Epigenetics

ISSN: 1559-2294 (Print) 1559-2308 (Online) Journal homepage: https://www.tandfonline.com/loi/kepi20

Epigenetics meets next-generation sequencing

Peter J. Park

To cite this article: Peter J. Park (2008) Epigenetics meets next-generation sequencing, Epigenetics, 3:6, 318-321, DOI: 10.4161/epi.3.6.7249

To link to this article: https://doi.org/10.4161/epi.3.6.7249

[Epigenetics 3:6, 318-321; November/December 2008]; ©2008 Landes Bioscience

Point of View

Epigenetics meets next-generation sequencing

Peter J. Park

Harvard Medical School Center for Biomedical Informatics; Harvard-Partners Center for Genetics and Genomics; HST Informatics Program at Children's Hospital Boston; Boston, Massachusetts USA

Abbreviations: ChIP, chromatin immunoprecipitation; ChIP-seq, ChIP-sequencing

Key words: chromatin, histone modification, nucleosome, ChIP-seq, ChIP-chip

Next-generation sequencing is poised to unleash dramatic changes in every area of molecular biology. In the past few years, chromatin immunoprecipitation (ChIP) on tiled microarrays (ChIP-chip) has been an important tool for genome-wide mapping of DNA-binding proteins or histone modifications. Now, ChIP followed by direct sequencing of DNA fragments (ChIP-seq) offers superior data with less noise and higher resolution and is likely to replace ChIP-chip in the near future. We will describe advantages of this new technology and outline some of the issues in dealing with the data. ChIP-seq generates considerably larger quantities of data and the most challenging aspect for investigators will be computational and statistical analysis necessary to uncover biological insights hidden in the data.

Introduction

One of the latest and most exciting scientific developments is the next-generation sequencing technology. "Next-generation sequencing" refers to a set of new DNA sequencing techniques that deliver dramatic improvements in sequencing capabilities by employing massively parallel reactions on millions of DNA fragments.[1,2] It allows sequencing of relatively short DNA fragments at a cost that is at least two orders of magnitude less than the traditional Sanger method. This progress has been enabled by convergence of several techniques including new chemistries, amplification methods, miniaturization and high-resolution microscopy.

For several years, microarray-based technologies have offered high-throughput whole-genome approaches that have brought important advances. The earliest and most prominent application of this technology was in profiling of gene expression, but it has had widespread uses with major impact in estimation of DNA copy number, profiling of microRNAs, identifications of genotypes in single nucleotide polymorphisms and mapping of the binding sites for DNA-binding proteins. Next-generation sequencing, however, offers important advantages and has the potential to replace many of the microarray-based platforms in the near future. The lack of availability in sequencing machines and the high cost are still prohibitive for many investigators at this time, but with more broad adoption of the technology and increased competition among manufacturers, the cost is expected to decrease substantially in the coming years. With a number of new platforms on the horizon with a promise of even higher throughput and lower cost, the field is evolving rapidly and its impact beyond the next couple of years is difficult to predict. One thing that is certain is that this technology has begun to have a transformative effect in many areas of molecular biology and genetics. Epigenetics is no exception. In this work, we briefly describe and then compare microarray-based and sequencing-based platforms. Then we highlight the main applications of the technology and the key challenges in the field.

Correspondence to: Peter J. Park; Harvard Medical School Center for Biomedical Informatics; 10 Shattuck St; Boston, Massachusetts 02115 USA; Email: peter_park@harvard.edu

Submitted: 08/05/08; Accepted: 10/22/08

Previously published online as an *Epigenetics* E-publication:
http://www.landesbioscience.com/journals/epigenetics/article/7249

ChIP-chip

A prominent genomics approach for epigenetics in the past few years has been chromatin immunoprecipitation followed by microarray hybridization (ChIP-chip). This technique has been used primarily to detect the locations where a protein of interest is bound to DNA in vivo. More recently, ChIP-chip has been used to profile the sites of DNA methylation or various covalent modifications to the histone tails. In a typical ChIP-chip experiment, the protein of interest is cross-linked with the DNA, generally with a gentle formaldehyde treatment; the DNA is sheared by sonication or micrococcal nuclease to small fragments, usually in the 200–800 base pair (bp) range; an antibody specific to the protein is used to enrich for the DNA fragments associated with the protein; the cross-links are reversed to release the DNA; the fragments are amplified, labeled, and hybridized on microarrays; and finally the signals are read using a scanner to generate the data. Statistical analysis of the data results in a list of binding sites with assessment of statistical significance.

The quality of ChIP-chip or ChIP-seq data depends on several factors. Experimentally, perhaps the most important factor is the specificity of the antibody. Antibodies for many of the histone modifications are currently unavailable and, even when they are available, might exhibit poor affinity or cross-reactivity with other forms of modifications. Obtaining the desired antibody is time-consuming and costly, and remains a major hurdle in epigenetics research regardless of the development in profiling technology. In terms of platforms, it is important to have sufficient probe resolution to map the locations of binding or modification with reasonable precision.

For microarray platforms, this is constrained by the total number of probes that can be fit onto an array. The first publications describing ChIP-chip used an array with several thousands PCR-amplified fragments in *Saccharomyces cerevisiae*.[3] Since then, there has been an enormous progress in the array manufacturing technology, especially with the use of oligonucleotide probes.[4] Nowadays, whole-genome tiling can be done inexpensively for organisms with small genomes. For *Drosophila melanogaster*, for instance, a single 'tiling' array manufactured by Affymetrix covers the entire genome with 25 mer probes at 38 bp resolution. For higher organisms, it is feasible but still remains expensive: for humans, a 7-array set is required to cover the genome at 35 bp.

Another important advance in ChIP-chip has been the capability to customize microarrays with minimal additional cost. This enabled the investigator to determine specific regions of interest to be interrogated with an array. In addition to whole-genome arrays, we, for instance, have used custom arrays to study the Drosophila X-chromosome[5,6] (tiled along with an autosome as a control) as well as the human HOX regions.[7] For the former, 100 bp resolution was chosen so that all the probes can be contained in a single array; for the latter, 5 bp resolution was specified in order to locate nucleosomes in the small region. For those interested in transcription factors, promoter arrays are commonly used. On those arrays, several hundred base pairs upstream of the transcription start sites are typically covered for all genes in the genome. The customized array approach was first pioneered by NimbleGen, with their ~60 mer oligonucleotide probes; currently Agilent and other companies also offer this solution, each with its differing manufacturing technique and probe selection criteria. There is obviously a trade-off between the spatial resolution between probes and the total size of the region covered, but the ability to customize the array has enabled much progress. Analytical tools for ChIP-chip data have matured over the years, and a variety of tools are available for processing the data and predicting binding sites.[8-12] In order to estimate binding locations, these methods use various smoothing techniques to reduce noise in the data and search for regions in which consecutive probes along the chromosomes give consistent signals.[13]

ChIP-seq

Rather than hybridizing the genomic DNA fragments enriched for a protein or histone modification, the ChIP-sequencing (ChIP-seq) approach aims to directly sequence them. There are several new techniques for the sequencing step. In the Solexa platform, for instance, a library of adapter-ligated ChIP DNA fragments is constructed and loaded onto a solid substrate. This is followed by cluster amplification that generates many clonal copies of each fragment. Each cluster then is subject to 'sequencing-by-synthesis' in which fluorescently-labeled reversible terminator nucleotides are added and high-resolution image is taken at each base pair. The nucleotide sequence for each cluster can be deduced from analysis of the fluorescent signal on the image. More details can be found elsewhere.[1,2]

Currently, there are three next-generation sequencing platforms available: (1) pyrosequencing by 454 Life Sciences, later acquired by Roche ("454"); (2) Genome Analyzer by Solexa, later acquired by Illumina ("Solexa"); and (3) Sequencing by Oligo Ligation/Detection ("SOLiD") by Applied Biosystems. 454 was the first to be introduced to the market, followed by Solexa, and then SOLiD. The Solexa and SOLiD technologies can sequence 35–50 bp fragments reliably at this time while 454 can sequence 200–400 bp fragments. But the first two are more recent technologies and are substantially less expensive for applications that do not require long read length. Between Solexa and SOLiD, Solexa became available a year or two before SOLiD and thus has been more widely used. Most of the published studies on ChIP-seq so far have used the Solexa platform. In the current Solexa configuration, a flow cell is divided into eight lanes, onto which different samples can be loaded. Thus, a single lane is the standard unit for a sample, and it can currently generate 8–12 million reads. A single run of the entire flow cell therefore produces close to 100 million tags or, at 35 bp per tag, more than thre gigabases. The actual numbers depend on the exact version of the sequencer, the technical ability of the person running the experiment and experimental variations that are difficult to control.

Compared to the Sanger sequencing, the main disadvantage of the Solexa and SOLiD platforms is their short read length. Although it is expected to increase to 80–100 bp soon, the current read length makes certain applications less amenable for direct sequencing. Whole-genome sequencing, for instance, has recently begun on next-generation sequencing, but genome assembly from short sequences poses significant difficulties and requires a high-fold coverage of the genome. On the other hand, 35 bp is sufficient to map a tag uniquely to non-repetitive regions of a genome and does not impose a limitation for ChIP-seq.

The cost of sequencing millions of the short fragments has been prohibitively expensive in the past, but next-generation sequencing has made it feasible. As of late 2008, the cost of ChIP-seq is still higher than that of ChIP-chip, but it offers higher-quality data, as described below. Just as big an obstacle as cost has been the availability of a sequencing machine and well-trained staff, and the lack of familiarity in the academic community. However, this is likely to change in the next few years, as more academic institutions adopt this technology and the cost of sequencing continues to decrease.

ChIP-chip vs ChIP-seq

There are three main advantages to ChIP-seq compared to ChIP-chip. The first is the single base-pair resolution of direct sequencing. For some factors, resolution of the binding sites on the order of 50–100 bp or larger on microarrays is sufficient; for others, however, more precise location is informative. In particular, determining the sequence motif responsible for binding can often be facilitated by sharper, more precise set of binding positions along the genome.[14] Second, ChIP-seq data are likely to have less noise or artifacts. The quality of microarray data relies on hybridization chemistry between the probe and the DNA target; this is followed by measurement of fluorescent signal using a scanner. In both steps, much noise and artifacts are introduced. Tiling the genome at an equal interval for ChIP-chip does not allow flexibility for selecting probes with desired characteristics, such as having a GC-content in a specified range and not having a similar sequence elsewhere in the genome. The GC-content of the probe is known to influence the hybridization chemistry substantially, and having a wide range of GC-content results in much non-specific binding. While sequencing is not immune to GC-content bias,[15] it is likely to be less than the bias observed on arrays. In terms of noise from signal quantification, the amount of signal captured through a scanner does not vary linearly with the number of fragments bound for ChIP-chip, especially at the high end of scale where one often observes signal saturation.

In contrast, ChIP-seq captures the absolute number of fragments mapping to a specified region, given the total number of fragments mapped in that experiment. The third advantage of ChIP-seq is that potential binding regions need not be specified prior to experiment. Heterochromatic regions, for instance, are generally not represented on microarrays. Binding in such regions can be detected with ChIP-seq, although incomplete genome assembly limits analysis of those regions.

ChIP-seq has some disadvantages as well, in addition to the current higher cost. As is the case with any nascent technology, results of a sequencing run can vary, both in the number of reads generated and their quality. Although the situation is improving, it was not uncommon for a new sequencing machine to be out of order half the time, due to various malfunctions. In terms of experimental design, the number of tags needed to characterize a protein binding or modifications is highly variable, while the number of sequence tags generated from a single sequencing run is relatively fixed. Thus, one must determine for each experiment whether enough tags have been sequenced and, if not, how much more should be sequenced. The required number of tags depends on two aspects: how widespread the binding or modification is and how large the genome is. For pervasive histone modifications, such as H3K36me3, a histone mark associated with transcription elongation, many more tags are required to see the enrichment over background over large regions. For organisms with large genomes, e.g., humans, the problem is compounded, as is the case with tiling arrays. Even worse, how many tags would be required cannot be estimated a priori except in cases where much is known about the protein or modification.

This large tag requirement becomes an immediate issue even for proteins with sharp binding features, because recent studies have made it clear that input DNA should also be sequenced as a control if some false positive sites are to be avoided. Data show that some segments of the genome exhibit enrichment even in the input DNA profile, most likely due to increased fragmentation in open chromatin regions or bias in amplification. But to see this effect clearly, enough tags are needed to cover most of the genome with some depth. Some large studies in progress for humans are sequencing 20–30 million input tags; this is likely to be the minimum required. The basic question of whether enough tags have been obtained in an experiment is a complex one: as more tags are sequenced, more regions of smaller fold-enrichment are generally found at a given statistical threshold. Indeed, in many cases, one does not observe a true 'saturation' of the peaks; instead one can only speak of saturation at a particular fold-enrichment. These issues are addressed in a recent manuscript.[16]

One important technology to alleviate the tag requirement is the 'capture-and-release' technology. In this approach, a microarray is designed with the probes matching to the regions of interest and it is used as a filter to capture the fragments that are to be sequenced. If one is only interested in the profiles over exons, for instance, those fragments can be captured, released and then sequenced using this approach. Some proof-of-concept work has been published[17-19] and it is expected to be widely available soon.

Nucleosome Positioning

One important area in which the high resolution of ChIP-seq has brought tremendous progress is in determining nucleosome positions in the genome. Nucleosomes play a fundamental role in gene regulation by packaging genomic DNA, altering chromatin structure and modulating accessibility of proteins to genomic loci. Profiling the nucleosome locations on a genome-scale finally became possible with tiled microarrays in 2005 for yeast,[20] but profiling them in humans has been limited to specific regions such as the promoters[21] or the HOX regions.[7] Even with 10 bp[21] or 5 bp[7] resolution, data are noisy and nucleosome positions cannot be determined precisely.

With ChIP-seq, precise estimation of nucleosome positions has become possible. This has resulted in genome-scale maps of nucleosomes containing histone variant H2A.Z in *S. cerevisiae*[22] and humans,[23] nucleosomes with various methylation and acetylation marks on the histone tails,[23,24] and, most recently, bulk nucleosomes in Drosophila,[25] *C. elegans*[26] and humans.[27] These profiles have identified nucleosome-free regions in the genome, highlighted the dynamic role of nucleosomes especially near the transcription start sites and allowed investigation of potential sequence elements that influence nucleosome positioning.

Consortium Projects

While next-generation sequencing has the potential to bring new understanding of chromatin structure and epigenetic mechanisms, the number of chromatin-associated proteins and histone modifications to profile is too large for any single laboratory. The US National Institutes of Health (NIH) has recognized this and has started a project that involves a consortium of laboratories across the country. The pilot Encyclopedia of DNA Elements (ENCODE), started in 2003, was an effort to comprehensively characterize 1% of the human genome by identifying all sequence-based functional elements using all available high-throughput platforms.[28] This project is being followed by two projects: model organism ENCODE and full-scale human ENCODE. The model organism project focuses on *D. melanogaster* and *C. elegans*, and its chromatin component involves mapping of nearly 100 chromatin factors and histone modifications across multiple cell lines and developmental stages. Other components include profiling transcriptome, transcription factor binding sites, microRNAs and histone variants. The human ENCODE also has a chromatin component, which covers a smaller number of factors on a larger number of cell lines. The latest addition to the consortium projects is the Roadmap Epigenomics Program[29] that aims to produce reference epigenomes of a variety of human cells, including embryonic stem cells, differentiating cells, and a subset of cell lines representative of human disease.

Summary and Challenges

While ChIP-chip has brought many successes, ChIP-seq allows one to map factors in a genome-wide manner at base pair resolution, with significant reduction in noise and without predefining regions of interest. This represents a new opportunity for understanding the role of nucleosomes, chromatin-associated proteins, and histone modifications in epigenetics. In particular, several consortium projects are set to generate a staggering amount of data, to be publicly available. As of late 2008, the cost for ChIP-seq is still too high for routine use in most laboratories, but this is expected to change rapidly. The lack of validated antibodies is still a major problem in comprehensive characterization of the epigenome, although it is being addressed to some extent in the consortium projects described above.

For most investigators, the main obstacle will soon be not in generation of data but in their interpretation. This is exactly as it was nearly a decade ago for microarrays: there was an initial period of excitement in which even cursory analysis of new data resulted in novel findings; gradually, it became clear that a simple listing of relevant genes gave little insight into the underlying mechanisms and that more sophisticated analysis and extensive validations were necessary. For ChIP-seq data, computational challenges are greater. A single run generates a terabyte of raw data and it is in most cases not even feasible to keep the data for a long period of time. Many basic issues in analysis, such as defining minimum depth of sequencing and accurate identification of both sharp and broad binding sites, are not fully resolved yet, although much progress is being made. More challenging will be integrative analysis to correlate features from multiple profiles and across multiple data types. Therefore, taking full advantages of the data will require close collaboration with those who have computational skills. It will be in the judicious mix of molecular biologists' knowledge in where to look and computational biologists' expertise in how to look that will result in novel hypotheses.

Acknowledgements

The author thanks P. Kharchenko, M. Tolstorukov and A. Alekseyenko for critical reading of the manuscript. Research in the author's laboratory is supported by grants from the National Institutes of Health. This work was supported by the National Institutes of Health (DK052356)

References

1. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008; 9:387-402.
2. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008; 26:1135-45.
3. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, et al. Genome-wide location and function of DNA binding proteins. Science 2000; 290:2306-9.
4. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, et al. A high-resolution map of active promoters in the human genome. Nature 2005; 436:876-80.
5. Alekseyenko AA, Larschan E, Lai WR, Park PJ, Kuroda MI. High-resolution ChIP-chip analysis reveals that the Drosophila MSL complex selectively identifies active genes on the male X chromosome. Genes Dev 2006; 20:848-57.
6. Alekseyenko AA, Peng S, Larschan E, Gorchakov AA, Lee OK, Kharchenko P, et al. A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. Cell 2008; 134:599-609.
7. Kharchenko PV, Woo CJ, Tolstorukov MY, Kingston RE, Park PJ. Nucleosome positioning in human HOX gene clusters. Genome Res 2008; 18:1554-61.
8. Buck MJ, Nobel AB, Lieb JD. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. Genome Biol 2005; 6:97.
9. Toedling J, Skylar O, Krueger T, Fischer JJ, Sperling S, Huber W. Ringo—an R/Bioconductor package for analyzing ChIP-chip readouts. BMC Bioinformatics 2007; 8:221.
10. Peng S, Alekseyenko AA, Larschan E, Kuroda MI, Park PJ. Normalization and experimental design for ChIP-chip data. BMC Bioinformatics 2007; 8:219.
11. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, et al. Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci USA 2006; 103:12457-62.
12. Du J, Rozowsky JS, Korbel JO, Zhang ZD, Royce TE, Schultz MH, et al. A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. Bioinformatics 2006; 22:3016-24.
13. Johnson DS, Li W, Gordon DB, Bhattacharjee A, Curry B, Ghosh J, et al. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. Genome Res 2008; 18:393-403.
14. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science 2007; 316:1497-502.
15. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 2008; 36:105.
16. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of protein binding experiments from ChIP-sequencing. Nat Biotechnol 2008; in press.
17. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, et al. Multiplex amplification of large sets of human exons. Nat Methods 2007; 4:931-6.
18. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 2007; 4:907-9.
19. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007; 4:903-5.
20. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, et al. Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science 2005; 309:626-30.
21. Ozsolak F, Song JS, Liu XS, Fisher DE. High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol 2007; 25:244-8.
22. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, et al. Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. Nature 2007; 446:572-6.
23. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell 2007; 129:823-37.
24. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 2008; 40:897-903.
25. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, et al. Nucleosome organization in the Drosophila genome. Nature 2008; 453:358-62.
26. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res 2008; 18:1051-63.
27. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell 2008; 132:887-98.
28. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447:799-816.
29. Moving AHEAD with an international human epigenome project. Nature 2008; 454:711-5.