Pacific Symposium on Biocomputing 15:302-304(2010)

PERSONAL GENOMICS

Can Alkan

Department of Genome Sciences, University of Washington.
and the Howard Hughes Medical Institute.
Seattle, WA, 98195, USA

Michael Brudno

Department of Computer Science, University of Toronto.
Toronto, ON, M5S 3G4, Canada

Evan E. Eichler

Department of Genome Sciences, University of Washington.
and the Howard Hughes Medical Institute.
Seattle, WA, 98195, USA

Maricel G. Kann

Department of Biological Sciences, University of Maryland Baltimore County.
Baltimore, MD, 21250, USA

S. Cenk Sahinalp

School of Computing Science, Simon Fraser University.
Burnaby, BC, V5A 1S6, Canada

1. Introduction

Until just recently, bioinformaticians and genomicists did not have access to genomic data from multiple human individuals or many different species due to labor intensiveness and limitations of targeted approaches such as array comparative genomic hybridization (arrayCGH), PCR, SNP microarrays and similar assays, and the cost of cloning based traditional sequencing technology. This restricted our understanding of the evolution of species as well as normal and disease-causing genetic variation between different individuals of the same species. Recent improvements in sequencing methods introduced high-throughput, low-cost and cloning-free (thus less labor-intensive) technologies such as *pyrosequencing* (Roche 454), *sequencing by synthesis* (Illumina), *sequencing by ligation* (AB SOLiD), *single-molecule sequencing* (HeliScope), and many others. The revolution in DNA sequencing opened many possibilities for researchers working in the fields of evolution, genetic variation, diseases of genomic origin, and even personalized medicine. The reduced cost for whole-genome resequencing prompted a large-scale genome variation study called the 1000 Genomes Project* that aims to sequence the genomes of approximately more than 1000 individuals from different populations to build the most extensive genetic variation database to date. The new sequencing technologies can also be employed to discover functional landscape of the human genome as part of the ENCODE Project†, such as epigenetic variation (methylation patterns and histone modification) and protein-DNA interaction. Further uses of the high-throughput sequencing technologies include transcriptome analysis, non-coding RNA discovery, gene expression profiling, rapid testing of genotype-phenotype associations, and identification of pathogens.

Our genetic identity not only determines our physical differences, but it also defines our susceptibility against diseases. Several groups are working on various methods to exploit the power of cost efficient se-

*http://www.1000genomes.org
†http://www.genome.gov/10005107

Pacific Symposium on Biocomputing 15:302-304(2010)

quencing technologies as well as more traditional genome analysis approaches (SNP microarrays, arrayCGH, etc.) to better perform genotype-phenotype associations, in particular to identify susceptibility to disease, and eventually diagnose disease at its early stages. The ultimate goal is to vastly improve the field of pharmacogenomics, which can broadly be defined as the study of the relationship between genotype and drug response and how the drugs affect our metabolism. The abundance of new sequence data gives many opportunities to advancing our understanding of how to optimize drug combinations for each individual's genetic makeup. The underlying computational tools for such studies analyze available sequence data to identify differences between a reference genome and high-throughput sequenced genomes and perform sequence oriented clustering and classification to obtain both normal and disease-related phenotype associations.

This session focuses on the development of novel computational methods in all aspects of Personal Genomics including genetic and epigenetic variation discovery, genotype-phenotype associations, indexing and cataloguing both normal and disease-related variation, exome capture and resequencing, and personalized medicine. This session has a broad target audience that includes algorithm developers working on sequence analysis, genomics researchers, pharmacogeneticists, and medical geneticists.

2. Session Summary

This session includes an invited talk, five reviewed oral presentations, two additional accepted papers and a tutorial. The studies presented in this session focus on the development of computational methods to analyze genomic data generated with various types of methods.

2.1. *Papers*

The paper by **Garten *et al.*** targets an important pharmacogenomics problem that is an essential first step of personalized medicine. This work presents methods to automatically curate a network of drug-gene relationships through text mining. The authors propose that accurate curation of drug-gene relationships together with their previously described algorithm that ranks potential pharmacogenes will help identify genes that can explain variation in drug response.

The paper by **Li, Chen, and Li** addresses the problem of detecting genome-wide haplotype polymorphism. The authors present a computational framework combining the efforts of recombination detection, zero-recombinant haplotype inference and haplotype local structure clustering to jointly use the pedigree and population information. The methods presented in this work accurately reveal the haplotype structure in human populations on a genome-wide level.

The paper by **Greene *et al.*** addresses the problem of test of interaction between genes (epistasis). The authors present a novel permutation test that allows the effects of nonlinear interactions between multiple genetic variants to be specifically tested in a manner that is not confounded by linear additive effects. This method for explicitly testing epistasis or gene-gene interaction effects will likely be complimentary to genome-wide association studies (GWAS) improving our understanding of biological and statistical epistasis and their roles in human health and disease.

Li, Iakoucheva, Mooney, and Radivojac investigate the influence of disease associated mutations on known post-translational modifications. The authors provide a statistical analysis method to estimate statistical confidence of the observed trends of post-translational modification sites and amino acid substitutions.

The paper by **Yavas *et al.*** describes a new software that utilizes SNP microarray data to predict rare copy number variants (CNVs) from raw copy number. Accurate and inexpensive detection of CNVs are particularly important in disease studies where large number of patients with genetic disease can be genotyped. The authors also describe an algorithm based on simulated annealing to refine the breakpoints of the detected CNVs.

In addition to the oral presentations, the Personal Genomics session also contains two more valuable studies published in the proceedings. The paper by **Grady *et al.*** describes a computational framework

Pacific Symposium on Biocomputing 15:302-304(2010)

composed of a variety of filters to identify a subset of interesting SNPs from a much larger set for efficient interaction analysis in genome-wide association study (GWAS) data. Finally, **Thomson *et al.*** presents the application of the "sequence feature variant type" (SFVT) method to analyze the HLA genetic association in Juvenile Idiopathic Arthritis.

We are excited by the breadth of research in the field of Personal Genomics, and are hopeful that our session will help bring together researchers in these areas. The five papers presented at our session, and the additional two papers in the proceedings were selected with the help of several reviewers, whose help we gratefully acknowledge.

3. Acknowledgments

We would like to thank all the authors who submitted their work to the Personal Genomics Session. We are also indebted to the anonymous reviewers who contributed their time and expertise to evaluate the submitted papers.