Pacific Symposium on Biocomputing 6:3-5 (2001)

HUMAN GENOME VARIATION: LINKING GENOTYPES TO CLINICAL PHENOTYPES

FRANCISCO M. DE LA VEGA
Applied 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

ISSAC S. KOHANE
Children's Hospital Informatics Program, Harvard Medical School,
300 Longwood Avenue, Boston, MA 02115, USA

Now that major milestones have been reached by both the public and private sectors in the quest for the human genome sequence, the focus is shifting to the genetic variability of our species. Lying buried in human genetic variability is the source not only of all genetic disease, but the entire range of normal phenotypic variation, including susceptibilities to pathogens and environmental factors, and individual differences in response to drug treatment. Emerging high-throughput technologies like the DNA microarray are enabling for the first time large-scale genotyping[1] and gene expression profiling[2] of human populations. Databases comprising large number of polymorphisms[3] and gene expression profiles of normal and diseased tissues or from different clinical states[4] are now thriving.

This is the second time a session devoted to human genome variation has been held at the Pacific Symposium on Biocomputing[5]. The focus of the session has been broadened this year to include the computational challenges in elucidating connections between genotypes and phenotypes using high-throughput technologies. Submissions more than doubled as compared to the previous edition, making the selection of papers a difficult job for both reviewers and organizers. Six accepted manuscripts comprise this year's original work presented at the conference.

The accepted papers demonstrate the increasing maturity of the science of linking the genotype to the phenotype. Rather than focusing on techniques that might come up with plausible associations, the submissions this year addressed "head on" many of the stumbling blocks involved in making robust the analytic techniques within formalized frameworks of the sources error and noise that are inherent to all genomic measurement systems. Furthermore, these manuscripts further the information theoretic foundations for many of the machine learning techniques developed for the investigation of functional genomics.

Many large-scale genetic association studies have been proposed to tackle the genetic origin of common disease. Due to the low penetrance of complex traits, these approaches rely on the genotyping of large numbers of biallelic polymorphisms in hundreds or thousands of subjects seeking to find linkage disequilibrium between a dense map of markers and the disease loci. The contribution of Gordon and Ott provides valuable advise to the practitioners through an assessment of the effect of genotyping errors in the power of detecting association in case-control studies. Since the high-throughput technologies for SNP-genotyping are still in emergence[1], a precise estimate of the impact of genotyping errors is crucial for the accurate interpretation of results.

Thanks to the advent of the microarray, molecular profiling is becoming the essential approach for the analysis of genotype and phenotype at the genomic scale. However, microarrays are still a rough tool where methodological and biological noise can obscure the analysis of the data[6]. Therefore, reproducibility is critical in microarray experiments and Butte *et al.* offer an assessment of the significance of fold differences in gene expression profiling that is very much needed. Typical gene expression profiling experiments produce data on thousands of genes for a dozen or less conditions (i.e., tissues, disease, time points). Furthermore, not all of the genes are being affected in all conditions but if used in subsequent analysis they contribute importantly to the noise of the data. Thus, efficient methods for dimensionality reduction are required before proceeding with further analysis. The manuscript of Park *et al*. describes a nonparametric scoring algorithm that find informative genes and that is robust to outliers and normalization schemes. In the same vein, Wahde *et al*. paper dealt with the task of finding consistently misregulated genes in gene expression matrices and provides a statistical assessment of their significance.

When confronted with the huge amount of data produced by microarray experiments, researchers frequently apply tools that find structure in the datasets, which in turn aim to reveal some inherent property of the condition under study. Typical algorithms involved in this process include hierarchical clustering, k-mean clustering, and self-organizing maps. Kim *et al.* present a novel approach for unsupervised learning from gene expression matrices which leverages the geometric properties of the data structure: the matrix incision algorithm. This approach promises to be a useful addition to the analysis tools being used in the field.

In cancer, gene expression changes are originated not only by alterations in the regulatory gene circuits of the malignant cells, but also by alterations in the chromosome counts or structure. With the imminent availability of the locations of most human genes within the genome sequence is natural to ask if changes in gene expression correlate with aberrations affecting a given chromosome. Klus *et al.* proposes such a method using a mutual information analysis to assess the effect of aneuploidy in differential gene expression observed in profiling experiments of cancer.

Massive genotyping and gene expression profiling studies are being undertaken by distinct groups of researchers, namely the human genetics and the functional genomic communities. These communities have different expertise and immediate goals, but at the end of the day what is sought is analogous: the connection between a variation in a group of genes or in their expression and observed phenotypes. There is an imminent need to link information across the huge data sets these groups are producing independently. What are the challenges in the integration of polymorphism and gene expression databases and their clinical phenotypic annotation? As high-throughput genotyping and expression-measurement methodologies are applied to large populations, the opportunity will soon arise to use existing clinical phenotypic annotations, i.e., the extended medical record. This poses several technical challenges. Among them: To what extent can clinical databases be used? Are existing clinical data models and vocabularies sufficient for the purposes of clinical annotations of genomic databases? If not, how can they be improved. Are genomic data models adequate in their present form to add to existing individual medical record systems? These upcoming questions in the field are still unanswered but we hope to explore some of them during the panel discussion of the session and in future editions of the meeting.

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. M. C. Ellis. "Spot-On SNP Genotyping" *Genome Res*. **10**, 895-897 (2000).
2. T. R. Hugues *et al*. "Functional discovery via a compendium of expression profiles" *Cell*. **102**, 109-126 (2000).
3. 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).
4. U. Scherf, *et al*. "A gene expression database for the molecular pharmacology of cancer" *Nature Genetics* **24**, 236-244 (2000).
5. F. M. De La Vega, and M. Kreitman. "Human genome variation" In: *Pacific Symposium on Biocomputing 2000*, R.B. Altman *et al*. (Eds.). World Scientific Press, Singapore (2000).
6. M. T. Lee, F.C. Kuo, G.A. Whitmore, and J. Sklar. "Importance of replication in microarray gene expression studies" *PNAS* **97(18)**, 9834-9839 (2000).