Statistical phylogeography: Admixture graphs and `sparg`

When I introduced the coalescent several weeks ago, I mentioned the "Out of Africa" hypothesis — the hypothesis that anatomically modern humans evolved in Africa and spread from there throughout the rest of the world. Three decades of research have strengthened support for that hypothesis, and it is now widely accepted that anatomically modern human populations left Africa and moved into other parts of the world about 50,000 years ago [4]. As they expanded, they interacted with archaic human populations, e.g., Neanderthals and Denisovans. And when human populations interact, interbreeding often occurs. The result is that 1-3 percent of human genomes from outside of sub-Saharan Africa show evidence of Neanderthal ancestry [11] and that as much as 5 or 6 percent of the human genomes from Oceania show evidence of ancestry from Denisovans [6]. When we visualize these relationships (Figure 1), the result no longer looks like a simple tree. There are lines connecting different branches representing times when there was some degree of interbreeding among populations that had previously diverged. You may have encountered methods for inferring phylogenies in previous courses. In this course we saw how `STRUCTURE` can be used to estimate patterns of admixture. How do we go about estimating trees that are admixed? Funny you should ask.

Admixture graphs

Pickrell and Pritchard [10] described the most widely used approach to estimating admixture graphs. It is implemented in `TreeMix`. At about the same time Patterson et al. [9] described a related method. I'm going to focus on the `TreeMix` approach because I am more comfortable with the underlying model.[1] Unfortunately, if you want to use `TreeMix`, you'll have to be comfortable with compiling C++ programs from source (or find a friend who can help you or who can share a copy).[2]

[1] If you're curious about why I'm more comfortable with the Pickrell and Pritchard approach, feel free to ask.

[2] The most recent version of the `TreeMix` manual notes that "TreeMix should run on any Unix or Unix-like (e.g., Linux or Mac OS X) system. It may be more difficult to get it compiled under Windows. Notice that

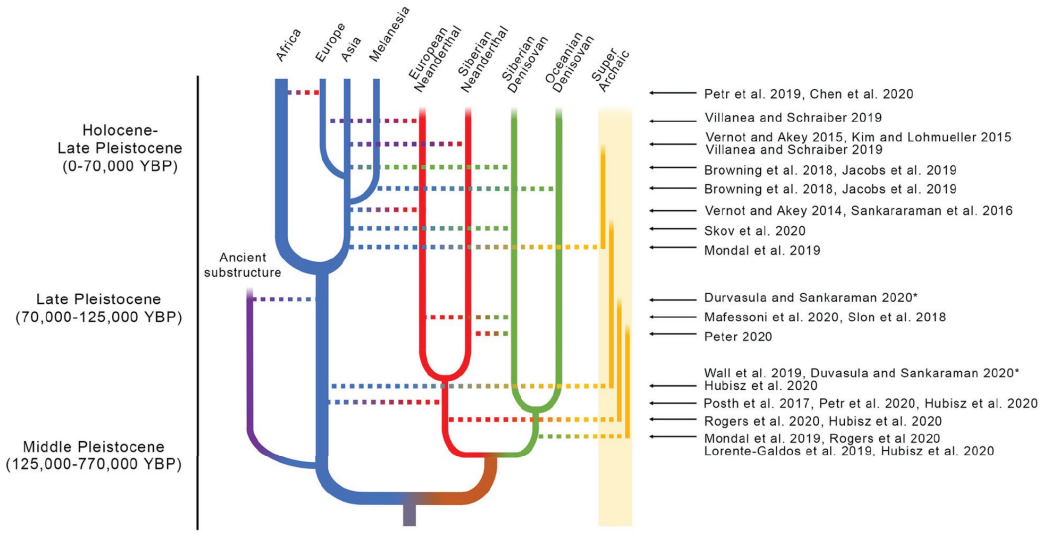

Figure 1: Periods of introgression in humah history (from [1]).

The basic idea behind `Treemix` is not too complicated, although it would be a stretch to say that it's simple. We start by assuming that the allele frequencies are changing as a result of genetic drift. Results going back to Kimura [5] tell us that the variance in allele frequency is

$$\mathrm{Var}(p_t) = p_o(1-p_0)\left(1-e^{-t/2N_e}\right) \quad ,$$

where p_t is the allele frequency in the population at time t, p_o is the initial allele frequency, t is the number of generations, and N_e is the effective population size. So long as the effective population size is large enough that allele frequency changes are relatively small from generation to generation and so long as p_o is not "too close" to 0 or 1, then we can approximate the probability distribution of allele frequencies at time t with a normal distribution:

$$\mathrm{P}(p_t|p_o,t,N_e) \sim \mathrm{N}\left(p_o, p_o(1-p_o)\left(\frac{t}{2N_e}\right)\right) \quad .$$

Now suppose we have a series of four populations related like those shown in Figure 2. As you can see, this example shows populations that have a simple tree-like relationship. Here's where the fun starts.

regardless of operating system, you'll also need to install the GNU Scientific Library and the Boost Graph Library."

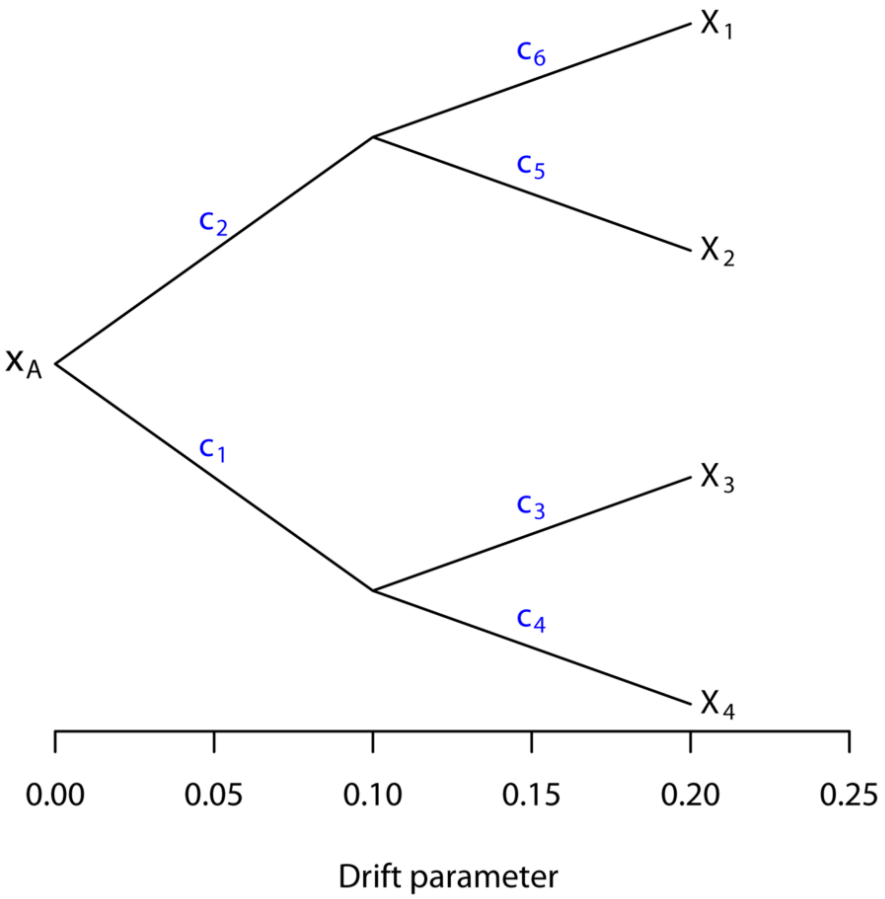

Figure 2: A purely tree-like relationship among four hypothetical populations. The allele frequencies in each population are represented by X_i. The drift parameter on the x-axis is $t/2N_e$, i.e., it's measuring time from the root of the tree to the tips in units of $1/2N_e$. A part of a figure in [10].

It's a well known fact [3] that the variance in allele frequencies (X_i in the figure) are simply

$$\begin{aligned}
\mathrm{Var}(X_1) &= (c_2 + c_6)X_A(1 - X_A) \\
\mathrm{Var}(X_2) &= (c_2 + c_5)X_A(1 - X_A) \\
\mathrm{Var}(X_3) &= (c_1 + c_3)X_A(1 - X_A) \\
\mathrm{Var}(X_4) &= (c_1 + c_4)X_A(1 - X_A) \quad ,
\end{aligned}$$

where $c_i = \frac{t_i}{2N_e^{(i)}}$, t_i is the time associated with branch i and $N_e^{(i)}$ is the effective size of the population associated with branch i. It's obvious from looking at the tree that populations 1 and 2 have been evolving independently from populations 3 and 4 from the start, while 1 and 2 have been evolving independently of one another for a shorter period of time. As a result, we expect allele frequencies in populations 1 and 2 to be more similar than those in populations 3 and 4. In fact, Pickrell and Pritchard point out that we can write the various covariances down pretty simply too:

$$\begin{aligned}
\mathrm{Cov}(X_1, X_2) &= c_2 X_A(1 - X_A) \\
\mathrm{Cov}(X_1, X_3) &= 0 \\
\mathrm{Cov}(X_1, X_4) &= 0 \\
\mathrm{Cov}(X_2, X_3) &= 0 \\
\mathrm{Cov}(X_2, X_4) &= 0 \\
\mathrm{Cov}(X_3, X_4) &= c_1 X_A(1 - X_A) \quad .
\end{aligned}$$

As a result, we can write down a multivariate probability distribution that describes all of the allele frequencies simultaneously, given the same caveats as above about the normal distribution.

$$\mathrm{P}(\mathbf{p_t}|\mathbf{p_0}, \mathbf{t}, \mathbf{N_e}) \sim \mathrm{MVN}(\mathbf{p_0}, \mathbf{\Sigma}) \quad ,$$

where boldface refers to vectors, MVN refers to the multivariate normal distribution, and $\mathbf{\Sigma}$ is the covariance matrix of allele frequencies. Since we can write down that probability distribution, you can probably imagine that it's possible to estimate the likelihood of our data given a particular tree. To get a maximum likelihood estimate of how our populations are related, assuming there's no migration, we simply have to compare the likelihoods across all possible trees and choose the one that's most likely.[3]

[3] If you know anything about estimating phylogenies, you know there is tremendous complexity buried in that "simply have to compare." Also notice that if we can get a maximum likelihood estimate, we can also get a full Bayesian posterior "simply" by providing the appropriate priors.

Now suppose we allow migration from one of our populations into another. The simple example Pickrell and Pritchard provide (Figure 3 shows a single migration from the lineage leading to population 2 into population 3, labeling the source population as Y and the destination population as Z. As you can see in Panel D of the figure, the migration event changes the structure of the covariance matrix. Since all the migration event does is to change the covariance matrix, we can once again explore parameter space and find the network that maximizes the likelihood. When we do so, not only do we have estimates for population relationships and effective population sizes but also for the timing and direction of migration events. Estimating admixture is, however, even more challenging than estimating a population phylogeny. The number of alternative configurations explodes rapidly with more than 4-5 populations, making heuristic searches necessary. Molloy et al. [7] recently described a new approach that builds on `TreeMix` and seems to avoid getting stuck in a local optimum. Since the basic approach is the same and this isn't a course in computational biology, we won't discuss it further, but you should investigate it if you use admixture graphs in any of your work.

Estimating dispersal and ancestral geography

As you can see, admixture graphs provide a very flexible approach to understanding the history of populations. But they do have one significant limitation. We have to know ahead of time which individuals belong in which populations, just as we did with F-statistics, and just as with `STRUCTURE` gave us a way to look at population structure without pre-assigning individuals to populations, there's a way of looking at ancestry that uses individuals rather than pre-defined populations [8]. As with admixture graphs, the mathematics lying behind the approach gets pretty hairy, but the basic idea is pretty simple (Figure 4).

- At any position along a genome, we can construct a phylogenetic tree showing the genealogical relationship among all chromosomes in the sample at that location.[4]
- Individuals disperse randomly through space with the distance of an offspring from its mother given by a bivariate normal distribution with a mean of 0 and a covariance matrix $\mathbf{\Sigma}$. In any real sample, glacial migrations, barriers to dispersal, or the opening of new habitat will cause some aspects of the dispersal history not to be well approximated by this model of Brownian motion, so we only use parts of the tree from the first step that are more recent than these events to estimate dispersal parameters.[5]

[4]Notice that I wrote "chromosomes", not individuals, because the different allele copies within an individual may have different genealogical histories.

[5]Notice that while this approach means that the Brownian motion model for dispersal is a better fit to

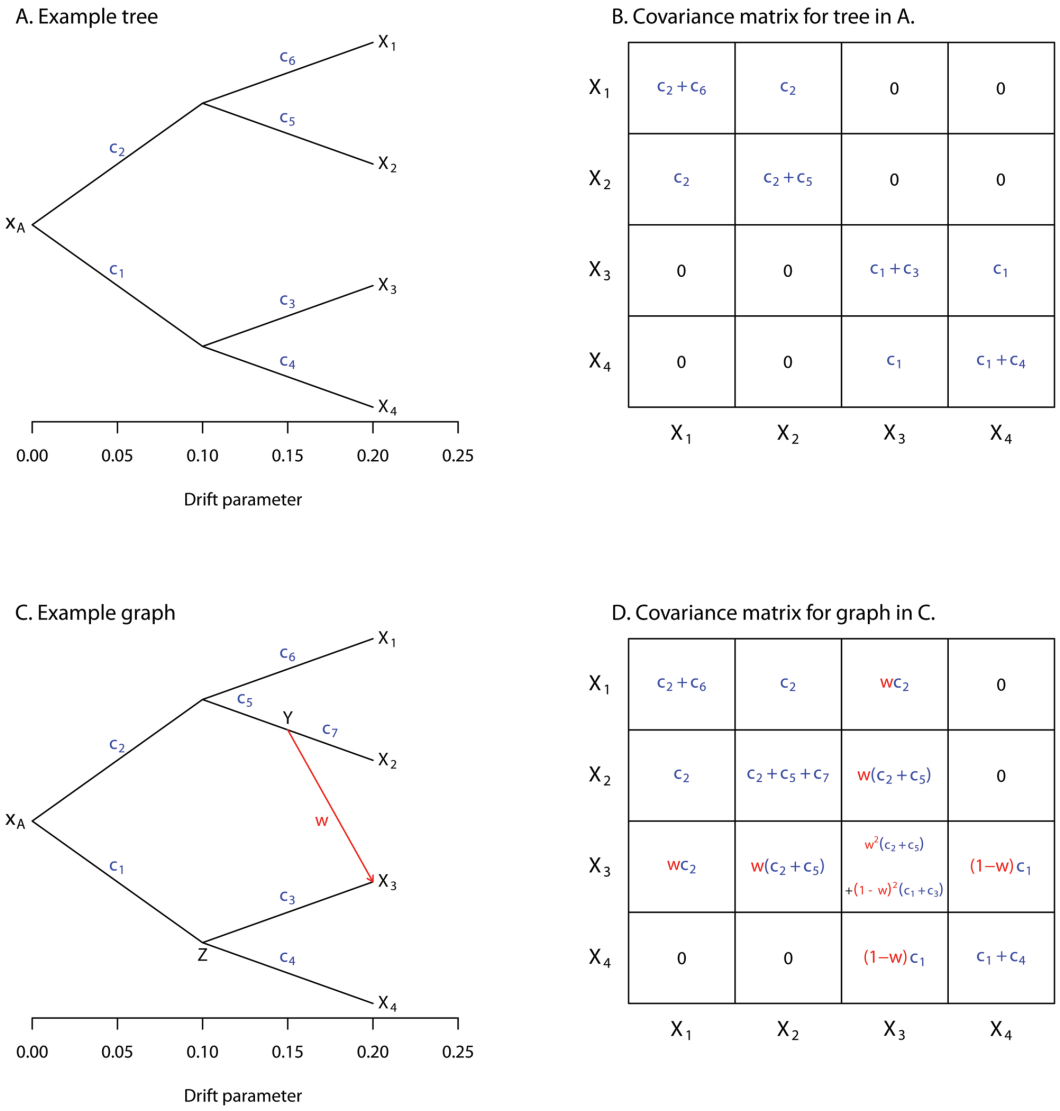

Figure 3: Illustrating the covariance matrices of admixed and unadmixed populations. From [10].

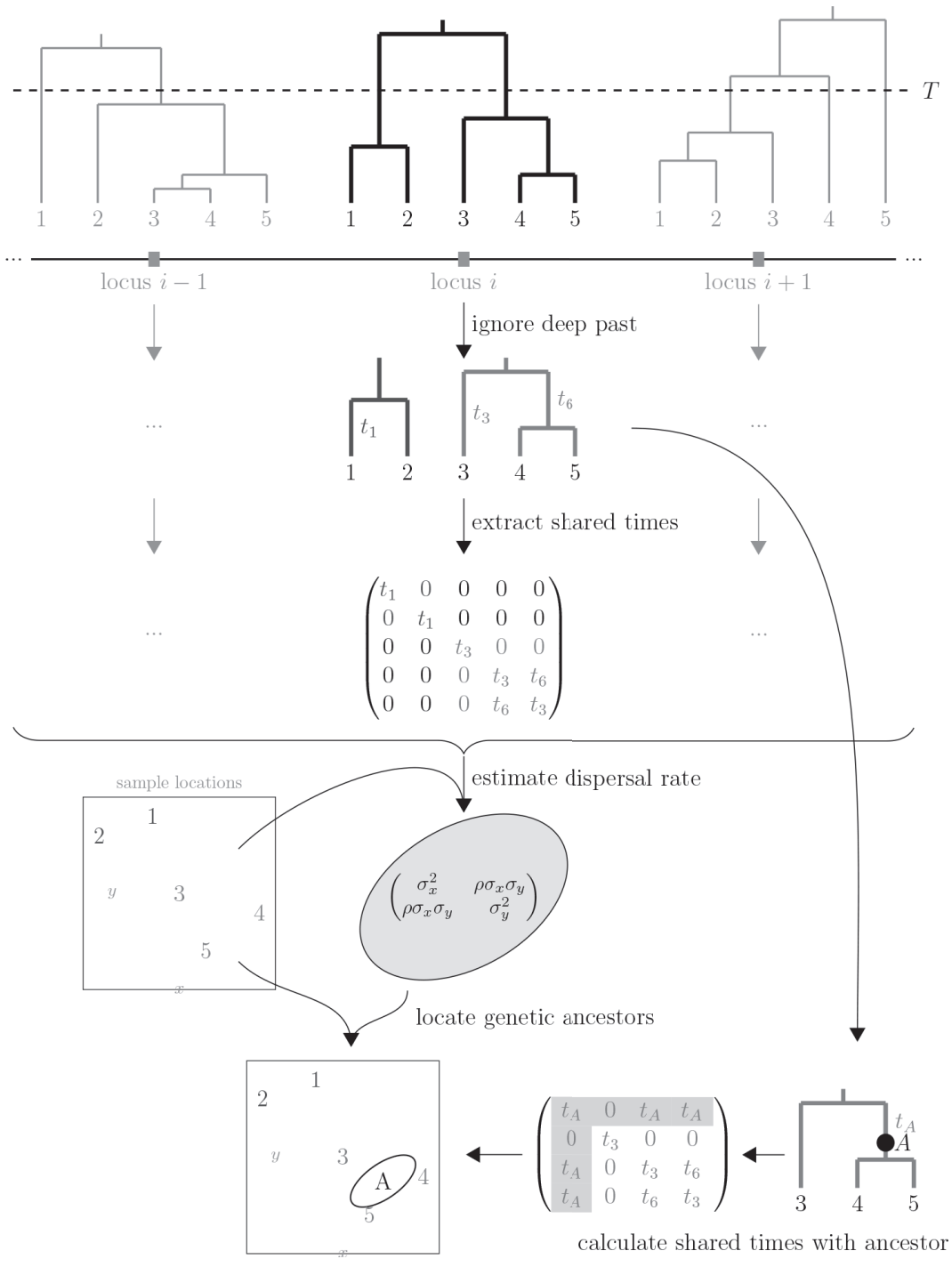

Figure 4: Conceptual overview of the process for estimating the spatial position of ancestors (from [8]).

- Given the estimates of time to a common ancestor between two individuals, the spatial location of those individuals, and the dispersal rate, we can estimate the spatial location of the ancestor.

This method implicitly assumes that differences are selectively neutral.[6] Although we could try this approach with data from only one locus, the results are unlikely to be informative for two reasons. First, there is a lot of uncertainty associated with our estimate of phylogenetic relationships at one locus. Second, because the coalescent history of unlinked loci will differ even though the effective population size and the patterns of migration that affect different loci are the same. But since the patterns of migration *are* the same across different loci and since the effective population size *is* the same across loci, we can combine information across loci to get better estimates of the dispersal rates. Since we estimate the location of ancestors at every locus, we end up with a distribution of ancestral locations rather than a single estimate. Osmond and Coop also point out that we can define different "epochs" in which to estimate dispersal rates and ancestors. This allows the dispersal rate to vary over time. All of this is available in a Python package, `sparg`, which should run on any platform with phython3 (`https://github.com/mmosmond/sparg`).

An example from *Arabidopsis*

Plant geneticists have studied *Arabidopsis thaliana* extensively. Alonso-Blanco et al. [2] reported results derived from sequencing 1135 different wild accessions derived from Eurasia and North Africa. Osmond and Coop used `sparg` to explore historical patterns of dispersal and the geographical location of ancestors using this data set. They first estimated dispersal rates in both a one-epoch model and in multi-epoch models. As you can see in Figure 5, the estimates of dispersal rates are very similar across all of the loci. In addition, the per-generation rate of east-west dispersal (σ^2_{long}) is about 10 times higher than north-south dispersal (σ^2_{lat}), and the correlation between the two rates (ρ) is relatively small. Comparison among the scenarios suggests that the 4-epoch model is the best fit to the data, suggesting that the rate of dispersal in the last 10 generations is substantially greater than it was earlier and that dispersal between 10 and 1000 generations ago is greater than it was more than 1000 generations ago.

Now that we have a good idea *when* dispersal happened, let's see *where* it happened. As you can see in Figure 6, much of the estimated dispersal over the last 10-100 generations

the data, it also means that we can't use this approach to study events that involve ancient dispersal, like early modern human movements out of Africa.

[6] Remember: This doesn't mean that there aren't any fitness differences, only that the product of the selection coefficient associated with any of those differences and the population size is less than one, implying that the evolutionary dynamics are rougly similar to those of a purely neutral locus.

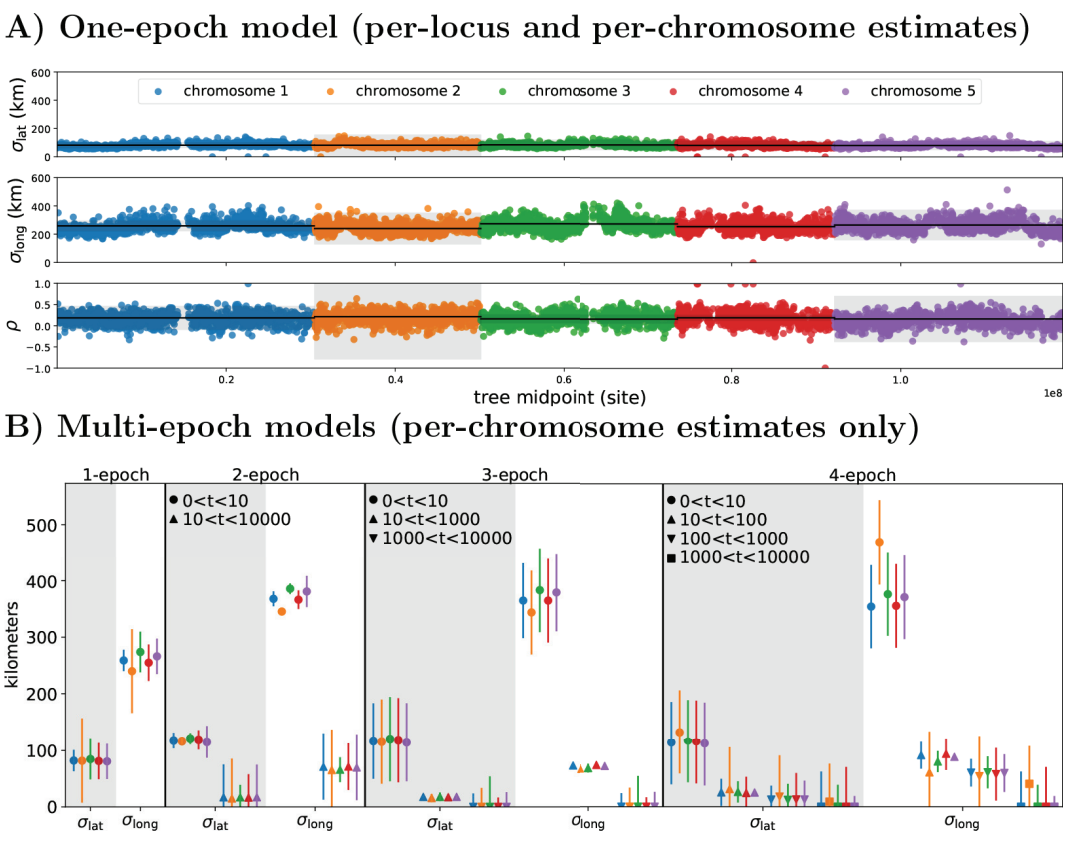

Figure 5: Estimates of dispersal rates in *Arabidopsis thaliana* in both one-epoch (panel A) and multi-epoch (panel B) models (from [8]).

didn't move individuals very far. In addition, it's a little hard to see, but if you zoom in on the figure and focus on the purple colors, you'll notice that most of the lines leading from the dots (current locations) point towards the center of Europe. This pattern is particularly clear in for the 100-generation ago ancestral location of samples from Scandinavia. There are, however, a few individuals that moved very long distances. Individual 9627, for example, seems to have an ancestor 10 generations ago that was more than 3000km to the east of its current location, and its ancestor seems to have been more than 4000km to the east 100 generations ago.

References

[1] K. D. Ahlquist, Mayra M. Bañuelos, Alyssa Funk, Jiaying Lai, Stephen Rong, Fernando A. Villanea, and Kelsey E. Witt. Our tangled family tree: New genomic methods offer insight into the legacy of archaic admixture. *Genome Biology and Evolution*, 13(7), 2021.

[2] Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M. Borgwardt, Jun Cao, Eunyoung Chae, Todd M. Dezwaan, Wei Ding, Joseph R. Ecker, Moises Exposito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G. Grimm, Angela M. Hancock, Stefan R. Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A. Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Cheng-Ruei Lee, Dazhe Meng, Todd P. Michael, Richard Mott, Ni Wayan Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Yu Novikova, F. Xavier Picó, Alexander Platzer, Fernando A. Rabanal, Alex Rodriguez, Beth A. Rowan, Patrice A. Salomé, Karl J. Schmid, Robert J. Schmitz, Ümit Seren, Felice Gianluca Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M. Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel, and Xuefeng Zhou. 1,135 genomes reveal the global pattern of polymorphism in ¡em¿arabidopsis thaliana¡/em¿. *Cell*, 166(2):481–491, 2016.

[3] L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis models and estimation procedures. *American Journal of Human Genetics*, 19:233–257, 1967.

[4] Monika Karmin, Lauri Saag, Mário Vicente, Melissa A. Wilson Sayres, Mari Järve, Ulvi Gerst Talas, Siiri Rootsi, Anne-Mai Ilumäe, Reedik Mägi, Mario Mitt, Luca Pagani, Tarmo Puurand, Zuzana Faltyskova, Florian Clemente, Alexia Cardona, Ene Metspalu, Hovhannes Sahakyan, Bayazit Yunusbayev, Georgi Hudjashov, Michael DeGiorgio,

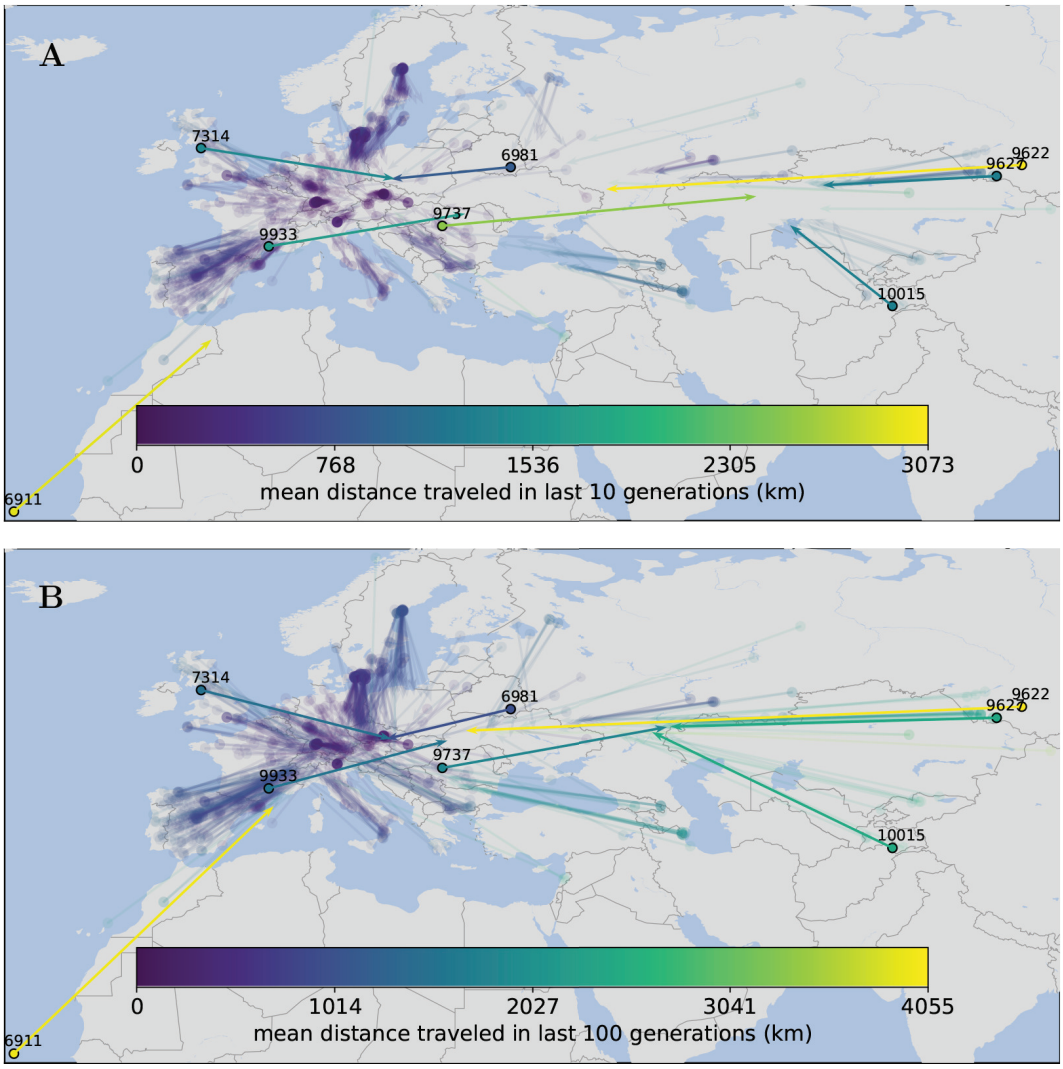

Figure 6: Estimates of the ancestral location of *Arabidopsis thaliana* accessions 10 and 100 generations ago (from [8]).

Eva-Liis Loogväli, Christina Eichstaedt, Mikk Eelmets, Gyaneshwer Chaubey, Kristiina Tambets, Sergei Litvinov, Maru Mormina, Yali Xue, Qasim Ayub, Grigor Zoraqi, Thorfinn Sand Korneliussen, Farida Akhatova, Joseph Lachance, Sarah Tishkoff, Kuvat Momynaliev, François-Xavier Ricaut, Pradiptajati Kusuma, Harilanto Razafindrazaka, Denis Pierron, Murray P. Cox, Gazi Nurun Nahar Sultana, Rane Willerslev, Craig Muller, Michael Westaway, David Lambert, Vedrana Skaro, Lejla Kovačević, Shahlo Turdikulova, Dilbar Dalimova, Rita Khusainova, Natalya Trofimova, Vita Akhmetova, Irina Khidiyatova, Daria V. Lichman, Jainagul Isakova, Elvira Pocheshkhova, Zhaxylyk Sabitov, Nikolay A. Barashkov, Pagbajabyn Nymadawa, Evelin Mihailov, Joseph Wee Tien Seng, Irina Evseeva, Andrea Bamberg Migliano, Syafiq Abdullah, George Andriadze, Dragan Primorac, Lubov Atramentova, Olga Utevska, Levon Yepiskoposyan, Damir Marjanović, Alena Kushniarevich, Doron M. Behar, Christian Gilissen, Lisenka Vissers, Joris A. Veltman, Elena Balanovska, Miroslava Derenko, Boris Malyarchuk, Andres Metspalu, Sardana Fedorova, Anders Eriksson, Andrea Manica, Fernando L. Mendez, Tatiana M. Karafet, Krishna R. Veeramah, Neil Bradman, Michael F. Hammer, Ludmila P. Osipova, Oleg Balanovsky, Elza K. Khusnutdinova, Knut Johnsen, Maido Remm, Mark G. Thomas, Chris Tyler-Smith, Peter A. Underhill, Eske Willerslev, Rasmus Nielsen, Mait Metspalu, Richard Villems, and Toomas Kivisild. A recent bottleneck of y chromosome diversity coincides with a global change in culture. *Genome Research*, 25(4):459–466, 2015.

[5] M. Kimura. Random genetic drift in multi-allelic locus. *Evolution*, 9:419–435, 1955.

[6] Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, Flora Jay, Kay Prüfer, Cesare de Filippo, Peter H. Sudmant, Can Alkan, Qiaomei Fu, Ron Do, Nadin Rohland, Arti Tandon, Michael Siebauer, Richard E. Green, Katarzyna Bryc, Adrian W. Briggs, Udo Stenzel, Jesse Dabney, Jay Shendure, Jacob Kitzman, Michael F. Hammer, Michael V. Shunkov, Anatoli P. Derevianko, Nick Patterson, Aida M. Andrés, Evan E. Eichler, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. A high-coverage genome sequence from an archaic denisovan individual. *Science*, 338(6104):222–226, 2012.

[7] Erin K. Molloy, Arun Durvasula, and Sriram Sankararaman. Advancing admixture graph estimation via maximum likelihood network orientation. *Bioinformatics*, 37(Supplement_1):i142–i150, 2021.

[8] Matthew M Osmond and Graham Coop. Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. *bioRxiv*, page 2021.07.13.452277, 2021.

[9] Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

[10] Joseph K. Pickrell and Jonathan K. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genetics*, 8(11):e1002967, 2012.

[11] Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L. F. Johnson, Hélène Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, 2014.

Creative Commons License