

BRC bioinformatics

The twelve prospective bioinformatics students who visited on February 20 greatly impressed us with their credentials, knowledge, and enthusiasm.

From the Director

New faculty, new and renewed grants, a new group of outstanding applicants, and a new computer cluster make Spring 2004 a welcome season at the BRC.

Dr. Philip Awadalla, a new faculty member in Genetics whose research is in evolutionary and population genetics, has joined the BRC and is sharing an office with Dr. Greg Gibson. Welcome again to recent arrivals Dr. Jung-Ying Tzeng (who you'll meet below) and Dr. Steffen Heber (introduced in last fall's newsletter).

Our congratulations to post-doc Garrick Skalski, who was awarded an NSF research grant for work on the population genetics of striped bass, and to his collaborator, Dr. Craig Sullivan in Zoology.

NSF and NIH have told us that our grants for the Summer Institute in Statistical Genetics will be renewed for another five years, allowing us to continue to support deserving students.

Twelve prospective bioinformatics students visited on February 20 and greatly impressed us with their

credentials, knowledge, and enthusiasm. We are hopeful that they will join us in the Fall and continue the tradition of excellence in our program.

Chris Basten has installed a new Mac cluster that is already running near capacity around the clock. With a new computer room across the hall, workstation users in room 1514 will be freed from the flashing lights, noise, heat and bulk of our two clusters.

While the bioinformatics students celebrated the Year of the Monkey at their annual Chinese New Year dumpling party, it seemed to be the Year of the Monkey Wrench in the office, what with water spilling into the BRC from upstairs labs, a heating system that roasted some and froze others, and a break-in that resulted in a loss of many of our computers. Our sincere thanks to Stan Martin, who was able to restore computer access within 24 hours to everyone affected by the break-in. Yes, we're glad it's Spring.

Bruce Weir
weir@stat.ncsu.edu

The grant application received the highest priority score of NIH proposals in its group.

NSF, NIH Award \$1 Million Grant to Summer Institute

NSF and NIH have renewed their funding of the Summer Institute in Statistical Genetics for another five-year period, awarding over \$1 million, nearly double that of the current funding period. The grant supports tuition and travel scholarships for student participants and instructor stipends.

The Institute grant application received the highest priority score of NIH proposals in its group, testimony to the program's scientific relevance, popularity, and expert faculty.

The Institute consists of a series of three-day workshops in current areas of statistical genetics. Now in its ninth year, the Institute has trained nearly 1500 students, with about 250 participants attending annually in recent years.

This year's Institute will be held at NC State May 24

through June 11. The 18 workshops include basic genetics and statistics, molecular and population genetics, quantitative genetics, forensic DNA analyses, bioinformatics, association mapping, microarray analysis, genetic data in clinical trials, Markov Chain Monte Carlo analyses, and coalescent theory.

International interest has always been strong, with a significant number of international participants attending the Institute in Raleigh. This year, the Institute will offer workshops in Faro, Portugal, from July 19 to 28, at the invitation of the University of Algarve. Previous international Institutes have been held in Melbourne, Australia; Dublin, Ireland; and Christchurch, New Zealand.

More information about the Summer Institute is available at <http://statgen.ncsu.edu>.

Now in its 9th year, the 2004 Summer Institute will offer workshops at NC State in Raleigh and in Faro, Portugal.

GSK Intern Helps Hone Gene Mapping Techniques

“Linkage and association studies may each miss some true disease-related genes or report some wrong genes. Our method [aims] to reduce the probability of misidentifications.”

Li Li,
Grad Student Intern

Li Li, a fourth-year doctoral student in bioinformatics, is working with scientists in the Genetic Data Science group at GlaxoSmithKline to increase the reliability of statistical techniques for detecting genes related to disease. Their method merges the results of two principal gene mapping techniques, called linkage studies and association mapping.

“Linkage and association studies may each miss some true disease-related genes or report some wrong genes,” said Li. “Our method is to combine results from these two analyses using statistical

tests to reduce the probability of misidentifications.” The project is led by Dr. Sharon Browning, who is also adjunct professor of statistics at NC State.

“The internship at GSK is a wonderful experience,” said Li. “My group manager, Meg Ehm, makes every effort to ensure that my work can be part of my thesis, which is really important. She creates lots of opportunities for me to participate in other projects, meetings, and conferences, such as the 2003 annual meeting of the American Society of Human Genetics.”

Li started her internship at GSK in May 2003. “In less than a year at GSK, I’ve learned a great deal of useful information about linkage and association studies that you can’t find in textbooks or the classroom,” she said. “Also, I am seeing what the real world is like, what kind of techniques companies need most.”

Strengths in statistics, programming, and communication are essential for working in the company, said Li, who has a bachelor’s in biology. She found her courses in statistical theory and genetic data analysis to be very useful, and at least one programming language is essential (she uses Java primarily, but Perl, UNIX shell, SAS, and Splus are also used).



Grad students celebrate the Year of the Monkey at their annual Chinese New Year dumpling party.

New DNAMix Software Improves Forensic Analyses

DNAMix was the first computer program to calculate probabilities for mixed DNA crime samples that took proper account of the number of contributors.

“We wanted to make the program easier for forensic scientists working in the lab to use.”

Gary Beecham,
Grad Student and
Programmer of DNAMix v.3

Crime samples that are mixtures of DNA from two or more people, as occurs in rape cases and some other crimes, are complicated to analyze because DNA from a number of unknown persons might yield the same DNA coding pattern that is observed in the mixture.

DNAMix, first developed in 1997 by NC State undergraduate John Storey, was the first computer program to calculate probabilities for mixed DNA samples in a way that took proper account of the number of contributors. Police in Victoria, Australia, began using DNAMix in criminal investigations five years ago.

Now, thanks to the efforts of first-year bioinformatics graduate student Gary Beecham, a new version of DNAMix (v.3) is available that is faster and easier to use, has greater error-checking capabilities, and calculates confidence intervals for the estimated probabilities.

“We wanted to make the program easier for forensic scientists working in the lab to use,” said Beecham, who worked closely with Dr. Bruce Weir, Director of the Bioinformatics Research Center, and with users in Australia, who are validating the new program.

“This software would have been helpful in the O.J. Simpson case,” said Weir, who testified in the trial in 1995. “The Simpson trial was in the early days when the statistical theory was not as widely accepted and no software was available. DNAMix uses a published methodology that is now accepted by many people, and it is an established and validated software program.”

Earlier versions of DNAMix did not calculate confidence limits because of the complexity of the equations, said Beecham. The methodology is from a now-classic 1999 paper by James Curran, Christopher Triggs, John Buckleton, and Bruce Weir, then all at NC State’s Statistics Department.

“The reliability of the estimated probabilities expressed by the confidence limits is important to convey in the courtroom, because there’s greater uncertainty when analyzing DNA mixtures or very small or deteriorated samples,” Weir said.

Beecham is submitting a paper on DNAMix v.3 to the Journal of Forensic Sciences. DNAMIX v.3 can be downloaded at <http://statgen.ncsu.edu/~gwbeecha/>.

Statisticians Cast New Light on Evolution

“We realized the molecular clock method of dating species divergence times didn’t work.”

Tae-Kun Seo,
BRC Post-doc

“This statistical approach offers valuable clues to the processes of evolution at work.”

Jeffrey Thorne,
Depts. of Statistics
and Genetics

A new approach for statistical analysis of genetic data is expected to improve scientists’ ability to date species divergence times and differentiate factors affecting the evolution of species ranging from *Homo sapiens* to viruses.

Developed by Dr. Tae-Kun Seo, a post-doctoral fellow at the Bioinformatics Research Center, with Dr. Jeffrey Thorne of the BRC and Dr. Hirohisa Kishino of the University of Tokyo, the method is the first that can infer how chronological rates of “synonymous” and “nonsynonymous” nucleotide substitutions in DNA change over time.

Nonsynonymous nucleotide substitutions alter the amino acid sequence of a protein and may alter its structure and function. Synonymous substitutions are “silent” and have no effect on proteins. Nonsynonymous substitutions are much more responsive to the forces of natural selection than are the “silent” synonymous substitutions.

Synonymous and nonsynonymous substitutions occur at different rates, which can change over time under the influence of different evolutionary factors. Changes in the rate of substitutions can be thought of as changes in the speed of the molecular clock.

“The conventional method of dating species divergence times using DNA was to assume a constant molecular clock, but as more genetic data became available, we realized the molecular clock method

didn’t work,” said Seo, whose doctoral research was on the evolution of HIV viruses.

Current statistical approaches for estimating rates of molecular evolution can infer the ratio of non-synonymous to synonymous substitutions. Seo’s method, which estimates synonymous and nonsynonymous rates separately, opens a new window on the dynamics of evolution.

“This statistical approach offers valuable clues to the processes of evolution at work,” said Thorne, whose earlier work was extended by Seo. Because this method can single out an acceleration in the nonsynonymous substitution rate, it can more easily reveal genes that are undergoing rapid evolution in response to forces of natural selection.

“This method of analysis allows us to evaluate how important natural selection is compared to other kinds of biological factors that affect evolution, such as mutation rates, generation length, and effective population size,” Thorne said.

Seo and his colleagues used the method to analyze the DNA segment that codes for cytochrome oxidase subunit I, a protein critical for energy production. Other scientists have shown that the protein is evolving faster in anthropoids (such as humans and gorillas) than in nonanthropoid primates. Seo’s analysis is the first to reveal that the protein is evolving more rapidly in a group of Old World monkeys than in humans and other anthropoids.

Recent Publications

- Aris-Brosou S. 2003. Least and most powerful phylogenetic tests to elucidate the origin of the seed plants in presence of conflicting signals under misspecified models. *Systematic Biology* 52(6): 781–793.
- Aris-Brosou S, and Yang Z. 2003. Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa. *Molecular Biology and Evolution* 20(12):1947–1954.
- Awadalla, P. 2003. The evolutionary genomics of recombination in pathogens. *Nature Genetics Reviews* 4(1):50–60.
- Barbash D, and Awadalla P. 2004. Functional divergence caused by ancient positive selection of a *Drosophila* hybrid incompatibility locus. Public Library of Science (in press).
- Charlesworth D, Mable BK, Schierup MH, Bartolome C, and Awadalla P. 2003. Diversity and linkage of genes in the self-incompatibility gene family in *Arabidopsis lyrata*. *Genetics* 164(4):1519–35.
- Dworkin I, Palsson A, Birdsall K, and Gibson G. 2003. Evidence that EGFR contributes to cryptic genetic variation for photoreceptor determination in natural populations of *Drosophila melanogaster*. *Current Biology* 13:1888–1893.
- Haydon D, and Awadalla P. 2004. Recombination and linkage disequilibrium in the foot and mouth disease virus. *J General Virology* (in press).
- Hill WG, and Weir BS. 2004. Moment estimation of population diversity and genetic distance from data on recessive markers. *Molecular Ecology* 13 (in press).
- Honeycutt E, and Gibson G. 2003. Use of regression methods to identify motifs that modulate germline transcription in *Drosophila melanogaster*. *Genetical Research* (in press).
- Hsieh WP, Chu TM, Wolfinger R, and Gibson G. 2003. Mixed model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics* 165:747–757.
- International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426:789–796. (Bruce Weir, member.)
- Morris RW, and Kaplan, NL. 2004. Testing for association with a case-parents design in the presence of genotyping errors. *Genetic Epidemiology* 26:142–154.
- Palsson A, and Gibson, G. 2004. Association between nucleotide variation in *Egfr* and wing shape in *Drosophila melanogaster*. *Genetics* (in press).
- Seo TK, Kishino H, and Thorne JL. 2004. Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences. *Molecular Biology and Evolution* (in press).
- Skalski GT. 2004. The diffusive spread of alleles in heterogeneous populations. *Evolution* (in press).
- Tzeng JY, Byerley W, Devlin B, Roeder K, and Wasserman L. 2003. Outlier detection and false discovery rates for whole-genome DNA matching. *J Am Statistical Assoc.* 98: 236–246.
- Tzeng JY, Devlin B, Wasserman L, and Roeder K. 2003. On the identification of disease mutations by the analysis of haplotype similarity and goodness-of-fit. *Am J Human Genetics* 72: 891–902.
- Wiegmann BM, Yeates DK, Thorne JL, and Kishino H. 2003. Time flies: a new molecular time-scale for Brachyceran fly evolution without a clock. *Systematic Biology* 52(6):745–756.

New Statistics Faculty Member Focuses on Genetic Signals of Disease

Tzeng has described new statistical measures for locating genetic differences that occur more frequently in diseased individuals than in healthy controls.

Dr. Jung-Ying Tzeng, who joined the Statistics Department in August 2003, is a statistician with a mission – to help identify genes that underlie human diseases. Her research consists of designing and testing new statistical tools that can decipher meaningful patterns in complex genetic data.

“My work is to develop methodologies that can detect an association between genetic content and disease more efficiently and precisely than current methods allow,” said Tzeng, in her office at the Bioinformatics Research Center.

The wealth of genetic data now available offers rich opportunities for correlating genetic differences with complex traits such as cardiovascular disease or psychiatric illnesses. But more advanced statistical tools are needed to unravel and interpret the masses of genomic data now being generated.

Tzeng’s current focus is on improving statistical methods for associating specific regions of the genome with a disease trait. She described new statistical measures for locating genetic differences that occur more frequently in diseased individuals

than in healthy controls in two recent publications (see page 3). Software for the statistical tests is available at her website <http://www4.stat.ncsu.edu/~tzeng>.

“I have been interested in identifying potential risk factors in human disease ever since my undergraduate major in epidemiology,” said Tzeng. After earning a master’s in biostatistics at National Taiwan University, she went on to earn a Ph.D. in statistics at Carnegie Mellon, where she participated in a genetic study of schizophrenia in kinship groups on the Pacific island of Palau. Singling out genetic patterns associated with the disease was confounded by the genetic similarity of affected individuals and the case-controls due to kinship. This problem formed the basis of her doctoral research and her continuing interest in interpreting complex genetic data sets.

Tzeng plans on collaborating with researchers in analyzing human genetic data. Her goal is to frame statistical questions “so we can reduce the ‘fuzziness’ in the data and detect meaningful signals” – signals that are likely to point to genetic risk factors for disease.

BRC
bioinformatics
is published by the
Bioinformatics Research
Center three times per year.

Director: Bruce Weir
Editor/Writer: Pat Westphal

Address correspondence and
suggestions to the editor
(pawestph@unity.ncsu.edu).

Bioinformatics Research Center
North Carolina State University
Campus Box 7566
Raleigh, NC 27695-7566

NON-PROFIT ORG.
U.S. Postage
PAID
Raleigh, NC
Permit No. 2353