Published in final edited form as:
Epigenomics. 2009 December 1; 1(2): 319–329. doi:10.2217/epi.09.29.

High-resolution mapping studies of chromatin and gene regulatory elements

Alan P Boyle and **Terrence S Furey**[†]

Abstract

Microarray and high-throughput sequencing technologies have enabled the development of comprehensive assays to identify locations of particular chromatin structures and regulatory elements. It is now possible to create genome-wide maps of DNA methylation, *trans*-factor binding sites, histone variants and histone tail modifications, nucleosome positions, regions of open chromatin, and chromosome locations and interactions. This review provides a summary of these new assays that are changing the way in which molecular biology research is being performed. While the generation of large amounts of data from these experiments is becoming increasingly easier, the development of corresponding analysis methods has progressed more slowly. It will likely be years before the full extent of the information contained in these data is fully appreciated.

Keywords

ChIP; chromatin; chromatin immunoprecipitation; DNA methylation; DNA nuclease I hypersensitive; DNaseI HS; histone; nucleosome

The rapid completion of the sequencing of many eukaryotic genomes is providing researchers with a necessary scaffold for advancing our knowledge of gene function and regulation. Annotating locations of transcribed protein-coding sequences within genomes has been a major initial focus, but simply knowing their locations does little to reveal just how their expression is regulated. The corresponding discovery of non-coding regulatory elements has lagged behind, owing to a much weaker or complete lack of signal in the primary DNA sequence. Previous attempts to identify regulatory elements have either focused on computational techniques to identify transcription factor binding sites and *cis*-regulatory modules, or were based on lower-throughput chromatin assays that looked at a small number of loci or functional assays that perturbed sequences within regulatory regions. Computational methods have a low specificity, producing many false-positives and providing no information about cell-specific or condition-specific regulation. Traditional wet laboratory experiments lack global sensitivity, since they were restricted to analyzing only small regions of the genome.

Chromatin structure has been known to play a critical role in gene regulation. DNA accessibility, largely governed through the global and local positioning of nucleosomes, can be altered, preventing transcription factors from binding DNA; histones and their modifications

[†]Author for correspondence: Institute for Genome Sciences & Policy, Duke University, Durham, NC, USA Tel.: +1 919 668 4728 Fax: +1 919 668 0795 terry.furey@duke.edu.

Financial & competing interests disclosure
APB and TSF were supported by NIH grant U54-HG004563 and the Institute for Genome Sciences & Policy at Duke University. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.
No writing assistance was utilized in the production of this manuscript.

influence chromatin structure and regulatory function in ways not well understood; and chromatin looping can allow regulatory elements to regulate transcription of genes long distances away on the same or even on a different chromosome. In addition, we know that methylation of DNA can alter local binding affinities to *cis*-regulatory sites and influence the recruitment of chromatin remodeling factors. These epigenetic structures and marks are crucial to cellular function, but previous studies have been limited, often providing an incomplete picture. For example, some of these chromatin modifications were initially considered axiomatic when studied at only specific loci, such as individual transcription start sites. However, more global studies revealed that some specific modifications were not as important as first theorized based on the limited data. An accurate and comprehensive characterization of chromatin has been difficult as this information cannot be accurately deduced from DNA sequence alone. In addition, chromatin organization can be highly variable among different cell types and under different cellular conditions, emphasizing the need to look beyond static DNA for its characterization.

The development of assays employing microarray and, more recently, high-throughput sequencing technologies now allow for high-resolution measurements of chromatin structure, modifications and locations of specific regulatory factors genome-wide. Based on analyses of these data, we are beginning to appreciate the complexity and nuances of epigenetic influences on gene regulation, indicating the importance of these more comprehensive assays. Researchers now have the tools to map DNA methylation, transcription factor binding sites, nucleosome positions, histone modifications and even interactions between distant chromosomal regions. Projects such as the ENCyclopedia Of DNA Elements (ENCODE) [1,2], model organism ENCODE (modENCODE) [3,4] and the Epigenomics Roadmap initiative [101] are catalyzing the genome-wide annotation of noncoding regulatory elements at a high resolution across a diverse range of cells types and conditions in multiple species. These consortia, as well as work from individual laboratories, have begun to reveal a more complete picture of chromatin, improving our understanding of its role in gene regulation. We can now observe larger scale phenomena that are proving to be much more complex than originally thought. An illustrative example of the quality of these data can be seen in Figure 1.

In this review, we provide a brief update on the current experimental techniques that are rapidly expanding our knowledge of chromatin structure and gene regulation. Importantly, new technologies have allowed for a rescaling of assays from analyzing small, individual loci to complete or nearly complete genome-wide studies. With the creation of diverse sets of whole-genome chromatin data, more comprehensive analyses to understand how chromatin organization changes across cell types and conditions can be considered. We focus specifically on assays that have been developed using microarray and high-throughput sequencing technologies. Details of specific experimental protocols will not be discussed. Instead, we will describe the types of information that these experiments can produce, include a brief comparison of microarray and sequencing technologies, and discuss the analytical challenges associated with these massive datasets. We begin our discussion with the assays that interrogate modifications at the lowest level, the DNA itself, followed by increasingly more general characterizations including transcription factor binding, histone modifications, nucleosome structure and chromosomal conformation.

DNA methylation

The most common form of DNA methylation occurs on cytosine bases that are 5' of guanine bases (CpGs). This modification is found in many species, though it is not ubiquitous, being absent in notable organisms such as *Caenorhabditis elegans*. DNA methylation can silence gene expression bi-allelicly or mono-allelicly (e.g., random inactivation or imprinting). This silencing is thought to occur in two ways: the disruption of transcription factor binding and/or

the recruitment of chromatin silencing factors (such as histone deacetylases) to the locus [5, 6]. The latter mechanism involves proteins with methyl-CpG-binding domains that recognize methylated CpGs. The loss of these proteins has been associated with several diseases and cancers [7]. In addition, hyper- and hypomethylation of key genomic regions have been linked to carcinogenesis [8].

There are a wide variety of assays to measure DNA methylation (review available [6]), but it has been difficult to measure on a genome-wide scale and at a high resolution. Recently, several different strategies have been employed to produce more comprehensive maps. Methylated DNA immunoprecipitation (MeDIP) uses chromatin immunoprecipitation of methylated cytosine in combination with microarrays (MeDIP-chip) [6] or high-throughput sequencing (MeDIP-seq) [9] to provide genome-wide methylation maps. Although this has been demonstrated to be highly accurate and comprehensive, two limiting factors of this assay are the quality of the methylated DNA antibody and the shearing size resolution. Even so, a recent study [9] was able to achieve a resolution of 100 bp giving good coverage of CpG islands. This technology has been used to create genome-wide DNA methylation profiles in several different human tissues, allowing for the identification of tissue-specific differentially methylated regions (tDMRs) [10].

An alternative to immunoprecipitation is the conversion of nonmethylated cytocines into uracil through sodium bisulfite treatment. The resultant DNA can be sequenced at individual queried sites (methylation-specific quantum dot fluorescence resonance energy transfer [MS-qFRET]) [11], at captured genomic regions [12-15] or genome-wide (bisulfite sequencing [BS-seq] [16] and MethylC-seq [17]). While bisulfite sequencing techniques can provide quantitative single base pair resolution, the initial alignments of these sequences to a genome can be confounded by the cytosine to uracil conversions resulting in a lower overall uniqueness of the genome. Whole-genome sequencing to identify methylated DNA has been practical at this point in *Arabidopsis thaliana* [17]; the task is more difficult in the 20-fold larger human genome and other mammalian genomes.

To avoid sequencing entire genomes, researchers have developed techniques to generate sequence reads only at specific cytosines (reduced representation bisulfite sequencing [18] and Methyl-seq [19]). Meissner *et al.* use a restriction enzyme MspI to cut at all CCGG positions regardless of methylation status, convert with bisulfate and then sequence [18]. Brunner *et al.* use MspI as a measure of all sites and also cut with HpaII, which only recognizes unmethylated instances of the same site to identify which sites have been modified [19]. While this restriction site represents only a fraction of all CpGs, it gives some signal in almost all CpG islands, and results in to-the-base resolution of measurable cytosine methylation.

These new experiments have led to the comprehensive identification and confirmation of relationships between DNA methylation, cellular function and their links to disease. For example, Meissner *et al.* mapped methylation states in most CpG islands in mouse embryonic stem cells at single-base resolution [18]. These experiments revealed the distribution and pattern changes of methylation that occur during differentiation, and demonstrated that they are strongly correlated to changes in histone modifications. Korshunova *et al.* comprehensively analyzed methylation patterns in normal and cancerous breast tissues, demonstrating not only the complexity of these patterns, but also the promise of using differences in normal and cancerous tissues as diagnostic and prognostic markers [12]. Brunner *et al.* used Methyl-seq to describe methylation state differences in human embryonic stem cell (hESC) differentiation and fetal liver development [19]. They found that hESC differentiation requires relatively small numbers of methylation changes, specifically at H3K27me3-occupied regions, bivalent domains and low-density CpG promoters, and consists primarily of demethylation *in vivo*. Ball

et al. found that highly expressed genes are characterized by low promoter methylation and high gene-body methylation [13].

Trans-acting factors

Proteins that interact directly with DNA can play a key role in determining the structure of chromatin. Many of these factors affect chromatin properties through the recruitment of chromatin modifying factors such as DNA methyltransferases and histone-modifying proteins, or through the steric displacement of nucleosomes. Other factors, such as polycomb and ATP-dependent chromatin remodeling complexes, can directly alter the chromatin structure when interacting with *cis*-elements or histone marks. The large number of possible factors and the large number of possible interacting sites emphasizes the need for more high-throughput methods of analysis.

Chromatin immunoprecipitation (ChIP) has been an important technique used to identify sites of *trans*-factor interaction [20-22]. ChIP involves cross-linking proteins to DNA and creating a DNA library enriched for sequences bound by a particular protein of interest using a specific antibody to that protein. This technique can be combined with microarrays (ChIP-chip) or high-throughput sequencing (ChIP-seq) to identify specific protein-DNA interaction sites in a limited number of genomic regions or genome-wide at approximately 50-100-bp resolution. This limitation in the resolution of ChIP is inherent in the procedure of fragmenting DNA and enriching for fragments with the bound factor. These fragments are of varying size, typically a couple of hundred bases in length, and do not accurately represent the typical 6–20-bp interaction site of *trans*-acting proteins. Nevertheless, a large number of factors have been assayed to date across many organisms, including RNA polymerase II (PolII) [23], STAT1 [24], CCCTC-binding factor (CTCF) [23,25], GABP [26], SRF [26], neuron-restrictive silencer factor (NRSF) [26,27], FoxA2 [28] and FoxA3 [29] in human cell lines. Kim *et al.* provided comprehensive maps of CTCF binding, a factor known to act as an insulator influencing both chromatin structure and gene regulation, in human primary fibroblasts [25]. Most sites were found in regions far from transcription start sites of genes, although their distribution correlated well with locations of genes. Preliminary analysis in 1% of the genome for multiple cell types suggested CTCF binding is not highly variable across different cells. Johnson *et al.* mapped approximately 2000 binding sites of neuron-restrictive silencer factor (NSRF) genome-wide [27]. As expected, they demonstrated that this factor regulates genes involved in neurons and their development. Surprisingly, though, they also found enrichment in genes that drive islet cell development in the pancreas.

Comprehensive identification of transcription factor binding sites has been, and still is, an active area in bioinformatics. Extending this type of research to study the interaction of transcription factor regulatory networks plays a major role in the field of systems biology. There are many tools to determine preferred binding sequences, or motifs, for a specific factor [30-32]. However, some proteins may not directly interact with DNA or may bind nonspecifically, nullifying the utility of motif identification. In addition, computational techniques that apply these motifs to the entire genome are plagued by numerous false-positives owing to such problems as not knowing the chromatin accessibility of specific regions, the unknown requirements of additional factors for binding, and inadequate information content in the motif. The development of tools that combine ChIP information with motifs have the potential to use strengths from both tools to allow for the more accurate prediction of exact binding sites.

Histone tail modifications & histone variants

Regulation of biological processes can also be directly affected by modifications to the core histone proteins that comprise the nucleosome. Each nucleosome consists of an octamer

comprised of two each of histones H2A, H2B, H3 and H4, making histones the most abundant protein component of chromatin. Histone variants, such as H2A.Z and CENPA, can replace one of the normal core histones and are involved in key cellular processes such as transcription, repair and replication [33]. Post-translational modifications to the histone tails have been shown to alter the structure of chromatin [34]. Modifications include mono-, di- and tri-methylation, acetylation, ubiquitination and phosphorylation of specific amino acids in histone tails. The list of these modifications is growing, and elucidating these is a major focus of the Epigenomics Roadmap initiative. Different histone modifications have been associated with many aspects of the genome, including transcriptional silencing, transcriptional activation, active transcriptional units, enhancers, DNA repair and other genomic features. For a full review, see [35].

Antibodies to histone variants and specific histone tail modifications have been developed, enabling the use of ChIP experiments to identify genomic locations of specific histones and histone modifications. As the histone is part of a larger nucleosome, the resolution of the positioning of the histone need not be on the single-base level. Unlike the ChIP experiments described above, fragments targeted by antibodies can be isolated by cleaving DNA in the linker regions between nucleosomes with micrococcal nuclease (MNase) or through sonication. The resulting locations of the modifications should indicate enrichment for a modification at a nucleosome-level resolution. However, these experiments cannot determine whether both or just one of a particular histone has been modified. In addition, as mentioned below, some issues limit resolution, such as imprecise nucleosome positioning and the large number of sequence reads required in organisms with larger genomes.

Recently, many groups have performed ChIP-chip [1,36] and ChIP-seq [23,37,38] to identify locations of histone methylations, acetylations and a limited number of variants. These have provided high-resolution maps across entire genomes for many common modifications and variants, allowing for the further study of their relationship to cellular function. The sheer number of known possible modifications and variants under all conditions has limited the comprehensiveness of these studies to modifications whose functions are better understood, but the roles for even these are more complex than previously understood.

Mikkelsen *et al*. mapped the locations of five modifications genome-wide in mouse pluripotent and lineage-committed cells [39]. They found that trimethylation of lysine 4 and 27 on histone 3 (H3K4me3 and H3K27me3) could effectively distinguish being expressed genes (H3K4me3 present), stably silent genes (H3K27me3 present) and those poised for expression (both marks present). In addition, another mark, H3K36me3, appears throughout actively transcribed coding and noncoding regions, allowing accurate gene annotation in a cell-type-specific manner. They also demonstrated an additional benefit of sequencing-based experiments by using heterogeneous polymorphic sites to identify allele-specific transcription. Barski *et al.* annotated locations of 20 histone lysine and arginine methylations and the histone variant H2A.Z in human T cells [23]. They found that monomethylations of H3K27, H3K9, H4K20, H3K79 and H2BK5 are associated with actively transcribed genes, while the trimethylation of H3K27, H3K9 and H3K79 are linked to silent genes. They also performed ChIP-seq for CTCF in these cells, and showed that CTCF is found at the edges of various methylation domains.

Nucleosome positioning & open chromatin

The combination of the histones within the nucleosome octamer and the genomic DNA wrapped around it allows for the steric regulation of transcriptional activity. Each nucleosome interacts with approximately 146 bp of DNA, rendering these bases essentially inaccessible by many factors. The prevention of access for these factors can allow transcriptional modulation through both proximal regulation (nucleosomes blocking the promoter region) and distal regulation (nucleosomes blocking enhancer elements).

Precisely how nucleosomes are positioned in a particular cell type is currently unclear. In general, it is thought that nucleosomes interact with DNA as a default state, and thus displacement of the nucleosomes is required for access by other factors. The act of displacement can be through direct factor interaction with its preferred binding site [40], mediated by an ATP-dependent complex such as switch/sucrose nonfermentable (SWI/SNF) [41-43], or through apparent acetylation prior to transcription [44]. It is generally found that at the promoter region of actively transcribed genes, nucleosomes are completely removed. The same is not true in the body of these genes. It has been hypothesized the RNA polymerase complex does not completely displace nucleosomes during transcription, somehow retaining and/or reinserting them once the polymerase has passed. This is supported in a study by Dion *et al.* in yeast where it was demonstrated that nucleosomes in the gene body of actively transcribed genes were not actively replaced with specially labeled nucleosomes that had been added [45]. Nucleosomes are also either removed or displaced to allow the binding of regulatory proteins. Therefore, regulatory elements in general can be identified by mapping the locations of nucleosomes or detecting where they are absent.

Several studies have demonstrated that certain DNA sequence patterns, such as the oscillation frequencies of particular dinucleotides, influence nucleosome positioning. Computational models based on these sequence characteristics can generate predictions in yeast that are highly correlated with *in vivo* nucleosome positions [46,47]. While these models seem to demonstrate some statistical accuracy, others postulate that these sequence patterns are primarily found when nucleosomes need to be precisely positioned and that other nucleosomes are placed through statistical packing [48]. For a review of nucleosome positioning, see [49]. In general, these models have not been able to provide accurate mappings genome-wide and are limited as they cannot show cell-type-specific changes in chromatin.

Positions of nucleosomes can now be identified using MNase, which has been shown to efficiently digest the linker regions between two nucleosomes. High-resolution genome-wide nucleosome maps can be generated by extracting nondigested DNA and employing tiled microarrays or sequencing. This was successfully carried out first in yeast, where nucleosome maps have now been created under various conditions and determined using both sequencing and array methods [47,50–52]. Subsequently, genome-wide maps have been generated for *C. elegans* [53], *Drosophila melanogaster* [54] and humans [55]. While the tiling arrays generally provide lower resolution annotations than sequencing, hidden Markov models have been used to generate maps with as high as 10-bp resolution [56]. These studies have demonstrated that there are both well-positioned nucleosomes and nucleosomes whose exact positions seem to vary across the cell population. The promoter regions of genes tend to have well-positioned nucleosomes that are phased with respect to each other.

While a sequencing approach to determine nucleosome positions in species with relatively small genomes such as yeast is feasible, equivalent sequencing depth in larger mammalian genomes requires significantly more work. In humans, it has been demonstrated that as few as 100 million short reads provide an accurate map of well-positioned nucleosomes such as those found at transcription start sites [55]. However, equivalent coverage in humans to generate nucleosome positioning maps similar to those in yeast may likely require over 10 billion sequence reads assuming similar genomic nucleosome occupancy.

In contrast to identifying the locations of nucleosomes, some researchers are interested in identifying regions that are nucleosome free, also referred to as open chromatin. DNA nuclease I (DNaseI) has been shown to preferentially digest DNA in nucleosome-depleted regions. These DNaseI hypersensitive (HS) sites have been used to annotate promoters, enhancers, silencers, insulators and locus control regions [57]. Genome-wide assays that comprehensively identify DNaseI HS sites have recently been developed using tiling microarrays and high-

throughput sequencing [58–64]. An additional method, formaldehyde-assisted isolation of regulatory elements (FAIRE), has also been shown to identify open chromatin regions in a completely different way. FAIRE is a rather straightforward experiment that isolates DNA not cross-linked by formaldehyde to bound proteins, primarily nucleosomes, and then determines the locations of these protein-free regions using tiled microarrays or sequencing [65,66]. FAIRE has been shown to be highly associated with DNaseI HS sites and other chromatin marks. In humans, approximately 2% of the genome is nucleosome-free in a given cell type; therefore, identifying nucleosome-depleted regions requires significantly less sequencing than when determining positions of all nucleosomes.

Identifying nucleosome-free regions provides clues as to the location of active regulatory elements, but this does not reveal the function of these elements, nor what factors may be bound. It has been previously shown that DNaseI experiments can identify precise locations of individual transcription factor-binding sites, referred to as DNaseI footprinting. This utilizes the fact that *trans*-factors also protect the DNA from digestion similar to nucleosomes, but at a much smaller scale. Recently, it has been demonstrated in yeast that the high-throughput DNaseI sequencing protocol can perform whole-genome DNaseI footprinting with single base pair resolution [67]. These footprints can be compared with known factor-binding motifs to predict the particular protein interacting within that segment of DNA, potentially providing an idea of the function of the putative regulatory element. As there are many factors whose binding motif is unknown or where binding is nonspecific, combining DNaseI footprints with ChIP data for specific transcription factors can provide the precise positioning of a factor's binding site, revealing more precisely the DNA binding characteristics. Alternatively, Kang *et al.* have proposed a protocol in which ChIP is performed prior to footprinting [68].

Nuclear localization of chromosomes

While the aforementioned experimental assays map chromatin along the strands of DNA (essentially 1D data), the true structure of chromatin resides in a 3D world with chromosome loops and folds. Interactions between distal regions of chromatin can explain how enhancers many kilobases away from a gene, and even on a different chromosome, can affect the expression of that gene. Within the nucleus there are compartments of regulation that are associated with expression or repression of genes. In mammals, active gene regulation tends to take place away from the nuclear envelope, while repressed regions tend to be sequestered to the nuclear envelope, although there are numerous exceptions to these tendencies. In *Saccharomyces cerevisiae*, these trends seem to be the opposite [69]. A better understanding of this higher-order structure of chromosomes will greatly enhance our understanding of data generated from experiments described above and gene regulation as a whole.

Fluorescence *in situ* hybridization assays have been the standard for mapping chromosomal locations for many years. While these experiments have led to important discoveries, they are restricted in the number of regions that can be examined simultaneously and have limited resolution. More recent techniques, such as chromosome conformation capture (3C [70]), can reveal characteristics of chromosomal positioning at high-resolution on a much larger scale. The 3C protocol first cross-links interacting segments of chromatin and then identifies the genomic locations of these interactions. This technology has rapidly progressed from initially being limited to assaying only one predetermined pair of sites (3C [70]), to revealing all interactions for one specific site (4C [71,72]), and can now reveal all interactions for all sites within a specific genomic region (5C [73]). This 5C approach relies on microarrays or high-throughput sequencing to map these interactions. The accuracy and utility of this method was demonstrated by mapping all interactions in a 400-kb region encompassing the human β-globin locus that clearly showed strong links between the locus control region (LCR) and globin genes that are separated by 10–60kb [73] (see review at [74] for further details).

High-resolution technologies

As shown above, microarray and sequencing technologies are actively being employed for creating high-resolution mappings of chromatin structure. Each of these technologies has benefits and concerns associated with it. The choice of the appropriate technology for a given research study depends on many factors, including number of samples, cost, amount of genome to be assayed, desired resolution and availability of informatics pipelines. Below, we briefly discuss the major strengths and weaknesses of each in their current form.

Microarrays represent a relatively mature technology that has well-developed analysis protocols. Important issues such as normalizing results with an appropriate experimental control have been extensively studied. Custom microarrays can be designed to query a subset of a genome, making them cost-effective for multi-sample analyses. For small genomes, high-density tiling arrays have been or can be designed, providing high-resolution results. However, for larger genomes this is less practical and generally results in lower resolution due to probe spacing. Probe design and cross-hybridization concerns are better understood, but these have not been completely resolved and can still result in experimental artifacts. The repetitive nature of many genomes affects what regions can be assayed and leads to uneven spacing of probes in many regions.

High-throughput sequencing of short sequence tags is a relatively new technology that theoretically allows for whole-genome experimental coverage of any organism with a reference genome sequence assembly. For many of the experiments described above, sequencing technologies can produce results with near single-base resolution. Higher genomic coverage can be achieved compared with microarrays owing to the ability to map tags to short unique regions that are largely repetitive and to regions that are polymorphic or duplicated. In addition, single allele phenomena can be described with the availability of informative heterogeneous SNPs and sufficient sequencing depth. However, as with any new technology, much more work is needed to understand how best to process these data and to identify experimental artifacts. Analyses tools are still relatively immature. While information is generated from the whole genome, short sequence tags cannot be confidently mapped in many instances, and it is not clear how to use sequences that align to multiple genomic locations. Polymorphic sites and single base sequencing errors are more problematic owing to the short length of many sequence tags. The appropriate design and use of experimental controls is not well understood.

Despite their limitations, both microarray and sequencing technologies have been demonstrated to produce very high quality and accurate data when used in many different experimental settings. For the basic identification of genomic regions of interest, one technology has not been clearly shown to be superior. The trend towards the use of sequencing technologies seems to be motivated by the ability to produce higher resolution results, and the promise of increased information from having actual sequence data.

Conclusion

Technology advances have spurred an explosion in the development of many high-resolution assays that are characterizing chromatin and regulatory elements at scales not previously imaginable. Data are being generated at an amazing rate from both individual laboratories and large-scale projects. Results for many of these experiments from the ENCODE and modENCODE projects are becoming available at online resources [102,103]. Public database resources, such as the Gene Expression Omnibus (GEO [104]) and the Short Read Archive (SRA [105]), are being inundated with these data. Most of the hurdles for generating these data have been cleared, and it is simply now a problem of time, money and resources for this production aspect of this type of research. The ability to perform comprehensive analyses of these data is another story.

Most experimental wet laboratories lack the computational infrastructure and trained personnel necessary to deal with these enormous datasets. Sequencing technologies that are still advancing and offering the promise of more accurate, more informative and cheaper data will almost surely compound this problem. Researchers are actively developing computational tools to assist in this task, but clearly there is much more work to be done. Results from each individual experiment hold a wealth of information, but many laboratories are unclear as to where to even start. Summary analyses can provide general characteristics, but may not be as useful when investigating individual loci. Where the most 'interesting' loci are is also unclear. Even greater insight is possible when results from multiple complementary assays are integrated, but methods to do this are largely still in the design phase. It may take years before the full impact of the data being created today is realized, and by then the amount of data will have increased many-fold. Nevertheless, being able to generate these types of data allows new types of questions to be addressed and invariably will further knowledge on many fronts.

Future perspective

New experimental protocols that accurately detect and map DNA methylation, *trans*-factor binding, histone modifications, nucleosome positions, open chromatin and nuclear localization of chromosomes are revolutionizing our ability to define and dissect cellular functions, especially gene regulation. Along with whole-genome sequencing and transcript mapping and quantification, these high-resolution chromatin mapping studies are producing a clearer picture of the complexity of life at the molecular level. Diverse data will continue to be generated in an increasing variety of cells from different lineages, at different stages of development, from individuals with different phenotypic traits, including diseased, after exposure to different environmental stimuli and from different species. These will almost surely lead to the discovery of novel molecular mechanisms and states responsible for each cell's identity. For example, two current areas of research that will see definite direct benefits are systems biology and disease research. A primary focus of systems biology is modeling whole transcriptional networks and regulatory systems. This is currently being done primarily through the analysis of gene-expression data in different cell types and conditions, and thus inferring relationships between genes. Results from chromatin and regulatory element experiments will provide direct evidence of gene interactions, allowing for more accurate and complete modeling of regulatory systems. In disease research, the availability of whole-genome expression data has both increased our understanding of disease and provided new avenues for biomarkers and targets for new therapeutic drugs. New diagnostic tests based on measuring expression levels of a few key genes are currently in human clinical trials. This line of research can only improve if not only differentially regulated genes associated with diseases are identified, but also the molecular causes of this differential regulation uncovered. Biomarker and drug target discovery will be aided by and may eventually be based on identifying chromatin states and a more complete picture of the regulatory program of a particular disease. As costs related to the key technologies involved keep decreasing, performing these experiments on large numbers of individuals will become feasible.

Executive summary

■ High-throughput technologies have enabled the development of comprehensive assays describing chromatin structure characteristics and locations of regulatory elements.

DNA methylation

■ DNA methylation status can be discerned in general regions and at a single base pair resolution.

Trans-*acting factors*

■ Binding sites of chromatin remodeling and transcription factors can be identified genome-wide with high accuracy.

Histone tail modifications & variants

■ Regions containing histone variants and/or many types of different histone tail modifications can be localized to better understand their role in cellular processes.

Nucleosomes & open chromatin

■ Positions of each individual nucleosome can be found where they are stably positioned and more variable positions can be described.

■ Nucleosome-free regions of open chromatin corresponding to regulatory elements can be comprehensively identified.

Nuclear localization of chromosomes

■ Interacting genomic loci separated by thousands or millions of bases on the same chromosome or even from different chromosomes can be characterized.

Conclusion

■ Microarrays and sequencing technologies both produce high-quality and accurate data, but each have their benefits and drawbacks.

■ Much work is still necessary to better standardize and understand how the processing of these data should be performed.

■ Generation of these data is becoming easier. The current challenge is the computational processing, integration and understanding of the information in these data, a process that will likely take years.

Acknowledgments

We would like to thank Greg Crawford for his discussions and critical review of this manuscript.

Bibliography

Papers of special note have been highlighted as:

* of interest

** of considerable interest

1■. Birney E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007;447(7146):799–816. Mapping and integrative analysis of 1% of the genome at high resolution. [PubMed: 17571346]

2. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. Science 2004;306(5696):636–640. [PubMed: 15499007]

3. Celniker SE, Dillon LA, Gerstein MB, et al. Unlocking the secrets of the genome. Nature 2009;459 (7249):927–930. [PubMed: 19536255]

4. Spencer G. Fly and worm models to teach researchers about biology and medicine. NIH News. May 14;2007

5. Weaver IC, D'Alessio AC, Brown SE, et al. The transcription factor nerve growth factor-inducible protein a mediates epigenetic programming: Altering epigenetic marks by immediate-early genes. J. Neurosci 2007;27(7):1756–1768. [PubMed: 17301183]

6. Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res 2009;19(6):959–966. [PubMed: 19273618]

7. Amir RE, Zoghbi HY. Rett syndrome: Methyl-cpg-binding protein 2 mutations and phenotype-genotype correlations. Am. J. Med. Genet 2000;97(2):147–152. [PubMed: 11180222]
8. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat. Rev. Genet 2007;8(4):286–298. [PubMed: 17339880]
9. Down TA, Rakyan VK, Turner DJ, et al. A bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat. Biotechnol 2008;26(7):779–785. [PubMed: 18612301]
10. Rakyan VK, Down TA, Thorne NP, et al. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (TDMRs). Genome Res 2008;18 (9):1518–1529. [PubMed: 18577705]
11. Bailey VJ, Easwaran H, Zhang Y, et al. MS-qFRET: a quantum dot-based method for analysis of DNA methylation. Genome Res 2009;19(8):1455–1461. [PubMed: 19443857]
12. Korshunova Y, Maloney RK, Lakey N, et al. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. Genome Res 2008;18(1):19–29. [PubMed: 18032725]
13. Ball MP, Li JB, Gao Y, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat. Biotechnol 2009;27(4):361–368. [PubMed: 19329998]
14. Deng J, Shoemaker R, Xie B, et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat. Biotechnol 2009;27(4):353–360. [PubMed: 19330000]
15. Rusk N. Capturing the human methylome. Nat. Methods 2009;6:320–321.
16. Cokus SJ, Feng S, Zhang X, et al. Shotgun bisulphite sequencing of the arabidopsis genome reveals DNA methylation patterning. Nature 2008;452(7184):215–219. [PubMed: 18278030]
17. Lister R, O'Malley RC, Tonti-Filippini J, et al. Highly integrated single-base resolution maps of the epigenome in arabidopsis. Cell 2008;133(3):523–536. [PubMed: 18423832]
18. Meissner A, Mikkelsen TS, Gu H, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 2008;454(7205):766–770. [PubMed: 18600261]
19■. Brunner AL, Johnson DS, Kim SW, et al. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res 2009;19 (6):1044–1056. High-quality mapping of DNA methylation. [PubMed: 19273619]
20. Collas P, Dahl JA. Chop it, chip it, check it: the current status of chromatin immunoprecipitation. Front. Biosci 2008;13:929–943. [PubMed: 17981601]
21. Aparicio O, Geisberg JV, Sekinger E, Yang A, Moqtaderi Z, Struhl K. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences *in vivo*. Curr. Protoc. Mol. Biol. 2005 Chapter 21, Unit 21.3.
22. Schmidt D, Wilson MD, Spyrou C, Brown GD, Hadfield J, Odom DT. Chip-seq: using high-throughput sequencing to discover protein–DNA interactions. Methods 2009;48(3):240–248. [PubMed: 19275939]
23■. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129(4):823–837. High-quality mapping of histone methylations. [PubMed: 17512414]
24. Robertson G, Hirst M, Bainbridge M, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods 2007;4(8):651–657. [PubMed: 17558387]
25. Kim TH, Abdullaev ZK, Smith AD, et al. Analysis of the vertebrate insulator protein ctcf-binding sites in the human genome. Cell 2007;128(6):1231–1245. [PubMed: 17382889]
26. Valouev A, Johnson DS, Sundquist A, et al. Genome-wide analysis of transcription factor binding sites based on chip-seq data. Nat. Methods 2008;5(9):829–834. [PubMed: 19160518]
27■. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. Science 2007;316(5830):1497–1502. High-quality mapping of transcription factor binding for neuron-restrictive silencer factor. [PubMed: 17540862]
28. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of chip-seq (MACS). Genome Biol 2008;9 (9):R137. [PubMed: 18798982]
29. Shah A. Chromatin immunoprecipitation sequencing (chip-seq) on the solid(tm) system. Nat. Methods 2009;6(4):i–iii.

30. Narlikar L, Gordan R, Hartemink AJ. A nucleosome-guided map of transcription factor binding sites in yeast. PLoS Comput. Biol 2007;3(11):E215. [PubMed: 17997593]

31. Pavesi G, Mereghetti P, Mauri G, Pesole G. Weeder web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res 2004;32(Web Server issue):W199–W203. [PubMed: 15215380]

32. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol 1994;2:28–36. [PubMed: 7584402]

33. Boulard M, Bouvet P, Kundu TK, Dimitrov S. Histone variant nucleosomes: Structure, function and implication in disease. Subcell. Biochem 2007;41:71–89. [PubMed: 17484124]

34. Henikoff S, Furuyama T, Ahmad K. Histone variants, nucleosome assembly and epigenetic inheritance. Trends Genet 2004;20(7):320–326. [PubMed: 15219397]

35. Kouzarides T. Chromatin modifications and their function. Cell 2007;128(4):693–705. [PubMed: 17320507]

36. Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE. Genome-wide analysis of mono-, di- and trimethylation of histone h3 lysine 4 in arabidopsis thaliana. Genome Biol 2009;10(6):R62. [PubMed: 19508735]

37. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell 2007;128(4):669–681. [PubMed: 17320505]

38. Wang Z, Zang C, Rosenfeld JA, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet 2008;40(7):897–903. [PubMed: 18552846]

39. Mikkelsen TS, Ku M, Jaffe DB, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 2007;448(7153):553–560. [PubMed: 17603471]

40. Yu L, Morse RH. Chromatin opening and transactivator potentiation by rap1 in *Saccharomyces cerevisiae*. Mol. Cell. Biol 1999;19(8):5279–5288. [PubMed: 10409719]

41. Varga-Weisz P. ATP-dependent chromatin remodeling factors: Nucleosome shufflers with many missions. Oncogene 2001;20(24):3076–3085. [PubMed: 11420723]

42. Tsukiyama T, Wu C. Purification and properties of an ATP-dependent nucleosome remodeling factor. Cell 1995;83(6):1011–1020. [PubMed: 8521501]

43. Sudarsanam P, Winston F. The SWI/SNF family nucleosome-remodeling complexes and transcriptional control. Trends Genet 2000;16(8):345–351. [PubMed: 10904263]

44. Reinke H, Horz W. Histones are first hyperacetylated and then lose contact with the activated pho5 promoter. Mol. Cell 2003;11(6):1599–1607. [PubMed: 12820972]

45. Dion MF, Kaplan T, Kim M, Buratowski S, Friedman N, Rando OJ. Dynamics of replication-independent histone turnover in budding yeast. Science 2007;315(5817):1405–1408. [PubMed: 17347438]

46. Segal E, Fondufe-Mittendorf Y, Chen L, et al. A genomic code for nucleosome positioning. Nature 2006;442(7104):772–778. [PubMed: 16862119]

47■. Kaplan N, Moore IK, Fondufe-Mittendorf Y, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 2009;458(7236):362–366. High-quality prediction of nucleosome positions and mapping of *in vivo* and *in vitro* positions. [PubMed: 19092803]

48. Mavrich TN, Ioshikhes IP, Venters BJ, et al. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res 2008;18(7):1073–1083. [PubMed: 18550805]

49. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: Advances through genomics. Nat. Rev. Genet 2009;10(3):161–172. [PubMed: 19204718]

50. Yuan GC, Liu YJ, Dion MF, et al. Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science 2005;309(5734):626–630. [PubMed: 15961632]

51. Lee W, Tillo D, Bray N, et al. A high-resolution atlas of nucleosome occupancy in yeast. Nat. Genet 2007;39(10):1235–1244. [PubMed: 17873876]

52. Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. PLoS Biol 2008;6(3):E65. [PubMed: 18351804]

53. Valouev A, Ichikawa J, Tonthat T, et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res 2008;18(7):1051–1063. [PubMed: 18477713]

54. Mavrich TN, Jiang C, Ioshikhes IP, et al. Nucleosome organization in the *Drosophila* genome. Nature 2008;453(7193):358–362. [PubMed: 18408708]

55. Schones DE, Cui K, Cuddapah S, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell 2008;132(5):887–898. [PubMed: 18329373]

56. Yassour M, Kaplan T, Jaimovich A, Friedman N. Nucleosome positioning from tiling microarray data. Bioinformatics 2008;24(13):I139–I146. [PubMed: 18586706]

57. Gross DS, Garrard WT. Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem 1988;57:159–197. [PubMed: 3052270]

58■. Boyle AP, Davis S, Shulha HP, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell 2008;132(2):311–322. High-quality open chromatin mapping in $CD4^+$ T cells. [PubMed: 18243105]

59. Sabo PJ, Hawrylycz M, Wallace JC, et al. Discovery of functional noncoding elements by digital analysis of chromatin structure. Proc. Natl Acad. Sci. USA 2004;101(48):16837–16842. [PubMed: 15550541]

60. Sabo PJ, Humbert R, Hawrylycz M, et al. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc. Natl Acad. Sci. USA 2004;101(13):4537–4542. [PubMed: 15070753]

61. Sabo PJ, Kuehn MS, Thurman R, et al. Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. Nat. Methods 2006;3(7):511–518. [PubMed: 16791208]

62. Crawford GE, Davis S, Scacheri PC, et al. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat. Methods 2006;3(7):503–509. [PubMed: 16791207]

63. Crawford GE, Holt IE, Mullikin JC, et al. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc. Natl Acad. Sci. USA 2004;101(4):992–997. [PubMed: 14732688]

64. Crawford GE, Holt IE, Whittle J, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 2006;16(1):123–131. [PubMed: 16344561]

65. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. Genome Res 2007;17(6):877–885. [PubMed: 17179217]

66. Giresi PG, Lieb JD. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (formaldehyde assisted isolation of regulatory elements). Methods 2009;48(3):233–239. [PubMed: 19303047]

67. Hesselberth JR, Chen X, Zhang Z, et al. Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. Nat. Methods 2009;6(4):283–289. [PubMed: 19305407]

68. Kang SH, Vieira K, Bungert J. Combining chromatin immunoprecipitation and DNA footprinting: A novel method to analyze protein-DNA interactions *in vivo*. Nucleic Acids Res 2002;30(10):E44. [PubMed: 12000849]

69. Lanctot C, Cheutin T, Cremer M, Cavalli G, Cremer T. Dynamic genomic architecture in the nuclear space: regulation of gene expression in three dimensions. Nat. Rev. Genet 2007;8(2):104–115. [PubMed: 17230197]

70. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science 2002;295 (5558):1306–1311. [PubMed: 11847345]

71. Zhao Z, Tavoosidana G, Sjolinder M, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet 2006;38(11):1341–1347. [PubMed: 17033624]

72. Simonis M, Klous P, Splinter E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet 2006;38(11):1348–1354. [PubMed: 17033623]

73■. Dostie J, Richmond TA, Arnaout RA, et al. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 2006;16(10):1299–1309. 3D conformation for a large genomic region. [PubMed: 16954542]

74. Dekker J. Gene regulation in the third dimension. Science 2008;319(5871):1793–1794. [PubMed: 18369139]

Websites

101. NIH Roadmap for Medical Research. http://nihroadmap.nih.gov/epigenomics/

102. ENCODE Data Coordination Center at UCSC. http://genome.ucsc.edu/ENCODE/

103. modENCODE. www.modencode.org/

104. NCBI Gene Expression Omnibus. www.ncbi.nlm.nih.gov/geo/

105. NCBI Short Read Archive. www.ncbi.nlm.nih.gov/Traces/sra/

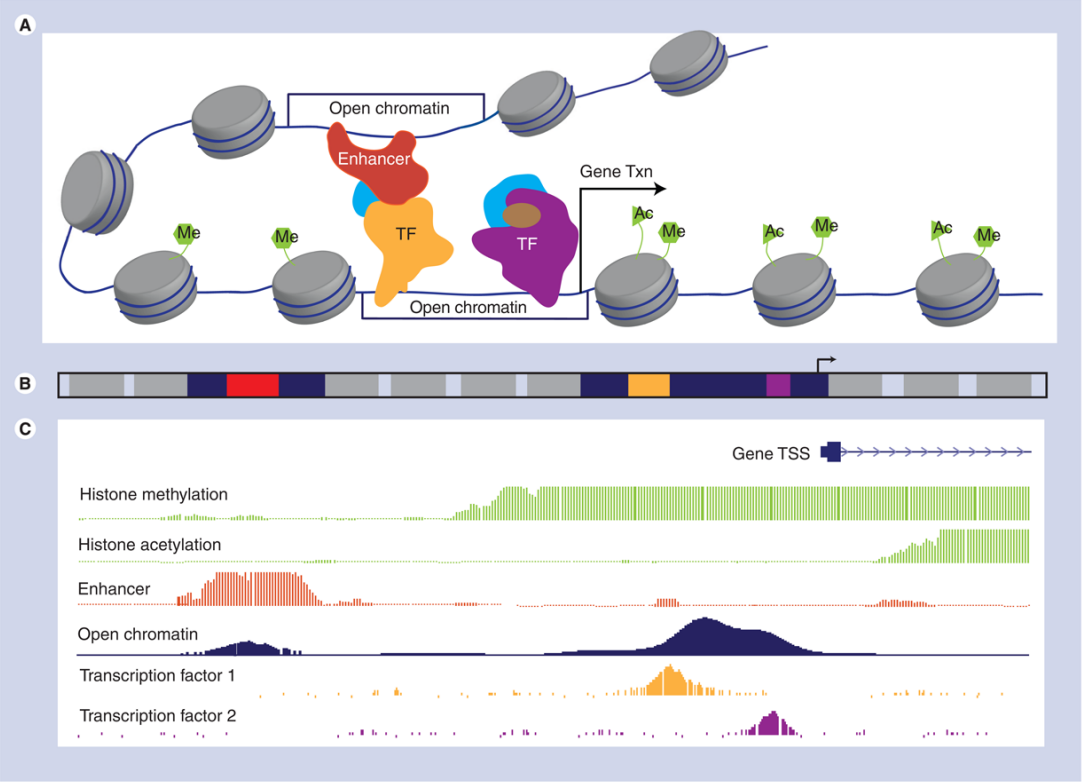

Figure 1. Demonstration of the resolution of data given an illustrated genomic locus

(A) Cartoon of the promoter region of a gene being transcribed from left to right with two upstream transcription factors, one of which is interacting with a distal enhancer. Me and Ac represent a methylation or acetylation of the histone tails. **(B)** To view these data, the DNA is depicted in a linear state, and here it is annotated with colors matching the above features. **(C)** A possible depiction of the view of these data as tracks from the University of California Santa Cruz (UCSC, CA, USA) Genome Browser. Each track demonstrates how these data may look given the region depicted in the cartoon. There is currently no better visualization of chromatin looping data, so the enhancer interaction would not be visible. Ac: Acetylation; Me: Methylation; TF: Transcription factor; TSS: Transcription start site; Txn: Transcription.