# Supplementary material for Part XY (Siepel lab analysis)

This page summarizes the analysis of Carl Zimmer’s genome by Adam Siepel’s laborotary, which is centered at Cold Spring Harbor Laboratory. The GPhoCS analysis was done by former lab member Ilan Gronau, who is now a faculty member in computer science at the Interdisciplinary Center in Herzliya, Israel. The ARGweaver analysis was done by Melissa Hubisz, who is currently a graduate student at Cornell University. **Continue reading 🡪** …

## 1) Data preparation

We received VCF files that had been processed at the Broad Institute, representing the variant calls for Carl Zimmer’s genome relative to the human reference genome (hg19). We used the VariantsToTable program from the Genome Analysis Toolkit (McKenna 2010) to annotate a list of all variants and non-variants called with genotype quality at least 30. All other regions were treated as unknown in the remainder of the analysis. We combined this data set with other public data sets which the lab had previously processed in a similar manner, including the genomes of the Altai Neanderthal, Denisovan, several modern humans, and the chimpanzee.

## 2) GPhoCS analysis

**Slides providing more details on the GPHoCS analysis:**[**GPhoCS\_presentation.pdf**](http://archive.gersteinlab.org/proj/zimmerome/part05/ppt/Slides.Variant_Calling-Examples_FN.pdf)

We ran GhoCS using the same pipeline used in Kuhlwilm 2016, substituting Carl Zimmer’s genome in the place of European individuals. The PDF document referenced above summarizes the methods and results. Overall, this analysis illustrates that Carl Zimmer’s genome is representative of modern European genomes, and contains sufficient information to make inferences about European population history, as well as to detect ancient interbreeding between the Neanderthal and ancient modern humans.

## 3) ARGweaver analysis

We ran ARGweaver (Rasmussen 2014) on a data set including Carl Zimmer’s genome, the Altai Neanderthal (Prufer 2014), the Denisovan (Meyer 2012), and 12 public modern human genomes. The modern humans are a subset of the “69 Genomes Data” published by Complete Genomics (Drmanac *et al*, 2010), and include 4 individuals each of European, Asian, and African descent (full list downloadable below). The genome was divided into 716 overlapping chunks of roughly 5Mb each. “Phase integration” was used in order to account for uncertainty in which allele came from each parent at heterozygous sites. This procedure, as well as other ARGweaver details such as recombination and mutation rates used, were the same as those used in Kuhlwilm 2016. ARGweaver was run for 3000 iterations, with the state recorded every 20 iterations starting from iteration 1020. This resulted in 100 ancestral recombination graphs (ARGs) representing the predicted ancestral relationship between all of the individuals in the sample, at every location along the autosomal genome (chromosomes 1 through 22).

These ARGs were then used to look at several statistics of interest, including:

*Pop assignment*: For a given individual and genomic location, a population assignment of either “European”, “Asian”, “African”, or “unknown” was made. This was done by tracing the two lineages coming from an individual (one for each parent) and determining which other individual either of these lineages shares the most recent ancestry with. No assignment was made in the case of a tie with multiple populations. It is important to note that this population assignment does not necessarily indicate the population of ancestry for the genomic location. For example, all non-African individuals appear to have a significant fraction of African ancestry; however this can be explained by lineages whose ancestry traces back to the ancestral population pre-dating the out-of-Africa event. Overall, Carl’s ancestry looks quite similar to that of the other European samples.

*Neanderthal ancestry:* To call regions of putative Neanderthal ancestry, we looked at the most recent time that one of Carl Zimmer’s parental lineages found ancestry with a Neanderthal lineage. If at least 98 out of 100 sampled ARGs show a time more recent than 300,000 years ago, then this region is flagged as possible Neanderthal introgression. The cutoff of 300,000 years was used as a very conservative estimate which helps correct for uncertainty in local mutation rates and keep false positives to a minimum. The actual split time is not known precisely but is roughly considered to be around 600,000 years ago.

*Denisovan ancestry:* This was done similarly to Neanderthal ancestry above, but with the Denisovan genome.

Neanderthal and Denisovan ancestral regions were called across several individuals, as shown in the presentation below. The amount of Neanderthal ancestry detected in African individuals was around 0.6% and is probably indicative of the error rate of this method. Carl Zimmer’s genome showed 2.0% Neanderthal ancestry; considerably less than the ~3% observed in European and Asian controls. However, without further analysis, it is difficult to say whether this is because he has less Neanderthal ancestry than average. It could also be explained by differences in the genetic sequencing, genotyping, filtering used in Carl Zimmer’s sample compared to others.

The longest regions appearing to be Neanderthal and Denisovan in Carl Zimmer’s genome can be browsed at the following URLs:

<http://compgen.cshl.edu/~mhubisz/carlZimmer/carl.altai.introgressed.html>

<http://compgen.cshl.edu/~mhubisz/carlZimmer/carl.denisova.introgressed.html>

***associated data files: .***

[**CompleteGenomicsSamples (PDF format)**](http://archive.gersteinlab.org/proj/zimmerome/part05/data/variants/Z.variantCall.SNPs.vcf)

[**Argweaver\_presentation (PDF format)**](http://archive.gersteinlab.org/proj/zimmerome/part05/data/variants/Z.variantCall.SNPs.vcf)

***References :***

*Drmanac, R. et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 327/5961. 78-81.*

*Gronau, I. et al. (2011) Bayesian inference of ancient human demography from individual genome sequences. Nature Genetics. 43/10. 1031-4.*

*Kuhlwilm M. et al. (2016) Ancient gene flow from early modern humans into Eastern Neanderthals. Nature 530/7591. 429-33.*

*McKenna A et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20/9. 1297-303.*

*Meyer M. et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science. 338/6104. 222-6.*

*Prufer K. et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 505/7481. 43-9.*

*Rasmussen M. et al. (2014) Genome-wide inference of ancestral recombination graphs. PLoS Genetics. 10(5):e1004342.*