A F F Y M E T R I X
MICROARRAY
BULLETIN

REGULATION
VOLUME 3 ■ ISSUE 2
AMB REPRINT

www.microarraybulletin.com
Spring 2007

High-resolution Tiling Arrays Used to Map Genome-wide Methylation Patterns in Arabidopsis

The Salk Institute's Joseph Ecker and Lawrence Livermore's Jutta Kollet discuss the first study to map the methylome in Arabidopsis

By Stacey Ryder

LOS ANGELES, May 28, 2007 — Researchers at the Salk Institute for Biological Studies and The University of California, Los Angeles, have created the first comprehensive, genome-wide map of DNA methylation in the plant model organism, *Arabidopsis thaliana*. Understanding DNA methylation in the genome may have far-reaching implications, from breeding stronger crop species to understanding and treating human cancer.

The team used custom-designed Arabidopsis whole-genome tiling arrays (now commercially available) to determine that more than one-third of genes are methylated in transcribed areas. Those genes with so-called "body methylation" were also those that were most highly expressed. Only about 5 percent of Arabidopsis genes displayed methylation in the promoter region, a condition associated with tissue-specific gene expression. The study was published in the September 2006 issue of *Cell*.

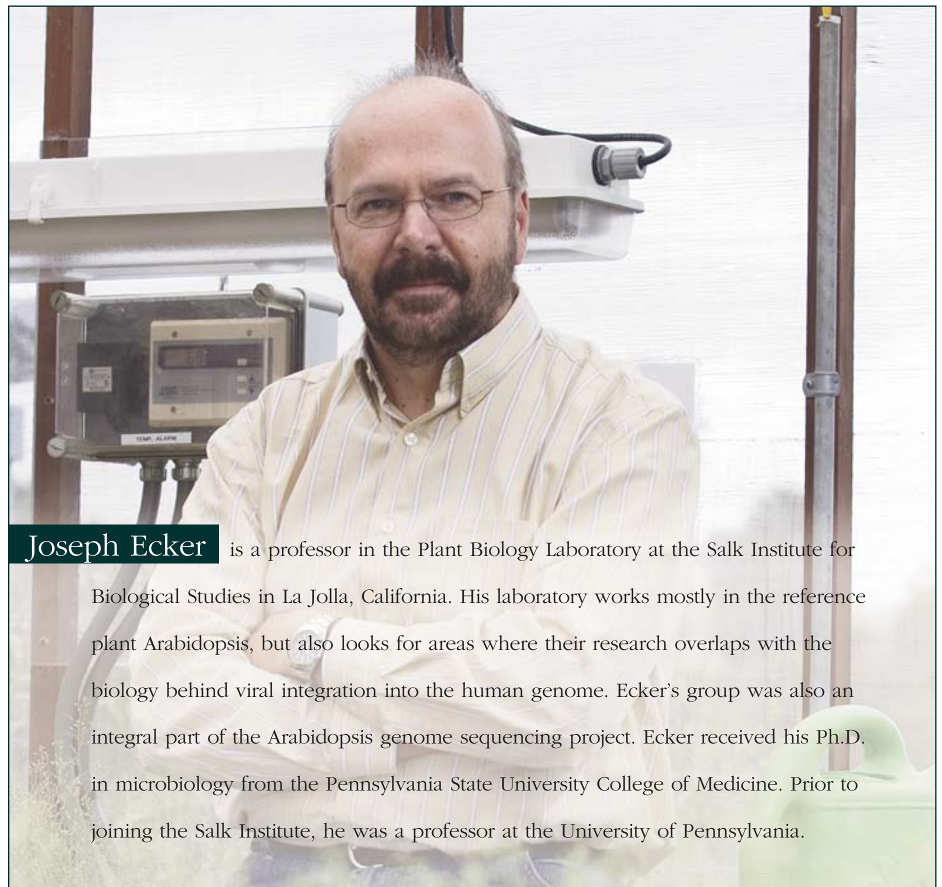
"From projects like the HapMap, we are beginning to understand how the phenotype relates to the genotype. However, we know very little about the 'epigenotype,' and this first whole-DNA methylome study gives a hint that it is more complex than we imagined," said Joseph Ecker, professor of plant biology at the Salk Institute and one of the primary investigators on the paper.

Ecker describes this study as an initial step in investigating methylation in Arabidopsis. The team's experiments used extracts from whole plants, so they could not make any inferences about tissue-specific methylation patterns; however, one of their next studies will focus on analyzing tissue-specific samples.

Ecker recently spoke to Jutta Kollet, a postdoctoral fellow in Lisa Stubbs' Genome Biology Group at Lawrence Livermore National Laboratory, about mapping the Arabidopsis methylome and how information about DNA methylation might change crop science and human medicine in the future.

The two discussed:

- Microarray design and data analsis methods
- How methylome data has changed the understanding of genome regulation
- Future applications of methylome knowledge for agriculture and medicine



Joseph Ecker is a professor in the Plant Biology Laboratory at the Salk Institute for Biological Studies in La Jolla, California. His laboratory works mostly in the reference plant Arabidopsis, but also looks for areas where their research overlaps with the biology behind viral integration into the human genome. Ecker's group was also an integral part of the Arabidopsis genome sequencing project. Ecker received his Ph.D. in microbiology from the Pennsylvania State University College of Medicine. Prior to joining the Salk Institute, he was a professor at the University of Pennsylvania.

Jutta Kollet is a postdoctoral fellow in Lisa Stubbs' laboratory in Genome Biology at the Lawrence Livermore National Laboratory in Livermore, California. Her work centers on transcriptional regulation during immune responses, organ development and genomic imprinting in mouse model systems. She utilizes ChIP-on-chip and gene expression microarray experiments to better understand epigenetic and global chromatin structure mechanisms that might be of relevance to human genome regulation. She received a Diplom (equivalent to MSc.) in Biology (technical orientation) from the University of Stuttgart in Germany and her Ph.D. in Medical Sciences from the University of Nebraska Medical Center.

Methodology and data analysis

Kollet: How did you design the microarrays for this study and what kind of correlations did you see with existing Arabidopsis annotation data?

Ecker: When we designed the array in collaboration with Affymetrix, we first used their algorithms to choose probes for a 35-base resolution.

Essentially, we picked the probes randomly across the genome. We aligned them according to the most current versions of genome annotation on a probe-by-probe basis, so we could look at correlations between existing annotations and what we see on the arrays.

This is our third generation of experiments related to transcription. We published some of the earlier work looking at gene expression in 2003 [Yamada, *et al.*]. This was the first paper in which tiling arrays were used to look at the whole Arabidopsis genome. We published some methods for looking at whether or not a gene was expressed and tried to define regions that were not an expressed part of the genome, like known promoter regions. These regions were compared to known genes to define a background.

In our recent publication, we compared an improved version of the annotation based on our previous data. The correlations were pretty good, although we found many new exons because of the differences in the arrays and the increased numbers of experiments that we could carry out on a single array, versus the previous experiments where we were using 14 arrays.

Kollet: In your data analysis, you compared the results obtained by previously published methods—the TileMap method published in 2005 by Ji and Wong and the Tiling Analysis Software from Affymetrix—with your own novel non-parametric approach using Kolmogorov-Smirnov statistics. Why did you look at these three methods and why did you choose TileMap?

Ecker: The KS statistics are the work of my collaborator at UCLA, Matteo Pellegrini. He is a computational biologist. The main reason for this three-approach comparison was that we weren't confident relying on any one methodology. It still isn't clear what the best approach will be. Differences between probe-probe and probe-signal intensities can be very dramatic.

Our goal was to use several published methods. We did an analysis of all three methods and the results were strikingly similar. Ultimately, we settled on using a very stringent cutoff and the TileMap method. Before the publication of this paper, there weren't any existing algorithms that were able to accommodate multiple comparisons. They all just made pair-wise comparisons.

Since the publication of our paper, there have been some improvements in this program. Also, a recent paper from Srinka Ghosh at Affymetrix has highlighted a new rank statistics method for looking at signals that are enriched in particular regions of the tiling array.

The bottom line is that we wanted to try everything that was out there. We were happy when a lot of the results showed a very high correlation. Ultimately, we used the method that gave us the best fit to the empirical data that we had already.

Kollet: You took a novel approach by assessing the entropy of expressed genes. How much information does that provide?

Ecker: Again, this is an idea of Matteo's. He had done some similar analysis previously. The entropy measures only whether a gene is expressed uniformly across various conditions.

In our case, we were looking at tissue specificity from data that was produced using gene expression arrays. It was a correlation. We were looking at whether or not genes that tended to be more tissue-specific versus more general also tended to be methylated.

In our study, we ground up the whole plant. It was a crude first pass analysis. Since that time, we have begun to look at tissue specificity. But the dogma when we were first doing these experiments was that there shouldn't be a whole lot of tissue-specific differences. I had not worked much in epigenetics before coming to the field, so it was striking to me that folks in the field would make this assumption based on a fairly small amount of data.

Since that time, we have done a number of experiments. Those experiments show very clearly that there are dramatic differences between cell populations. We haven't taken this to the level of looking at single cells in specific organs or tissues, flowers, leaves, etc. But I think that is where we need to look. When we look at cell populations, there is a mixture of cell types. I believe we are missing a lot.

Kollet: In the entropy analysis you identified genes with promoter methyla-

tion as predominantly tissue-specific genes. But do you expect to see different methylation patterns of these tissue-specific genes in specialized cells?

Ecker: I think it will be at least as complex as that. There are correlations, such as whether a gene expresses any antisense transcripts and whether it expresses those transcripts in various methylation-mutant backgrounds.

> "These studies show a clear correlation between increased methylation on the body of a gene and increased expression... genes with more promoter methylation were more tissue-specific."

Looking at whole plant RNA and methylation profiles, these effects for specific genes might easily be masked. Where there might be a correlation is between reduced methylation, for example, and increased antisense expression. This may not hold true across all tissues, but there could be a significant number of cases where there is such a correlation. We would miss that when using whole plant extracts.

Kollet: How could tissue-specific genes be targeted for promoter methylation? Did you analyze for sequence motifs?

Ecker: We didn't. We think that many of the methylation events we see using the arrays are directed by small RNAs. There is RNA-directed DNA methylation machinery that my collaborator, Steve Jacobson, has been studying for many years. This is a homology-dependent DNA methylation machinery. So, small RNAs directed to particular regions of genes can direct DNA methylation there.

The other mechanism one can imagine is that there are protein-directed DNA methylation events. Depending on what chromatin modification state exists, there is a specific DNA methyl transferase in plants that has a chromodomain that binds to modified histones and directs DNA methylation to those targets.

The spectrum of how DNA methyl transferases get directed to their target site is quite interesting. As far as I know, there is no DNA sequence motif that anyone has associated directly with the methyl transferase, but there are other proteins and RNAs that seem to direct the methylation machinery to specific targets.

New mechanisms of genome regulation

Kollet: Were you surprised to see "body methylation" mainly in constitutively expressed genes?

Ecker: We were quite surprised when we began to look at correlate locations of methylation marks and gene annotation. There was some evidence already in the literature from Steve Henikoff that there might be some 3'-located DNA methylation marks in Arabidopsis. I think those results may have been a consequence of not having high-resolution arrays.

Henikoff's group just had a paper in *Nature Genetics* that largely confirms most of the data in our paper, that the body of the gene can be very highly methylated. We found that this methylation is more associated with genes that are highly expressed, based on Shannon entropy analysis, than the average gene.

We originally thought that methylation was involved in silencing the antisense strand and would prevent internal initiation of transcription of antisense transcripts if it occurred in promoters or introns.

Based on these initial experiments, that model didn't hold up for most genes. Time will tell whether we see different patterns in different cell types. We will be looking at particular cell types to determine whether methylation changes antisense expression.

These studies show a clear correlation between increased methylation on the body of a gene and increased expression. But genes with more promoter methylation were more tissue-specific. This correlates with the gene expression data from the Affymetrix Arabidopsis ATH1 Array.

That said, we know from the literature that most antisense transcripts in plants and animals are largely non-polyadenylated. When you look at northern blots of polyadenylated and non-polyadenylated RNA fractions, the antisense transcripts are largely enriched in the non-polyadenylated fractions.

If that holds up, what we are seeing on these arrays is just a small fraction, maybe even an insignificant fraction, of the amount of antisense transcripts that are actually expressed. We really need to look at fractionated RNAs that were non-polyadenylated to see whether or not there is a correlation with methylation.

Joseph Ecker

Applications for methylome data

Kollet: How could the Arabidopsis methylome and transcriptome data be used in agricultural plant-breeding applications?

Ecker: Out in the wild, there are probably lots of epialleles for genes. We're just beginning to look at the genetic and epigenetic diversity of Arabidopsis. We have some papers that are submitted that use Affymetrix technology, through a company called Perlegen, to look at natural variation.

Our next step will be to look at how changes in gene expression and changes in DNA methylation patterns associate with regions that have SNPs or do not have SNPs. Then we can begin to look at regions that are linked to large changes in gene expression, but have very little polymorphism.

A substantial fraction of what we currently view as genetic variation might actually be epigenetic. I know that there are some groups out there that are looking at treating real crop plants with drugs to look at alterations in phenotypes. I have heard that they are seeing some changes. So some of the differences in traits that people have been selecting for could be epigenetic traits, as well as genetic traits. Time will tell if this is correct.

Kollet: Could you transfer your findings of differential methylation patterns of active or silenced chromatin regions to your previous work on HIV and MLV retroviral integration sites in the mammalian genome?

Ecker: My collaborator Rick Bushman is quite interested in this. Future extensions of these kinds of studies to correlate HIV and other retroviral integration events into the human genome and other genomes is actually something that is in progress and could be quite exciting, especially regarding the mechanisms directing retrotransposons to specific sites. Rick already has some interesting data about targeting using those specific transcription factors that HIV interacts with. It would be nice to know the different chromatin states, as well as different methylation states. This is something that we are planning to look at.

Kollet: You mentioned the use of drugs to alter epigenetic traits. What do you think about the application of DNA methylating drugs used in therapeutics and cancers?

Ecker: I don't have a lot of experience with that, but it is a fascinating area. Certainly, changes in the levels of expression of tumor suppressors have been known for a long time to be correlated with DNA methylation events. There are many interesting studies that indicate that this could be a promising approach for altering the state of expression in somatic cells. It could have an effect on whether tumor suppressor genes are expressed or not.

Ultimately, you need to understand the global effects of such compounds in order to understand how they will work. Being able to use a technology like arrays may allow you to expand your understanding of what target genes are affected. Using whole-genome tiling arrays of the human genome to look at cell types and different drug treatments is going to be very, very powerful!

AFFYMETRIX
MICROARRAY
BULLETIN

Editorial Staff

Rachel Shreter, *Editor-in-Chief*
rachel_shreter@affymetrix.com

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

Bronwyn Barnett, *Managing Editor*
bronwyn_barnett@affymetrix.com

Stephen Antonopoulos, *Photo Editor*

Joel Fellis, *Assistant Editor*

Sven Brown, *Associate Editor*

Stacey Ryder, *Associate Editor*

Sabrina Banwait, *Associate Editor*

Marva Maida, *Graphic Designer*

Jim Butler, *Graphic Designer*

FOR MORE INFORMATION

Contacts

- Joseph R. Ecker, Ph.D.
Genomic Analysis Laboratory
The Salk Institute for Biological Studies
10010 N. Torrey Pines Rd.
La Jolla, CA 92037
ecker@salk.edu
- Jutta I. Kollet, Ph.D.
Biosciences and Biotechnology Division
Chemistry Materials and
Life Sciences Directorate,
Lawrence Livermore National Laboratory
L-452
7000 East Avenue
Livermore, CA 94550
kollet3@llnl.gov

Companies

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

People

- Joseph Ecker - http://signal.salk.edu
- Steve Jacobsen - http://epigenomics.mcdb.ucla.edu/DNAmeth/ http://hhmi.org/research/investigators/jacobsen_bio.html
- Lisa Stubbs - http://genome.llnl.gov
- Matteo Pelligrini - http://mcdb.ucla.edu/MCDBfacultyindiv.php?FacultyKey=816
- Robert Marteinssen - http://cshl.edu/public/SCIENCE/martien.html
- Frederic Bushman - http://med.upenn.edu/micro/faculty/bushman.html
- Steven Henikoff - http://hhmi.org/research/investigators/henikoff.html

Software

- TileMap - http://biogibbs.stanford.edu/~jihk/TileMap/index.htm
- SIGnAL - http://signal.salk.edu

Further Reading

- Zhang X., *et al.* Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* 126(6):1189-201 (2006).
- Yamada K., *et al.* Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* 302(5646):842-6 (2006).
- Ji H., Wong W. H. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics* 21(18):3629-36 (2006).
- Ghosh S., Hirsch, H. A., Sekinger E., Struhl K., and Gingeras T. R. Rank-statistics based enrichment-site prediction algorithm developed for chromatin immunoprecipitation on chip experiments. *BMC Bioinformatics* 7:434 (2006).
- Schroder A. R. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* 110(4):521-9 (2002).