# Petljak Lab Mapping and alignment pipeline


## Author: Luka Culibrk


## July 2024

We more or less follow the GATK best practices guidelines for mapping and processing of raw sequence data.

1. FASTQ is produced by one of the FASTQ modules available in the pipelines - at time of writing, this was SRA, EGA, and local BAM. One pair of R1/R2 fastqs come from a single sequencing run - ie. if one sample was multiplexed on multiple lanes, then each sample/lane is processed separately.
2. FASTQ is converted to unaligned BAM (uBAM)
3. uBAM is adapter-marked by Picardtools to make downstream methods aware of illumina adapters contaminating raw data
4. SamtoFastq is used to convert the uBAM back to FASTQ (required for alignment). Marked adapter bases are set to base qual 2 and positions are noted with the XT tag
5. Reads are mapped to the reference genome using BWA 0.7.17 - latest version
6. The BWA output SAM is missing unaligned reads - this would prevent round-trip reproducibility of the final CRAM. The SAM is processed using MergeBamAlignment. One convenient side effect of this is that the reads become coordinate sorted in the process.
7. Next we run GATK's MarkDuplicatesSpark. This marks duplicate reads, using sequence data, and coordinate data if available. This is non-destructive - reads are marked but retained if they are duplicates. This marks ~20% of reads typically, as PCR (duplicated during library prep) or optical (single spot called as 2 by the image processing algorithm during sequencing) duplicates.
8. We merge all runs from a single library/sample using GATK's MergeSamFiles into a sample-specific bam
9. To optimize for space, and in some cases, speed, we compress BAM losslessly to CRAM. Here, samtools is used, which imparts a default compression level of 6, found to be an optimal trade-off in decompression speed and space savings. The MD5 checksum of the reference genome is also generated, as well as a short readme to point to the path of the reference. This is done to make it explicit which reference genome was used to generate the CRAM.