

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

By Rachel Shreter

MOUNTAIN VIEW, October 27, 2005

— 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."

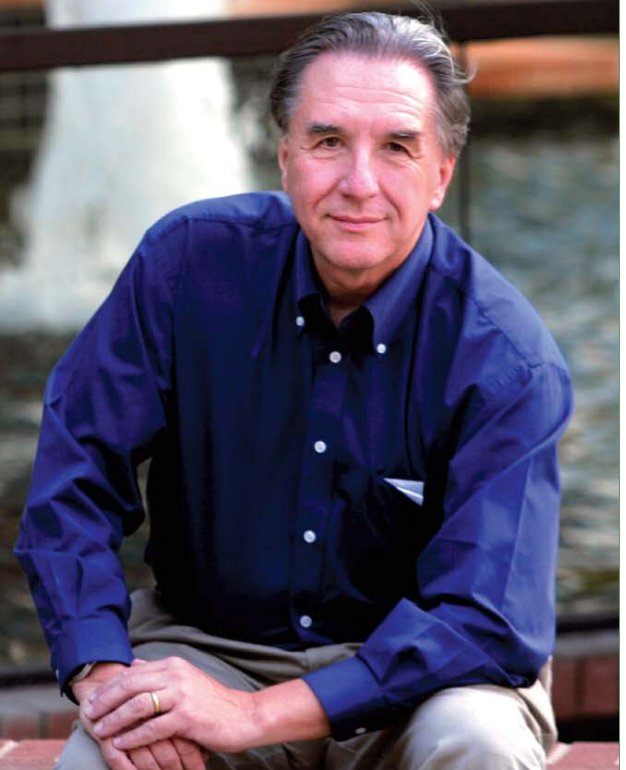
Perlegen used the same Affymetrix whole-wafer technology that had yielded 1.6 million SNPs in its initial 2002 haplotype study. That study was based on a set of 71 Americans of European, African, and Chinese ancestry (Hinds, *Science* 2005), and the data from that study was released to the public in 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

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.



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 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

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

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



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].

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.

Craig: Originally Perlegen used custom Affymetrix microarrays for genotyping. Affymetrix is now producing arrays to genotype 500,000 SNPs and Perlegen is starting to use these standard arrays in its 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.

AFFYMETRIX MICROARRAY BULLETIN

Editorial Staff

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*
Daniel Noble, *Copy Editor*
Michelle Majewski, *Contributing Designer*

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, Kershensobich 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-resolution scanning of human chromosome 21. *Science*. 2001; Nov 23;294(5547):1719-23.