Surveying the Chromatin Landscape with Next-Generation Sequencing

Researchers develop novel sequencing methods with the MiSeq® and HiSeq® Systems to understand the epigenome and its impact on cancer and immune disease.

Introduction

Every cell in the human body has long strands of deoxyribonucleic acid (DNA) compactly folded inside its nucleus. That folding is made possible by chromatin, the complex of macromolecules that package each cell's DNA into that small, condensed volume—an architecture necessary to protect its structure and sequence. Understanding chromatin and this dynamic architecture are crucial to understanding how the genome works. Its tightly packed grooves and folds provide a unique physical landscape for gene transcription—one that has profound implications for our understanding of gene regulation, replication, and expression. Scientists are now finding new ways to delve into chromatin's many biochemical mysteries.

William Greenleaf, PhD, an assistant professor in Stanford University's renowned genetics department, is focused on understanding how the 2 meters of DNA in each cell nucleus are folded and stored. "About 95% of the genome is folded and sequestered away in the chromatin," Dr. Greenleaf said. "Only a small percentage is accessible to the transcription machinery. Deciphering how that all works is intriguing and important."

iCommunity spoke with Dr. Greenleaf about his team's development of 2 new next-generation sequencing (NGS) methods to better survey the enigmatic chromatin landscape: assay for transposase-accessible chromatin sequencing (ATAC-seq)[1] and single-cell ATAC-Seq (scATAC-seq).[2] He believes that these approaches might one day provide new insights into the development and treatment of cancer and autoimmune disease.

Q: What sparked your interest in applied physics?
William Greenleaf (WG): I was always interested in molecular biology—particularly DNA and the molecular machinery of the genome. But as an undergrad, I wanted to avoid chemistry, so I studied physics instead. I ended up getting my PhD in applied physics with a focus on single-molecule biophysics, because I was interested in understanding the mechanics by which individual molecules carry out tasks within the cell. During my postdoc, I was bitten by the high-throughput sequencing bug. We were thinking a lot about new ways to approach these different complex biological questions. A sequencer can make hundreds of millions or even billions of measurements across the genome and that's what is needed to understand the complexity of this biology.

Q: What does high-throughput sequencing provide over the other methods you used previously?
WG: As a grad student, I performed experiments on individual molecules. It's labor-intensive work—and you have to deal with a lot of handcrafted data. After a few years, I wanted to find a different way. I wanted to do the exact opposite—take an enormous number of measurements very quickly. So we've been working to repurpose the infrastructure associated with high-throughput sequencers to do massive scale biochemistry on nucleic acids.

Q: What inspired you to develop new tools to study chromatin?
WG: We have a great understanding of the structure of DNA—and a good understanding of a single nucleosome. However, that's where our high-resolution understanding of the nucleus ends. The question of how DNA is organized at the kilobase length scale remains a fundamental question to be answered. We don't know all that much about how the nucleosomes that bind to DNA tightly are shifting around, how the transcription factor binding sites might be competing for DNA, and how different transcription factors may cooperate to build enhancers. These things touch and interact mechanically to make things happen. We need to understand the logic of the physical regulatory landscape—the regulome, if you will—to see what makes a gene fire or not.

One of the significant questions is how a cell can mark and use these different elements to change their biological state. We know that all the different cells in a body have the same genome effectively, yet they do incredibly different things. I like to think of chromatin as a physical landscape that tells the cell which parts of the DNA to use and which parts to ignore. In a sense, it's a major organizational principle of biology.

Q: Has the data from the Encyclopedia of DNA Elements (ENCODE) Consortium and Epigenetics Roadmap provided a glimpse into the regulome?
WG: Recent work from the ENCODE consortium and the Epigenomics Roadmap have tried to illustrate how different elements in DNA are functional, and how they can be marked and used. That initial

Dr. William Greenleaf is an assistant professor in the Stanford University Genetics Department.

glimpse into the full richness of the epigenomic landscape has been fascinating. It's like discovering a new continent. But much of that data are observational—there's much less of a functional understanding of how these elements conspire to allow the development of a single egg into a full human being. That's what we're trying to develop methods to understand.

> "We need to understand the logic of the physical regulatory landscape—the regulome, if you will—to see what makes a gene fire or not."

Q: What methods were used previously to assess the epigenetic information encoded in the chromatin structure?
WG: Certainly, there has been some deep, important foundational work at the level of individual loci. For more than 35 years, researchers have used DNase hypersensitivity and run resulting patterns on polyacrylamide gels. That work has formed the foundation for our understanding of how DNA accessibility changes within the chromatin and how these individual loci function. There have been other biochemical assays of transcription factors that provided information on what's necessary and sufficient in very defined systems for them to bind.

Q: What are the limitations of those older methods and how can the tools you're developing answer those challenges?
WG: Most of the limitations of previous methods have to do with throughput and sensitivity. We're trying to understand comprehensively how these different markings and genetic elements are used across the whole genome. We want to follow these things over time in unperturbed native states. For example, high-throughput sequencing is effectively a single molecule measurement. It's a digital measurement that can have ultimate sensitivity if used in the appropriate way. Current sequencers allow you to read out billions of sequences at a time, and you can make a billion measurements at a time for a reasonable cost.

The old methods are brilliant methods, really. What's wonderful is that they, too, can be moved into this high-throughput realm with incredible sensitivity and broad application across the genome in a straightforward manner.

Q: What sparked the development of your new method, ATAC-seq?
WG: Early on, I was interested in pushing sensitivity for epigenomic methods to the single cell level so we could ask questions about populations versus individual cells and how individual cells within even a biologically similar population might vary. We've been interested in transposase because it has an ability to insert sequencing adapters in a single step. It captured our imagination and Jason Buenrosto, a member of my lab, thought it was an interesting reagent and wanted to see if we could use it to read out open chromatin. He got it going very quickly—and I credit Jason and my colleague, Professor Howard Chang, and members of his lab for working to make it happen.

Q: How can the transposase, Tn5, be used to identify regions in accessible chromatin?
WG: Tn5 is a hyperactive transposase that's meant to more or less insert sequencing adapters, or adapters that can be used to make a sequencing library, everywhere in naked DNA. It's an efficient and simple way to generate a sequencing library for whole-genome sequencing. But we thought if only a few percent of the genome is "open" in a nucleus, and therefore accessible to the Tn5 transposase, then likely these regions would be the only ones that would be accessible. So it acts, in some sense, like a test transcription factor and it generates reads from regions that are accessible. The length of the fragments it generates also tells you about some of the finer structure of the chromatin. For example, we see that there's an enrichment for reads that are on the order of 200 base pairs long—reads that are being protected by a single nucleosome. You can determine whether the reads that you're getting are from nucleosomal or nucleosome-free regions based on the predicted length.[3] So not only do the reads tell you something, but also the length of the fragments provides information about the local structure.

> "Technical reproducibility between replicates using ATAC-seq was very high—with an R value of ~0.93."

Q: What can we identify when sequencing the DNA fragments at locations of open chromatin using ATAC-seq?
WG: There are 3 different things that you can get out of the ATAC-seq primary data analysis. The first is information about the regions of open chromatin, those regions that are accessible to the machinery of transcription. The second is information about whether those regions of increased accessibility have nucleosomes that are bound to them. Finally, in an analogy to DNase hypersensitivity methods, you can see when a transcription factor is tightly bound to a specific locus because we see a depletion of reads where that transcription factor protects the DNA to which it is bound. Using the footprinting observed from the binding of specific transcription factors, you can infer the presence of a factor that binds to that specific DNA motif. It's valuable because it's giving you a sense of what the molecular factors are that are potentially driving gene regulation. We and others have also been working on algorithms that use information about the distribution of fragments to call a chromatin state, or determine whether regions are poised or in active transcription.

Q: How much of a sample was required to perform this method?
WG: Most of our replicates in the initial ATAC-seq study were generated from about 50,000 cells. In comparison, most published protocols for DNase require 10 million cells or more. We've also now published data sets that have been generated from about 250 cells in our single-cell protocols. How many cells you need depends on your goal and the type of data analysis you want to perform. We've created some relatively deep, complex libraries from only 50,000 cells.

Q: What was the quality of the results for ATAC-seq?
WG: Technical reproducibility between replicates using ATAC-seq was very high—with an R value of ~0.93. We also had a correlation with 2 other DNase hypersensitivity data sets. Correlation with those data sets was similar to the correlation of the 2 data sets with each other. That said, it's going to take a lot more data from many different cells to compare to the DNase hypersensitivity bulk data to fully characterize how many peaks ATAC-seq observes and DNase doesn't, and what peaks DNase observes and ATAC-seq doesn't, and why that might be the case.

> "ATAC-seq allows you to ask questions about the epigenetic variability in complex or rare tissues and epigenomic landscape in populations of cells that haven't been observable at the genome-wide level before."

Q: What did the ATAC-seq paired-end reads provide regarding nucleosome packaging and positioning information?
WG: There's no reason not to do paired-end sequencing when generating ATAC-seq libraries. Both ends tell you the insertion point of a transposase—so it really doubles your data. It also tells you the length of each of the fragments generated from your library. When you align each of the ends to the genome, you can determine how long that fragment is. That length can be used to identify, for example, nucleosome positions. We recently reported a method for using the length distribution of the ATAC-seq fragments to call nucleosome positions at high resolution.

The fragment length distribution provides another dimension of information. You have both positions where the insertions occur and the lengths of the fragments generated. That length distribution can be valuable in understanding the local chromatin structure.

Q: What does ATAC-seq offer that other methods can't?
WG: ATAC-seq gives you a bit of a hybrid between DNase hypersensitivity and MNase concentrated in regions of accessibility that are likely regulatory regions. It can be applied to relatively low numbers of cells—and the workflow is simple. ATAC-seq allows you to ask questions about the epigenetic variability in complex or rare tissues and epigenomic landscape in populations of cells that haven't been observable at the genome-wide level before.

Q: How does the multidimensional portrait of gene regulation that ATAC-seq provides help us to understand disease?
WG: One of the interesting applications of this methodology is to look at open chromatin landscapes over time or in a developmental hierarchy. This gives you an understanding of what DNA elements are changing or becoming active during an immune response or during the differentiation of blood cells. It gives you a map of the landscape, a picture of DNA elements that are controlling the biological state of the cell. When you pair that with RNA-Seq information, you can make inferences about regulatory elements that are potentially controlling RNA expression. This allows us to ask questions about what factors are driving chromatin changes that result in disease phenotypes.

We can use ATAC-seq on a few cells, which you can get from a standard blood draw, and assay their epigenomic signatures.[4] If there's a patient with an autoimmune problem, we can look at what molecular factors are driving gene expression of a particular cytokine responsible for the faulty immune response. We can then rationally nominate drugs that might antagonize those molecular factors. One day, it might help us nominate a first course of treatment in a clinical context.

Q: Why did you choose the MiSeq and HiSeq Systems to develop this method?
WG: The MiSeq System is great for method development because of its relatively low cost per run and rapid turnaround time. You can sequence something in a day—get an answer and then optimize your protocols. The HiSeq System provided the least expensive cost for the bases sequenced and, like the MiSeq System, offered the ability to perform paired-end reads. Illumina sequencing systems generally can't be matched in terms of raw throughput or cost.

> "The MiSeq System is great for method development because of its relatively low cost per run and rapid turnaround time."

Q: Why did you choose the Nextera® DNA Library Prep Kit for this method?
WG: The Tn5 aspect of the Nextera DNA Library Prep Kit was the most important thing. This Tn5 enzyme provided us with the assay and the library prep step all in one. The assay is identifying regions of the genome that are accessible and then the transposase, with its cut and paste mechanism, gives you a sequenceable library after some PCR steps. The Nextera kit is also commercially available, we felt like it would be a method that others could perform on their own.

Q: What led you to the development of scATAC-seq?
WG: There have been powerful methods for assessing cell-to-cell variation in both the genomic sequences and the transcriptome. This has led to a better fundamental understanding of standing variation in subsets of mammalian cells. Much single-cell imaging work suggests that that transcription factors and regulatory proteins that bind to DNA and initiate transcription are dynamically expressed at varying levels or dynamically localized into the nucleus. The question is, how do these observed heterogeneities couple into the chromatin itself and how does this affect gene expression? In addition, if you had the ability to look at cell chromatin at the individual cell level, you might also be able to take a complex tissue and build a regular hierarchy understanding of the underlying elements performing gene regulation by trying to understand what cells look like other cells—and then building up your understanding of the population from there.

Q: What kind of cell-to-cell variation do you see using scATAC-seq?
WG: With scATAC-seq, we saw variation in accessibility associated with the binding sites of specific transcription factors—and the transcription factors that were associated with this noise varied from cell type to cell type. We also observed that regions of the genome tended to be correlated in their accessibility noise if they were in the same compartments of chromatin organization. Regions that looked to be close together in three-dimensional space tended to be open together as well.

> "The HiSeq System provided the least expensive cost for the bases sequenced and the ability to perform paired-end reads."

Q: Would scATAC-seq information be useful in a clinical setting?
WG: This is one of the things we're excited about. If we could scale up our single cell methods to more than 100 cells at a time, it would be exciting to try to link transcription start sites to their promoters based on correlated accessibility. There's an ongoing question of what the linkage between enhancers in the genome and the specific genes they regulate might be. Because each individual cell we look at is an independent measurement, we might expect that transcription start sites that are open together in a correlated manner might be functionally linked. That is, the genetic element might be regulating that gene of interest. That wiring is not only interesting from a basic biological perspective, it also offers the possibility of clinical applications. Even if there are relatively rare circulating tumor cells, for example, we might be able to capture those and assess the regulatory landscape. That might allow us to nominate specific molecular drivers that drive that dysfunction and provide hypotheses for drug targeting downstream. It's an exciting possibility.

Q: How did using Illumina sequencing systems help you develop scATAC-seq?
WG: We've used the MiSeq System to pioneer a lot of these studies. It enables a rapid assessment of whether the libraries we generate are of high quality and are ready to be sequenced deeper. We can multiplex samples onto a MiSeq flow cell and, for a reasonable cost, obtain a rapid turnaround on the library quality. We then use the HiSeq System to sequence deeper when needed.

Q: What's next for your lab?
WR: We're interested in applying these methods to specific biological contexts. We want to understand the differences in the epigenomic landscape in cancers, and also in immune response and autoimmune disorders. There are many directions we can go in here. With these techniques, we can draw a blood sample and interrogate the genome of specific subsets of cells to learn more about the general biology of the genome, and also the pathology involved in disease states. It's an exciting time to be studying biology.

References

1. Buenrostro JD, Giresi PG, Zaba LC, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013; 10(12): 1213-1218.
2. Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. 2015; 523(7561): 486-490.
3. Schep AN, Buenrostro JD, Denny SK, et al. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res*. 2015; August 27, pii: gr. 192294. 115.
4. Qu K, Zaba LC, Giresi PG, et al. Individuality and variation of personal regulomes in primary human T cells. *Cell Syst.* 2015; 1(1):51–61.

Learn more about the Illumina products and systems mentioned in this article:

MiSeq System
www.illumina.com/systems/miseq.html

HiSeq System
www.illumina.com/systems/hiseq_2500_1500.html

Nextera DNA Library Prep Kit
www.illumina.com/products/nextera_dna_library_prep_kit.html

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

illumina®