

New Bioinformatics Algorithm Predicts Forkhead Protein Plays a Role in Breast Cancer Estrogen Response

Dana-Farber's Shirley Liu and Lawrence Berkeley National Laboratory's David Nix weigh in on improved algorithms for ChIP-on-chip experiments

By Stacey Ryder

BOSTON and BERKELEY, October 1, 2005 — Dr. Shirley Liu, assistant professor of biostatistics at the Dana-Farber Cancer Institute, has developed a new algorithm for analyzing ChIP-on-chip data which enables the unbiased mapping of estrogen receptor (ER) binding sites along chromosomes 21 and 22. Sequence analysis shows that a forkhead protein works together with estrogen receptor to mediate steroid response in breast cancer cells.

Dr. David Nix, a computational scientist working on the Berkeley *Drosophila* Transcription Network Project at Lawrence Berkeley National Laboratory has also been using ChIP-on-chip technology to look at binding sites for transcription factors that are expressed in early *Drosophila* development.

“Eventually, our goal is to be able to predict the expression of genes based on sequence,” said Nix. “To do that,

we need to understand the rules that govern cis-regulatory modules. Based on sequence so far, we have found it most difficult to make meaningful predictions in many areas. Our hope is that we can take an unbiased approach by using ChIP-chip technology to look at where things are actually binding in the embryo.”

Liu recently spoke to Nix about the complementary roles of ChIP-on-chip

A close-up portrait of Shirley Liu, a woman with short dark hair, smiling warmly at the camera. She is wearing a light-colored top and a pearl earring.

Shirley Liu is an assistant professor in the Department of Biostatistics at the Dana-Farber Cancer Institute. Dr. Liu's research is focused on using computational methods to predict sites of transcription factor binding. Recently, Dr. Liu has been collaborating with Dr. Myles Brown's laboratory, using their microarray data to help make accurate predictions of transcription factor binding sites for the estrogen receptor.

experiments and bioinformatic algorithms. The two discussed:

- Methodologies used by Liu and collaborators
- Developing standards for processing and interpreting tiling microarray data
- Commercial and open source software available for motif finding

Analysis of ER Chromatin IP on Tiled Microarrays

Nix: Your algorithm tried to estimate probe behavior by obtaining an empirical measurement of noise. How well did the algorithm work?

Liu: It worked pretty well at least for our analysis of estrogen receptor ChIP-chip, and re-analysis of the p53 data from Cawley's *Cell* paper, out of Affymetrix. There were 48 sets of microarray data from Cawley's paper alone. Dana-Farber generated another 20 datasets on chromosomes 21 and 22. One thing we tried to improve on the Cawley analysis, is to consider all probe pairs, instead of only those with higher perfect match (PM) than mismatch (MM). Sometimes the mismatch might hybridize to other sequences in the genome, so it could have a much higher value. Half of the probes on the array have higher mismatch than perfect match values, so we could lose a lot of information if we don't consider them.

Nix: We've tried to use mismatch probes in our analysis, but find that they introduce a lot of noise into the data. In cases where you don't have a good control, like in some of the transcriptome experiments where you are putting mRNA/cDNA onto a chip, you don't have a good control that can be used to normalize for differential probe response. With ChIP-chip data, we can use input chromatin. By taking a simple ratio of treatment to input control chromatin, we can avoid the use of mismatch oligos for the ChIP-chip data.

Liu: We used PM-MM for the analysis just published, but you can get pretty good results from PM alone. Once we

identified the ChIP-enriched regions, we tried to look for the transcription factor binding motifs in these sequences. These regions are theoretically bound by transcription factors *in vivo*, so every sequence that's pulled down should have a binding site. We performed *de novo* motif finding, without any knowledge of what the motif might look like or checking any motif databases, to find enriched sequence patterns of widths 5 to 17 bp in the ChIP-enriched regions.

From the ER ChIP-regions, we found a very strong palindrome motif of width 15 which turned out to be the canonical ER binding motif. When we masked

“We are finding that we have to go back to ground zero to try to figure out how to process the data. We do stuff very differently than what folks have done before.”

out this motif to look again, another motif came out very strong around 9 to 10 bases. From my experience in analyzing yeast ChIP-chip experiments, I told our collaborator that there seemed to be a forkhead protein working with the ER to bind these regions.

The biologists checked a cancer microarray database called Oncomine, and found FoxA1—a member of the forkhead protein family—to be highly co-expressed with ER. They used an antibody against FoxA1 to pull down ChIP DNA and did site-directed quantitative PCR to check for the ER ChIP regions that contain the FoxA1 binding site. This procedure verified that those sites are indeed bound by FoxA1. When they used siRNA to knock out the FoxA1 protein, and conducted ChIP-chip against ER, the ER binding site was lost. So our recent *Cell* paper not only located the ER-bound region on chromosomes 21 and 22, but found a cofactor FoxA1 which is required by ER to bind to these regions.

Analysis of p53 Chromatin IP on Tiled Microarrays

Liu: We used a similar method to reanalyze the p53 ChIP-chip data in Cawley's *Cell* paper. We were able to identify many p53 bound regions, including all 10 previous qPCR-validated regions. In addition, we found a motif from the predicted p53-bound regions similar to the canonical p53 binding motif. This motif and the canonical p53 binding motif are both highly enriched in the predicted p53 ChIP regions compared to the genome background. They are also more enriched in our predicted p53 ChIP regions than in Cawley's pre-

dicted p53 ChIP regions. We thought this was pretty reassuring.

We found that p53 binding may not be conserved between humans and rodents. A lot of the regions we identified including some of the qPCR-validated regions do not align to the rodent genome at all. We talked to some colleagues, and they were not surprised. Rodents may have quite different p53 regulation from human.

Nix: There is not a whole lot of conservation in the bound regions we've been looking at either. They are not conserved and that's actually fitting with what people are reporting elsewhere. Several groups have looked at the *eve* locus in several *Drosophila* species, and really teased apart the enhancers. There's not a heck a lot of conservation at the sequence level, but functionally, they still work. It suggests that there are very different mechanisms by which these things are evolving.

A Question of GC-rich Content

Nix: We've noticed disturbingly in the last few weeks that our identified regions are enriched for GC content. And this is even found in the mock IP/IgG controls. The problem is that if you have a non-specific enrichment for GC right under the binding peaks, it's really going to skew your ability to detect novel "words." Have you seen this? And if you have, have you tried to control for it in any way?

Liu: GC-rich probes often give higher probe intensity, which might cause them to look "enriched." But these probes should also give high intensity in genomic input or mock IP controls, so the algorithm would know they are not really enriched.

Nix: Well, it does. For some reason, we are pulling down regions that are GC-rich. This may very well be an experimental artifact. We are using formaldehyde as a crosslinker and that is known to have a GC bias. Maybe these regions amplify better by PCR. They certainly hybridize better. We can get rid of most of the differences in hybridization by doing a ratio of treatment versus control. But it still doesn't get around the fact that we are seeing those increases.

Liu: GC-rich sequences and repeats often interfere with downstream motif finding. After retrieving the ChIP sequences, we ran three rounds of repeat masking. The first is RepeatMasker from the University of Washington, which most people use. The second is the tandem repeat masker, for which we downloaded the repeat coordinates from UCSC. The third is our own simple repeat masker which gets rid of single base (e.g. AAAAAAA) or double bases (e.g. ACA-CACACACA) repeats. With these repeats removed, we haven't had a problem with repeats or GC-rich sequences, at least in the ER and p53 data.

Creating Standards to Compare Methods

Nix: I think we need to have some standards by which we can compare different

data analysis methods. We are finding that we have to go back to ground zero to try to figure out how to process the data. We do stuff very differently than what folks have done before.

We recently produced some spike-in data from an experiment where we took BACs containing large 200 kb inserts and mixed these at varying concentrations with input chromatin. These were then hybridized to the tiling microarray. Then we measured the output of various tests to see how well they differentiated between calling a true positive and a true negative.

We think that by using such spike-in experiments, we can also get a handle on some of the pre-processing steps that might help with the data analysis.

We are hoping that the Hidden Markov Model that you folks have developed will be a step up. Because right now, we are not using a traditional statistical test. We are basically looking at simple ratios. And that gets into a difficult issue of how to establish a false positive rate. We create a list where we have all these regions of the genome that we think are bound and we rank them by the highest intensity ratio. Transcription factor binding sites tend to cluster at

the top, but we don't know how far down the list we can go.

Liu: The only concern I have is that spike-in DNA may not represent real ChIP DNA. In the real experiment, we are dealing with the whole genome DNA population and sonication is random.

Nix: I should have mentioned that we did everything the same with the BACs as with an IP. They were sonicated, random primed, and amplified. There's going to be bias for fragment sizes, amplification, etc.

I think such BAC spike-in experiments are a good first step towards developing standardized datasets for use in evaluating different tests and data processing methods, but clearly they are not the end.

Liu: Eric Lander's group published a ChIP-chip paper using tiled arrays (Bernstein et al.) this spring, in which they did qPCR to validate their ChIP-enrichment predictions. Many regions below the prediction cutoff still have pretty high qPCR fold enrichment. So current methods might still give many false negatives, but we will see many improvements on correct fold change prediction in the future. The most important part now is probably to



David Nix is a computational scientist in the Department of Genome Sciences at Lawrence Berkeley National Laboratory. He is part of a large multidisciplinary group called the Berkeley *Drosophila* Transcription Network Project (<http://bdtnp.lbl.gov>) that is currently working with scientists at Affymetrix to use ChIP-on-chip technology to identify binding sites for transcription factors involved in early *Drosophila* development.

estimate probe behavior by considering probe sequence.

Open-source Bioinformatics Software

Nix: Is your Hidden Markov Model (HMM) software available?

Liu: We have a package already which people could download. At the same time, Stratagene, a designated third party software developer for Affymetrix chips, is working to incorporate our analysis algorithm into their software. The light version of the Stratagene software is free, but they charge some money for the advanced version. A third effort is to work with John Quackenbush, a professor DFCI recently recruited from TIGR, to incorporate our algorithm into TM4, an open source Java software package. Compared to the R open-source package BioConductor, TM4 has a more graphical interface and does not require any coding from users. It's more for biologists to use but informatics people can look at the source code and modify it to do whatever they want.

Nix: We also have developed a large Java open-source package. It's called

TiMAT for Tiling Microarray Analysis Tools. The whole package is available as open-source. I built it to process our *Drosophila* tiling data but haven't had an opportunity to test it with other datasets. It would be great to add an HMM option into TiMAT.

Liu: We would love to incorporate our software into your system. The only concern right now is we may not have enough experiments to estimate the probe behavior for the algorithm to work in *Drosophila* or the whole human genome. Instead of using other experiments, since the whole human genome is 44 million probes, we hope to estimate probe behavior by looking at other probes in the same experiment that have similar probe sequence or hybridization quality. This way, we can just do the estimate from a single experiment instead of using other experimental data that may or may not be there. Once this is ready, it'll be easier for us to integrate our software into TiMAT.

Nix: We can also provide you with a lot of *melanogaster* chip information. We've produced a data set from 30 to

40 individual chips from several different ChIP-chip experiments. Data from another 30 odd chips are also available from a set of experiments performed in collaboration with Tom Gingeras' group. The whole genome of *Drosophila* is covered on a single chip featuring about 3 million oligos. It's nice because it covers the entire spectrum of where transcription factors bind on the genome level.

AFFYMETRIX MICROARRAY BULLETIN

Editorial Staff

Wes Conard, *Editor-in-Chief*

wes_conard@affymetrix.com

Tommy Broudy, *Managing Editor*

thomas_broudy@affymetrix.com

Rachel Shreter, *Editor*

rachel_shreter@affymetrix.com

Kamalia Dam, *Associate Editor*

Stacey Ryder, *Associate Editor*

Daniel Noble, *Copy Editor*

Michelle Majewski, *Contributing Designer*

FOR MORE INFORMATION

Contact Information

■ Xiaole Shirley Liu, Ph.D.
Department of Biostatistical Science
Harvard School of Public Health /
Dana Farber Cancer Institute
44 Binney Street, Mayer 1B22,
Boston, MA 02115
xslu@jimmy.harvard.edu

■ David Nix, Ph.D.
Lawrence Berkeley National Laboratory
Department of Genome Sciences
1 Cyclotron Road, Mailstop 84-171
Berkeley, CA 94720
DANix@lbl.gov

Companies

■ Affymetrix Inc. –
<http://www.affymetrix.com>
■ Stratagene
<http://www.stratagene.com>

Organizations

■ Dana Farber Cancer Institute –
<http://dana-farber.org>
■ Lawrence Berkeley National Laboratory –
<http://www.lbl.gov/>

■ Berkley *Drosophila* Transcription Network
Project – <http://bdtnp.lbl.gov/>
■ UCSC Genome Bioinformatics –
<http://genome.ucsc.edu>

Tools

■ Oncomine Cancer Microarray Database –
<http://www.oncomine.org>
■ Bioconductor
<http://www.bioconductor.org>
■ TM4
<http://www.tm4.org>
■ TiMAT
<http://bdtnp.lbl.gov/TiMAT>
■ RepeatMasker
<http://www.repeatmasker.org>

Further Reading

■ Cawley, S. et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of non-coding RNAs. *Cell* 2004; 116(4): 499-509.
■ Hong P, Liu XS, Zhou Q, Lu X, Liu JS, Wong WH. A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics*. 2005; 21(11):2636-43.

■ Li W, Meyer CA, Liu XS. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics* 2005; 21 Suppl 1:i274-i282.
■ Nix, D.A, Li, X.Y, Moses, A.M., Eisen, M.B., Biggin, M.D. Development and optimization of tiling microarray analysis tools for chromatin immunoprecipitation. Manuscript in preparation. 2005
■ Bernstein B. et al. Genomic Maps and Comparative Analysis of Histone Modification in Human and Mouse. *Cell* 2005; 120(2): 169-181.

People

■ Myles A. Brown
<http://www.dana-farber.org/abo/danafarber/detail.asp?PersonID=22&RD=True>
■ Eric Lander
<http://web.mit.edu/biology/www/facultyareas/facresearch/lander.shtml>
■ John Quackenbush
<http://www.hsph.harvard.edu/faculty/JohnQuackenbush.html>