Processing of Raw Genome Data

The sequencing of the human genome is one of the biggest advances in genetics over the last four decades. It enables researchers and even clinicians to understand mutations, to identify markers for diseases and to take personalized treatment decisions. First developed by Sanger et al. in 1977, DNA sequencing technologies have improved rapidly since. With the emergence of Next Generation Sequencing (NGS) machines, the time spent on sequencing DNA strings has been increased from 100 min to only 0.002 min for sequencing 1 kilobase pairs (Kb). This rapidly emerging technology has both its benefits and drawbacks. On the one hand, NGS machines sequence the genome at low cost and a high throughput. On the other hand, the high throughput results in relatively short DNA sequences, i.e. reads, which are typically 30–400 base pairs long, and a high per base sequencing error rate. Both the high error rates and the large amount of data through the short reads require the integration of new algorithms into the DNA analysis workflow.
For processing raw genome data, alignment and variant calling algorithms play an important role in the workflow. Therefore, this presentation concentrates on these two topics.

Alignment is an indispensable requirement for variant calling in the workflow. As the actual positions for all short reads in the genome are not known, they have to be aligned against a reference, e.g. the human reference genome. This is crucial because without alignment the function of certain bases and the effects of variants cannot be examined. Alignment algorithms must not only be time efficient to deal with the large amount of data but also accurate at aligning, because errors can propagate into later stages of the pipeline. Reads containing errors and variants are challenges for the accuracy of the algorithms. To deal with these challenges, a variety of algorithms, such as SOAP3, Bowtie2, or MAQ, have been proposed.
In this presentation, we will discuss Bowtie2, one of the most popular alignment algorithms. We will take a closer look at the essential parts of this algorithm, such as the Burrows-Wheeler-Transformation and the dynamic programming approach used to deal with Insertions and Deletions (Indels).

In the variant calling stage, identifying variants like Single-Nucleotide-Polymorphisms (SNP), Indels and larger structural variants is a big computational challenge. Sequencing errors and the resulting misaligned reads lead to false-positive variants. The false-positive variants compound distinguishing real variants from artificially introduced ones. Furthermore, the lower the number of available values for a position, so called coverage, the bigger is the computational complexity for identifying extensive variants.
For this presentation, we focus on SNP calling, which is the simplest form of variants. Popular tools for SNP calling are SOAPsnp, SAMtools and the HaplotypeCaller. The latter will be reviewed in detail as it is able to minimize the false-positive variants. We will discuss the use of Pair Hidden Markov Models (PairHMM) and the Bayesian Model to show the pursued strategy and the inner mechanics of the algorithm.