Copy number haplotype inference with Hidden Markov Model and localized haplotype clustering

Wen-Ping Hsieh*
National Tsing Hua University, Hsinchu, Taiwan wphsieh@stat.nthu.edu.tw

Yen-Jen Lin
National Tsing Hua University, Hsinchu, Taiwan d928307@oz.nthu.edu.tw

Copy number alterations have been reported to be associated with disease and various cancer types. Hence, identifying the accurate position and the type of copy number alterations in the human genome is currently a critical issue. There are many tools targeting on detecting CNV regions, constructing haplotype phases on CNV regions, or estimating the numerical copy numbers. However, none of them can do all of the three tasks at the same time.

This study proposed a method based on Hidden Markov Model to detect parent specific copy number change on both chromosomes with signals from SNP arrays. A haplotype tree is constructed with dynamic branch merging to model the transition of the copy number status of the two alleles assessed at each SNP locus. The emission models are constructed for the genotypes formed with the two haplotypes. The proposed method can provide the segmentation points of the copy number alteration regions as well as the haplotype phasing for the allelic status on each chromosome. The estimated copy numbers are provided as fractional numbers, which can effectively accommodate the somatic mutation in cancer specimens that usually consist of heterogeneous cell populations. The algorithm is evaluated on the previously published regions of copy number variation on the 270 HapMap individuals and simulated data.

Key Words: Copy number variation, haplotype phasing, localized haplotype clustering