

## Thousands of Previously Unknown Transcription Factor Binding Sites Mapped Using Tiling Microarrays

Harvard's Kevin Struhl and Dana-Farber's Myles Brown discuss ChIP-on-chip as an unbiased method for transcription factor binding site identification

By Stacey Ryder

**BOSTON, October 1, 2005** — Scientists at the Dana-Farber Cancer Institute and Harvard Medical School are using chromatin immunoprecipitation (ChIP) combined with tiled microarrays, or ChIP-on-chip, to discover thousands of previously unknown transcription factor binding sites that regulate gene expression. Their findings reveal new information about the basic processes of transcription and potential regulatory targets for breast cancer treatments.

Myles Brown's research team in the Department of Medical Oncology at the Dana-Farber Cancer Institute is using ChIP-on-chip to define the gene expression pathways controlled by steroid hormones and their receptors, including the estrogen receptor in breast cancer and the androgen receptor in prostate cancer.

"When you get data using this technique, it's much more powerful than purely computational methods that can

be based on mistaken assumptions," said Brown. "I think this technique tells us a lot more about how proteins find their targets in DNA and shows that the previous models were oversimplifications."

Dr. Kevin Struhl, in the department of Biological Chemistry and Molecular Pharmacology at Harvard Medical School, studies gene regulation in yeast and has recently begun to focus his studies in humans. He believes that



**Myles Brown** is chief of the

Division of Molecular and Cellular  
Oncology at the Dana-Farber Cancer

Institute. His research focuses on the roles  
of estrogen receptor in breast cancer and  
androgen receptor in prostate cancer.

Researchers in his laboratory have been  
using ChIP-on-chip technology to study  
the details of gene regulation in cancer cells.

ChIP-on-chip methods lead to relatively unbiased results.

“I think a lot of the previous predictions were based on data from quite biased experiments,” said Struhl.

“When you look at things in an unbiased fashion with a relatively unbiased technique, you find a lot of things that you didn’t expect.”

The AMB recently spoke with Brown and Struhl about ways in which ChIP-on-chip technology are changing ideas about binding sites and gene regulation. The two discussed:

- Discrepancies between the predicted number of binding sites and what they find in microarray experiments
- The role of DNA motifs in transcription factor binding
- Limitations and advantages of different technologies for studying binding sites

#### Surprising Numbers for Real Binding sites

*Both Struhl and Brown have found major differences between the actual and predicted number of transcription factor binding sites using ChIP-on-chip technology.*

**Struhl:** Chromatin immunoprecipitation is a powerful technique for locating binding sites, especially when combined with unbiased, tiled microarrays. It

allows you to see where the proteins are associating and it’s quite open to detecting interactions even if they’re just associated indirectly with the target region. The nice thing about a tiled microarray, is that it allows you to ask where the proteins are associating in an unbiased fashion. We found that for some factors, like Sp1 and Myc, there are a large number of sites. We didn’t find that many sites for p53. Until you do the experiment you really don’t know.

**Brown:** What we found for the estrogen receptor is similar to what you found for p53. I think it may suggest that when different transcription factors have their authentic sites mapped by this technique, they will fall into different categories: those that seem to have relatively few targets and those that have a lot more targets.

**Struhl:** The classical view is that you have a DNA sequence motif and that motif is recognized by a protein, and once that protein is bound there, it does something transcriptionally. The reason that classic view exists is because people have studied genes, during which they mapped protein-binding sites and found motifs. So that’s the view from a very biased perspective. Only a few motifs look like the “good sequences.” These are sequences that have a good-looking motif and bind *in vitro*. But they don’t bind *in vivo*. Only

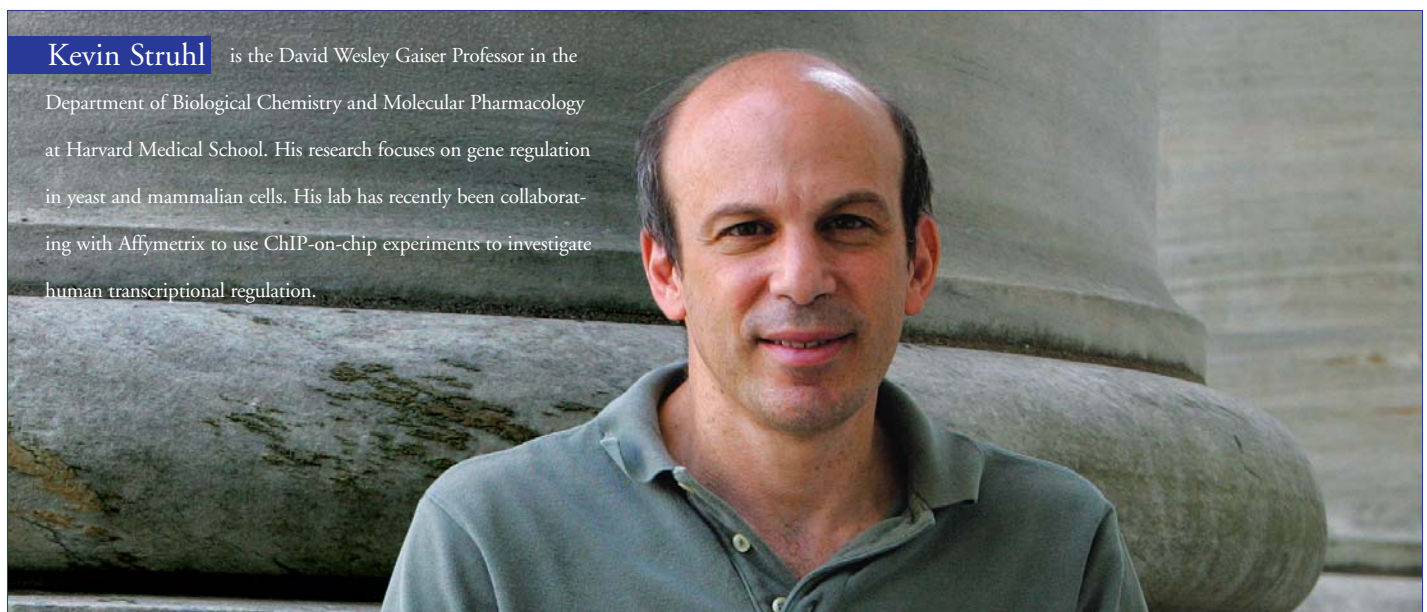
a small percentage of good sequences are bound *in vivo*. The flip side of that is that not every real *in vivo* binding site actually has the motif. So the original dogma is dramatically wrong on both sides.

#### Binding Requires More than Motifs

**Struhl:** We’ve done a little bit of work with *E. coli*. Unlike in eukaryotes, in *E. coli*, if you have a good motif, you are bound virtually 100 percent of the time. Interestingly, even in *E. coli*, it turns out that about half the sites that we mapped don’t have the motif at all. They don’t bind *in vitro*, but they are clearly bound *in vivo*. And so even in an organism that you think would be the bioinformaticist’s dream, roughly half the sites for the one factor we’ve studied in detail don’t have the motif, but still bind.

There is an overrepresentation of motifs in the real *in vivo* targets. So, it’s not like the motif is irrelevant, but it certainly is only part of the story.

**Brown:** That’s what we found as well. In fact, we found that most of our sites had either perfect estrogen receptor binding sites or half sites, suggesting that the motif is a necessary component of the binding we saw, or nearly necessary component, but it’s not sufficient information to find where the true binding sites are. And that’s probably going to depend on the factor. Some



**Kevin Struhl** is the David Wesley Gaiser Professor in the Department of Biological Chemistry and Molecular Pharmacology at Harvard Medical School. His research focuses on gene regulation in yeast and mammalian cells. His lab has recently been collaborating with Affymetrix to use ChIP-on-chip experiments to investigate human transcriptional regulation.

“When you have a new, powerful technique, it reveals the biases you had in your old experiments.”

factors are going to care more about their motif than others. For ER, we found that it required not only its own response element, but also the forkhead factor FoxA1 to define a true binding site.

#### Searching the Whole Genome for Regulatory Elements

*Struhl and Brown discuss the types of experiments needed to further elucidate the role of DNA binding and the contribution of microarray analysis to those studies.*

**Struhl:** We are currently part of the ENCODE Consortium, which was formed to identify all of the functional elements of the human genome. The first stage of the Human Genome Project was, of course, sequencing the human genome, but you can now imagine a project where you basically did ChIP-chip on a lot of transcription factors to map the regulatory elements of the human genome. I'm sure both of us believe we could do it right now. But convincing the powers that be that one can do this and getting them to devote serious funding to it, is a different matter. So, this ENCODE Consortium includes a number of labs that hope these studies can be done on a pretty large scale. There aren't all that many transcription factors. There may be a thousand or so. I don't think we could do a thousand. That's probably too expensive. We don't have all the antibodies, but certainly one could do a couple of hundred at this point. And that would be incredibly useful.

**Brown:** I think the technology is available to do the whole genome, and so we are in the middle of a collaboration with Tom Gingeras at Affymetrix on ER binding across the whole genome. It is the time to do one's transcription factor of choice. I think this, coupled

with expression arrays, is going to be the direct output of transcription factor action. If we want to understand regulation, this is what's needed.

#### Comparing Approaches to Binding Site Studies

*ChIP-on-chip experiments generate detailed genome-wide data that was previously impossible to obtain. Struhl and Brown weigh in on the advantages of using microarrays with complementary technologies like siRNA for whole-genome discovery of gene regulation.*

**Struhl:** With polymerase, one of the nice things this array tells you is which mechanism is used more often and which genes use which mechanism. That information would be difficult to get otherwise.

**Brown:** And that's been a problem in the estrogen receptor field. People have made large conclusions studying a very small number of genes. For any given transcription factor, only a few targets have been studied in any detail. This array allows us to study all the targets in detail.

**Struhl:** Something that's nice about chromatin IP, is that it can be used in any organism. And what's particularly nice is that it's a quantitative assay for the wild type cell. That's a rare thing to be able to do. One of the problems in the mammalian field is that, unlike in yeast, doing good genetics is hard. In yeast, making a gene knockout is trivial and it's a powerful approach. It is doable in mammalian cells, but it's difficult.

**Brown:** I think siRNA has allowed that to be done more. Approaches like you're using—doing siRNA followed by an expression profile—is also extremely useful for mammalian cells. It's just surprising how few transcription factors have many known direct target genes.

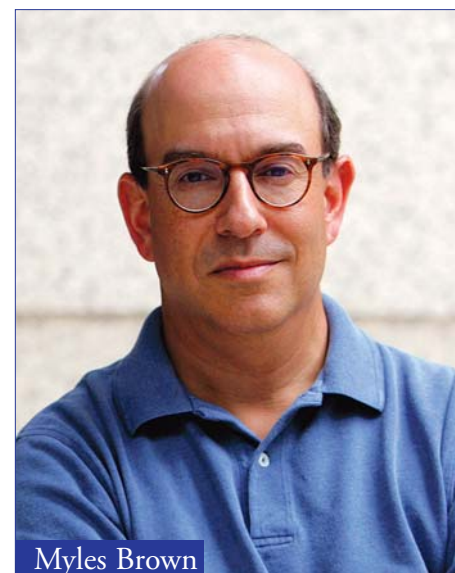
**Struhl:** I think most of the other *in vivo* assays that have been used are things like transient transfections, and those are prone to artifacts for two reasons. First, the template you're putting in doesn't look like a normal template—it's not real chromatin. And second, when people are looking at the effect of their favorite factor, they will toss in a plasmid that expresses their favorite factor, and they're generally grossly over-expressing it. So you see what your protein can do if you grossly over-express it, but not what it does in the real world.

**Brown:** I think transient transfection is an important technology, but to a large degree, chromatin immunoprecipitation and ChIP-chip certainly gives superior answers in terms of actual targets and the actual proteins involved in regulation. Chromatin IP has allowed people to look at endogenous proteins at their normal levels, and people have even looked at them from tissues, so you really can get a true look at what's happening under normal physiology.

#### Limitations of Using Antibodies

*While ChIP-on-chip studies have provided high-resolution views into the genome, a major limitation is the quality of antibodies available for these studies.*

**Brown:** Unfortunately, having good antibodies is often the limiting factor in IP experiments. But I think a lot of companies and a lot of people are



Myles Brown

developing good antibodies to mammalian transcription factors. You can use tagged proteins, and we have done that for chromatin IP. In mammalian biology in general, it's still surprising how often having good antibodies to your protein is a rate limiting step.

**Struhl:** Yes. I think there's another issue with antibodies, which is a difficult one to assess. You assume an antibody is specific for your protein, but it's difficult to know that. You can certainly show that it pulls your protein down nicely and that it works for a ChIP, but it's difficult to know whether you're pulling something else down that you don't know about.

**Brown:** We've been concerned about that issue too. Early on, we used multiple different antibodies for the estrogen receptor. Another limitation of ChIP is the possibility that certain epitopes might be masked under certain conditions, because these proteins are in large complexes. You could imagine that an antibody that can recognize estrogen receptor might not work well in ChIP because the epitope is masked.

#### Advantages of Using Microarrays

**Struhl:** I think reviewers often mention these theoretical possibilities to give you trouble. The technique is not perfect. It does have its theoretical issues about what you're really measuring. But by and large, I would say that compared to most techniques, it's quite robust.

**Brown:** I agree completely. For us it's revolutionized what we can look at.

**Struhl:** We're probably among the first people to do ChIP-chip for transcriptional regulation. And in our yeast work, every time we did a new experiment, just using a ChIP on a standard protein, we found new things. Either we answered a major question or we found something completely unexpected. That happened frequently. When you have a new, powerful technique, it reveals the biases you had in your old experiments.

**Brown:** People have made good use of the promoter arrays, but the unbiased whole-genome tiled arrays eliminate the bias you have when you look only at particular regions of the genome. And what we found, for example, is that ER is mostly not at the promoters. So promoter arrays would have missed most of the sites of estrogen receptor binding and functional regulation.

**Struhl:** And for us, we found a similar thing in a slightly different way. Especially for things like Sp1 and Myc, which are often thought of as fairly proximal factors, we found that only about 20 percent of all the sites were in classical promoters. Now a lot of that is turning out to be because there is a lot more transcription than people thought. And a lot of these sites are in places that drive these so-called unusual transcripts. There's no question that if you just use a biased array, you're going to confirm your own biases. You're going to miss a lot of important stuff.

#### High Background in ChIP-on-chip Experiments

*Scientists have consistently been faced with high background problems in ChIP-on-chip experiments. Figuring out ways to lower that background and increase sensitivity are critical to discovering regulatory sites that are currently dismissed as noise.*

**Brown:** We were lucky to have wonderful collaborators who helped us enormously with these studies. We have an extremely good Microarray Core at the Dana-Farber headed by Ed Fox that provided us with very high quality data. Pam Silver's group in our Department of Cancer Biology has been a pioneer in the use of ChIP-chip in both yeast and mammalian cells and were instrumental in getting us started in this area. In addition, Shirley Liu's group in the biostatistics department at Dana-Farber has done a lot of work on developing methods for data analysis for ChIP-chip experiments. Using her methods, we came up with a list of binding sites. Then we did some experiments, because I'm an experimentalist. I like having a way of validating the list, so

we got an ordered list of sites and then we determined where we thought the boundary between real and background might be. We did directed ChIP experiments to define where the cut-off should be. But, ultimately in these types of experiments, you're always going to have some false discovery rate. It's inevitable when you go to these genome-wide analyses.

**Struhl:** I think this issue is important, but it makes you too conservative. The real problem is in the control sample. You're trying to get correct hybridization, so you must be able to pick it up at one part in 108, and that's asking a lot for a biochemical experiment. That's what causes a lot of the background. As a result, you miss more things. In other words, when you see something that's significant, it's usually pretty significant. So a lot of times we'll see something that passes a cut-off, at least by a computational method. Then you do an experiment, you find its 30-fold enriched. It is a non-trivial issue. It's relatively easy to pick out the best sites. If you want to go down on the list, then it gets a little more tricky. And I think the most serious issue occurs when you want to compare different samples and you want to do more quantitative stuff. So, if you have a situation where you're looking at condition X and condition Y and want to know whether something is down 3-fold here or there, that's where it gets a little more difficult.

## 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*

**Contact Information**

■ Kevin Struhl, Ph.D.  
 Department of Biological Chemistry and  
 Molecular Pharmacology  
 Harvard Medical School  
 Building C1, Room 315  
 240 Longwood Avenue  
 Boston, MA 02115  
[Kevin@hms.harvard.edu](mailto:Kevin@hms.harvard.edu)

■ Myles A. Brown, M.D.  
 Department of Medical Oncology  
 Dana-Farber Cancer Institute  
 44 Binney Street  
 Dana 730  
 Boston, MA 02115  
[myles\\_brown@dfci.harvard.edu](mailto:myles_brown@dfci.harvard.edu)

**Companies**

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

**Organizations**

■ Harvard Medical School  
<http://hms.harvard.edu>  
 ■ Dana-Farber Cancer Institute  
<http://dana-farber.org>  
 ■ The ENCODE Project  
<http://www.genome.gov/ENCODE>  
 ■ The Human Genome Project  
[http://www.ornl.gov/sci/techresources/Human\\_Genome](http://www.ornl.gov/sci/techresources/Human_Genome)

**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.  
 ■ Geisberg JV, Struhl K. Quantitative sequential chromatin immunoprecipitation, a method for analyzing co-occupancy of proteins at genomic regions in vivo. *Nucleic Acids Res.* 2004; 32(19):e151.

■ Grainger DC, Overton TW, Reppas N, Wade JT, Tamai E, Hobman JL, Constantinidou C, Struhl K, Church G, Busby SJ. Genomic studies with *Escherichia coli* MelR protein: applications of chromatin immunoprecipitation and microarrays. *J Bacteriol.* 2004; 186(20):6938-43.  
 ■ Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoutte J, Shao W, Hestermann EV, Geistlinger TR, Fox EA, Silver PA, Brown M. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 2005; 15;122(1):33-43.

**People**

■ Xiaole Shirley Liu  
[xsliu@jimmy.harvard.edu](mailto:xsliu@jimmy.harvard.edu)  
 ■ Tom Gingeras  
[Tom\\_gingeras@affymetrix.com](mailto:Tom_gingeras@affymetrix.com)  
 ■ Ed Fox  
<http://chip.dfc.harvard.edu/lab/services.php>  
 ■ Pam Silver  
<http://research.dfc.harvard.edu/silverlab/>