White Paper 23-14

Ancestry Timeline

Authors:

Katarzyna Bryc `kbryc@23andme.com`
Eric Y. Durand
Joanna Mountain

Created: 1 December 2016
Updated: 10 March 2017
Summary:

Ancestry Timeline is a 23andMe feature that enables customers to find out, for each of the ancestries they carry, when they may have had an ancestor in their genealogy who was likely to be a non-admixed representative of that population. This document is a technical description of the statistical methodology supporting this feature.

1 Ancestry Timeline context

Each variant in a person's genome can be traced back to ancestors in previous generations. Often nearby pairs of variants (genotypes) trace to the same ancestor. In some cases the geographic location or cultural affiliation of that ancestor can be inferred by comparing the variants to the variants of reference individuals from around the globe. In Ancestry Composition, 23andMe has a proprietary system for making that geographic or cultural inference by resolving the ancestral origins of chromosomal segments of different ancestries. This document describes an additional interpretation that takes advantage of information on the lengths of the chromosomal segments from an ancestry, that trace back to a particular geographic location or cultural group. The number of segments and their lengths reflect the number of generations since the particular ancestor of that group or location lived.

2 Ancestry Timeline background

Human population history, including population splits, migrations. and mixture events, were primary shapers of current patterns of genetic variation. Admixture, or the mixture of populations that previously were long separated, has recently been shown to have occured pervasively throughout human history. Recent literature highlights the many examples that illustrate the prevalence of admixture events in human history, including: gene flow from Neanderthals (Green *et al.*, 2010), archaic population mixtures in Europe (Lazaridis *et al.*, 2014), gene flow across Austronesia (Lipson *et al.*, 2014), and modern admixture in the Caribbean (Moreno-Estrada *et al.*, 2013).

Population-based methods for inferring admixture times

With the advances in genome-wide high-density genotype data, population genetics has witnessed an increase in the number of statistical methods for inferring not just proportions of admixture from each parent population (Falush *et al.*, 2003; Alexander *et al.*, 2009), but also the timing of these population mixture events. These methods, which estimate the date of population mixing, or "admixture date" rely on one of several signals visible through genetic data:

1. The size and number distribution of migrant segments, eg. Pool & Nielsen (2009); Gravel (2012)

2. The decay in correlations between nearby markers, eg. Loh *et al.* (2013)

3. The variance in proportions of ancestry among individuals, eg. Goldberg *et al.* (2014)

These methods have been shown to perform well in reconstructing the history of admixture events, typically using genotype data from a set of individuals from a population of interest, and one or more sets of individuals from populations that serve as proxies for the parent populations.

Adaptation of population genetic methods to individual estimates

Here, we wish to learn about the history of a single individual leveraging their genetic data. We aim to estimate their "admixture" timing, namely, when their ancestors from divergent populations began to mix. Put another way, estimating the date of an individual's admixture is similar to asking how many generations ago an individual most recently had an ancestor that was fully from a particular population. This estimate may be of interest, either as a tool for learning about one's genealogy, in figuring out which ancestors a particular ancestry may have been inherited from, or for piecing together the history of their likely migrations.

In our method, we use the signals of (1.) above. We leverage our existing Ancestry Composition results, which provide estimates of ancestry along each position of the genome from 31 world-wide populations (Durand *et al.* (2014)). Unlike many previous methods that use ancestry segments, our method was developed to provide inference for a single genome.

Challenges to development

By estimating admixture dates for a single individual rather than a population, fundamentally, the single biggest challenge in implementation is the reduction in data available to generate the estimate. As with previous population-based methods, the number of possible population histories is infinite, and therefore, all previous studies implement some parameterization or reduced set of possible histories over which to search (for example, a single "pulse" model of admixture, or a "continuous" model of constant migration).

Likewise, though there are many possible ways that one can inherit ancestry (from any number of genealogical ancestors going back in time), to reduce our

possible parameter space, we assume a model where exactly one ancestor contributed an ancestry. Though simplistic, this model allows us to provide an estimate using just a single genome, and we make some allowances for violations of our model assumptions in implementation.

Brief guide to interpretation of results

In the simplest case, when an ancestry is indeed introduced by one ancestor g generations ago, and is well-captured by a single population ancestry, the admixture date is designed to capture that date. However, there are several caveats to this direct interpretation. In many cases, individuals from some world-wide population may themselves be highly admixed, obfuscating the time to when this ancestry may have first been introduced. The admixture date provided is based on the ancestry segments estimated by Ancestry Composition, and is, consequently, dependent on their accuracy and specificity for accurate date estimation. Any genealogical history or ancestries that are not well captured by Ancestry Composition estimates may result in poor admixture date estimation, which typically results in earlier estimated dates of admixture.

Secondly, the admixture date is based on all segments of a particular ancestry. If multiple genealogical ancestors contributed independently, the admixture date may reflect these multiple ancestors in a complex way. If many segments, from independent ancestors, recombine to form longer segments, the estimated admixture date may be shifted towards a more recent date. This is especially likely in the case when segments cover over 50% of a genome. On the other hand, if many older genealogical ancestors contribute discrete, shorter segments, the estimated admixture date may be pushed back, reflecting a weighted average over the multiple ancestors' generations.

Lastly, it is important to note that the inheritance of segments in one genome from a genealogical ancestor is a highly stochastic process, resulting in overlapping inheritance patterns that are not distinguishable the further back in time you go, even under otherwise ideal conditions. [For a great discussion, see the recent Coop Lab blog post[1]]. Because some amount of uncertainty is inherent in the data, we present admixture date estimates as ranges that take into account for some of this inherent uncertainty.

[1] http://gcbias.org/2013/11/11/how-does-your-number-of-genetic-ancestors-grow-back-over-time/

A note on translating generations to years Population geneticists have estimated that the average generation time, or the number of years, on average, between the birth of an individual and their child's birth, is about 29 to 30 years Jobling *et al.* (2013). Contrary to popular belief, this estimate seems to be true even going back in time hundreds and thousands of years. So if we wish to crudely translate between the admixture date in generations to years, we often multiply by 30. This of course, represents an average, and may not be accurate for any particular genealogy, and to translation to absolute dates (e.g. 1790) is dependent on an indvidual's present age. (When someone's great-grandparents lived is quite different whether that someone is 9 or 90.) However, this translation from generations to years may be helpful in providing a timeframe for events (e.g. 10 generations ago is likely after the Mayflower sailed but before George Washington wass born).

3 Method overview

At each generation, during meiosis, homologous chromosomes recombine, shuffling up the material that gets passed along to offspring. Likewise, the ancestry associated with those chromosomes gets shuffled, and "after migrant chromosomes enter a population, they are progressively sliced into smaller pieces by recombination. Therefore, the length distribution of 'migrant tracts' contain information about historical patterns of migration," or, in our case, the time since a genealogical ancestor introduced that ancestry (Pool & Nielsen, 2009). Figure 1 illustrates the recombination process and the breakdown of segments that get passed from one generation to the next over eight generations.

Some useful definitions:

Ancestry segment a section of the genome, measured in genetic distance, or centiMorgans (cM), defined by the start and end of a continuous ancestry assignment.

Admixture date shorthand notation for the time when you last had an ancestor fully from a population.

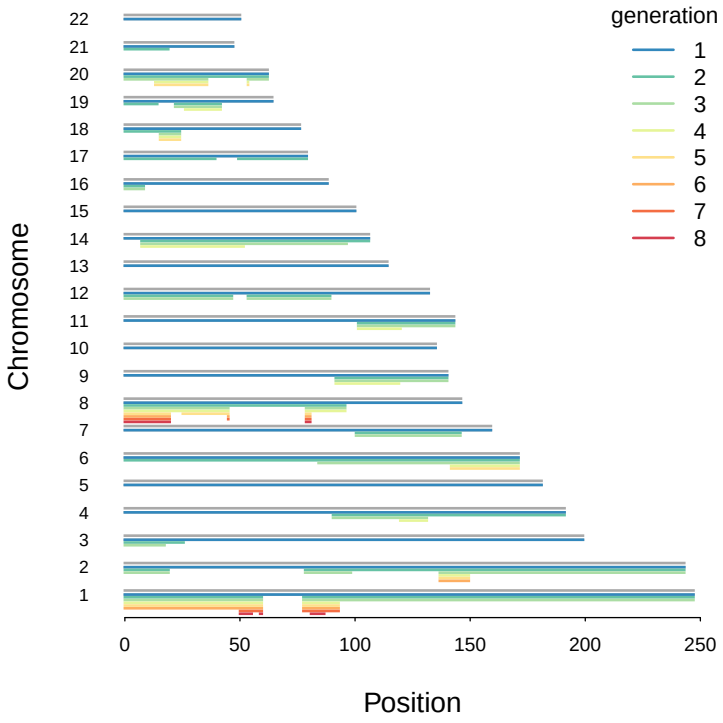

Figure 1: **Segments from an ancestor traced over eight generations.** Simulated results illustrate that in the first generation, whole chromosomes are inherited, and over each subsequent generation the segments become progressively shorter and fewer in number. Segments for each subsequent generation are shown below the previous painted segments, and are colored by shade indicating generation.

Summary

We used forward simulations to generate expected segment distributions, assuming a Poisson model of recombination events and no recombination interference. We calculated summary statistics from the data generated viasimulations, which we then used to limit the feasible generation range, based on summary statistics that are more than twice the simulated statistic.

We then found the maximum likelihood estimate (MLE) of the most like generation using an Exponential distribution model of segment lengths, and extrapolated to obtain a range of admixture dates from the MLE based on simulated estimate accuracy. The mathematical details of generating our estimates follow below.

Model assumptions

1. Ancestry Composition proportions and segment lengths capture the true levels of ancestry from each population.

2. Each ancestry is introduce by a single ancestor g generations ago. Though obviously not the case for most complex admixture events (or for any ancestry inherited from both parents), this assumption allows for the simplification of statistical calculations.

Technical challenges that violate model assumptions

- Poor recall for some Ancestry Composition populations or imperfect phasing may lead to subsequent errors in the date estimates. Likewise, ancestry that is "broadly" assigned rather than population-assigned may similarly impact estimates. Typically this results in an underestimate of ancestry, which pushes back the admixture date.

- "Choppy" segment data, as a result of uncertain ancestry assigments. When the assignment score drops below the threshold, a segment may break. When a region of the genome is not matched particularly well by a reference population, frequently this results in many such drops in scores, leading to broken, or choppy, short segments of ancestry. This in turn will lead to an older admixture date, as the segments appear shorter.

- Violations of the "one genealogical ancestor" assumption. If, instead, multiple independent ancestors in your genealogy contribute an ancestry, the segments of the same ancestry may recombine to create longer, not shorter, segments that suggest a more recent admixture date. It also typically results in a greater number of segments, invalidating any method that uses count of segments without accounting for multiple ancestors.

Similarity to inference of relationship from segments shared Identical-By-Descent (IBD) The size and number distribution of ancestry segments, under the single ancestor scenario, is akin to identifying kinship via IBD sharing. This problem has most recently been worked on by Huff *et al.* (2011) and Hill & White (2013).

4 Method steps

For each ancestry carried by an individual, we generate a distribution of ancestry segment genetic lengths, based on "Best Guess" Ancestry Composition estimates that use a threshold of 0.

1. Reduce feasible admixture date range using simulated summary statistics.

Forward recombination simulations generate summary statistics We use forward simulations, based on a Poisson-model of recombination events and genetic lengths of autosomal chromosomes to estimate summary statistics for segments inherited from a single ancestor g generations in the past. We run forward simulations using an R algorithm, first proposed by Luke Jostins [2], to generate statistics from 10,000 forward simulations. The statistics we tracked are:

- Proportion of genome covered by inherited segments
- Number of segments
- Number of chromosomes bearing segments
- Length of longest segment

Reduce feasible generation space For each customer, we reduce the feasible range, by excluding a generation from the feasible dates, if the customer's summary statistic falls outside the range of the simulated data for that generation. Since multiple ancestors violates the simulation model, and increases the amount of ancestry an individual carries, we only reduce the feasible dates using the lower bound of each metric. Note that in our simulation, the first admixed generation is deterministic, so we allow for a factor of two difference from the summary statistic, to relax our algorithm to allow room for imperfect ancestry assignments, and to conservatively reduce the feasible dates. Multiple older genealogical ancestors are expected to have lesser impact on the longest segment, so we exclude a generation from the feasible dates if the length of the longest segment is greater than twice the largest simulated length. We allow for twice the length to conservatively reduce the feasible date space.

[2] http://www.genetic-inference.co.uk/blog/2009/11/how-many-ancestors-share-our-dna/

2. Estimate the admixture date

Estimate the maximum likelihood estimate (MLE) in feasible range We calculate the likelihood of observing the segment data as a function of the number of generations since admixture, modeling the lengths of the segments as an exponential, namely

$$L(g|X) = \prod_{x_i} g * e^{(-g*x_i)}$$

Generate admixture date range from MLE As you go further back in time, the stochasticity of what you inherit from an ancestor increases, making it more and more difficult to pinpoint the generation that they were present in your genealogy. For example, you may have inherited a long segment by chance from a distant ancestor, or gotten very short segments from a recent ancestor. As a result the MLE is often not the true generation. We accommodate this randomness by providing a range of generations that a person with this ancestry may have been present in your genealogy.

5 Evaluation of method

We evaluated our algorithm using segment data created by an independent set of forward simulations. Using segment data from 1,000 simulated individuals, we find that the correct bin (containing the true generation time) is estimated 94.5% of the time. We use the 2.5 and 97.5 quantiles of the distributions of statistics as our thresholds for reducing the feasible range. A table of accuracy per generation is shown in Table 1.

6 Frequently Asked Questions

Why can't we just use the proportion of ancestry for inference? Often, people make a back-of-the-envelope calculation, where the amount of ancestry you inherit is about $\frac{1}{2^g}$, where g is the number of generations since your ancestor. This works well generally, and on average. There are two reasons why we use a more complex model. First, the amount of ancestry you inherit from an ancestor is very stochastic, and we need to take this into account. Second, and perhaps more importantly, having multiple ancestors that carry an ancestry will greatly influence the estimates from ancestry proportions,

Table 1: **Admixture date estimates of simulated data.** Using segment data from simulations, we tested the ability of the admixture mapping algorithm to estimate the correct bin of generations over 2,000 simulated individuals for each generation. On average, across generations, we witness an average of 94.5% accurate identification.

True generation	Accuracy (%)
1	100.0
2	99.6
3	99.0
4	93.0
5	91.1
6	99.7
7	99.8
8	98.1
9	86.0
10	89.5

making an ancestry appear more recent. Using segment lengths allows us to better infer the true generation time, and is more robust to multiple contributing ancestors.

Why isn't the X chromosome used in calculations? The X chromosome has a particular inheritance pattern that differs between men and women, which results in recombination patterns and proportions of ancestry that are quite complex depending on the male and female line of descent of your ancestry. Rather than model these inheritance patterns, we simply exclude the X chromosome for both simulations and likelihood calculations. Hence, any ancestry that is only found on the X chromosome is excluded from inference. We aim to improve our estimation method in the future to allow modeling the X chromosome.

Why do the date estimates not match what I know about my X ancestry? As noted above, we assume generations are of length about 30 years. If some of your ancestors had their children at much younger or older ages, then our estimate of the number of generations converted into years will not match dates you know to be the case from family or historical records. An important caveat of the Ancestry Timeline feature is that it assumes that your ancestry from each population originally comes from a single (recent or dis-

tant) ancestor. It also says nothing about where a particular ancestor was bornonly their genetics. Therefore, a British/Irish ancestor born in Philadelphia would look the same as one born in Dublin. Ancestry Timeline also has the same limitations as Ancestry Composition, in that it doesn't report contributions by recently admixed populations (e.g. "Mexican" ancestry is reported as a mixture of European, Native American, and African ancestry) and trace ancestries (e.g. less than 0.5%) should be taken with a grain of salt.

References

Alexander, David H, Novembre, John, & Lange, Kenneth. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**(9), 1655–1664.

Durand, Eric Y, Do, Chuong B, Mountain, Joanna L, & Macpherson, J Michael. 2014. Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution. *bioRxiv*, 010512.

Falush, Daniel, Stephens, Matthew, & Pritchard, Jonathan K. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**(4), 1567–1587. Comparative Study.

Goldberg, Amy, Verdu, Paul, & Rosenberg, Noah A. 2014. Autosomal admixture levels are informative about sex bias in admixed populations. *Genetics*, genetics–114.

Gravel, Simon. 2012. Population genetics models of local ancestry. *Genetics*, **191**(2), 607–619.

Green, Richard E, Krause, Johannes, Briggs, Adrian W, Maricic, Tomislav, Stenzel, Udo, Kircher, Martin, Patterson, Nick, Li, Heng, Zhai, Weiwei, Fritz, Markus Hsi-Yang, Hansen, Nancy F, Durand, Eric Y, Malaspinas, Anna-Sapfo, Jensen, Jeffrey D, Marques-Bonet, Tomas, Alkan, Can, Prüfer, Kay, Meyer, Matthias, Burbano, Hernán A, Good, Jeffrey M, Schultz, Rigo, Aximu-Petri, Ayinuer, Butthof, Anne, Höber, Barbara, Höffner, Barbara, Siegemund, Madlen, Weihmann, Antje, Nusbaum, Chad, Lander, Eric S, Russ, Carsten, Novod, Nathaniel, Affourtit, Jason, Egholm, Michael, Verna, Christine, Rudan, Pavao, Brajkovic, Dejana, Kucan, Zeljko, Gusic, Ivan, Doronichev, Vladimir B, Golovanova, Liubov V, Lalueza-Fox, Carles, de la Rasilla, Marco, Fortea, Javier, Rosas, Antonio, Schmitz, Ralf W, Johnson, Philip L F, Eichler, Evan E, Falush, Daniel, Birney,

Ewan, Mullikin, James C, Slatkin, Montgomery, Nielsen, Rasmus, Kelso, Janet, Lachmann, Michael, Reich, David, & Pääbo, Svante. 2010. A draft sequence of the Neandertal genome. *Science*, **328**(5979), 710–22.

Hill, William G, & White, Ian MS. 2013. Identification of pedigree relationship from genome sharing. *G3: Genes— Genomes— Genetics*, g3–113.

Huff, Chad D, Witherspoon, David J, Simonson, Tatum S, Xing, Jinchuan, Watkins, W Scott, Zhang, Yuhua, Tuohy, Therese M, Neklason, Deborah W, Burt, Randall W, Guthery, Stephen L, *et al.* 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome research*, **21**(5), 768–774.

Jobling, Mark, Hurles, Matthew, & Tyler-Smith, Chris. 2013. *Human evolutionary genetics: origins, peoples & disease*. Garland Science.

Lazaridis, Iosif, Patterson, Nick, Mittnik, Alissa, Renaud, Gabriel, Mallick, Swapan, Sudmant, Peter H, Schraiber, Joshua G, Castellano, Sergi, Kirsanow, Karola, Economou, Christos, *et al.* 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, **513**, 409–13.

Lipson, Mark, Loh, Po-Ru, Patterson, Nick, Moorjani, Priya, Ko, Ying-Chin, Stoneking, Mark, Berger, Bonnie, & Reich, David. 2014. Reconstructing Austronesian population history in Island Southeast Asia. *Nature Communications*, **5**(4689).

Loh, Po-Ru, Lipson, Mark, Patterson, Nick, Moorjani, Priya, Pickrell, Joseph K, Reich, David, & Berger, Bonnie. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, **193**(4), 1233–1254.

Moreno-Estrada, Andrés, Gravel, Simon, Zakharia, Fouad, McCauley, Jacob L, Byrnes, Jake K, Gignoux, Christopher R, Ortiz-Tello, Patricia A, Martínez, Ricardo J, Hedges, Dale J, Morris, Richard W, *et al.* 2013. Reconstructing the population genetic history of the Caribbean. *PLoS genetics*, **9**(11), e1003925.

Pool, John E, & Nielsen, Rasmus. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, **181**(2), 711–9.