

Evolutionary Analysis of Single Nucleotide Polymorphisms in Mammalian Genomes

Zhongming Zhao

Department of Psychiatry and CSBC

Single nucleotide polymorphisms (SNPs) are the most abundant genetic variants in mammalian genomes. A detailed examination of the non-random sequence context where the genetic polymorphism occurs is important for understanding the mechanism of mutation (e.g., hotspot), protein-DNA interaction and genome sequence evolution. The early analyses of sequence variations in mammalian genomes and their influence of neighboring nucleotides has been limited to pseudogenes and functional regions, this is largely due to the limited data that was available at that time. The recent availability of millions of SNPs in the public SNP databases presents us with a wealth of opportunities to examine the sequence compositions of SNPs genome-wide or chromosome-wide. In our recent analyses of 2.6 million human SNPs and 0.4 million mouse SNPs, we revealed a large bias relative to the genome average at the two adjacent sites, as well as a small bias that could extend further from the polymorphic site. Our results demonstrate a non-random sequence fashion of SNPs that are surviving in today's genomes. However, these studies are based on the observation of whole genomic regions. A detailed and comparative analysis of the sequence context patterns in specific genomic categories, such as regulatory and exonic regions, would reveal more important biological information on this non-random sequence fashion of SNPs. Moreover, it is possible to infer the mutational spectrum by comparing the SNPs with their ancestor alleles. The ongoing projects in my lab are:

1. To examine the sequence context and neighboring-nucleotide biases of SNPs.
We are investigating the distribution and identity of short sequences surrounding the polymorphic sites. By comparing the distribution patterns in several genomic categories, we aim to reveal novel context-sensitive mechanisms of mutation.
2. To develop statistical method for evaluation of the effective SNP size.
Similar bias patterns at the neighboring sites were observed in the human and mouse genomes. One may ask how many SNPs are sufficient to represent the bias patterns in a genome or a chromosome. We are developing statistical method and computational algorithm to efficiently obtain the effective SNP size.
3. To infer the mutational spectrum in the mammalian genome.
We are comparing human and mouse SNPs with their corresponding outgroup sequences. The genome-wide analysis should provide a reliable estimation of the mutation spectrum in the mammalian genome.
4. To develop bioinformatics tools for the SNP analysis and psychiatric genetics.

Large-scale data collection, management and computer program development are required for genome-wide SNP analysis. The computer programs are generally written in Perl, C or Java in our lab. More information can be found at <http://bioinfo.vipbg.vcu.edu/>.