Pacific Symposium on Biocomputing 5:633-635 (2000)

HUMAN GENOME VARIATION: ANALYSIS, MANAGEMENT AND APPLICATION OF SNP DATA

FRANCISCO M. DE LA VEGA
PE Biosystems, 850 Lincoln Centre Drive,
Foster City, CA 94404, USA

MARTIN KREITMAN
Department of Ecology and Evolution, University of Chicago,
1101 East 57th Street, Chicago, IL 60637, USA

Recently there has been considerable interest in the use of single nucleotide polymorphisms (SNP) for understanding the genetics of complex human diseases. The reasons are multiple. Firstly, SNPs are much more abundant than microsatellite polymorphisms (about once every 500-1000 base pairs[1]) and hence are potentially more powerful in detecting linkage disequilibrium around disease loci. Secondly, due to the binary nature of SNPs, high throughput genotyping of large number of markers is feasible with the advent of microarray technologies. Thirdly, some SNP mutations may be causative of the disease phenotypes. Finally, the imminent conclusion of the human genome reference sequence[2] opens the door to the discovery of many, if not all, of the common polymorphisms in the next several years.

In contrast with the perceived benefits of the application of SNPs to genetic studies, the analysis of SNP data pose a number of challenges. This is the first time a session devoted to human genome variation has been held at the Pacific Symposium on Biocomputing. The outcome of this year's submissions is very encouraging, with four timely manuscripts accepted for publication in the conference proceedings. The contributions accepted to this session are excellent examples of the critical topics in this field.

While the abundance of SNP markers is generally seen as an advantage, a consequence of testing large number of markers is the increase in the numbers of false positive errors. Thus, new statistical approaches to cope with the inevitable reduction in power are needed, as well as methods to better detect and quantify errors. The submission of Gordon *et al.* to this session provides a formal approach to SNP error-detection via Mendelian inconsistencies in nuclear families.

If considered alone, SNPs may carry insufficient information for detecting linkage or association since they are biallelic. Therefore, it may be necessary to consider haplotypes constructed from several such markers. A step in that direction is presented in the paper of Xiong *et al.* in this section. Here, the authors propose a new method, the Haplotype Linkage Disequilibrium Test, to increase the power of mapping disease genes by considering haplotypes, as opposed to individual markers.

The extension of the linkage disequilibrium associated with a disease locus along a human chromosome segment is not yet known. This information is important to deduce the resolution of SNP markers required for a whole genome scan and the number of patient samples needed to detect association, and is currently a debated topic of many theoretical studies. In his submission to this conference session, A. Collins provides a novel approach to this problem. Rather than using simulations, Collins uses real data on classical polymorphisms to look at linkage disequilibrium between tightly linked markers. Collins' paper thus provides insight that will be useful when more experimental data obtained by SNP typing is available for human populations.

Beyond the analytical challenges, there are the problems of handling the required sample throughput, and the storage for the SNP data that will be necessary to carry out large candidate gene or whole genome scan studies. It has been estimated recently that as many as 500,000 SNPs would be required for whole genome scans[3], and upwards of 500 patient samples would be required for finding association in candidate gene scans[4]. New, specialized SNP databases are being designed and implemented to capture the impending flood of polymorphism data. The National Center for Biotechnology Information recently created an archival repository to collect variation information focused in batch submissions of newly discovered SNPs[5]. Several other polymorphism databases will increasingly be available from several institutions and companies. What is the information that should be stored together with the SNP data to ensure maximum exploitation of its content? How comprehensive should this information be? How should we validate this data and provide quality control of the database contents? The answer to these questions will be central to the effective use of SNP data in population and human genetics studies. The contribution of Cheung *et al.* to this session provides an excellent example of possible solutions. In this paper, a new allele frequency database accessible through the Internet is described. Emphasis is given to data quality and the needs of the population genetics community in this domain specific database.

These are interesting times for the study of human genetic variation, which are being characterized by the convergence of several disciplines: human and population genetics, statistics, and now computational biology. Taken together, the increasing availability of genomic sequence information, high-throughput analysis technology, and the interest in the pharmacogenetic applications of SNPs, signals that we are indeed witnessing a major development in the field.

Acknowledgments

We would like to acknowledge the generous help of the anonymous reviewers that supported the selection process for this session, as well as the panelists that joined us to discuss the challenges in this field.

References

1. A. Chakravarti "Population genetics – making sense out of sequence" *Nat. Genet.*. Supp **21**, 56-60 (1999).
2. J.C. Venter, M.D.Adams, G.G. Sutton, A.R. Kerlavage, H..O. Smith, and M. Hunkapiller. "Shotgun sequencing of the human genome" *Science* **280**, 1540-2 (1998).
3. L. Kruglyak. "Prospects for whole-genome linkage disequilibrium mapping of common disease genes" *Nat. Genet.* **22**, 139-44 (1999).
4. A.D. Long and C.H. Langley. "The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits" *Genome Res*. **9**, 720-731 (1999).
5. S.T. Sherry, M. Ward, and K. Sirotkin. "dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation" *Genome Res.* **9**, 677-9 (1999).