

Review

ChIP'ping Away at the Next Layer of Genomic Information: ChIP-Seq

New studies use the Illumina® Genome Analyzer for ChIP readout.

by Ivan Garcia-Bassets, Ph.D.

Beyond the essential DNA sequence information decoded by the Human Genome Project, it is now fundamental—in the so-called postgenomic era—to understand how genomic information is used by normal cells, and even more importantly, how it is *misused* by anomalous cells. DNA binding factors (e.g., histones and transcription factors) and their associated cofactors (e.g., coactivators and corepressors) are the dynamic regulators responsible for utilizing genomic information by controlling the transcriptional gene regulation that is behind cellular processes, including cell growth, proliferation, differentiation, and

“...it is now fundamental to understand how genomic information...is *misused* by anomalous cells”

death¹. Despite the key relevance of these regulators, and the extensive biochemical and functional characterization for at least some of them, we are still missing genome-wide mapping of their binding sites. A global binding map would allow us to determine which, when, and how genes might be regulated by these factors at a genomic scale.

ChIP-on-chip is a powerful technology that permits genome-wide localization analysis of DNA binding factors, cofactors,

and histone marks. It combines specific immunoprecipitation of genomic DNA fragments that *in vivo* are directly or indirectly associated with specific proteins or histone marks (ChIP) and DNA microarray analysis (chip)². However, serious technical challenges are associated with whole genome ChIP-on-chip analysis in its different variants. These include potential bias introduced by a global PCR amplification step, low resolution and low sensitivity by some approaches, high input material requirements in most approaches, uninformative results on repetitive sequences, highly sophisticated analysis, and expensive microarrays. These drawbacks are likely the main reasons for its extremely limited use by researchers. Indeed, no more than half a dozen factors and histone marks have been mapped at a whole genome level, whereas there are predicted to be more than 3,000 human DNA binding factors³. When considered in combination with thousands of cofactors, histones and non-histone marks, different cell types, and huge numbers of possible cellular conditions, this field is left virtually unexplored.

Now, however, three independent groups have almost simultaneously reported the combination of conventional ChIP assays with the Illumina Genome Analyzer, using massively parallel Solexa® DNA sequencing technology (aptly named the ChIP-Seq assay). This technological breakthrough



Dr. Ivan Garcia-Bassets is an Assistant Research Scientist in Dr. Michael G. Rosenfeld's Lab at the University of California, San Diego (UCSD), School of Medicine. He received his Ph.D. in Biochemistry from Universitat Autònoma de Barcelona, Spain. Dr. Garcia-Bassets' current scientific interests focus on defining and understanding the molecular mechanisms that regulate gene transcription at the genome-wide level, particularly their integration with cellular signaling and their relevance in cancer. Email: ibassets@ucsd.edu

seems to finally permit high resolution, highly specific, highly sensitive, simple, and relatively inexpensive genome-wide mapping of DNA regulators and histone marks. These innovative studies could represent the technical advance necessary to chip away at this missing essential layer of genomic regulation^{4,6}.

In a May article in the journal *Cell*, Barski and collaborators reported an impressive number of high resolution maps for the whole genome distribution of 20 histone methylation marks (mono-, di-, and tri-methylated H3K4, H3K9, H3K27, and H3K79; mono- and di-methylated H3K36, H4K20, and H3R2; mono-methylated H2BK5; and di-methylated H2AR3/H4R3), the histone variant H2A.Z, RNA polymerase II, and the insulator binding protein CTCF by combining a ChIP assay with the Illumina Genome Analyzer in purified CD4⁺ T cells from human blood⁴. In addition to successfully proving the utility of this technology for whole genome location analysis and providing the most comprehensive mapping of histone marks and H2A.Z ever reported, it provided an important new observation. The results showed unprecedented accumulative patterns of specific histone marks that seem to distinctly differentiate active enhancers from promoters. High occupancy by H2A.Z, mono-methylated H3K9, and mono-, di-, but also surprisingly, tri-methylated H3K4, was associated with both active enhancers and promoters. Additional high occupancy by mono-methylated H3K27, H3K9, H4K20, H2BK5, and tri-methylated H3K36 downstream of the transcription start site (TSS) was only associated with the presence of active promoters. The authors propose that these patterns can be widely used to confirm annotated promoters and identify new TSSs. Overall, this paper shows not only the validity of this new assay, but reports a whole genome-wide map of enhancers and promoters in the human genome that includes the position of multiple undescribed promoters. This information sheds new light onto our understanding of global regulation by promoters and enhancers.

In a second paper published in the June

issue of *Science*, Johnson and collaborators reported a practically identical technical approach to build an *in vivo* location (“interactome”) map of the neuron-restrictive silencer factor, NRSF/REST, in Jurkat human T lymphoblast cells⁵. NRSF is a well-described zinc finger repressor that negatively regulates gene expression of neuronal genes in non-neuronal cells. The authors selected this factor mainly because of the ~80 validated binding sites previously reported for NRSF. ChIP-Seq revealed up to 1946 NRSF binding sites in the genome of T cells, exhibiting 87% sensitivity (successful detection of true positives) and 98% specificity (successful rejection of true negatives). Very interestingly, virtually all near-optimal canonical NRSF sites found in this study ($\geq 90\%$ match to consensus NRSF motif) were occupied *in vivo* by NRSF, whereas a new noncanonical class of NRSF sites and other less-optimal NRSF sites were only partially occupied in the genome. These results confirm the importance of identifying sequence binding sites to determine occupancy, and might be used to further study factors that determine occupancy rates in the genome. In addition, a third new group of NRSF sites was described that only contained half NRSF sites. In terms of transcriptional status, NRSF was strongly associated with repression of neuronal genes, as expected. However, the finding of 110 predicted DNA binding factors potentially regulated by NRSF was unexpected. These factors include important neuroendocrine developmental regulators known to control pancreatic β cell development. Overall, these results reveal an unprecedented program of potentially NRSF-regulated genes, including a detailed description of NRSF binding associated with the presence of canonical and noncanonical NRSF motifs, and predict an important role of NRSF in pancreatic development.

Finally, in a third paper from *Nature Methods*, Robertson and collaborators reported an unprecedented binding map for the signal transducer and activator of transcription protein 1, STAT1, in human HeLa S3 cells using the Illumina Genome

“Illumina’s massively parallel Solexa Sequencing technology will be remembered as a key catalyst of the coming wave of astonishing discoveries.”



Analyzer⁶. STAT1 is a cytoplasmic transcription factor that translocates to the nucleus upon stimulation by an extracellular signal, such as interferon γ (IFN γ), and regulates transcription of genes involved in cell differentiation, survival, proliferation, and apoptosis. Multiple STAT1 binding locations were already known, making STAT1 a good test case for Illumina's DNA sequencing approach on ChIP'ed DNA. Impressively, > 11,000 and > 41,000 putative STAT1 binding sites were revealed in unstimulated and IFN- γ -stimulated cells (> 70% sensitivity, > 95% specificity). For comparison, ChIP-on-chip and other ChIP-sequencing approaches revealed from four-fold⁶ to several thousand-fold⁷ fewer potential STAT1 locations in the genome. This valuable STAT1 genomic mapping also revealed that 15% of locations disappeared upon stimulation. STAT1 mainly peaked at -100bp when close to a TSS, and approximately 50% of locations were intragenic, compared to 25% intergenic. Repetitive sequences contained 16–18% of STAT1 sites and pericentromeric sequences contained 1–3% of sites. Overall, this extensive location map of STAT1 reveals not only new gene programs potentially regulated by STAT1, but suggests new paradigms of STAT1 function in genomic regulation that have been missed by other technologies.

It is rare that three manuscripts almost simultaneously describe the development of the same new technology, as has been done with the ChIP-Seq assay of massively parallel sequencing of ChIP'ed DNA. Even more impressively, these three papers are not limited to validation of the ChIP-Seq technology, but also, aligned with the ultra high-throughput philosophy of this technology, provide massive amounts of new and interesting data. It is clear that new

experimental approaches will be necessary to extract and understand the biological relevance of the insights provided by these studies. However, the development of a technology for genome-wide mapping of regulators and histone marks—described in these three studies as highly resolutive, highly specific, highly sensitive, simple, requiring reasonable amounts of input material, and relatively inexpensive—represents an important milestone that will shed crucial new light in our long road to deciphering the complex regulation of transcriptional networks in normal and anomalous cells. The ENCODE Project Consortium recently initiated a journey down this road with the identification and analysis of functional elements in 1% of the human genome (ENCODE pilot project)⁸. A major surprising conclusion derived from this project is very well stated by Dr. Steven Henikoff "... the ENCODE pilot project, representing a detailed and comprehensive characterization of 1% of the human genome, has demonstrated how little we truly understand about how our genes are regulated. Transcripts are nearly everywhere, regulatory sequences remain poorly defined, and evolutionary conservation is a surprisingly inadequate predictor of transcriptional features⁹." Thus, what is predictable is that Illumina's massively parallel Solexa Sequencing technology will be remembered as a key catalyst of the coming wave of astonishing discoveries.

REFERENCES

- (1) Rosenfeld MG, Lunyak VV, Glass CK (2006) Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes & development* 20: 1405-1428.
- (2) Kim TH, Ren B (2006) Genome-Wide Analysis of Protein-DNA Interactions. Annual review of genomics and human genetics 7: 81-102.
- (3) Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. *Nucleic acids research* 34: D74-81.
- (4) Barski A, Cuddapah S, Cui K, Roh TY, Schones DE et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823-837.
- (5) Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* (New York, NY 316: 1497-1502.
- (6) Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*.
- (7) Bhinge AA, Kim J, Euskirchen GM, Snyder M, Iyer VR (2007) Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome research* 17: 910-916.
- (8) (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799-816.
- (9) Henikoff S (2007) ENCODE and our very busy genome. *Nature genetics* 39: 817-818.

ADDITIONAL INFORMATION

To learn more about the next generation of sequencing products by Illumina with Solexa Technology, visit our website at www.illumina.com under Products & Services.

We are committed to providing you with the content you want as a member of the Illumina community. Please email us with comments and suggestions for topics at icommunity@illumina.com.

FOR RESEARCH USE ONLY

© 2007 Illumina, Inc. All rights reserved.

Illumina, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, and CSPro are registered trademarks or trademarks of Illumina. All other brands and names contained herein are the property of their respective owners.
Pub. No. 370-2007-017 26Jul07

