Benjamin Strauch, Moritz Kiekeben 20.06.14

Exercise3 MHC

Outline

1. Preprocessing the reads
   - Adapter removal
     - We will check if there are overrepresented sequences contained in the reads and if so, remove them, since they are likely to be adapter sequences. Quality control will be done with software such as fastqc.
   - Quality trimming
     - Apart from adapter removal, we will also cut bases of insufficient quality from the reads, so that the mapping steps can be made more reliable. We might try different software, such as Trimmomatic.
   - Trimming of unidentified bases (N's) or complete removal of the read
     - If a read is found to have insufficient quality (too much cut away) or contains too many unidentified bases which cannot be cut, the read will be discarded.

2. Assembly of the DRB region of each individual (62 & 64)
   - Two approaches were suggested in the problem statement:
     - **De novo assembly**: We will try to assemble the genomic reads purely *de novo* using common assembly software such as *Velvet*, but we might try different tools.
     - **Reference-based assembly:** We will also consider using the genomic regions of related species as a basis for reference-driven assembly using *bwa*, for instance.
     - Ideally, we will compare the performance of both approaches with regard to the quality of the assembly and our ability to infer information about alleles and such.

3. Evaluation of the assembly
   - After the assembly is finished, we will try to identify contigs or regions where different sequences might represent parts of the genome originating from different chromosomes
   - Much of the sequence might be assembled uniquely, but those contigs that are distinct, but share lots of similarities might be alleles from different chromosomes
   - Searching for paralogs, we might employ similarity searches, such as alignment-based searches using large admissible edit-distances (to allow for large dissimilarities) or approximate string searching algorithms. This way, our pattern would be the DRB gene and any matches are potential candidates for copies or paralogous genes.