

# Mapping of genetic and epigenetic regulatory networks using microarrays

Bas van Steensel

**The highly coordinated expression of thousands of genes in an organism is regulated by the concerted action of hundreds of transcription factors and chromatin proteins, as well as by epigenetic mechanisms. Understanding the architecture of these vastly complex regulatory networks is one of the main challenges in the postgenomic era. New microarray-based techniques have become available for the genome-wide mapping of *in vivo* protein-DNA interactions and epigenetic marks. Data sets obtained with these techniques begin to offer the first comprehensive views of genetic and epigenetic regulatory networks.**

One of the most amazing phenomena in biology is the precisely coordinated expression of thousands of genes in an organism: every gene must be expressed at the right time, in the right place. Not surprisingly, all forms of life devote a considerable part of their protein repertoire to gene regulation. In eukaryotes, ~3–5% of all genes encode transcription factors, which are proteins that bind to specific regulatory DNA sequences and direct the activation or repression of nearby genes.

In addition to transcription factors, the packing of DNA into chromatin has a key role in gene regulation. The basic building block of chromatin is the nucleosome, which consists of a histone octamer with 146 bp of DNA wrapped around it in two turns. Nucleosomes can block or enhance the access of transcription factors to DNA and thereby modulate gene expression. Several protein complexes are able to modify the positioning of nucleosomes along the DNA or the folding of nucleosomal arrays into higher-order chromatin structures. Many transcription factors are in turn able to recruit some of these proteins and thereby indirectly control chromatin structure. Thus, an intricate network exists of interactions between transcription factors, nucleosomes and proteins that control nucleosomal organization, and this network is largely responsible for the gene expression program in any given cell (Fig. 1).

In most cases, gene expression programs need to be maintained during cell division and propagated to the daughter cells. For this purpose, epigenetic ‘memory’ mechanisms have evolved. One such mechanism involves methylation of the C5 position of cytosine (<sup>5m</sup>C) in CpG dinucleotides in DNA. <sup>5m</sup>C provides a chemically stable mark and serves as a potent signal for gene silencing in many eukaryotes. Upon replication,

the DNA methyltransferase Dnmt1 copies each methylation mark of the parental DNA strand to the newly synthesized DNA, ensuring faithful transmission of the genomic <sup>5m</sup>C pattern to both daughter cells. Thus, <sup>5m</sup>C has an important role in mitotically heritable silencing<sup>1</sup>.

Modification of histones could provide an analogous system for the epigenetic inheritance of information. Histones are subject to extensive post-translational modifications, such as acetylation, methylation, phosphorylation and other covalent marks<sup>2,3</sup>. Many of these modifications are linked to either activation or repression of transcription. As the DNA-replication machinery passes, nucleosomes on the parental DNA are randomly distributed to the two daughter DNA molecules, and the remaining gaps are filled in with new nucleosomes. Thus, the genomic pattern of histone modifications may be inherited partially by the daughter cells, and this may facilitate the re-establishment of a particular gene-expression program after cell division<sup>2</sup>.

There are probably other epigenetic mechanisms that do not make use of covalent marks on DNA or nucleosomes. Feedback loops involving freely diffusing factors can also create stable, heritable states. For example, if a transcription factor activates its own gene, then a brief stimulus to trigger its expression will lead to a stable presence of the transcription factor even after the stimulus has subsided. This active state is heritable, because the transcription factor molecules are distributed equally to the daughter cells during mitosis. The DNA-binding of transcription factors is highly dynamic<sup>4</sup>; hence, there is no stable mark on the DNA in this model. Therefore, because of their potential epigenetic roles, feedback loops in regulatory networks are of particular interest.

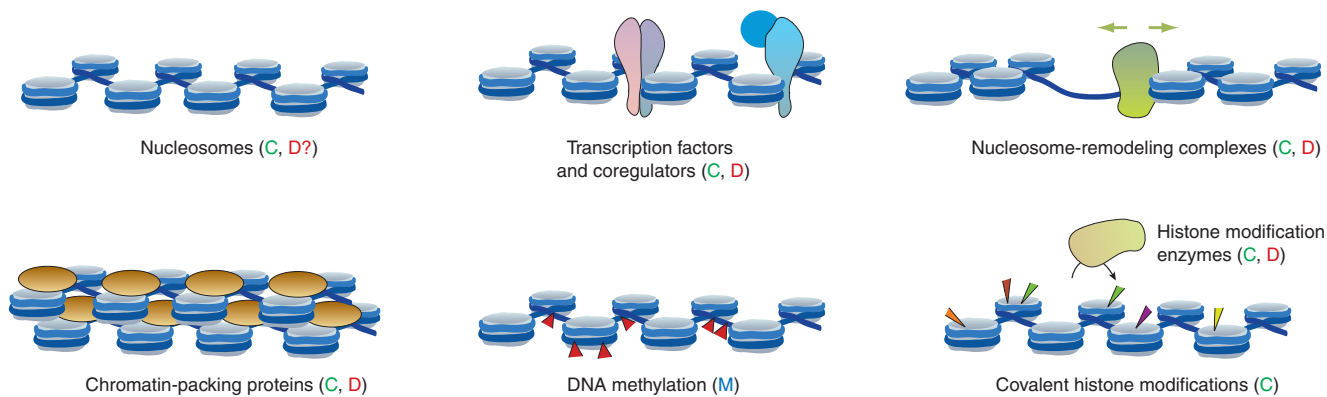
Transcription factors, nucleosomes, chromatin-modifying proteins and epigenetic marks together form extremely complex regulatory systems. Although detailed studies of individual genes have identified many of the components and basic principles that control transcription, we still lack understanding of the global architecture of genetic and epigenetic networks. During the past five years, new microarray-based approaches have been developed that allow us to obtain broader views of gene regulation. In particular, methods have been developed for the genomic mapping of *in vivo* binding sites of regulatory proteins and the distributions of histone modifications and DNA methylation. Here, I will outline these powerful methods and some of the exciting insights that they have provided into genetic and epigenetic regulatory networks.

## Genomic mapping of protein location: ChIP-chip and DamID

Two microarray-based methods for genome-wide mapping of *in vivo* protein-genome interactions have contributed considerably to our

Bas van Steensel is at the Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands. e-mail: b.v.steensel@nki.nl

Published online 26 May 2005; doi:10.1038/ng1559



**Figure 1** The various types of regulatory factors and covalent modifications that can be mapped using microarray-based methods (C, ChIP-chip; D, DamID; M, various methods for methylation mapping).

current knowledge of gene regulatory networks (Fig. 2). A commonly applied method is chromatin immunoprecipitation (ChIP) combined with microarray detection<sup>5–7</sup>. In this technique, cells are treated with a cross-linking reagent (typically formaldehyde), which is thought to covalently link protein complexes *in situ* to DNA. The cross-linked chromatin is then isolated and fragmented, and immunoprecipitation is used to purify the protein of interest together with the attached DNA fragments. To identify these DNA fragments, the cross-links are reversed, and the DNA fragments are labeled with a fluorescent dye and hybridized to microarrays with probes corresponding to genomic regions of interest (ChIP-chip).

DamID is an alternative method based on an entirely different principle<sup>8,9</sup>. Here, a transcription factor or chromatin-binding protein of interest is fused to DNA adenine methyltransferase (Dam). When this fusion protein is expressed *in vivo*, Dam will be targeted to the native binding sites of its fusion partner, resulting in local methylation of adenines in DNA in the immediate vicinity of the binding sites<sup>8</sup>. To identify these sites, the methylated regions are purified or selectively amplified from genomic DNA, fluorescently labeled and hybridized to a microarray. Because adenine methylation does not occur endogenously in the DNA of most eukaryotes, the sites of targeted methylation can be deduced from the microarray signals<sup>9</sup>.

ChIP-chip and DamID have distinct advantages. ChIP-chip requires a highly specific antibody against the protein of interest, whereas DamID does not. On the other hand, DamID is not suitable for detection of post-translational modifications, such as histone modifications (Fig. 1). Discussions of the theoretical merits and drawbacks of the two methods can be found elsewhere<sup>10,11</sup>. Practical side-by-side comparisons have so far been limited, but studies of the regulatory protein GAGA factor indicate that the two methods can yield very similar results<sup>12</sup> (C. Moorman, L. Sun, K.P. White & B.v.S, unpublished data). Although more rigorous cross-validation is needed, ChIP-chip and DamID are both very powerful tools to identify the genome-wide distribution patterns of regulatory proteins.

### Methods for genomic mapping of DNA methylation

Several microarray-based methods have been developed to map <sup>5mC</sup> patterns in genomes<sup>10</sup>. One set of methods uses methylation-sensitive restriction endonucleases. For example, the enzyme McrBC cuts methylated, but not unmethylated, DNA. Size-fractionation of genomic DNA digested with this enzyme will lead to under-representation of methylated DNA sequences in the high-molecular-weight

fraction. Microarray hybridizations can then be used to identify sequences that were methylated<sup>13</sup>. Many variations of this approach have been reported<sup>14–17</sup>.

An entirely different method uses treatment of genomic DNA with sodium bisulfite, which converts cytosine, but not <sup>5mC</sup>, to uracil. In a subsequent PCR reaction, the DNA polymerase reads uracil as thymidine. Thus, C is converted to T, whereas <sup>5mC</sup> remains C. Specially designed oligonucleotide arrays are then used to quantify the bisulfite-induced C-to-T changes at defined genomic positions<sup>18,19</sup>.

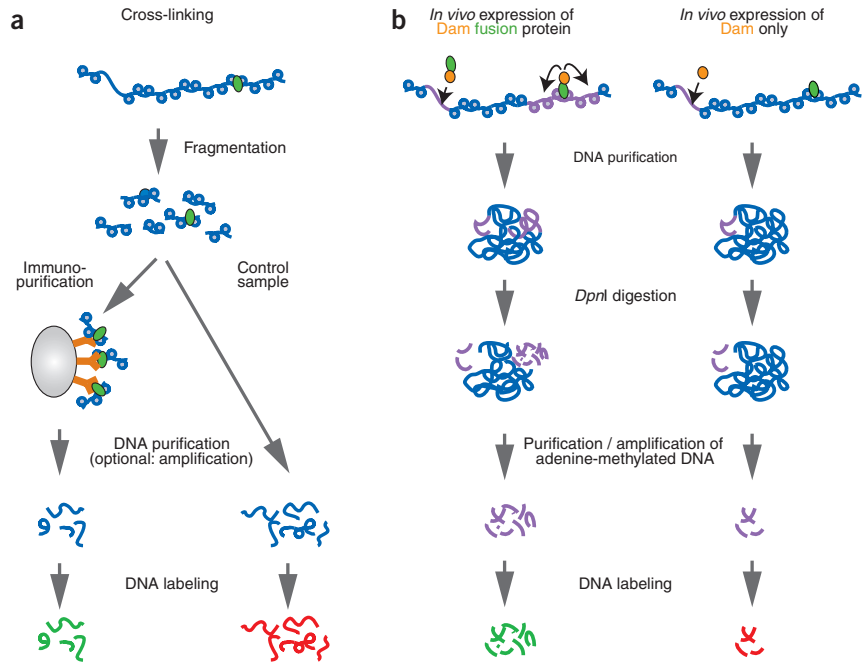
These new microarray-based methods for the mapping of protein-genome interactions and DNA methylation patterns are beginning to revolutionize the research of gene regulatory networks. Below, some of the initial biologic insights obtained with these methods will be discussed.

### The emerging complexity of transcription networks

How many genes are typically under direct control of a transcription factor? To answer this question, true whole-genome maps of transcription factor binding are needed. Such complete maps have been generated in *Saccharomyces cerevisiae*<sup>20–24</sup>. A systematic ChIP-chip survey of 106 different transcription factors<sup>22</sup> showed that the number of target promoters for these factors ranged from 0 to ~200. These numbers should be taken only as crude estimates because they depend on an arbitrarily chosen statistical significance threshold. Nevertheless, these data suggest that, in yeast, a transcription factor may occupy up to ~3% of all promoters.

Is this density of promoter binding by transcription factors similar in higher eukaryotes? Several groups have used promoter arrays to map transcription factor-binding sites in murine and human cells. The cell cycle regulator E2F4 was found to bind to 2–3% of ~13,000 tested promoters in quiescent cells<sup>25–27</sup>, and the regulatory factors HNF1 $\alpha$  and HNF6 could be detected on 0.8–1.6% of the promoters in human liver and pancreas islets<sup>28</sup>. Other mammalian transcription factors seem to associate with a much larger proportion of promoters. This has been most extensively documented in case of the Myc-Max-Mad family of transcription factors, which have key roles in the control of cellular growth. ChIP-chip analysis of human cells<sup>29,30</sup> showed that c-Myc binds to ~10–15% of the probed promoters, as does its dimerization partner Max<sup>29</sup>. This is in agreement with DamID experiments in *Drosophila melanogaster* cells, which showed that Myc, Mad and Max have binding sites near ~15% of all genes<sup>31</sup>. Another transcription factor with ample promoter binding

**Figure 2** Principles of two microarray-based methods for mapping protein–genome interactions: (a) ChIP–chip and (b) DamID. In both protocols, two-color hybridizations are typically done. For ChIP–chip, DNA from total (nonimmunoprecipitated) chromatin is generally used as reference sample. For DamID, the reference sample is obtained from cells expressing Dam alone and serves to correct for untargeted Dam activity<sup>8,9</sup> (adapted from ref. 10).



is HNF4 $\alpha$ , which was detected on 11–12% of all tested promoters<sup>28</sup>. Although only a few transcription factors have been mapped in higher eukaryotes, it seems that some transcription factors interact with large sets of promoters.

The estimates of genomic binding sites based on promoter arrays probably represent only the tip of the iceberg. Intergenic regions and introns in higher eukaryotes form a large part of the genome and create enormous protein-binding potential. Initial DamID mapping of GAGA factor in *D. melanogaster* showed abundant binding to introns and regions 3' of genes<sup>12,32</sup>. A more recent ChIP–chip study with high-density tiling arrays that covered all nonrepetitive regions of human chromosomes 21 and 22 detected large numbers of transcription factor–binding sites not only in 5' regulatory regions, but also in and downstream of genes<sup>33</sup>. Extrapolation of these whole-chromosome data leads to stunning estimates of the total number of transcription factor–binding sites in the human genome: ~12,000 binding sites for SP1 and ~25,000 for c-Myc. Similar experiments with chromosome 22 tiling arrays suggest that there are ~19,000 loci in the human genome bound by CREB<sup>34</sup> and roughly the same number by NF- $\kappa$ B<sup>35</sup>. These data suggest that metazoan regulatory networks are much more complex than their yeast counterpart.

**Functions of transcription factor–binding sites**

Why do some transcription factors bind to so many locations? Are all these binding sites involved in gene regulation? This has been addressed in the case of NF- $\kappa$ B, taking advantage of the fact that its activity can be stimulated by the cytokine TNF- $\alpha$ . Nearly half of the genes with nearby NF- $\kappa$ B–binding sites showed no change in gene expression after treatment with TNF- $\alpha$  treatment<sup>35</sup>. Along the same lines, 85% of the mapped CREB-binding sites are located near genes that show no response to forskolin, a known activator of CREB<sup>34</sup>. These observations suggest that many transcription factor–binding sites may have no regulatory role under a given condition, although subtle effects cannot be ruled out.

Perhaps these 'inactive' transcription factor sites are conditional *cis*-acting elements whose regulatory activity depends on the presence or absence of other factors. It is also conceivable that many transcription factor–binding sites do not control specific genes but rather serve as storage sites that buffer the free pool of transcription factors. Furthermore, it has been suggested that a substantial fraction of transcription factor–binding sites is involved in the regulation of noncoding transcripts<sup>33</sup>. Finally, there may be fortuitous binding sites with no function at all. The elucidation of the functional roles of each transcription factor–binding site will pose a challenge in the future.

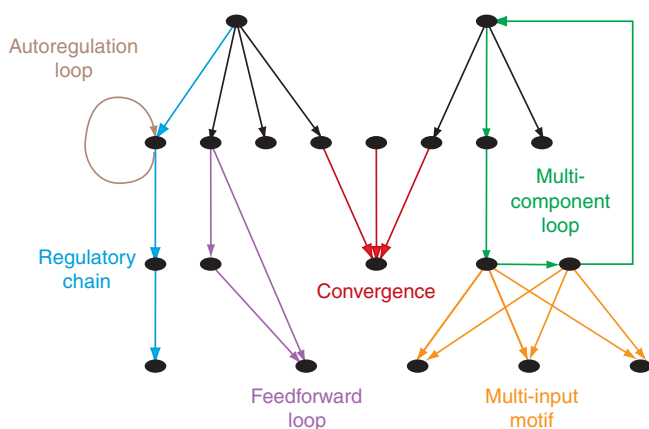
**Plasticity of transcription factor binding**

A few genome-wide studies have shown that the binding patterns of some transcription factors can depend considerably on cell type, stage of the cell cycle or environmental conditions. The yeast factor Ste12, which regulates mating and filamentous growth, binds to different sets of target genes depending on growth conditions<sup>36</sup>. This differential behavior is controlled by the transcription factor Tec1 as well as by the MAP kinases Fus3 and Kss1. A more extensive study in yeast documented the binding patterns of a large set of transcription factors under 12 different culture conditions<sup>24</sup> and showed that most transcription factors had altered binding patterns under certain conditions.

Similarly, the binding patterns of transcription factors in human cells can depend on the cellular differentiation state. For example, the myogenic factor MyoD binds to distinct (though overlapping) sets of promoters in myoblasts and myotubes<sup>26</sup>, and the binding patterns of HNF1 $\alpha$ , HNF4 $\alpha$  and HNF6 differ partially between hepatocytes and pancreatic island cells<sup>28</sup>. Furthermore, the genomic binding pattern of pRb changes during the cell cycle<sup>37</sup>. Thus, the binding patterns of many transcription factors are highly dynamic. Most likely, the changes in these patterns are the result of altered interactions with other factors. Obviously, the plasticity of transcription factor–binding patterns adds another layer of complexity to regulatory networks.

**Combinatorial interactions in transcription factor networks**

Genome-wide binding maps can help to unravel the interactions between transcription factors (and other regulatory proteins). For example, the genome-wide binding pattern of the yeast histone deacetylase Hst1 was nearly identical to that of the previously mapped transcription factor Sum1. Experiments with a *Sum1* deletion mutant confirmed that the binding of Hst1 to nearly all its target sites is dependent on Sum1 (ref. 38). Another study addressed the interactions between the *D. melanogaster* repressor protein Hairy and its presumed cofactors Groucho, dCtBP and dSir2. Based on comparisons of DamID maps of all four proteins, the authors conclude that Hairy is generally accompanied by dCtBP and dSir2 but much less frequently by Groucho<sup>39</sup>. These examples illustrate the power of comparative genome-wide mapping of regulatory proteins.



**Figure 3** Graph representation of a hypothetical transcription factor regulatory network. Each black oval represents a transcription factor. If a transcription factor binds to the promoter of a gene encoding another transcription factor, then the two transcription factors are linked by an arrow. In this illustration, various connectivity motifs can be identified, as indicated in different colors.

Other whole-genome mapping studies further emphasize the prominent roles of protein-protein interactions in targeting transcription factors to specific genomic loci. Many *in vivo* target sites of c-Myc, SP1, p53 and NF- $\kappa$ B lack the respective consensus binding sequences of these factors<sup>33,35</sup>. This suggests that these transcription factors are often recruited to their target loci by other proteins rather than through direct protein-DNA interactions.

### The organization of transcription factor networks

Despite the uncertainty over whether individual transcription factor-binding sites are functional, several laboratories have begun to use transcription factor-binding maps to construct models of regulatory networks. Some of these models are simply directed graphs that place transcription factor A directly upstream of transcription factor B if A binds to the promoter of the gene encoding B (Fig. 3). Such an analysis was done for the set of nine transcription factors known to regulate the cell cycle in yeast<sup>40</sup>. This showed that these transcription factors together may form a closed regulatory loop: transcription factors active in one phase of the cell cycle bind to genes encoding transcription factors that are active in the next phase, thus forming a serial network that is a cycle in itself<sup>40</sup>. By the same approach, a graph representation of the yeast regulatory network was constructed based on a compendium of binding profiles of more than 100 transcription factors<sup>22</sup>. In this network, many transcription factors are associated with promoters of genes encoding other transcription factors. Initial network graphs have also been constructed for myogenic regulators<sup>26</sup> and members of the HNF family of regulators<sup>28</sup>. Analysis of such graphs may help to identify the basic building blocks of regulatory networks, such as feedback and feedforward loops and other motifs (Fig. 3).

### Genome-wide regulatory roles of DNA methylation

DNA methylation profiling with microarrays has become an important tool for the dissection of the basic mechanisms and functions of DNA methylation. The plant *Arabidopsis thaliana* has become a popular model for this purpose. This plant has methylation of not only CpG but also CNG motifs. Microarray mapping of the changes in methylation after mutation of the CNG methyltransferase CHROMOMETHYLASE3 showed that this protein is primarily involved in the methylation of

transposable elements<sup>15</sup>. Abundant methylation was also found at most transposable elements in a heterochromatic region of the genome of *A. thaliana*, consistent with a role for DNA methylation in the inactivation of such elements<sup>13</sup>. Another methylation-profiling study identified dense CG methylation clusters that were preferentially located in genes. On the basis of a statistical analysis of dinucleotide distributions in the genome of *A. thaliana*, the authors suggested that these methylation clusters might serve to silence cryptic promoters that arise sporadically in transcription units<sup>41</sup>. These mapping studies contribute to our understanding of the control and functions of DNA methylation.

DNA methylation has been implicated in the epigenetic silencing of tumor-suppressor genes in cancer cells<sup>1</sup>. Microarray-based mapping techniques have been adopted to study the links between DNA methylation and cancer in more detail. Some tumors can be classified on the basis of DNA methylation patterns alone<sup>18</sup>. Other mapping studies have provided new insights into the role of epigenetic silencing in cancer<sup>16,42,43</sup>. For example, methylation profiling suggests that loss of estrogen receptor- $\alpha$  (ER $\alpha$ ) in breast cancer leads to the progressive epigenetic silencing of genes that are normally regulated by ER $\alpha$ <sup>43</sup>. Mapping of <sup>5mC</sup> has also been combined with ChIP-chip analysis of proteins that specifically bind to <sup>5mC</sup>CpG sites (MeCP2, MBD1, MBD2 and MBD3) to identify new targets of epigenetic inactivation in human cancer<sup>44</sup>. These genome-wide studies help to elucidate the roles of DNA methylation in cancer.

### Genome-wide studies of histone modifications

Histone modifications have important roles in transcription regulation<sup>2,3</sup>. Some modifications, such as acetylation of lysine residues on histones H3 and H4, have been primarily associated with gene activation. Others, such as methylation of Lys9 and Lys27 on histone H3, seem to be linked to gene repression. The multitude of possible modifications on histones could form a complex combinatorial 'histone code' that would direct the activation or repression of genes<sup>2</sup>. In such a combinatorial code, the effect of each modification could be dependent on the presence or absence of other modifications. The numerous possible combinations could render such a code difficult to decipher. But recent genome-wide mapping studies suggest that at least some aspects of the histone code are rather simple.

Several histone modifications previously linked to gene activation (di- and trimethylation at Lys4 of H3, dimethylation at Lys79 of H3, and acetylation of H3 and H4) were systematically mapped in ~6,000 coding regions of a *D. melanogaster* cell line<sup>45</sup>. This study identified strong correlations between the distributions of all these modifications. Moreover, these modifications each had a strong correlation with gene expression levels. This suggests a rather simple histone code in transcribed regions: several 'active' modifications may all be linked to one another and to transcription levels. The parallel action of several modifications could ensure robustness of signaling or promote switch-like behavior<sup>46</sup>.

Similar strong correlations, in both intergenic and coding regions, were observed in a systematic ChIP-chip study of 11 different histone acetylation marks in budding yeast<sup>47</sup>. After using a data-normalization procedure that highlighted subtle local differences between the levels of the 11 marks, however, the authors identified several groups of biologically related genes that shared these subtle differences. In some cases, genes with related functions may therefore share regulatory fine-tuning mechanisms involving specific histone modifications. Further experimental validation of this model is needed.

Genome-wide mapping of the binding sites of histone acetyltransferases (HATs) and histone deacetylases (HDACs) complements these studies and suggests that histone acetylation patterns can be controlled both globally and on a gene-specific basis. Promoter binding of the



**Figure 4** Distribution of dimethylation at Lys4 of H3 (H3K9me2) in the human genome. Most parts of the genome show focal sites of dimethylation at Lys4 of H3, as illustrated by a region of chromosome 22 (**a**). Location of dimethylation at Lys4 of H3 over extended regions was only observed in the Hox gene regions, such as the HoxA cluster (**b**). The regions shown in **a** and **b** are both 140 kb in length (data from ref. 51).

yeast HATs Gcn5 and Esa1 was tightly correlated with gene expression levels, suggesting that virtually all active promoters recruit these HATs<sup>38</sup>. In contrast, several HDACs show a preference for distinct groups of genes<sup>38,48–50</sup>. Mapping of histone acetylation patterns after deletion of the HDACs Rpd3, Hda1, Sir2, Hos1-Hos3 and Hos2 also indicates that each of these proteins controls the acetylation levels of specific sets of genes<sup>48</sup>. Taken together, these global mapping data show that, in *S. cerevisiae*, most HDACs have a preference for distinct gene classes, whereas HATs may bind to all active promoters.

Detailed genomic maps of histone modifications in mammalian cells are being constructed by several laboratories. Recently, Bernstein *et al.*<sup>51</sup> mapped di- and trimethylation at Lys4 of H3 and acetylation of H3 using large arrays covering all nonrepetitive parts of human chromosomes 21 and 22 and some additional regions. Most sites of trimethylation at Lys4 of H3 coincided with transcription starts, whereas regions with strong dimethylation at Lys4 of H3 were mostly located close to genes. Acetylation of H3 generally coincided with di- or trimethylation at Lys4 of H3, indicating that these modifications are linked, as observed in *D. melanogaster*<sup>45</sup>.

Histone modifications on one or a few adjacent nucleosomes would make poor epigenetic marks owing to the stochastic mitotic inheritance of individual nucleosomes by daughter cells<sup>52</sup>. Di- and trimethylation at Lys4 of H3 and acetylation of H3 in human cells all show predominantly punctate distributions: the median size of the modified regions was 500–700 bp<sup>51</sup>. This is close to the expected mapping resolution of ChIP-chip, suggesting that only between one and three nucleosomes carry the modifications at each site. An exception to this was the Hox gene clusters, for which long contiguous stretches (up to 60 kb) of dimethylation at Lys4 of H3 were observed (**Fig. 4**). Hence, perhaps dimethylation at Lys4 of H3 serves as an epigenetic ‘memory’ mark at only a limited number of loci, such as the Hox gene clusters. At the punctate sites, histone modifications may have nonepigenetic regulatory functions.

### Mapping of nucleosome occupancy

In addition to histone modifications, the positioning and spacing of nucleosomes can be important for gene regulation because nucleosomes may modulate the binding of transcription factors as well as the passage of transcribing RNA polymerases. Two groups carried out global mapping of nucleosome occupancy in yeast by ChIP-chip using an antibody against histone H3 or epitope-tagged H2B or H4 (refs.

53,54). These experiments showed that promoters and coding regions of transcribed genes generally had fewer (more widely spaced) nucleosomes. In human cells, however, the nucleosome distribution along the DNA seems to be relatively homogeneous<sup>51</sup>.

### Genomic maps of heterochromatin complexes

Various proteins interact with nucleosomes and are able to modify chromatin structure. Heterochromatin is a specialized type of chromatin, traditionally thought of as a compacted structure that represses transcription. A ChIP-chip study in *A. thaliana* compared the distributions of several putative epigenetic marks in a 1.5-Mb heterochromatic chromosomal region<sup>13</sup>. The results indicate that trimethylation at Lys9 of H3 (thought to be a repressive mark) is predominantly associated with transposable elements and related repeats in this region. The distribution of <sup>5mC</sup> closely followed that of trimethylation at Lys9 of H3, consistent with reported functional links between these two marks<sup>55</sup>. In contrast, dimethylation at Lys4 of H3 is primarily located in gene islands between transposon-dense regions, in agreement with its role as an active mark<sup>13</sup>.

Another study used DamID in *D. melanogaster* cells to generate genome-wide binding maps of HP1 and Su(var)3-9, two proteins implicated in the formation of heterochromatin<sup>56</sup>. HP1 binds specifically to methylated Lys9 of H3, whereas Su(var)3-9 is a methyltransferase for Lys9 of H3 that is able to generate a binding site for HP1. As expected, comparison of the genomic binding maps showed that these proteins often bind together, primarily at transposable elements and in pericentromeric regions. In addition, several genes, mostly located on the chromosome arms, strongly bind Su(var)3-9 but not HP1. Most pericentromeric target genes of HP1 and Su(var)3-9 were actively transcribed, whereas genes bound by Su(var)3-9 alone were mostly silent<sup>56</sup>. These molecular maps show that heterochromatin is not a single entity but rather exists in different variations with distinct compositions and regulatory functions.

Polycomb group (PcG) proteins form specialized complexes that may create heterochromatin-like compacted chromatin structures. Several human target genes of PcG proteins were identified by a combination of ChIP-chip and expression profiling<sup>57</sup>. Some of these genes seem to be activated rather than repressed by PcG proteins. This suggests that, like heterochromatin proteins, PcG proteins have target-dependent regulatory functions. An important future goal will be to understand the molecular determinants of this context-specific regulatory behavior.

## Future challenges

As outlined above, enormous progress has been made with the generation of genomic maps of regulatory protein binding and histone modifications, thanks to technological advances such as ChIP-chip and DamID. In the next few years, these approaches will undoubtedly profit further from new improvements in microarray technology that allow for the rapid and flexible construction of very large arrays<sup>58</sup>. Other high-throughput methods for mapping DNA methylation<sup>59</sup> and protein-genome interactions<sup>60</sup> may complement microarray-based methods.

As mentioned earlier, it will be essential to move from the descriptive (yet important) mapping of genomic distributions to understanding the functions of protein binding and epigenetic marks at each locus. This functional information may be obtained by combining genomic location maps with maps of gene expression changes after selective removal of the proteins or epigenetic marks of interest.

The biggest future challenge, however, will be the unraveling of the combinatorial interactions that control gene regulatory networks. This will require the integration of genome-wide location maps of large numbers of transcription factors, histone modifications, chromatin-modifying proteins, DNA methylation and gene expression. Global identification of noncoding transcripts<sup>33,61,62</sup> and new microarray-based methods for mapping chromatin structure<sup>63,64</sup> may also provide crucial information. The number of parameters that can be mapped is enormous: in mammals, up to 2,000 regulatory proteins, each in hundreds of cell types and tissues. Concerted efforts, for example by the ENCODE consortium<sup>65</sup>, may help to ensure the use of uniform model systems, experimental conditions and data formats.

The modeling of gene regulation networks will require not only large amounts of high-quality data but also new computational approaches. Great progress has been made with the computational analysis of large gene-expression data sets. Methods have been developed to use microarray expression data sets to identify *cis*-acting motifs<sup>66–68</sup> and modules of coexpressed genes<sup>69–71</sup>. A promising Bayesian algorithm can even be trained to predict gene expression patterns crudely on the basis of promoter sequences alone<sup>72</sup>. Integration of microarray mRNA expression data and regulatory protein-binding maps<sup>73,74</sup> will undoubtedly enhance the accuracy of such computational predictions and provide insight into the combinatorial interactions of transcription factors and other regulatory factors.

Further down the line, quantitative rather than qualitative data may become essential. Often, we think of protein-DNA interactions in terms of 'to bind or not to bind'. But virtually all proteins interact with the genome in a highly dynamic fashion, with average residence times on the order of minutes<sup>4</sup> and with occupancy rates that are likely to vary from locus to locus. Subtle quantitative changes in transcription factor binding can have pronounced nonlinear effects on transcription<sup>75</sup>. Therefore, for accurate modeling of gene regulatory networks, quantitative protein location maps may be needed. It is unclear whether such quantitative data can be generated with DamID or ChIP-chip; new techniques might need to be developed for this purpose.

Despite these challenges, research in the field of transcription regulation has entered an exciting new era. Genome-wide mapping approaches, together with computational analyses, are beginning to expose genetic and epigenetic regulatory networks.

## ACKNOWLEDGMENTS

I thank M. Wolkers, F. van Leeuwen, D. Schübeler and members of my laboratory for suggestions. This work was supported by the Human Frontier Science Program, NWO-Genomics, the Dutch Cancer Society and an EURYI Award.

## COMPETING INTERESTS STATEMENT

The author declares competing financial interests (see the *Nature Genetics* website for details).

Published online at <http://www.nature.com/naturegenetics/>

1. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33** Suppl., 245–254 (2003).
2. Jenuwein, T. & Allis, C.D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
3. Peterson, C.L. & Laniel, M.A. Histones and histone modifications. *Curr. Biol.* **14**, R546–R551 (2004).
4. Phair, R.D. *et al.* Global nature of dynamic protein-chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins. *Mol. Cell Biol.* **24**, 6393–6402 (2004).
5. Weinmann, A.S. & Farnham, P.J. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods* **26**, 37–47 (2002).
6. Robyr, D., Kurdistani, S.K. & Grunstein, M. Analysis of genome-wide histone acetylation state and enzyme binding using DNA microarrays. *Methods Enzymol.* **376**, 289–304 (2004).
7. Ren, B. & Dynlacht, B.D. Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Methods Enzymol.* **376**, 304–315 (2004).
8. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat. Biotechnol.* **18**, 424–428 (2000).
9. van Steensel, B., Delrow, J. & Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.* **27**, 304–308 (2001).
10. van Steensel, B. & Henikoff, S. Epigenomic profiling using microarrays. *Biotechniques* **35**, 346–350, 352–354, 356–357 (2003).
11. Lodén, M. & van Steensel, B. Whole-genome views of chromatin structure. *Chrom. Res.* (in the press).
12. Sun, L.V. *et al.* Protein-DNA interaction mapping using genomic tiling path microarrays in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **100**, 9428–9433 (2003).
13. Lippman, Z. *et al.* Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**, 471–476 (2004).
14. Hatada, I. *et al.* A microarray-based method for detecting methylated loci. *J. Hum. Genet.* **47**, 448–451 (2002).
15. Tompa, R. *et al.* Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. *Curr. Biol.* **12**, 65–68 (2002).
16. Yan, P.S. *et al.* Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Res.* **61**, 8375–8380 (2001).
17. Wei, S.H. *et al.* Methylation microarray analysis of late-stage ovarian carcinomas distinguishes progression-free survival in patients and identifies candidate epigenetic markers. *Clin. Cancer Res.* **8**, 2246–2252 (2002).
18. Adorjan, P. *et al.* Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res.* **30**, e21 (2002).
19. Gitan, R.S., Shi, H., Chen, C.M., Yan, P.S. & Huang, T.H. Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. *Genome Res.* **12**, 158–164 (2002).
20. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
21. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.* **28**, 327–334 (2001).
22. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
23. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
24. Harbison, C.T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
25. Cam, H. *et al.* A common set of gene regulatory networks links metabolism and growth inhibition. *Mol. Cell Biol.* **16**, 399–411 (2004).
26. Blais, A. *et al.* An initial blueprint for myogenic differentiation. *Genes Dev.* **19**, 553–569 (2005).
27. Ren, B. *et al.* E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.* **16**, 245–256 (2002).
28. Odom, D.T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).
29. Li, Z. *et al.* A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl. Acad. Sci. USA* **100**, 8164–8169 (2003).
30. Fernandez, P.C. *et al.* Genomic targets of the human c-Myc protein. *Genes Dev.* **17**, 1115–1129 (2003).
31. Orian, A. *et al.* Genomic binding by the *Drosophila* Myc, Max, Mad/Mnt transcription factor network. *Genes Dev.* **17**, 1101–1114 (2003).
32. van Steensel, B., Delrow, J. & Bussemaker, H.J. Genomewide analysis of *Drosophila* GAGA factor target genes reveals context-dependent DNA binding. *Proc. Natl. Acad. Sci. USA* **100**, 2580–2585 (2003).
33. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
34. Euskirchen, G. *et al.* CREB binds to multiple loci on human chromosome 22. *Mol. Cell Biol.* **24**, 3804–3814 (2004).



35. Martone, R. *et al.* Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci. USA* **100**, 12247–12252 (2003).
36. Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395–404 (2003).
37. Wells, J., Yan, P.S., Cechvala, M., Huang, T. & Farnham, P.J. Identification of novel pRb binding sites using CpG microarrays suggests that E2F recruits pRb to specific genomic sites during S phase. *Oncogene* **22**, 1445–1460 (2003).
38. Robert, F. *et al.* Global position and recruitment of HATs and HDACs in the yeast genome. *Mol. Cell* **16**, 199–209 (2004).
39. Bianchi-Frias, D. *et al.* Hairy transcriptional repression targets and cofactor recruitment in *Drosophila*. *PLoS Biol.* **2**, E178 (2004).
40. Simon, I. *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697–708 (2001).
41. Tran, R.K. *et al.* DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. *Curr. Biol.* **15**, 154–159 (2005).
42. Shi, H. *et al.* Triple analysis of the cancer epigenome: an integrated microarray system for assessing gene expression, DNA methylation, and histone acetylation. *Cancer Res.* **63**, 2164–2171 (2003).
43. Leu, Y.W. *et al.* Loss of estrogen receptor signaling triggers epigenetic silencing of downstream targets in breast cancer. *Cancer Res.* **64**, 8184–8192 (2004).
44. Ballestar, E. *et al.* Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. *EMBO J.* **22**, 6335–6345 (2003).
45. Schübeler, D. *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* **18**, 1263–1271 (2004).
46. Schreiber, S.L. & Bernstein, B.E. Signaling network model of chromatin. *Cell* **111**, 771–778 (2002).
47. Kurdistani, S.K., Tavazoie, S. & Grunstein, M. Mapping global histone acetylation patterns to gene expression. *Cell* **117**, 721–733 (2004).
48. Robyr, D. *et al.* Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* **109**, 437–446 (2002).
49. Kurdistani, S.K., Robyr, D., Tavazoie, S. & Grunstein, M. Genome-wide binding map of the histone deacetylase Rpd3 in yeast. *Nat. Genet.* **31**, 248–254 (2002).
50. Humphrey, E.L., Shamji, A.F., Bernstein, B.E. & Schreiber, S.L. Rpd3p relocation mediates a transcriptional response to rapamycin in yeast. *Chem. Biol.* **11**, 295–299 (2004).
51. Bernstein, B.E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181 (2005).
52. Henikoff, S., Furuyama, T. & Ahmad, K. Histone variants, nucleosome assembly and epigenetic inheritance. *Trends Genet.* **20**, 320–326 (2004).
53. Bernstein, B.E., Liu, C.L., Humphrey, E.L., Perlstein, E.O. & Schreiber, S.L. Global nucleosome occupancy in yeast. *Genome Biol.* **5**, R62 (2004).
54. Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D. & Lieb, J.D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.* **36**, 900–905 (2004).
55. Jackson, J.P., Lindroth, A.M., Cao, X. & Jacobsen, S.E. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416**, 556–560 (2002).
56. Greil, F. *et al.* Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location. *Genes Dev.* **17**, 2825–2838 (2003).
57. Kirmizis, A. *et al.* Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev.* **18**, 1592–1605 (2004).
58. Nuwaysir, E.F. *et al.* Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.* **12**, 1749–1755 (2002).
59. Rakyán, V.K. *et al.* DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.* **2**, e405 (2004).
60. Kim, J., Bhinghe, A.A., Morgan, X.C. & Iyer, V.R. Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat. Methods* **2**, 47–53 (2005).
61. Stolc, V. *et al.* A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**, 655–660 (2004).
62. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
63. Gilbert, N. *et al.* Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555–566 (2004).
64. Weil, M.R., Widlak, P., Minna, J.D. & Garner, H.R. Global survey of chromatin accessibility using DNA microarrays. *Genome Res.* **14**, 1374–1381 (2004).
65. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
66. Liu, X.S., Brutlag, D.L. & Liu, J.S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* **20**, 835–839 (2002).
67. Bussemaker, H.J., Li, H. & Siggia, E.D. Regulatory element detection using correlation with expression. *Nat. Genet.* **27**, 167–171 (2001).
68. Roth, F.P., Hughes, J.D., Estep, P.W. & Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939–945 (1998).
69. Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. & Teichmann, S.A. Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* **14**, 283–291 (2004).
70. Chinnaiyan, A. Integrative analysis of the cancer transcriptome. *Nat. Genet.* **37**, 31–37 (2005).
71. Koller, D. From signatures to models: understanding cancer using microarrays. *Nat. Genet.* **37**, 38–45 (2005).
72. Beer, M.A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
73. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **21**, 1337–1342 (2003).
74. Gao, F., Foat, B.C. & Bussemaker, H.J. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* **5**, 31 (2004).
75. Setty, Y., Mayo, A.E., Surette, M.G. & Alon, U. Detailed map of a cis-regulatory input function. *Proc. Natl. Acad. Sci. USA* **100**, 7702–7707 (2003).