Genome-Wide Analysis of Single Nucleotide Polymorphism (SNP) Data

Zhongming Zhao
Department of Psychiatry

Single nucleotide polymorphism (SNP) is the result of single base change in a population (sample) of DNA sequences. SNPs are the most abundant genetic variants in the mammalian genomes. At present, approximately a total of 12.2 million SNPs have been registered and annotated in the dbSNP database (build 120), including 9.1 million human SNPs, 0.5 million mouse SNPs, and 1.0 million dog SNPs, representing the largest public collection of the single nucleotide variants in the mammalian genomes. In the private domain, about 5.0 million human SNPs and 3.1 million mouse SNPs have been identified and well annotated by Celera Genomics. The nucleotide variations observed in the today's genomes result from the combinatory evolutionary processes such as genetic drift and selection on the naturally occurring mutations. The substitution patterns at polymorphic sites and the sequence context in a local environment of SNPs reflect the mutability of the sequence, therefore, are important for understanding the mechanisms of mutation and the evolution of genome sequences. In our lab, we are investigating the distribution and density of SNPs in the human and mouse genomes, examining the sequence context and neighboring-nucleotide bias on SNPs, and inferring the mutational spectrum in the mammalian genomes. These projects take advantage of the recently available genomic sequence and SNP data.

Bioinformatics analysis and computational tool development is required for genome-wide data analysis or comparative genomics. The computer programs were generally written in Perl, C, C++, or Java in our lab. The primary interests are:

1. To examine the sequence context at single nucleotide polymorphism sties. We investigate the distribution and identity of short sequences surrounding the polymorphic sites and reveal novel context-sensitive mechanisms. One of the hypotheses is that CpG dinucleotides are overrepresented in the short sequences that overlap polymorphic sites and are suppressed in other part of the genome sequences.
2. To estimate the mutational spectrum in the mammalian genomes. The major hypothesis is that the mutation direction is not symmetric. Two unique sets of genome data will be used to infer the direction of mutations (e.g., A → G vs. G → A): human SNP data using outgroup chimpanzee sequences and mouse SNP data using outgroup rat sequences. The genome-wide analysis should provide a reliable estimation of the mutation spectrum in the mammalian genomes. Since more SNP data will be available in the near future, this study will promote our understanding on the mutational analysis in mammalian genomes as well as in other organisms.