# Downloaded data

For this experiment, raw Illumina whole genome sequencing (WGS) 2X250bp reads from the NIST's Genome in a Bottle (GIAB) project were used. For each sample, 1.72 \* 10^9 forwards (R1) reads were obtained. The DNA has been obtained from immortalized lymphoblastoid cell lines.

Table 1: Sample information for the two homo sapiens individuals

|  |  |  |
| --- | --- | --- |
|  | **HG002** | **HG005** |
| **Sex** | Male | Male |
| **Populations** | **Ashkenazim** | **Chinese** |
| **More info** |  |  |

# Simulation of fragmented DNA templates

First, we want to remove any reads with adapter sequences. The downloaded Illumina R1 reads are trimmed with AdapterRemoval to remove adapter sequences present in the data and filtered by length (min. length 250bp), so that any trimmed reads are discarded.

Now, the reads trimmed to different lengths to represent ancient DNA fragment lengths. For this the lengths of 1 Mio DNA fragments have been sampled from a Homo sapiens neanderthalensis individual Vi33.19 (figure 1). Using gargammel fragSim, the reads from the two modern DNA samples are trimmed to match those ancient DNA fragment lengths. Lengths longer than 250bp had to be changed to 250bp, as this is the read length. This was only the case for a tiny fraction (I think only 3 lengths out of 1 mio were >250).

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 templates. Because reads are used to generate the fragments, there will be some sequencing errors in the generated DNA templates.

# Simulation of reads

Using the DNA fragments from the previous step, paired-end Illumina HiSeq 2500 reads with a length of 125 bp/read were simulated using gargammel adptSim and ART. Substitutions, but no indels, are included in the reads as sequencing 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 GRCh37/hg19 (1000 genomes) reference genome using bwa mem. PCR duplicates are marked with Picard.

The alignment file (bam file) is subsampled to archive the following mean depth of coverage: 0.25X, 0.5X, 1X, 2X, and 4X. For each depth, 10 different subsamples are taken.

# PCR

Using smartPCA from Eigensoft. Using data from the human origins dataset with contains 3,902 individuals, each represented by 593,124 single nucleotide polymorphisms (SNPs). Those SNPs have exactly two different alleles, and each individual has one of four possible values at each genotype: homozygous reference, heterozygous, homozygous alternative, or missing. Those four values are encoded 2, 1, 0 and 9 respectively.

To generate the PCA input data for our 2 samples we run samtools mpileup with pileupCaller to generate a haploid genotype for each position of the above mentioned SNPs. For each SNP position, one read will be taken at random.

**Benchmarking best practices:**  
To establish best practices for using GIAB genomes for benchmarking, we have worked with the Global Alliance for Genomics and Health Benchmarking Team:  
[Benchmarking tools](https://github.com/ga4gh/benchmarking-tools/), [Manuscript](https://doi.org/10.1038/s41587-019-0054-x), [GitHub](https://github.com/ga4gh/benchmarking-tools/)  
[Stratification Bed Files for Difficult Regions (cite precisionFDA manuscript below)](https://github.com/genome-in-a-bottle/genome-stratifications)  
[precisionFDA Truth Challenge V2 Manuscript](https://doi.org/10.1016/j.xgen.2022.100129) and [data/vcfs](https://doi.org/10.18434/mds2-2336) are an example of small variant benchmarking with v4.2 and stratifications

TO DO:

1. Start snakemake with the new input data (HG002/HG005) until merging
2. Fai index the bgzipped reference genome
   1. Test alignment and mpileup/pileupCaller with the bgzipped ref genome
3. Try smartPCA: merge eigenstrat files and run PCA