Integrated studies of copy number and genotype

Wei Sun, Fred Wright, Zhengzheng Tang, Silje H. Nordgard, Peter Van Loo, Tianwei Yu, Vessela Kristensen, and Charles Perou

Some background, what is SNP, what is CNV/CNA?

One major interest in genetic studies is to dissect the genetic factors that cause phenotypic variations, especially certain diseases. For any two individuals, although the vast majority of their genomes are the same, there could be several types of DNA polymorphisms. One of the most frequent polymorphisms is single nucleotide polymorphism (SNP) (Figure 1) that occurs when a single nucleotide (A, T, C, or G) differs between two individuals. Many array techniques have been developed to identify genotypes of millions of SNPs simultaneously, for example, Affymetrix oligonucleotide arrays, or Illumina bead arrays.

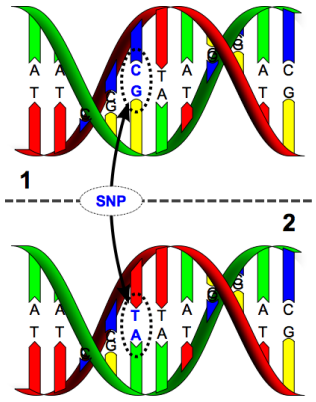

Figure 1. Illustration of one SNP in DNA sequence

In this study, we are interested in another type of genetic variation: copy number alterations, which include deletions and amplifications of DNA sequence, ranging from less than one kilobase to multiple megabase pairs. The copy number alterations in tumor and normal tissues have dramatically different characteristics. Those in normal tissues, which are often refereed to be copy number variations (CNVs), are inheritable and tend to be short and sparsely located in the genome. Those in tumor tissues, which are often referred to as copy number aberrations (CNAs), are acquired somatic alterations, tend to be longer, and occupy a significant proportion of the genome. SNP arrays can also detect copy number changes. Most current generation of SNP arrays are designed to enhance such abilities. We have developed a statistical method named genoCN, which jointly dissects copy number states and genotype calls, using SNP array data.

Why do we care about CNVs or CNAs? Because numerous studies have shown that CNVs/CNAS can cause gene expression changes, phenotype variations, and diseases. In fact, CNAs is one of the characteristics of tumor.

Illumina SNP array data

For each SNP, let X and Y be the normalized intensity measurements of allele A and B of one SNP, respectively. X and Y are first transformed to be R = X+Y and θ = arctan(Y/X)/(π/2) so that R measures the total copy number and θ measures the allelic contrast. R and θ are further normalized to remove some systematic bias to obtain LRR and BAF (Figure 2).

$$\mathrm{LRR} = \log_2(R_{\mathrm{observed}}/R_{\mathrm{expected}})$$

$$\mathrm{BAF} = \begin{cases} 0 & \text{if } \theta < \theta_{AA} \\ 0.5(\theta - \theta_{AA})/(\theta_{AB} - \theta_{AA}) & \text{if } \theta_{AA} \le \theta < \theta_{AB} \\ 0.5 + 0.5(\theta - \theta_{AB})/(\theta_{BB} - \theta_{AB}) & \text{if } \theta_{AB} \le \theta < \theta_{BB} \\ 1 & \text{if } \theta \ge \theta_{BB} \end{cases}$$

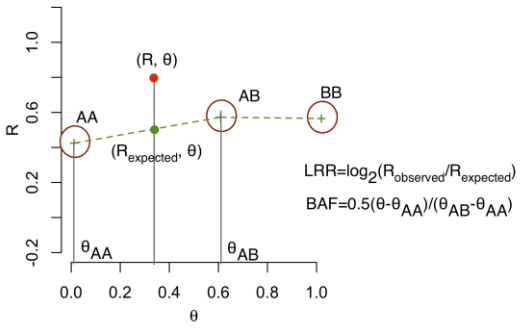

Figure 2. A cartoon of the data normalization process for Illumina SNP

GenoCNV

GenoCNV employs a hidden Markov model (HMM) to dissect copy number changes. This HMM has 6 different states (Table 1). Different from existing methods, genoCNV estimates the parameters of the HMM from the input data (instead of

State	Copy Number	(General) Genotype
1	2	AA, AB, BB
2	2	AA, BB
3	0	Null
4	1	A, B
5	3	AAA, AAB, ABB, BBB
6	4	AAAA, AAAB, AABB, ABBB, BBBB

Table 1. Six States of the HMM in genoCNV

assuming they are known or assigning a strong prior distribution) and outputs the posterior probabilities of both copy number and genotype calls.

Application: Genome-wide allele-specific copy number associations

In genome-wide association studies, we compare genotype AA, AB, and BB to detect the difference between allele A and B. In most copy number studies, however, we only compare the trait with total copy number, but ignore the allele-specific effects. For example, AAB and ABB may have different effects on a trait. The general genotype calls from genoCNV, which reflect copy number alterations, can be used for genome-wide allele-specific copy number associations (Figure 3).

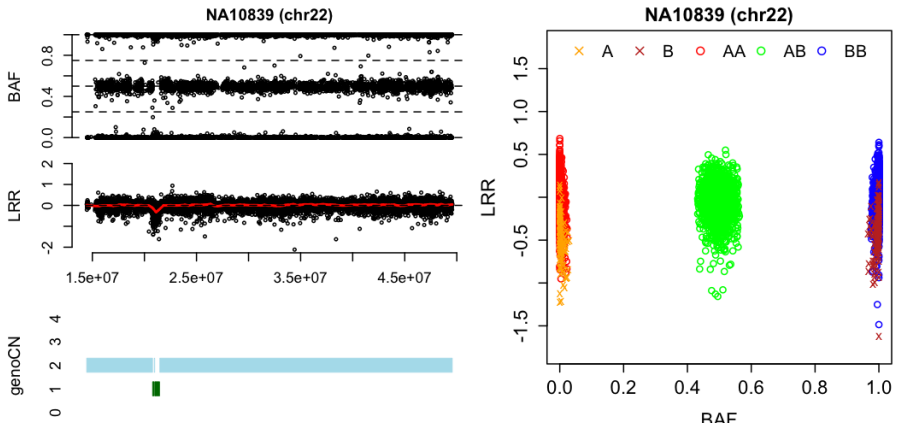

Figure 3. Copy number states and general genotype calls in chromosome 22 of a HapMap individual. The left panel shows the BAF, LRR and the most likely copy number state call. The right panel shows the LRR and BAF, grouped by genotypes (genotypes with

GenoCNA

GenoCNA extends genoCNV to analyze CNA data from tumor. There are two important improvements. First, genoCNA explicitly models the mixture of tumor tissue and normal tissue by changing the emission probabilities of the HMM. Such tissue contamination is often inevitable in tumor studies. Second, if SNP arrays are performed in both tumor and normal tissues of the same individual, we incorporate the genotype calls in normal tissue into genoCNA, again, by updating the emission probability of the HMM. Figure 4 deomonstrate the obvious advantage of genoCNA vs. typical software designed for CNV studies.

In the future, we will bring both messenger RNA and protein into the picture and develop statistical methods to help us understand the living organism from a system biology point of view.

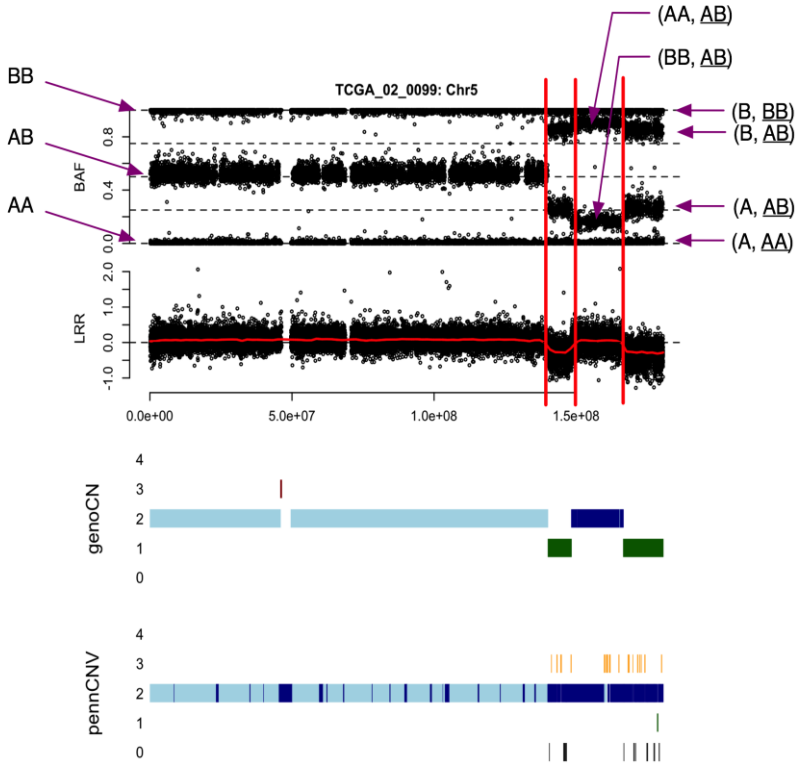

Figure 4. BAF and LRR of chromosome 5 of a tumor sample, as well as results of genoCNA and PennCNV (a popular CNV study software). The y-axis of the results of genoCNA/PennCNV corresponds to copy number. For a certain copy number, there may be different states, which are distinguished by different colors.

This is a joint work by faculties and students from UNC Chapel Hill, Department of Biostatistics, Department of Genetics, and collaborators from Emory University and Norway. The following is a picture of four co-authors from UNC-Chapel Hill. They are, from left to right, Fred Wright, Chuck Perou, Wei Sun, and Zhengzheng Tang.