

SNP-VISTA: An Interactive SNPs Visualization Tool

Single Nucleotide Polymorphisms (SNPs) are established genetic markers that aid in the identification of loci affecting quantitative traits and/or disease in wide variety of eukaryote species. A strategy that is employed today is the re-sequencing of a large set of appropriate candidate genes in individuals with a given disease to screen for causative mutations. Such a strategy is beginning to prove fruitful in the area of cancer genetics and is likely to contribute to our understanding of gene mutations responsible for sporadic forms of congenital disorders [1]. In addition, SNPs have been used extensively in efforts to study the evolution of microbial populations. Such efforts have largely been confined to multi-locus sequence typing of clinical isolates of bacterial species such as *Neisseria meningitidis* and *Staphylococcus aureus*. However, the recent application of random shotgun sequencing to environmental samples makes possible more extensive SNP analysis of co-occurring and co-evolving microbial populations. Tools for visualization and interactive exploration of ecogenomics data are still in their infancy. An intriguing finding reported in the Tyson et al. study [2] was the mosaic nature of the genomes of an archaeal population, inferred to be the result of extensive homologous recombination of three ancestral strains. This observation was based on a manual analysis of a small subset of the data (ca. 40000 base pairs) and remains to be verified across the whole genome.

We have developed and present a new interactive data visualization and exploration tool, called SNP-VISTA, to aid in analyses for the following types of data:

1. Large-scale resequence data of disease-related genes for discovery of associated and/or causative alleles
2. Massive amounts of ecogenomics data for studying homologous recombination in microbial populations

The main features and capabilities of SNP-VISTA are: 1) Mapping of SNPs to gene structure; 2) classifying SNPs based on their location in the gene, frequency of occurrence in samples and allele composition; 3) performing clustering based on user-defined subsets of SNPs, highlighting haplotypes as well as recombinant sequences; 4) integrating protein conservation information in the visualization; and 5) displaying automatically calculated recombination points that are user-editable.

The main advantage of SNP-VISTA is its graphical user interface and visual representation of the data. SNP-VISTA supports interactive data exploration and hence leads to a better understanding of large-scale SNPs data. The applicability of our new visualization tool to various data sets and the tool's relevance for biological data analysis will be demonstrated.

SNP-VISTA is a Java application that can be executed on any Windows, Linux or MacOS X system with Java version 1.4 or higher. SNP-VISTA uses clustering software, see Levenshtein [3], which is incorporated in the package. SNP-VISTA is available for download under GPL at <http://hazelton.lbl.gov/~teplitski/dtree/>

References

- [1] Reider M. J., Taylor S. L., Clark A. G. and Nickerson D. A., Sequence variation in the human angiotensin converting enzyme, *Nature Genetics*, **22**, 59-62, 1999.
- [2] Tyson *et al.*, Community structure and metabolism through reconstruction of microbial genomes from the environment, *Nature*, **428**, 37 – 43, 2004.
- [3] <http://odur.let.rug.nl/~kleiweg/levenshtein/index.html>