Poster: A Hidden Markov Model for Copy Number Variant Prediction from Whole Genome Resequecning Data

Yufeng Shen
Department of Computer Science, Columbia University, New York, New York 10027
Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032
yshen@c2b2.columbia.edu

Yiwei Gu
Department of Computer Science, Columbia University, New York, New York 10027
yg2181@columbia.edu

Itsik Pe’er
Department of Computer Science, Columbia University, New York, New York 10027
Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032
ip2169@columbia.edu

Keywords-component; Copy Number Variants, Whole Genome Sequencing, Hidden Markov Model, High-throughput sequencing

Motivation: Copy Number Variants (CNVs) are important genetic factors for studying human diseases. While high-throughput whole genome re-sequencing provides multiple lines of evidence for detecting CNVs, computational algorithms need to be tailored for different type or size of CNVs under different experimental designs.

Results: To achieve optimal power and resolution of detecting CNVs at low-depth coverage, we implemented a Hidden Markov Model that integrates both depth of coverage and mate-pair relationship. The novelty of our algorithm is that we infer the likelihood of carrying a deletion jointly from multiple mate pairs in a region without the requirement of a single mate pairs being obvious outliers. By integrating all useful information in a comprehensive model, our method is able to detect medium-size deletions (100-2000bp) at low depth-coverage (<10× per sample). We applied the method to simulated data and demonstrate the power of detecting medium-size deletions is close to theoretical values.

Availability: a program implemented in java, Zinfandel, is available at (http://cs.columbia.edu/~ys2411/zinfandel/)