CNV detection from exome sequencing data using a generative probabilistic model

Günter Klambauer and Sepp Hochreiter

Next generation sequencing (NGS) has emerged to one of the key technologies for analyzing genome variations. In particular exome sequencing is widely used as a cost and time efficient technology to identify disease-causing genetic variants as about 85% are located around coding regions. One important category of genetic variants are copy number variants (CNVs) typically detected by whole genome sequencing (WGS). However, most methods finding CNVs in WGS data are not applicable to exome sequencing data, since their read distributions differ substantially due to enrichment effects. The problem of read variations across targeted regions can be circumvented by locally modeling the read counts.

Our recent method "Copy Number estimation by a Mixture Of PoissonS" (cn.MOPS) for CNV detection constructs a local model across samples at each genomic position. cn.MOPS locally decomposes read variations across samples into integer copy numbers and noise by its mixture components and Poisson distributions, respectively. Model selection in a Bayesian framework is based on maximizing the posterior by an expectation maximization algorithm. Most importantly, a Dirichlet prior on the mixture components prefers constant copy number two for all samples and thereby controls the FDR in detecting CNVs. cn.MOPS excelled at CNV detection in WGS data [Klambauer, 2012]. We adapted cn.MOPS to exome sequencing data by a specialized preprocessing pipeline and incorporating information about targeted regions.

We artificially implanted CNVs in simulated exome sequencing read count data at different coverages and compared methods with respect to their performance in identifying these implanted CNVs. Besides precision and recall also the accuracy in break point detection (start and end of CNVs) was assessed. Further we compared the methods on data from the 1000 Genomes project. Exomes of 22 individuals were sequenced at high coverages. The task was to rediscover CNVs that were found previously by microarray studies and confirmed on different platforms.
Our method significantly outperforms competing methods on both data sets with respect to precision and recall.

[Klambauer, 2012] Günter Klambauer, Karin Schwarzbauer, Andreas Mayr, Djork-Arné Clevert, Andreas Mitterecker, Ulrich Bodenhofer, Sepp Hochreiter. "cn.MOPS: mixture of Poissons for discovering copy number variations in next generation sequencing data with a low false discovery rate." **Nucleic Acids Research** 2012 40(9): e69.