LABORATORY REPORT

Introduction to the Galaxy platform: Quality Control and Mapping

László Szalai (gusszalla@student.gu.se)

Department of Molecular Biology, Faculty of Science, University of Gothenburg, Gothenburg 41 390, Sweden

Introduction

Many errors and impurities can occur in a sequencing mechanism due to numerous factors like amplification caused replicative mistakes, accidental polyA- and adapter region reads. These errors could contribute to false positives, mapping disturbances and other alignment errors, however modern bioinformatics have developed different analytical methods to filter and trim many impurities, enabling us to create better quality alignments providing more certain results. The goal of the laboratory was to get familiarized with the tools provided by the Galaxy platform and/or Unix as well as to use and understand the processes of quality control (QC) and mapping/alignment tools.

Trimming and Filtering

Our initial sequence file (moo.fastqsanger.gz) contained 812 individual 296 base pair long sequences with low read quality on positions above the 105 base pair mark, GC distribution artifacts likely caused by polyA reading, high sequence duplication levels and detected adapter reading content matching Nextera Transposase Sequences. These results were calculated and visualized by the original FastQC tool, emphasizing the need for QC tools, most prominently trimming but also filtering.

In Galaxy these tools are used to exclude low quality reads from the sequences, reducing bad mapping quality downstream. While trimming only excludes anomalies from the ends of sequences, filtering can exclude anywhere in the sequence. Depending on the method and settings, the tools can either exclude whole reads if they contain impurities above a threshold or exclude certain parts of reads to keep good quality parts of them. There are also possibilities often to use paired-end or single reads. By utilizing 4 different QC tools we were able to find key differences between them and form an opinion on which would be preferred in an industrial setting.

My initial experience was that ***PRINSEQ*** and ***fastp*** were more customizable than ***Trim Galore!*** and ***Trimmomatic*** but the latter two had better mean quality scores when compared in a MultiQC plot. In my opinion this is caused by not having the time to experiment on giving different settings for the 4 tools, which only indicates that the latter two have better default settings. This led to a preference for PRINSEQ in overly specific situations but using Trimmomatic in a general setting. PRINSEQ and Trimmomatic are also favorable because they utilize filtration and trimming without excluding whole reads while also considering retaining a certain minimum length of the reads providing a good balance. Overall, every tool increased significantly the read quality and cut most adapter reads.

Mapping and Alignment

Before variants can be called, we need to match the different reads to the corresponding places of the reference genome. We used BWA-MEM, BWA-MEM2 and Bowtie2 to complete the alignment with paired-end initial data and we used ***Samtools*** (flagstat, idxstats, depth) and ***Picard*** (CollectInsertSizeMetrics, Collect Alignment Summary Metrics) toolkits to compare the two mapping methods. Both mapping tools were able to do a good quality alignment, however I prefer the original BWA-MEM mapping tools because they were slightly better at aligning reads and out of MEM and MEM2 I prefer MEM as it took half the processing time as MEM2 did.

Assignment 1 workflow:

https://usegalaxy.eu/u/laddze/w/assignment-1

Assignment 2 workflow:

https://usegalaxy.eu/u/laddze/w/assignment-2