**The following are abstracts from the October 2004 grantee meeting for the NHGRI/NIMH program:**
**Centers of Excellence in Genomic Science (CEGS)**

**Roger Brent, Molecular Sciences Institute** **P50 HG02370**
**Center for Genomic Experimentation and Computation** **Awarded September 2002**

The mission of the Center for Genomic Experimentation and Computation at the Molecular Sciences Institute (MSI) is to combine genomic and computational research to make predictive models of biological systems. Its flagship activity, the "Alpha Project", was established to model a prototype signal transduction pathway in the budding yeast, *S. cerevisiae.* The Alpha Project is being conducted at four sites (MSI, Caltech, MIT and PNNL):

Molecular Sciences Institute (Berkeley, CA) – Project Core, Experimental biology, Theory, and Computation
Bruck Lab (Caltech, CS) – Theory and Computation
Endy Lab (MIT, Biology & BE) – Experiment/Model Coupling and Theory
Smith Lab (PNNL, Chemistry) – Mass Spec. Development & Application

The Alpha Project is a systematic attempt to develop the experimental and computational tools necessary to predict the quantitative behavior of the pheromone pathway in individual yeast cells and in populations of cells over time and in response to defined perturbations. In the process of modeling this pathway, we are learning how to perform certain kinds of multidisciplinary work. We expect that researchers trained at this interface will nucleate laboratories at new sites that will combine experimental and computational science to understand genome function.

**(a) Aims**

The Aims of the Center are:

1. Develop means to measure pathway output and key intermediate quantities from individual cells and populations of cells.
2. Develop better computational means to simulate the behavior of individual cells and populations of cells.
3. Build quantitative models that predict the cellular response to mating pheromone and defined perturbations.
4. Combine experimental and computational work and training of scientists in a multidisciplinary genomic research environment.

## (b) Recent Key Findings

Colman-Lerner, A., Gordon, A., Serra, E., Chin, T. and Brent, R. (2004) Cell-to-cell variation in the pheromone response in yeast is dominated by epigenetic differences in the global ability to express genes into proteins. Submitted to Nature. (Please see lightning talks).

Burbulis, I., Yamaguchi, K., Gordon, A., Carlson, R. and Brent, R., (2004) Using protein-DNA chimeras to detect and count small numbers of molecules. Submitted to Nature Biotechnology. (Please see poster).

Lok, L. and Brent, R. (2004) Simulating Cellular Reaction Networks with Moleculizer 1.0. In press, Nature Biotechnology. (Please see poster).

Soergel, D., Choi, K., Thomson, T., Doane, J., George, B., Morgan-Linial, R., Brent, R. and Endy, D. (2004) MONOD, a Collaborative Tool for Manipulating Biological Knowledge. Submitted to PLOS. (Please see poster).

Benjamin, K., Thompson, T., Lok, L. and Endy, D. (2004) Constraining computational models of the yeast pheromone pathway by measuring the average number of molecules per cell of key pathway proteins. (Please see poster).

Riedel M. and Bruck J. "Cyclic Combinational Circuits: Timing Analysis and Synthesis for Delay," Thirteenth International Workshop on Logic and Synthesis (IWLS), June 2004.

Cook, M. and Bruck, J. "The Expressive Power of Relational Circuits", Technical Report, Caltech 2004.

**George M. Church, Harvard University** **P50 HG03170**
**Molecular and Genomic Imaging Center (MGIC)** **Awarded August 2004**

George Church (HMS Genetics, Biophysics, Health Sciences & Technology), James Sherley (MIT Biological Engineering), Rob Mitra (Wash U Genetics) & David Gottlieb (Wash U Anatomy & Neurobiology).

Our MGIC web page is: http://arep.med.harvard.edu/P50_03/

This Center will focus on new technologies to make "personal genomics" affordable. Genome and transcriptome resequencing will be applied to cancer and stem cell RNA measurements. The core technology is based on single-molecule polymerase colonies (nicknamed "polonies") and fluorescent base extension. A strong emphasis will be on collaborations and transfer of the technology to commercial and clinical settings.

In addition to continuation of work described in the recent articles below, we have pushed polony haplotyping to 40-fold higher densities, expanded (from a previous 2) to 7 loci per molecule and spanning up to 60 Mbp. We have developed a new method for increasing accuracy in homopolymer runs based on multiple nested primers and single base extension called "wobble sequencing". We have shown sequencing rates up to 360 kbp/min (about 200X faster than current best methods) and substitution raw error rates (i.e. 1X coverage) of 4e-5. We are developing automatable methods for in vitro libraries for polony genome sequencing and applying these to *E.coli* and human genomes. We applied for HMS IRB approval (16-Sep-2004) for a human subject protocol for non-anonymous genome-phenotype analyses to be available on the MGIC web site.

Zhang K, Reppas NB, Church GM (2004) Amplification and sequencing of genomes from single cells. (submitted).

Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet. 2004 May;5(5):335-44.

Mikkilineni V, Mitra RD, Merritt J, DiTonno JR, Church GM, Ogunnaike B, Edwards JS. Digital quantitative measurements of gene expression. Biotechnol Bioeng. 2004 86:117-24.

Buhler JD, Souvenir RM, Zhang W, Mitra RD. Design of a high-throughput assay for alternative splicing using polymerase colonies. Pac Symp Biocomput. 2004:5-16.

Zhu J, Shendure J, Mitra RD, Church GM. Single molecule profiling of alternative pre-mRNA splicing. Science. 2003 301:836-8.

Xian H, Gottlieb DI. Dividing Olig2-expressing progenitor cells derived from ES cells. Glia. 2004 47:88-101.

Adams LD, Choi L, Xian HQ, Yang A, Sauer B, Wei L, Gottlieb DI. Double lox targeting for neural cell transgenesis. Brain Res Mol Brain Res. 2003 110:220-33.

Xian HQ, McNichols E, St Clair A, Gottlieb DI. A subset of ES-cell-derived neural cells marked by gene targeting. Stem Cells. 2003;21(1):41-9.

Sherley JL. Human Embryonic Stem Cell Research: No Way Around a Scientific Bottleneck. J Biomed Biotechnol. 2004;2004(2):71-72.

Lee HS, Crane GG, Merok JR, Tunstead JR, Hatch NL, Panchalingam K, Powers MJ, Griffith LG, Sherley JL. Clonal expansion of adult rat hepatic stem cell lines by suppression of asymmetric cell kinetics (SACK). Biotechnol Bioeng. 2003 83:760-71.

Sherley JL. Embryos aren't essential to stem-cell research. Nature. 2003 423:381.

**Andrew Feinberg, Johns Hopkins University** **P50 HG03233**
**Center for the Epigenetics of Common Human Disease** **Awarded September 2004**

Andrew Feinberg, Hans Bjornsson, Patrick Onyango, M. Daniele Fallin, Karl Broman, James Potash (Johns Hopkins); Villmundur Gudnasson (Icelandic Heart Foundation); Eric Green, Matthew Portnoy (NHGRI); Webb Miller (Penn State); Oliver Ryder (San Diego Zoo); Kurt Berlin (Epigenomics); Roland Green (NimbleGen)

Epigenetics is the study of information in a cell, heritable during cell division, other than the DNA sequence itself. Epigenetics has attracted a great deal of recent attention because more of the genome that is conserved represents noncoding sequence than coding sequence, and much of this appears susceptible to epigenetic modification. Furthermore, several human disorders, including autism and cancer, appear to involve epigenetic effects. Most studies of the role of epigenetics in human disease have focused on monogenic disorders, particularly those involving imprinted genes, i.e. with parental origin-specific gene silencing. Studies of the epigenetic basis of common human disease, however, have been limited to more conventional genetic approaches, because of the absence of technical, statistical, and epidemiological resources and methods necessary for such investigations. Our long term goals are to develop tools for determining the epigenetic basis of common human disease, and to take the first steps in applying these tools to specific important illnesses. First, we plan to develop high throughput tools for epigenome analysis. This component involves development and application of methods for characterizing epigenetic marks at specific loci, including: comparative sequencing to identify critical nonexonic regulatory sequences; high throughput allele-specific gene expression; high throughput methylation analysis; and computational approaches to analysis and integration of this information. Second, we will develop the novel field of quantitative epigenetics. This includes: a novel epigenetic transmission test; quantitative epigenotype-quantitative phenotype association; and genetic linkage in which the epigenotype is treated mathematically as a quantitative phenotype to identify conventional genetic variants that influence epigenetic variation. Third, we will apply these tools to the epigenetics of common human disease, using epidemiological approaches that interface between the quantitative epigenetic tools and defined populations: a large Icelandic population that has been "digitally phenotyped" and followed for decades, including multiple blood samplings, allowing assessment of stability of epigenetic marks over time, as well as epigenetic analysis of autism and bipolar disease patients to assess disease association of specific candidate gene regions in autism and bipolar disease patients.

Our CEGS is new and our efforts in the past few months have been focused on developing the theoretical foundation of an epigenetic approach to human disease and in several proof of principle experiments. We have proposed a hypothesis termed "Common Disease Genetic and Epigenetic", or CDGE, that incorporates genetic and epigenetic variation in disease etiology, and relates this variation to the biochemistry of epigenetics. CDGE predicts common variation in the population that can be measured at several levels, including quantitative allele-specific gene expression and methylation. As a test for such variation, we modified experimental and analytical methods for dye-based single-nucleotide primer extension (SNuPE) and applied this approach to neonatal samples, reasoning that these would be unaffected by postnatal environment or age. There was considerable variation in allele-specific expression of some genes, consistent with CDGE. Interestingly, stringency of imprinting of *IGF2* was related to low birth weight, a phenotype of great public health importance.

Also, given substantial budget constraints, we have placed more emphasis on technology development as a means of partly circumventing brute force methylation analysis. We have developed a "Chromachip" on a NimbleGen array, which incorporates on the same array multiple expressed SNPs per gene, for allele-specific expression analysis, as well as a panel of epigenome modifying genes identified by a bioinformatics approach. We are also developing a ChIP on chip approach to imprinted gene identification. As an additional effort to narrow the domain of potentially methylated sites for analysis, we have begun to perform comparative sequence analysis, including armadillo, galago, bat, hedgehog, cow, dog, and rat. The methylation data obtained in these other organisms will be useful in training computational algorithms that we ultimately intend to utilize as a methylation predictor.

**Jingyue Ju, Columbia University** **P50 HG02806**
**Genomic Approaches to Neuronal Diversity and Plasticity** **Awarded September 2003**

The objectives of the proposed research are the development of new genomic technologies for massive parallel DNA sequencing and large-scale gene expression analysis from single living cells, and applications of these technologies to study neural function. We will pursue the development of three new genomic approaches: (i) *Massive Parallel DNA Sequencing Chip System* for gene discovery and expression analysis; (ii) *Nanoscopic DNA Arrays* for global gene expression profiling at the level of individual cells and subcellular compartments, and (iii) *Real-time monitoring of multiple mRNA species* in living neurons and defined cellular microdomains with high spatial resolution and fast temporal resolution. Each of these technologies will be rigorously tested and validated using the memory-forming network of *Aplysia*, a unique model organism for neurobiology. The technologies will then be implemented to explore three fundamental brain mechanisms: (1) the molecular basis of neuronal identity, (2) the molecular signals controlling the formation of the precise pattern of interconnections, which underlie behavior and, (3) the molecular basis of synapse-specific neuronal plasticity and neuronal growth. Using identified neuronal networks in *Aplysia* as experimental models we will study the role of asymmetric mRNA distribution in integrative functions and phenotypes of eukaryotic cells. We will use a hierarchical design to achieve structural resolution of single-cell profiling in a descending fashion, where a parallel genomic and functional analysis within the same memory-forming networks will be performed according to the scheme: *single neuron–single axon–single synapse*. The gene expression profiling will be correlated with functional imaging at functionally characterized neurons and synaptic terminals in a simple network during memory consolidation. The combined approach based on genomics, photochemistry, nanoscience, engineering, biochemistry, and neuroscience will be used to understand how neurons and synapses operate in the context of learning and memory. The technologies developed and the biological discoveries made in the project will have a broad impact including application to study how genes regulate cellular and organism behavior on the scale from simple nervous systems in invertebrates to the human brain.

We have made the following progress in each project in year-1.
*(1) Development of a Parallel DNA Sequencing Chip System*. The ultimate goal of this project is to develop a chip-based platform for DNA sequencing by synthesis. Four photocleavable (PC) fluorescent nucleotide analogues were synthesized, and successfully used for DNA sequencing by synthesis on a glass chip constructed by 1,3-dipolar azide-alkyne cycloaddition coupling chemistry. These results demonstrate that the PC nucleotide analogues can be accurately incorporated into a growing DNA strand during a polymerase reaction on a glass chip, and the fluorophores can be detected and then cleaved with high efficiency using near UV irradiation thereby allowing base-by-base identification of the template sequence. Nucleotides whose 3'-OH group is capped by several small chemical groups have been synthesized and some of which are shown to function as reversible terminators in polymerase reaction. We have also discovered that when using a laser at 355 nm to conduct photolysis, the fluorescent dye attached to the nucleotide can be removed completely in 10 seconds in solution. The unique coupling chemistry and the library of photocleavable fluorescent nucleotides will facilitate the development of single molecule DNA sequencing and digital gene expression analysis approaches. (2) *Development of Molecular Probes for Ultrasensitive Detection of Multiple mRNA Species on Solid Surfaces.* The research effort for this project was dedicated to the synthesis of fluorescent probes and optimization of the fluorescence detection systems. Factors that limit the sensitivity and selectivity of fluorescence-based DNA-target detector systems were analyzed and optimized. A one-dimensional gene chip prototype was tested. Highly fluorescent extremely photostable nanoparticles for signal amplification in DNA analysis were developed, and gene expression analysis in single neurons from *Aplysia* was validated. (3) *Molecular Beacons and Combinatorial Fluorescence Energy Transfer (CFET) Tags for Monitoring Multiple mRNA Species in Living Neurons and Defined Cellular Microdomains.* Evaluation of real-time imaging of mRNA in single living cells using molecular beacons was performed. A novel approach using two-photon excitation of CFET tags for multiplex genetic analysis was tested to increase the number of distinguishable fluorescence signatures by employing multi-photon absorption techniques. (4) *Genomic Basis of Neuronal Identity and Plasticity in the Simple Nervous System of Aplysia.* Single-neuron genomic protocols for gene expression analysis were developed. The characterization of sub-cellular transcriptome and real-time monitoring of dynamics of multiple mRNA species were pursued using *Aplysia* neuronal transcriptome database. We have made and tested *Aplysia* specific microarrays to identify genes specifically expressed in serotonergic and cholinergic neurons, as well as to identify genes regulated by modulatory transmitters in our model network.

**Deirdre Meldrum, University of Washington** **P50 HG02360**
**CEGS Microscale Life Sciences Center (MLSC) ("Life-on-a-chip")** **Awarded July 2001**

Deirdre Meldrum and Mary Lidstrom, Co-Directors

Increasingly, it is becoming apparent that understanding, predicting, and diagnosing disease states is confounded by the inherent heterogeneity of *in situ* cell populations. The Microscale Life Sciences Center (MLSC) at the University of Washington is focused on solving this problem, by developing microscale technology for genomic-level and multi-parameter single cell analysis, and applying that technology to fundamental problems of biology and health. The MLSC addresses two of the NHGRI Grand Challenges: under Genomics to Biology, we pursue the link between genomics and phenotype (Grand Challenge I-2), and in a logical progression, under Genomics to Health, we pursue the link between genomics, phenotype, and predictors of illness (Grand Challenge II-3).

The MLSC is developing microscale instrument modules to measure multiple parameters in living cells in real time to correlate cellular events with genomic information (e.g. gene expression and genomic rearrangements). These modules comprise a low-cost, flexible, reconfigurable, benchtop toolbox with state-of-the-art detection and analysis features to enable scientists to pursue and solve scientific questions that require analysis of heterogeneous cell populations. The microsystem modules are used for real-time quantitative assessment of expression of different genes and the resulting phenotypes as a function of environmental (and cell-to-cell) interactions. Center microsystem designs take advantage of recent developments in manufacturing processes for microscale biological microsystems, along with major advances in sensor technology, which enable the detection of minute changes in cellular properties.

The MLSC progress to date includes the development of microsystems with multiple sensors, environmental control, and cell manipulation capability. To ensure broad applicability, experiments have been performed with yeast, macrophages, T-cells, and bacteria. Current capabilities in live cells include measurement of substrate-dependent $O_2$ consumption rates and measurement of expression from multiple genes using fluorescent reporters, while in fixed cells we are developing the ability to carry out qPCR and qRT-PCR on multiple genes simultaneously and to generate single cell proteomics profiles. Other technologies under development will allow measurement of cell state, viral production, small molecules, membrane function, and enzyme activities, and the incorporation of all cell handling and analysis technologies into a single integrated bio-systems-on-chip.

As we expand technology capabilities, we are focusing this technology on a set of interconnected problems involving *in situ* cellular heterogeneity that link genomics to phenotype to health. These interrelated challenges all have the common theme of pathways to cell damage and cell death and include pro-inflammatory cell death (pyroptosis), programmed cell death (apoptosis), and avoidance of cell death (neoplastic progression). The disease states represented by this suite of problems include cancer, heart disease, and stroke.

**Maynard Olson, University of Washington** **P50 HG02351**
**Center for the Study of Natural Genetic Variation** **Awarded November 2001**

Our CEGS focuses on the development and application of methods for the long-range, haplotype-resolved, targeted resequencing of segments of previously sequenced genomes. The major applications all involve human genetic variation. "Long range" typically refers to genome segments of 50-500 kbp. The core technology is based on recombinant-DNA methods rather than PCR.

Most data have been generated using fosmid-based methods. A strength of this cloning system is that complex libraries can be constructed readily from the genomes of many individuals. The clones, which have inserts of approximately 40 kbp, are quite stable and pools of clones can be propagated with minimal changes in the relative representation of different clones. However, the modest insert sizes are a limitation since 500-kbp targets must be pieced together from a considerable number of overlapping clones.

Major progress has been made in implementing fosmid-based resequencing on a production scale. A project in the HLA Class II region, involving 20 haplotypes sequenced across approximately 80 kbp, has been completed. A *BRCA1* study involving 14 haplotypes across approximately 150 kbp is nearly complete. Finally a study in the *CFTR* region of human chromosome 7 that involves 30 haplotypes across approximately 500 kbp is at an advanced stage. Even the smallest of these projects is a broader and deeper haplotype-resolved-resequencing study than any other yet carried out in any organism. Our CEGS has both major R&D and data-production components. The scale of the R&D effort has required center-level support, particularly because of the data-tracking demands of these large resequencing projects. A significant proportion of the effort has been in development of the software tools that are required to manage these projects. In addition, the scale of the data production itself requires the resources of a vertically integrated genome center.

Analysis of the data from the *BRCA1* and *CFTR* projects will allow assessment of both methodological and substantive issues surrounding haplotype structure in the human genome. Because the data involve complete SNP acquisition they are ideal for determining the tradeoffs between SNP density and haplotype ascertainment. They will also allow rigorous assessment of the accuracy of computationally based deconvolution of genotypes into haplotypes and the sources of errors in this procedure. Substantively, they will allow a clearer view of the way that haplotypes have been shaped by the interplay between mutation, recombination, selection, and genetic drift. Related studies are underway characterizing the most variable segments of the human genome, acquired without regard to function.

Future work is focused on applications to human genetics. In particular, we are collaborating with human geneticists on mutation detection in candidate intervals, defined by linkage studies, in which standard mutation-detection methods have failed. This application has two dimensions: one involves complete detection of variants in the candidate intervals, while the other involves distinguishing between SNPs that are unlikely to affect phenotype and the recent-family-specific mutations that account for the linkage data.

**Michael Snyder, Yale University** **P50 HG02357**
**Human Genome Array Technology and Its Applications** **Awarded August 2001**

Michael Snyder, Sherman Weissman, Richard Lifton, Mark Gerstein, Perry Miller

The sequence of the human genome has been determined. The research goal of our CEGS is to develop human genomic DNA tiling arrays to elucidate the functions and properties of the encoded nucleotides.

Toward this end we have generated a genomic tiling array of human chromosome 22 and a variety of methods for its analysis. An array was constructed that contains 21,000 PCR product s encoding nearly all of the nonrepetitive regions of the chromosome. To identify transcribed regions the array was probed with human placental polyA+ RNA. Over 1/2 of the fragments that hybridized were not previously annotated, indicating that there is over twice as much RNA coding elements are previously appreciated; the functions of these novel RNAs are not known. We have also extended methods we have developed for mapping transcription factor binding sites from yeast to humans. Using this procedure termed chIP chip (for chromatin immunoprecipitation and DNA microarrays) we have mapped the binding sites for a number of transcription factors including NF-KappaB, CREB and STAT1. In collaboration with Dirk Schubeler we have used our chromosome array to determine the timing of DNA replication for a human chromosome in two cell types. We discovered many features including the observation that chromosomal regions with many transcribed segments tend to replicate early in S phase. Finally, we have been developing methods to identify mismatches in large segments of genomic DNA.

In collaboration with Viktor Stolc, we have also begun studies to analyze the entire human genome and have recently used maskless photolithography to prepare a tiling array of 46 bp resolution for the entire nonrepetitive sequence of the human genome. Transcribed sequences were mapped across the genome via hybridization to complementary DNA samples, reverse-transcribed from human liver polyA+ RNA obtained from human liver tissue. In addition to identifying many known and predicted genes, we found 10,595 novel transcribed sequences not detected by other methods. A large fraction of these are located in intergenic regions distal from previously annotated genes and exhibit significant homology to other mammalian proteins. The results of these different studies will be presented.

**William Talbot, Stanford University** **P50 HG02568**
**Genomic Basis of Vertebrate Diversity** **Awarded April 2002**

David Kingsley, Richard Myers, William Talbot

Our CEGS is using a combination of approaches to address two fundamental questions in genome biology: What do our genes do, and where did we come from? Rapid progress in genomics has provided nearly complete sequences for several vertebrates. Comparative analysis suggests many fundamental pathways and gene networks are conserved between organisms. And yet, the morphology, behavior, physiology, and disease susceptibility of different species are obviously and profoundly different. What are the mechanisms that generate new functions for genes, new physiological traits, and the unique form and functions of different species? Has the great variety of life forms been created by changes in gene number, by alterations in the functional attributes of particular proteins, or by diversification of the regulatory mechanisms that control where and when genes are expressed?

We are analyzing vertebrate diversity using a combination of techniques from structural and functional genomics and traditional genetics in stickleback fish and zebrafish. The unique experimental advantages of these two models make it possible to take complementary approaches. The “bottom-up” approach is examining the diversification in expression and genetic function of duplicated gene pairs, a major hallmark of the vertebrate genome. To define what factors determine the outcome of gene duplication, we are working to generate a database of gene expression patterns and loss-of-function phenotypes for duplicated genes in zebrafish. We have used in situ hybridization to examine the expression patterns of more than 1500 zebrafish genes, and we are working to generate loss-of-function phenotypes with a combination of antisense morpholino injections and a reverse genetic approach in which mutations are identified by a rapid screening procedure in yeast.

The complementary “top-down” approach exploits naturally occurring species of sticklebacks that show profound differences in size, anatomy, and physiological traits. Genetic crosses have been used to identify the number and location of genetic changes that create the anatomical and physiological differences between recently evolved species from different regions around the world. We are developing genetic and physical mapping resources for sticklebacks that will make it possible to identify the actual genes and mutations responsible for evolutionary change. These resources include a BAC-based physical map of the genome, a meiotic map based on SNPs and microsatellites, and collections of EST and BAC-end sequences. In addition, we have developed methods for producing transgenic sticklebacks. In recent work, we have identified genes responsible for two traits, variation in pelvic fin (hind limb) structure and armor plate number.

**Michael S. Waterman, University of Southern California** **P50 HG02790**
**Implications of Haplotype Structure in the Human Genome** **Awarded September 2003**

Michael Waterman, Norman Arnheim, Simon Tavaré

This CEGS focuses on a number of topics relevant to the search for genes associated with common complex diseases. We have five broad themes this year: Polymorphism detection, algorithms for tagSNP discovery and haplotyping, estimation of recombination rates using single sperm experiments, inference about relative rates of recombination and gene conversion using coalescent methods, and development of statistical methods for association studies.

We have had a number of successes. We developed methods to identify polymorphisms from shotgun sequencing data (including a new method for calculating Phred quality scores), and for reconstructing haplotypes from shotgun sequencing data. We extended our dynamic programming (DP) algorithms for SNP selection in a number of ways. We combined the haplotype inference algorithm with DP algorithms for haplotype block partitioning from genotype data, and developed a maximum parsimony approximation algorithm to infer haplotypes from genotypes. These and other methods are implemented in a software package. Single sperm long-range PCR methods proved feasible for estimating recombination rates. Preliminary estimates of recombination rates (relative to genome average) in 5 intervals of chromosome 21 were obtained. We developed new methods for estimating the relative rates of gene conversion and recombination from variation data.

Much of our effort has been devoted to developing methods for haplotype association analysis in case-control designs using unrelated individuals, with immediate applications to the Multiethnic Cohort Study. Our approach to selecting haplotype tagging SNPs and their analysis using an extension of the E-M algorithm for joint estimation of haplotype frequencies and relative risks received its first test on data from the multiethnic cohort study. We devised a Bayesian approach to haplotype associations using spatial smoothing and clustering methods to exploit the similarity in risks across genetically similar haplotypes. This was tested on Arabidopsis data generated in our center.

Over the next year, we will continue methodology development for tagSNP studies, the Eulerian alignment methods and optical mapping data. We will explore alternative technologies for experimental determination of recombination rates, and will continue development of spatial smoothing and clustering methods for analyzing association studies, with a particular focus on their application to dense whole-genome scans. We are also implementing efficient designs of multistage gene-association studies, in which a substudy of a larger case-control study is used for characterizing haplotype block structure and picking tag SNPs, and then only these tag SNPs are genotyped in the main study. This has direct application to the multiethnic cohort study. Finally, we are developing methods for the analysis of expression and genotyping data coming from new technologies such as Illumina, with a view to studying the correlation between SNP variation and expression variation.