Identifying wrongly mapped reads via majority vote of several scoring schemata

Fritz Sedlazeck, and Arndt Von Haeseler
CIBIV, Vienna, Austria
fritz.sedlazeck@univie.ac.at

The advent of next generation sequencing provides access to important information on genes, gene function and genetic variation in genomes. Nonetheless, a critical step during the analysis is the mapping of reads to a reference genome. All subsequent analysis and normalizations are based on the mapping. Currently, all assembly programs share a certain amount of heuristics that may affect the mapping outcome. When choosing the mapping method the tradeoff between speed and alignment sensitivity has to be taken into account. A too high stringency of the mapping method leads to an inflated number of not mapped reads and a short runtime. Whereas a too flexible alignment strategy with respect to mismatches, insertion or deletions may results in a large number of reads mapped to the false genomic region.

Here we suggest an alignment quality measure that copes with the following challenges: It detects wrongly mapped reads independent of the read length (e.g. DNA-seq, ChIP-seq) and sequencing technology (HiSeq, GaII, 454), while allowing for various sources of variability in the data due to e.g. sequencing error or biological divergence.

Our method is based on the theory of suboptimal pairwise alignments and applies two alignment scoring functions and two alignment strategies (local Smith Waterman alignment and semi-global Needleman Wunsch alignment). In summary four alignments per read are calculated. To solve the additional challenge of computing four alignments per read we implemented alignment algorithms on an high performance hardware like graphic cards.

Each optimal alignment of a read points in the worst case to four different genomic regions in the reference genome. We only map a read, if all four alignments point to the same genomic region. Otherwise the read is discarded and called not-mapped read.

A large simulation study shows that this strategy improves the number of correctly mapped reads considerably. Thus, our method outperforms other frequently used mapping programs (e.g. Bowtie, BWA, SSaha2). We will discuss the details of the simulation and the improved quality of the method.