Jaesub Kim, M. Sc. bioinformatics, University of Potsdam

Username: kim14

Summary of the 6th Computer-Lab Exercise: Read Mapping

In the 6th computer-lab excercise, we conducted practice on read mapping. bwa, samtools, and bcftools were utilized in this process.

Mapping reads in this exercise involves two essential types of data - you need your read data and some reference data. The read data we used consisted of paired-end sequencing data, comprising two fastq files representing forward and reverse reads. Since our chosen read mapper processes *.gz files, these two fastq files were handled in a compressed state as .gz files.

Before initiating mapping, indexing of the reference sequence was necessary. We used BWA, which provides indexing using the Burrows-Wheeler Transform (BWT) method, to index the given fasta file as the reference.

Then, using bwa, we mapped our forward and backward reads to the reference sequence. Through this process, we obtained mapping results in the sam file format. This data provides information sorted into a total of 11 columns, we focused on the first (QNAME), second (FLAG), and sixth (CIGAR) columns in this exercise. Since we mapped paired-end sequencing data, all query templates in the first column were aligned with pairs having the same ID. The FLAG in the second column represents information about these alignments, with pairs sharing the same ID also having specific FLAG pairs. Most pairs have a 77-141 pair, indicating that the read pair is unmapped. In cases where mapping is successful, for example, a 99-147 flag pair may be present. The sixth column represents the summary of the alignment, including M: match, I: insertion, D: deletion, among other information. Unavailable information is denoted by *.

We also practiced finding the most frequent flags. However, we could not check the results, whether an error occurred or the data was too large and the processing time was too long.

To further analyze the mapping results, we utilized samtools. Samtools allows reading, writing, and manipulating SAM files. We converted the SAM file to a BAM file for subsequent computer processing. To simplify data processing, we sorted the result according to the reference's position and indexed the BAM file again to quickly access the mapping results.

Samtools provides a well-organized table with information on the coverage of the BAM file. Additionally, it offers a viewer that allows an intuitive checking of the positions of mapped reads on the reference sequence, letting us to evaluate the read mapping results.

Next, using bcftools, we converted the BAM file to a pileup file simultaneously transforming it into an uncompressed vcf file. The vcf file provides information about genotypes and the variations present.

Finally, utilizing the previously created sorted BAM file and the indices of both the BAM file and the reference file, we generated a consensus sequence seamlessly integrated into one cohesive result.