Genomic Variant Calling from RNA-Sequencing Data

Genomic Variant Calling is a promising technique that allows understanding the relationship between genotype and phenotype of individuals. The goal of variant calling is to find *Single-nucleotide polymorphisms (SNPs)*, i.e., variants of single nucleotides in the genomes of individuals within a population. Variants that are only found in a minority of individuals in the population may often be an indicator or cause for a specific trait or a genetic disease.

Traditionally, SNPs are found by using whole-genome sequencing (WGS), which is based on the DNA. WGS results in high cost financially, computationally and in increased analytical complexity, but many algorithms are available for the analysis. However, in RNA-sequencing (*RNA-Seq*) the RNA of a tissue or sample is analyzed in order to quantify genes expressed in specific cells. Information on variants is also contained in the data, but as RNA is extracted from specific cells, the variants only resemble genetic properties that are relevant for that specific cell. Due to its limitation to expressed genes, it is seldom used for variant calling. However, RNA-Seq is much cheaper than WGS, and its limitation might not be relevant if variants are present in the expressed genes. Because of the low cost RNA-Seq data is highly abundant while only a limited number of approaches exist to use RNA-Seq data for variant calling. Much valuable information that could be gained from variant calling is currently not used in further analyses.

In my expert session I will present two RNA-Seq variant calling approaches: *GATK*'s best practices – a pipeline of well-known variant calling and filtering steps and the earlier published *SNPiR,* which adds additional filtering steps and metadata that allows for a higher confidence of produced variants. RNA-Seq-based variant calling pipelines generally consist of the following steps: (1) Mapping of RNA reads to a reference genome, (2) deduplication and filtering of unmapped reads and reads with low quality, (3) variant calling on mapped reads, and (4) postprocessing including filtering of false-positive variants. In my presentation I will go through these steps and explain measures taken in each step to focus on generating especially high quality variants.

Besides the variant calling pipeline, I will provide you with some biological background that will help to understand RNA-Seq based variant calling and its implications. Possible approaches to gain practical information from variants are statistical procedures such as cluster analysis and combination with further medical data. This part will be covered by Paul's presentation in the same expert session.