AFFYMETRIX
MICROARRAY
BULLETIN

PREMIER ISSUE: ADVANCES IN DNA ANALYSIS
COPY NUMBER ■ ASSOCIATION STUDIES ■ BIOINFORMATICS

www.microarraybulletin.com

April 2005

IN THIS ISSUE

DNA Analysis
University of Tokyo's Hiroyuki Aburatani discusses his new Genomic Imbalance Map algorithm for analyzing data from 100K genotyping arrays in copy number studies

Jurilab's Jukka Salonen details one of the first whole-genome association studies ever undertaken—finding 700 genes associated with coronary heart disease, hypertension and diabetes with 100,000-SNP microarrays

Bioinformatics
Affymetrix' Jing Huang and Earl Hubbell talk about copy number algorithm design and ways to increase detection limits for chromosomal amplifications and deletions

RNA Analysis
Dana Farber's John Quackenbush and Yale's Paul Bertone discuss new tools and methodologies for analyzing the flood of data from new whole-genome transcript mapping studies

Special Section: Standards in the lab
The CDC's Joe Boone on standardization and harmonization initiatives and developing a quality systems approach to laboratory tests

Affymetrix Presentations at HGM 2005
Using high-density microarrays for population genetics studies ■ Developing a microarray genotyping assay for 500,000 SNPs ■ Automating microarray workflow with 96-array plates ■ Transcriptome mapping of the genome at 5 bp resolution ■ Tracking cell lines with 10K arrays ■ Genetic screening and public health management

SECTION 1: DNA ANALYSIS

Serono Identifies 80 Genes Involved in Multiple Sclerosis Using 100,000 SNPs

Dr. Daniel Cohen discusses whole-genome association study to find new genes involved in disease and new drug targets

Dr. Daniel Cohen

On March 17th Daniel Cohen announced the results of a whole genome SNP scan for multiple sclerosis, an experiment he's been waiting to do since he first thought of it over 20 years ago. Cohen said he remembers when he first came up with the idea, sitting with a colleague in his office, trying to figure out how they could possibly generate enough markers to do the experiment they were both imagining.

"We said we need markers for association, not linkage, because that is more precise," said Cohen, who went on to co-found the Human Polymorphism Study Center (CEPH) with Nobel Laureate Jean Dausset in 1983. "Our plan was to use linkage disequilibrium in the genome—to us it was obvious that we should find many markers across the genome— *cont'd on p. 2*

500,000-SNP Study Scans Thousands Of Japanese Patients for Genes Associated with Graft-Versus-Host Disease

GVHD experts Dr. Seishi Ogawa and Dr. Takanori Teshima discuss the challenges of whole-genome association studies

Graft-versus-host disease (GVHD) is one of the greatest risks patients undergoing allogenic hematopoietic stem cell transplantation face; it results in serious and sometimes fatal complications in the approximately 30 percent of patients who reject the transplant.

Traditionally, doctors have relied on methods to match major human leukocyte antigens (HLA) in patients and donors to determine compatibility, which work in most cases. However, differences in minor antigens—the number and identity of which is unclear—often generate histocompatibility complications that result in GVHD.

Dr. Seishi Ogawa, associate professor, Graduate School of Medicine at the University of Tokyo, director, Core Research for Evolutional Science and Technology (CREST) Program of the Japan Science and Technology agency (JST), thinks the most practical way to discover the minor antigens involved *cont'd on p. 5*

I can tell you that with 100,000 SNPs, I found a lot of things. With 500,000 SNPs, I will find many more things! That is the point. People want to find everything, but they should start by finding something.

but at that time it was impossible, financially, to generate 100,000 markers. So we decided to focus on linkage, using only a few thousand markers—an experiment we could do at the time."

Over twenty years later, Cohen's vision became reality—he and a team of scientists from the Serono Genetics Institute conducted an multiple sclerosis (MS) association study scanning 1,800 people for 100,000 SNPs each. The result was the identification of 80 genes involved in the inflammatory and neuro-degenerative pathways of MS.

Cohen discussed this experiment and the challenges of association studies during a break at Serono Symposia International's *Genetics of Autoimmune Diseases* conference, which was held in Cannes March 18th - 20th. Cohen was interviewed by Greg Yap, vice president of DNA products at Affymetrix, who was instrumental in development of the 100K SNP arrays used in the study.

Yap: How many samples did you study?

Cohen: We generated 300 million genotypes from 3,000 samples—two-thirds of them were MS and the rest were lupus—using 100,000 SNPs. We had 900 cases for MS, plus 900 controls, and 600 cases for lupus, plus 600 controls. Controls were matched for quality, gender and geography.

Yap: Why did you choose MS and lupus as the first diseases to study?

Cohen: MS is a disease in which Serono is a leader and which I also have a personal interest. This was the first disease I worked on as a student, so it was good to come back to it. MS is a very bad disease that affects one person out of every thousand. There are drugs like Rebif that help, but there is a need for new good drugs.

Yap: How well understood are the genetics that lead to MS?

Cohen: It is well known through previous studies using twins that there is a hereditary component to MS: twin risks are 30 to 40 percent (i.e., if one twin has MS, the other twin has a 30 to 40 percent risk of MS). To give you an idea, schizophrenia's twin risk is 50 percent. So there is a strong genetic component and a strong environmental component. You can calculate all the data published on the genetics of MS in terms of genes that are found. There are many reports, but they are not all validated. Let's say there are about seven genes known for sure, and this includes one major one: HLA, which plays a central role in immune function.

Yap: HLA is interesting in that it has been implicated in many diseases but no one seems to know how and why it acts in any disease, including MS. Why is this?

Cohen: Nobody knows. HLA has been known for 35 years, but nobody knows how it acts in MS. One problem is that there is a lot of LD in HLA, so the association to HLA is with a very large region. There are maybe 200 genes there!

Yap: How did you design your study?

Cohen: In our scan for MS, one thing we did was to look for association in three different populations for more robustness. When you find genes giving associations in three populations they are more robust. You get two types of results. One rather frequent situation is there are no genes in the region of association. You call this region a "desert."

Yap: If you find an association in a place you thought was a gene desert, does that mean there is something in the desert?

Cohen: Yes, especially when you replicate the association three times! We have experienced this in the past, and in the past we looked carefully at this region by generating one PCR amplicon every kilobase (kb) across 500 kb and screening cDNA very carefully. This is very time consuming and expensive, but we have found not one gene, but several genes. If you take any desert in the genome, and there are many deserts, you find new genes!

This is terribly hard, and this why we want to use tiling arrays instead.

Yap: Are these truly new genes?

Cohen: Yes, with good cDNA, ATG sites—everything. They are just not conserved, and so the computer is not able to predict them. The homology across species is very small and so the computer fails. We have found many things like this. The trend is to think that genes for mouse and man are identical, and some are, but you don't know how many are different. People should not be dogmatic about the current belief in 30,000 genes. There are more rather than less. There is a very good paper from Tom Gingeras [vice president, biological science at Affymetrix Research Laboratories] using Affymetrix tiling arrays suggesting that there are twice as many genes just on chromosome 21 as expected. If you follow this [premise] across the genome, you go from 30,000 to 60,000 genes, and that is probably only for a few tissues in the body, not every tissue. I would not be surprised if we really have 100,000 or more true genes. Eventually the computers will become more clever and pick these up.

Yap: Do these genes make proteins?

Cohen: Show me! This is fine, but you should have to prove that there is no protein made. Then I will be convinced. It could be a protein-coding gene, but you need to prove it either way. I have no problem with RNA that is not coding. But the attitude that any new gene you find today cannot code for a protein—it is amazing! This is where the tiling arrays really work to show transcription in these gene deserts. I don't think people get it. They don't believe it yet.

Yap: What fraction of associations do you find in desert regions?

Cohen: 10 to 20 percent. For other associations that we find in all populations, sometimes the situation is very clear, because there is only one gene. But many times you find in the region of association two genes that are overlapping, one in one strand and one in the other. So in the end you cannot make a conclusion, you have to do more experiments to determine which gene is truly associated.

Yap: I thought one of the most interesting aspects of your recent results was your comment about building a comprehensive catalog of genes involved in MS. As you put more pieces in the puzzle, I'm assuming it becomes easier to fit in or reject additional candidate pieces; is that correct?

People should not be dogmatic about the current belief in 30,000 genes. There are more rather than less. I would not be surprised if we really have 100,000 or more true genes. Eventually the computers will become more clever and pick these up.

Cohen: Yes, that's what we have started to do with MS. We selected our best associations and some were coherent with the nature of the disease. This disease is autoimmune, so some of these genes are known "immune genes," but not all of them. So you take 80 genes and you ask which genes are known to be involved in immune response, like HLA, but there are many others. Some are known inflammatory genes, and that is good. We know 2,000 inflammatory genes, but we don't know which ones to study. Now we know there are about five or 10 that may be a sort of bottleneck of the disease. So now we know which ones to study: instead of 2,000, it is only five. The second good thing is that some are not inflammatory genes. You find more genes than expected involved in neurodegeneration and in very clear pathways. Many of these genes are involved for example in Alzheimer's or schizophrenia.

Yap: What stage are you at in the process of analyzing your results from the initial set of genes you discovered?

Cohen: You have to read the literature—80 genes takes a lot of work! We are doing more sophisticated statistical analysis on these genes. Some types of analyses require two days of very powerful computer work per gene, so you cannot do this for every gene you find. You cannot do this on the whole genome. Even 80 genes take a lot of time, and we have very powerful computers. It takes time to build pathways with the genes that are found. Some are already in pathways, but not all, so you have to rebuild the pathways. I don't want to elaborate too much, as some of these are trade secrets, but there are many methods to build such pathways. Now with 100,000 SNPs, we found 80 genes, and with 500,000 SNPs, we will find more genes, but this is not a problem. Let's say you find 200 genes, but these are not 200 independent problems, you can organize them into five to 10 pathways. It's not that complex.

Yap: What was magical about 100,000 SNPs?

Cohen: With this level of SNPs, you don't need to make any prior assumptions and hypotheses about associations. You start from scratch that is the most important thing. When you use other methods, you make a hypothesis and you look in one place for association. Sometimes there will be one gene, sometimes two, in the same region. Sometime there will be five genes, sometimes no gene at all. The magic of whole-genome scans is that you can look at the whole genome for all the genes—you don't need to know exactly where to look. It was a shock when Affymetrix announced that it would provide 100,000 SNPs. We had calculated we would need at least a few million genotypes over 10 years and

with this we could now do 200 million— a factor of 100 more. We could do two logs more within one year!

Yap: How many samples do you need to do this type of study? Are you saying that 900 cases and 900 controls were enough to look at MS in general, but not enough to study all the different subtypes of the disease?

Cohen: This is a good question. There are a number of genes to be found in these studies, and some have a strong effect and then there is a continuum of others with weaker effects. Your sample size gives you the threshold: the larger the effect you are looking for, the smaller the sample size you need. But weaker effects are also very important and you need more samples to find those. If you ask me what size sample to use for any experiment, I don't know. The more you do, the more you find, it is that simple.

Yap: In terms of number of SNPs, how do you see the value of 500,000 SNPs versus the 100,000 you are already using?

Cohen: There is a debate as to whether you want to use more and more random SNPs, which is basically what we do, or whether we need to use carefully selected SNPs. I would put this a different way: my idea 10 years ago was to use 50,000 selected SNPs, very frequent SNPs that are well distributed at one every 50 to 60 kb, but my point then and now, is not to find all genes, but to find what I can, enough to assemble the pathways and find drug targets. With 50,000 SNPs using haplotype analysis you can find some genes. If you use random SNPs, you probably need more SNPs, and I have found that 100,000 SNPs corresponds with 30 to 50 percent genome coverage in our scans. That's the point that people don't understand: the more you add, the better. But you can already find genes today, and you can also keep adding SNPs. How important is being complete? What Lon Cardon [from Wellcome Trust Center of Human Genetics in Oxford] says is very true: we don't know how many SNPs you need for a complete scan. Is 10 million enough? We don't know. But I can tell you that with 100,000 SNPs, I found a lot of things! With 500,000 SNPs, I will find many more things! That is the point. People want to find everything, but they should start by finding something.

Yap: [laughs] That is a very pragmatic approach!

Cohen: This is why we wanted to get started with 100K and we found many things!

FOR MORE INFORMATION

Contact Information

■ **Daniel Cohen, M.D., Ph.D.**
Serono Genetics Institute S.A.
Vice-President, Worldwide Head of Genetics
Serono Genetics Institute
Route Nationale 7
91030 Evry Cedex, France
Phone : +33 1 60 79 66 66
Fax : +33 1 60 79 66 40

■ **Greg Yap**
Affymetrix Inc.
Vice President of DNA Products
3380 Central Expressway
Santa Clara, CA 95051
Phone: 408-731-5000
greg_yap@affymetrix.com

Companies:

■ **Serono Genetics Institute S.A.** http://www.genetics.serono.com/index.jsp

■ **Affymetrix Inc.** http://www.affymetrix.com/

Conferences:

■ Serono Symposia International's "Genetics of Autoimmune Diseases" https://www.seronosymposia.org/ms/event_descrip.ihtml?id=242

Further Reading

■ Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G and others. Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. *Science* 2005.

■ Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 2002;296(5569):916-9.

People:

■ **Lon R. Cardon, Ph.D.**
Professor of Bioinformatics at Wellcome Trust Center of Human Genetics http://bioinformatics.well.ox.ac.uk/~lon/people.html

■ **Thomas Gingeras, Ph.D.**
Vice President for Biological Research at Affymetrix
http://www.affymetrix.com/transcriptome/index.affx

Background

■ Serono Symposia International Learning Center on Multiple Sclerosis
http://www.seronosymposia.org/ms/home.ihtml

Dr. Seishi Ogawa is an associate professor in the Department of Regeneration Medicine for Hematopoiesis, University of Tokyo, Graduate School of Medicine. He received his M.D. from the University of Tokyo in 1988 and his Ph.D from the same institution in 1994. Over the next decade, Dr. Ogawa focused his research on the blood-borne cancer, leukemia. He discovered a number of genetic alterations that contribute to hematopoietic malignancies, including inactivation of the p16 gene in leukemias and genetic alterations in t(1;7)(q10;p10) in myeloid dysplasia. Ogawa's early genetics leukemia research naturally led to his current study searching for the genetic variations associated with graft-versus-host disease—a common consequence of bone marrow transplant therapy for leukemia patients.

in GVHD may be to use high-density genotyping microarrays. Ogawa is performing a 2000-patient case-control association study with an early version of the Affymetrix 500K array to genotype over 500,000 SNPs from each participant.

Traditionally, scientists have searched for minor antigens by correlating incidence of GVHD with polymorphisms in genes suspected to be minor antigens. This candidate-gene approach requires prior knowledge of what genes are thought to be involved in GVHD—often an impossible undertaking. An alternative approach is to identify molecules that recognize T-cell lines established from GVHD patients. Ogawa points out that these molecules often cannot be clinically proven to contribute to GVHD.

Ogawa recently spoke with Dr. Takanori Teshima, associate professor at Kyushu University and world-reknowned GVHD expert, who has published nearly 100 articles focused on bone-marrow transplantantion and GVHD development. Most recently, Teshima and his laboratory discovered an natural killer T-cells (NKT) signaling pathway that regulates GVHD that may provide new targets to treat the disease.

Study Design: Finding Genes Associated with GVHD

Teshima: How does your method of antigen discovery differ from other candidate-gene types of studies?

Ogawa: Our current approach is a genome-wide association analysis, where we define compatibility between each donor/recipient combination according to the genotypes of each SNP marker, and then conduct association tests in a case-control design. We analyze the correlation between random SNPs and GVHD status. Of course, if a random SNP is not associated with a minor antigen, then this is meaningless. However, if it happens to be located very close to an unidentified minor antigen, the r^2 value will be sufficiently large, and we can show a statistical correlation between the marker and GVHD.

In order for a high r^2 value to be expected between an unknown target and a marker SNP, a very large number of SNPs need to be genotyped. This has been one of the problems associated with the genetic epidemiological approach.

Teshima: How many markers will you need?

Ogawa: Generally speaking, this is thought to require between several hundred thousand to a million markers. In our study, we are using the GeneChip® system developed by Affymetrix which allows for genotyping of 500,000 SNPs at an extremely high throughput and a reasonable cost. Using these arrays, we genotype as many as 1000 donor/recipients in GVHD positive transplants and 1,000 cases without GVHD. We then perform a genome-wide association analysis for the genetic factors responsible for GVHD, not only minor antigens but other genetic factors that the donors and recipients have.

Teshima: In your current study, you have included a variety of diseases. For example, I imagine acute myeloid leukemia alone will make up quite a number of cases. I think you can also study genetic predisposition to these diseases, in addition to the analysis of GVHD. Is that your intention?

Ogawa: It's a good point. Since this is a large epidemiological study in which 500,000 SNPs are genotyped for as many as 4000 samples at a great cost, I think our study has a favorable design in terms of its cost-effectiveness. Our control group contains 2000 normal volunteers who passed a strict health check before marrow donation on one hand and the case group consists of a number of hematological disorders including leukemias, lymphomas, and aplastic anemia patients. This means we can investigate a genetic predisposition to these disorders also using the same genotyping data. In addition, when comparing relapse and non-relapse cases or those who suffered from severe adverse reactions to the preparative transplant regimens and those who did not, we could analyze tumor antigens that mediate graft versus leukemia effects or development of those adverse reactions.

Sample Population

Teshima: I am struck by the fact you mentioned that the healthy donors have to pass rigorous screening tests. So your study allows you to acquire SNP data on a very healthy population?

Ogawa: Yes, you are right. Our study will analyze the largest number of polymorphisms from the largest number of samples from healthy subjects so far. I think this will be interesting data to analyze the genetic origin of the Japanese people.

Teshima: Where did the samples come from that you are analyzing?

Ogawa: This study is made possible thanks to the Japan Marrow Donor

Program (JMDP) several thousand donor/recipient DNA samples and almost complete clinical information on these bone marrow transplantations.

Teshima: Many SNPs are identified in the Caucasian population. How do they relate to the Japanese population? Are there racial differences in SNPs?

Ogawa: Of course, there are many SNPs specific to certain races. However, compared to the total number of SNPs, this makes up a surprisingly small fraction. Among the 10 million SNPs with minor allele frequencies of more than one percent, 90 percent are common to all populations.

Although the distribution of the minor allele frequencies in the Japanese population is not reported for the 500K array we are using in our study, the SNPs with less than 10 percent minor allele frequencies account for about 20 percent of all the SNPs in 100K arrays. So, we think it is quite feasible to perform a genome-wide association analysis with these arrays.

Statistical Analysis

Teshima: By genotyping so many SNPs, even a very low error rate will generate many false positives. How did you control for this?

Ogawa: This is known as a problem of multiple analyses. In this type of association study, the possibility of having an association is excluded at a predetermined significance level. However, if we set this to 0.05, which is used for many clinical studies, and tested association for a total of 500,000 SNPs, approximately 25,000 become positive for association only by chance. In other words, setting the significance level like this results in a tremendous inflation of type I errors. In order to suppress the overall type I error rate to less than 0.05, the individual significance level needs to be set to one ten millionth using the most conservative correction which, needless to say, leads to an increased type II error rate. Because, according to the principles of statistics, it is impossible to reduce both type I and type II error rates without increasing sample numbers, these type of studies are usually performed in two steps. At the first step screening, the individual type I error rate is set higher, say 0.01. Then, in the second step, the positive SNPs are tested for another set of case-control groups. This manner of analysis could reduce the total number of genotypes with the sensitivity and specificity levels being maintained. In our studies, a similar kind of consideration must be taken to increase sensitivity and specificity of the analysis.

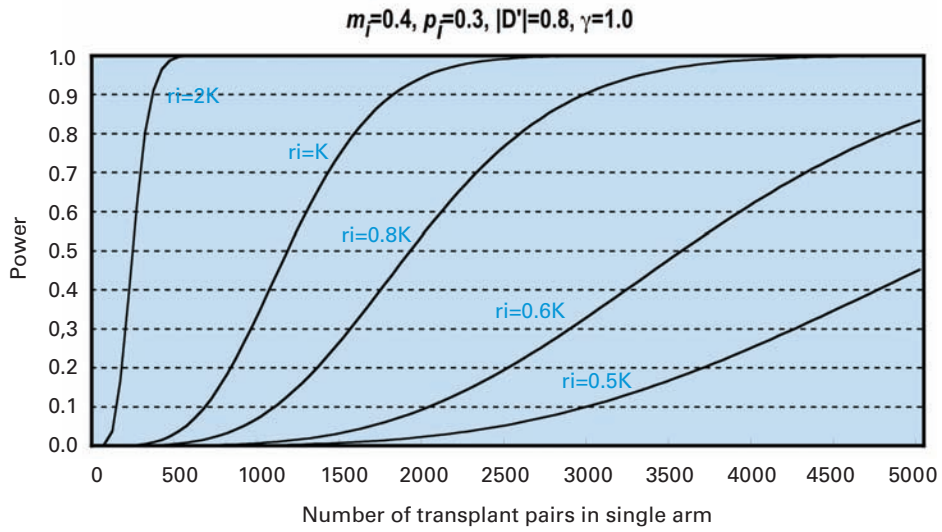


Figure 1. Power estimation for indirect association analysis for minor antigens using Bonferroni's correction for 500K SNP testing. K is the overall GVHD rate in the case-control group and ri denotes a probability of developing GVHD when the minor antigen is incompatible in a donor/recipient pair. Minor allele frequency (MAF) for minor antigens is assumed to be 0.3 and that for marker SNP is 0.4.

Teshima: With regard to the statistic analyses you employ in your study, what is the smallest relative risk you can detect?

Ogawa: It is very difficult to tell with this kind of analysis. We have to consider many factors. For example, the strength of linkage disequilibrium, frequency of the minor antigen allele, the difference between the relative frequencies of the marker allele and that of the minor antigen allele, and most importantly, how much risk is there to develop GVHD at the gene locus in question; all these parameters are not known prior to the study.

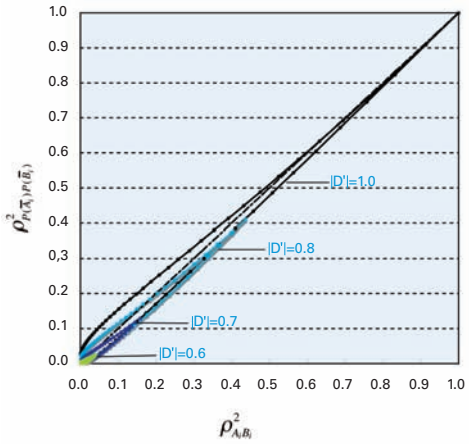
By using the 500K array, the average distance between a given SNP and a minor antigen is estimated to be less than several kilobases, so based on our knowledge of linkage disequilibrium among the human population, we expect on average, larger than 0.8 linkage disequilibrium (LD) values.

If we assume the LD value is relatively large, that the relative risk of the minor antigen mismatch is similar to known minor antigen mismatches—about 1.3 to 2.1—and that the minor antigen is not very rare, but it is a relatively common allele, we can estimate the power of our statistical analysis.

Under such assumptions, if we have a sample size of 500 to over 1000 cases per group, we currently think we can identify minor antigens or polymorphisms related to GVHD with a high probability.

Teshima: Studying GVHD is difficult because you have to take into account two individuals, and the possible mismatches that can result. Considering this, how does your study design and analysis differ from other association studies of common complex disease?

Ogawa: We have two scenarios here. The first is discovering associated polymorphisms that are independent from the donor/recipient combination. In this case, the polymorphism in either donor or recipient is sufficient to confer



risk for GVHD. This is essentially the same as other studies for common disease. For GVHD, polymorphisms in the IL-10 gene are an example of this.

We also have to take the combination of two people (donor/recipient) into account. As far as I know, I don't think there are many reports on the theoretical basis of such indirect association analysis. We are of course evaluating how much analytical power we can get out of this kind of study. As I mentioned earlier, the problem here is to define the allele combination as match or mismatch and based on this definition we have to analyze our data. This sets our study apart from previous ones.

At first thought, in this case, you might expect the minor antigen compatibility to be doubly obscured from the genotype of both donor and recipient. However, this is not always the case. In indirect association analysis, the kai square value for the test (D) at the marker SNP (M) is reduced by the factor of r^2 or $\rho^2_{D,M}$. In the association test for minor antigens, it is reduced by the correlation coefficient between compatibility of a minor antigen locus and SNP marker locus ($\rho^2_{P(\overline{mHA})\,P(\overline{M})}$), where N denotes number of samples to be tested for each locus. On the other hand, r^2 and correlation coefficient of compatibility both show good correlation with the usual definitions of compatibility. So, we've found taking an indirect approach results in only a mild reduction in power.

Teshima: What software are you using to analyze these associations?

Ogawa: There aren't many software programs available to perform this kind of analysis. Therefore, we think we must develop our own software program to perform our analyses. Of course we are going to take advantage of what is already available as much as we can.

Teshima: Related to the previous question, transplant patients have various disease states, going through various conditioning regimens and GVHD preventive regimens, so you need to match their background as much as possible. This might make it difficult to enroll a large enough number of patients. Taking this into consideration, how many samples do you think you need for your study?

Ogawa: Of course, in order to perform statistical analyses, the two groups must be comparable except for the GVHD status. Appropriate stratification is necessary, as you pointed out. So, we have to match at least some of the most important factors. I believe that at least the extent of HLA match and the GVHD prevention methods has to be the same between groups.

As we try to match more and more conditions, the number of samples we can use for analysis will become limited. We looked into the existing GVHD data in the Bone Marrow Bank. Based on this, we think we can secure between 1500 and 2000 cases, this is the number for both groups (GVHD and non-GVHD), even if we match background conditions.

Matching Donor/ Recipient Antigens

Teshima: Based on studies on mice, GVHD is caused not by the mismatch of a single minor antigen but by multiple antigens. Is it possible to analyze multiple complex combinations?

Ogawa: I think the general consensus today is that there is no single minor antigen that confers a significant risk for GVHD. This is essentially the same problem that has been discussed frequently in the context of so-called common diseases such as hypertension and diabetes, where each individual disease-related allele has only a minor effect.

It is quite difficult to analyze the correlation between combinations of gene polymorphisms and GVHD disease. I don't think there is a good model for a case control study like this one, that is, a study to analyze how certain combinations increase the risk of a disease.

Still, a case control study using patient data is the only way to prove whether a certain minor antigen mismatch is responsible for clinically significant GVHD. I do not think there is much clinical significance in identifying all such antigens and identifying all bad combinations. In reality, it is difficult to use this knowledge in choosing donors that pose no GVHD risk. Of course, we'll never know until we do this study.

Teshima: Lastly, minor antigens may be restricted by HLA, so is it necessary to pay attention to HLA subtype combinations?

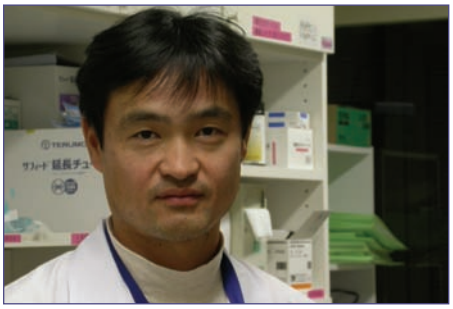
Ogawa: In order to analyze minor antigens, we need to take HLA restriction into account. By definition, minor antigens are recognized by certain HLA subtypes. Therefore, in order to perform a case control study, we may have to limit our correlation analysis to a certain HLA subtype combination. This of course will make it very difficult to collect a large number of samples. However, since we can test for associations with a number of HLA types, we could show associations of some minor antigens that are restricted by some HLAs.

Since Merikangas and Kruglyak pointed out the possibility in 1996, it is still debatable whether we can identify disease susceptibility genes by using the genome-wide association analysis. There is still no consensus among genetic epidemiologists, but we cannot deny that many researchers are banking on this possibility. The ultimate goal of the international HapMap project is to identify disease susceptibility

genes through genetic epidemiology. Genetic epidemiological approaches have proven to be enormously useful for identifying a large number of disease-genes inherited for Mendelian disorders. In the future, we hope to identify susceptibility genes for diseases including GVHD that are inherited in a more complex manner. We hope this whole-genome association study of GVHD will be the first of many such studies to follow.

Teshima: Thank you very much. Good luck with your projects.

NOTE: The previous interview was edited and condensed from a full-length interview on February 28.



Dr. Takanori Teshima

FOR MORE INFORMATION

Contact Information

■ **Seishi Ogawa, M.D., Ph.D.**
Department of Regeneration Medicine for Hematopoiesis
Graduate School of Medicine,
University of Tokyo Hongo 7-3-1,
Bunkyo-ku, Tokyo 113-8655, Japan
sogawa-tky@umin.ac.jp

■ **Takanori Teshima M.D., Ph.D.**
Kyushu University Hospital
Center for Cellular and Molecular Medicine
3-1-1 Maidashi, Higashi-ku,
Fukuoka 812-8582, Japan
tteshima@cancer.med.kyushu-u.ac.jp

Organizations

■ **The University of Tokyo**
http://www.u-tokyo.ac.jp/index_e.html

■ **Kyushu University, Japan**
http://www.kyushu-u.ac.jp/

■ **Japan Society and Technology Agency**
http://www.jst.go.jp/EN/

■ **Core Research for Evolutional Science and Technology (CREST)**
http://momo.tokyo.jst.go.jp/kisoken/crest/

■ **Japan Marrow Donor Program (JMDP)**
http://www.jmdp.or.jp/

Companies

■ **Affymetrix Inc.**
http://www.affymetrix.com/

Further Reading

■ Hashimoto D, Asakura S, Miyake S, Yamamura T, Van Kaer L, Liu C, Tanimoto M, Teshima T. Stimulation of host NKT cells by synthetic glycolipid regulates acute graft-versus-host disease by inducing Th2 polarization of donor T cells. *J Immunol* 2005;174(1):551-6.

■ Ichikawa M, Asai T, Chiba S, Kurokawa M, Ogawa S. Runx1/AML-1 ranks as a master regulator of adult hematopoiesis. *Cell Cycle.* 2004;3:722-724.

■ Tanaka Y, Kami M, Ogawa S, Machida U, Takahashi T, Ichikawa M, Yuji K, Izutsu K, Asai T, Kanda Y, Honda H, Mitani K, Chiba S, Hirai H, Yazaki Y, Sasaki M, Sasaki K, Mineishi S. Hyperacute graft-versus-host disease and NKT cells. *Am J Hematol.* 2000;63:60-61

New Copy Number Algorithm Developed at University of Tokyo

Dr. Hiroyuki Aburatani discusses the Genomic Imbalance Map software

One of the first researchers in Japan to use high-density expression microarrays in cancer research five years ago, has developed one of the first algorithms to analyze data generated by new high-density genotyping arrays in copy number and loss of heterozygosity (LOH) studies.

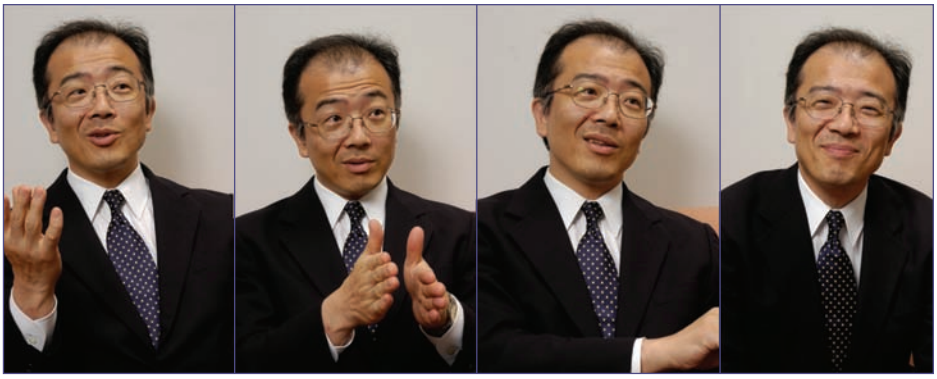
Professor Hiroyuki Aburatani, who runs a laboratory within the Research Center for Advanced Science and Technology at the University of Tokyo, is using a new Genomic Imbalance Map (GIM) algorithm in copy-number studies of liver and lung cancer. Aburatani and his colleague, Shumpei Ishikawa, report that the new algorithm reduces noise and can readily detect two to three copy changes with ease.

"I started studying cancer about 20 years ago in my postdoctoral work, but, we didn't have any high resolution technology to look at cancer cells; we could only estimate or predict what was happening," Aburatani said. "Now, we are using array technology for high-resolution studies of gene expression, gene copy number, LOH, and even for allelic variation with expression data."

Dr. Aburatani recently spoke with Dr. Kwong-Kwok Wong of Baylor College of Medicine about the development of the GIM algorithm and the challenges of copy number analysis. Dr. Wong studies copy number changes in pediatric gliomas using the 10K array.

Wong: What are your current methods of measuring genomic copy numbers?

Aburatani: We use the 100K genotyping microarray from Affymetrix. We started off with the 10K and initially used that for LOH analysis. Although we didn't expect it to be very accurate for copy number prediction because it's a short oligonucleotide array, it was both very accurate and very reproducible for copy number prediction. We are currently doing validation experiments using Fluorescence In Situ Hybridization (FISH).



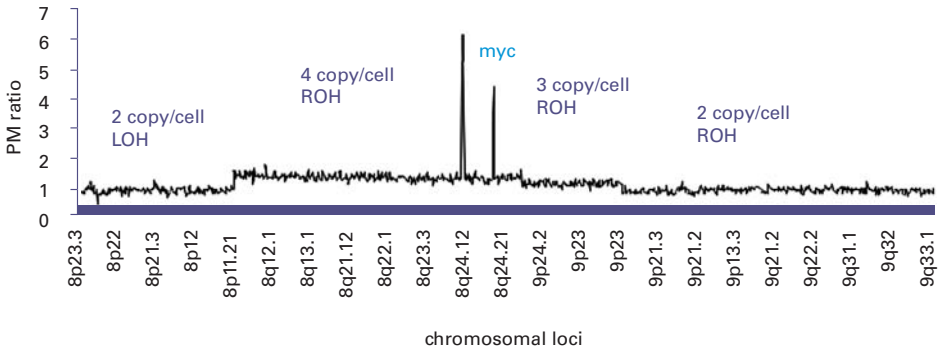
Dr. Hiroyuki Aburatani

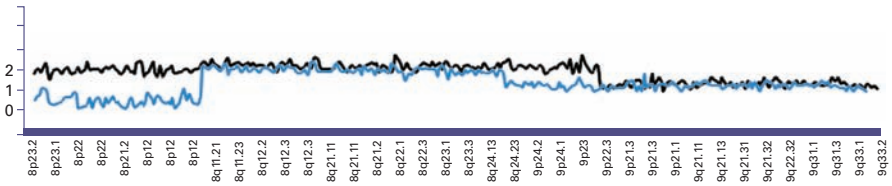
Wong: So when you do the copy number measurement, do you use the Affymetrix copy number tool?

Aburatani: No, we developed a new analysis tool called GIM, it's like the Expression Imbalance Map (EIM) tool we previously developed to detect genomic losses and gains by visualizing expression imbalance regions from microarray expression data. EIM gave us a much higher resolution than conventional technologies such as comparative genomic hybridization (CGH). With GIM, we measure copy number from either the total or allelic intensity, and through several steps of normalization and statistical procedures, we can accurately estimate copy number. Because we look at SNP genotype data, we can measure the signal intensity from each allele and that allows us to judge the LOH status or copy number increase or decrease very easily. Detecting two copies versus three copies is not a big problem.

Wong: What about cases where one allele is lost but the other allele is duplicated. In such cases, you cannot tell the copy number unless you separate the two alleles. Are you able to do that?

Aburatani: Yes, that's the powerful thing with the genotyping array because with a regular Bacterial





Artificial Chromosome (BAC) array or a long oligonucleotide array, you cannot see the LOH status—the cancer cells show the two copies just as normal. And if the normal copy was lost and the mutated copy remained—like what happens with uniparental disomy—the cancer cells already have a homozygous inactivation. Genotyping array analysis of cancer cells gives you the copy number and the allelic information together, so I think it's very powerful.

Wong: Have you used your EIM and your GIM tools together?

Aburatani: Yes. Compared to EIM, copy number analysis provides more data points because with the 100K array we have 100,000 data points scattered throughout the genome. With EIM, you can only measure copy number from genes expressed in particular tissues or cells. That typically gives us at most, 10,000 data points, maybe 50% of the genes.

EIM uses clusters of expression data from adjacent genes. But, in areas of the chromosome where genes are very sparse, you don't have a certain number of genes in the neighborhood, therefore you cannot create a good estimate of copy number.

Wong: Have you tried to evaluate the concordance between GIM and EIM estimations?

Aburatani: Overall correlation of copy number versus gene expression level is very good from our calculations. Of course, as we discussed already, gene density is a big issue. The GIM data covers much more of the genome than EIM data, so it is hard to compare for those regions. But, sometimes, we see discordance—false positive and false negatives. Some genes are regulated very specifically by other regulatory mechanisms regardless of the copy number change.

Wong: Is EIM available?

Aburatani: We published the program a couple of years ago in the journal of *Physiological Genomics,* so it is freely available.

One disadvantage with EIM, in regard to expression profiling, is that you need to compare it to normal expression patterns. With some types of tumors, it's more difficult to analyze normal tissue. For example, breast tumors originate from acinar or ductal cells. But most of the cells in normal breast tissue are fat cells and that makes it difficult to isolate and analyze normal acinar or ductal cells to make the comparison between cancer and normal expression levels.

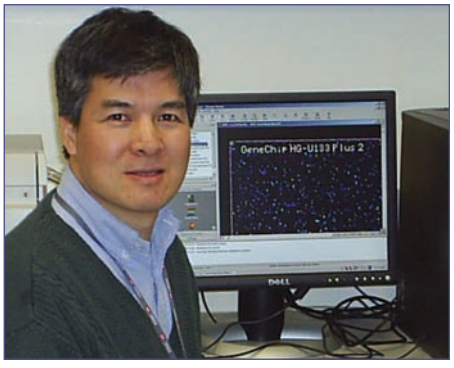
Wong: Typically in translational research we want to find biomarkers for prediction or early detection. Do you think copy number will be able to determine whether a patient has a high chance of progression or better prognosis?

Aburatani: We are analyzing liver cancer and lung cancer in our pilot research studies. Once we are able to say this chromosomal change is very important for cancer progression, we can correlate it with treatment response. To obtain such results that can be used for clinical diagnosis, we need to analyze at least a couple of hundred cases, so we are still collecting materials and of course we should run many more experiments. Because DNA is more stable than RNA, copy number analysis could be a good diagnostic tool.

Wong: I think we also need to know the fundamental polymorphisms in the general population. In Japan, for instance, you can genotype a lot of normal individuals so that you know the background of copy number polymorphism in the general public. Then you can develop a better algorithm with your GIM tool. Can you comment on that?

Aburatani: Copy number polymorphism research is a very new area. Mike Wigler's pioneering paper in *Genome Research* used 78K arrays from NimbleGen and, as the authors noted in their paper, an increase in the density of probe coverage would reveal additional copy number polymorphisms. The current Affymetrix typing system we are using is the 100K, but still the probes are 20 kb or 30 kb apart on average and not evenly spaced. So we are really looking forward to having the next generation of genotyping arrays from Affymetrix—like the 500K. Then, you will have 3 or 5 kb resolution. It can be very difficult to estimate copy number variation among normal individuals with one single probe, so if you can see similar signal changes in adjacent probes, then you can be confident that the copy number changed. If you only have a 20 kb or 30 kb resolution, probably you can detect no more than 100 kb genomic changes. But our genome may have smaller genomic duplications or deletions. So with the next generation, higher-density genotyping array, hopefully we will see more copy number variation throughout the human genome. At this point I don't think anybody really knows the relevance in predisposition of diseases.

NOTE: The previous interview was edited and condensed from a full-length interview conducted on March 2.



Dr. Kwong-Kwok Wong

FOR MORE INFORMATION

Contact Information

Hiroyuki Aburatani, M.D., Ph.D.
University of Tokyo
Professor, Genome Science Division Research Center for Advanced Science and Technology
4-6-1 Komaba, Meguro, Tokyo 153-8904
haburata-tky@umin.ac.jp

Kwong-Kwok Wong, Ph.D.
Assistant Professor, Department of Pediatrics
Baylor College of Medicine
6621 Fannin St., MC 3-3320
Houston, TX 77030
kkwong@bcm.tmc.edu

Companies

- **Affymetrix Inc.**
http://www.affymetrix.com/
- **NimbleGen Systems Inc.**
http://www.nimblegen.com/

Further Reading

- Kano M, Nishimura K, Ishikawa S, Tsutsumi S, Hirota K, Hirose M, Aburatani H. Expression imbalance map: a new visualization method for detection of mRNA expression imbalance regions. *Physiol Genomics* 2003;13(1):31-46.
- Midorikawa Y, Tsutsumi S, Nishimura K, Kamimura N, Kano M, Sakamoto H, Makuuchi M, Aburatani H. Distinct chromosomal bias of gene expression signatures in the progression of hepatocellular carcinoma. *Cancer Research* 2004; 64(20): 7263-70.
- Wong KK, Tsang YT, Shen J, Cheng RS, Chang YM, Man TK, Lau CC. Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. *Nucleic Acids Res* 2004;32(9):e69.
- Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J and others. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res* 2003;13(10):2291-305.
- Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR and others. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res* 2004;14(2):287-95.

Tools for Download

- **Expression Imbalance Map**
http://physiolgenomics.physiology.org/cgi/content/abstract/13/1/31
- **Affymetrix Chromosome Copy Number Analysis Tool v2.0**
http://www.affymetrix.com/support/developer/tools/affytools.affx

Jurilab Scientists Find Hundreds of Genes Associated with Cardiovascular Disease

Principal Scientist Jukka T. Salonen Discusses Challenges of Population Study

Finnish biotechnology company Jurilab Ltd. recently completed one of the first whole-genome association studies ever undertaken, discovering over 700 genes associated with coronary heart disease, hypertension and diabetes.

The company's Chief Scientific Officer, Jukka T. Salonen, led the research team that scanned 100,000 SNPs from each of the study participants. Jurilab studied cardiovascular disease in a group of 250 people from the isolated population of Eastern Finland. This population is very homogenous because of the small number of founders, approximately 600 to 800, who settled there 15 generations ago.

On March 2, Salonen spoke about the challenges of the experiment with Giulia Kennedy Ph.D., the Affymetrix scientist who developed the whole-genome sampling assay for the Mapping 100K Set. Among the topics the two discussed were:

- The discovery of new associations in the three cardiovascular and metabolic diseases in the study, the significant overlap of genes between these disease phenotypes and his strategy for classifying these various phenotypes.
- The large linkage disequilibrium (LD) block sizes Salonen encountered in such an isolated population and calculating relative risk detection limits using the 100K array.
- New linkage mapping software developed at the Helsinki Institute of Information Technology and how Salonen dealt with the multiple testing problem.

Discovery of 25 New Haplotypes Associated with Cardiovascular Disease

Kennedy: How many genes did you find associated with cardiovascular disease?

Salonen: The number of associations depends on our criteria, but for myocardial infarction, we did find according to our criteria, over 400 associated genes and more than a 1,000 associated markers, about 25 strongly associated haplotypes. For hypertension, we found 300 genes associated with the disease, and with diabetes it was somewhat below 300. So we knew, of course, in advance that these are all very polygenic complex diseases, and so that's no surprise. Myocardial infarction, as expected, is the most complex of these diseases, and in the sense of complexity, hypertension and type II diabetes are similar.

> Geneticists sometimes say it would be good to have very narrowly defined phenotypes. I don't agree with that, because if you take very narrowly defined phenotypes, you can't generalize the findings to anything else.

Kennedy: Is there overlap between the genes found in myocardial infarction, hypertension and diabetes?

Salonen: Yes, there is some overlap—about one quarter to one third of the genes are the same and we expected this. We know that hypertension and diabetes both lead to coronary heart disease—that's why some of the genes for diabetes and coronary heart disease were the same. And that is also the case for diabetes and hypertension. We assume that some pathways are shared by both diseases.

Kennedy: All three are complex diseases; can you stratify those disease populations using stricter criteria?

Salonen: We put a lot of resources into classifying the diseases. For instance, for myocardial infarction we have a special classification system—we don't just accept clinical diagnoses. Our diagnostic classification system is based on chest pain, cardiac enzymes, and electrocardiography changes for

example. So it's based on the use of conventional stuff, but using rigid criteria and a rigid classification system.

It's extremely important that the classification of the phenotype is as accurate as possible. Geneticists sometimes say it would be good to have very narrowly defined phenotypes. I don't agree with that, because if you take very narrowly defined phenotypes, you can't generalize the findings to anything else.

Our strategy is to use molecular tests to determine the sub-diagnosis of the disease. For that reason, we needed to have in our genome wide scan a fairly broadly defined disease. That concerns more than myocardial infarction; this concerns two other diseases that we are studying: hypertension and type II diabetes.

Hypertension and type II diabetes are quite heterogeneous diseases, and for that reason it's very valuable to be able to classify those patients into multiple sub-diseases based on the genetic test.

New HIIT Analysis Method for LD Mapping

Kennedy: For your association studies, did you use commercial software or homegrown algorithms or a combination?

Salonen: We use both. Actually for the association study analysis we have a bunch of programs that have been written here in Finland at the University of Helsinki. The institute is called HIIT—Helsinki Institute of Information Technology. They have a project to develop LD mapping methods and they've written a haplotype sharing software, which they called HPM—Haplotype Pattern Mining.

Kennedy: Is this publicly available?

Salonen: It became recently available through Helsinki University and the licensing services of Licentia.

Kennedy: How do you plan on dealing with the multiple testing problem without diluting the power of what you really are seeing?

Salonen: It is an eternal problem and I don't think that there is a good solution to it. We have followed the literature and what others are doing very carefully and we haven't found anything very satisfactory. The software we are using is based on permutation tests, which does take into account the number of tests carried out.

As a philosophical approach, what I like the most is the replication of findings in multiple populations. You could take on Bonferroni correction or less conservative corrections. But whatever arbitrary p-value you use, it is arbitrary. I believe you first have to repeat the finding, and see multiple methods working in the same study population. And then secondly, and more importantly, repeat it in other study populations.

With that sort of a sample size and in our kind of population we can detect relative risk or odds ratios of two or more...In practice, when we are interested in diagnostic products or drug targets, we are actually very interested only in odds ratios of four or more.

In our case, there is an additional reason. In our population the LD blocks are so large that some of those contain many genes. To then be able to pinpoint individual genes with positional cloning requires a more heterogeneous population. And that also serves as a replication of the original findings.

Large Blocks of LD in an Isolated Finnish Population

Kennedy: My guess is that with the small number of founders and generations and the isolated nature of the population, you're expecting to find extensive LD in this population.

Salonen: Well, our average LD block size is between 30 and 50 kb. That's based on r^2 larger than 0.3. This is all dependent on the criteria that you use, but these would be typical criteria for this purpose.

Kennedy: When you use an isolated population it's great at the beginning because you've got these nice big, large blocks of LD, but in the end the positional cloning becomes that much more daunting because of how many genes are in these large regions.

Salonen: It's not as laborious as it would be in a family study where the blocks are even larger. This is a game of cost consideration for us. The homogenous population enables us to use a somewhat smaller sample size in the first step. Then we can repeat it in a more heterogeneous population, where the sample size needs to be quite a bit larger. But, of course, at that stage we don't need to genotype as many markers as in the first stage, so that would be cost effective.

Kennedy: Did you estimate the lowest genetic relative risk that you would be able to detect for any particular locus using the 100K in your particular population?

Salonen: We chose to start with the acute myocardial infarction program that we are running. We took the cases from a prospectively examined cohort because we wanted to have cases that were incident cases, and so the number would be limited. The second requirement was that every case must also have an affected family member, because we wanted to be sure that the disease is due to genetic reasons in every case.

For that reason, we took the rock bottom smallest sample size that we could use. And the sample size chosen was 125 cases and 125 controls. That's only 250 subjects in total. That's much smaller than almost anybody else has claimed to be sufficient. And, of course, what follows from that is that we are not observing weak associations. With that sort of a sample size and in our kind of population we can detect relative risk or odds ratios of two or more. That's the absolute lowest limit. However, we wouldn't be even interested in weaker associations than that. In practice, when we are interested in diagnostic products or drug targets, we are actually very interested only in odds

ratios of four or more. Anything below four hasn't really got biological importance. We are finding lots of those.

NOTE: The previous interview was edited and condensed from a full-length interview conducted on March 2.

FOR MORE INFORMATION

Contact Information

Jukka T. Salonen, M.D., Ph.D.
Jurilab Ltd.
Chief Scientific Officer
Microkatu 1
FIN-70210 Kuopio, Finland
jukka.salonen@jurilab.com

Giulia Kennedy, Ph.D.
Affymetrix Inc.
Director, Genomic Collaborations
3380 Central Expressway
Santa Clara, CA 95051
giulia_kennedy@affymetrix.com

Organizations

- **University of Helsinki**
http://www.helsinki.fi/university/
- **Helsinki Institute of Information** Technology (HIIT) - http://www.hiit.fi/

Companies

- **Jurilab Ltd.**
http://www.jurilab.com
- **Affymetrix Inc.**
http://www.affymetrix.com/

Further Reading

- More Than 300 New Hypertension Genes Discovered by Jurilab Ltd. (Press Release)
http://www.jurilab.com/default.asp?toc=69
- Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G and others. Dynamic model-based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics* 2005.
- Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J and others. Large-scale genotyping of complex DNA. *Nat Biotechnol* 2003;21(10):1233-7.

Tools for Download

- **Haplotype Pattern Mining Software** (pseudocode download)
http://www.cs.helsinki.fi/group/genetics/alg.html
- **Haplotype Pattern Mining Software (licensed version)**
http://www.licentia.fi/english/)

Advances in Developing Genetic Testing Standards

The CDC's Joe Boone weighs in on developing standardization and harmonization initiatives

Dr. Joe Boone

As the world's economies grow closer together, uniform standards, protocols and procedures are becoming the norm for the manufacture of everything from cars to computers and even in service industries like banking and tourism. One field that has struggled to come together on such standards is the life sciences, particularly in areas like laboratory practices and genetic testing.

Dr. Joe Boone began wrestling with these issues in 1977 when he joined the Laboratory Improvement Program Office at the Centers for Disease Control and Prevention (CDC). Nearly 30 years later, Boone is the associate director for science in the Division of Public Health Partnerships at CDC and is playing a key role in numerous standards programs that are currently underway. Boone says researchers need to look to the same sorts of processes that industry has adopted to ensure both safety and efficacy.

"You need to use a quality systems approach to thinking about this, like the ISO process, where you look at the technical and management issues and try to determine what the likely sources of error are and how you can manage to control sources within a quality system framework," said Boone.

For researchers, one of the most important initiatives currently underway to help develop that kind of framework is the Institute for Quality in Laboratory Medicine (IQLM), said Boone. Formed in 2003, the institute comprises representatives from the entire health care industry who have come together to set quality, safety and measurement standards for laboratory services. The group's 2005 conference will take place on April 28-30 in Atlanta.

Joe Boone talked with AMB Editor Tommy Broudy about some of his efforts to develop fully standardized testing processes.

Standardized Controls

Broudy: In March of 2004 at the 9th Conference on Quality in the Spotlight in Antwerp, you posed the question, "Is it safe to have a laboratory test?" At the time you said it was, but not as safe as it should be or could be. A year later what efforts have you and the CDC been involved with to create a more harmonized and integrated approach to laboratory medicine?

Boone: Well, a big part of this is that we're helping to create IQLM, which I hope will be the mechanism for integration both within institutions and across institutions and across state boundaries and across international boundaries, to really protect patients and also to enhance laboratory services.

Broudy: How could the development of standard controls for microarrays advance the field today, both in the clinical field and the research lab?

Boone: Assuming that you have validated control materials, I think controls are extremely important both for verifying the accuracy of a test and for validating assays that are being put into place. Use of controls helps assure that the whole analytical process is being done correctly.

Broudy: What are your thoughts on synthetics versus naturally-occurring

samples, and where do you think we might be heading with standardized controls?

Boone: The ideal material for proficiency testing, or internal quality control, is something that is actually going to mimic the patient sample and can be put through the whole testing process. If there's an extraction step in the preparation of the sample that the synthetic material can't go through, then there's always going to be a question about whether or not that really is a reliable measurement. Synthetic materials certainly have value though, and if there is no other source of real mutations or materials available, then synthetics will have to be used.

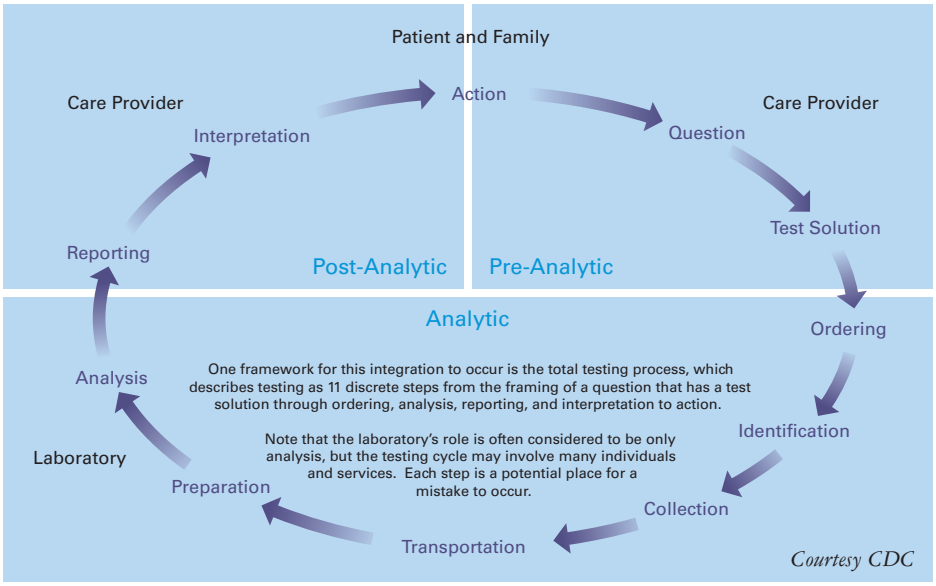
Broudy: And for microarrays specifically?

Boone: There's the question of whether it's feasible or cost-prohibitive to try to acquire patient material for all the things that you would have to QC. I think the whole issue of how you QC a multi-test profile, is still something that professional organizations need to provide better guidance on.

Broudy: How critical is it for there to be a single source that distributes the controls to everyone at a reasonable cost?

Boone: Well, I think you've hit on a really important point. We've been working with a number of groups like Affymetrix for several years trying to make sure that there are quality control materials available. The CDC has funded two research projects to try and develop cell lines or synthetic materials that have known mutations in them. These materials could then be stored at a biobank and be made available for quality control purposes. We now have a number of materials and validated cell lines stored at the Coriell cell repository in Camden, New Jersey. Coriell is marketing those at a reasonable cost, so what we'll be able to make available will either be the cell line itself, or DNA from that cell line.

Broudy: What steps are you taking to expand the collection?



Boone: We've funded a national coordinator to expedite the process of collecting cell lines that are priorities for the genetic testing community. In addition, we're working on a process to help contributors deal with potential roadblocks, like informed consent and Institutional Review Board concerns, for donating cell lines to Coriell. We're also listing the synthetic materials on our new website, which should be up in a few weeks. The site will have all the materials we're aware of that are available for control purposes.

The recent *Science* article, "Getting the Noise Out of Gene Arrays," was a good illustration of the kinds of things that can go wrong, and you hate to see a whole technology get branded as being a problem if we can head some of these things off in advance.

Standardized Software

Broudy: What other areas of microarrays need to be standardized?

Boone: Well, it's not just the materials—it's also the algorithms. The software that goes into the interpretation of the microarray technology needs to be standardized. It would help if we could get that standardization, particularly in the cancer field. We know that this is a critical factor. You can get completely different results if you use different statistical approaches.

The recent *Science* article, "Getting the Noise Out of Gene Arrays," (http://www.sciencemag.org/cgi/content/summary/306/5696/630) was a good illustration of the kinds of things that can go wrong, and you hate to see a whole technology get branded as being a problem if we can head some of these things off in advance.

Broudy: Algorithms are complex and there is still some debate about what works best for certain situations. What are some of the basic issues at hand when comparing genetic tests?

Boone: One of the things that CDC has been working on fairly directly is working with the National Quality Forum to try to develop a patient safety nomenclature and taxonomy that everyone could use. That's a very important aspect of being able to talk to each other in the same terminology.

Broudy: Can you give us an example?

Boone: Sure. Something as basic as defining what constitutes a testing error is a terminology issue. There's a whole array of both inputs and outputs that could occur as a result of some sort of incident occurring for a patient. What would be considered an error in one setting might not be considered an error in another. We need very clear definitions.

Broudy: How can the terminology be developed for a new technology like microarrays?

Boone: It takes time to develop because it really needs to go through a consensus process, like that used by the Clinical and Laboratory Standards Institute (CLSI) or the National Quality Forum. Stakeholders, such as insurers and health care providers need to reach some sort of agreement about what the common terms are going to be.

Working with Physicians

Broudy: What about the end user—how will health care practitioners deal with these tests?

Boone: It could be very difficult for primary care physicians to get results that they can understand from a microarray, where you have maybe 10,000 results coming to them. They have a difficult time right now sorting through a much smaller set of results. So, trying to find a way to interpret results, using appropriate algorithms is essential.

We recently had a conference on cystic fibrosis (CS) where we looked at some case studies of some reports generated from various laboratories. The people who were receiving these reports felt that the nomenclature being used by the genetics community was totally unintelligible to them, including many medical doctors. If they can't interpret the results, then it's going to be very hard for them to apply those appropriately for the patients.

Broudy: How can the report problem be addressed?

Boone: Some of this is very simple. I have a scientist in our group, who's working very closely with OECD, the Organization for Economic Cooperation and Development, to take a look at what an ideal report would really look like—what is essential information and what is secondary information. When we looked at reports for CF the problem was that a lot of the really important information that the clinician needed to know was buried in the report; it wasn't clearly stated.

Broudy: In five years, where do you see quality management of laboratory systems, and genetic testing services?

Boone: Oh, I think we'll make quite a bit of progress in the next five years, because people are beginning to recognize that we need to do a better job of harmonization, standardization of the whole field. That's the first thing that has to happen. There has to be some awareness that things do need to change and that is certainly coming about. There's a number of different groups, the ISO group, the OECD group, the Affymetrix-sponsored conference that's coming up in Paris—all these things are pointing in the right direction.

NOTE: The following text is an edited and condensed version ferom a full-length interview conducted March 3.

FOR MORE INFORMATION

Contact Information

D. Joe Boone, M.S., Ph.D.
Associate Director for Science,
Division of Public Health Partnerships
Centers for Control and Prevention (CDC)
4770 Buford Highway, MS G-25
Atlanta, GA 30341
djb2@cdc.gov

Organizations

- **Centers for Disease Control and Prevention**
  http://www.cdc.gov/
- **International Organization for Standardization**
  http://www.iso.org/iso/en/ISOOnline.frontpage
- **Institute for Quality in Laboratory Medicine**
  http://www.phppo.cdc.gov/dls/iqlm/default.aspx
- **Coriell Institute for Medical Research**
  http://coriell.umdnj.edu/
- **National Quality Forum**
  http://www.qualityforum.org/
- **Organisation for Economic Co-operation and Development**
  http://www.oecd.org/home/

Conferences

- **10th Quality on the Spotlight Meeting (March 7-8, 2005; Antwerp, Belgium)**
  http://www.bioanalytics.co.uk/html/quality_2005.html
- **International Meeting on Clinical and Laboratory Genomic Standards**
  (May 3-5 2005; Paris, France)

Companies

- **Affymetrix Inc.**
  http://www.affymetrix.com/index.affx

Further Reading

- Boone, J. Is it safe to have a laboratory test? *Accred. Qual. Assur.* 10:5-9, 2004.
- Boone, J. Quality in laboratory medicine. *Accred. Qual. Assur.* 10:76-7, 2005.

Improving Algorithms for Microarray-based Chromosomal Copy Number Analysis

Affymetrix' Jing Huang and Earl Hubbell discuss increasing detection limits for chromosomal amplifications and deletions

Dr. Earl Hubbell and Dr. Jing Huang

New high-density microarrays are enabling researchers to assess copy number and genotype across the entire genome in just one experiment, providing quicker detection of amplifications and deletions—offering more detail than comparitive genomic hybridization (CGH) and more markers than loss of heterozygosity (LOH) studies using microsatellites. However, using 100,000 SNPs poses a new challenge as well—developing an algorithm that can deal with an unprecedented amount of data and identify chromosomal changes at a resolution never before possible.

Increasing sensitivity and stability, finding a way to work around the multiple testing problems, and developing better ways to correlate chip intensity with copy number are among the hurdles that must be cleared to make this new approach work.

AMB Editor Dr. Tommy Broudy recently sat down with two Affymetrix scientists who have been among the first to face these new challenges and discussed some approaches that they are trying. Dr. Jing Huang is a bioinformatics data analyst who has written copy number algorithms and works in genotyping microarray R&D. Dr. Earl Hubbell is the company's principal statistician and has developed a number of low-intensity signal extraction algorithms and microarray probe hybridization models.

Algorithm Sensitivity & Stability

Broudy: Like all other types of analysis, copy number analysis is bound by the sensitivity, specificity and stability of the algorithm. Stochastic models, like the Hidden Markov Chain Model (HMC), form the basis of these algorithms, but detecting small changes in copy number remains a challenge. In what way do you see improving the models to increase sensitivity and stability of the algorithm?

Huang: First of all, whenever we increase marker density, the correlation between the markers definitely will increase, and we can use stochastic models like the Markov Chain Model with them to make better predictions. Also, when you are looking at SNP probe and fragment characteristics, you use those characteristics to predict which SNP will really perform well and which will not. So, in a perfect world, you can select probes that are more sensitive and specific, making the algorithm even better. We also look at large samples from sets of normal people, so we can better estimate the distribution or variation in normal people.

Broudy: What other stochastic models might be implemented?

Hubbell: One typical set of models would be some sort of hierarchical Bayes approach, in which you're modeling not just the intensity data you see on the chip, but something about the processes likely to lead to different copy number polymorphisms. So, for example, you might have, from previous data or from biological understanding, a feeling about how often break points occur, or perhaps that there are hot spots for break points where it is more likely to have a copy number polymorphism than not. Similarly, you could have something about the amount of amplification you're likely to see—whether it's a run-away process that leads to a huge number of amplifications or a complete deletion, or whether you're likely to just lose one copy or add a few copies.

Broudy: What models—hierarchical, HMC or otherwise—are available for copy number analysis today?

Huang: This is a relatively new area, and I estimate that there are about 10 to 20 groups really working on developing algorithms for copy number. Many of them are from academic institutions, such as Stanford. The Cluster Along Chromosome (CLAC) algorithm that uses hierarchical clustering on

CGH array data and microarray data to predict copy number provides a decent way to estimate false positive rates, and there is also Cheng Li's dCHIP model developed at Harvard and Dana Farber. That is also publicly available. These institutions do LOH and copy number estimation incorporating HMC. There's also a Genomic Imbalance Map algorithm developed largely by Shumpei Ishikawa in Dr. Hiroyuki Aburatani's laboratory at the University of Tokyo.

The Affymetrix copy number tool uses a regression model to predict copy number and it works quite well when we test it on one to five-fold changes. The main advantage is we have a relatively large reference pool consisting of more than 100 normal individuals. For each SNP marker, we can specifically describe the range of random variation in normal samples. So, when a test sample comes in, you can easily tell if this test sample was in the range of normal variation or outside. If it's outside, you can determine how far it deviates and you can directly derive a confidence measurement. For many other algorithms out there, it's just one test sample against one control sample and you have to make the assumption that all the marker across the genome behave similarly. Then based on that assumption, you detect signal.

Multiple Testing Problem

Broudy: At a traditional p-value cutoff of 0.01, you get lots of errors because there are so many measurements being made. We don't always have 10,000 samples to analyze. What are you doing with algorithm development to help address this multiple-testing problem?

Huang: The most straightforward and direct way is to increase your threshold of p-values. Instead of 0.001, you can do 10^{-10} or 10^{-20}. You gain true positives, but you also lose moderate signals. So that goes back to your previous question. The main aspect for improvement is again the sensitivity and the stability of your algorithm, so that you can confidently detect moderate changes that stretch across small regions.

When you are using high-marker densities, you can use neighboring information to boost the confidence of your findings. So, for example, you can measure the similarity between nearby markers and group them using a bottom-up classifying algorithm. For the 10K and 100k algorithm, we also use some kernel smoothing and meta-analysis to borrow strength from the neighboring point. In a region where all the markers are consistently showing the same pattern of change, although the change may not be dramatic, you will have more confidence that the change is true.

You can also use the probe characteristics to better predict dosage response and to better describe what is happening between the relationship of chip intensity and real target concentration, i.e. copy number.

Hubbell: The two main approaches Dr. Huang has described have been improving the significance of the results by improving the quality of the data, and altering the hypothesis from, this marker has changed two-fold, to, this segment of the genome has changed two-fold. There's also another approach of going for more false discovery rates.

The idea there is not that you choose a p-value for a certain number of false positives, but that you choose for a certain proportion of false positives among those things you find significant. This tends to give you a much less conservative result than altering your p-value to an incredibly high threshold and usually gives you more power to detect. Provided, of course, all the assumptions that go into making that proportion actually hold true for your experimental system—delicate and tricky to get right.

Correlating Chip Intensity with Copy Number

Broudy: You and Dr. Huang just talked about correlating probe intensity with copy number; it seems like there could be a number of criteria that go into designing a probe for copy number specifically. What does an ideal copy number probe look like?

Hubbell: The natural things you want for any probe in this sort of hybridization-based assay, you want a good affinity, you want it to be relatively unique in the genome to be sure it is not cross-hybridizing all over the place. If your assay is one that reduces the complexity of the genome, such as the whole genome sampling assay we use here, you can design stuff to be unique in the fraction of the genome that's amplified. That gives you a little more flexibility. In general, however, we built some sort of predictive model from the base composition of the probe—not just GC content and length, but position within probe of each base pair. We found in expression analysis that dimer-based models, nearest neighbor models, and nucleotides worked better than single nucleotides alone, although they both have considerable predictive power over a sort of naïve GC content approach. As with expression analysis, you can build a straight-forward regression model on both the expected background of the probe—how likely it is to promiscuously hybridize—and you can also build a model for how likely it is to respond to a perfectly matched target. Both of those models together will tell you what a good probe is, or allow you to predict what probes are likely to be good. Naturally, of course, hybridization is complicated. Empirical verification is always best, but you can get a long way towards filtering out things that are probably going to be bad.

Broudy: What specific areas do you see that can be tailored, that can be improved?

Huang: In five years, we're going to have higher-density microarrays. So, that for all the neighboring markers, we will be able to very confidently estimate how highly they correlate. And then, it will be very important to

link copy number signatures with expressions—how the DNA abnormality factors into RNA function.

Hubbell: In five years, the chip density will be much larger. Probably there won't be as many SNPs as there will be probes at that point. So we'll be talking copy number, rather than loss of heterozygosity. It also seems likely that our ability to do multiple testing correction and joint modeling of the probes on a chip will be improved. First, we'll have a greater understanding of hybridization assay conditions and be able to normalize a lot more of the assay and sample-to-sample variation, to get very precise estimates of just how much things are different and just where the change points are. We will be able to take, for example, in a cancer sample not just a clonal set of deletions, but perhaps a mixture of deletions leading from area to area in the sample. The sensitivity of the assay could possibly be increased to single cells, rather than a population-based assay like we have currently.

NOTE: The following interview was edited and condensed from a full-length interview conducted on March 2.

FOR MORE INFORMATION

Contact Information

Earl Hubbell, Ph.D.
Affymetrix Inc.
Principal Statistician
3380 Central Expressway
Santa Clara, CA 95051
earl_hubbell@affymetrix.com

Jing Huang, Ph.D.
Affymetrix Inc.
Bioinformatics Data Analyst
3380 Central Expressway
Santa Clara, CA 95051
jing_huang@affymetrix.com

Companies

- **Affymetrix Inc.**
http://www.affymetrix.com/

Further Reading

- Huang J, Wei W, Zhang J, Liu G, Bignell GR, Stratton MR, Futreal PA, Wooster R, Jones KW, Shapero MH. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics* 2004;1(4):287-99.
- Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR and others. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res* 2004;14(2):287-95.
- Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen MM, Lu G, Fang J, Liu WM, Ryder T and others. Probe selection for high-density oligonucleotide arrays. *Proc Natl Acad Sci U S A* 2003; 100(20):11237-42.
- Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G and others. Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics* 2005.

Tools for Download

- Affymetrix Chromosome Copy Number Analysis Tool v2.0
http://www.affymetrix.com/support/developer/tools/affytools.affx
- dChip - http://www.dchip.org/
- CLAC algorithm - http://www-stat.stanford.edu/~wp57/CGH-Miner/index.html

SECTION 2: RNA ANALYSIS

Developing Tools and Methodologies for Whole-genome Transcript Mapping

A Discussion with Dana-Farber's John Quackenbush and Yale University's Paul Bertone

New microarrays that tile probes across the complete genome—both coding and non-coding regions—are challenging scientists' most fundamental assumptions about how the genome works and even the very meaning of the word "gene." They are also posing huge challenges for the statisticians and bioinformaticians who must create tools to analyze and interpret this flood of new data.

Dr. John Quackenbush from the Dana-Farber Cancer Institute and the Harvard School of Public Health and Dr. Paul Bertone from the Center for Genomics and Proteomics at Yale University recently sat down with Affymetrix Microarray Bulletin to discuss the data analysis implications of this new research.

Quackenbush focuses on developing open source software tools to manage this type of complex microarray data. Bertone recently published the first high-resolution tiling array scan of all human chromosomes, which validated and expanded the efforts of earlier scientists.

Dr. John Quackenbush

Among the topics the two discussed were:

- Validation efforts for transcription mapping beyond the coding regions, including expressed sequence tags (ESTs) and tiling array discoveries

- Improving ChIP-on-Chip methodologies to reduce technological artifacts and increase signal-to-noise ratios

- New software tools to analyze the complex data generated from high-information capacity tools like tiling microarrays

Transcription Beyond Coding Regions

Bertone: A number of groups, including ours and Tom Gingeras' research group at Affymetrix laboratories, have undertaken large-scale surveys of transcription across the human genome. A surprising finding early on has been the relatively high degree of transcription beyond that predicted by existing gene annotation data. What are your impressions of these results?

Quackenbush: I was thrilled beyond belief when Tom's paper and the work that you've published first appeared. At first a lot of people looked at it and said, "Oh, this is garbage, you know, this is obviously noise from the hybridization." But, it really made people step back and think about what exactly we mean by a gene and what we mean by a transcript. The bias has always been towards protein-coding genes and it's clear from the data that there's a lot more going on than we actually anticipated.

My group has worked with EST data for quite some time, and one of the things that we've been criticized for is the fact that EST data is very noisy, and people have said, "Look, these aren't all real transcripts. These aren't genes." But we could look in multiple libraries, we could map them back to the draft genome, we'd see support from many tissues, many places, and many independently made libraries that suggested that the genome was transcribed in regions in which there weren't protein-coding genes. And it was just phenomenally exciting to see this kind of data validated by somebody else with a completely independent technique.

Bertone: I recently visited the lab of Yoshihide Hayashizaki, who is running the FANTOM project at the RIKEN center, and he told me that they got a lot of EST data from what they initially assumed were partial transcripts. After getting the same results over and over again, he concluded that those were full-length transcripts; it's just being transcribed from within the gene locus. So what they were seeing were alternate spliced isoforms, where the gene was showing up with promoters internal to the sequence. This was resulting in some of the exons being omitted.

> An absolute gene count is like asking, "How many angels can dance on the head of a pin?" It doesn't really matter.

Quackenbush: Well, I've seen some of their data too and I love what they've been doing, because if you look at it carefully, you see a lot of truncated transcripts—either at the 5-prime or 3-prime end—and a lot of these do look real. You see a lot of evidence for trans-splicing, where one gene is actually spliced to the next gene downstream. You see evidence for multiple alternative splice forms, exon-skipping, but you also see a lot of evidence for non-coding and antisense RNAs as well. And a lot of that data was dismissed as noise early on; now the RIKEN group is in a position to address whether or not those transcripts are really biologically meaningful because the transcripts are real.

Changing Gene Counts and Definitions

Bertone: Mapping new transcripts is one thing, but people always ask, "How many genes are there in the human genome?" And I think to myself, "Who cares?" Alternate splicing and other complex models are conferring unimaginable amounts of variation; it doesn't really make a big impact on me that there are 19,000 genes in *Drosophila* and probably 24 or so thousand in the human genome.

Quackenbush: I think you're right. The 24,000 number doesn't account for splice variants; it doesn't account for trans-splicing; it doesn't account for non-coding RNAs, or even anti-sense RNAs. An absolute gene count is like asking "How many angels can dance on the head of a pin?" It doesn't really matter. The fundamental question is how all these various mechanisms affect the phenotypes we see. I don't think there's a high-throughput, large-scale way to assess the function of these individual transcripts. This is a situation where we have to think about over-expressing or knocking out these transcripts and seeing if they have a phenotypic effect. That's really what understanding function comes down to—doing careful analyses.

If we think about what a gene is, we have to go back to the fundamental definition. A gene is a unit of heredity; it's not necessarily something that encodes a protein. And if we want to identify the genes in the genome, we have to ask whether or not these non-coding RNAs are units of heredity—do they influence phenotype? And that's a really basic biological question, and it's a hypothesis that only comes from generating these holistic data sets and carefully analyzing them.

Improving ChIP-on-Chip Methods

Bertone: One of the things we're trying to be very careful about is reducing the technology-based artifacts in the ChIP-on-Chip method. For example, imagine that you have a lot of secondary structure just in the chromosomal sequence. When cross-linking time is extended, you can imagine cases where the associating DNA could be looping around from quite distal areas of the chromosome that have nothing to do with the actual interaction. That could cause some problems. We're working with the ENCODE Cosortium and with Bing Ren's lab (University of

California, San Diego) and Peggy Farnham's lab (University of California, Davis) to really standardize the protocol so that we minimize those effects.

Also, spatial artifacts have to be taken into account when you are doing this analysis. When we immunoprecipitate our fragments we sonicate them and end up with a population of IP'd chromatin fragments that are around 500 base pairs and up. As a result, when you hybridize the DNA sample to the array, you have a slightly larger footprint of binding than you would normally want to see—in an exon mapping experiment, for example. When you then score this data, there are cases where the results don't appear to agree; but then when you look at the binding profiles by eye, the sites line up very well. It turns out there may be a slight shift due to the size of the fragments that you are hybridizing to the array.

Quackenbush: I think one of the things we have to realize in all of this genome-scale data is that as the technology improves, the signal-to-noise improves. The work you've been doing to try to improve the protocols is incredibly useful, but that has to go hand-in-hand with methods to analyze the data, and to allow for variation. Some of the variation is variation we are just going to see because of biological noise, not necessarily technical noise.

Beating Down the Signal-to-Noise Problem

Bertone: In terms of the ChIP-on-Chip experiment, we're seeing a lot more signal from our early yeast experiments than we do in our current human ones. When you're doing this assay, you're hybridizing two different samples: one is the immunoprecipitated sample with all the transcription factor-bound DNA and the other is a non-specific sample. Because you are hybridizing in a genomic DNA background, you're always going to get some signal. The point is that you are looking for fold enrichment and it's not like a transcript mapping experiment where some features might not light up. Almost every feature lights up because of the nature of the hybridization. What we find in yeast is that that fold enrichment can be quite high, making the analysis rather straightforward. The outliers just stand out, no matter what analysis method you want to do—for example, you could calculate p-values or normalize fluorescent intensity ratios and just take outliers outside a standard distribution from the mean.

In the human system, we find fold enrichment to be much lower than in yeast—like on the order of three- to four-fold. So, in terms of signal-to-noise ratio that's a big problem. What people have done to try to compensate is PCR amplify their immunoprecipitated fragments and they get better enrichment that way, as you might imagine. However, PCR is going to distort your population of DNA fragments represented. For that reason, we've been trying to avoid PCR amplifying our samples. Instead, we pile on a lot of replicate data—about 8 to 10 replicates. The cumulative effect is that we can intersect all of these data sets and the fairly random false positives tend to fall away. So even though the enrichment from any one experiment might not be very high, we do get clear binding signals from the sites that we know are high confidence.

Quackenbush: You know the best way to beat down signal-to-noise is to make more measurements. And if you make an infinite number of measurements eventually the signal will stand out. A lot of what we've seen change in biology has nothing to do with a more fundamental understanding. It really is a revolution driven by technology. And as the cost of the technology falls, our understanding is going to improve, because we are going to be able to generate much more accurate data just by doing more and more replicates.

Data Analysis

Quackenbush: It is important to realize that we are generating a lot of data, but if you look at the quantity of data that's coming out of the Hubble Space Telescope, or if you look at the amount, the shear volume of data that's coming out of any high energy physics experiment—like out of the Large Hadron Collider at CERN—we're really not talking about vast quantities of data from these biological systems. But what makes analyzing the biological data a challenge is that the structure of the biological system that we're looking at is very complex.

We're constantly trying new algorithms to deal with this complexity. One of the tools that we recently developed, which has turned out to be incredibly useful for interpreting some of this data, is an algorithm called EASE that was developed by Doug Hosack and Glynn Dennis at National Institute of Allergy and Infectious Diseases (NIAID). Thanks to them, it's now been incorporated into the MeV software package that my group has developed. But what Evolutionary Algorithms Scripting Environment (EASE) does is it looks at representation of Gene Ontology terms, and does a Fisher Exact Test, looking at whether or not certain classes of genes are over-represented in any "significant" gene set that you select using some criteria from an experiment. And things like that can often give you really interesting insight into the underlying biological questions.

Quackenbush: And, you know, we're working on other things, like tools to mine the literature and pull out over-represented MeSH terms. Anytime you can take what we know—like what is in the literature—and use it as a biological filter on the data, on top of the statistical filters that people are using, you gain a perspective that can really be useful in trying to interpret the results of your experiment.

NOTE: The following interview was edited and condensed from a full-length interview conducted on March 2.

FOR MORE INFORMATION

Contact Information

John Quackenbush, Ph.D.
Professor of Computational Biology and Bioinformatics
Harvard School of Public Health
Biostatistics, Bldg 2-451
655 Huntington Ave
Boston MA 02115
johnq@jimmy.harvard.edu

Paul Bertone, Ph.D.
Yale University
Molecular, Cellular, and Developmental Biology
944 Kline Biology Tower
Yale University
New Haven, CT 06520
paul.bertone@yale.edu

Organizations

- **ENCODE Consortium** http://www.genome.gov/ENCODE/
- **National Institute of Allergy and Infectious Diseases (NIAID)** http://www.niaid.nih.gov/
- **Dana-Farber Biostatistics and Computational Biology** http://www.dana-farber.org/res/departments/biostatistical/
- **Yale University Center for Genomics and Proteomics** http://cgp.yale.edu/
- **FANTOM Project** http://fantom.gsc.riken.jp/

People

- **Bing Ren laboratory** http://licr-renlab.ucsd.edu/
- **Peggy Farnham laboratory** http://som.ucdavis.edu/departments/pharmacology/departmentlabs/Peggy%20Farnham/Peggy%20Farnham

Companies

- **Affymetrix Inc.** http://www.affymetrix.com

Further Reading

- Kapranov, P. et al. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916-9 (2002).
- Cawley, S. et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116, 499-509 (2004).
- Bertone, P. et al. Global identification of human transcribed sequences with genome tiling arrays. *Science* 306, 2242-6 (2004).

Tools for Download

- EASE - NIAID software that facilitates the biological interpretation of gene lists derived from the results of microarray, proteomic, and SAGE experiments http://apps1.niaid.nih.gov/david/ease1.htm
- TM4 - Software developed by Quackenbush and his group for the management and analysis of microarray data http://www.tm4.org
- ExpressYourself - web-based microarray data analysis platform with a variety of normalization and scoring implementations, compatible with several contact-printed microarray file formats - http://bioinfo.mbb.yale.edu/expressyourself/

AMB NEWS BRIEFS

Harvard to Use Affymetrix GeneChip® 500K SNP Arrays in Atherosclerosis Study

Affymetrix will provide Harvard Medical School - Partners HealthCare Center for Genetics and Genomics (HPCGG) with early access to its new Affymetrix GeneChip® Human Mapping microarrays capable of scanning up to 500,000 SNPs. The arrays will allow HPCGG researchers to investigate the genetic origins of atherosclerosis in more than 1,200 patient samples from the Women's Health Study, an ongoing 10-year National Institutes of Health-funded investigation of cardiovascular events in 38,000 women. Investigators in this project include David Kwiatkowski, Director for Genotyping for HPCGG and principal investigator within the Donald W. Reynolds Center for Cardiovascular Research at Brigham and Women's Hospital (BWH); and Paul Ridker, Director of the Center for Cardiovascular Disease Prevention at BWH.

Karolinska Institutet and Affymetrix to Enter into Strategic Alliance

Karolinska Institutet and Affymetrix have agreed to participate in a five-year strategic alliance, which includes projects involving DNA and RNA analysis in patients with atherosclerosis, breast cancer, rheumatoid arthritis, asthma and dyslexia. The goal of the alliance is to develop diagnostic and prognostic tools.

Institut Curie and Affymetrix to Study Cancer Signatures

Institut Curie and Affymetrix announced a strategic alliance designed to produce genetic signatures for different types of cancer, which could then be used to develop diagnostic tests. The first two projects will focus on identifying important genetic markers for cancer prognosis, including markers to help predict breast cancer relapse and to determine the likelihood of uveal melanoma tumors (a form of eye cancer) spreading further.

Stratagene and Affymetrix Announce Strategic Software Alliance

As part of a strategic alliance, software maker Stratagene plans to develop a new data analysis software package for Affymetrix called ArrayAssist® Lite, which will be offered as a standard statistical analysis solution for Affymetrix gene expression GeneChip® microarrays. In April 2005, ArrayAssist Lite will be made available via the Internet for existing Affymetrix GeneChip users and Affymetrix will begin pre-installing it on new GeneChip workstation desktops. ArrayAssist Lite will support RMA, GC-RMA, MAS5, and also PLIER, a new advanced signal estimator developed by Affymetrix statisticians. Affymetrix customers will also be given the option of upgrading to Stratagene's standard ArrayAssist software, which provides advanced downstream analysis capabilities such as gene clustering, pattern detection and interactive visualizations.

HUGO'S TENTH INTERNATIONAL HUMAN GENOME MEETING, HGM2005, WILL BE HELD AT THE KYOTO INTERNATIONAL CONFERENCE HALL, KYOTO, JAPAN, FROM THE 18TH TO THE 21ST APRIL 2005.

Hidden Complexities of the Human Transcriptome

Philipp Kapranov, Jill Cheng, Jorg Drenkow, Sujik Dike, Sandeep Patel, Jeffrey Long, Shane Brubaker, Ian Bell, Gregg Helt, Hari Tammana, Antonio Piccolboni, Srinka Ghosh, Stefan Bekiranov, Dione Kampa, Victor Sementchenko, John Manak, Ganesh Madhavan and Thomas R. Gingeras.
Affymetrix Inc.
Presentation number: 63
Poster number: 174
Corresponding author:
philipp_kapranov@affymetrix.com

Dr. Philipp Kapranov of Affymetrix Research Laboratories will present data describing the unbiased transcriptome mapping of the human genome at 5 bp resolution. The study presents the most detailed picture of human genome transcription, describing previously unknown levels of transcriptional complexity.

- A significant fraction of the transcribed genome lies outside of the known annotations. Transcripts of un-known functions or TUFs appear to represent a diverse group of transcripts that are located both within and outside of the bounds of known genes, often overlapping with existing annotations on either sense or anti-sense strand.

- A single base pair in the genome can be part of multiple isoforms of coding or non-coding transcripts, overlapping sense/antisense transcripts, transcripts that join together exons of separate genes and trans-spliced transcripts.

- An additional facet of the complexity of the human transcriptome is contributed by intracellular compartmentalization of transcripts, with a significant fraction of cellular polyA+ and polyA- transcripts found exlusively in the nucleus or cytosol.

- A large fraction of the human transcriptome is composed of polyA- transcripts.

These observations provide some pause as to the state of our understanding concerning where and how the information from the human genome is organized and how it should be interpreted in regards to the linkage and sequence polymorphisms studies.

Ultra-high Density Microarray-Based SNP Genotyping

Jones, Keith W.; Matsuzaki, Haj; Chen, Joyce; Tom, Jennifer; Zhang, Jane; Loi, Halina; Law, Jane; Ho, Tam; Webster, Teresa; Di, Xiaojun; Yang, Geoff; Hui, Henry; Mitra, Swati; Hubbell, Earl; Walker, Wynn; Liu, Guoying; Cawley, Simon; Walsh, Sean; Mei, Rui and Shapero, Michael H. *(cont'd on p. 17)*
Affymetrix Inc.
Presentation number: 88
Poster number: 253
Corresponding author:
keith_jones@affymetrix.com

Dr. Keith Jones will discuss the development of a microarray genotyping assay which allows greater than 500,000 single nucleotide polymorphism (SNP) markers to be interrogated.

The emerging role of SNP analysis in clinical association and pharmacogenetic studies has created a need for high-throughput genotyping technologies. Synthetic DNA microarrays are an ideal format for high throughput genotyping because they are capable of generating data for more than 100,000 SNP markers in a highly parallel fashion. The principle challenge is to reproducibly present a subset of the genome to an array such that accurate allelic information is obtained. Kennedy *et al.* (2003) and Matsuzaki *et al.* (2004) recently described a methodology which utilizes restriction enzyme digestion and PCR to convert genomic DNA into a form that is amenable to highly multiplexed amplification using a single generic primer. The method is highly accurate, reproducible and gives consistently high call rates; it serves as the basis for the newly developed 500,000 SNP genotyping methodology.

An Appropriate Use of Funds? Genetic Research, Screening and Public Health Management

Katie Tillman-Buck
Affymetrix Inc.
Presentation number: 45
Poster number: 78
Corresponding author:
katie_buck@affymetrix.com

Katie Tillman-Buck of Affymetrix Corporate Affairs will discuss the public health challenges and opportunities of genetic screening from the perspective of a technology innovator that has enabled whole genome analysis.

Genetic testing and the debates surrounding its harms and benefits have, thus far, focused on tests of rare and highly-penetrant single-gene disorders. However, modern information-rich technologies continue to make genetic testing cheaper, broadening the scope of diagnostic inquiry beyond traditionally defined "treatable disease."

In the near future, individuals could readily demand information regarding their genetic predispositions, drug metabolism, environmental influencers or the genetic severity of a given disease.

Public health labs involved with newborn screening are already incorporating new technologies in order to provide expanded access to screening. Given a resource-limited public health system; however, expanding one program often translates to shrinking another, making the "value" of a broader genetic testing program difficult to measure. And when you consider proactive testing, such as drug metabolism or predisposition testing, the cost/benefit analysis becomes even more difficult.

While more questions than answers remain, advances in technology are forcing the question of what we can test for, what we should test for, and ultimately, what we will test for.

Automation of Applications Using High Density GeneChip® Probe Arrays

Thomas Ryder
Affymetrix Inc.
Poster number: 271
Corresponding author:
tom_ryder@affymetrix.com

Dr. Tom Ryder will present recent progress and validation studies in the development of a 96-array high throughput system that automates the most labor-intensive steps in GeneChip® microarray processing—ample preparation, hybridization, washing and scanning.

- Advances in photolithographic manufacturing have allowed the researchers to pack tens of thousands of DNA probes onto 6 mm x 6 mm arrays small enough to fit into the well of a standard 8 x 12, 96-well microtiter tray.

- The multi-step biochemical process of target preparation—the conversion of mRNA into amplified cRNA applied to the array—has been automated on a Affymetrix GCAS™ robot system, which integrates a Sciclone fluid handler and

Twister II robotic arm from Caliper Life Sciences with other peripherals and consumables that performs the sample prep with a minimum of user intervention.

■ Methods have been developed for the GCAS system to automate the post-hybridization processing of the 96-array plates of an 8-micron version of the GeneChip U133A probe array.

The performance of this automated assay system has been tested in a number of carefully calibrated experimental systems and has been shown to yield expression metrics and assay performance that meet or exceed the high standards attained with the cartridge-based GeneChip 2.0 platform expression assay system.

Using 100,000 SNPs on High-density Oligonucleotide Arrays to Map the History of Human Populations

[1]Giulia C. Kennedy, [1]Chun Zhang, [1]Manqiu Cao, [2]Margaret Taub, [1]Simon Cawley, [1]Hajime Matsuzaki, [1]Dione K. Bailey
[1]**Affymetrix Inc.**
University of California, Berkeley
Poster number: 16
Corresponding author:
giulia_kennedy@affymetrix.com

Dr. Giulia Kennedy will present data using GeneChip® Mapping 100K arrays to discover new genotypic markers that can be used as ancestry-informative markers for admixture mapping of human disease genes.

■ The researchers determined allele frequencies across African, East Asian and European populations and identified hundreds of SNPs with high inter-population frequency differences. Additionally, Kennedy's team identified DNA regions that have undergone recent positive selection by looking for extremely rare recombinations and then performing a long-range haplotype test (Sabeti *et al.,* 2002).

■ The research team also determined ancestral state by genotyping chimpanzee and gorilla DNA and used this information to identify non-ancestral haplotypes that have risen to high frequency. These regions contain many genes involved in human health and reproduction, as well as genes involved in important biological processes, such as ion channels, receptors and growth factors.

■ Because the Mapping 100K arrays contain 2334 SNPs on the X chromosome, the team undertook the first high-density genome-wide comparison of the X chromosome with autosomes. As expected, they found heterozygosity of the X chromosome to be significantly lower than that of autosomes, in all three populations; the effect was most pronounced in Asian and European populations, suggesting unique factors that may have affected variation differently amongst the populations.

Fidelity of Tumor Cell Lines Using a 10,000 SNP Whole-genome Scan

[1]John Palma, [2]David Judd, [1]Mamatha Shekar, [1]Chunmei Liu, [1]Janet Warrington

[1]**Affymetrix Inc.**
[2]**Invitrogen Corporation, Gibco R and D**
Poster number: 268
Corresponding author:
john_palma@affymetrix.com

Dr. John Palma will present a collaborative study with Invitrogen Corporation that used GeneChip® Mapping 10K arrays to discriminate between the cell lines HeLa S3 and 293F (training sample), and to distinguish the identity of 10 unknown cell lines (blinded samples). Tracking cell lines has remained a challenge for clinical researchers, largely because conventional cytogenetic methods are limited in resolution and higher-resolution methods, such as CGH, are not widely available.

■ The training set comprised five replicates of each cell type; after comparing 10,000 SNP genotypes from each, the team discovered 2389 SNPs that discriminated the five HeLa samples from the five 293F cells.

■ Technical replicates of one HeLa and one 293F sample showed greater than 99 percent concordance between the overlapping SNPs that passed the discrimination filter for both cell lines. Similarly, concordance within the five replicates of HeLa or 293F cells was 99.8 percent and 99.9 percent, respectively.

■ Five of the 10 blinded samples had more than 80 percent concordance with either HeLa or 293F cells. Two blinded samples had absolute identity (more than 97 percent) to the HeLa or 293F profiles which were consistent with their true identity. In addition, the discriminatory SNPs could correctly distinguish all of the other 8 blinded cell lines as resembling or unique from Hela S3 or 293F cells.

AFFYMETRIX
MICROARRAY BULLETIN

Affymetrix Microarray Bulletin (AMB) is a quarterly newsletter published by Affymetrix Inc. that highlights the work of thought leaders in molecular biology and microarray research from around the world. AMB is a scientist-to-scientist publication, written for researchers, statisticians, and software and information technology engineers. Our goal is to create a forum where the research community can discuss its work and share methods, tools, insights and information.

The online version of this newsletter can be accessed at http://www.microarraybulletin.com

Wes Conard, *Editor-in-Chief*
wes_conard@affymetrix.com

Tommy Broudy, *Managing Editor*
thomas_broudy@affymetrix.com

Lisa Osborne, *Copy Editor*

Tri Hoang, *Associate Editor*

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