
# White Paper 23-05

## Neanderthal Ancestry Estimator

*Authors:*

Eric Y. Durand `edurand@23andme.com`

*Created:* 5 December 2011
*Last Edited:* 8 January 2012
*Summary:*

Neanderthal ancestry estimator is a 23andMe feature that enables customers to find out how much of their genome is of Neanderthal ancestry. This document is a technical description of the feature.

# Building a genome-wide estimator of Neanderthal ancestry

A fascinating question in human evolution is the relationship between us and our closest evolutionary relatives, the now (virtually) extinct Neanderthals. One of the most debated aspects of this relationship was whether Neanderthals and modern humans interbred during the tens of thousands of years we cohabited in Europe and Western Asia.

As a member of the Neanderthal Genome Analysis Consortium, I participated in the analysis of the first draft of the Neanderthal genome that was published in 2010 (Green *et al.* , 2010). More specifically, I was involved in the analysis that led to the discovery that Neanderthals did indeed interbreed with modern humans. We found that 1-4% of the genomes of all modern humans outside of Africa is of Neanderthal ancestry.

I was also involved in the analysis of another archaic human genome, the Denisova genome. We named the Denisovans after the Denisova cave, Siberia, where the bones we used to extract DNA were found. Using the same methodology, we showed that the Denisovans may have contributed 4 to 6% of the genome of present day Melanesians (Reich *et al.* , 2010). This added to the – yet incomplete – picture of a complex relationship between us, modern humans, and the other human species.

When I joined 23andMe in October 2011, we decided to implement the methodology I helped develop to enable our customers to look at their Neanderthal ancestry. In brief, our method involves the direct comparison of two modern human genomes with the Neanderthal genome. If Neanderthals did not interbreed with any modern human populations, then theory tells you that its genome should match any modern human genome at equal frequency. The observation that the Neanderthal genome matched non-African genomes at a significantly higher rate than African genomes led us to develop our Neanderthal ancestry estimator. The reader is refered to (Durand *et al.* , 2011) for more details.

## Effect of ascertainment bias

The methodology I just described was developed on whole genome data, and its implementation on the genotype data we have here at 23andMe turned out to be tricky. A pervasive issue when using genotype data for demographic inference is ascertainment bias, also known as sampling bias. Ascertainment bias arises because of the way the SNPs on our platform were chosen – or ascertained. Typically,

an ascertainment scheme consists of two phases. First, the SNPs are discovered from the genetic material of a small group of individuals, often called the discovery panel – in our case, a panel of individuals of European ancestry. Then, the discovered SNPs are typed in a larger panel – in our case, the 23andMe customer database. The major effect of ascertainment bias is an over-representation of SNPs with variants that are common in the population from which the discovery panel was sampled – in our case, Europeans. Ascertainment bias shifted the estimator towards zero (technical reasons are briefly explored in Durand *et al.* (2011)).

## PCA estimator

Although some methodologies exist to alleviate the effect of ascertainment bias in some ideal situations, we felt it was safer to turn to another way to estimate Neanderthal ancestry with genotype data: Principal Component Analysis (PCA). PCA is a very powerful tool to represent high dimensional, possibly correlated data (such as one million SNPs!) onto a much smaller, uncorrelated set of variables called principal components. The first thing we needed to do was to compute principal components that were representative of the variation in customer Neanderthal ancestry. To do so, we performed PCA on three individuals: Neanderthal, Denisova, and the Chimpanzee reference individual (named Clint). The used the chip set of SNPs for these three individuals.

This analysis resulted in two principal components; the first one, PC1, describes general genetic similarity to archaic humans (represented by the Neanderthal and the Denisova genomes) (see Figure 1). The second component, PC2, shows to contrast between Neanderthal and Denisova ancestries. We then projected the customer genotypes on the plane defined by PC1 and PC2.

Customers who have no Neanderthal or Denisova ancestry in their genomes are expected to be centrally distributed between the Neanderthal and the Denisova genomes in the PCA plot (Reich *et al.* , 2010) whereas people with some Neanderthal ancestry are expected to be projected closer to the Neanderthal. Figure 2 illustrates the projection onto PC1 and PC2 for customers of European, East Asian, South Asian and African American ancestry. We can see that the Europeans and East Asians are shifted towards Neanderthals compared to African Americans. It is also striking that the distance from an African American to the Neanderthal strongly correlates with his/ her inferred percentage of European ancestry, in accordance with previous observation that there is no Neanderthal ancestry in the genome of Africans (Figure 3).

To convert customer coordinates into a genome-wide estimate of Neanderthal

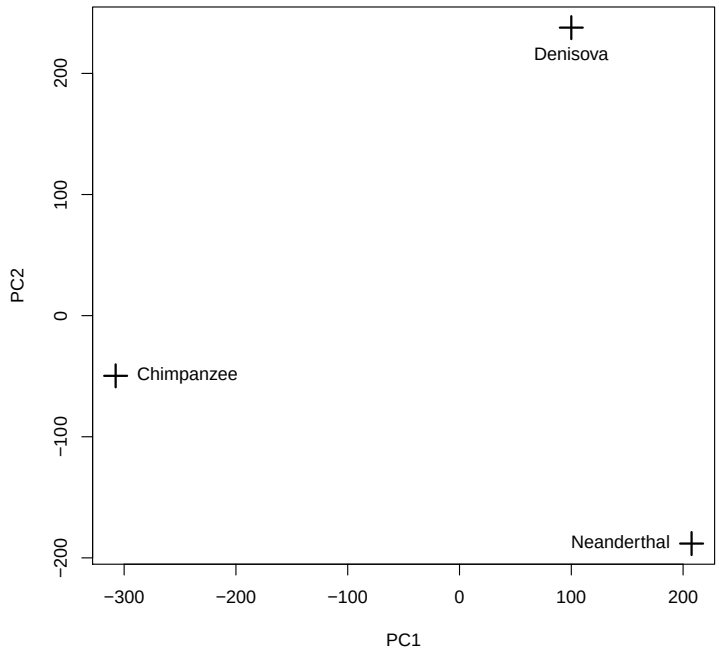


Figure 1: PCA on the Neanderthal, Denisovan and Chimpanzee genomes. We used the subset of the genomes that was defined on the 23andMe genotyping platform. PC1 differentiates archaic humans from the Chimpanzee while PC2 separates Neanderthal from Denisova.

ancestry, we projected each customer onto the Neanderthal axis, defined by the line joining the Neanderthal point with the origin of (PC1, PC2). However, the effect of ascertainment bias on the PCA is to shift all customers away from the origin. Using (0,0) as the origin to measure distance to Neanderthal will therefore bias our estimates. To correct for that, we used 246 whole African genomes from the 1000 genomes project. Whole genome data is not affected by ascertainment bias. We checked using the genome wide estimator that these African individuals showed no evidence of gene flow from Neanderthals or Denisovans. We then restricted the 246 genomes to the SNP positions defined on our genotyping platform (V3) and projected them on PC1 and PC2. We then used the centroid of the resulting 246 points as our corrected origin to measure distance from Neanderthals.

There is one shortcoming with this method; we applied the same correction

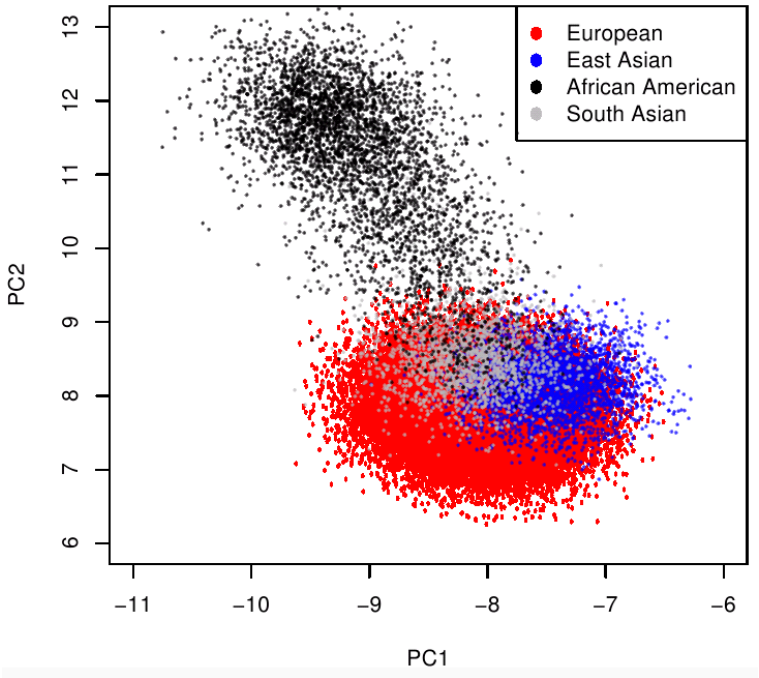


Figure 2: Projection of customer genotypes onto PC1 and PC2. Customers of European, East-Asian and South-Asian ancestry are shifted towards Neanderthal compared to African-Americans.

for each population, and ascertainment bias may affect different populations differently. For this reason, we do not completely trust between-populations comparison (ie. East Asian vs. European).

## Tag SNPs lookup

There is another, more direct way of looking for evidence of Neanderthal ancestry in our genomes. In the original publication (Green *et al.* , 2010), we identified regions in the genome of modern humans that were likely to be of Neanderthal origin. One can identify these regions in modern humans using tag markers. These markers are those SNPs where the Neanderthal variant is common in non-Africans but absent in Africans. In our paper, we identified a set of 180 such SNPs, tagging a total of 13 regions likely to be of Neanderthal origin. We could simply count the number of Neanderthal variants at these SNPs in our customers, and

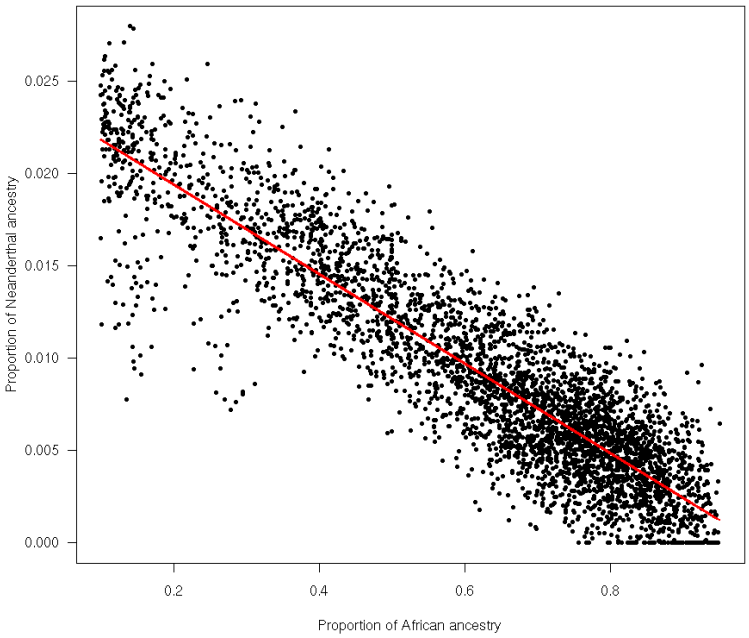


Figure 3: Neanderthal ancestry versus African ancestry in African Americans. The red line is the fitted corresponding linear model ($R^2 = 0.85$).

report this as a Neanderthal score.

However, we believe there are a number of shortcomings with this approach. First, there is no formal guarantee that these variants are indeed of Neanderthal origin. Then, even in the ideal case where all of the 180 variants are indeed of Neanderthal origin, they identify only 13 regions, the longest of which spans 160,000 bases. This length is two orders of magnitude lower than the 2.5% of Neanderthal ancestry in the average genome. Therefore, the number of tag SNPs where one carries the Neanderthal variant provides very little information regarding the total amount of Neanderthal ancestry one may have.

# References

Durand, Eric Y, Patterson, Nick, Reich, David, & Slatkin, Montgomery. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, **28**(8), 2239–52.

Green, R. E, Krause, J, Briggs, A. W, Maricic, T, Stenzel, U, Kircher, M, Patterson,

N, Li, H, Zhai, W, Fritz, M. H. Y, Hansen, N. F, Durand, Eric Y, Malaspinas, A. S, Jensen, J. D, Marques-Bonet, T, Alkan, C, Prufer, K, Meyer, M, Burbano, H. A, Good, J. M, Schultz, R, Aximu-Petri, A, Butthof, A, Hober, B, Hoffner, B, Siegemund, M, Weihmann, A, Nusbaum, C, Lander, E. S, Russ, C, Novod, N, Affourtit, J, Egholm, M, Verna, C, Rudan, P, Brajkovic, Dejana, Kucan, Z, Gusic, I, Doronichev, V. B, Golovanova, L. V, Lalueza-Fox, C, Rasilla, M De La, Fortea, J, Rosas, A, Schmitz, R. W, Johnson, P. L. F, Eichler, E. E, Falush, D, Birney, E, Mullikin, J. C, Slatkin, M, Nielsen, R, Kelso, J, Lachmann, M, Reich, D, & Paabo, S. 2010. A Draft Sequence of the Neandertal Genome. *Science*, **328**(5979), 710–722.

Reich, David, Green, Richard E, Kircher, Martin, Krause, Johannes, Patterson, Nick, Durand, Eric Y, Viola, Bence, Briggs, Adrian W, Stenzel, Udo, Johnson, Philip L F, Maricic, Tomislav, Good, Jeffrey M, Marques-Bonet, Tomas, Alkan, Can, Fu, Qiaomei, Mallick, Swapan, Li, Heng, Meyer, Matthias, Eichler, Evan E, Stoneking, Mark, Richards, Michael, Talamo, Sahra, Shunkov, Michael V, Derevianko, Anatoli P, Hublin, Jean-Jacques, Kelso, Janet, Slatkin, Montgomery, & Pääbo, Svante. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**(7327), 1053–1060.