**Supplementary material for part 10 (Shapiro lab analysis)**

This page summarizes the analysis performed on the Zimmerome at the UCSC Paleogenomics Lab (Shapiro Lab).

**Estimating the human population history from the Zimmerome using PSMC**

**1) Background**

**a) What? Why?**

We used the PSMC software package (PSMC is an acronym for Pairwise Sequentially Markovian Coalescent) to learn the history of population-size changes in the human populations that included the Zimmerome's ancestors. This approach shows how one can learn the history of an entire population by analyzing the genome sequence of a single individual from that population.

**b) What’s the logic behind it?**

Coalescent theory correlates the genetic diversity with the demographic history of a given population. For simplicity, let’s consider a single gene. Two individuals in a population may differ in the sequence of a single gene by a few mutations. However, that gene sequence was the same at some point in history and, over time, different changes will occur on each version of that gene. It is possible to use information about the molecular clock (how quickly mutations accumulate) to estimate when the two different genes were the same. This is said to be the “coalescence time,” of those two genes.

PSMC extends this in two important ways. First, every person’s genome is actually two genomes: one from the chromosomes that that person gets from their mother, and the other from the chromosomes that they get from their father. In this way, his genome contains two copies of nearly every part of the genome. Second, rather than look at just one gene, PSMC uses all of the information in the genome at the same time. Specifically, Li & Durbin introduced a method in 2011 that uses the density of heterozygous sites (where the chromosome from mom differs from the chromosome from dad) to infer periods of time in the history of the population when that population was small or large (drawing from coalescent theory).

**Reference**:

Heng Li & Richard Durbin (2011) Inference of human population history from individual whole-genome sequences. Nature 475, 493–496. doi:10.1038/nature10231

**2) PSMC methods**

**a) Create a consensus sequence**

We used the provided BAM files with the alignment of the Zimmerome sequence short reads mapped to the reference human genome (hg19). We then created a diploid consensus sequence from this alignment using the pileup command from the Samtools package.

**b) Create the PSMC input file**

We transformed this consensus sequence into a different file format. This step ‘bins’ the genome into 100 bp non-overlapping windows indicating if inside that window there was at least one heterozygote.

**c) Run PSMC**

PSMC will now infer the population size history. It will look into the input file for sequences of 0’s and 1’s (0 = homozygous, 1 = heterozygous), and calculate the transition probability of states using a Hidden Markov Model, also using other information like the mutation rate, and recombination rate. After that it will output the variation of population size over time. All parameters are scaled by a constant that is determined by a neutral mutation rate (µ = 2.5 x 10-8) and generation time (g = 25 years), giving you the information that can be then plotted as effective population size x time in years.

**References for…**

**Hg19**: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/>

**Samtools**: Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. doi: 10.1093/bioinformatics/btp352

**PSMC**: Heng Li & Richard Durbin (2011) Inference of human population history from individual whole-genome sequences. Nature 475, 493–496. doi:10.1038/nature10231

**Neutral mutation rate**: Michael W. Nachman & Susan L. Crowell (2000) Estimate of the Mutation Rate per Nucleotide in Humans. Genetics, 156 (1) 297-304. PMCID: PMC1461236

**3) Command lines**

**a) Creating a diploid consensus sequence**

samtools mpileup -C50 -uf referencegenome carl.bam | bcftools call -c -| vcfutils.pl vcf2fq -D 100 | gzip > carl.psmc.fq.gz

**b) Converting the diploid consensus sequence to PSMC input format**

fq2psmcfa -q20 carl.psmc.fq.gz > carl.psmcfa

**c) Running PSMC**

psmc -N25 -t15 -r5 -p "4+25\*2+4+6" -o carl.psmc carl.psmcfa

**d) Creating the PSMC plot**

psmc\_plot.pl carl carl.psmc

**References for the programs used:**

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. doi: 10.1093/bioinformatics/btp352

Heng Li & Richard Durbin (2011) Inference of human population history from individual whole-genome sequences. Nature 475, 493–496. doi:10.1038/nature10231

**4) Conclusions**

Unsurprisingly, the results obtained from the Zimmerome reflect the same results obtained by Li & Durbin (2011). [Reference figure file: nature10231-f3.2.jpg] It demonstrates that the European population suffered a severe bottleneck between 20 and 50 thousands of years ago, that could be explained by the Last Glacial Maximum. It also shows that human populations had a high effective population size between 60 and 200 thousands of years ago, which could be interpreted as the time as the origin of the modern humans\* or an effect of population substructures\*.

**References:**

Behar, D. M. et al. (2008) The dawn of human matrilineal diversity. Am. J. Hum. Genet. 82, 1130–1140 doi: 10.1016/j.ajhg.2008.04.002

**Associated data files:**

Diploid consensus sequences carl.psmc.fq.gz

PSMC input: carl.psmcfa

PSMC output: carl.psmc

PSMC plot: Zimmerome.pdf