2015). Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, 522(7555):207–211.

Haller, B. C. and Messer, P. W. (2023). Slim 4: Multispecies eco-evolutionary modeling. *The American Naturalist*, 201(5):E127–E139. doi: 10.1086/723601.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, 12(5):1–22. doi: 10.1371/journal.pcbi.1004842.

Kelleher, J., Thornton, K. R., Ashander, J., and Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, 14(11):1–21.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338. doi: 10.1038/s41588-019-0483-y.

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLOS Genetics*, 8(1):1–16. doi: 10.1371/journal.pgen.1002453.

Lewanski, A. L., Grundler, M. C., and Bradburd, G. S. (2024). The era of the arg: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. *Plos Genetics*, 20(1):e1011110.

Lewis, A. C., Molina, S. J., Appelbaum, P. S., Dauda, B., Fuentes, A., Fullerton, S. M., Nanibaa'A, G., Ghosh, N., Green, R. C., Hammonds, E. M., et al. (2023). An ethical framework for research using genetic ancestry. *Perspectives in Biology and Medicine*, 66(2): 225–248.

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., and Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319 (5866):1100–1104. doi: 10.1126/science.1153717.

Maddison, W. P. (1991). Squared-change parsimony reconstructions of ancestral states for continuous-valued characters on a phylogenetic tree. *Systematic Zoology*, 40(3):304–314.

Mathieson, I. and Scally, A. (2020). What is ancestry? *PLOS Genetics*, 16(3):1–6. doi: 10.1371/journal.pgen.1008624.

Mattila, T. M., Svensson, E. M., Juras, A., Günther, T., Kashuba, N., Ala-Hulkko, T., Chyleński, M., McKenna, J., Pospieszny, Ł., Constantinescu, M., et al. (2023). Genetic continuity, isolation, and gene flow in stone age central and eastern europe. *Communications biol-*

ogy, 6(1):793.

Osmond, M. M. and Coop, G. (2021). Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. *bioRxiv*. doi: 10.1101/2021.07.13.452277.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959. doi: 10.1093/genetics/155.2.945.

Racimo, F., Woodbridge, J., Fyfe, R. M., Sikora, M., Sjögren, K.-G., Kristiansen, K., and Vander Linden, M. (2020). The spatiotemporal spread of human migrations during the european holocene. *Proceedings of the National Academy of Sciences*, 117(16):8989–9000.

Ralph, P. and Coop, G. (2013). The geography of recent genetic ancestry across Europe. *PLoS Biol*, 11(5):e1001555.

Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42.

Sankoff, D. and Rousseau, P. (1975). Locating the vertices of a steiner tree in an arbitrary metric space. *Mathematical Programming*, 9:240–246.

Speidel, L., Forest, M., Shi, S., and Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature genetics*, 51(9):1321–1329.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, 375(6583):eabi8264. doi: 10.1126/science.abi8264.

Wong, Y., Ignatieva, A., Koskela, J., Gorjanc, G., Wohns, A. W., and Kelleher, J. (2023). A general and efficient representation of ancestral recombination graphs. *bioRxiv*.

Zaidi, A. A. and Mathieson, I. (2020). Demographic history mediates the effect of stratification on polygenic scores. *Elife*, 9:e61548.

Zhan, S. H., Ignatieva, A., Wong, Y., Eaton, K., Jeffery, B., Palmer, D. S., Murall, C. L., Otto, S., and Kelleher, J. (2023). Towards pandemic-scale ancestral recombination graphs of sars-cov-2. *bioRxiv*, pages 2023–06.

Supplementary Information

Appendix

A1. Overview

We first review generalized parsimony as it applies to a single gene tree (Sankoff, 1975; Sankoff and Rousseau, 1975). For each node u we maintain three cost functions. The *node cost* function $g_u(x)$ assigns a cost to each geographic state x (when u is in state x) that gives the minimum migration cost required to explain the geographic states of all sample nodes whose most recent common ancestor is u. If u is a sample node, we set $g_u(x) = 0$ if $x = x_u$ and $g_u(x) = \infty$ if $x \neq x_u$. When u is not a sample node, g_u is formed by the sum of the *stem cost* functions of its children.

The stem cost function for a node u is given by $h_u(x) = \min_z \left[\Delta_u(x,z) + g_u(z)\right]$. The function $\Delta_u(x,z)$ assigns a cost to the migration from state x to state z over the branch leading to node u. Here, we assume that Δ_u can be factored as: $\Delta_u(x,z) = \phi(\tau_u) * \Delta(x,z)$, where τ_u is the length of the branch leading to node u and Δ is a time-independent transition cost function. Typical choices for the time-dependent component are $\phi(\tau_u) = \frac{1}{\tau_u}$ or $\phi(\tau_u) = 1$, the latter making the whole cost function independent of branch lengths.

The node cost and stem cost functions can be computed in a single post-order tree traversal. A second pre-order tree traversal can then be used to compute the *final cost* functions for each node. The final cost $f_u(x)$ gives the minimum migration cost required to explain the geographic states of *all* sample nodes when u is in state x. When u is the root of the tree, $f_u(x) = g_u(x)$. Otherwise, $f_u(x) = \min_z \left[f_{\pi_u}(z) - h_u(z) + \Delta_u(z,x)\right] + g_u(x)$, where π_u represents the immediate ancestor of node u.

A2. Algorithm

Here we describe how we use the succinct tree sequence encoding (Kelleher et al., 2016) to efficiently compute node cost and stem cost functions for each local genealogy. Under the coalescent with recombination, nearby gene trees are highly correlated. Moving between adjacent local trees typically requires only a small number of subtree-prune-and-regraft (SPR) operations. By keeping an index of the edges involved in these SPR operations and the order in which they need to be applied, the succinct tree sequence encoding allows us to efficiently maintain the state of parsimony calculations as we iterate over the local genealogies.

A detailed description of the tree sequence data structure can be found in Kelleher et al. (2016). For our purposes, it suffices to know that relationships among all N nodes in the tree sequence are recorded in the edge table E. Each row in the edge table records an ancestor-descendant relationship, and the indices of nodes involved in that relationship can be accessed as $E[k]$.`parent` and $E[k]$.`child`. The relationship encoded by an edge applies to the half-open genomic interval between $E[k]$.`left` (inclusive) and $E[k]$.`right` (exclusive). As we move along the genome from left to right, the index vectors I and O give the insertion and removal order of the edges needed to build each local gene tree topology (recorded in the vector π). The following algorithm is based on the branch statistic algorithm that appears in Ralph et al. (2020) but differs in its treatment of sample weights and how they are propagated along a genealogy.

Algorithm P (generalized parsimony). Given a set of georeferenced samples related by a tree sequence with length L, compute the genome-wide average final cost function $F_u(x)$ and the genome-wide average migration rate statistic σ_P. The function $F_u(x)$ returns the average minimum migration cost required to explain the observed geographic states of all sample nodes when u is in state x, where the average is taken over all local genealogies (weighted by their genomic span) where node u appears. The statistic σ_P is the average per-branch migration cost in a most parsimonious migration history averaged over all local genealogies weighted by their genomic span.

S1. [Initialization.] For $0 \leq u < N$ set $\pi_u \leftarrow -1$, $s_u \leftarrow 0$, $g_u(x) \leftarrow 0$, $F_u(x) \leftarrow 0$. Then, if u is a sample node and $x \neq x_u$, set $g_u(x) \leftarrow \infty$ and $F_u(x) \leftarrow \infty$

Finally, set $j \leftarrow 0$, $k \leftarrow 0$, $\sigma_P \leftarrow 0$, $t_l \leftarrow 0$, $s_P \leftarrow 0$.

S2. [Terminate.] If $j = |E|$ terminate.

S3. [Edge removal loop.] If $k = |E|$ or $t_l \neq E[O_k]$.`right` go to S6.

S4. [Remove edge.] Set $u \leftarrow E[O_k]$.`parent`, $v \leftarrow E[O_k]$.`child`, $w \leftarrow \pi_u$, and $k \leftarrow k+1$. Then, if $w \neq -1$, set $g_w(x) \leftarrow g_w(x) - h_u(x)$. Finally, set $g_u(x) \leftarrow g_u(x) - h_v(x)$, $\pi_v = -1$, $v \leftarrow u$, $u \leftarrow w$.

S5. [Update node and stem costs.] While $u \neq -1$, set $w \leftarrow \pi_u$ and if $w \neq -1$ set $g_w(x) \leftarrow g_w(x) - h_u(x)$. Then set $h_v(x) \leftarrow \min_z \left[\Delta_v(x,z) + g_v(z)\right]$, $g_u(x) \leftarrow g_u(x) + h_v(x)$, $v \leftarrow u$, $u \leftarrow w$. Afterward, go to S3.

S6. [Edge insertion loop.] If $j = |E|$ or $t_l \neq E[I_j]$.`left` go to S9.

S7. [Insert edge.] Set $u \leftarrow E[I_j]$.`parent`, $v \leftarrow E[I_j]$.`child`, $w \leftarrow \pi_u$, and $j \leftarrow j+1$. Then, if $w \neq -1$, set $g_w(x) \leftarrow g_w(x) - h_u(x)$. Then set $h_v(x) \leftarrow \min_z \big[\Delta_v(x,z) + g_v(z)\big]$. Finally, set $g_u(x) \leftarrow g_u(x) + h_v(x)$, $\pi_v = u$, $v \leftarrow u$, $u \leftarrow w$.

S8. [Update node and stem costs.] While $u \neq -1$, set $w \leftarrow \pi_u$ and if $w \neq -1$ set $g_w(x) \leftarrow g_w(x) - h_u(x)$. Then set $h_v(x) \leftarrow \min_z \big[\Delta_v(x,z) + g_v(z)\big]$, $g_u(x) \leftarrow g_u(x) + h_v(x)$, $v \leftarrow u$, $u \leftarrow w$. Afterward, go to S6.

S9. [Genomic span of tree.] Set $t_r \leftarrow L$. If $j < |E|$ set $t_r \leftarrow \min\big(t_r, E[I_j]$.`left`$\big)$. Then, if $k < |E|$ set $t_r \leftarrow \min\big(t_r, E[O_k]$.`right`$\big)$. Set $s \leftarrow t_r - t_l$.

S10. [Update average migration costs.] Set $\sigma \leftarrow 0$, $n \leftarrow 0$, $s_P \leftarrow s + s_P$. Then visit each node u in a pre-order traversal and set $s_u \leftarrow s_u + s$. If $\pi_u = -1$ set $f_u(x) \leftarrow g_u(x)$ and set $\sigma \leftarrow \sigma + \min_z f_u(z)$; otherwise, set $f_u(x) \leftarrow \min_z \big[f_{\pi_u}(z) - h_u(z) + \Delta_u(z,x)\big] + g_u(x)$ and set $n \leftarrow n+1$. Then, set $F_u(x) \leftarrow F_u(x) + s \times \frac{f_u(x) - F_u(x)}{s_u}$.

S11. [Update average migration rate.] Set $\sigma_P \leftarrow \sigma_P + s \times \frac{\sigma/n - \sigma_P}{s_P}$.

S12. [Tree loop tail.] Set $t_l \leftarrow t_r$. Go to S2.

We begin in S1 by initializing the cost functions for each node to zero except in the case of sample nodes, for which we set the cost function to positive infinity for all states not equal to the observed state. We then set the parent of each node to -1 (signifying the null element) so that the initial state of the tree sequence is a forest of disconnected nodes. The average cost functions (denoted by the corresponding capital letters) are also initialized to 0 as these will be updated as we iterate over the tree sequence.

The meat of the algorithm occurs in steps S4 and S5 of the edge removal loop and in steps S7 and S8 of the edge insertion loop. Removal of an edge from node u to v will alter the node and stem cost functions along the path from u back to the root of the genealogy. Step S4 prepares for this by first subtracting the stem costs of u and v from the node costs of their respective parents. In step S5, we walk back along the path from u's parent to the root and recompute new stem costs and node costs given the updated node costs at the head of the path. Insertion of an edge from node u to v will similarly alter the node and stem cost functions along the path from u back to the root of the genealogy. Step S7 prepares for this by first substracting the stem cost of u from the node cost of its parent and then computing the stem cost of v and adding it to the node cost of u. In step S8, we walk back along the path from u's parent to the root as before and recompute new stem costs and node costs given the updated node costs at the head of the path.

Upon reaching step S9, we have finished constructing the tree together with its node cost and stem cost functions. We record the genomic span of the tree in the variable s, which will be the weight applied to the current tree in the weighted average. In S10, we perform a preorder traversal of the newly constructed tree and compute the final cost function for each node. At the same time, we increment the total weight s_u of each node by s and update the weighted averages $F_u(x)$ with the cost functions for the current tree. When we begin the traversal at the root(s) of the tree we also record the minimum migration cost σ needed to explain the sample distribution. We use this cost in S11, together with the number n of edges in the tree, to update the weighted average migration rate statistic σ_P.

Correctness of the algorithm requires that edges are removed in order of nondecreasing right genomic coordinate and decreasing time and that edges are inserted in order of nondecreasing left genomic coordinate and increasing time (time is measured backward from the present). We assume that the index vectors I and O are constructed to satisfy these conditions; these assumptions form the basis of the conditional checks in steps S3 and S6 for determining when to exit the edge removal and insertion loops.

Relation to existing work

Several existing approaches to geographic inference with tree sequences merit discussion in relation to our own work. Wohns et al. (2022) introduced a nonparametric approach that estimates ancestor locations by successively averaging the coordinates of sample locations in a postorder traversal of the ARG to their most recent common ancestor. The resulting estimates are local estimates in the sense that the inferred location of an ancestor depends only on the locations of samples that trace some portion of their ancestry to that ancestor and on the topology of the corresponding subset of the ARG. By contrast, our approach estimates the location of an ancestor using information from all samples. Because all samples share common ancestry at some time in the past, even those samples that are not direct descendants of an ancestor can be informative about that ancestor's location.

Osmond and Coop (2021) describe a likelihood method for locating genetic ancestors and estimating migration rates that is based on a model of branching Brownian motion. Their approach also uses information from all samples to

estimate ancestral locations and can optionally estimate separate migration rates for deep and shallow time horizons. Unlike our method, inference is carried out on a sample of widely spaced genealogies rather than on the full tree sequence. We note that when squared Euclidean distance weighted by inverse branch length is used as the transition cost function, the maximum parsimony reconstruction on each marginal genealogy has highest posterior probability under a Brownian motion dispersal process. In this sense, the genome-wide average reconstruction produced by our method can be viewed as a weighted average of posterior modes. Deraje et al. (2024) recently extended the model of branching Brownian motion to work with the full ARG rather than a sample of local gene trees.

Bibliography

Deraje, P., Kitchens, J., Coop, G., and Osmond, M. M. (2024). Inferring the geographic history of recombinant lineages using the full ancestral recombination graph. *bioRxiv*. doi: 10.1101/2024.04.10.588900.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, 12(5):1–22. doi: 10.1371/journal.pcbi.1004842.

Osmond, M. M. and Coop, G. (2021). Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. *bioRxiv*. doi: 10.1101/2021.07.13.452277.

Ralph, P., Thornton, K., and Kelleher, J. (2020). Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics*, 215(3): 779–797. doi: 10.1534/genetics.120.303253.

Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42.

Sankoff, D. and Rousseau, P. (1975). Locating the vertices of a steiner tree in an arbitrary metric space. *Mathematical Programming*, 9:240–246.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., and McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, 375(6583):eabi8264. doi: 10.1126/science.abi8264.

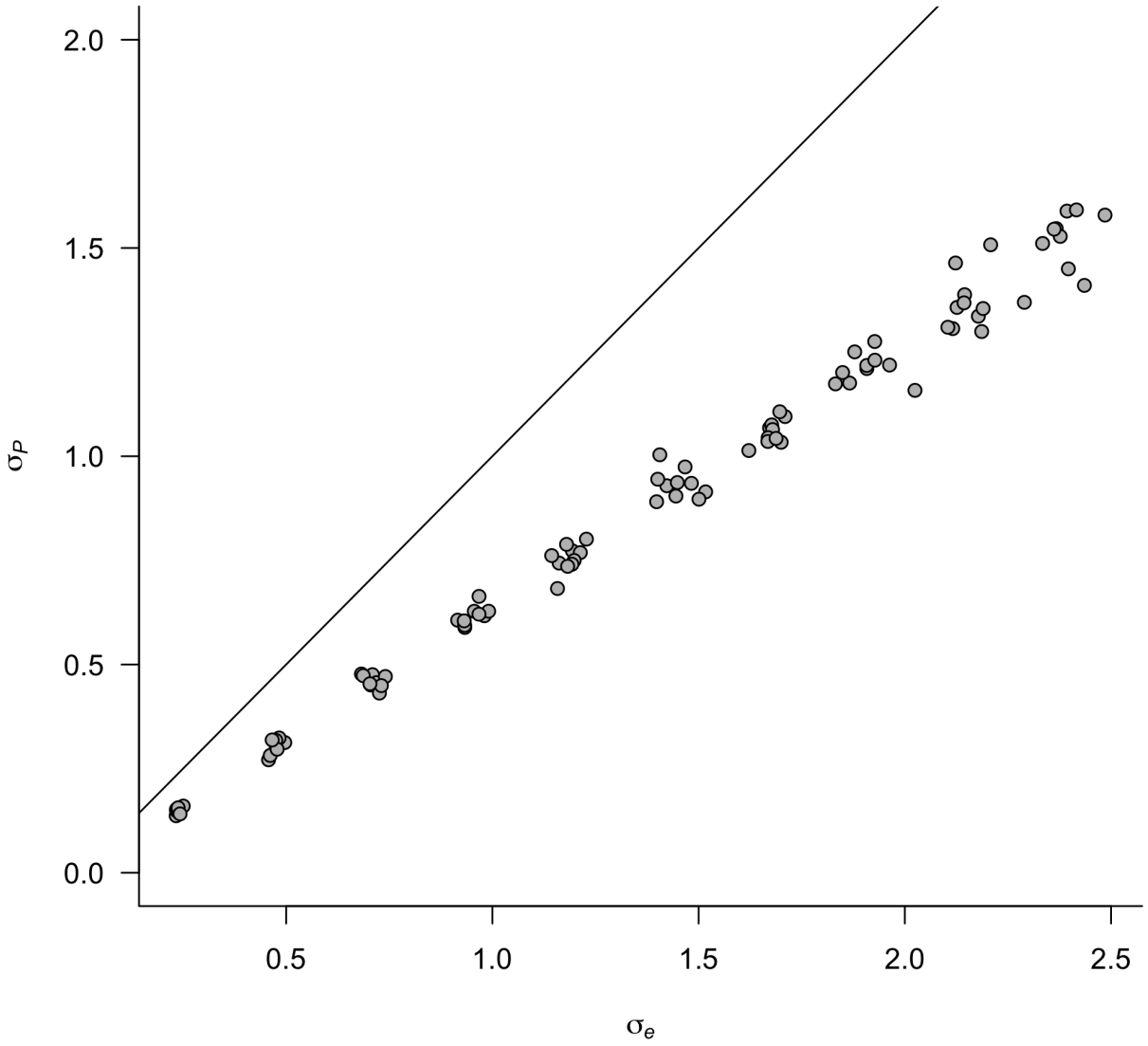

Figure S1. Migration rate estimates for a Gaussian dispersal kernel. Each point represents a single simulation generated under Gaussian dispersal with effective migration rate given on the x-axis and a parsimonious genome-wide estimate of that rate on the y-axis. The inset line shows a 1:1 relationship.

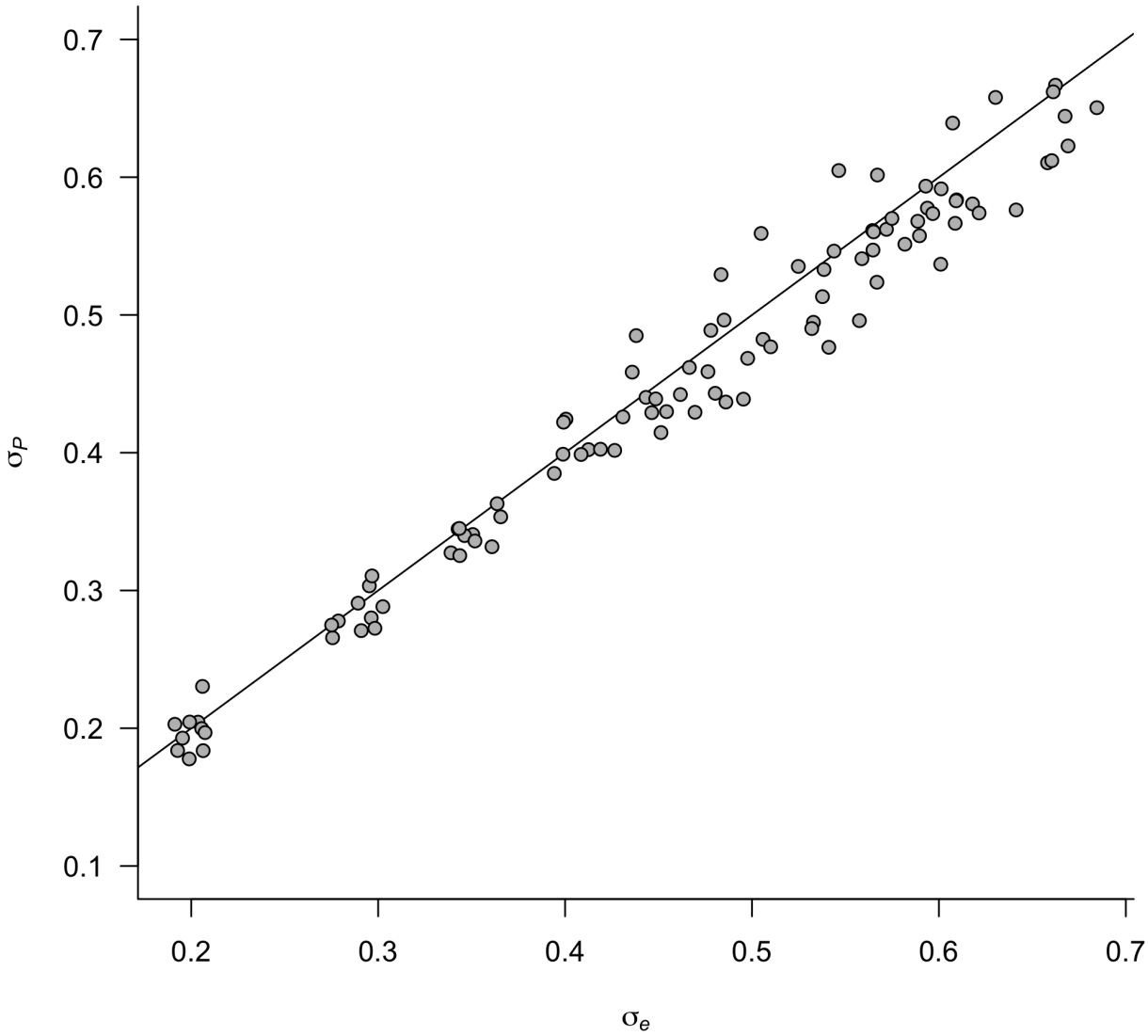

Figure S2. Migration rate estimates for a Laplace dispersal kernel. Each point represents a single simulation generated under Laplace dispersal with effective migration rate given on the x-axis and a parsimonious genome-wide estimate of that rate on the y-axis. The inset line shows a 1:1 relationship.

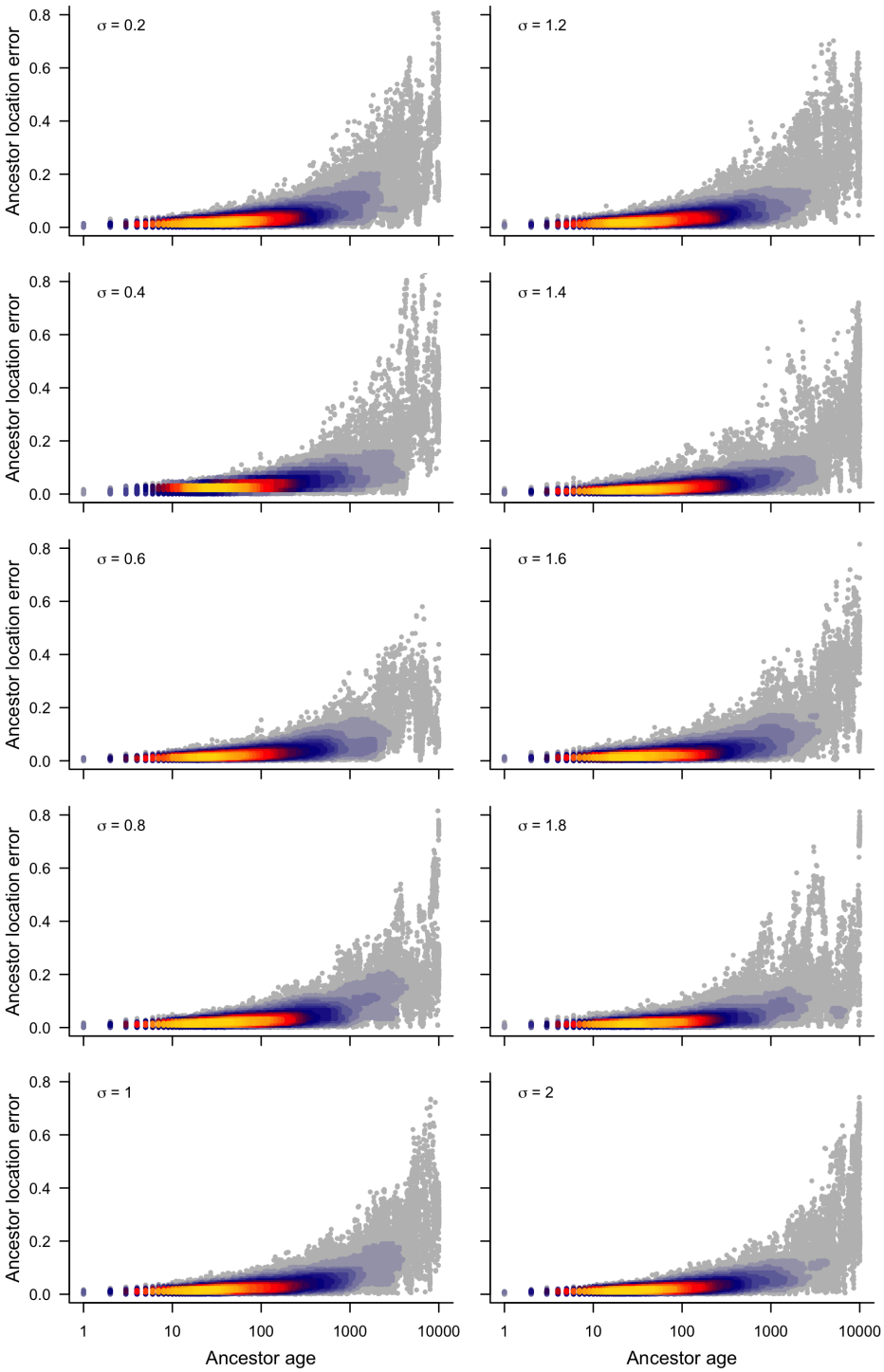

Figure S3. Ancestor location error for a Gaussian dispersal kernel. Each point represents a single genetic ancestor. Ancestor location error is measured as the distance between the estimated and the true location divided by the greatest distance separating any pair of samples. Warm colors signify a greater density of points.

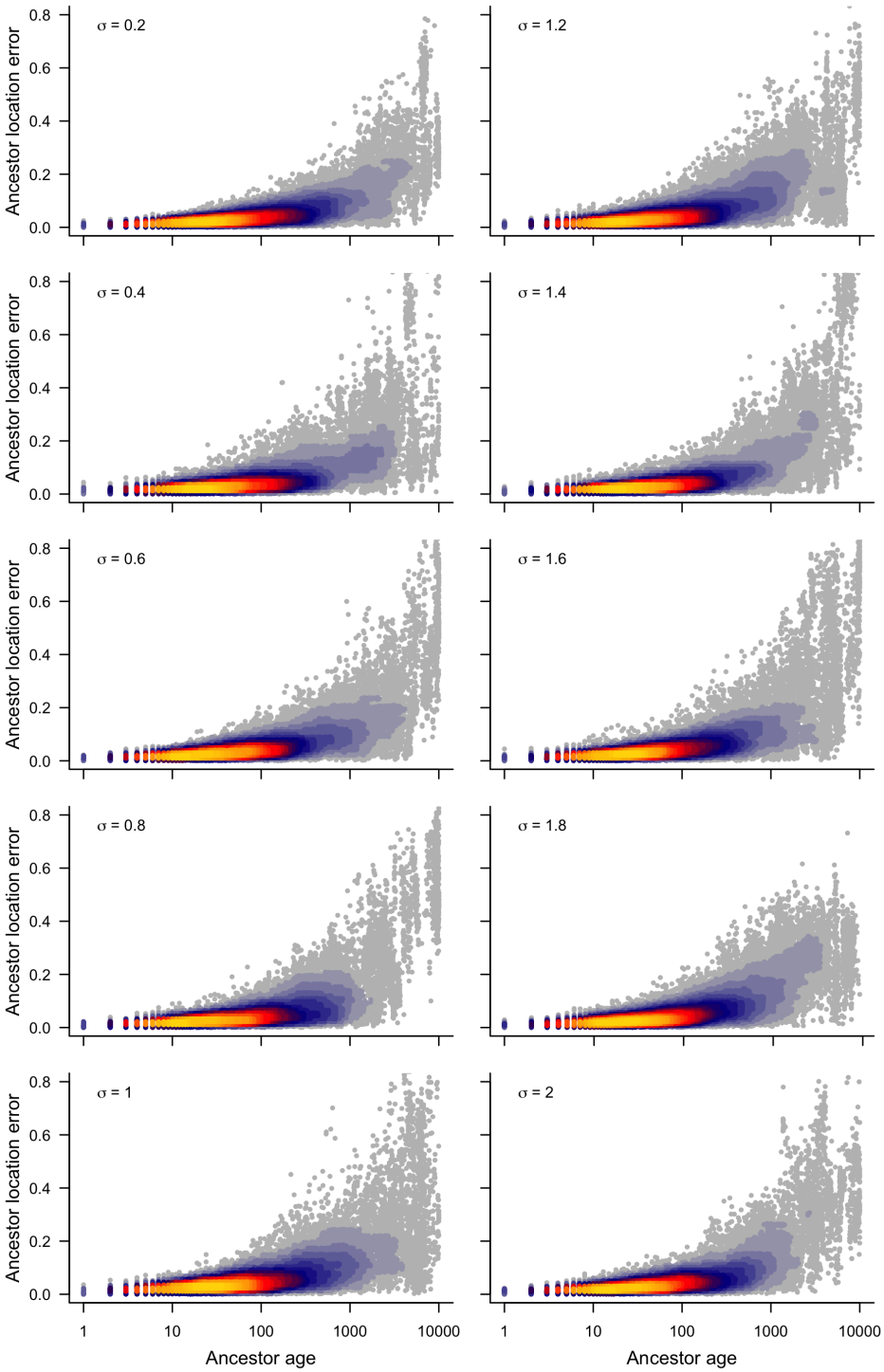

Figure S4. Ancestor location error for a Laplace dispersal kernel. Each point represents a single genetic ancestor. Ancestor location error is measured as the distance between the estimated and the true location divided by the greatest distance separating any pair of samples. Warm colors signify a greater density of points.