Statistical Analysis of High-Throughput Genetic Data

Jiahua Chen (University of British Columbia),
Yuejiao Cindy Fu (York University),
Mary Lesperance (University of Victoria),
David Siegmund (Stanford University),
Heping Zhang (Yale University),
Hongyu Zhao (Yale University)

June 25–June 29, 2007

1 Overview of the Field

Recent years have seen the rapid accumulation of various types of genomic information due to concerted efforts by the scientific community and advances in molecular technologies. The publication of the human genome sequences and the sequences of many other species represent a great milestone in our scientific history. In addition, a large number of genetic variants responsible for the diversity seen in a given organism have also been identified. For example, more than 10 million "common" single nucleotide polymorphisms (SNPs) are estimated to be present in humans, and a many of these have been discovered and documented in the literature and public databases. Parallel to SNP discovery, microarrays have made it possible to examine gene expression levels at the genome level, to study genomic-wide DNA copy number changes, which are ubiquitous in human DNA, but especially in cancer tissue, to identify essentially all the binding targets of a transcription factor under different conditions, to evaluate epigenetic controls of genetic regulation, and to collect other types of information across the whole genome. All of these great successes in knowledge and data acquisition have created opportunities and challenges for statisticians, mathematicians, computer scientists, engineers, physicists, and other quantitative scientists to work closely with biologists and biomedical researchers. Mathematical scientists can assist biologists to utilize efficiently such enormous amounts of data, to identify genetic variants underlying human diseases, to dissect biological pathways and address many other scientific inquiries.

One immediate and apparent challenge in statistical genetics is the computational load of any statistical method when applied to high-throughput data set. One particular problem, for example, is that human error can occur at any stage of data collection. To identify inconsistency and sometimes make necessary corrections can greatly reduce the potential bias in the subsequent statistical analysis. This is an important area still under intensive research. An important research product is computer software that provides convenient tools to biological researchers.

Another area to which this workshop devoted substantial attention is the variable selection problem for the large parameter spaces that are most relevant in statistical genetics. A surge of publications has started to appear in many statistical journals. This workshop foresaw this tread and had many presentations in this hot research area.

Finite mixture models have long found applications in statistical genetics. There has been substantial progress in developing statistical methodology in recent years. A vast proportion of them have implications

to the way some genetic data should be analyzed. Some provide completely new means of analysis. This workshop seized the opportunity to present these new statistical ideas to biologically oriented researchers who may not yet be following these developments due to their different research focus.

2 Outline of the Workshop

The workshop has invited active statistical researchers with diverse expertise in statistical genetics to meet and exchange ideas. We strived to invite scientists from all over the world, scientists working on a wide variety of genetic problems, and scientists that are in all stages of their careers. The workshop participants include scientists from the National University of Singapore, University of Hong Kong, Hebrew University, as well as a large concentration of scientists from North America. Many participants are already well known in statistical genetics such as Professor David Siegmund who is a member of the National Academy of Sciences in the USA and renowned for his advanced research in statistical genetics, Professor Jun Liu who was the winner of the prestigious award of the Committee of Presidents of Statistical Societies (COPSS) and who has made outstanding contributions in Bayesian methods among others. Both Professor Heping Zhang and Professor Hongyu Zhao from Yale University have their research strongly supported by research grants from the National Science Foundation, and supervise large research labs on statistical genetics. Professor Shelly Bull from the University of Toronto is one of the leading researchers in statistical genetics in Canada and directs the Samuel Lunenfeld Research Institute of Mount Sinai Hospital. Dr. Dongsheng Tu, from Queens University, Canada, and Professor Benny Zee from Chinese University of Hong Kong and many other participants have close collaborative relationship with clinical geneticists. They discussed their first hand experience on the relevance of research progress in genetics, drug development, and treatment improvements.

This workshop also includes researchers who excelled in other areas of statistics, and look forward to extending their research impact to statistical genetics. Some participants are at the early stage of their research career but are attracted to this very promising area and its abundant opportunities to to make a significant impact. The workshop also provided opportunities to many post-doctoral scholars and current graduate students. For many, this is their first time to have distinguished listening to their research ideas and results in detailed presentations and in after presentation social activities.

The workshop takes the form of inviting participants to submit their latest research for possible presentation and discussion. Over 28 researchers presented their work in either one-hour plenary talks or in 30-minute invited talks. Lengthy discussions followed each plenary talk, and sessions of invited talks. Many researchers exchanged their email addresses for more specific future one-to-one correspondence.

3 Presentation Highlights

3.1 Linkage Mapping and Association Studies

Gene mapping, either by linkage or by association, is concerned with identifying genomic regions that harbor genetic polymorphisms that contribute to phenotypes of interest. It is practiced by scientists studying plants, animals, animal models of human genetics, and humans. Because changes in the technology of high throughput genotyping introduce new experimental possibilities, the subject is in a state of rapid change.

The opening plenary talk by Professor D. Siegmund posits an intriguing question to all statisticians working on linkage analysis in experimental or in human genetics: “Do complex statistical methods help in mapping complex and quantitative traits?” Striving to find a statistical model that match various characteristics of genetic data, statisticians may want to develop more and more sophisticated models. One may choose to use (i) standard statistical methods for genome scans, or (ii) more complex methods designed to take advantage of the possibilities of gene-gene and gene-environment interactions. The trade-off is between not taking necessary factors into proper consideration, and paying the price of devoting a portion of data (in some sense) to capture complex modeling features. The presentation included enlightened theoretical developments and well thought-out computer simulations.

The presentation by Professor Daniel Weeks introduces “linkage statistics that model relationship uncertainty.” In linkage analysis, reported family relationships are assumed to be correct, and thus misspecified

relationships can lead to erroneous results. Often, studies either discard individuals with erroneous relationships or use the best possible alternative pedigree structure. The linkage statistics developed by Professor Weeks and his collaborators model relationship uncertainty by properly weighing different possible true relationships. Using simulated data containing relationship errors, statistical methods are compared for maximum likelihood statistic and non-parametric LOD scores for small pedigrees and for large pedigrees.

A second plenary talk by Professor Warren Ewens: "The transmission-disequilibrium test (TDT) and its generalizations" addressed family based association tests, The transmission-disequilibrium test was developed as a test of linkage between a marker locus and a purported disease susceptibility locus. It is however more frequently used today as a test of association between the alleles at these respective loci. Complications to the test arise with the current availability of hundreds of thousands of marker loci. There hence have been substantial new developments to handle this and other novel situations. Professor Ewens' presentation provided a timely update and provoked participants to think carefully about the test.

The presentation of Professor Josee Dupuis ("Mapping quantitative trait genes using high density SNP scans in extended families: challenges and partial solutions") also addresses the challenge in statistical analysis of high-throughput data with familial relationship. High density SNP scans are currently being performed on several family-based population samples with multiple phenotypes available. There are several statistical challenges related to finding genes influencing quantitative traits in such rich datasets. Is there any value to performing linkage analysis using dense SNPs, when it is believed that genome-wide association (GWA) analysis is the answer to many questions? With the huge number of tests that result from a GWA scan with multiple phenotypic outcomes, how does one control type-I error and still retain some power to detect effects of modest size? Should one perform analyses to safe-guard against false positive results arising from population stratification, at a cost of a possible reduction in power, or should one rely on replication studies to limit false positive associations? This presentation provided some partial solutions to these questions in the context of the Framingham Heart Study and urged other researchers to think hard for further improvements.

While many new statistical methods presented in this workshop deals with straightforward association studies, Professor Shelly Bull reminded us that genome-wide association studies(GWA) are typically designed with multiple stages. Whether the second stage involves an independent sample of individuals or a family-based design, genetic effect estimates at a second stage will be less optimistic that those obtained at the first stage, due to selection bias arising from genome-wide screening. Motivated by a GWA study of the genetics of complications of type I diabetes, Professor Bull and her research team presented their research result on evaluating implications for power and bias in alternative designs and analytic strategies for detection and mapping of gene regions using high-density SNP arrays. The bias of genetic effect estimates depends on sample size, true effect size, minor allele frequency, and screening stringency. They found that the application of computationally-intensive bootstrap estimation yields less biased effect estimates, and hence more realistic specifications for replication in a subsequent stage.

3.2 Statistical Genomics and Computational Biology

Information on the parental origin of each of two alleles at each locus on chromosome cannot be measured directly but must inferred by statistical means. Such haplotype information is very useful in genetic research. Haplotypes have hence become a central topic in genetics analysis in recent years.

It is a rule rather than an exception, that the data do not contain the enough information to completely determine haplotypes. Utilizing information from genotypes of a good sized neighborhood loci, a probability distribution with good certainty is possible. Yet the sheer size of the number of possible configurations makes complete numeration beyond the power of modern computers. Instead, a Bayesian approach via stochastic exploration or some other cleverly designed approaches can provide a solution. Dr. Jun Liu from Harvard University gave a review of his recent work on Bayesian Inference of Haplotypes and Epistasis. These included Bayesian models that have been developed over the past few years for haplotype inference and included a new hierarchical Bayes model and a Bayesian approach to detect multi-locus interactions (epistasis) for case-control association studies.

The talk by Dr. Fei Zou from the University of North Carolina at Chapel Hill focused on a very timely topic, the identification of genetic variants affecting gene expression levels (eQTL). She investigated use of Bayesian methods and focused on the computation of posterior distributions. A common feature of her research and other work presented at the workshop is a massive data set that poses a serious numerical

challenges. In her talk titled "Fast Bayesian eQTL Analysis," Dr. Zou discussed a Bayesian linkage model that offers highly interpretable posterior densities for linkage. Instead of direct numerical computation the likelihood functions, which is very costly, she has developed Laplace approximations that are highly accurate and efficient in numerical implementation so that the computation of posterior densities for over 30,000 transcripts becomes feasible.

In another eQTL talk, Dr. Liang Chen from University of Southern California focused on "Considering Dependency among Genes for the False discovery Control of eQTL mapping. In most studies, the dependency among genes is largely ignored in consideration of multiple comparison adjustments. However, such dependency may be strong for eQTL data, and it can have significant impact on the outcome of the data analysis and its interpretation. Dr. Chen and colleagues introduced a weighted version of false discovery control to improve the statistical power to identify eQTL. The relative performance of the new method in eQTL studies was illustrated through simulations and data analysis.

Motivated by high throughput genotyping platforms, Dr. Ingo Ruczinski from Johns Hopkins University spoke on "An integrated approach for the assessment of chromosomal abnormalities using SNP chip estimates of genotype, copy number, and uncertainty measurements". Copy number variations have become a focus of human genetics research in the past several years, yet the detection and quantification remain a challenging statistical problem. Dr. Ruczinski and colleagues developed Hidden Markov Models (HMMs) based on SNP array data. Their approach can simultaneously integrate gene copy number estimates, genotype calls, and the corresponding confidence scores when available. They have further implemented their methods in the R programming language.

Data integration is an intensively researched area. In his talk on "Bayesian methods for reconstructing transcriptional regulatory networks" Dr. Hongyu Zhao described a Bayesian error analysis model to integrate protein-DNA binding data and gene expression data to reconstruct transcriptional regulatory networks. There are two unique aspects to this proposed model. First, transcription is modeled as a set of biochemical reactions, and a linear system model with clear biological interpretation is developed. Second, measurement errors in both protein-DNA binding data and gene expression data are explicitly considered in a Bayesian hierarchical model framework. Model parameters are inferred through Markov chain Monte Carlo. The usefulness of this approach was demonstrated through its application to infer transcriptional regulatory networks in the yeast cell cycle.

With a network focus, Dr. Steve Horvath from the University of California at Los Angeles discussed "Weighted Gene Co-Expression Network Analysis and Other Systems Genetic Approaches for finding Complex Disease Genes". His talk covered several theoretical topics, including network construction, module definition, network-based gene screening, and differential network analysis. The usefulness of these methods were illustrated using several applications, including i) screening for biomarkers of kidney transplantion success ii) finding obesity related genes in mice, and iii) complex disease gene mapping in humans.

3.3 Variable Selection and Data Mining in Genomics

One goal of the analysis of high throughput data is to identify genetic variations that are responsible for phenotypic variations such as cancer and diabetes. As the name suggests, the high throughput data provides hundreds of thousands candidate genetic variations to be screened. Paradoxically, the large volume of data has the potential to make the statistical inference more powerful, but it poses serious challenges to develop statistical methods that are solid in theory and efficient in computation. Genetically, it is not realistic to chase after hundreds of suspects. Further, with so many candidates, many subsets of these genetic variations may appear to be strongly associated with the phenotypic variation. How do we identify a small set of true culprits, quantify the associated uncertainty, and statistically measure the degree of success?

This workshop provides an ideal platform for participants to report their latest achievements in developing novel statistical methods in identifying candidate genes or biomarkers for diseases.

The first presentation in this area was given by Professor Heping Zhang from Yale University, entitled "Tree and forest based approaches to genomewide association studies". He compared the statistical analysis in this area to the practice of gold mining. The key is not how pure the final product obtained in the initial screen, but concentration of gold for further purification. In this presentation, Professor Zhang presented tree and forest based analyses in genomewide association studies to identify high risk genes and gene-gene interactions. Simulation studies were used to demonstrate the potential of the proposed method. A re-analysis

of existing data from a genetic study of age-related macular degeneration also revealed the advantages of the proposed method in identifying a possibly protective variant.

Professor Jiahua Chen from University of British Columbia spoke on "Extended Bayesian Information Criteria for Model Selection with Large Model Space". It is observed that the model space for analyzing high throughput data sets is extremely large. The special terminology of "large-n-small-p" has recently been coined to describe statistical models of this nature. Many participants of the workshop have make research contribution in this area, and it is generally recognized that the ordinary Bayes information criterion is too liberal for model selection when the model space is large. In this talk, Professor Chen re-examined the Bayesian paradigm for model selection and proposed an extended family of Bayes information criteria. Unlike the original Bayes information criterion, which balances the log likelihood by a penalty on the number of unknown parameters, the extended Bayes information criteria take into account both the number of unknown parameters and the complexity of the model space. The consistency of the extended Bayes information criteria is established, and its performance in various situations is evaluated by simulation studies. It is also compared with the original Bayes information criterion in terms of positive selection rate and false discovery rate in problems of variable selection. It is demonstrated that the extended Bayes information criteria incurs some loss in positive selection rate but tightly controls false discovery rate, a desirable property in many applications. The extended Bayes information criteria are extremely useful for variable selection in problems with moderate sample size but very large numbers of covariates, especially in genome-wide association studies.

"Penalized Methods for Variable Selection and Estimation with High Dimensional Data" delivered by Professor Jian Huang from the University of Iowa, discussed some non-Bayesian penalized approaches for variable selection, including "lasso" and bridge penalties. The new type of variable selection represented by lasso is one of the latest techniques developed in the statistical literature. It suggests to fit models according to some likelihood or least squares criteria, but penalizes complex models with a non-smooth function. By adjusting a tuning parameter, the severity of the penalty can be increased. A very interesting property, as a consequence of the non-smooth penalty, is that the fitted values of many regression coefficients are likely to be exactly zero, rather than simply small. When the number of variable subjects to selection remains low compared to the sample size, the procedure has been found to have desirable properties, as well as being consistent in general. Since this type of procedure can be directly applied to the "small-n-large-p" problems and is numerically efficient, it is of great interest whether these methods also possess desirable statistical properties under new circumstances. Due to its importance in genetic applications, there have been intensive research on the properties of this kind of variable selection procedures. Professor Huang is one of the first few made significant contributions in this respect. He presented some preliminary results concerning variable selection consistency and asymptotic oracle properties of lasso and bridge methods in high-dimensional settings. Under some sparsity assumptions, he showed that the majority of the true variables will survive the first round of screening under the lasso and other methods. He also illustrated applications of these methods to the analysis of binary and censored outcomes with high-dimensional genomic covariate data.

The importance of the variable selection procedures for small-n-large-p situations is further discussed in other presentations. Professor Jie Peng from University of California-Davis presented results on "Model Selection for QTL Mapping," where she considered the lasso, but her application is to QTL mapping. In addition, she also proposed another method called the "fence method," which is useful for model selection in both regression and random effects models and enjoys some computational advantages over the BIC and AIC type of methods in searching of the model space. The idea is first to build a statistical fence to eliminate incorrect models and then among the correct models, to select optimal one in a defined sense.

The workshop concluded with a presentation by Professor Zehua Chen from the National University of Singapore, entitled "A tournament approach to model selection with applications in genome-wide association studies". Ultimately, no matter how innovative a statistical model selection procedure is, to make an impact in genetic application, it must be computationally feasible. His presentation echoes the fact that the sheer amount of the covariates (genetic markers) and a relatively small sample size make many existing model selection methods infeasible. Even for newly developed methods that are renowned for their computational efficiency, additional measures make be taken to enhance their applicability. To this end, he presented a novel tournament approach to model selection for the situation that the number of covariates under consideration far exceeds the sample size. The approach consists of a stage-wise screening procedure, which mimics the rounds of competitions in a tournament, and an extended Bayes information criterion.

Although the speakers considered different methods, they all recognized one of the common challenges in genomic data analysis, traditionally called variable selection in statistics. Whatever the name, the methods have a defined pool of models. The selection of a model or models from the pool is then based on a penalization criterion that considers the size of the model pool as well as the complexity of the selected model. While the progress is promising, it is also evident from all presentations that the challenge has not been fully resolved. The participants were very pleased to find so much common ground, and surprised to see so many others are working on the same problem. Needless to say, heated discussions and exchange of references and emails were carried out all the time.

3.4 Mixture Models

A unique feature of this workshop is to explore the use of specific, advanced statistical methodologies in analysis of high-throughput genetic Data. Mixture models have a long history of being applied to statistical genetics. In classical linkage analysis, it is often suspected that only a subgroup of patients have a disease gene which is linked to the marker. Detecting the existence of this subgroup amounts to the acceptance of a mixture model description of the whole population. Consider the analysis of high-throughput genetic data where hundreds of thousands gene expression levels are obtained, it is plausible to assume that only a proportion of these levels are elevated. Again, finite mixture models can be naturally employed.

This workshop invited several researchers working on statistical theory of finite mixture models. Bruce Lindsay (Distinguished Professor and head of the Department of Statistics, Penn State University) gave a plenary talk on "Modes, mixtures, and diffusion kernels - Building a three way theory". Lindsay's talk was a report on a new type of model for genomic sequence data as well as inference for that model. Interest is focused on data sampled from some population, and a tree of relationships for those sequences is created. Each sampled sequence is modelled as having first been sampled from a population of ancestral sequences; that population is modelled by an unknown distribution Q on sequence space. The chosen sequence then undergoes T time units of evolution using Markov Chains that describe mutation and recombination processes before it is observed. The goal is to go backward in time T units and estimate the ancestral sequence distribution Q, as well as which modern sequence maps to each ancestor (or ancestors in the case of recombination). One methodology involves using maximum likelihood inference on the model for each fixed T on a grid, then linking the estimated ancestors together over T to create an ancestral tree. A second method is to use modal inference. This new method, which can be motivated by a reverse diffusion argument, seems to give results similar to maximum likelihood with much less programming and computation time.

Professor Ji-Ping Wang from Northwestern University presented his research entitled "Statistical method for nucleosome DNA sequence alignment and linker length preference prediction in Eukaryotic cells". Eukaryotic DNAs exist in a highly compacted form known as chromatin. The nucleosome is the fundamental repeating subunit of chromatin, formed by wrapping a short stretch of DNA, 147bp in length, around four pairs of histone proteins. Nucleosome DNA obtained by experiments however varies in length due to imperfect digestion. Wang developed a mixture model that characterizes the known dinucleotide periodicity probabilistically to improve the alignment of nucleosomal DNAs. To further investigate the chromatin structure, he obtained experimentally cloned and sequenced di-nucleosome sequences from yeast, chicken and human. Each dinucleosome sequence roughly covered two nucleosomes (located toward the two ends) with a linker DNA in between. A Hidden Markov Model model was trained based on the nucleosome sequence alignment for prediction of nucleosome positioning. Results show that Eukaryotic cells do favor periodic linker length in chromatin forming on a roughly 10 bp basis, however with two different forms, i.e. with peaks around 5 bps or 10bps.

A Ph.d student, Pengfei Li from the University of Waterloo, presented yet another new development in finite mixture models. As pointed out by Professor Hangfeng Chen, it is highly desirable to develop a theory that does not rely on the artificial compactness assumption. At the same time, it is also noticed that some widely used finite mixture models in statistical genetics, have infinite Fisher information, which are also excluded from most existing theory for finite mixture models. To overcome these shortcomings, a new class of hypothesis tests is proposed. It is show the new method not only has broader range of applications, but also provide a more efficient procedure.

The analysis of directional data occupies a special place in statistics. Such data not only raise as measurement of directions, but also as measurement of phases for physical processes with periodic behaviors.

They are useful in genetic in modeling the circles of gene expression levels among others. The presentation by Professor Yuejiao Fu from York University discusses inference problems related to finite mixtures of von Mises distributions and its applications to statistical genetics. Fu proposed the use of the modified likelihood ratio test and the iterative modified likelihood ratio test in general two-component von Mises mixture with a structural parameter. Two accuracy enhancing methods are developed. The limiting distributions of the resulting test statistics are derived. Simulations show that the test statistics have accurate type I errors and adequate power.

4 Outcome of the Meeting

We would like to thank BIRS for the excellent facilities and great support. Our workshop is a big success. It provided a forum for statisticians to mix with biologists allowing both groups to identify important scientific questions. The talks were of high quality. Participants expressed their great interest for further discussion. The continuation of already existing collaborations and the creation of new ones is an important outcome of the 5-day workshop. We also received a lot of feedback from the participants:

"It's a wonderful meeting and I enjoyed it very much!"

"Thank you all for organizing such an excellent workshop! I really enjoyed it and was very stimulated by many of the talks."

"I wanted to thank you all for the fantastic job you did in organizing the workshop. I have learned a lot."

"I have no doubt that this workshop is very beneficial and rewarding experience. Thank you very much!"

"Overall this is one of THE BEST workshops I have attended."

5 Press Release

From Mendel's agricultural experiments on peas to human genome projects on chromosomes, geneticists as well as the general public have been fascinated by factors that inherently define the vastly diverse characters of the living world. While only most obvious traits such as the color of flowers were observed in the old days, modern techniques enable geneticists to measure hundreds of thousands of genes of an organism on a single microarray chip. A high amount of high-throughput data are thus generated routinely. The task of identifying important genes out of tens of thousands, which are associated with traits such as cancer and diabetes, demands serious effort in designing effective statistical analysis procedures.

From June 24 -29, 2007, a group of statisticians/geneticists from all over the world will come to Banff International Research Station to exchange ideas and report their advances on the analysis of high-throughput data. This event is co-organized by Professor Jiahua Chen from the University of Waterloo, who has recently been awarded the tier I Canada Research Chair in Statistical Genetics at the University of British Columbia. Other organizers include Professor Mary Lesperance from the University of Victoria, Professor Yuejiao Fu from York University, Canada; Professor David Siegmund from Stanford University who is a member of the National Academy of Sciences in the USA and renowned for his advanced research in statistical genetics; and Professors Heping Zhang and Hongyu Zhao from the School of Public Health, Yale University. Professor Heping Zhang is director of the Collaborative Center for Statistics in Science, and Professor Hongyu Zhao is the director of the Center for Statistical Genomics and Proteomics.

6 List of Participants

Allison, David (University of Alabama at Birmingham)
Baglivo, Jenny (Boston College)
Bryan, Jennifer (University of British Columbia)
Bull, Shelley (University of Toronto)
Chen, Jiahua (University of British Columbia)
Chen, Zehua (National University of Singapore)
Chen, Hangfeng (Bowling Green State University)
Chen, Liang (University of Southern California, Los Angeles)

Dupuis, Josee (Boston University School of Public Health)
Ewens, Warren (University of Pennsylvania, Philadelphia)
Fan, Guangzhe (University of Waterloo)
Fu, Yuejiao Cindy (York University)
He, Xuming (University of Illinois at Urbana-Champaign)
He, Wenqing (University of Western Ontario)
Horvath, Steve (University of California, Los Angeles)
Huang, Jian (University of Iowa)
Lesperance, Mary (University of Victoria)
Li, Pengfei (University of Waterloo
Lindsay, Bruce (Pennsylvania State University)
Liu, Ching-Ti (Yale University)
Liu, Jun (Harvard University)
Liu, Lei (Yale University)
Molinaro, Annette (Yale University)
Pelizzola, Mattia (Yale University School of Medicine)
Peng, Jie (University of California, Davis)
Rao, J. Sunil (Case Western Reserves University)
Ruczinski, Ingo (Johns Hopkins University)
Shao, Yongzhao (New York University)
Siegmund, David (Stanford University)
Song, Peter (University of Waterloo)
Tu, Dongsheng (Queen's University)
Wang, Ji-Ping (Northwestern University)
Wang, Huixia (North Carolina State University)
Wang, Hsiao-Hsuan (York University)
Weeks, Daniel (Univesity of Pittsburgh)
Yakir, Benjamin (Hebrew University Mount Scopus)
Zee, Chung-Ying (Chinese University of Hong Kong)
Zhang, Heping (Yale University)
Zhang, Meizhuo (Yale Univesity, School of Public Health)
Zhao, Hongyu (Yale University)
Zou, Fei (University of North Carolina at Chapel Hill)

References

[1] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle, in *Second International Symposium on Information Theory*, eds. B.N. Petrox and F. Caski. Budapest: Akademiai Kiado, 267, 1973.

[2] Y. Benjamini, and Y. Hochberg, Controlling the false discovery rate — A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* B, **57**, (1995) 289-300.

[3] K. W. Broman, and T. P. Speed, A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Statist. Soc. B* **64**, (2002) 641-656.

[4] S. C. Chen, and B. G. Lindsay, Building mixture trees from binary sequence data. *Biometrika*, textbf93, (2006) 843-860.

[5] J, Fan and J. Lv Sure Independence Screening for Ultra-High Dimensional Feature Space. *Annals of Statistics*. To apear.

[6] A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E. E. Schadt, T. A. Drake, A. J. Lusis, and S. Horvath, Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* **18**, (2006) 2:e130

[7] J. N. Hirschhorn , and M. J. Daly Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6** (2005), 95-108.

[8] H. Matsuzaki, H Loi, S. Dong, Y. Y. Tsai, J. Fang, J. Law, X. Di, W. M. Liu, G. Yang, G. Liu, J. Huang, G. C. Kennedy, T. B. Ryder, G. A. Marcus, P. S. Walsh, M. D. Shriver, J. M. Puck, K. M. Jones, R. Mei, Parallel genotyping of over 10,000 SNPs using, a one-primer assay on a high-density oligonucleotide array. *Genome Res* **14** (2004a) 414-425.

[9] H, Matsuzaki, S. L. Dong, H. Loi, X. J. Di, G. Y. Liu, E. Hubbell, J. Law, T. Berntsen, M. Chadha, H. Hui, G. R. Yang, G. C. Kennedy, T. A. Webster, S. Cawley, P. S. Walsh, K. W. Jones, S. P. A. Fodor, and R. Mei, Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* **1** (2004b), 109-111.

[10] S. S. Murray, A. Oliphant, R. Shen, C. McBride, R. J. Steeke, S. G. Shannon, T. Rubano, B. G. Kermani, J. B. Fan, M. S. Chee, and M. S. T. Hansen A highly informative SNP linkage panel for human genetic studies. *Nat Methods* **1** (2004) 113-117.

[11] G. Schwarz, Estimating the dimension of a model, *Annuals of Statistics*, **6**, (1978), 461-464.

[12] D. Siegmund, Model selection in irregular problems: Application to mapping quantitative trait loci. *Biometrika* **91** (2004), 785-800.

[13] R. S. Spielman, R. E. McGinnis, W. J. Ewens, The transmission/disequilibrium test detects cosegregation and linkage. *Am. J. Hum. Genet.*, **54**, (1994), 559-60.

[14] S. Ray and B. G. Lindsay, The topography of multivariate normal mixtures. *Ann. Statist.*, **33** (2005), 2042–2065.

[15] R. Tibshirani, Regression shrinkage and selection via the LASSO, *J. Roy. Statist. Soc. Ser. B*, **58** (1996), 267-288.

[16] W. Y. Wang, B. J. Barratt, D. G. Clayton, J. A. Todd (2005) Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* **6** (2005) 109-118.

[17] C. H. Zhang and J. Huang The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Annals of Statistics*. To appear. Manuscript.

[18] The International HapMap Consortium. The International HapMap Project. *Nature* **426** (2003), 789-796.

[19] Y. Zhang, and J.S. Liu (2007) Bayesian inference of epistatic interactions in case-control studies. *Nat Genet.* **39**, (2007)1167-73.