

Jurilab Scientists Find Hundreds of Genes Associated with Cardiovascular Disease

Principal Scientist Jukka T. Salonen Discusses Challenges of Population Study

By Tommy Broudy

PHILADELPHIA, April 1, 2005 — 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.

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 home-grown 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.

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

“This is a game of cost consideration for us.”

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 homogeneous 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.

AFFYMETRIX MICROARRAY BULLETIN

Editorial Staff

Wes Conard, *Editor-in-Chief*
wes_conard@affymetrix.com
Tommy Broudy, *Managing Editor*
thomas_broudy@affymetrix.com
Rachel Shreter, *Editor*
rachel_shreter@affymetrix.com
Kamalia Dam, *Associate Editor*
Stacey Ryder, *Associate Editor*
Daniel Noble, *Copy Editor*
Jim Butler, *Contributing Designer*

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/>