Making sense of chromatin states

Monya Baker

Researchers find new pieces in the puzzle of genome regulation.

Inside the cell, DNA is never without an entourage of proteins. Stretches of ~150 base pairs are wrapped around octets of histone proteins to form nucleosomes. These and other DNA-associated proteins make up chromatin, a structure that may be the most complex molecular assembly in the cell[1]. Once considered a straightforward packaging system for unused DNA, chromatin is becoming recognized as a dynamic genome organizer, a scaffold that directs DNA activity.

Shortly after the turn of the century, researchers began cataloging chromatin proteins and their modifications. Now, they are applying computational analysis to these genome-wide studies in an effort to segregate chromatin's complexity into discrete numbers of chromatin states. These exercises have already revealed regulatory elements across the genome. In time, chromatin-state mapping promises to reveal many secrets of genome function, how cells inherit acquired states, how chromatin directs functions such as transcription and RNA processing, and, crucially, how chromatin biology contributes to disease.

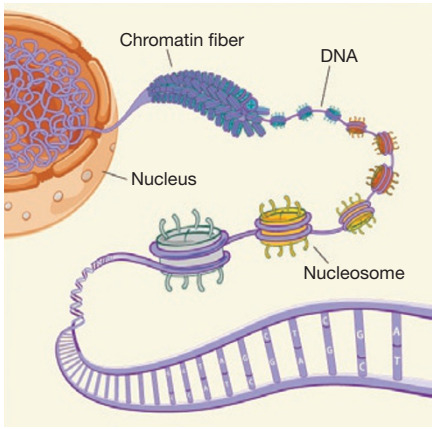
From genome-wide lists to states

Scientists have long known that nuclear DNA exists in different conformations. As far back as the days before television, microscopes revealed different types of DNA in the nucleus: the densely staining heterochromatin, also called 'closed' chromatin, in which genes are packed securely away from transcription machinery, and the lighter-staining euchromatin, or 'open' chromatin, in which accessible DNA allows for active transcription. A closer look at chromatin has revealed considerably more complexity.

Even without considering the DNA sequence, a single nucleosome can exist in trillions of trillions of potential variations. The four basic types of histone proteins can be exchanged with variants and chemically modified in a bewildering variety of ways. Amino acids on histones' tail-like extensions can be singly methylated, dimethylated, trimethylated, acetylated, phosphorylated, ubiquitinated or otherwise modified. More than a hundred chromatin modifications, or 'marks', have already been identified. Together, these epigenetic modifications create patterns that correlate with various functional elements in the genome.

Table 1 | The many states of chromatin (so far)

Cell source	Marks and proteins surveyed	States or groups identified	Analysis	Reference
Arabidopsis thaliana seedlings	11 histone marks plus DNA methylation	4 major states	Heat map and hierarchical clustering	7
Caenorhabditis elegans embryos and larvae	28 histone modifications, variants and chromosome proteins	5 groups	Hierarchical clustering	13
C. elegans at all stages of development	33 genome-wide maps, mostly histone marks	3 combinations	Integrative analysis, using chromatin marks to predict function	14
C. elegans cells, isolated tissues and whole organisms at several developmental stages	700 datasets that profile transcripts, histone modifications and nucleosomes	9 or 30 states, depending on model, used to find cell- and tissue-specific regulators and predict gene expression	Integrative analysis	15
D. melanogaster at several developmental stages	22 histone modifications and chromosomal proteins	15 clusters grouped into 5 states	Cluster and principal component analysis	16
D. melanogaster cell lines and organisms at several developmental stages	18 histone marks	9 or 30 states, depending on the algorithm	Machine learning; combinatorial model considering probability of presence of certain marks	6
D. melanogaster cell lines	53 proteins	5 states	Integrative analysis of genome-wide binding maps	8
Homo sapiens lymphocytes	38 histone marks	51 states, grouped into 5 classes	Multivariate hidden Markov model	9
Nine *H. sapiens* cell types	9 histone marks	15 states used to identify regulatory elements in the genome	Multivariate hidden Markov model	10



Broad Communications

DNA is wrapped around protein complexes called nucleosomes. Proteins comprising nucleosomes contain hundreds of different modifications, which together serve to regulate gene expression.

Soon after the discovery of histone-modifying enzymes in the mid-1990s, researchers began probing where on the genome modifications occurred. In these attempts they mainly looked for regions containing a particular mark or perhaps a combination of two or three marks. In one study, for example, researchers identified over 50,000 potential human enhancers by mapping where different trios of marks occurred[2]. More recently, researchers have begun to take another approach: identifying dozens of marks across the genome, computationally finding recurring combinations and grouping these combinations into states.

Algorithms for defining chromatin states from genome-wide datasets are the "most interesting advance" in understanding chromatin modification in the past three years, says Keji Zhao at the US National Institutes of Health. Researchers in his laboratory were one of the first groups to map the genome-wide methylation and acetylation patterns in human histone proteins. Such maps have a variety of applications, he says. They could help match genes with their regulatory elements or assess the differentiation potential of cells.

Understanding chromatin states will be harder than finding them, says Gary Karpen at Lawrence Berkeley National Labs. "You can identify combinatorial patterns that are common and then start to think about what they mean biologically. That's the hard part."

"Chromatin-state maps can help us out a lot," says Jason Lieb at the University of North Carolina at Chapel Hill. He and others have developed techniques to assess whether chromatin is in open or closed conformations across the genome and used these to identify cell-specific regulatory elements. Last year, his team identified a genetic polymorphism associated with

type-2 diabetes in open chromatin of insulin-producing islet cells[3]. (Commercial kits for evaluating chromatin states at discrete locations are available. EpiQ from Bio-Rad uses quantitative PCR to assess how accessible chromatin is across researcher-selected locations. "A researcher can analyze the chromatin structure of over 100 different genomic loci from a typical EpiQ sample," says Steven Okino, staff scientist at Bio-Rad. EpiTect ChIP from SA Biosciences uses a similar approach to look at 84 preselected genomic locations.)

C. David Allis and colleagues at the Rockefeller University are studying a histone protein called histone H3.3, which sometimes takes the place of the standard histone H3 in nucleosomes. His team found, surprisingly, that histone H3.3 seems to have its own set of chaperone proteins that place it, and it alone, into the genome[4,5]. The chaperone proteins for histone H3.3 had not previously been identified as chromatin remodelers, but researchers in other laboratories had implicated these proteins in mental retardation, cell death and pancreatic cancer.

The distribution of histone H3.3, which is often associated with active genes, was also surprising. It occurred in euchromatin as well as in quintessential regions of heterochromatin such as telomeres—the regions at the end of chromosomes. Allis turned to the literature and was intrigued by work that analyzed histone modifications and DNA-binding proteins across the *Drosophila melanogaster* genome, mapping recurring combinations into nine chromatin states[6]. Histone H3.3 was not a focus of this analysis, says Allis, but the pattern was obvious. The genomic locations of one of the nine states and its associated histone modifications matched what Allis had observed with histone H3.3. "All the clusters they defined as chromatin state 3 tracked very nicely," says Allis. "It was a near-perfect match." Work in his and other labs is now well underway to understand pancreatic cancer and other diseases in terms of chromatin biology, says Allis. "An awful lot of work came together remarkably fast, and the genome-wide maps contributed."


Maria Nemchuk

Mapping more histone marks may identify additional chromatin states and potentially new regulatory principles, says Bradley Bernstein at Massachusetts General Hospital. "Part of the excitement of the states field is looking across more and more marks and asking, 'where are there things we are missing?'."

Sorting out chromatin states

Mapping chromatin states is very much a work in progress. Various statistical techniques have been applied so far to datasets cataloguing different marks and proteins from several species, detecting fewer than five states or more than 50 (**Table 1**). That does not mean that one study is right and another is wrong. "There's not a magic number of states," says Lieb. "The whole point of these is just to distill down the data into something that's interpretable."

The data appear to be distillable: observed combinations are only a tiny fraction of the total possibilities. A study published in April 2011 examined the occurrence of DNA methylation and 11 histone marks across the *Arabidopsis thaliana* genome[7]. Though over 4,000 combinations are theoretically possible, only 38 occurred frequently. These could be collapsed even further, says François Roudier of the Institut de Biologie de l'Ecole Normale Supérieure in Paris, who co-led the study. "Clustering analysis indicates that the 38 combinations correspond actually to four main chromatin states with distinct functional properties."

So far, the repertoire of observed chromatin signatures seems to be limited, not only in plants but also in fruit flies, roundworms and human cells. "It's encouraging that you have a limited number of chromatin types," says Bas van Steensel, a chromatin biologist at the Netherlands Cancer Institute. "It makes life a little more manageable." Recurrent combinations reflect a redundancy that makes biological sense, says Bradley Bernstein at the Massachusetts General Hospital. "The cell uses it to ensure robust regulation, but the computational biologist can use it to obtain robust annotations of the genome."

Though algorithms provide a systemic, unbiased approach to find states, scientists themselves decide, roughly, how many states algorithms identify. "If you want to find dozens or hundreds of states, you can, but the datasets as they are now are not comprehensive enough to do such


Tamara Lackey

Chromatin states distill overwhelming amounts of information, says Jason Lieb at the University of North Carolina, Chapel Hill. "It's bound to be an oversimplification, but it's also bound to be a useful way to understand genome organization."

fine-grained classification," says van Steensel. "The point at this stage is to obtain the big picture of chromatin." van Steensel and colleagues analyzed 53 proteins to segregate the *D. melanogaster* genome into five states, which they designated using colors: put simply, yellow designates active housekeeping genes, red designates active tissue-specific genes, blue designates genes covered in the gene-repressing polycomb proteins, green designates proteins also found around centromeres, and black designates nearly two-thirds of silent genes[8]. Though flies and roundworms lack DNA methylation, many mapping projects use these species. In addition to the advantages of genetics, the genomes of these species are a fraction the size of the human genome, providing a better ratio of signal-to-background noise.

Researchers led jointly by Gary Karpen at Lawrence Berkeley National Laboratory and Peter Park at Harvard Medical School looked at 18 histone marks in *D. melanogaster* and applied different algorithms to the same data to segregate the genome into 9 or 30 states[6]. The more states there are, the more complicated follow-up experiments become, but too few states can be even more confusing, says Karpen. "There is a point at which you lose meaning because you go too low. You're lumping things together that don't belong together."

Manolis Kellis and his postdoc Jason Ernst at the Massachusetts Institute of Technology used 38 histone marks to find 51 states in human lymphocytes, which they grouped into five classes: active intergenic states, large-scale repressed states, promoter-associated states, repetitive states and transcription-associated states[9]. A subsequent study with Bernstein across nine cell types identified cell-specific regions, linking distal regulatory elements to putative target genes and hinting at the functional relevance of disease-associated genetic polymorphisms[10].

One of the most important next steps is to figure out which combinations of marks encode biologically distinct states of chromatin, says Kellis. In general, the more marks are examined, the more subtle are the distinctions that can be discerned. Some marks carry more information than others, however, and the functional meaning of some marks changes depending on the presence of other marks, Kellis explains. "In English, just seeing the letter 'e' in the middle of the word does not tell you anything about its pronunciation if you don't look at the context of what other letters are there." Similarly, some marks may indicate a repressed state when combined with one mark but an active state when combined with another.

To get a better handle on chromatin states, researchers still have to answer several questions, Kellis says: "How many marks do we need to experimentally map in each new cellular condition? Which marks should be prioritized to capture different subsets of states and which are redundant? And given a set of genome-wide maps of individual chromatin marks, how many biologically meaningful chromatin states can be distinguished reliably?"

And the discovery of new histone marks and chromatin proteins will likely reveal new states. "To be sure that you've covered all the states, you have to include all the proteins that are representative of the states, and since we don't know what the states are, it's always possible we are missing something," says van Steensel.

Consistent measuring

Validating techniques for mapping marks and chromatin proteins is a challenge. van Steensel's technique marks DNA that comes into contact with a protein of interest by fusing it with a *D. melanogaster* protein that methylates the nucleotide adenosine. This can reveal which sequences encounter, even transiently, a variety of chromatin-associated proteins, including those that modify histones or line the nuclear envelope.

Manolis Kellis of MIT says chromatin states reveal sophisticated organization: "You're only finding a small number of chromatin states compared to the huge plethora of possibilities."

Most chromatin mapping studies, however, rely on chromatin immunoprecipitation (ChIP), which uses antibodies to particular histone modifications or DNA-binding proteins to purify associated DNA, which can then be analyzed by sequencing or microarrays. But these antibodies do not always perform as expected. Although genome-wide ChIP studies have been around for several years, biologists must be careful about these datasets, says Karpen. "It's not clear what you can and cannot trust from the literature." It is clear, he adds, that researchers need data from several types of control studies to trust the antibodies.

Because these reagents are so crucial, researchers in several laboratories recently banded together to characterize the performance of 246 commercially available antibodies directed to 3 unmodified histones and 57 distinct histone modifications. They evaluated each antibody in three ways: ChIP studies to make sure the antibodies would pull down the desired mark; dot blots (using synthetic peptides with an array of histone modifications) to assess whether antibodies ever 'mistook' one mark for another, and western blots to assess whether antibodies cross-reacted with other cellular components[11]. Success in one assay does not guarantee success in others, says Lieb. "The fact that it works on a western [blot] doesn't mean that you should stop testing it."

About a quarter of the antibodies failed tests for specificity. For example, antibodies to a triply methylated lysine on a histone might also bind to singly or doubly methylated versions. In three cases, antibodies were completely specific, but for different modifications than the ones that they had been sold to detect, says Lieb. Antibodies also sometimes pulled down unmodified histones or proteins besides histones. As protein content varies by condition, cell type and species, antibodies that demonstrate exquisite specificity under one set of conditions may perform less well under another. And polyclonal antibodies, which comprise the majority of commercial preparations, can vary considerably between batches. "You have to test every lot," says Karpen. (One of the corresponding authors, Peter Park, has created a website (http://compbio.med.harvard.edu/antibodies/) where researchers can post test results for antibodies by lot number, which hopefully will save researchers the hassle of replicating control studies.)

Commercial manufacturers are responding to demands for reliable tools for genome-wide ChIP, says John Rosenfeld, who manages epigenetic product development at EMD Millipore. His company, for example, is now adopting the same series of tests that the consortium applied. They are also trying to make sure that researchers always have more than one antibody to use for a particular histone modification. More fundamentally, he says, a growing understanding of the context of histone marks in the cell is changing the antigens used to stimulate antibody production. This typically starts with synthetic peptides containing only a single histone modification, but in actual chromatin, histone modifications generally occur together with others, so manufacturers such as Millipore are adopting techniques to ensure that antibodies can bind the histone modification of interest in the presence of other modifications.

Just as crucial as antibodies for understanding chromatin states are the type and quantity of the cells. "If you mix cell types, you're in danger because there could be one region that's marked with one histone modification in one cell type but not another," explains Lieb. van Steensel is working around this difficulty by genetically engineering flies so that the DNA-marking protein is active only in certain conditions or certain tissues. "Once this is operational, we won't even have to sort the cells," he says. Only DNA from the labeled cells will be amplified.

Much research is done in cultured cells, which can be produced in large quantities. Scientists in Bernstein's lab and others are developing techniques to conduct ChIP studies with a tenth or less of the normally required numbers of cells. The work, he says, is "nothing glamorous." It consists of titrating antibodies, optimizing how DNA is fragmented and amplified, and eliminating unnecessary steps[12]. The tedium is paying off. Bernstein and colleagues now can perform ChIP studies on cells derived from tissue samples and clinical biopsies. Preliminary results, he says, show tantalizing differences between cultured cells and *ex vivo* samples.

Eventually, says Bas van Steensel of the Netherlands Cancer Institute, chromatin states will show how genome regulation works. "What you want to know is whether the states correlate with function, and then it becomes interesting."

Drilling down to function

The hardest work will not be in identifying the chromatin states but in figuring out how they are maintained and how they regulate the genome. "The classical way is you take a piece of functional DNA, and you ask if it helps with the expression of transgenes. You

can do that for 20 or 50 genes, but it's a lot of work. And to really test for functionality, you probably need to do 1,000 genes and the controls," says Karpen. "We don't have the tools to do this in a high-throughput way."

Instead, chromatin maps are opening up the field for researchers who tend to focus in on specific genes, proteins and patterns, says Allis at Rockefeller University, one of the scientists who introduced the idea that combinations of histone marks could have distinct meanings. "It's staggering what the researchers have learned from the genome-wide methods," says Allis, "but at the end of the day it's going to be important to take that information and ask what a particular chromatin state means mechanistically."

1. van Steensel, B. *EMBO J.* **30**, 1885–1895 (2011).
2. Heintzman, H.D. *et al. Nature* **459**, 108–112 (2009).
3. Gaulton, K.J. *et al. Nat. Genet.* **42**, 255–259 (2010).
4. Goldberg, A.D. *et al. Cell* **140**, 678–691 (2010).
5. Lewis, P.W., Elsaesser, S.J., Noh, K.M., Stadler, S.C. & Allis, C.D. *Proc. Natl. Acad. Sci. USA* **107**, 14075–14080 (2010).
6. Kharchenko, P.V. *et al. Nature* **471**, 480–486 (2010).
7. Roudier, F. *et al. EMBO J.* **30**, 1928–1938 (2011).
8. Filion, G.J. *et al. Cell* **143**, 212–224 (2010).
9. Ernst, J. & Kellis, M. *Nat. Biotechnol.* **28**, 817–825 (2010).
10. Ernst, J. *et al. Nature* **473**, 43–49 (2011).
11. Egelhofer, T.A. *et al. Nat. Struct. Mol. Biol.* **18**, 91–93 (2011).
12. Adli, M., Zhu, J., & Bernstein, B.E. *Nat. Methods* **7**, 615–618 (2010).
13. Liu, T. *et al. Genome Res.* **21**, 227–236 (2011).
14. Gerstein, M.B. *et al. Science* **330**, 1775–1787 (2010).
15. modENCODE Consortium *et al. Science* **330**, 1787–1797 (2010).
16. Riddle, N.C. *et al. Genome Res.***21**, 147–163 (2011).

Monya Baker is technology editor for *Nature* and *Nature Methods* (m.baker@us.nature.com).