Human Genome Variation: Haplotypes, Linkage Disequilibrium, and Populations
Session Introduction

F.M. De La Vega, K. Kidd, and I. Kohane

Pacific Symposium on Biocomputing 8:463-465(2003)

HUMAN GENOME VARIATION: HAPLOTYPES, LINKAGE DISEQUILIBRIUM, AND POPULATIONS

FRANCISCO M. DE LA VEGA

Applied Biosystems, 850 Lincoln Centre Dr., Foster City, CA 94404, USA
E-mail: delavefm@appliedbiosystems.com

KENNETH K. KIDD

Department of Genetics, Yale University School of Medicine
333 Cedar Street, New Haven, CT 06520, USA
E-mail: kidd@biomed.med.yale.edu

ISAAC S. KOHANE

Children's Hospital Informatics Program & Harvard Medical School
320 Longwood Avenue Boston, MA 02115, USA
E-mail: Isaac_Kohane@Harvard.edu

The working draft of a reference sequence of the human genome is nearing completion and is providing a basis for studies in a variety of domains. Computational challenges exist in all of these domains because of the massive amounts of data, the multiple often complex relationships among the types of relevant data, and the need to make the data accessible to researchers approaching the data from different perspectives. One aspect of genomic data of broad relevance is the variation in the DNA sequence among the billions of separate copies existing in the several billion living humans. Some of that variation is "abnormal" and the basis for inherited diseases. However, most of the variation is normal and simply makes each of us unique. Yet this common, normal variation is of great biomedical relevance because it can alter disease susceptibility, physiologic reactions to drugs, and response to environmental stimulus. The variation is also relevant to anthropology and understanding human evolution. In fact, several aspects of DNA sequence variation are consequences of recent human evolution: the amount of variation, the distribution of variation among human populations, and the organization of variation along the DNA sequence. This last issue has become increasingly interesting as millions of single nucleotide polymorphisms (SNPs) have been identified and mapped. With multiple SNPs mapped to every small segment of DNA the focus has shifted from the individual SNP to considering groups of SNPs as haplotypes (haploid genotypes) with the common finding that for n SNPs in a small segment of DNA there are usually far fewer than the 2^n haplotypes expected by chance. This non-randomness, commonly referred to as linkage disequilibrium (LD), is adding an additional level of complexity to genetic databases and analytic programs.

In previous years this session of the Pacific Symposium on Biocomputing has dealt with many of the issues related to DNA sequence variation. Last year the focus was relating genotype to phenotype and the handling of the data generated by high-throughput genotyping technologies. This year the focus has shifted to haplotypes and their computational challenges. Nine accepted manuscripts comprise this year's original work presented at the conference.

The recent empirical findings about the patterns of LD across samples of the genome have generated enormous interest, since the extent and intensity of this parameter relates to the statistical power that a set of SNP markers can have in association studies. In particular, the observation of large "blocks" of strong LD and low haplotype diversity [1,2] predicts that, at least in those areas, power would be adequate and that a highly dense SNP map could be redundant if the aim is to detect the common haplotypes. The contribution of Avi-Itzhak *et al.* to this volume is to present a simple algorithm to select minimum subsets of SNPs that can "tag" all the common haplotypes of a block. Utilizing Shannon Entropy as the metric of the haplotype diversity within the block, the authors find that the algorithm can efficiently reduce the number of SNPs required for genotyping, the extent to which depends on the population studied. Is the very concept of haplotype block what is studied in the manuscript of Koivisto *et al.* Here, the authors describe a new method for finding haplotype blocks based on the use of the minimum description length principle, paying special attention to the robustness of the block boundary predictions. The authors compare the accuracy of this new methodology with other published methods.

Many of the theoretical frameworks developed for studying the patterns of genetic variation in the Human genome and their application for mapping complex disease assume that in a given experiment, the researcher can obtain accurate data about the variants present in a set of DNA samples. However, as is clear for those who have spent some time as experimentalists, genotyping error is inherent in routine laboratory work. In their contribution, Gordon and colleagues describe what the interaction between genotyping error and LD is, and how this affects statistical power and the required sample size for case-control studies. Their report suggest that high LD can mitigate the reduction in power induced by genotyping errors, and that marker selection for genetic studies needs to take this additional factor into account. Comeron *et al.,* deal with the requirements necessary to detect quantitative trait loci for complex disease from a theoretical point of view. Using results from coalescent simulations, the authors suggest that the use of a highly dense SNP map does not necessarily results in increased power, unless recombination is high in the region under study. Now that empirical data on LD is becoming available, the comparison between the theoretical predictions the authors' present and the experimental data, would test our understanding on the origin of genomic variation. On the other hand, Rannala and Reeve outline a new Bayesian method to jointly estimate the position of a disease mutation and its age utilizing LD. The work of these authors is a valuable addition to the analytical tools needed for analysis of the impending volumes of genotyping data coming from the many large scale projects that are underway.

Novel methods to visualize and analyze large amounts of genotyping data that are robust and can be applied in the clinic are urgently needed. Tsalenko *et al.* present a number of such methodologies with the aim of selecting subsets of SNPs that can predict disease state and can be used for patient classification. By using data available in the public domain from a previous study, the authors demonstrate the utility of the proposed methods. In another contribution to the conference, Lancaster *et al.* address the issue of the availability of tools for analysis of genetic studies, by describing a software framework for large scale analysis of Human population data that undoubtedly would be extremely useful for those in the community just getting started with efforts of such magnitude. Finally, Stryke *et al.* remind us that these large genetic studies are in many cases product of multidisciplinary collaborations, where easy and efficient access to the results of the data analysis is crucial. These authors present their implementation for handling the data analysis flow of one of such projects.

As the papers in this session illustrate, the biocomputing aspects of human genome variation are numerous and diverse. There is ongoing need for innovation and implementation of new approaches to help individual researchers manage their own data and to make these complex data readily accessible to the broad scientific community.

Acknowledgements

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. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. *Nat Genet* 29:229-232 (2001).
2. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Altshuler, D. *Science* 296:2225-2229 (2002).