PODKAT: Large-Scale Association Testing Utilizing Rare and Private Variants

Ulrich Bodenhofer
Institute of Bioinformatics
Johannes Kepler University, Linz, Austria
bodenhofer@bioinf.jku.at

High-throughput sequencing technologies have facilitated the identification of large numbers of single-nucleotide variants (SNVs), many of which have already been proven to be associated with diseases or other complex traits. Since association tests considering individual SNVs independently are known to be underpowered, different collapsing strategies have been proposed that consider multiple SNVs occurring in a region simultaneously. Such strategies can be classified into burden tests and non-burden tests.

Several large sequencing studies, such as, the 1000 Genomes Project, the UK10K project, or the NHLBI-Exome Sequencing Project, have consistently reported a large proportion of private SNVs, that is, variants that are unique to a family or even a single individual. The role that private SNVs play in diseases and other traits is currently poorly understood — which is largely due to the fact that it is statistically very challenging to consider private SNVs in association testing. While it is generally impossible to use single-marker tests for private SNVs, burden tests are potentially able to deal with private SNVs, but only if the number of SNVs occurring in a region is correlated with the trait under consideration. Non-burden tests like the popular SNP-set (Sequence) Kernel Association Test (SKAT) are typically utilizing correlations between SNVs — a strategy that is not applicable to private SNVs either, since singular events are generally uncorrelated.

We propose the *Position-Dependent Kernel Association Test (PODKAT),* which is designed for detecting associations of rare and private SNVs with the trait under consideration even if the burden scores are not correlated with the trait. PODKAT assumes that, the closer two SNVs are on the genome, the more likely they have similar effects on the trait under consideration.

This talk first explains the basic principles of association testing in general and, subsequently, explains the basic idea of PODKAT. We particularly focus on the use of PODKAT for large whole-genome studies. We will discuss issues related to data handling, computational complexity, and statistical significance and highlight the upcoming R/Bioconductor package `podkat` which implements PODKAT along with various tools for association testing. Finally, we will present results for the UK10K whole-genome cohorts.