# Downloaded data

The DNA from three individuals have has been obtained from immortalized lymphoblastoid cell lines.

Table 1: Sample information for the three homo sapiens individuals

|  |  |  |  |
| --- | --- | --- | --- |
|  | **HG01109** | **HG03098** | **HG03492** |
| **Sex** | Male | Male | Male |
| **Populations** | [Puerto Rican in Puerto Rico](https://www.internationalgenome.org/data-portal/population/PUR), American Ancestry (PUR) | [Mende in Sierra Leone](https://www.internationalgenome.org/data-portal/population/MSL), African Ancestry | [Punjabi in Lahore, Pakistan](https://www.internationalgenome.org/data-portal/population/PJL), South Asian Ancestry (PJL) |
| **More info** | [Link](https://www.internationalgenome.org/data-portal/sample/HG01109) | [Link](https://www.internationalgenome.org/data-portal/sample/HG03098) | [Link](https://www.internationalgenome.org/data-portal/sample/HG03492) |

For this experiment only the Illumina forwards reads are needed. The read length for all the samples is 150 bases. The data was download from <https://github.com/human-pangenomics/hpgp-data>.

# Simulation of ancient DNA fragments

The downloaded Illumina reads are first trimmed with AdapterRemoval to remove adapter sequences present in the data. The reads are filtered by length (min. length 150bp), so that any trimmed reads are discarded.

From the Homo sapiens neanderthalensis individual Vi33.19, 1 Mio fragments lengths have been sampled (figure 1). Using gargammel fragSim, the modern DNA reads are trimmed according to the ancient DNA fragment lengths. Fragment lengths longer than 150bp were changed to 150bp, as this is the maximum length that can be generated from the Illumina forward reads. As can be seen in figure 1, this is only the case for a small fraction of reads.

Chart, histogram

Description automatically generatedFigure 1: Fragment length distribution of an ancient DNA sample from the neanderthal individual Vi33.19

The reads are then transformed to fasta format, as they are representing DNA fragments. Because reads are used to generate the fragments, sequencing errors are included in the “true” sequences of the simulated fragments.

# Simulation of reads

Paired-end Illumina HiSeq 2500 reads with a length of 125 bases per read were simulated using gargammel adptSim and ART. Only substitutions, but no indels, are simulated as sequences errors.

# Preprocessing of the simulated reads

The simulated paired-end reads are trimmed and merged with the 7 different tools using standardized settings. Reads that cannot get merged are discarded.

# Alignment to reference genome

The reads are aligned to the GRCh38 reference genome using bwa mem. PCR duplicates are marked with Picard.

TODO: Check the coverage depth and adjust the number of fragments to archive 3X coverage.

# Variant calling

TODO: Use gatk HaplotypeCaller for identification of candidate sites or regions at which one or more samples differ from the reference sequence. Create a file that contains only the variant sites with gatk Genotype GVCFs.

I wonder if I should filter the variants to remove false positives. Hard filtering would be difficult with 3X coverage, but I could for example try to restrict variant calling to the ‘accessible genome’, defined as that portion of the reference sequence that remains after excluding regions with many ambiguously placed reads or unexpectedly high or low numbers of aligned reads

# Genotyping

Estimation of the alleles present in each individual at the variant sites or regions. Diploid.

# PCR

To generate the input matrix, I need to find out how to encore diploid variant information into a matrix. A simple 1-out-of-k encoding would not work for this.