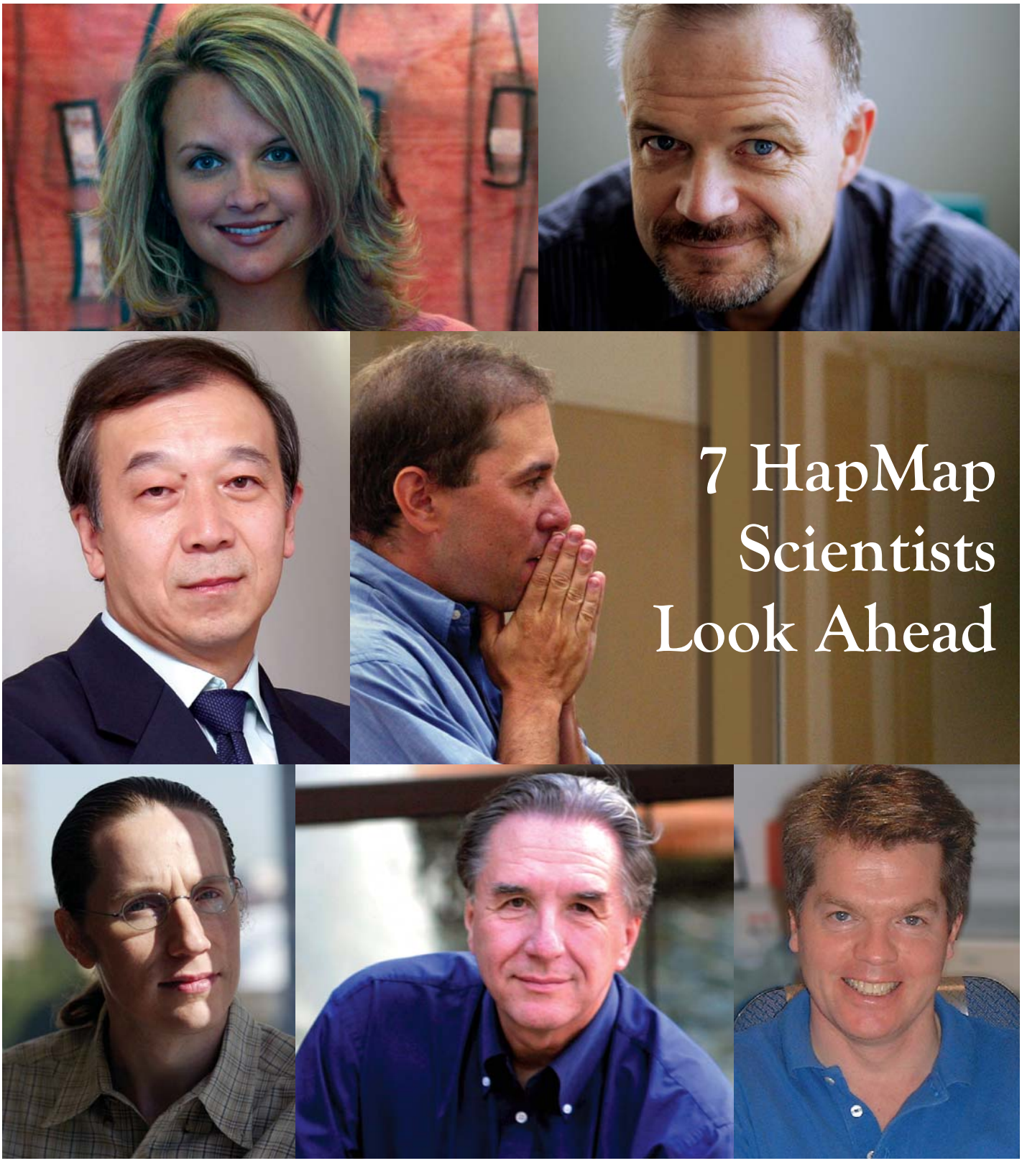
AFFYMETRIX
MICROARRAY
BULLETIN

SPECIAL HAPMAP ISSUE
VOLUME 1 ■ ISSUE 5 ■ $10.00

www.microarraybulletin.com

November 2005



7 HapMap Scientists Look Ahead

EDITOR'S FOREWARD

HapMap Project and New Technology Shaping the Future of Genetic Research and Medicine



The completion of the International HapMap Project, an international effort to identify and catalog genetic similarities and differences between human beings, marks a monumental step forward in our understanding of human genetic variation. The results provide scientists with a fundamental knowledge base to explore how genetic variation affects health, disease and individual responses to medication and environmental factors.

This HapMap Special Edition of the *Affymetrix Microarray Bulletin (AMB)* features interviews with seven scientists whose teams collectively contributed more than 90 percent of the data for Phases 1 and 2 of the HapMap Project. These researchers, from Baylor College of Medicine, the Broad Institute, Perlegen Sciences, RIKEN/University of Tokyo, and the Cambridge University/Wellcome Trust Sanger Institute, discuss their roles in the project, the lessons learned and how they are applying the information to new studies.

Like the Human Genome Project that preceded it, the HapMap project demonstrates the power and potential of harnessing the world's scientific resources in a cooperative effort. Scientists from Japan, the United Kingdom, Canada, China, Nigeria, and the United States worked together across borders and cultures to make this project a success. The project similarly demonstrates the key role that technology plays in helping scientists penetrate the complexity of the genome.

Just as advances in sequencing technology accelerated progress on the Human Genome Project, high-density microarrays and new multiplex assays enabled the HapMap project to reach completion in remarkably little time and for a relatively low cost. This technology evolution is unprecedented in its speed and scale. When the HapMap Project began in October of 2002, it would take one scientist months to genotype 1,000 to 2,000 single nucleotide polymorphisms (SNPs). Now, three years later, using only two microarrays, any researcher can look at 500,000 SNPs that cover the entire genome in just two days.

The seven researchers we interviewed are applying this technology to the natural next step—large-scale association studies. The HapMap Consortium predicted that by documenting the structure of the genome, scientists would be able to successfully design comprehensive association studies. As you will read in this issue, preliminary work by these scientists and others around the world is validating this prediction. Dozens of new association studies are underway, enabling researchers to study genetic variation in common diseases such as cancer, stroke, and heart disease and also in therapeutic response.

For Affymetrix scientists, the successful completion of the HapMap is the gratifying conclusion of a long-standing commitment. Affymetrix Founder Steve Fodor committed the company to large-scale haplotype mapping research in 1997, after a visit by Serono's Daniel Cohen, who suggested that GeneChip® microarray technology would be perfect for creating a haplotype map.

David Cox, a Stanford genetics professor, joined the effort, developing the experimental design necessary to construct the first-generation haplotype map by scanning 50 human genomes using whole–wafer GeneChip microarrays, with up to 60 million probes each. In 2001, Fodor and Cox spun off a company called Perlegen and brought in Brad Margus to act as CEO to help complete the project and apply the resulting haplotype data to pharmacogenomic research.

In two hallmark publications in 2001 and 2005, the Perlegen scientists used Affymetrix microarray technology to establish a first genome-wide haplotype map of the human genome. They found surprising genetic similarity between people by sequencing 21,676,868 bases of Chromosome 21 from 20 individuals; they also captured common genetic differences that likely underlie disease predisposition and variable drug response by genotyping 1,586,383 SNPs from 71 people. In September 2004, Perlegen began working with the International HapMap Consortium on Phase 2 of the project, and has contributed over 4 million SNPs to the HapMap effort.

The combination of all of the HapMap participants—dozens of public and private organizations and hundreds of individual scientists—has overwhelmingly exceeded the goals of the original HapMap project. Their efforts have given the global scientific community the tools to study the link between health and genetic variation. These tools are shaping a new model of medicine: one that looks not only at a patient's medical history, but also at his or her genetic history, to diagnose and prescribe treatment.

We will follow the progress of work emanating from the HapMap project in future issues of the *AMB*; we will also continue to feature discussions on microarray research with leading scientists in every field. The goal of our publication is to give members of the microarray research and diagnostic communities a forum where they can share their methods, tools and insights.

In our premier April issue, Daniel Cohen discussed conducting an association study he'd been planning for 20 years—and finding 80 genes potentially involved in multiple sclerosis. In our July Cancer Issue,

Johns Hopkins' Anirban Maitra talked about finding mutations linked to cancer in the mitochondrial genome. In our most recent October issue, Dana-Farber's Myles Brown discussed ChIP-on-chip as an unbiased method of identifying estrogen receptor binding sites. There are now more than 30 *AMB* stories on our website (**www.microarraybulletin.com**) on everything from sequence analysis to translational medicine and bioinformatics. We welcome your thoughts and suggestions and offer our congratulations to all of the International HapMap Project participants.



Sincerely,
Rachel Shreter
Editor, AMB

AFFYMETRIX
MICROARRAY BULLETIN

Wes Conard, *Editor-in-Chief*
wes_conard@affymetrix.com
Tommy Broudy, *Managing Editor*
thomas_broudy@affymetrix.com
Rachel Shreter, *Editor*
rachel_shreter@affymetrix.com
Kamalia Dam, Associate Editor
Stacey Ryder, Associate Editor
Lisa Osborne, Copy Editor
Reprints email:
reprints@microarraybulletin.com
www.microarraybulletin.com

Affymetrix Inc.
3420 Central Expressway
Santa Clara, CA 95051
408.731.5000

Penelope E. Bonnen began her scientific career with a W.M. Keck Computational Biology fellowship at Baylor College of Medicine in Houston, Texas. Working in Dr. David Nelson's laboratory, she studied SNP haplotype diversity in cancer susceptibility genes.



As a postdoctoral fellow, Bonnen is continuing to pursue her interest in genome variation and its role in susceptibility to disease in Dr. Jeffrey Friedman's laboratory at The Rockefeller University, New York. Bonnen, a Rockefeller University Women & Science fellow, used microarray technology to develop a HapMap for the genome of an isolated population in the South Pacific (Kosrae, F.S.M.) and is currently utilizing this HapMap to pursue mapping genes that contribute to Metabolic Syndrome.

Further Reading

■ Bonnen, PE, Pe'er, I, Plenge, R, Salit, J, Breslow, J, Stoffel, M, Daly, MJ, Jones, K, Altshuler, D, and Friedman, J. Haplotype map for an isolated population in Micronesia: evaluating potential for whole-genome studies (in preparation).

■ Bonnen, PE and Nelson, DL. 2004. SNPs and Functional Polymorphisms in Cancer, Oncogenomics: Molecular Approaches to Cancer. Edited by Charles Brenner and David J. Duggan, Wiley Inc.



Dr. David Craig is an associate investigator and faculty member within the Neurogenomics Program at the Translational Genomics Research Institute (TGen) in Phoenix, Ariz. His lab is working to develop cost-effective design strategies and accompanying analysis tools for high-density, genome-wide SNP genotyping scans. These include design of multi-staged SNP association studies, such as pooling-based SNP microarray studies.

Craig is applying these technologies to his research in autism, bipolar disorder and multiple sclerosis.

Craig completed his post-doctoral training at TGen and received his Ph.D. from the University of Washington.

Further Reading

■ Craig DW, Stephan DA. Applications of whole-genome high-density SNP genotyping. *Expert Rev Mol Diagn.* 2005, Mar;5(2):159-70. Review.

■ Strauss KA, Puffenberger EG, Craig DW, Panganiban CB, Lee AM, Hu-Lince D, Stephan DA, Morton DH. Genome-wide SNP arrays as a diagnostic tool: Clinical description, genetic mapping, and molecular characterization of Salla disease in an Old Order Mennonite population. *Am J Med Genet A.* 2005, Sep 12 [Epub ahead of print].

IN THIS ISSUE

The Broad Institute's David Altshuler, Stacey Gabriel and Mark Daly discuss the lessons learned from the HapMap Project, high-throughput genotyping and the data analysis infrastructure required to search for genes influencing common diseases p.4

Perlegen's David Cox discusses genotyping 4.6 million SNPs in Phase 2 of the HapMap Project using array-based technologies p.9

Baylor's Richard Gibbs talks about multiplexed genotyping assays used for the HapMap Project and performing other large-scale epilepsy and cancer genome studies p.12

RIKEN's Yusuke Nakamura discusses plans to genotype more than 10,000 individuals to study 47 common diseases p.16

Cambridge Medical School's John Todd discusses controls, samples, and collaborations needed to find new associations between gene variants and disease p.19

AMB Featured Research: Association Studies

- Genetics of graft versus host disease investigated with 500,000 SNP study p. 10
- Rockefeller scientists make progress in genetic study of obesity; create haplotype map for Kosrae's population p. 12
- 100K SNP scan helps Finnish scientists uncover 700 new associations to cardiovascular disease p. 14
- Finding a common variant associated with increased risk to macular degeneration p. 16
- Discovering the genetic variations behind multiple sclerosis p. 18
- Researchers combine linkage analysis and whole-genome association studies to investigate autism and Alzheimer's disease p. 22

Broad Institute Opens HapMap Genotyping and Data Analysis Infrastructure to Researchers for Association Studies

The Broad Institute of Harvard and MIT's David Altshuler, Stacey Gabriel and Mark Daly discuss the lessons learned from the HapMap Project, high-throughput genotyping, and the data analysis infrastructure required to search for genes influencing common diseases

Researchers at the Broad Institute, who generated millions of genotypes for the International HapMap Project, believe the map—and the technology and analysis methods that made it possible—have opened the door for individual researchers to finally undertake comprehensive association studies. To that end, they're making the high-throughput infrastructure and data analysis capabilities they developed available to other researchers and are working on their own association studies in areas like type 2 diabetes, bipolar disorder and cardiovascular disease.

"The argument can always be made to wait for the next technology, or wait for it to be cheaper," said Gabriel, director of the Genetic Analysis Platform. "But, if you're convinced by the data and tools that well-powered studies are for the first time possible—and you have a large enough and worthwhile collection of well-phenotyped patients—I do think now is the time to begin these studies."

In July, the team announced that it would undertake genome-wide association studies of diseases such as cancer and metabolic and neuropsychiatric disorders using new 500K single nucleotide polymorphism (SNP) arrays from Affymetrix. Researchers will genotype more than 12,000 DNA samples in the attempt to identify the genetic variations underlying disease. Additionally, the group is collaborating with several large disease consortiums that are pooling patient samples across different labs and genotyping them at the Broad Institute as part of a National Institutes of Health (NIH)-funded, National Center for Research Resources (NCRR) genotyping center directed by Gabriel.

"The possibility of making important discoveries through comprehensive genotyping is tremendously exciting," said Altshuler, director of Medical and Population Genetics and a founding member at the Broad. "But the potential to get into trouble by overanalyzing the data or overstating the significance of a finding is a real danger. It's going to take great discipline to interpret these data. This is true with microarray data and it's true in any large data test that generates billions of data points."

"All of the data collected and evaluated to date from the HapMap Project suggests that a selection of polymorphisms at high density affords the opportunity to take a very reasonable look at the majority of the common genetic variation in the genome," said Daly, statistical geneticist and assistant professor at the Massachusetts General Hospital and Harvard Medical School. "This year is the first point in time where we can design an efficient and cost-effective study of a large fraction of the common variation of the genome."

Altshuler, Gabriel and Daly spoke with Dr. David Craig, associate investigator at the Translational Genomics Research Institute (TGen) about the following topics:

- Defining haplotype structure and genetic variation across the genome
- Designing infrastructure for sample, genotyping and informatics management

- Developing local data analysis infrastructures to analyze unprecedented volumes of data from large-scale association studies
- Integrating and sharing data to more quickly discover the genetic underpinnings of disease

Defining Haplotype Structure and Genetic Variation Across the Genome

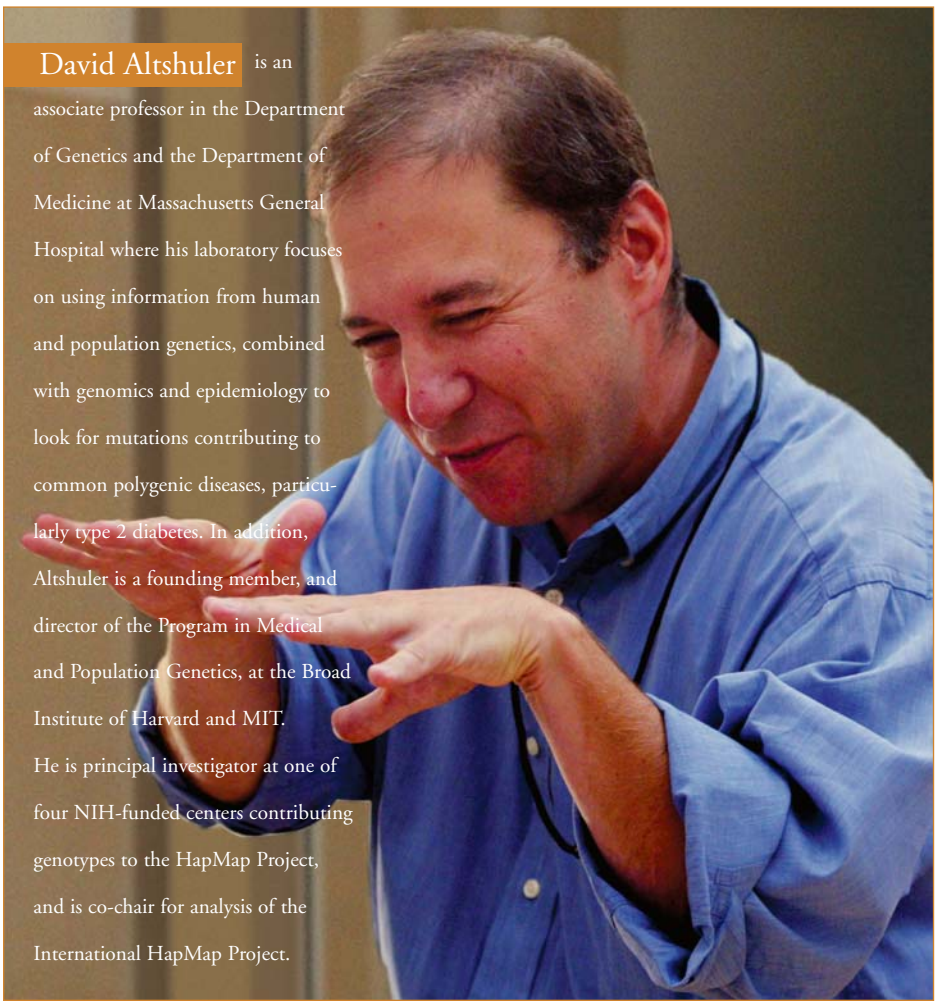
Craig: The HapMap Project is nearing its first key publication following a massive multi-center effort to genotype millions of SNPs across hundreds of individuals. Are there any big surprises that came out right away?

Altshuler: What's been very gratifying in analyzing the HapMap data, as well as the data published by Perlegen in *Science* earlier this year, is how they unambiguously show that the patterns previously suggested by earlier small-scale, less complete studies are, in fact, general across the genome. They show that that long haplotypes and limited haplotype diversity are characteristics of the entire human genome. While there was a degree of confidence that this would be the case, it's very gratifying that it came out so unambiguously in support of those models and of the utility.

Craig: I'm curious about your thoughts on the common-variant versus multiple-variant models. Does this dictate which diseases you would avoid or target for association studies?

Altshuler: The question is: what are the opportunities to make progress? Common variants—because they are finite in number—can be catalogued and systematically tested with new technologies. Whereas finding rare variants requires, by definition, dedicated resequencing of each sample, which is not yet practical for large numbers of samples on a genome-wide scale.

The genetic variant in complement factor H gene that was discovered recently in age-related macular degeneration [by Josephine Hoh at Yale and collaborators] is perhaps the most striking example of a disease influenced by a common variant (Klein et al, *Science* 2005), and it was found by [linkage disequilibrium] LD-mapping in genetic-association studies. It's worth remembering that it was very far from clear a year ago, whether age-related macular degeneration was a common-variant disease or rare-variant disease. HapMap-type approaches were useful in getting there, but that doesn't mean there aren't rare variants as well. But, it underscores the fact that the field needs to move forward by collecting data, seeing what's where and then refining it, and that successes will occur.



David Altshuler is an associate professor in the Department of Genetics and the Department of Medicine at Massachusetts General Hospital where his laboratory focuses on using information from human and population genetics, combined with genomics and epidemiology to look for mutations contributing to common polygenic diseases, particularly type 2 diabetes. In addition, Altshuler is a founding member, and director of the Program in Medical and Population Genetics, at the Broad Institute of Harvard and MIT. He is principal investigator at one of four NIH-funded centers contributing genotypes to the HapMap Project, and is co-chair for analysis of the International HapMap Project.

Building Infrastructure for Large Genetic Studies

Craig: Stacy, you were recently funded by NCRR to build an infrastructure allowing the genetics community to access genotyping technology without requiring each lab to build their own. What have you built, and how is it being made available to the outside community?

Gabriel: We're leveraging the system that we were building already for our internal studies at Broad in such a way that external users can access that same infrastructure. The front-end will be a Web interface that researchers from the outside can use. It will let researchers use database tools to manage phenotypes, samples, check the status of their studies, send us samples, and view and download results. On the lab side, we have three SNP genotyping technologies that cover the range of different studies from whole genome to following up a specific region. Eventually, we'll move all the way into resequencing as well.

Craig: Are you ready to handle the informatics of all this if three or four groups ask you to run a few thousand Affymetrix 500K arrays and generate a few terabytes of data?

Gabriel: Yes, in fact, that's happened already. We are working with several large, external disease consortiums that have come together to pool patient samples across different labs and they send them to us to do the genotyping.

Craig: How do you see clinical groups working together with large genotyping centers like yours?

Gabriel: To have statistical power for a lot of these studies, people may need to combine sample collections. Groups which have collected samples from 500 to 1,000

patients in diseases like diabetes, lupus and multiple sclerosis are finding advantages to banding together with other groups with similar numbers of patients, achieving the power required to identify the needle of a true disease gene in the very large haystack of the genome.

And because there is no one single technology platform that can enable genotyping for the whole range of genetic studies, substantial capital investment is required. In addition to lab hardware, considerable investment in equipment, informatics infrastructure, and software engineers who can build the kind of laboratory information management systems (LIMS) to link together these technologies is also required. In that way, it's made sense for us to try to combine these all into one center.

Craig: You mentioned different strengths and weaknesses of different platforms. Is there an area where you wished the technology was a little bit further on?

Gabriel: The technology has come a huge way, but it's certainly true that, in some ways, it's now much less efficient to genotype one SNP than 500,000—but sometimes that's what an investigator needs to do. There's room for improvement in all of these technologies in terms of automation, ease of sample handling, consistency across different sample sets, and robust behavior with samples of varying quality, but definitely in the lower multiplex area where we follow up results of these whole-genome scans.

Craig: Once you finish a genome-wide association study with 500,000 SNPs, do you anticipate following up on candidate genes or regions in the second validation phase with tens, hundreds, or thousands of SNPs?

Gabriel: That's going to depend on the size of the available sample cohort. Certainly, depending on the magnitude of the initial effect you observed, you might want to genotype hundreds to thousands of SNPs to follow up initial association results.



Stacey Gabriel is director of the Genetic Analysis platform and director of the NCRR-funded National Center for Genotyping and Analysis at the Broad Institute where she oversees all genotyping, SNP discovery and production activities related to human genetics. Dr. Gabriel is also the scientific director of the Broad Institute's contribution to the International HapMap Project and a member of the HapMap steering committee. She has contributed to the design of protocols for genotyping, methods of quality assessment and quality control, data sharing and transfer protocols, human studies approvals and all production genotyping for the HapMap Project.

Altshuler: We think it is unlikely that initial scans will allow us to distinguish true positives from noise—our design (as is typical in the field) will be to analyze 1,000 to 3,000 patients as a genome-wide screen, and then to follow up the best hits in subsequent and larger populations. In our type 2 diabetes scan (performed in collaboration with Leif Groop in Sweden, and the Novartis Institute for Biomedical Research), we will be screening 3,000 patients, and then following up the top 1,500 SNPs (perhaps 0.3 percent) in another set of 4,000 samples. The power is limiting, and it is unrealistic to hope for hugely significant results in smallish samples although we wouldn't mind being lucky!

Developing a Data Analysis Infrastructure for Association Studies

Craig: With thousands of samples and 500K SNP arrays, you're generating tons of data. What worries you more from an informatics perspective, infrastructure or analysis?

Daly: I think the infrastructure may be the more pressing need. People have for a long time been thinking about how to analyze genetic association data, and how to deal with the massive set of results as a statistical entity rather than the traditional way of just studying association, gene by gene. The infrastructure, on the other hand, is more challenging because of the technology-specific components. Two gigabytes of data times thousands of samples requires a considerable amount of attention on figuring out exactly how to handle, store, and pass around all that information. Fortunately, I think there's a great wealth of experience in the Human Genome and the HapMap Projects, as well as other projects. We're not starting from ground-zero, we're actually starting from a position of considerable experience.

The challenge is developing a system that seamlessly processes and stores the data while facilitating creative and complex analyses—it's not simply a matter of extracting genotypes and passing them over to genetic analysis. We need to incorporate quality metrics on those genotypes and enable additional analyses of the raw hybridization data. This is already popular in the cancer field, with respect to [loss of heterozygosity/comparative genomic hybridization] LOH/CGH analysis, but



Mark Daly is assistant professor of medicine at Harvard Medical School at the Center for Human Genetic Research at Massachusetts General Hospital. Over the past 18 years, Dr. Daly has contributed to statistical genetics and understanding patterns of variation in the human genome with the goal of creating more effective approaches for detecting the genetic variation responsible for disease. He developed the GENEHUNTER software for performing statistical analyses to identify genomic regions containing disease risk factors in families, and his group now develops and distributes Haploview as a tool for LD and association analysis. He is a co-leader of the Broad's contribution to the HapMap Project, and a member of the Steering Committee and Analysis Group for the International project.

will be brought more broadly into large-scale studies of almost every disease.

Altshuler: While the analytical approaches have been widely discussed, there remain issues of over-enthusiastic interpretation. It's going to take great discipline to actually interpret these data, and not be overly impressed by a p value of 0.0001 in a dataset with 1,000,000 hypotheses. This is true with microarray data, and it's true in any large data test that generates billions of data points.

Craig: Mark, you spent a lot of time developing analysis tools and packages which have gone into hundreds of labs. Do you think individual labs will require a specialized statistical team that develops their own custom tools or do you see a need for widely distributed software programs that are accessible to the broader statistical genetics community?

Daly: It is going to be difficult to do these types of projects early on without local statistical expertise. As David said, the concern is applying improper statistical practice to the results—people will be overwhelmed with exciting p values, when some of these are really nothing more than the chance occurrences expected in running hundreds of thousands of trials. The real challenge in this research is appropriate judgment and evaluation of these data in order to drive subsequent research. Off-the-shelf packages will be limited in their ability to incorporate investigator hypotheses and resources in judging which regions of the genome to follow up with additional genotyping, samples, or which secondary analytic methods may be appropriate. In addition, statistical significance of individual results after these secondary analyses and experiments is complicated by the more opaque process by which they were generated.

Altshuler: There will be a lot of other information that's not in and of itself definitive. And you can make it more definitive by doing further genetic studies with larger samples, more sequencing, more information, or integrating the data with other kinds of data—mouse models, expression data, RNAi perturbations and so forth. It will certainly be a huge advantage to be in a collaborative environment.

The ability to share data in digitized form obviously makes things possible that couldn't be done if all you've done is write in your notebook, and walked over to the other person to talk about what you got. Potentially there are things that can be done to share your entire data set and integrate with others. That's fundamentally a computational problem and an analytical problem.

Integrating and Sharing Different Types of Molecular Data in Large-scale Genetic Studies

Craig: Are you concerned about being able to integrate and analyze all these data types at once?

Daly: We now have technology that's enabling us to study how genetic variation affects disease in a much more comprehensive fashion; not just SNP variation, but structural variation, what regions of the genome might be under selection, and expression studies in particular disease areas and tissue types. All of this information needs to be looked at in a rigorous and comprehensive fashion, so there's a great opportunity for the development of computational tools and infrastructure to perform these studies.

Altshuler: With RNAi, for the first time it's possible in cell lines and in different models to manipulate each gene, in turn, and measure its phenotype in a relevant setting. This enables researchers to annotate each gene with regard to its phenotype when perturbed in humans—in model systems like mice, worms, and flies—and using that information to make networks of relatedness among genes.

There isn't any formula that takes a SNP association value, an expression trait and a knockout phenotype and says you should be very interested. What is clear is that if you can find out that a particular gene is involved in the trait of interest, then knowing what other genes are involved in the same process, the same pathway, or the same multi-protein system can lead you very rapidly to the other genes that are involved.

Bardet-Biedel syndrome is a spectacular example of this. It's a Mendelian disease, and it took many years to find the first gene. It turns out that this gene is involved in a subcellular structure (the cilia). Finding one gene, and knowing the genes involved in that structure, led very rapidly to identifying other mutations causing Bardet-Biedel syndrome.

Craig: Many of the technologies that were used to genotype the SNPs for the

HapMap Project can also assess copy number variations or structural variations. There have been a number of exciting papers exploring structural variations found in normal individuals. What is the importance of being able to analyze both SNPs and copy number variations with, fundamentally, the same data set?

Gabriel: The field has come to recognize that we need an integrated picture of SNPs, haplotypes and structural variation in the same samples. We have a close collaboration with Matthew Meyerson and Bill Sellers from Dana Farber who have done really exciting work in this area. We are using Affymetrix arrays to detect copy number changes in cancer samples using tumor DNA. Something that should be hugely advantageous to this area of research, is having a bank of known information from HapMap samples, for example, to tell you there's a long stretch of homozygous in all your cancer samples. Often in cancer samples, you don't necessarily have a matched normal sample for comparison. So comparing your result to information from the HapMap samples, or other sources, is really useful. This information also gives you technical comparisons for expected chip intensity across these regions.

Altshuler: We're undergoing a shift from an information-poor analytical approach, where you analyze your data as if it's the first such data ever collected in the world. You design a linkage or association study, collect data, then run a T-test, and look at your [logarithmic odds] LOD score or chi square test.

We're moving to a world where you get results from a genome-wide association or a chromosome copy number study, and you interpret them in light of everything you know about the genomic region. This includes functional annotation, other variants in the population you're studying, their frequencies and other information. The genome is a scaffold on which all information can be hung. You can reference back different kinds of data to their location in the genome. This wasn't possible five years ago.

FOR MORE INFORMATION

Contacts

- David Altshuler, Ph.D.
Department of Molecular Biology
Center for Human Genetic Research
Massachusetts General Hospital
Richard B. Simches Research Center
185 Cambridge St., Room 6806
Boston, MA 02114
altshuler@molbio.mgh.harvard.edu

- Stacey Gabriel, Ph.D.
Director, Genetic Analysis Platform
Director, National Center for Genotyping and Analysis
Broad Institute of Harvard and MIT
320 Charles Street
Cambridge, MA 02141
stacey@broad.mit.edu

- Mark Daly, Ph.D.
Assistant Professor of Medicine
Center for Human Genetic Research
Massachusetts General Hospital
Richard B. Simches Research Center
185 Cambridge St., Room 6806
Boston, MA 02114
mjdaly@chgr.mgh.harvard.edu

Companies

- Affymetrix Inc. – http://www.affymetrix.com
- Perlegen Sciences – http://www.perlegen.com
- Novartis – http://novartis.com

Organizations

- The International HapMap Project – http://www.hapmap.org
- Broad Institute of Harvard and MIT – http://www.broad.mit.edu
- Massachusetts General Hospital – http://www.mgh.harvard.edu
- Harvard Medical School – http://hms.harvard.edu/hms/home.asp
- National Center for Research Resources (NCRR) - http://www.ncrr.nih.gov/
- Dana-Farber Cancer Institute – http://www.dana-farber.org

Further Reading

- Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P, Altshuler D. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*. 2005;308(5718):107-11.
- Daly MJ, Altshuler D. Partners in crime. *Nat Genet*. 2005;37(4):337-8.
- Florez JC, Burtt N, de Bakker PI, Almgren P, Tuomi T, Holmkvist J, Gaudet D, Hudson TJ, Schaffner SF, Daly MJ, Hirschhorn JN, Groop L, Altshuler D. Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes*. 2004 May;53(5):1360-8.
- Florez JC, Hirschhorn J, Altshuler D. The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu Rev Genomics Hum Genet*. 2003;4:257-91. Review.
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. *Science*. 2002 Jun 21;296(5576):2225-9. Epub 2002 May 23.
- Altshuler D, Clark AG. Genetics. Harvesting medical information from the human family tree. *Science*. 2005;307(5712):1052-3.
- Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*. 2005 Feb;6(2):95-108. Review.
- Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005 Jan 15;21(2):263-5.
- Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. Assessing the impact of population stratification on genetic association studies. *Nat Genet*. 2004;36(4):388-93.
- Gabriel SB, Salomon R, Pelet A, Angrist M, Amiel J, Fornage M, Attie-Bitach T, Olson JM, Hofstra R, Buys C, Steffann J, Munnich A, Lyonnet S, Chakravarti A. Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nat Genet*. 2002;31(1):89-93.
- Gabriel et al. The structure of haplotype blocks in the human genome. *Science*. 2002 Jun 21;296(5576)2225-9.
- Klein et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005 Apr 15;308(5720)385-9.

People

- Matthew Meyerson, M.O., Ph.D.
Dana-Farber Cancer Institute
http://research.dfci.harvard.edu/meyersonlab

- William R. Sellers, M.D.
Dana-Farber Cancer Institute
http://research.dfci.harvard.edu/sellerslab

Software

- GENEHUNTER- ttp://www.broad.mit.edu/ftp/distribution/software/
- HAPLOVIEW – http://www.broad.mit.edu/mpg/haploview

Perlegen Scientists Genotype 4.6 million SNPs in Phase 2 of the HapMap Project Using Array-based Technologies

Perlegen's David Cox discusses genotyping Phase 2 of the HapMap project and implications for association studies and patient care

This year Perlegen genotyped 4.6 million SNPs in the second phase of the HapMap, almost twice as many as expected; the overwhelming success of the project shows the tremendous potential of leveraging high-density microarrays and large-scale collaborations, said Chief Scientific Officer David Cox.

"The phase 2 grant allowed us to put into the public databases an even more dense SNP resource than Perlegen published on its own," said Cox. "It's a perfect example of how this public and private collaboration was able to produce a much better product than either could have alone."

Perlegen used the same Affymetrix whole-wafer technology that had yielded 1.6 million SNPs in its initial 2002 haplotype study. That study was based on a set of 71 Americans of European, African, and Chinese ancestry (Hinds, *Science* 2005), and the data from that study was released to the public in 2005. Now, Perlegen is using that data and the HapMap to help scientists explain and predict the effects of prescription drugs in clinical trials.

Cox spoke with Dr. David Craig from the Translational Genomics Research Institute (TGen) about the current and future prospects of whole-genome associations studies using the data generated from the HapMap. The two discussed:

- The way HapMap data confirmed hypotheses underlying the project
- Pooling strategies and association study design
- The future role of genetics in patient care

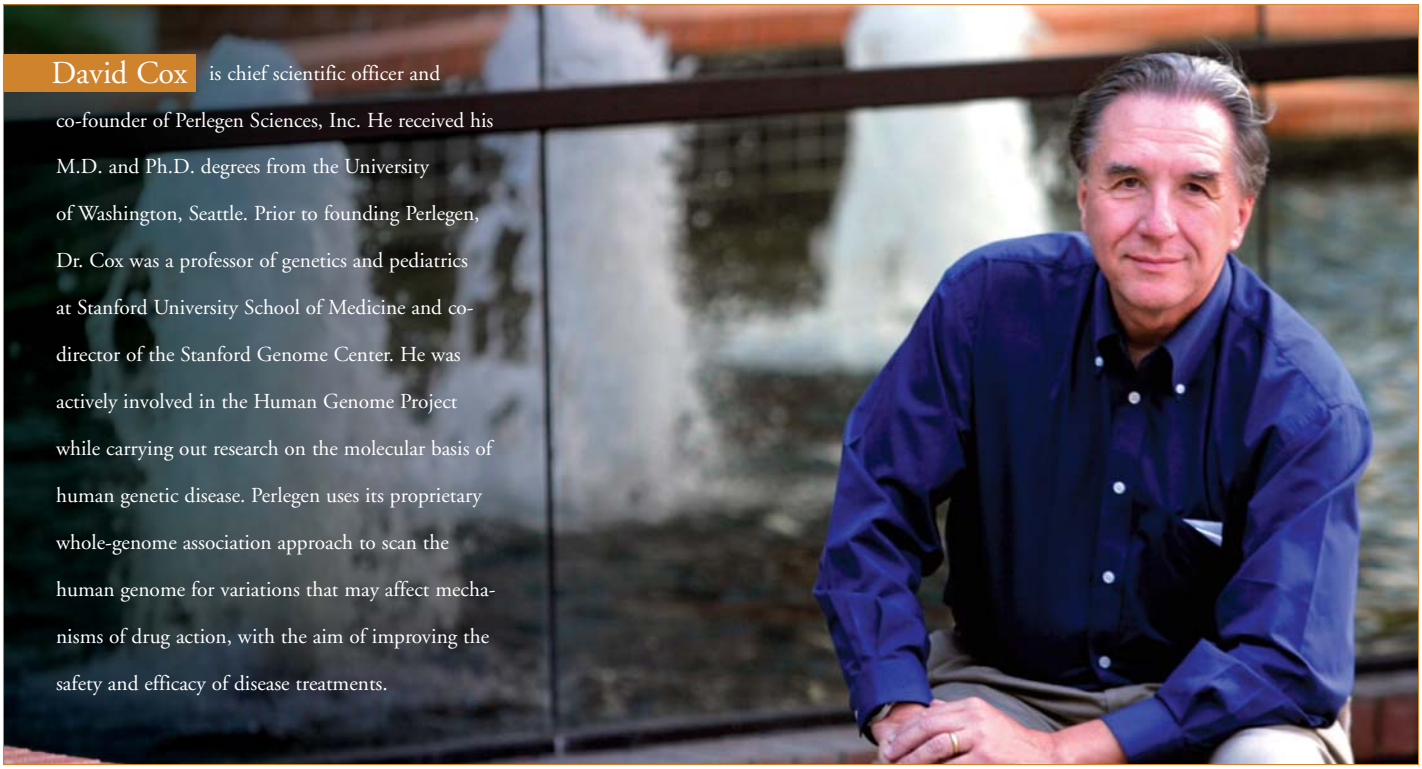
Genome Structure Confirmed by HapMap

Craig: The HapMap is nearing one of its first major publications. I am wondering what you think the biggest surprise is thus far? Perhaps, something that maybe you didn't expect a year or two ago?

Cox: The best news is that there are not very many surprises. The whole basis for beginning a human haplotype map, both originally at Perlegen and in the International HapMap Project, was the belief that a common set of variants would give you useful information about people from different geographic or ethnic origins. I think that is turning out to be the case.

Some SNPs are unique to a population, but it remains to be determined how important they will be as the basis for differences in disease or drug response compared to SNPs that are shared across populations. The HapMap allows us to actually test common sets of SNPs instead of resequencing everybody's genome each time we want to understand differences. It would be great to do complete resequencing, but right now that's still cost prohibitive.

The second important assumption of the HapMap project was that due to the correlation structure, it would be possible to select subsets of SNPs that would give much of the information from all common SNPs. I think that the data in



David Cox is chief scientific officer and co-founder of Perlegen Sciences, Inc. He received his M.D. and Ph.D. degrees from the University of Washington, Seattle. Prior to founding Perlegen, Dr. Cox was a professor of genetics and pediatrics at Stanford University School of Medicine and co-director of the Stanford Genome Center. He was actively involved in the Human Genome Project while carrying out research on the molecular basis of human genetic disease. Perlegen uses its proprietary whole-genome association approach to scan the human genome for variations that may affect mechanisms of drug action, with the aim of improving the safety and efficacy of disease treatments.

AMB Featured Research: Association Studies

Genetics of Graft Versus Host Disease Investigated in 500,000 SNP Study

Seishi Ogawa of the Graduate School of Medicine at the University of Tokyo has started the most comprehensive genetic study of graft versus host disease (GVHD) ever undertaken. His research team is in the process of genotyping 500,000 SNPs from each participant in a 2,000-patient, case-control association study.



Stage 1 of the project uses Affymetrix 500K microarrays to genotype 1,000 donor/recipients in GVHD-positive transplants, and 1,000 cases without GVHD. In Stage 2, Ogawa's team will perform a genome-wide association analysis to pinpoint the genetic factors—including minor antigens—responsible for disease.

Ogawa told the *AMB* in April that he designed his study to include GVHD patients diagnosed with different diseases so that he could use the same genotype data to uncover genetic predispositions to diseases such as leukemia, lymphoma, and aplastic anemia.

Further Information:

- GVHD Experts Dr. Seishi Ogawa and Dr. Takanori Teshima Discuss the Challenges of Whole-genome Association Studies (*Affymetrix Microarray Bulletin*, April 2005) – http://www.microarraybulletin.com/ogawa
- Power estimation of genome-wide association analysis of minor histocompatibility antigens relevant to pathogenesis of GVHD (Presentation, Human Genome Meeting 2005) - http://hgm2005.hgu.mrc.ac.uk/Abstracts/Publish/WorkshopOrals/Workshop02/hgm010.html

Perlegen's publication in *Science* in February 2005, and now denser data from Phase 2 of the HapMap Project, suggest that is the case. With a relatively small subset of all of the common human SNPs, you can encompass much of the information content of the complete set of SNPs; that's good news because otherwise studies would be cost prohibitive.

Craig: One goal of the HapMap was to identify a minimal number of SNPs, or "tag SNPs" that when genotyped, can sufficiently characterize an individual given the overall genetic heterogeneity of the population. There has been a lot of debate over the minimum number of tag SNPs you need to retrieve the majority of information in the genome. How many SNPs do you need to cover the underlying genomic structure in most populations?

Cox: Covering the entire genome is a relative issue. The SNPs studied in HapMap and studied by Perlegen now are common SNPs, these are variants with a minor allele frequency of five percent or greater, which means that they're relatively common in the population. There are many, many more SNPs that are very rare. So when people are talking about coverage of the genome today, they are really talking about coverage of the common SNPs. It remains to be seen how important rarer SNPs are in common human disease. It's known that in certain situations rare SNPs can absolutely impact disease.

To really know your chances of finding the variants that are important in disease, you also need to understand the genetic architecture of the disease. How many different genes are involved? How big is their effect? The more SNPs you look at, the more information you can get about other SNPs that are missing from your set. Without typing all of the SNPs, you are never going to have all of the information. But the good news is that by typing 300,000 to 500,000 selected SNPs, you can get in the range of 80 percent of the genetic information encompassed by common SNPs. If you looked at more SNPs would you get more information? What coverage is sufficient? Until we as a community do more of these studies and better understand the genetic architecture, no one really knows for sure.

In general, for complex disease, it's estimated that 20 to 50 independent genetic components add together to account for the genetic variation underlying differences in disease or drug response. So, if there are 50 components in a disease, if you cover half of the genome with common SNPs, then you would expect that you should be able to find 25 of them or at worst, 10 of them. That's pretty good since people aren't finding any right now.

You would like to find as many of these genes as you can because each one is likely to have a small effect. If you don't find many of them, then you aren't going to be able to use them in any kind of predictive way for choosing between existing treatment options or regrouping people. So, that's why in the context of what Perlegen is interested in, which is being able to use existing drugs more effectively, finding as many of the genes as possible that play a role in a drug response or disease is very important.

Designing Association Studies Using Pooling Strategies and 500K Microarrays

Craig: Your laboratories at Perlegen have had the technology and capabilities to conduct genome-wide SNP association studies for a couple of years now. Based on your experience, what advice do you have for designing association studies, especially as the chip-based technology reaches the broader genetics community?

Cox: There aren't a lot of tricks you can use—you have to do a well-powered experiment. If people are looking for a gene involved with complex traits using 5,000 random markers across the genome, the chance that they are going to find it is pretty small. Different experimental questions in different populations require different numbers of markers, and that also has to do with the cost point. As cost comes down and you can look at more and more markers for a cheaper price, everyone will look at more and more SNPs. For now, it has to be based on the question you are trying to address.

Craig: I know Perlegen has used sample pooling strategies for genome-wide association studies with the genotyping assays you have developed [using multiplex Polymerase Chain Reaction (PCR) on Affymetrix microarrays]. Is pooling still a key component of your strategy for doing high-density, genome-wide association

studies or have you moved to individual genotyping?

Cox: Pooling works, but the decision is mostly dictated by the question you want to address. If you have a rich dataset of phenotypic information with many different components, pooling to study one phenotype keeps you from studying other phenotypes in the dataset. If you individually genotype each person it would be more expensive, but you can use the genotype information to assess a number of different phenotypes. You should try to do the best design that you can, given economic constraints, and even if you are not covering the whole genome, as long as you are covering enough of the genome so that you have a reasonable chance of getting some answer to some gene, then it's not a waste of time. We are mostly genotyping individual samples, not pooling, in our current and recent studies.

> "We always need better technology, but it won't matter unless we do something useful with the amazing technology we have today"

Craig: Originally Perlegen used custom Affymetrix microarrays for genotyping. Affymetrix is now producing arrays to genotype 500,00 SNPs and Perlegen is starting to use these standard arrays in it's studies. How is what you are both doing complementary?

Cox: Perlegen's goal has always been to improve treatment options for people with disease. When Perlegen started, we had to develop new technology to do these novel genome-association studies. What's happening now though is that Affymetrix and other genotyping platforms are providing people around the world with access to tools that a couple of years ago didn't exist. Perlegen's business is not to make tools. If commercial technologies get better than what we can do internally, then Perlegen can focus more on its primary goal which is using technology to make existing treatment options for people with disease better.

Role of Genetics in Patient Care

Craig: What will the next three years be like for the genetics community?

Cox: Now that technology to do association studies exists, it's going to become painfully evident how badly we need to pay more attention to the phenotypes themselves and to clinicians who understand the specific traits. One of the things that is most difficult is the lack of ongoing epidemiology on outcomes in both disease and drug response. Existing and upcoming sample sets will be typed with more and more markers, but many lack a clinical outcome phenotype assessment program equivalent to the technology. This will be a real bottleneck to further understanding.

Initial association studies will identify causal components of complex disease that haven't been seen in the past. Those will be, alone and in combination, the basis for tremendous amounts of ongoing research to understand the mechanisms by which those genetic components contribute to the disease or drug response. As a nation, it won't be the science that drives studies to track response to treatment, but the economic cost of medical care and drugs, and the absolute necessity to use the resources applied to health in a more cost-effective way. These discoveries are going to be part of general medical care and as a result, it's not going to just be the scientists involved, but the politicians, the health care professionals, and the insurers, too. Although it won't be easy and it won't happen fast, my belief is that genetic medicine will have to happen and there will be significant progress over the next five years.

Finally, it's difficult to take new technology and do something really useful with it. It's much easier to keep making technology better. Much of the discussion today focuses around getting better technology. We always need better technology, but it won't matter unless we do something useful with the amazing technology we have today. Over the next few years, given the tremendous improvement in our technology, I think people are going to demand that we actually do something that will impact their lives, health and medicine. As a genetics community, we need to keep our eye on that particular ball because if we just focus on technology over the next five years, then no one is going to let us do it anymore.

FOR MORE INFORMATION

Contact

■ David R. Cox, M.D., Ph.D.
Chief Scientific Officer and Co-founder
Perlegen Sciences, Inc.
2021 Stierlin Court
Mountain View, CA 94043
david_cox@perlegen.com

Companies

- Affymetrix Inc. – http://www.affymetrix.com
- Perlegen Sciences Inc. – http://www.perlegen.com

Organizations

- International HapMap Project – http://www.hapmap.org
- National Human Genome Research Institute (NHGRI) – http://www.genome.gov
- Stanford University – http://www.stanford.edu

Further Reading

- Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. Whole-genome patterns of common DNA variation in three human populations. *Science*. 2005;307(5712):1072-9.
- Hinds DA, Seymour AB, Durham LK, Banerjee P, Ballinger DG, Milos PM, Cox DR, Thompson JF, Frazer KA. Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. *Hum Genomics.* 2004;1(6):421-34.
- Hinds DA, Stokowski RP, Patil N, Konvicka K, Kershenobich D, Cox DR, Ballinger DG. Matching strategies for genetic association studies in structured populations. *Am J Hum Genet.* 2004;74(2):317-25.
- Frazer KA, Chen X, Hinds DA, Pant PV, Patil N, Cox DR. Genomic DNA insertions and deletions occur frequently between humans and non-human primates. *Genome Res.* 2003;13(3):341-6.
- Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT,Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR,Frazer KA, Fodor SP, Cox DR. Blocks of limited haplotype diversity revealed by high-resolutionscanning of human chromosome 21. *Science*. 2001; Nov 23;294(5547):1719-23.

AMB Featured Research: Association Studies

Rockefeller Scientists Make Progress in Genetic Study of Obesity

Scientists at The Rockefeller University are performing one of the world's largest association studies of obesity using Affymetrix microarrays to genotype over 600,000 SNPs from more than 3,200 individuals—almost the entire adult population of the Micronesian island of Kosrae. They've recently completed a critical first step of their study, using 100K microarray data to create a haplotype map of the Kosraen population, nearly 30 percent of which suffer from obesity, high blood pressure, or diabetes.



While the study is ongoing, Dr. Penelope Bonnen, a geneticist on the Rockefeller team says that they've generated 100K data for nearly the entire population, and are about to start 500K genome scans on the same samples—a total of 600,000 SNPs per individual.

Bonnen also says a publication describing their initial haplotype mapping results is forthcoming.

Further Information:

- Genetic Basis of Obesity in Humans - http://www.rockefeller.edu/abheads/friedman/research.htm
- Haplotype Map for an Isolated Population in Micronesia (Presentation, GeneChip Forum 2005) - http://www.affymetrix.co.jp/GeneChipForum2005/Abstract.Dr.Penelop.pdf

HapMap-tested Genotyping Assays and Microarrays Used for Association Studies

Baylor's Richard Gibbs talks about genotyping for the HapMap Project and performing other large-scale epilepsy and cancer genome studies

Baylor College of Medicine's Richard Gibbs, genotyped 5 percent of the genome, or 65,000 SNPs, for the HapMap project using microarrays and a new multiplex assay. This molecular inversion probe (MIP) assay enabled Gibbs' laboratory to genotype up to 10,000 targeted SNPs in a single experiment, and is now being used in follow-up association studies.

"Microarrays have brought about a revolution in cost and have given us the ability to look at large-scale genomic changes at the same time," said Gibbs. "I see developments in microarrays in the past year or so as being analogous to the developments in sequencing we saw 10 years ago that brought us down the road of genomics so quickly and so successfully."

Gibbs, who also played a key role in completing the public Human Genome Project and whose center has sequenced more than 10 percent of the human genome in addition to multiple other organisms, recently sat down with Dr. Penelope Bonnen, a postdoctoral fellow in the laboratory of Dr. Jeffrey Friedman, of The Rockefeller University to discuss future directions in genetic analysis. The two discussed:

- Baylor's role in the HapMap Project and use of new multiplexing assays
- Application of microarray technology in genome-scale cancer studies
- Progress in studying genetic variants associated with epilepsy in the Ion Channel Project

Multiplexed Genotyping Assays and the HapMap

Bonnen: Why did the Baylor College of Medicine's Human Genome Sequencing Center (BCM HGSC) choose this MIP genotyping technology for its contribution to the HapMap project?

Gibbs: You are speaking, of course, about the ParAllele MIP (molecule inversion probe) technology. If we wind the clock back three years, you will remember that the potential for doing very large-scale and accurate genotyping was not clear; and there were many challenges that accompanied the very idea of the HapMap project.

We were drawn to the ParAllele technology because of its great potential for high accuracy and its potential to be multiplexed. That was happily facilitated by a chance meeting with Mostafa Ronaghi, one of ParAllele's founders, who told us about the technology and invited us into larger discussions with the other ParAllele founders. Those discussions ultimately led to us submitting an application together to participate in the HapMap project and really test this technology out and see if it could evolve from a fairly specialized implementation at ParAllele, to something that was truly production-ready and robust technology.

Bonnen: How did you help develop the technology as you were working towards genotypes for the HapMap?

Gibbs: The fundamental molecular strategy was in place, but there was an awful lot of empirical testing—road testing—of variations and conditions to find the optimal configuration for accuracy, reproducibility and conversion rates. The majority of that work was done by ParAllele; then various versions of the technology were transferred to us as the alpha test site and implemented at scale in the HapMap Project. It was the performance of those assays in the HapMap application, at 10,000 SNPs per assay, that validated the new protocols and their readiness for wider distribution.

Studying Coding SNPs and the Cancer Genome Project

Bonnen: With a draft of the human genome sequence complete, and the HapMap nearing completion, what do you see as the next large-scale public scientific

effort with regard to using genetics towards the advancement of medicine and public health?

Gibbs: Well, my own bias is that we should be doing a very large project to identify and study coding SNPs, and this would be the germline exon sequencing project. In other words, we should be looking at very large numbers of samples and drilling in deep, finding out the variation in the population in the regions that are obviously of functional importance—coding regions and some of the promoter regions. I think this dataset will really be invaluable for those who want to follow up different diseases that may not be fully tractable even with the HapMap tools and the other reagents out there.

Bonnen: Please talk about the cancer genome project and your involvement in it.

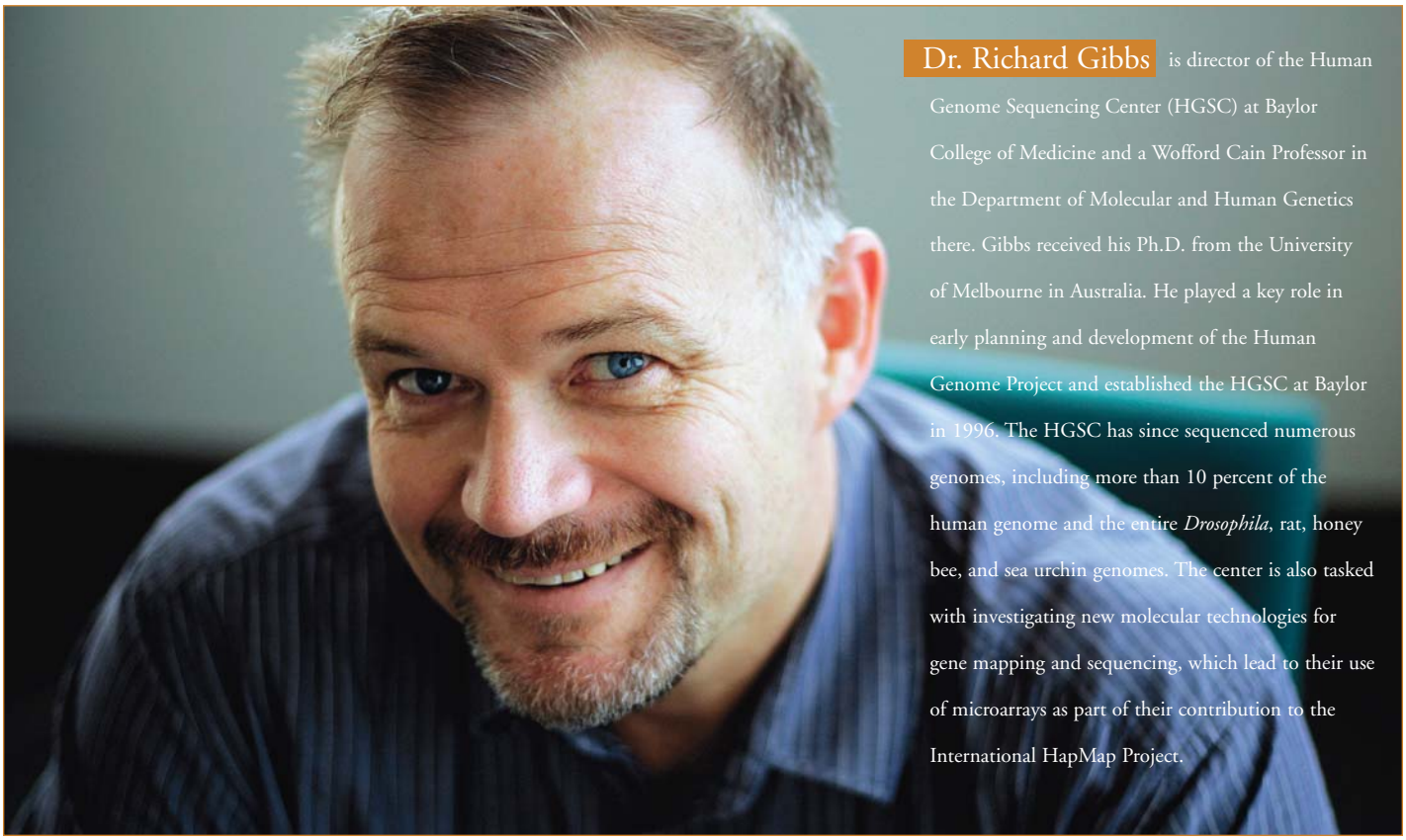
Gibbs: This is another exciting initiative that gets back to the next steps after the human genome sequencing and the HapMap. This is a somatic exon sequencing project for one of the most compelling health problems of humanity—cancer. The idea is to find both single nucleotide mutations in genes and larger-scale structural changes in the genome that may correlate with cancer. As a genome center, our current involvement is to perform pilot studies that show the feasibility of doing this kind of work at large scale. And as a member of the Texas Medical Center community, our role is to coordinate a rich sample collection from a very large number of cancer researchers into a pipeline where we can generate this genomic data and give it back to the researchers in the form of a public data release.

> "I see developments in microarrays in the past year or so as being analogous to the developments in sequencing we saw 10 years ago that brought us down the road of genomics so quickly and so successfully"

Bonnen: Can you describe the study design and how many samples you are interested in ultimately sequencing?

Gibbs: A lot of this is in flux. The initial project talked about the number of tumor plus constitutive DNA pairs that would be necessary to find mutations present at the five-percent level. That would probably be 250 or so tumor samples. The most recent plans, in a beautiful white paper produced by our colleagues at [the National Cancer Institute] NCI, aims for complete analysis of that large number of tumor samples, but that would probably not be achievable in under five years.

So, if we can draw our vision back to what is really going on at the bench and what the pipelines are achieving, I think it's realistic that in the next year we will see several hundred or even several thousand genes sampled in a couple of hundred



Dr. Richard Gibbs is director of the Human Genome Sequencing Center (HGSC) at Baylor College of Medicine and a Wofford Cain Professor in the Department of Molecular and Human Genetics there. Gibbs received his Ph.D. from the University of Melbourne in Australia. He played a key role in early planning and development of the Human Genome Project and established the HGSC at Baylor in 1996. The HGSC has since sequenced numerous genomes, including more than 10 percent of the human genome and the entire *Drosophila*, rat, honey bee, and sea urchin genomes. The center is also tasked with investigating new molecular technologies for gene mapping and sequencing, which lead to their use of microarrays as part of their contribution to the International HapMap Project.

AMB Featured Research: Association Studies

100K SNP Scan Helps Finnish Scientists Uncover 700 New Associations in Cardiovascular Disease

Finnish biotechnology company Jurilab Ltd. recently completed one of the first whole-genome association studies, discovering more than 700 genes associated with coronary heart disease, hypertension and diabetes. Its scientists identified the disease genes using Affymetrix 100K microarrays to genotype 100,000 SNPs from 125 cases and 125 controls in the isolated population of Eastern Finland. Through collaborative partnerships, Jurilab plans to develop the first molecular diagnostic for hypertension, followed by other diagnostic tests and drug development products.

In June, Jukka Salonen, the company's chief scientific officer, and his team published some of their first results in the journal of *Twin Research & Human Genetics*. They reported that linkage disequilibrium (LD) varies across the genome, but generally correlates with local intra- and inter-chromosomal recombination rates. Salonen spoke to *AMB* in April about these findings and the challenges he faced studying a mixture of common complex diseases through large-scale population studies.

Further Information:

- Principal Scientist Jukka T. Salonen Discusses Challenges of Population Study (*Affymetrix Microarray Bulletin*, April 2005) – http://www.microarraybulletin.com/salonen
- Genome-wide Linkage Disequilibrium From 100,000 SNPs in the East Finland Founder Population (*Twin Research & Human Genetics*, June 2005) – http://www.australianacademicpress.com.au/Publications/Journals/Twin_R/TResearch.htm

tumor samples. There will have to be some judicious selection of those and there have already been discussions about it. During this early period, there are many fundamental technical questions we need to ask and answer:

1. What is the role of whole-genome amplification?
2. What is the role of genome screens for other kinds of lesions?
3. What methodologies would you prefer to use for various types of genomic analyses?

While it's a five-year plan, this next year's mission is to bring together these different technologies and design the full experiment.

Bonnen: Some criticisms or concerns have been voiced about the cancer genome project. How does the group plan to address those concerns?

Gibbs: Oh, I love it. What we are hearing is all very reasonable as a response to the proposal, but it's pretty similar to what we heard about the Human Genome Project 10 to 15 years ago. So, the criticism really is that the project is naïve and that's true, but I think that's actually its strength. It is certainly true that we don't have a comprehensive catalog of these mutational changes, and these genotype data will greatly facilitate that.

I think the critics are really just opposing the idea of a broad and comprehensive, versus a targeted and more hypothesis-driven approach. The real answer is we need both. We need this broad, naively-generated catalog that gives a technically excellent description of these fundamental changes in all of these disease types, and then we need to be able to analyze and follow up those data.

The Ion Channel Project at BCMHGSC

Bonnen: What is the Ion Channel Project and how was it conceived?

Gibbs: The Ion Channel Project is a collaboration between the HGSC and Dr. Jeff Noebels, Dr. Dan Burgess and others here at Baylor College of Medicine. The Ion Channel Project is kind of a candidate-gene disease discovery project on steroids. In other words, we are looking at a group of ion channel genes likely to be involved in idiopathic epilepsy. Our collaborators see patients in the clinic and they determine cases that are truly idiopathic—that is, there is no known reason for the disorder. They then collect samples from affected individuals and controls, and we resequence them by [Polymerase Chain Reaction] PCR and direct sequencing of exons and 5' regions of about 250 ion channel genes. From that, we are able to discover putative functioning variants and then we enter that into a pipeline of informatic analysis to determine if the variation might truly be of functional significance.

When we have all the samples accumulated, we will be able to do a statistical study where we look to see if the cumulative variation in the individual genes or some subset of the individual genes correlates strongly with the phenotype and has a frequency over and above what we see in the control groups.

We are up to 130 patients right now. It's a five-year project, and we are now just finishing the first year. But, we have more than one-fifth of our patients—maybe a quarter of them—and the rate of accumulation of the patient samples is increasing, so we are in good shape there. We have also built the sequencing assays for every one of these genes, and we have got different combinations of all genes completed in some of the patients. The project has accumulated several hundred thousand resequencing reactions and we've discovered a couple of hundred potentially interesting mutations.

Bonnen: And are you planning a second-tier analysis where you will look at the effect of combinations of mutations or combinations of genes?

Gibbs: Absolutely. The study is tailored for that. One very workable hypothesis is that we need multiple hits in the same individual to get the phenotype. We will be able to see that, as long as the hits are in ion channel genes. Then we can see if there needs to be a certain mutational burden or mutational pattern in different genes to get the effect. I don't think we can find that any other way.

Applying Microarray Technology to Disease Studies

Bonnen: How have technological advances affected the scope of your

projects, and specifically, how does microarray technology fit into large-scale, high-throughput studies you're working on?

Gibbs: The theme here is to generate more data at the same or less cost to make more discoveries and to develop more powerful assays to answer fundamental questions. We can look at DNA sequencing as an example of how technological advances impact discovery and advance science. It used to be a real luxury to sequence just a kilobase of DNA; now, we have sequenced multiple genomes and we can track the impact of that on disease.

Now, if we turn to genotyping, microarrays have brought about a revolution in cost and have given us the ability to look at large-scale genomic changes at the same time. I see developments in microarrays in the past year or so as being analogous to the developments in sequencing we saw 10 years ago that brought us down the road of genomics so quickly and so successfully.

Bonnen: Are these types of large-scale scientific efforts only possible in large genome centers like yours or will major genetic studies be done in smaller labs as well?

Gibbs: Both. That's a great question, and it should guide us somewhat in how we, as a community, choose major projects. To a large extent, the measure of success of a large genome project should be its ability to enable activities of individual researchers. Once again, microarrays beautifully illustrate the principle. You can imagine that the pipeline consists of large-scale discovery in the conext of a genome-scale project in a large center, and that knowledge is delivered to the individual investigator by tailor-made reagents—microarrays. Microarrays are kind of the building blocks of discovery.

Bonnen: What will be the impact of these large-scale studies?

Gibbs: Many and varied. So, when we talk about genome projects, we really are beginning to understand the function of the human genome; and for the DNA that we used to wonder about, we are starting to find out whether it does or does not do anything through these comparative analyses. And in genetics, I believe that we are going to find all of these disease-causing alleles in the next couple of years, and it's going to be because of the genotyping and sequencing afforded by microarrays.

Bonnen: How will HapMap data be used in the study of disease?

Gibbs: I see two main categories. The initial public reason for doing the HapMap project was largely to generate a database of tag SNPs. Tag SNPs are SNPs that are in linkage disequilibrium with other SNPs, which allow you to get more genetic information with less genotyping and help you avoid doing redundant assays. The idea at the outset was that the presumed million or so markers that were going to be needed for genome-wide association studies might be reduced by a factor of two or more. There was some uncertainty about how useful HapMap data would be, but as it turned out, HapMap data looks like it will be very useful in helping to identify SNPs and analyze genetic results.

More generally, what we have done is created a first broad description of the genetic architecture across three major human populations, across the whole genome. We have really turned on the lights here and really started to get a good indication of the basic data that we have got out there.

Bonnen: What is the biggest overall accomplishment of the HapMap project?

Gibbs: Well, pragmatically, the project is helping drive down genotyping to under a penny per genotype for big experiments. That is a pretty substantial achievement considering when the project started, genotyping costs were 25 cents or more per genotype. Some people might consider it inappropriate for academics to focus on economics, but in fact economics is what enables large studies to be done and it is large studies that are going to lead us to the Holy Grail—finding the alleles that actually cause changes in human phenotype and disease.

Bonnen: How will this work be translated into diagnostics or treatments in the clinic?

Gibbs: I see a big question mark there. In other words, we have to translate discovery into clinical relevance—clinical utility. I think discovery can be done with the existing tools, but if we want to see these tests go out there in the clinics so that every single person that goes to the doctor's office can get a whole-genome sequence, then we really are going to need to see the technologies take a big leap and some new revolutionary methods come along to enable that.

FOR MORE INFORMATION

Contacts

- Richard A. Gibbs, Ph.D.
Director, BCM-Human Genome Sequencing Center
Wofford Cain Professor
Department of Molecular and Human Genetics
Baylor College of Medicine
One Baylor Plaza
Houston, TX 77030
agibbs@bcm.edu

Companies

- Affymetrix Inc. – http://www.affymetrix.com
- ParAllele BioScience – http://www.parallelebio.com

Organizations

- The International HapMap Project – http://www.hapmap.org
- Baylor College of Medicine Human Genome Sequencing Center – http://www.hgsc.bcm.tmc.edu/
- Baylor College of Medicine – http://www.bcm.edu
- Texas Medical Center – http://www.tmc.edu
- National Cancer Institute – http://www.cancer.gov

Further Reading

- Belmont JW, Gibbs RA. Genome-wide linkage disequilibrium and haplotype maps. *Am J Pharmacogenomics*. 2004;4(4):253-62.
- Hardenbol, P., et al. Highly multiplexed molecular inversion probe genotyping: Over 10,000 targeted SNPs genotyped in a single tube assay. *Genom Res*. 2005;15:269-275.

People

- Jeff Noebels, M.D., Ph.D.
Baylor College of Medicine
http://imgen.bcm.tmc.edu/molgen/facultyaz/noebels.html

- Dan Burgess, Ph.D.
Baylor College of Medicine
http://www.bcm.edu/neurology/faculty/burgess.html

- Jeff Friedman, M.D., Ph.D.
Rockefeller University
http://www.rockefeller.edu/labheads/friedman/research.htm

AMB Featured Research: Association Studies

Finding a Common Variant Associated With Increased Risk of Macular Degeneration

Josephine Hoh of Yale University and her colleagues at The Rockefeller University discovered one of the first genes associated with age-related macular degeneration (AMD) using microarrays to genotype hundreds of thousands of SNPs. They found that a common variant of the complement factor H (CFH) gene increases a person's risk of developing AMD by 800 percent.

Hoh's recent study, published in the April 2005 issue of *Science*, used Affymetrix 100K arrays to genotype 146 people to find the mutation. Following the SNP mapping study, the researchers sequenced the CFH gene from both AMD cases and controls and discovered that the SNP accounted for the disease-associated form of the gene.

CFH is a known inhibitor of the complement-based immune response; the mutation may make it impossible to control complement activation, resulting in chronic inflammation and damage to healthy tissues. The research group found both CFH and complement protein deposits in the retina and damaged eye tissue of AMD patients.

Further Information:

- Complement Factor H Polymorphism in Age-Related Macular Degeneration (*Science*, April 15 2005) – http://www.sciencemag.org/cgi/content/abstract/308/5720/385
- Understanding Macular Degeneration in Seniors (Affymetrix Research Highlight, August 2005) – http://www.affymetrix.com/community/wayahead/macular_degeneration.affx

Scientists Create Japan's Largest BioBank for Genetic Studies of 47 Common Diseases

RIKEN's Yusuke Nakamura discusses plans to genotype more than 10,000 individuals

RIKEN and University of Tokyo scientists, led by Yusuke Nakamura, contributed more than 24 percent of the genotype data to the International HapMap Project and are now genotyping 10,000 individuals as part of the largest set of association studies ever performed in Japan—the Personalized Medicine Project. The group is collaborating with Perlegen to study 47 diseases using Affymetrix 250K custom arrays; in just the first year scientists have discovered four genes that could help physicians more effectively dose anti-coagulant therapy.

Nakamura's team gained much of its experience in large-scale sample handling and data analysis during the five-year Japanese Millennium Project, as well as the three-year HapMap project in which they genotyped seven chromosomes.

"Before the HapMap era, we didn't have information about how different we are," said Nakamura. "But now, we have a set of [single nucleotide polymorphisms] SNPs that can be used in large-scale genetic studies aimed at finding disease-susceptibility genes or drug-response-related genes. The information should be applied to development of novel therapeutic tools, and also for preventing disease or preventing disease progression."

Sixty-six hospitals participating in a BioBank Japan initiative have already collected more than 115,000 DNA samples that will be used as part of the Personalized Medicine Project to genotype patient samples in disease areas such as atherosclerosis, rheumatoid arthritis, epilepsy, heart failure, chronic obstructive pulmonary disease, asthma, arrhythmia, cerebral infarction, and many types of cancer.

Scientists at RIKEN's SNP Research Center plan to conduct fine-mapping in candidate regions as well as genotyping the associated SNPs in additional sample sets to confirm the results. Furthermore, researchers at the SNP Research Center and their colleagues plan to conduct follow-up analysis to understand the function of the genetic associations they discover.

Nakamura spoke recently to *AMB* Contributor Dr. Penelope Bonnen, from The Rockefeller University about this monumental genotyping effort made possible through the HapMap and BioBank Projects. The two discussed:

- The design and goals of the BioBank Japan Project
- Challenges of genotyping BioBank samples to discover correlations between genetic variation and drug response
- Implementation of whole-genome microarray analysis for disease gene discovery

Design and Goals of the BioBank Japan Project

Bonnen: Please tell us more about the design of the biobank?

Nakamura: The BioBank Japan project started two years ago. Our goals are to build the BioBank with DNA and sera from 300,000 individuals, and to analyze a huge body of SNPs for those samples. We aim to identify the genes of medical importance and provide some information about the development of disease, with an additional goal of establishing personalized medicine.

At present, we have collected DNA from 115,000 individuals. The total number of cases is already beyond 160,000. There are more cases than individuals because one patient often has multiple diseases. I think these samples provide a huge number of valuable resources to identify genetic factors associated with common diseases.

Bonnen: The number of biobanks is growing. What are some of the ethical and practical issues a biobank needs to consider?

Nakamura: One of the important issues is the protection of individual privacy. In Japan, social discrimination based on genetic diseases still exists. So many Japanese people are concerned about genetic discrimination as a result of information in our BioBank project. It is not easy to convince patients to give us a blood sample. But so far, 87 percent of the individuals whom we ask to provide blood have agreed to do it. It's strikingly higher than we expected.

We still need to convince the Japanese public that the BioBank Project can lead to significant progress in medical treatment. So we made a BioBank Japan website which explains the benefits of the project to lay people. In addition, to eliminate genetic discrimination, we are doing our best to educate young Japanese people that we should be equal even though we are different. A strong sense of bioethics is important to the future of medicine, and for that, the Japanese government needs to make a law to prevent genetic discrimination.

Genotyping Samples from the BioBank; Taking on Personalized Medicine

Bonnen: BioBank Japan is a very large project. How can you get value out of it and how can it be used?

Nakamura: We just started SNP analysis for 47 different diseases and plan to have genotype data within the next three years. But, that is only the beginning of understanding the genetic cause of various diseases. We will need to do a deeper analysis in the following years.

We are now further investigating some genetic results that may increase a risk of common diseases. Identification of genetic factors associated with susceptiblity to diseases or to drug response should contribute to improvement of our medical care.

For example, warfarin is an anti-coagulant. Dose adjustment is very important and this is one of the examples where there is a big difference between Caucasians and Asians. In the United States, a patient usually takes 10 to 15 mg per day. But in Japan, it's 1 to 10 mg. There is a big difference between the two ethnic groups, but there is also a big difference among Japanese.

> "We now have four genes that may indicate the appropriate dose for individual patients"

If a patient overdosed on the drug, the patient would have an adverse reaction of bleeding, so adjustment of dose is very important. We now have four genes that may indicate the appropriate dose for individual patients and provide safer management of the anti-coagulant therapy. Although this is just one example, we hope to provide this information to our patients and to our hospital with the hope that doctors will provide a better grade of treatment.

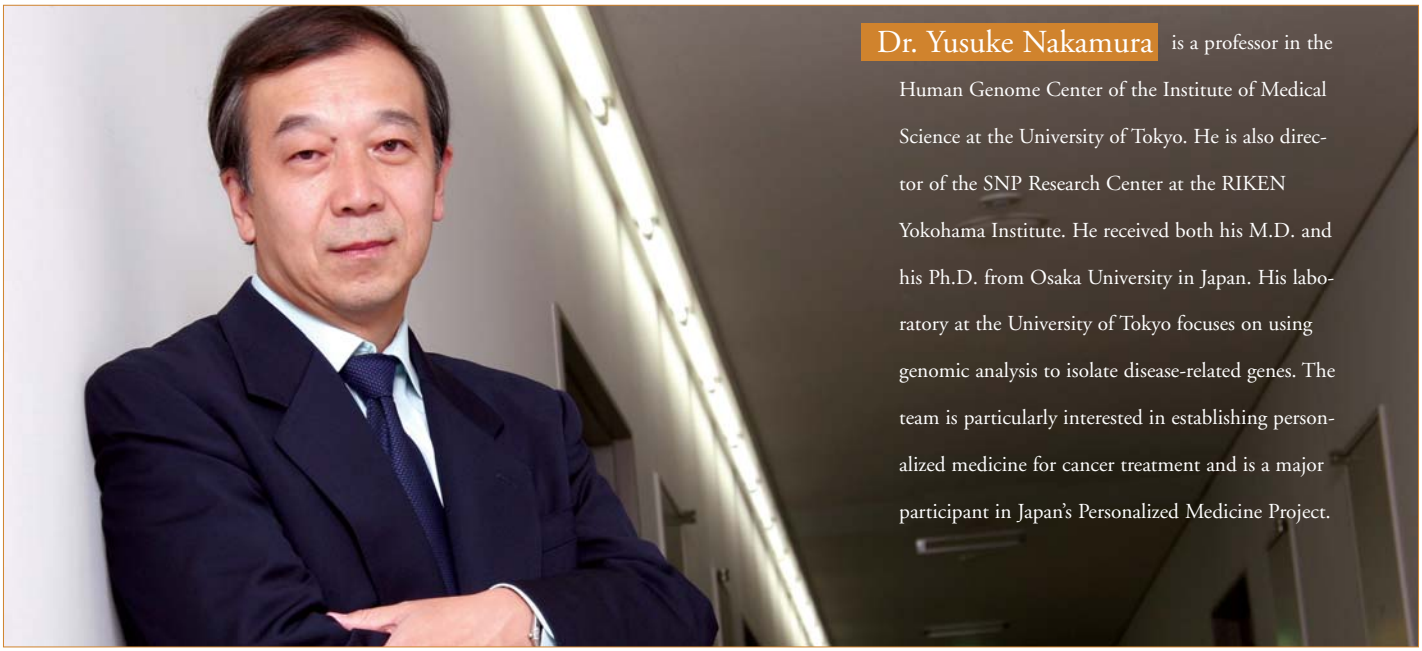
Bonnen: What is the Personalized Medicine Project and why was it conceived?

Nakamura: Actually, the BioBank project is a part of the Personalized Medicine Project.

The dose estimation of warfarin is one example of why this project is important. In addition, I think anti-cancer drugs are another good example. Efficacy of anti-cancer drugs is usually 20 to 30 percent. That means a majority of patients who take an anti-cancer drug do not respond well to that treatment.

At present, we cannot predict who will respond well and who will not. Patients have to take a chance when using these drugs. In the twentieth century, we did not have any tools to distinguish responders from non-responders. But now we have better tools to examine the expression profile of cancer cells and look at differences in characteristics of cancer cells in individual patients.

Using a combination of genomic information, expression profiles and genetic variations like SNPs, we can predict which patients will respond and which will have



Dr. Yusuke Nakamura is a professor in the Human Genome Center of the Institute of Medical Science at the University of Tokyo. He is also director of the SNP Research Center at the RIKEN Yokohama Institute. He received both his M.D. and his Ph.D. from Osaka University in Japan. His laboratory at the University of Tokyo focuses on using genomic analysis to isolate disease-related genes. The team is particularly interested in establishing personalized medicine for cancer treatment and is a major participant in Japan's Personalized Medicine Project.

AMB Featured Research: Association Studies

Discovering the Genetic Variations Behind Multiple Sclerosis

Scientists studying multiple sclerosis at the Serono Genetics Institute have used high-density genotyping microarrays to discover more than 80 genes associated with the disease. They replicated their findings in three different populations using Affymetrix 100K arrays to scan more than 100,000 single nucleotide polymorphisms in each of 3,000 samples.

Daniel Cohen, the research leader at Serono, told *AMB* in April that 10 to 20 percent of the genetic associations they discovered were in gene deserts or regions of "junk DNA." Now Cohen wants to use tiling microarrays to figure out the function of those associations present in parts of the genome that were not supposed to do anything.



Further Information:

- Dr. Daniel Cohen Discusses Whole-genome Association Study to Find New Genes Involved in Disease and New Drug Targets (*Affymetrix Microarray Bulletin*, April 2005) – http://www.microarraybulletin.com/cohen
- Serono Announces Major Milestone in Identifying the Genes Involved in Multiple Sclerosis (Press release, March 17, 2005) – http://www.seronousa.com/news/archives/20050317.jsp

a higher risk of adverse reaction. We can provide a drug only to the patients who are likely to show good response with a very small risk of adverse reaction. That's the kind of the personalized medicine we would like to establish.

Bonnen: As a part of this project, you are investing in large-scale association studies. How were these studies designed? How many SNPs and how many DNA samples will you be typing?

Nakamura: As I mentioned before, we currently have 160,000 samples for 47 diseases. Our design calls for 200 individuals for each disease in the first screen. We are screening 250,000 SNPs. According to the HapMap data, 250,000 SNPs will cover 98 percent of the genome information. That should be enough for Asian samples, but the same number of SNPs covers only 70 percent of the genome in Africans. Being in Japan, we are in a better position because we need fewer SNPs to cover the entire genome.

As a first screening, as I mentioned, we will analyze 250,000 SNPs. Then we do a preliminary association analysis and select SNPs which show a p value of 0.01 or less.

The number of SNPs being positive in the first screening is expected to be between 4,000 and 6,000 from our previous experience. Those roughly 5,000 SNPs will be analyzed for a larger number of cases—we are now planning around 2,000 cases per SNP. Through the additional finer SNP mapping of the disease loci, we will be able to narrow down the candidate genes to 10 or 20. Subsequently, we will analyze the biological or medical effect of the genetic variation using various analytical tools. We already have a lot of experience from the Japanese Millennium Project, and reported a number of genes susceptible to common diseases. So we are quite confident in this kind of approach to find the gene-associated risk with various diseases.

Bonnen: What are some of the analytical tools that you are considering for dealing with the data?

Nakamura: The most difficult part in analyzing the data is combining two or more SNPs together from the data of 250,000 SNPs. We need to develop a better algorithm to analyze the combined effect of multiple loci; we do not have an effective algorithm to handle thousands of loci.

We also need better bioinformatics and biostatistics to understand the interactions between genes and environmental factors. Although we at present don't have permission to obtain it, we hope to collect environmental or lifestyle information on the patients in order to better understand the etiology of disease. If we can do that, the analysis will become much more complicated because of the multiple genetic variations and environmental and lifestyle factors.

SNP Analysis and Microarrays

Bonnen: How have technological advances affected your project and specifically, how does microarray technology fit into your work?

Nakamura: It is important to remember that five years ago, we could work on a limited number of SNP loci at a time because that was the maximum we could get information for at once. But microarray technology provides us with tools to examine almost the entire genomic region in one experiment.

Limited genetic information provides us with a very small amount of information about the basis of disease. But now, we have a whole set of information at the expression level or genetic variation level. This information will help us understand disease better. That could change our medical system entirely.

Bonnen: How did you choose to work with Perlegen and Affymetrix?

Nakamura: Our budget is limited, but we wanted to examine SNPs covering an entire genome. Perlegen offered to construct custom-made analysis of 250,000 SNPs that cover most of our genome.

Affymetrix has given us tools to examine expression patterns of the entire set of exons in our genome. That is pretty important for my life-time interest of cancer research. It is widely known now that there is a very complex expression pattern of transcripts in cancer cells. If we can identify cancer-specific transcripts or cancer-specific exons, we will be able to generate treatments that target cancer cells specifically. That is very important in order to develop a drug with maximal efficacy and minimal risk of adverse reaction. Although our experiments in the past several years provided useful information for development of diagnosis and treatment of

cancer, these novel tools will add much more information for improvement of medical treatment.

I started as an oncologist, so my biggest hope is to develop a good drug for cancer patients and find a way to predict the efficacy of anti-cancer drugs prior to treatment. I think if we can get there through detailed expression information in cancer cells, we will be able to distinguish responders from non-responders, and provide the treatment only to those who will benefit from it.

FOR MORE INFORMATION

Contacts

■ Yusuke Nakamura, M.D., Ph.D.
Human Genome Center
Institute of Medical Science
The University of Tokyo
4-6-1 Shirokanedai
Minato-ku, Tokyo 108-8639
Japan
yusuke@ims.u-tokyo.ac.jp

Companies

■ Affymetrix Inc. – http://www.affymetrix.com
■ Perlegen Sciences Inc. – http://www.perlegen.com

Organizations

■ The University of Tokyo, Institute of Medical Science – http://www.ims.u-tokyo.ac.jp/imswww/index-e.html
■ The International HapMap Project – http://www.hapmap.org
■ BioBank Japan– ttp://www.biobankjp.org

Further Reading

■ Tsuge M, Hamamoto R, Silva FP, Ohnishi Y, Chayama K, Kamatani N, Furukawa Y, Nakamura Y. A variable number of tandem repeats polymorphism in an E2F-1 binding element in the 5' flanking region of SMYD3 is a risk factor for human cancers. *Nat Genet*. 2005 (in press).
■ Kato T, Daigo Y, Hayama S, Ishikawa N, Yamabuki T, Ito T, Miyamoto M, Kondo S, Nakamura Y. A novel human tRNA-dihydrouridine synthase involved in pulmonary carcinogenesis. *Cancer Res.* 005;65(13):5638-46.
■ Hamamoto R, Furukawa Y, Morita M, Limura Y, Silva FP, Li M, Yagyu R, Nakamura Y. SMYD3 encodes a novel histone methyltransferase involved in the proliferation of cancer cells. *Nature Cell Biology*. 2004;6:731-740.
■ Ozaki K, et al. Functional variation in LGALS2 confers risk of myocardial infarction and regulates lymphotoxin-α secretion in vitro. *Nature*. 2004;429:72-75.

British Geneticists Use HapMap in Major Study of Common Diseases

Cambridge Medical School's John Todd discusses controls, samples and collaborations needed to find new associations between gene variants and disease

More than a dozen scientists from Great Britain, including researchers from Cambridge University's Medical School and the Wellcome Trust Sanger Institute (WTSI), are using high-density genotyping arrays to screen for genome-wide markers associated with common diseases like heart attack, cancer and diabetes as part of the Wellcome Trust Case-Control Consortium (CCC).

The study builds on Phase 2 of the HapMap project and the consortium has partnered with Perlegen and Affymetrix to genotype more than 600,000 SNPs from 19,000 samples to discover genes associated with eight diseases.

John Todd of Cambridge University's Juvenile Diabetes Research Foundation (JDRF) and the Wellcome Trust Diabetes and Inflammation Laboratory (WT DIL) has been studying type 1 diabetes for more than 18 years. He uses a combination of case-control and family studies, statistics, genome informatics and data mining, and gene expression and function studies to identify and examine the consequences of disease-associated variants.

"HapMap is a crucial component in answering a big question for disease: What proportion of the heritability, of the familial clustering, is contributed by common variants down to those present at only five percent or so?" said Todd. "We couldn't begin to address that question without the coordinated effort and the availability of data from the HapMap project."

Todd, who will soon take a joint appointment at the WTSI, is currently engaged in one of the world's largest case-control whole-genome association studies of type 1 diabetes, using high-density microarrays and multiplexed genotyping assays to test a panel of more than 7,000 non-synonymous single nucleotide polymorphisms



Dr. John Todd is a university chair in the Department of Medical Genetics at the Cambridge Institute for Medical Research. His laboratory in the JDFR/WT DIL studies the molecular basis for type 1 (insulin-dependent) diabetes using an integrated combination of genetics, statistics, genome informatics and data mining, and gene expression and function studies. Todd is also involved in the international type 1 diabetes Genetics Consortium and in the WT CCC.

(SNPs)—SNPs that change the sequences of proteins.

Todd spoke recently to Affymetrix Vice President of DNA Products Greg Yap about his collaborative work and visions for future research. The two discussed:

- Medical Genetics and the HapMap at the WTSI
- The challenges of using historical samples and common controls in the CCC
- The distribution of work across the genetic community in multi-centered studies

Medical Genetics and HapMap at the WTSI

Yap: Could you describe your new role at WTSI in medical genetics?

Todd: I want to help the WTSI take some leadership in disease genetics, exploiting the scaled genotyping, sequencing and informatics expertise, the platforms and pipelines at the institute, coupled with building an equally powerful statistical analysis group. We will launch this major initiative by focusing, in house initially, on type 1 diabetes and obesity. The obesity genetics project involves a full collaboration between Nick Wareham at the Medical Research Council Epidemiology Unit, Inés Barroso at the WTSI, and Steve O'Rahilly in Cambridge. These in-house studies will be complementary to the WT CCC projects.

Yap: Would you characterize this as something new for WTSI or an extension of what the institute has been doing?

Todd: The new part is the focus on a very few, specific clinical phenotypes with epidemiology and statistical analysis that provides unique resources, expertise and opportunities. From the medical school side, we have a long track record of expertise in analyzing and studying human disease and associated phenotypes. We have an ongoing collaboration in metabolic disease (type 2 diabetes and obesity) between the WTSI and Cambridge called Genes, Energy and Metabolism, or GEM (Barroso, Wareham and O'Rahilly). We want to accelerate this collaborative study. We've also been collecting samples for and studying type 1 diabetes for the past 10 years in the JDRF/WT DIL. We have significant track records in expertise and sample collections in both diseases. When you put that together with informatics and statistical analysis expertise and the genotyping and sequencing technologies we're building at the WTSI, our intention is to make a major contribution to these two disease areas. We are consolidating, in an active and effective way, two well-established collaborations between the Cambridge Medical School and the WTSI. The follow-up studies at the WTSI will include collaborations in gene expression, epigenetic analyses and animal model studies.

Yap: The WTSI has made a major contribution to the HapMap project. Can you talk a little bit about how the HapMap project will influence downstream studies, including your own work?

Todd: HapMap will allow us to analyze common variants in a comprehensive way. It is a crucial component of answering the question, 'What proportion of the heritability of disease is contributed by common variants?'

Now, of course common disease is caused by common and rare variants. In a review that we wrote in *Nature Reviews Genetics*, we take a neutral position towards this. My view is that the determination of the allelic architecture of a common disease is limited by cost, feasibility and technology. We now have cost-effective SNP genotyping technology, we have the HapMap, and we have case and control samples. So there's an experiment that can be done now that will tell us a lot about what to do next. Clearly there are rarer variants contributing to the familial clustering of common disease. How do you find those? Some of them will come out of these genome-wide association scans, because looking at certain haplotypes will indicate that there's a rare variant present. And, as sequencing technologies improve, we'll be able to dig deeper into sets of candidate genes, some of which will come out of these genome-wide association scans, and some of which will come out of experimental models of disease or knowledge of gene function. At the WTSI, DIL and GEM collaborations there is already significant activity in deep resequencing of candidate genes, looking for rare variants. I'm not sure that the sequencing technologies are really at the level of cost effectiveness, throughput and feasibility we want yet; it's difficult to conceive doing even 100 genes on 4,000 people right now, and one aim is to make this routine in the future.

The Wellcome Trust Case-Control Consortium

Yap: Could you give a brief overview of the consortium? Why is it coming together and who are the major participants?

Todd: The Wellcome Trust CCC covers 12 diseases—inflammatory bowel disease, type 2 diabetes, type 1 diabetes, coronary heart disease, bipolar disorder, tuberculosis, rheumatoid arthritis and hypertension, with multiple sclerosis, breast cancer, and autoimmune thyroid disease in a secondary study. The investigators who collected the samples are members of the CCC and co-principal investigators (PIs), as well as experts in statistical analysis. We want to take 2,000 cases for each disease and 3,000 shared controls, the latest genotyping technology, HapMap data, and the statistics we've worked out over the past few years, and apply all those things together to see what we can find out. Do you find susceptibility genes? Are the diseases different in terms of the yield of susceptibility genes? What is the optimal approach for association studies? How many susceptibility genes are there? How can we design the experiment next time to find many more genes for less money? What is the relative distribution of rare versus common variants? All of those features are part of the goal of this consortium.

The consortium is funded by the Wellcome Trust and driven by a management committee chaired by Peter Donnelly from Oxford. It's comprised of a strong mixture of analysis researchers and disease-oriented PIs who have collected the samples for each of the disease areas. There is a very wide representation from the overall U.K. genetics community.

Yap: Could you talk about your choice of technology in these experiments in the CCC? Why did you choose to work with Perlegen and Affymetrix? And how do you see the combination of Affymetrix' 500K array with additional SNPs from Perlegen helping your work?

Todd: We chose to outsource the first phase—the genome-wide association study phase—to Perlegen and Affymetrix, because we felt that was the platform that could

generate genotypes today at a very large scale. We felt that Perlegen had a unique track record. And we wanted to begin to gain insights before the end of 2005.

We felt the combination of 500K and additional 'fill-in' SNPs from Perlegen would provide adequate coverage of common SNPs in the genome, specifically HapMap Phase 2. We recognized that the 500K chip was a product that could be run in different labs, including here at the WTSI. We liked the combination between a product that could be run outside Perlegen with the existing throughput and expertise of Perlegen.

Sample Collection and Common Controls

Yap: How difficult is it to collect samples in a standardized way from many different medical institutions?

Todd: Most of these samples have been collected historically under different protocols, so we are susceptible to the variation in sample bases between collections. That's one of the interesting challenges. But we have done our best to make the samples representative of the British population and to match cases and controls appropriately by geographical region. Almost all the samples have now been received by the WTSI-DIL joint collaborative DNA laboratory, and re-quantified using picogreen, for shipment to Perlegen. Re-quantification has been very successful with almost every sample passing the quality control (QC) process, which includes some limited agarose gel analysis and SNP genotyping.

Yap: In the CCC, you are planning to compare 3,000 common shared controls with 2,000 cases in each of your diseases—can you comment on that?

Todd: In studies with David Clayton (also a member of the DIL and CCC), we have discovered that if you match British people within broad regions of Great Britain, you can address, to a large degree, the issue of population stratification. Control samples come from all over Britain: 1,500 from the population-based 1958 British Birth Cohort, and a CCC-specified collection of 1,500 blood donor samples from the Blood Transfusion Services of England, Scotland and Wales, which is being coordinated by Willem Ouwehand in Cambridge. We already know from our own work in type 1 diabetes that common controls from the 1958 British Birth Cohort should be okay, provided the cases are from similar geographic regions in Britain. There are collections of controls being talked about in the U.S. and in other countries and I think they will all provide some very valuable information. This sample size—5,000 individuals for each disease and comparisons of genotype frequencies across case series—should provide a good starting point in terms of statistical power.

I can see enormous benefits in sharing individual genotype data among researchers. I don't think that information can just be put on a web site, because of confidentiality and ethical considerations, but I see huge benefits to figuring out how to share data between *bone fide* researchers as quickly as possible. How else will we assess compatibility and extract information from these data sets that will enhance each of the studies?

Yap: The utility and extent of utility for common controls is a topic of some debate lately. How do you see this evolving?

Todd: I think we can work it out empirically using the large-scale genotyping data itself to 'genomically-match' subjects. Once we do these experiments, we can estimate the loss of power in less-than-well-matched controls. We need large numbers of markers typed on different populations to adjust association test statistics by genetic stratification. There will be some loss in power, but we will be able to weigh the pros and cons of how much power we're losing, versus cost.

Impact of These Studies on Genetics Research

Yap: How will these studies impact genetic research in academic labs in the future?

Todd: Going forward, I think we will have major genotyping and sequencing pipelines in institutes such as the WTSI. We will make absolutely sure to release genotype data to researchers as quickly as possible.

I think this will create a whole industry of biology focused on a reduced part of the genome and its association with common disease. Some of these biological experiments have to be done in small labs by dedicated, small groups of people.

These early results will have a huge impact. This is the beginning of being able to say, 'Here's the human genome. These bits of it are important in disease and have functional variants.' We need to get that data out to the world. Then, I think hundreds of labs will be conducting genetic and functional studies. When a few susceptibility genes get discovered and published, there is always a massive amount of activity by a variety of labs, following up the biology and genetics of newly-discovered genes. I think dozens of genes will be discovered in the next few years. There are going to be hundreds of biologists looking at those genes and saying, 'Hmm, I know what that gene does' and providing a leap in our understanding of disease mechanisms. So for me, it's a question of how can we get good quality information out as fast as possible, right now?

FOR MORE INFORMATION

Contacts

- John Todd, Ph.D.
University Chair
JDRF/WT DIL
Cambridge Institute for Medical Research
Wellcome Trust/MRC Building
Addenbrooke's Hospital
Hills Road
Cambridge
CB2 2XY
United Kingdom
john.todd@cimr.cam.ac.uk

Organizations

- Cambridge Institute for Medical Research – http://www.cimr.cam.ac.uk
- Wellcome Trust Sanger Institute- http://www.sanger.ac.uk/
- International HapMap Project – http://www.hapmap.org
- Wellcome Trust (WT) Case-Control Consortium – Audrey Duncanson, WT, London; Peter Donnelly, Department of Statistics, Oxford

Further Reading

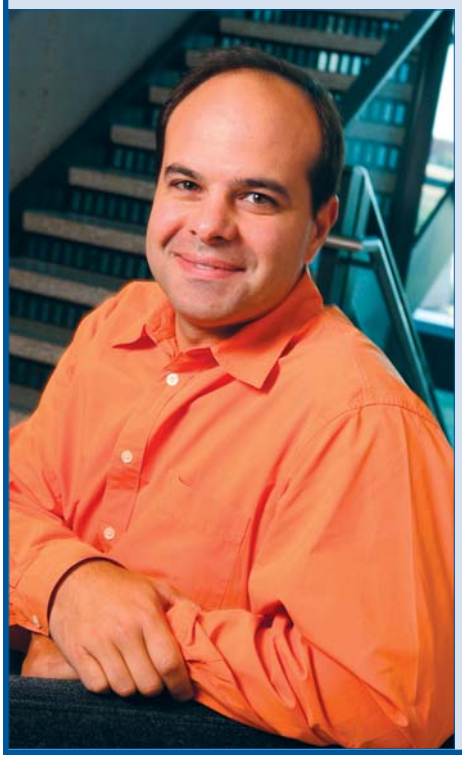
- Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*. 2005; 6(2):109-18. Review.
- The International HapMap Project. The International HapMap Consortium. *Nature*. 2003 Dec 18;426(6968):789-96.
- Burren OS, Healy BC, Lam AC, Schuilenburg H, Dolman GE, Everett VH, Laneri D, Nutland S, Rance HE, Payne F, Smyth D, Lowe C, Barratt BJ, Twells RC, Rainbow DB, Wicker LS, Todd JA, Walker NM, Smink LJ. Development of an integrated genome informatics, data management and workflow infrastructure: a toolbox for the study of complex disease genetics. *Hum Genomics*. 2004; 1(2):98-109.

AMB Featured Research: Association Studies

Researchers Combine Linkage Analysis and Whole-genome Association Studies to Investigate Autism and Alzheimer's disease

Researchers studying the genetic roots of autism at the Translational Genomics Research Institute (TGen) in Phoenix, Ariz., recently completed linkage analysis of 7,500 samples from 1,500 families using the Affymetrix 10K microarray and are now performing whole-genome association studies on a 1,000-sample subset with the 500K microarray. Their work is part of the National Alliance for Autism Research (NAAR) Autism Genome Project, a worldwide collaboration involving 43 institutions in the largest genetic study of autism ever conducted.

A multitude of other complex genetic disorders are also being deciphered at TGen using the 500K array. They include Alzheimer's disease, multiple sclerosis, progressive supranuclear palsy, memory defects, and bipolar disease. Dietrich Stephan, director of the Neurogenomics Division at TGen, spoke with *AMB* Editor Rachel Shreter about combining newly-published HapMap data with 500K genotype data to complete the analysis of their whole-genome association studies and pinpoint the disease genes.



Shreter: Could you describe the autism association study you are conducting?

Stephan: We're pinpointing new regions of the genome which harbor autism-susceptibility genes. Using a multi-pronged approach of performing linkage analyses in a huge set of multiplex pedigrees, as well as a high-density, whole-genome association study, we hope to arrive at a set of allelic variants of genes that cause a heritable predisposition to the trait. We know there are probably multiple genetic lesions that either alone, or in combination, predispose someone to autism. In order to get around the issue of genetic heterogeneity, we've done a huge linkage study, powered such that subgroups that are caused by specific individual genetic lesions fall out. This is a collaboration that was spearheaded by a group called NAAR, but really encompassed the entire world's collection of autism families—1,500 families—all of which had more than two individuals affected. We then took one proband from each of those families that we obtained through linkage analysis using the 10K chip, and did whole-genome association using the 500K chip. The data from the linkage study and the association study will eventually be integrated.

Whole-genome association studies using the 500K platform—or any high-density SNP genotyping technology—are going to require a lot of samples, probably more samples than any one person can collect. We've had to turn, in every case, to large national or international clinical consortia. For the autism study, we're using samples from the entire Rutger's sample repository, and nine other repositories. For Alzheimer's disease, we drew on the national network of 30 Alzheimer's disease centers, each of which has a brain bank that's been collecting samples for the past 15 years. In addition, we obtained samples from the Netherlands, a German brain bank, and a Swiss sample cohort. Collectively, we have 1,000 cases and 1,000 controls meeting the clinical and histopathologic criteria that we set forth. Granular and accurate phenotyping are crucial. So that's the same scenario that we've seen over and over again across disease areas—scientists banding together to find alleles of genes for common human diseases. The key to these things, at least on the screening side, is to get enough samples so that the study is powered correctly and you can get a real answer out of your training set.

We're also using pooling-based approaches that yield allelic imbalances of more than five percent within months, as opposed to years. Of course, it becomes impossible to shuffle the comparison cohorts to find, for example, associations with multiple clinical parameters. So rather than do linkage in pedigrees and watch how things are inherited, we're just going to see if there is an allelic imbalance between hundreds of cases, versus hundreds of controls at any locus.

Shreter: What about the analysis of these data?

Stephan: The analysis portion of the project is something that the whole field is wading through right now. That's where the HapMap comes into play. We're in the middle of trying to overlay all of the haplotype blocks out of the HapMap project over the 500K SNP map to see which SNPs define which blocks, to determine the coverage that we are getting with a 500K chip, and whether there is redundant information. So if one SNP is sitting next to another one in the genome, and they're both significant, you'd throw one out when you're doing your statistical calculations to reduce false positives. Additionally, the HapMap overlay will allow us to identify regions of the genome that are not covered with a specific genotyping panel and help identify regions which potentially contain false negative signals. Also, it's very difficult to do multipoint linkage analysis, because the SNPs are so close together. If one has a positive logarithmic odds [LOD] score, the one next to it—even if it's 200 kilobases away—is going to have a positive LOD score. You're going to get these massive multipoint LOD peaks that don't make sense with respect to sparsely spaced, relatively independently segregating microsatellite marker

scans. The HapMap project is going to show that those SNPs always move together and you should use either one or the other, but not both, and hopefully eliminate artificially-inflated multipoint LOD scores.

With the analysis hurdle solved by the HapMap, the other consideration is that you have to validate your data in a second population or a second sample set. We generally find that from one of these big studies we get several associated intervals that are each between 30,000 and 100,000 bases long and they usually only have between one and three genes in them.

Now the question is: which of those genes causes your predisposition to Alzheimer's or autism? And then, which base change within that real estate is the causative base change?

That's something that's going to be a little tricky to determine. The only sure way to prove that base causes the predisposition is to test to see if it causes the disease in tissue culture, or in an animal model, and some of these variables have to act in combination with other loci across the genome. The worst case scenario is that you may have to make one animal, with three different gene variants in it, from three different genes, in order to recapitulate that human predisposition.

The genetics have been revolutionized with this chip, and the data analysis has been revolutionized with the HapMap project. Now it's a matter of proving causality of any variant within an associated region of the genome.

Further Information:

- National Alliance for Autism Research Launches Largest Study of Autism Genetics (Press Release, July 19, 2004) - http://www.naar.org/news/render_pr.asp?intNewsItemID=176
- The Autism Genome Project: goals and strategies (Am J Pharmacogenomics, 2005) - http://pt.wkhealth.com/pt/re/pcg/abstract.00129785-200505040-00004.htm

Don't miss these in-depth interviews in recent issues of AMB. Subscribe today.

Quarterly October 2005 Issue

Dana-Farber Cancer Institute's Myles Brown on the roles of estrogen receptor in breast cancer and androgen receptor in prostate cancer.



"People have made large conclusions studying a very small number of genes. For any given transcription factor, only a few targets have been studied in any detail."

John Hopkins' Joseph Califano discusses the link between mitochondrial mutations and oral cancer.



"It requires interesting mental gymnastics to figure out what results in increased selection of mitochondrial proliferation as opposed to cellular proliferation."

Special Stem Cell Issue

"Everybody is hoping to use these cells in some therapeutic protocol one day, so we need to know exactly what the cells are, what kinds of genetic changes they have, and whether those changes make any biological difference or not."



Johns Hopkins Medical Center's Aravinda Chakravarti on finding mutations in federally approved embryonic stem cell lines.

"Our recent results are just the tip of the iceberg. There are many, many, many more molecules playing interesting roles in stem cells."



Princeton's Ihor Lemischka on using microarrays in embryonic stem cell research.

www.microarraybulletin.com

Request your complimentary copy of the Quarterly October 2005 or the Special Stem Cell issue at reprints@microarray.com

UPCOMING INTERVIEWS

Michael Zwick, Ph.D. and Timothy Read, Ph.D.
Emory University and the Naval Medical Research Center

Dr. Zwick and Dr. Read discuss the use of anthrax resequencing microarrays for biodefense pathogen detection

Eric Blomme, D.V.M., Ph.D. and Heidrun Ellinger-Ziegelbauer, Ph.D.
Abbott Laboratories Inc. and Bayer Healthcare AG Inc.

Dr. Blomme and Dr. Ellinger-Ziegelbauer discuss toxicogenomics in the drug discovery and development process

Lajos Pusztai, M.D., Ph.D. and Thomas Goralski, Ph.D.
MD Anderson Cancer Center and Expression Analysis Inc.

Dr. Pusztai and Dr. Goralski discuss the development and use of gene expression profiles to diagnose breast cancer and to predict chemotherapy response

Elizabeth Winzeler, Ph.D. and Celine Carret, Ph.D.
The Scripps Research Institute and the Wellcome Trust Sanger Institute

Dr. Winzeler and Dr. Carret discuss the design and use of microarrays to characterize the life cycle of the malaria parasite, *Plasmodium falciparum*

To receive these upcoming interviews by email or RSS news feed subscribe at ***www.microarraybulletin.com***

UPCOMING SYMPOSIA



Peter J. van der Spek, Ph.D.
Erasmus MC, The Netherlands

Dr. van der Spek will talk about the ways that genomic technologies are changing the view of molecular pathways from static snapshots to dynamic networks



David Wong, D.M.D., D.M.Sc.
University of California, Los Angeles

Dr. Wong will discuss using expression profiles of the salivary transcriptome as diagnostics for major human diseases, including oral squamous cell carcinoma



Linda Strausbaugh, Ph.D.
University of Connecticut

Dr. Strausbaugh will discuss her research investigating the suitability of high-density mitochondrial resequencing arrays in forensic applications



Anindya Dutta, M.D., Ph.D.
University of Virginia

Dr. Dutta will talk about his discovery of pan S-phase replication in cancer cells using high-resolution tiling microarrays for unbiased mapping of the human genome



Hubertus (Bert) J.M. Smeets, Ph.D.
Maastricht University

Dr. Smeets will present a "bedside to genome to bedside" approach to studying and treating mitochondrial disease

For more information on dates and times for AMB Symposia, go to "Upcoming Events" at ***www.microarraybulletin.com***

INFORMATION ABOUT AMB

WHAT'S THE *AFFYMETRIX MICROARRAY BULLETIN* (*AMB*)?

AMB is a quarterly magazine published by Affymetrix that highlights leading microarray research and diagnostic developments around the world. *AMB* features interviews by and for scientists, statisticians and software engineers—one expert interviewing another. Our goal is to create a forum where this community can discuss research and share new resources and information.

WHAT'S THE *AMB* COMMUNITY?

AMB Community is the online home of the *Affymetrix Microarray Bulletin* (**www.microarraybulletin.com**). The site features new stories that have not made it to print yet and individual microarray research communities devoted to areas like association studies and transcriptome analysis. The communities feature all of the relevant *AMB* interviews, and will soon feature daily updates of all papers published in that field. The online site also links to *AMB* quarterly issues and special issues.

WHAT'S THE *AMB* SYMPOSIUM SERIES?

AMB interviews are read by thousands of people in the scientific community and often raise many questions, which readers don't get the opportunity to ask. To help foster dialog between scientists, *Affymetrix Microarray Bulletin* has launched the *AMB* Symposium Series. These dial-in conference calls give scientists an opportunity to present their work—perhaps published just days before—and take questions from the scientific community. The *AMB* calls are very similar to traditional scientific conference symposia, offering a great opportunity for people working in the same field to share their methods, tools and insights.

HOW DO I GET *AMB* INTERVIEWS ADDED TO MY NEWS SITE (YAHOO!, MSN, ETC.)?

You can have *AMB* interviews, information about *AMB* events, and abstracts from new papers in your field sent directly to your news reader (such as Yahoo! or MSN). One method is to simply click on the RSS, Yahoo! or MSN icon in the top right corner of the *AMB* Community site. To learn more about creating your own news feed, click on the "about RSS" icon, also at the top right corner of the site.

HOW DO I SUBSCRIBE TO *AMB*?

Go to **www.microarraybulletin.com** to order your free annual *AMB* subscription (includes four *AMB* Quarterly issues and two special editions), or sign up for your free email subscription to the quarterly issues, special issues and online interviews.