

## ABSTRACTS

*Saturday, 1/27/2007*

---

---

### **Andrew Clark**

Cornell University, Department of Molecular Biology and Genetics

#### **Association Testing with Phased Haplotypes vs. Unphased Diplotypes**

The usual state of genotype data derived from SNP typing platforms lacks haplotype phase information. Association testing with more than one SNP at a time can be done either without regard to phase, or phases may be inferred or identified directly by experiment, and the association testing can be done with the phased data. It remains an important question whether the latter is worth the effort, or instead whether the unphased diplotype analysis carries most of the information for successful association testing. We present data on 448 SNPs genotyped in roughly 4000 individuals in the CARDIA cohort, and show that for HDL-C phenotypes, linkage phase information does not improve the ability to identify associations. Simulations of the simplest case of two SNPs each with two alleles will be presented to identify conditions when phase information might be expected to provide improved inference. These simulations make it clear why phase information is so rarely crucial to inference of association.

---

### **Carlos D. Bustamante**

Cornell University, Department of Molecular Biology and Genetics

#### **Bayesian Inference for Whole Genome Association Studies**

A major challenge in whole-genome association mapping is dealing with simultaneous testing of the same null hypothesis across hundreds of thousands of SNPs. Classical approaches (e.g., Bonferroni) are likely to be overly conservative and will increase penalty as the number of tests grows. I will present on a novel approaches for multiple hypothesis testing that makes use of the inferred correlation structure among SNPs in the data to detect regions of highest association and has the desirable property of increasing in accuracy as the number of tested SNPs grows. I will also discuss a Bayesian hierarchical model (and corresponding MCMC scheme) which uses a mixture prior on the effect sizes with a strong prior on no effect to tackle the same problem.

---

**Magnus Nordborg**

USC, Molecular and Computational Biology

**Association Mapping in *Arabidopsis***

I will describe progress on a large collaborative project to develop genome-wide association mapping in *Arabidopsis thaliana*, a species that is almost ideal for such studies. Most importantly, because it is highly self-fertilizing, it exists in the form of naturally occurring inbred lines that can be brought into the lab, genotyped once, and phenotyped many times. The high levels of inbreeding have also elevated linkage disequilibrium to unusually high levels for an organism with a large effective population size, making it linkage disequilibrium mapping feasible.

I will discuss results from a pilot study in which we re-sequenced 1% of the genome (one sequence read every 50–100 kb) in a panel of 96 inbred lines in order to evaluate the prospects for association mapping and design a full-scale study. I will in particular focus on the problem of spurious associations due to population structure, a potentially important obstacle to association mapping in any organism. The problem turns out to be severe in our sample, and I describe our evaluation of statistical methods to circumvent it. I also discuss the sampling strategies we are using to minimize the problem in the second, full-scale phase of the project.

Using data from our pilot study and from a Perlegen whole-genome re-sequencing study, we also learned that the marker density of the pilot study was woefully inadequate, and I describe the custom Affymetrix genotyping array that we are developing for the second phase of the project.

---

**Hua Tang**

Stanford University, Statistics

**Reconstructing Ancestry Blocks in Admixed Individuals Using High Density Genotype Data**

A chromosome in an individual of recently admixed ancestry resembles a mosaic of chromosomal segments, or ancestry blocks, each derived from a particular ancestral population. We consider the problem of inferring ancestry along the chromosomes in an admixed individual and thereby delineating the ancestry blocks. Using a simple population model, we infer gene-flow history in each individual. Compared with existing methods, which are based on a hidden Markov model, the Markov–hidden Markov model (MHMM) we propose has the advantage of accounting for the background linkage disequilibrium (LD) that exists in ancestral populations. When there are more than two ancestral groups, we allow each ancestral population to admix at a different time in

history. We use simulations to illustrate the accuracy of the inferred ancestry as well as the importance of modeling the background LD; not accounting for background LD between markers may mislead us to false inferences about mixed ancestry in an indigenous population. The MHMM makes it possible to identify genomic blocks of a particular ancestry by use of any high-density single-nucleotide – polymorphism panel. One application of our method is to perform admixture mapping without genotyping special ancestry-informative–marker panels.

---

**Russell Schwartz**

Carnegie Mellon University, Department of Biological Sciences

**Near-perfect Phylogenies and the Human Genome**

The vast quantities of genotype data now being gathered from ever larger and more diverse population samples will create an unprecedented opportunity for learning about the history of the human genome and the forces that have shaped it. Making the best possible inferences from these larger datasets will, however, also create significant computational challenges, particularly for the field of phylogenetics.

Optimal phylogenetic tree inference can be performed efficiently for small data sets and for some special cases, most notably when the data are consistent with a perfect phylogeny. Practical applications on real data, however, generally require heuristic methods that provide no bounds on solution quality and could potentially give trees far from optimal. This talk will survey work from our group on extending the class of problems for which provably optimal maximum parsimony tree inferences are possible. We focus particularly on finding “near-perfect phylogenies,” those requiring only a small number of additional mutations beyond a perfect phylogeny. We have shown that these problems can be solved optimally and efficiently for small fixed degrees of imperfection (extra mutations) on both phased and unphased data, even on large data sets. We have further developed an integer linear programming method that extends the practical bound on provable optimal solutions well beyond what is possible with the provably efficient methods. After covering these computational methods, we will then examine their application to genome-wide phylogenetic analysis through an empirical study of human variation data from the HapMap.

This work was joint with Srinath Sridhar, Kedar Dhamdhere, Guy Blelloch, Eran Halperin, R. Ravi, and Fumei Lam.

---

**Kenneth Lange**

UC-LA, Biomathematics

**Bayesian Gaussian Mixture Models for High Density Genotyping Arrays**

Affymetrix's SNP (single nucleotide polymorphism) genotyping chips have increased the scope and decreased the cost of gene mapping studies. Because each SNP is queried by multiple DNA probes, the chips present interesting challenges in genotype calling. Traditional clustering methods distinguish the three genotypes of a SNP fairly well given a large enough sample of unrelated individuals or a training sample of known genotypes. The present paper describes our attempt to improve genotype calling by constructing Gaussian penetrance models with empirically derived priors. The priors stabilize parameter estimation and borrow information collectively gathered on tens of thousands of SNPs. When data from related family members are available, Gaussian penetrance models capture the correlations in signals between relatives. With these advantages in mind, we apply the models to Affymetrix probe intensity data on 10,000 SNPs gathered on 63 genotyped individuals spread over eight pedigrees. We integrate the genotype calling model with pedigree analysis and examine a sequence of symmetry hypotheses involving the correlated probe signals. The symmetry hypotheses raise novel mathematical issues of parameterization. Using the BIC criterion, we select the best combination of symmetry assumptions.

Compared to the genotype calling results obtained from Affymetrix's software, we are able to reduce the number of no-calls substantially and quantify the level of confidence in all calls. Once pedigree analysis software can accommodate soft penetrances, we can expect to see more reliable association and linkage studies with less wasted genotyping data.

---

**Ting Chen**

USC, Computational and Molecular Biology

**Sequence-based Prioritization of Non-synonymous Single Nucleotide Polymorphisms for Study of Disease Mutations**

The increasing demand of identifying genetic variation responsible for inherited diseases has translated into a need for sophisticated methods to effectively prioritize mutations occurring in disease-associated genetic regions. Here we prioritize candidate non-synonymous single nucleotide polymorphisms (nsSNPs) through a bioinformatics approach, which includes a set of improved numeric features derived from protein sequence information and a newly designed learning model named multiple selection rule voting (MSRV). The sequence-based feature set can maximize the application scope of our approach, and the MSRV learning model can capture subtle characteristics of individual mutations. Systematic validation of the approach demonstrates that it is

capable of prioritizing causal mutations for both simple monogenic diseases and complex polygenic diseases. Applications of the approach to currently unclassified mutations suggest that 10 suspicious mutations are likely to cause diseases with strong literature supports.

---

**Paul Marjoram**

USC, Preventive Medicine

**Fast 'Coalescent' Simulation**

The amount of genome-wide molecular data is increasing rapidly, as is interest in developing methods appropriate for such data. There is a consequent increasing need for methods that are able to efficiently simulate such data. We discuss algorithms that allow rapid simulation of data according to a model that is almost, but not quite, the standard coalescent model. The algorithm ignores a class of recombination events known to affect the behavior of the genealogy of the sample, but which do not appear to affect the behavior of generated samples to any substantial degree. We show that, using this scheme, we are able to simulate large chromosomal regions, such as those appropriate in a consideration of genome-wide data, in a way that is several orders of magnitude faster than existing coalescent algorithms. As such, the algorithms provide a useful resource for those needing to simulate large quantities of data for chromosomal-length regions using an approach that is much more efficient than traditional coalescent models, or who would benefit from more rapid simulation of coalescent data for shorter regions.

---

**Justin Kennedy**

University of Connecticut, Computer Science and Engineering

**Genotype Error Detection using Hidden Markov Models of Haplotype Diversity**

The presence of genotyping errors can invalidate statistical tests for disease association, particularly for methods based on haplotype analysis. Becker et al. have recently proposed a simple likelihood ratio test approach for detecting errors in trio genotype data. Under this approach, a SNP genotype is tagged as a potential error if the likelihood associated with the original trio genotype data increases by a multiplicative factor exceeding a user selected threshold when the SNP genotype under test is deleted. In this paper we give improved error detection methods using the likelihood ratio test approach in conjunction with likelihood functions that can be efficiently computed based on a Hidden Markov Model of haplotype diversity in the population under study. Experimental results on both simulated and real datasets show that proposed methods achieve improved detection accuracy compared to previous methods with a highly scalable running time.

---

**Alexander Zelikovsky**

Georgia State University, Department of Computer Science

**Design and Validation of Methods Searching for Risk Factors in Genotype Case-Control Studies**

Accessibility of high-throughput genotyping technology allows genome/chromosome-wide association studies for common complex diseases. This paper addresses two challenges commonly facing such studies: (i) searching enormous amount of possible gene interactions and (ii) finding reproducible associations. These challenges have been traditionally addressed in statistics while here we apply computational approaches -- optimization and cross-validation.

A complex risk factor is modeled as a subset of SNP's with specified alleles and the optimization formulation asks for the one with the maximum odds ratio. To measure and compare ability of search methods to find reproducible risk factors, we propose to apply cross-validation scheme usually used for prediction validation. We have applied and cross-validated known search methods with proposed enhancements on real case-control studies for several diseases (Chron's disease, autoimmune disorder, tick-born encephalitis, lung cancer, and rheumatoid arthritis). Proposed methods are compared favorably to the exhaustive search -- they are faster, find more frequently statistically significant risk factors and have significantly higher leave-half-out cross-validation rate.

---

**Lakshmi Matukumalli**

George Mason University, Bioinformatics and Computational Biology

**Simultaneous Prediction of Multiple Traits Using Dense Genome-wide SNP Markers**

Most human whole genome association studies currently under consideration are attempting to fine-map a given complex trait from multi-stage case-control studies with the final goal of being able to use a small marker dataset for detecting disease associations. In livestock species such as cattle new algorithms are being developed for application of genome-wide SNP markers to predict the genetic values and phenotypes. In this article we contrast the methods for marker-disease association approaches in biomedical studies against the genome prediction algorithms being developed in livestock and then propose application of the genome prediction methods to studies of complex disease in humans to better enable the development of novel preventive treatments and personalized medicine.

*Sunday, 1/28/2007*

---

**Dan Gusfield**

UC Davis, Computer Science Department

### **Progress on Combinatorial Haplotyping Algorithms**

We are concerned with the Haplotype Inference (HI) Problem of computationally inferring the pairs of unobserved haplotypes that likely gave rise to the observed genotypes of the people in a sampled population, or to find partial information about the underlying haplotypes. There is a large literature that addresses the HI problem by using detailed probabilistic models of haplotype evolution, but the subsequent computational methods are much less precise in following those models. The result is that the semantics of the computations and programs used in those approaches are not always well defined, so it is difficult to understand why certain methods accurately (as some do) or inaccurately (as other do) solve the HI problem. In contrast, in the combinatorial approach, one casts the HI problem as an optimization problem with a relatively simple, precise objective function, which one can solve exactly, either in a guaranteed sense or in practice. Since the objective functions are precise and the computations exact, the semantics of those computations are precise and one can then study the results of computations to try to learn which element(s) of the haplotype model are critical in obtaining accurate HI solutions. In the First SNP/HAP meeting, we reported on such results using the Pure Parsimony objective function.

Since then, my group at UC Davis (and in collaboration with people elsewhere) have studied roughly ten additional combinatorial objective functions which either have polynomial-time solutions, or can be solved in practice (usually by using integer linear programming) on problems of sizes of current interest in biology. These evolving formulations have sought to increasingly capture biological complexity. In this talk I will discuss several of these, and in particular, formulations that relate haplotype evolution to questions about perfect phylogeny and recombination.

---

**Weixiong Zhang**

Washington University, Department of Computer Science

### **How Robust is Pure Parsimony Haplotype Inferencing?**

The pure parsimony haplotype inference (PPHI) method seeks a smallest set of unique haplotypes for explaining a set of genotypes. PPHI is supported by the observation that the number of distinct haplotypes in natural populations is small due to natural selection, genetic drift, and other evolutionary forces. Thanks to its biological support and due to its combinatorial clarity, PPHI has attracted a great deal of attention since its introduction in early 2000's; a variety of computational approaches have been developed for the problem. However, two complications arise for this model when recent gene flow, mutations,

recombination, gene conversions, and other factors yield a subset of haplotypes with low frequencies. These infrequent haplotypes tend to create large numbers of PPHI solutions and may cause the number of haplotypes to be greater than the minimum possible. In our research, we considered the robustness of PPHI in terms of finding real biological solutions. Using seven human genotype datasets, to which real haplotype information is available, and a large set of human genotype data sampled from the International HapMap project, we found that the majority of real human genotype data have many pure parsimony solutions, some of which are in the order of thousands, and PPHI performs poorly for these problem instances. Moreover, we observe that three of the seven sets with known haplotypes have more than a parsimonious number of haplotypes. Our study also revealed some intrinsic structures of human genotype data and their pure parsimony solutions.

---

**Steven Orzack**

UC-Berkeley, Fresh Pond Research Institute

**Exact and Algorithmic Methods for Haplotype Frequency Inference: What do they tell us?**

We compare an exact likelihood method and various algorithmic methods for inferring haplotype frequency from phase-unknown two-site genotypic data. We show that the exact method is preferable to the EM algorithm when estimating haplotype frequency via maximum likelihood since it allows one to readily detect multiple likelihood maxima, it is quicker (especially when many pairs of sites are analyzed), and it is easy to implement. We also show that there can be substantial differences among the algorithms with respect to the frequency estimate they generate. In addition, the frequency estimate derived from stochastic methods can differ among sample paths even when there is a single maximum of the likelihood function. We conclude that an investigator should compare the results of several inference algorithms before deciding upon an estimate of haplotype frequency and that multiple sample paths should be assessed for any stochastic algorithm. If different sample paths result in different frequency estimates, one possibility for generating a single estimate is the use of a consensus method; further research is needed to assess the usefulness of this approach.

---

**Yufeng Wu**

UC Davis, Computer Science Department

**Algorithms for Association: Mapping of Complex Diseases with Ancestral Recombination Graphs**

Association, or LD (linkage disequilibrium), mapping is an intensely-studied approach to gene mapping (genome-wide or in candidate regions) that is widely hoped to be able to efficiently locate genes influencing both complex and Mendelian traits. The logic

underlying association mapping implies that the best possible mapping results would be obtained if the genealogical history of the sampled individuals were explicitly known. Such a history would be in the form of an "ancestral recombination graph (ARG)". But despite the conceptual importance of genealogical histories to association mapping, few practical association mapping methods have explicitly used derived genealogical aspects of ARGs. Two recent papers, Zollner and Pritchard (Genetics 2005) and Minichiello and Durbin (AJHG, in press) made significant progress in exploiting genealogical history to map traits.

However, both approaches have deficiencies. In this talk, I will present new results on algorithmic problems in association mapping with full minARGs (ARGs that minimize the number of recombinations) or near-minimum ARGs. I will present an ARG sampling method that provably samples minARGs uniformly at random, and that is practical for moderate sized datasets. I will also describe a different, faster, ARG sampling method that still samples from a well-defined subspace of ARGs, and that is practical for larger sized datasets. I will also introduce novel efficient algorithms on several extensions of the "phenotype likelihood" problem, a key step in the method in Zollner and Pritchard.

Finally, I will show practical results in mapping simulated and biological data, and examine how well our methods perform, compared to methods of Zollner and Pritchard's and Minichiello and Durbin's approaches. The empirical results show our methods are much more efficient than Zollner and Pritchard's method, and can be more accurate than both Zollner and Pritchard's and Minichiello and Durbin's methods.

---

## **Gad Kimmel**

UC-Berkeley, International Computer Science Institute

### **Evaluating Disease Significance in Genome Wide Association Studies**

Due to the rapid progress in genotyping techniques, many large-scale, genome-wide disease association studies are now under way. Typically, the disorders examined are multi-factorial, and therefore researchers seeking association must consider interactions among loci and between loci and other factors. One of the challenges of large disease association studies is obtaining accurate estimates of the significance of discovered associations. The linkage disequilibrium between SNPs makes the tests highly dependent, and dependency worsens when interactions are tested. The standard way of assigning significance (p-value) is by a permutation test. Unfortunately, in large studies it is prohibitively slow to compute low p-values by this method.

We present here a faster algorithm for calculating accurately low p-values in case-control association studies. Unlike several previous methods, we do not assume a specific distribution of the traits given the genotypes. Our method is based on importance sampling and on accounting for the decay in linkage disequilibrium along the

chromosome. The algorithm is dramatically faster than the standard permutation test. On datasets mimicking medium to large association studies, it achieves a speed-up of 5,000 to 100,000, sometimes reducing running times from years to minutes. Thus, our method significantly increases the problem size range for which accurate, meaningful association results are attainable.

---

## **Vineet Bafna**

UC-San Diego, Computer Science Department

### **Detecting Structural Variations in the Genome**

Knowledge about structural variation in human genome is increasingly important in our understanding of genotype-phenotype relationships. The variations (in the form of deletions, inversions, translocations of chromosomal segments) were known to be prevalent in the genomes of tumors, but are increasingly being discovered in normal human populations. However, it is a challenge to detect such variations. We present algorithms to help detect variations in normal and tumor genomes.

First, we present a statistical method to identify large inversion polymorphisms using unusual Linkage Disequilibrium patterns from high density SNP data. Application of this method to the data from the first phase of the International HapMap project resulted in 176 candidate inversions ranging from 200 kilobases to several megabases in length. Next, we describe algorithms that help detect cancer rearrangements through an analysis of BAC end-sequence profiling data. Finally, we will present algorithms for designing primers in a multiplex PCR based protocol for detecting large deletions in patient tumor samples.

(joint work with Vikas Bansal, Ali Bashir (UCSD), Ben Raphael (Brown). Experimental collaborators include Colin Collins, and Stas Volik (UCSF), and Dennis **Carson, and Y. T. Liu (UCSD Moores cancer center)**)

---

## **Jing Li**

Case Western Reserve University, Electrical Engineering and Computer Science  
Department

### **Prioritize and Select SNPs for Association Studies with Multi-stage Designs**

Large-scale whole genome association studies are increasingly common, due in large part to recent advances in genotyping technology. With this change in paradigm for genetic studies of complex diseases, it is vital to develop valid, powerful, and efficient statistical tools and approaches to evaluate such data. Despite a dramatic drop in genotyping costs, it might be still too expensive for many researchers to genotype thousands of individuals for hundreds of thousands SNPs for large-scale whole genome association studies. A two-stage (or multi-stage) design has been a promising alternative: in the first stage, only

a fraction of samples are genotyped and tested using a dense set of SNPs, and only a small subset of markers that show moderate associations with the disease will be genotyped in the second (or later) stage(s). Multi-stage designs have also been used in candidate gene association studies, usually in regions that have shown strong signals by linkage studies. To decide which set of SNPs to be genotyped in the next stage, a common practice is to utilize a simple test (such as a  $\chi^2$  test for case-control data) and a liberal significance level without corrections for multiple testing, to ensure that no true signals will be filtered out. In this paper, I have developed a novel SNP selection procedure within the framework of two-stage (or multi-stage) designs. Based on data from stage one, the method explicitly explores the correlation (linkage disequilibrium) among SNPs and their possible interactions in determining the disease phenotype. Compared with a regular two-stage design, the approach can select a much reduced set of SNPs with high discriminative power for stage two. Therefore, it not only reduces the genotyping cost in stage two, but also may increase the statistical power. Simulations have been performed on genome wide association studies, as well as on candidate gene association studies. Test results have shown that the procedure can further reduce the number of SNPs required in later stage(s) with improved power to detect association.

---

**Charles Kooperberg**

University of Washington, Fred Hutchinson Cancer Research Center

### **Identifying Interactions in Genetic Association Studies**

It is becoming increasingly common to conduct association studies involving thousands of Single Nucleotide Polymorphisms (SNPs). For many of these studies, interest will not be limited to just the characterization of individual SNPs or haplotypes that are associated with a disease outcome, but will include the identification of interactions between SNPs within a gene (as in haplotype effect), between genes (epistasis), or between gene and environment (e.g., drugs, smoking, and alcohol consumption).

The potential enormous number of interactions leads to serious problems, both related to multiple comparisons (power), and related to computation that require the development of new statistical methodology. In this talk I will about multi-stage approaches to identify epistasis, and when such approaches are more powerful than direct searches.

---

---

END