Associating complex traits with rare variants identified by NGS: improving power by a position-dependent kernel approach

Ulrich Bodenhofer and Sepp Hochreiter

Current high-throughput sequencing technologies have allowed for an easy and cost-efficient identification of rare single-nucleotide variations (SNVs), many of which have already been proven to be associated with diseases or complex traits. Despite these successes, genome-wide association studies involving rare variants remain statistically challenging. Classical single-SNV association studies particularly suffer from poor statistical power, as the potentially large number of SNVs often leads to poor significance upon false discovery rate (FDR) correction. To overcome these difficulties, approaches have been proposed that do not consider all SNVs individually; instead, they group SNVs and perform tests on those groups. This can either be done by grouping SNVs that are in the same genomic region of interest (e.g. the same transcript or exon) or by windowing along each chromosome. The choice of the groups/windows is crucial: FDR correction does not pose a serious problem if there are large, and consequently fewer, windows, but the local tests have poor power for large windows. If smaller windows, and consequently a larger number thereof, are chosen, the local tests perform well, but FDR correction nullifies this advantage.

The windowing approach is based on the implicit assumption that accumulations of SNVs that are associated with the outcome (disease or trait) are unlikely to be caused by chance. In other words, the closer two SNVs are on the genome, the more likely they have similar effects on the outcome. Following this line of thought, we propose the Position-Dependent Kernel Association Test (PODKAT). PODKAT employs ideas similar to the acclaimed *Sequence Kernel Association Test (SKAT)* by Wu et al. with the crucial difference that the pairwise genomic distances of SNVs are explicitly taken into account in a way similar to position-dependent sequence kernels. Thereby, PODKAT is potentially able to be used for larger window sizes without drastically sacrificing statistical power. This claim is validated using a large test bed of data sets.

Currently, many studies are underway that will identify many more rare SNVs, thereby also making the aforementioned statistical difficulties more severe. We strongly believe that PODKAT will provide an even stronger advantage over competing methods for further increased numbers of SNVs.