論文の内容の要旨

論文題目　Genome-wide detection of human copy number variations using high density DNA oligonucleotide arrays
（和訳　DNA マイクロアレイを用いたヒトゲノムコピー数多型の網羅的解析）

氏名　河村 大輔

In the last several years following completion of the human genome sequence, new progress in unraveling the complexities of the genome's architecture has revealed a remarkable degree of copy number variations (CNVs) present among normal individuals. Just as the effort to build a genome-wide haplotype map is already providing the framework for new studies designed to identify the underlying genetic basis of complex diseases, pathogen susceptibility, and differential drug responses, a thorough map cataloguing and indexing CNVs in the human genome is a necessary prelude to understanding their role in the context of both the normal and disease state. Although there are increasingly clear examples of how CNVs can, for example, influence susceptibility to HIV infection, modulate drug responses, or contribute to genomic micro-deletion and duplication syndromes, a comprehensive biological understanding of the roles of CNVs is not yet currently available.

To this end, a number of different molecular techniques can conceivably be used for CNV detection, but array-based experimental approaches, in contrast to more focused techniques such as quantitative PCR (QPCR) and fluorescence in situ hybridization (FISH), offer the greatest potential for global, high resolution scans of CNVs in the genome. However, they do not provide direct information about copy numbers. Signal of DNA microarray includes significant noise. Additionally, in contrast to the detection of copy number changes in tumor samples, where DNA from the same individual can be used as a reference, the use of matched samples is not possible for CNV detection in normal individuals. Similarly, the use of a single reference, as is often used in BAC-array CGH, is limited by the inability to determine whether a copy number change is from the test or the reference sample. For these reasons, computational method for the accurate detection of CNV are very much needed.

In this thesis, we study the problem of detection of copy number changes using high density DNA oligonucleotide SNP arrays, including detecting presence of CNVs and identifying the boundary and the absolute copy number of the CNV. The algorithm described in this thesis contains two major parts as shown in Figure 1.

Intensity pre-processing includes probe selection, noise reduction, scaling. CNV detection begins with pair-wise comparisons of probe intensities for all possible pairs of samples which are then merged to extract candidate CNV regions for each sample. Homozygous deletions are detected separately using an alternative approach which relies on the discrimination ratio between alternate SNP alleles in lieu of SNP genotypes. Then signal ratios and SNP information are utilized to more precisely define CNV boundaries and the copy number within each region. Finally a maximum clique algorithm is used to define the diploid samples for any given region based on the results from the large reference data. Through a comparison of the test sample to the diploid sub-set, precise boundaries and accurate copy number inferences can be drawn.

In Chapter 2 presented a method of pre-processing for microarray data in CNV analysis. We aimed to improve S/N ratio of microarray signals for subsequent CNV analysis. Probes that can be affected by cross-hybridization or sequence variation of recognition sites are removed. Skews of signal ratios due to probe affinity difference and properties of experimental conditions are reduced. We showed that our algorithms improve S/N ratio of microarray signals and lead to more accurate CNV detection.

In Chapter 3, we attempted to idenitfy CNVs using pre-processed microarray data. We aimed to detect CNVs and identify the boundary and the absolute copy number of the CNV accurately. This was achieved by summarizing pair-wise comparisons of probe intensities for all possible pairs of samples. Homozygous deletions are detected separately using an alternative approach which relies on the discrimination ratio between alternate SNP alleles in lieu of SNP genotypes. CNV boundaries and the copy number within each CNV region are estimated using signal ratios and SNP information. We showed that by using out approach, we can detect CNVs more accurately than conventional algorithms.

In Chapter 4 we applied the proposed method described in chapter 2 and 3 to the large scale real dataset and attempted to create a global map of CNVs in the human genome with high accuracy. We identfied 1203 CNVs in the dataset, spanning a large size range from less than 1 kb to greater than 3 Mb. The CNVs identified using this algorithm provides the framework for the comprehensive global map of CNVs in the human genome.

In this thesis, we presented a series of algorithms for addressing problems in detection of copy number variations in the human genome using high density SNP arrays. It is evident that in the upcoming years, much more data will become available. We hope our algorithms will be used to detect CNVs in such large amount of data and contribute to the future research of genomic variation in the human genome.