**Genome Informatics for High-Throughput Sequencing Data Analysis**
**Methods and Applications**

Dissertation, 2014

Steve Hoffmann

This thesis introduces three different algorithmical and statistical strategies for the analysis of high-throughput sequencing data. First, we introduce a heuristic method based on enhanced suffix arrays to map short sequences to larger reference genomes. The algorithm builds on the idea of an error-tolerant traversal of the suffix array for the reference genome in conjunction with the concept of matching statistics introduced by Chang and a bitvector based alignment algorithm proposed by Myers. The algorithm supports paired-end and mate-pair alignments and the implementation offers methods for primer detection, primer and poly-A trimming. In our own benchmarks as well as independent bench- marks this tool outcompetes other currently available tools with respect to sensitivity and specificity in simulated and real data sets for a large number of sequencing protocols. Second, we introduce a novel dynamic programming algorithm for the spliced alignment problem. The advantage of this algorithm is its capability to not only detect co-linear splice events, i.e. local splice events on the same genomic strand, but also circular and other non-collinear splice events. This succinct and simple algorithm handles all these cases at the same time with a high accuracy. While it is at par with other state- of-the-art methods for collinear splice events, it outcompetes other tools for many non-collinear splice events. The application of this method to publically available sequencing data led to the identification of a novel isoform of the tumor suppressor gene p53. Since this gene is one of the best studied genes in the human genome, this finding is quite remarkable and suggests that the application of our algorithm could help to identify a plethora of novel isoforms and genes. Third, we present a data adaptive method to call single nucleotide variations (SNVs) from aligned high-throughput sequencing reads. We demonstrate that our method based on empirical log-likelihoods automatically adjusts to the quality of a sequencing experiment and thus renders a „decision" on when to call an SNV. In our simulations this method is at par with current state-of-the-art tools. Finally, we present biological results that have been obtained using the special features of the presented alignment algorithm.