

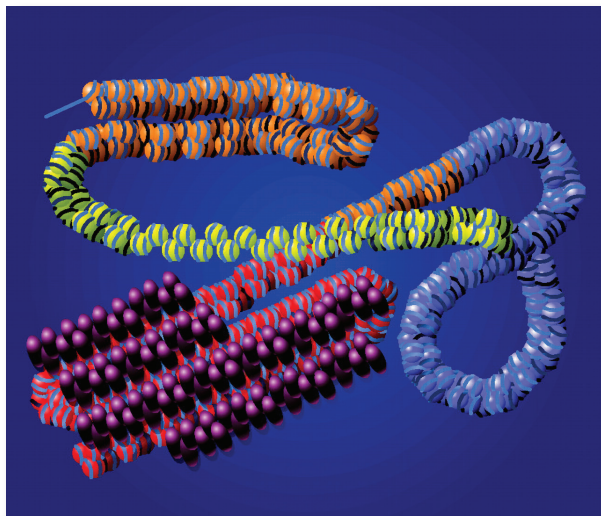
Genomics

Epigenomics:

The New Technologies of Chromatin Analysis

Multicellular organisms are essentially clonal. Every cell possesses the same DNA as every other. So what distinguishes a liver cell from a neuron? Epigenetics, that constellation of noncoding RNAs, protein-DNA interactions, and molecular modifications that govern which genes are expressed and which stay silent. Epigenetic mechanisms influence processes from stem cell differentiation to cancer, and researchers are keen to understand how these events differ at the genomic scale—the so-called epigenome. The problem is daunting, but the research community is resourceful. The epigenome has never seemed closer.

By Jeffrey M. Perkel



“We’ve got the technology, we’ve got the need, people are starting to do this, the lack of reference sets and new technologies are holding the field back. That was why [epigenomics] was identified as a good investment.”

In early 2008, the **U.S. National Institutes of Health (NIH)** announced that it was earmarking \$190 million over five years to study the problem of epigenomics. The effort, part of the NIH Roadmap Initiative, had several overarching goals, including creating a series of epigenomic reference maps for normal human cells and tissues and developing novel technologies to aid in that process.

According to James Anderson, director of the Division of Program Coordination, Planning, and Strategic Initiatives, the unit within the NIH’s Office of the Director that oversees the Common Fund (and hence, the Roadmap Initiative), epigenomics was a natural fit for the Roadmap, which is a cross-NIH funding mechanism that essentially concerns itself with grand challenges in the biological sciences.

Previously, he explains, researchers were attacking the epigenome piecemeal, but nobody could put it all together. After consulting with experts, NIH realized the field was fundamentally stymied by the lack of one essential resource: a reference dataset, an epigenomic metric against which other datasets might be measured. Without such a reference, a complete cataloging of all epigenetic marks and how they vary across development and disease could not possibly be completed. Yet at the same time, new technologies had been developed that for the first time meant the problem was not actually intractable, simply vast.

NIH decided to pull the trigger. “It all came together,” Anderson says. “We’ve got the technology, we’ve got the need, people are starting to do this, the lack of reference sets and new technologies are holding the field back. That was why [epigenomics] was identified as a good investment.”

Today, that labor is beginning to bear fruit. The NIH Common Fund, along with individual institutes and centers, has awarded 68 grants under the Epigenomics Program, which according to Anderson have yielded some 52 reference epigenomes—maps of DNA methylation and histone modifications across multiple cell types. (Those datasets join the fruits of an earlier, parallel effort, the National Human Genome Research Institute-funded ENCODE project (Encyclopedia of DNA Elements), which in September 2012 released 30 papers map-

ping not just DNA methylation and histone modifications, but also transcription-factor binding sites, higher-order chromatin structure, transcribed regions, and more across the human genome in nearly 150 cell lines; both those and the NIH Roadmap Epigenome Project datasets are freely accessible online.) But perhaps just as importantly, they have led to a slew of new epigenetic and epigenomic technologies that are providing researchers the tools to gain an increasingly clearer picture of what is really going on in cells at the genomic level.

Indeed, says Anderson, that’s really the point of spending all these millions. “Our intent is not to finish the epigenome. It is to transform individual investigators’ ability to do their work.”

ATTACKING THE METHYLOME

One researcher supported under the Epigenomics Program is Bing Ren, a member of the **Ludwig Institute for Cancer Research** in San Diego. Ren is principal investigator (PI) of a grant to establish one of four epigenome mapping centers charged with compiling the critical epigenomic maps. His center focuses on embryonic stem cells. The **San Diego Epigenome Center** has been awarded \$15.7 million since 2008, which it has used to map both DNA methylation and some 20 histone modifications in both human embryonic stem cells (hESCs) and four hESC-derived cell types.

The significance of the Epigenome Project “is equivalent to sequencing the human genome,” Ren says. “When you have the human genome, then you have a blueprint to understand human development. But without a detailed understanding of the epigenome we can’t read that blueprint.” *continued>*

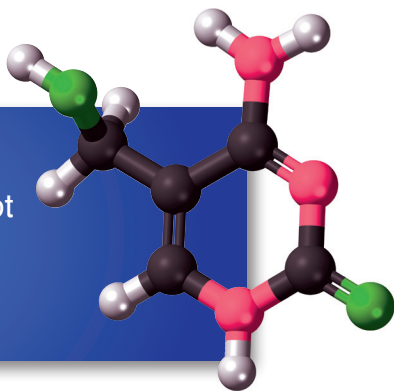
UPCOMING FEATURES

Tissue Engineering: 3-D/Scaffolding—December 7

The Connectome—January 18, 2013

Genomics—February 15, 2013

Bisulfite sequencing,
as it turns out, cannot
distinguish between
5-mC and 5-hmC.



The San Diego Epigenome Center builds its maps with the two key technologies of epigenomics: chromatin immunoprecipitation (ChIP)-Seq, which uses next generation DNA sequencing technology to identify the location of specific histone modifications across the genome, and MethylC-Seq, a genome-wide method for determining the position of 5-methylcytosine modifications.

MethylC-Seq is basically an optimized version of bisulfite sequencing for today's blazing-fast next-gen DNA sequencers. The problem it solves is this: Standard DNA sequencing methods cannot distinguish cytosine from 5-methylcytosine (5-mC). But if the DNA is first treated with sodium bisulfite they can, because bisulfite converts unmodified cytosines to uracil, which appears in DNA sequencer reads as thymine (T). By comparing bisulfite-treated samples against an untreated control, researchers can determine which bases were methylated and which were not.

Researchers have been using bisulfite conversion to interrogate methylation at the nucleotide level for decades, and in 2008 Joseph Ecker's team at the Salk Institute in San Diego (Ecker is also an investigator in the San Diego Epigenome Center) updated the method for the **Illumina** Genome Analyzer. That's MethylC-Seq. But in 2009 a new wrinkle appeared. That year, teams led independently by Nathaniel Heintz at the Rockefeller University in New York and Anjana Rao at Harvard Medical School reported that mammalian DNA contains a previously undiscovered methylated base, 5-hydroxymethylcytosine (5-hmC).

Bisulfite sequencing, as it turns out, cannot distinguish between 5-mC and 5-hmC, meaning that at least some sites reported as containing the former, may in fact contain the latter.

"What it means to the scientific community is that whatever information we had before is not true, because we don't know what percentage of the apparent 5-methylcytosines are actually 5-hydroxymethylcytosine," says Sriharsa Pradhan, head of the RNA Biology division at **New England Biolabs**, which sells restriction enzyme-based kits to distinguish between the two bases.

This year, researchers finally developed strategies to circumvent this problem. The first, developed by a team in Cambridge, UK, and called oxidative bisulfite sequencing (oxBS-Seq), uses an oxidizing reagent (potassium perruthenate) to oxidize 5-hmC residues to 5-formylcytosine (5-fC), which reads as T after bisulfite conversion.

The second method, developed in a collaboration between Ren's lab, Chuan He at the University of Chicago, and Peng Jin at Emory University, uses an enzyme to selectively protect 5-hmC residues. Called Tet-assisted bisulfite sequencing (TAB-Seq, commercialized by a Chicago-area firm named **WiseGene**), this method uses a ten-eleven translocation (Tet)-family oxidase enzyme to convert 5-mC to 5-carboxylcytosine (5-caC), which also reads as T after bisulfite

treatment. (The Tet enzyme progressively oxidizes 5-mC to 5-hmC, and then to 5-fC, and finally to 5-caC.)

First though, TAB-Seq uses β -glucosyltransferase to couple a glucose moiety to 5-hmC, protecting it from Tet. Thus, the only residues that should appear as cytosines during sequencing should be 5-hmC. Comparison with standard bisulfite-converted and sequenced DNA should reveal the balance of 5-mCs. (New England Biolabs' EpiMark 5-hmC and 5-mC Analysis Kit is based on a similar principle; it uses β -glucosyltransferase to render a sequence resistant to a restriction enzyme.)

Ren and He's team used TAB-Seq to decipher the methylome of human embryonic stem cells, identifying some 691,000 5-hmC sites. Based on the distribution of that epigenetic mark, Ren says, it appears that 5-hmC plays a role in regulating transcriptional enhancers. "This type of element has a high abundance of hydroxymethylcytosine," he says, "and a correspondingly lower level of methylcytosine in the same sequence."

New England Biolabs is working on an alternative method to interrogate 5-hmC directly. The company recently described the enzymatic properties of the PvuRts11 family of proteins, which binds 5-hmC (or its glucosylated form, 5-(β -glucosyloxymethyl)cytosine) and cleaves 9 to 13 bases on either side, releasing a 24-base fragment with the modified base in the center. These fragments can then be sequenced directly, an approach the company calls "ABASeq," ("like the musical group, but only one B," Pradhan quips) in honor of AbaS1, the PvuRTS11 family member used in the assay.

"You don't need a bisulfite conversion; you don't need any kind of Tet-based approach or oxidation-based approach," Pradhan says. "Your sequence output is just going to align with the genome sequence." According to Pradhan, the team has already used this approach to map 5-hmC residues in a mouse embryonic stem cell line, though those data are not yet published.

CATALOGING HISTONE MODIFICATIONS

Another recipient of NIH Epigenome Project funding is Brian Strahl, associate professor of biochemistry and biophysics at the **University of North Carolina (UNC) School of Medicine**. With UNC colleague Xian Chen, Strahl submitted an application focusing on the discovery of novel epigenetic marks.

"One of the questions we wanted to address is whether there were novel sites of histone modification that had gone undetected," Strahl explains. "This is relevant because to really understand epigenomics, or even epigenetics, you need to know first what are all the modifications on histones to begin with."

Put another way, you cannot map modifications you don't know exist. Those can be of two types: known modifications in novel locations, and novel modification types.

To find both types, many researchers turn to mass spectrometry. Strahl and Chen, for instance, have used top-down proteomics analyses on a **Bruker Daltonics** 12-Tesla Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer to show that histone H2B lysine 37 in *Saccharomyces cerevisiae* contains a previously unknown modification.

"One of the peaks that came out ... was, as far as we can tell, dimethylated on one particular lysine that had not been reported elsewhere," Strahl says. "Unfortunately, we couldn't link any particular biology to it; it's just too new."

That's not to say the modification isn't important, he says. "If the cell cares that much to burn so many ATPs to get **continued**>

Genomics

a particular modification on a residue, it's got to be there for a reason," he says.

Researchers are also discovering entirely novel modifications. One team that has made several such discoveries is led by Yingming Zhao, a professor in the **Ben May Department for Cancer Research at the University of Chicago** and another Epigenome Project grant recipient.

Using high-resolution mass spectrometry, Zhao has discovered several new posttranslational modifications on histone proteins, including lysine propionylation and butyrylation in 2007, lysine crotonylation in 2011, and earlier this year, lysine succinylation and malonylation.

Zhao's discovery of lysine crotonylation is actually a case study in why researchers should always verify what the computer tells them. In this case, that due diligence yielded a high-profile paper in *Cell*.

At the time, Zhao's lab had already discovered lysine butyrylation. Now, using a high-end **Thermo Scientific** LTQ Orbitrap Velos system, they were trying to map sites of that modification. Normally in this type of study, researchers rely on computers to chew through the data and map observed ion masses against possible modifications. It's simply too laborious to do it manually. But computers can make mistakes, so Zhao's team double-checks the computer's math.

When they checked the spectral assignments in this case, they noticed that some didn't quite match up—they were off by 2 daltons (Da). Looking more closely, they were able to narrow down the modification's molecular formula to C_4H_5O , a crotonyl group.

Using a homemade "pan-crotonyl" antibody, Zhao's team used ChIP-Seq to tackle the mark's distribution throughout the genome, and found that it is associated with transcriptional start sites, enhancers, and active genes, and also "plays a role in the reprogramming of gene expression in postmeiotic male germ cells," he says.

OF READERS AND DOWNSIZING

Of course, a histone modification is just that: a modification. It's like a genomic street sign, and signs don't exist in a vacuum. There must also be proteins that add and remove those signs, and "reader" proteins that interpret what they mean.

To find those readers, researchers like C. David Allis, head of the Laboratory of Chromatin Biology and Epigenetics at **Rockefeller University**, sift through protein extracts, looking for activities that can recognize, add, or remove a given modification. The key, says Allis: "Fractionate, fractionate, fractionate." Using that strategy, Allis says his team has begun to home in on what they believe are a family of enzymes that can add a crotonyl group to histones—that is, histone crotonylases.

The results are not yet published, so Allis is fairly tight-lipped. But he did reveal that "it has a functional sort of twist to it, some

FEATURED PARTICIPANTS

Ben May Department for Cancer Research, University of Chicago
benmay.uchicago.edu

Bruker Daltonics
www.bdal.com

Cornell University
www.cornell.edu

Illumina
www.illumina.com

Ludwig Institute for Cancer Research, UCSD
www.ludwigsd.org

New England Biolabs
www.neb.com

NIH Common Fund Office of Strategic Coordination Epigenomics Program
commonfund.nih.gov/epigenomics

The Rockefeller University
www.rockefeller.edu

San Diego Epigenome Center
epigenome.ucsd.edu

Stanford University
www.stanford.edu

Thermo Fisher Scientific
www.thermoscientific.com

University of North Carolina at Chapel Hill School of Medicine
www.med.unc.edu

University of Pennsylvania Perelman School of Medicine
www.med.upenn.edu

WiseGene
www.wisegeneusa.com

personality ... that looks very exciting and different from what has been well-accepted for acetyl-lysine."

Or Gozani, associate professor of biology at **Stanford University**, another Epigenome Project grant winner, uses an alternate strategy for reader identification, probing microarrays of modified histone peptides with purified candidate reader proteins. Currently, Gozani's arrays contain about 100 peptides, and in one recent study his team, in collaboration with Dinshaw Patel at Memorial Sloan-Kettering Cancer Center in New York, used them to determine that a protein associated with DNA replication called ORC1 binds specifically to dimethylated lysine-20 on histone H4.

"There's a lot of room left to discover new readers," Gozani says.

And there are a lot of new methods in the epigenomics application space to study them. But that doesn't mean the field has achieved technological maturity, says Kenneth Zaret, codirector of the epigenetics program at the **University of Pennsylvania School of Medicine**. "Base technologies" like ChIP-Seq work best with immortalized cell lines that can provide the hundreds of thousands or even millions of cells required to make that technique work; when sample size is limited, during stem cell development or embryogenesis, for instance, these techniques are harder to pull off. What is needed, Zaret says, is a way to apply epigenomics approaches to smaller cell populations.

Already, he and others are making headway. **Cornell University** Professor Paul Soloway, with colleague Harold Craighead, has developed a nanofluid approach called SCAN (single chromatin analysis at the nanoscale) to monitor groups of modifications simultaneously on anywhere from one to 10 nucleosomes—asking, for instance, whether a single nucleosome contains both H3K27-trimethyl and methylated DNA.

Zaret is using fluorescence-activated cell sorting to isolate discrete cell populations, which he then analyzes using a modified ChIP protocol. Applying that approach to nine transcriptionally silent genes in a few thousand mouse stem cell progenitors, Zaret's team discovered distinct "prepatterns" that appear to position different sets of genes in different ways. Now the team is scaling this approach up to the genomic level.

Look for these data and more from the NIH Roadmap Epigenome Project in the months and years ahead. In the meantime, those hoping to mine the epigenome datasets can do so today at the Project's official data-coordination website, www.genboree.org/epigenomeatlas.

Jeffrey M. Perkel is a freelance science writer based in Pocatello, Idaho.
DOI: 10.1126/science.opms.p1200069