Journal of Cellular Biochemistry 88:684–694 (2003)

Chromatin Remodeling as a Guide to Transcriptional Regulatory Networks in Mammals

Fyodor D. Urnov*

Sangamo Biosciences, Inc., Pt. Richmond Tech. Centre, 501 Canal Blvd., Suite A100, Richmond, California 94804

Abstract An important challenge of genome biology is a dissection of transcriptional regulatory networks that operate inside the nucleus during ontogeny and disease (Wyrick and Young [2002] Curr. Opin. Genet. Dev. 12:130). Limitations of existing experimental tools greatly complicate such analysis in the human genome: for example, genome-wide expression profiling of cells responding to a stimulus fails to reveal a majority of the genes involved in the functional network of responding to that stimulus [Giaver et al., 2002; Birrell et al., 2002]. This article discusses recent advances in analyzing mammalian transcriptional regulatory circuits [Nikiforov et al., 2002; Weinmann et al., 2002; Ren et al., 2002]. As evidenced by these and other data, paucity of information about the location of regulatory DNA elements in the human genome presents an obstacle to comprehensive transcription network analysis. It has been known since the late 1970s that chromatin over active regulatory DNA stretches is stably remodeled into ''nuclease hypersensitive sites'' [Elgin, 1988; Gross and Garrard, 1988]. Massively parallel analysis of such remodeling in cell nuclei identifies regulatory DNA that is difficult to map comprehensively using other approaches, reveals genes poised for rapid activation, and offers a novel perspective on the ''epigenome''—the regulatory program being executed by the genome in a given cell type. J. Cell. Biochem. 88: 684–694, 2003.

Key words: chromatin remodeling; genome-wide expression profiling; transcriptional regulation; c-myc; E2F; regulatory DNA; DNAse I hypersensitive site

''. . .MORE THAN MEETS THE MICROARRAY''?

An important challenge of genome biology in the post-sequencing era is to reveal the transcriptional regulatory networks that operate within the nucleus during normal ontogeny and disease [Wyrick and Young, 2002]. Unique biological properties of genomes in higher organisms simultaneously complicate and abet this task. Mammalian gene regulation is a very different phenomenon both quantitatively and qualitatively from such circuits in *E. coli* as regulation of the *lac* operon by glucose and lactose, or of the "lysis vs. lysogeny" decision in phage λ [Ptashne and Gann, 2002]. As this article will argue, the stunning complexity introduced into mammalian genome control by the multitude of chromatin-based regulatory processes [Wolffe, 1998; Wolffe and Hansen, 2001] can, in fact, be used as a discovery tool. The chromatin-based "epigenome"—the functional state imposed onto the genome by its assembly into the nucleus—can be studied experimentally, and such analysis may resolve existing quandaries in mammalian genome research.

Genome-wide expression profiling—massively parallel analysis of cellular mRNA levels via the use of microarrays with gene-specific probes—is a major tool in transcriptional regulatory network analysis [Wyrick and Young, 2002]. Extensive datasets have resulted from the application of this technology to a wide variety of phenomena, including cellular response to environmental stimuli, aberrations of transcription in diseases such as cancer, and consequences on the genome of genetic lesions in chromatin and molecular machines that remodel it. Converting all these data into knowledge—to borrow a phrase from Sydney Brenner—is a major challenge that is yet to be adequately met. For example, using expression

*Correspondence to: Fyodor D. Urnov, Sangamo Biosciences, Inc., Pt. Richmond Tech. Centre, 501 Canal Blvd., Suite A100, Richmond, California 94804.
E-mail: furnov@sangamo.com

Received 26 September 2002; Accepted 27 September 2002
DOI 10.1002/jcb.10397

profiling of tumor cells to gain insight into mechanisms of cancer pathogenesis has proven to be difficult because of an inability to distinguish between genes whose misregulation *caused* the disease from those genes that are misregulated *by* the disease process. Recent data [Birrell et al., 2002; Giaever et al., 2002] indicate that the same predicament hampers interpretation of genome-wide expression data even in budding yeast, an organism with a small, stable genome and a transcriptome of 6,000 genes, i.e., ~5 times less than in *H. sapiens*.

Ron Davis and colleagues developed a system for high-throughput *S. cerevisiae* reverse genetics in which a collection of strains each carrying a deletion of a single gene is assembled [Shoemaker et al., 1996]. To allow rapid subsequent discrimination between strains, each gene is deleted by insertion of a drug resistance marker gene flanked by a unique, identifiable sequence motif (thus, each strain now carries a "molecular bar code" at the deleted gene locus). This arrangement allowed for the following "natural selection in a test tube" experiment [Giaever et al., 2002]: 5,916 individual strains representing deletions in 96.5% of all budding yeast open reading frames were mixed together and grown under defined conditions, for example, medium containing galactose and not glucose, or medium containing 1 M NaCl. At given timepoints, the "fitness" of each strain under these conditions relative to its 5,915 siblings was assayed by isolating bulk genomic DNA from this mixed culture, PCR-amplifying all the "molecular bar codes," and analyzing the distribution of strains in the culture by hybridization of the output of this PCR amplification to a custom microarray containing probes complementary to the 5,916 individual bar codes. Strains that cannot grow on galactose are expected to be lost from the population much more rapidly than those that can, and this is reflected in a decrease of signal from the position on the microarray corresponding to that strain's "bar code."

This analysis yielded two unexpected observations. The first was the discovery of nine new genes the products of which are required for efficient galactose utilization. This was surprising because carbon metabolism by budding yeast—in some part due to the role of this pathway in baking and brewing—is one of the best studied phenomena in all of biology, and the genetic circuitry of *GAL* gene regulation was thought to be known in exquisite detail. In general, budding yeast has the smallest genome of any model system in eukaryotic biology, and genetic analysis tools in this organism are superb—we understand the mechanistic details of genome regulation in budding yeast more than in all the other eukarya combined [Gregory, 2001]. The incompleteness of this knowledge is illustrated by the unexpected identification of a number of new loci involved in a circuit vigorously studied for some 50 years—loci that were discovered *only* after a locus-by-locus saturation mutagenesis screen was performed with each strain assayed for a deficiency in the phenotype of interest!

The second surprising result was the discordance between data from genome-wide expression profiling and information on a given gene's requirement for growth under certain conditions. All budding yeast genes that become activated or repressed upon exposure to a stimulus such as galactose or high salt were identified in previous genome-wide expression profiling experiments, and it was reasonable to assume that genes upregulated by stimulus X are somehow required for the cell to respond to that stimulus. Refuting this assumption, experimental analysis showed that only 0.9% of genes in the budding yeast genome upregulated by high salt treatment were required for growth in that condition. The cognate value for the galactose pathway, while higher, was still only 7% [Giaever et al., 2002]. Simple arithmetic indicates, therefore, that 99.1% of budding yeast genes upregulated by high salt, and 93% of genes induced by galactose, are not required for the cell's response to the cognate stimulus. Confirming the general validity of this surprising finding, in a separate study, the authors analyzed genetic requirements for survival in the presence of DNA damaging agents, and discovered that there was little to no correlation between a gene's *genetic* requirement for survival and its upregulation during DNA damage [Birrell et al., 2002].

These data indicate that even in an organism with a small transcriptome, expression profiling yields many false negatives, i.e., it fails to identify the overwhelming majority of genes required for a given pathway. The authors offer two explanations for these data: (i) such genes may already be expressed in the cell prior to the stimulus; (ii) signaling relevant to the response

to a stimulus may have a major non-transcriptional component. More importantly, expression analysis appears to identify a large number (~90% in the case of the galactose pathway) of false positives, i.e., genes whose upregulation is either functionally irrelevant to the stimulus under study or whose relevance cannot be revealed by deleting that gene and measuring the fitness of the resulting strain.

Many published studies on the use of expression profiling use the words "transcriptional program" in reference to the data obtained [Wyrick and Young, 2002]. From the data presented by Davis and colleagues [Birrell et al., 2002; Giaever et al., 2002], it appears that genome-wide expression profiling in budding yeast illuminates only a small portion of the underlying regulatory *program* and instead reveals the transcriptome's *phenotype*, the relationship of which to the underlying regulatory program remains poorly defined.

In general, the existence in the best-studied yeast regulatory pathway of no fewer than nine previously unknown components, and the failure of expression profiling in budding yeast to illuminate >90% of the key genes required for major genomic pathways, such as change in carbon source or DNA damage, are ominous signs to scholars of mammalian genome regulatory networks. The human genome is ~275 times larger than that of budding yeast and contains ~6 times as many genes, of which at least 2,000 code for transcriptional activators. Furthermore, human genomics lack practically all the high-throughput tools used by yeast geneticists. Faced with this predicament, one option may be turning to recently developed chromatin-based approaches (see below).

A CASE IN POINT: THE "ENDURING ENIGMA" OF c-MYC

The unique challenges associated with analyzing gene control networks in mammals are best appreciated by analysis of a representative example, such as offered by the HLH-bZIP transcription factor, the protooncogene c-myc. A recent review on this protein written by R. Eisenman, one of its leading scholars, began with the following statement: "Although myc was among the very earliest oncogenes identified and the subject of intense study, it has nonetheless proven to be an enduring enigma" [Eisenman, 2001].

The key puzzle is the peculiar gap between our data about the protein itself, and about the genes it regulates. Myc is a transcription factor that belongs to the "immediate early response" gene class. It is induced in quiescent cells that encounter a growth signal, and is critical for normal progression through the cell cycle: cells deficient in myc grow exceedingly slowly, while deregulated expression of myc causes Burkitt's lymphoma (OMIM # 113970), and avian viruses successfully usurp mutated allelic forms of myc for their own oncogenic purposes [Eisenman, 2001]. Myc is capable of functionally engaging in vivo a diverse array of chromatin modifying, remodeling, and other multiprotein complexes that function in transcription [e.g., Frank et al., 2001].

In contrast to these extensive data about mechanisms of myc function in transcriptional control, a search for myc target genes the deregulation of which causes such striking phenotypic effects proved to be a Herculean effort, and bona fide c-myc target genes have been difficult to identify (see Dang [1999]) for a review of this complex issue). The sequence bound by this protein in vitro (CACGTG) is too simple to be by itself a useful predictor of direct myc response. Most genes identified as myc "targets" by differential display or expression profiling failed a simple criterion: their expression was unchanged in cells lacking myc [Bush et al., 1998]. These data offer an interesting parallel to the earlier discussed data from R. Davis and coworkers on the overwhelming lack of concordance between genes that are upregulated by a signal and genes functionally involved in responding to that signal [Birrell et al., 2002; Giaever et al., 2002], and further highlight the challenges of interpreting information obtained from genome-wide expression profiling experiments. Attempts to directly identify myc-bound DNA by cloning the output of a chromatin immunoprecipitation were stymied by the complexity of the output material [Eisenman, 2001].

Intricate experimental arrangements have proven necessary to identify bona fide myc targets. For example, a recent study used a functional complementation strategy [Nikiforov et al., 2002]: a cDNA library depleted for myc per se was screened for its ability to rescue the acute proliferation defect of c-myc null fibroblasts, and the sole clone identified coded for serine hydroxymethyltransferase. This gene

thus joins the very few others, such as the previously identified ornithine decarboxylase, that are known to function in core aspects of cellular metabolism and directly activated by myc. This is an important finding, but fails to account for the striking phenotypic consequences of myc deregulation on cells and the organism: Burkitt's lymphoma is unlikely to be caused by overexpression of ornithine decarboxylase and serine hydroxymethyltransferase! Puzzlingly, genes more immediately related to cell cycle progression, such as cyclin D2, that are thought to be direct myc targets fail to complement the growth defect of c-myc deficient cells [Berns et al., 2000].

Twenty years after its discovery, myc has been shown to be a transcription factor by every biochemical and functional criterion applied to it, and is known to be essential for normal cell cycle progression, and yet barely a rudiment of the genome's response to this protein has been assembled.

A BROADER PUZZLE

The problems associated with uncovering myc targets extend to most other "well-studied" mammalian transcription factors, for example, NF-κB [Ghosh and Karin, 2002], or such members of the NHR superfamily as receptors for estradiol, thyroid hormone, or corticosteroids [Urnov and Wolffe, 2001]. For some of these proteins, a list of direct targets has been assembled that is more extensive than that of myc, but, remarkably, for none of these proteins has an even quasi-comprehensive "genome response map" been identified (Fig. 1). A major disparity exists between the very extensive amount of information on the biochemical partners of nuclear hormone receptors or NF-κB in transcription control [Urnov and Wolffe, 2001; Ghosh and Karin, 2002], and the paucity of data on what these proteins do to the genome once inside the nucleus. This limited state of our knowledge is unfortunate given the clinical

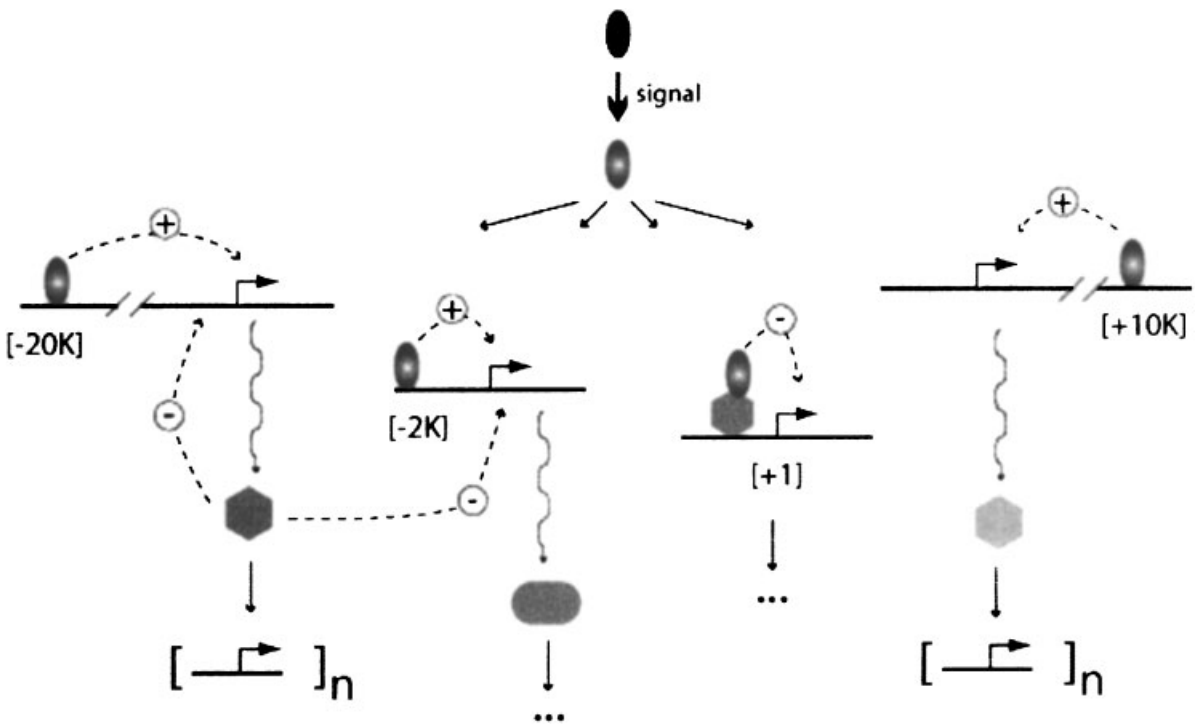

Fig. 1. A simplified representation of one possible architecture of a regulatory network. In this scenario, a signal, such as a signal transduction event or a rise in titer of a small molecule ligand, activates a transcription factor (ellipse). This protein, in turn, activates three genes, by binding to a target in an enhancer, distal promoter, or downstream enhancer, and represses another gene via transrepression (as has been shown for GR). The product of the first gene is a transcription factor—it regulates a number of downstream targets, and represses transcription of two genes directly induced by the initial signal (as has been shown for the ecdysone receptor cascade). The products of the second and third gene are protein that has some regulatory function (i.e., a kinase). The product of the 4th gene is also a transcription factor that regulates a number of downstream targets.

relevance of all these transcriptional regulators. It would be useful, for example, to develop a pathway-selective agonist for the glucocorticoid receptor (GR) that would lack the side effects of existing such agents, but these development efforts are stymied by the rudimentary knowledge we have on the regulatory network of liganded GR action in the nucleus. The same predicament hampers development of selective estrogen receptor modulators.

While intranuclear action by myc, NHRs, and NF-κB is little understood, even less is known about transcription-based cell lineages in mammals where the core regulators are less studied and cannot be induced by simple signals such as addition of a small-molecule ligand to the tissue culture medium. For example, many key transitions in stem cell fate are driven by transcription factors, and in some cases, a simple change in level of the regulator is enough to cause a remarkable change in phenotype or the developmental potential of the cell [Smith, 2001]. It is obvious that such transcription factors as Oct-3/4 or HoxB4 evoke these dramatic effects by engaging and regulating a set of targets in the genome, but at present, our understanding of this process is, at best, exceedingly limited: the targets are unknown, and the functional interplay between them, if any, remains unexplored.

Broadly speaking, the lack of information about transcriptional regulatory networks in mammalian genomes does not stem from a lack of effort, but reflects fundamental technical challenges, of which three are salient.

1. It is nontrivial to make a transition from the amino acid sequence of a transcriptional regulator to a "consensus binding sequence" that it engages. For the majority of the ~2,000 transcription factors the human genome contains, little to no information is available about even a rudimentary "consensus" binding site, and in vitro site-selection SELEX-type experiments—at present, the only meaningful way to determine what sequence a protein can bind in vitro—are labor-intensive. Further complicating the matter is the fact that some well-studied regulators—the thyroid hormone receptor, for example [Zhang and Lazar, 2000]—can tolerate a broad range of variation in their DNA binding site, and bona fide genomic targets for such regulators frequently contain receptor response elements that bear little to no resemblance to consensus sites defined in vitro. This is aptly characterized as a "phenomenon that defies ready conceptualization" [Judelson and Privalsky, 1996], and indicates that it is simply impossible to scan the regulatory DNA elements of a given gene and state that a transcriptional regulator such as the thyroid hormone receptor will or will not bind to it in vivo. Contrasted with such abject failure of sequence analysis is the obvious fact that all such proteins exhibit a very narrow target range in vivo: thus, these proteins are highly selective in vivo, but for reasons we do not understand and have, at present, no way of modeling.
2. Even for proteins that are intolerant of point mutations in their target site in vitro, in vivo experiments firmly indicate that the primary DNA sequence makes only a small and poorly understood contribution to target site selection in vivo [Urnov, 2002]. Such transcriptional regulators as budding yeast Gal4p [Ren et al., 2000], and mammalian GATA-1 [Horak et al., 2002] use a mechanism for selecting in vivo targets that we do not understand—it may include both a positive and negative role for chromatin structure [Urnov, 2002], and for cooperative interactions between different transcription factors to form multiprotein assemblies on the DNA [Thanos and Maniatis, 1995], but it does *not* include a search in the genome for the DNA stretch with the highest density of "consensus" binding sites per unit DNA length. For example, the budding yeast genome contains 1286 perfect matches to the Gal4p consensus binding site, of which only 10 *[sic!]* are bound by the protein in vivo [Ren et al., 2000]. These experimental data firmly indicate that, in interesting contrast to transcriptional regulation in prokarya, the mere presence or absence of a regulatory "module" for a given transcription factor in the promoter or enhancer of a gene is clearly insufficient to enable an in vivo response of that gene to that factor. For this reason, while it is an established practice to use large databases of transcription factor binding sites such as TRANSFAC (http://transfac.gbf.de/TRANSFAC/) to scan regions of interest for "putative binding sites," the near-overwhelming majority of "hits" such analysis yields are false positives [Pennacchio

and Rubin, 2001], as clearly indicated by in vivo analysis of protein binding to DNA. At present, only two experimental techniques are available that can be used to formally prove that protein X is bound to DNA stretch Y in vivo: chromatin immunoprecipitation, and comparison of in vivo DMS footprinting with data from in vitro methylation interference analysis. While both techniques are challenging and labor-intensive, they have the important advantage of providing data, rather than predictions.

3. Last but not the least, a considerable proportion of *cis*-acting regulatory DNA elements in the human genome remain unidentified. The functional behavior of genes in all metazoa studied, from *Drosophila* to humans, emerges from a poorly understood interplay between a large number of short DNA stretches dispersed over many 1000s of bp, including core promoters, distal promoter elements, enhancers, locus control regions, insulators, etc. [Pennacchio and Rubin, 2001]. Information about the core promoter sequence of a gene—the only annotation available for >90% of the human transcriptome—is insufficient to analyze the regulatory circuitry this gene is subject to, as amply evidenced by the very few genes, such as those found in the β-globin locus, where comprehensive analysis of this issue has been performed.

The first problem is a technically very challenging one, and it remains to be seen if recent advances in introducing elements of massively parallel architecture into protein-DNA binding studies [Bulyk et al., 2001; Roulet et al., 2002] can be used to scale such experiments up to include even a small fraction of the ~2,000 transcription factors found in the human genome. This effort, even if successful, will have to contend with the difficult issue of cooperativity between distinct transcription factors in determining target site selection.

The second problem has been solved by developing technologies for genome-wide profiling of in vivo binding sites for transcription factors in budding yeast and, more recently, in human cells. This approach analyses DNA output from a chromatin immunoprecipitation done against a regulator of interest on a microarray containing a panel of DNA probes corresponding to regions of interest. The budding yeast genome is small enough for its intergenic regions—i.e., the entirety of its regulatory DNA component—to be tiled onto a single microarray. Use of such "ChIP on a chip" illuminated elegant regulatory circuits—it was discovered, for example, that upon transition from glucose to galactose in the medium, Gal4p not only activates the galactose transporter, but also represses expression of the gene coding for the glucose transporter [Ren et al., 2000]. Over the past 3 years, "ChIP on a chip" studies of transcription factors such as SBF/MBF, the transcriptional regulator Rap1p, proteins involved in initiation of DNA replication such as ORC and the MCM complex, chromatin modifying and remodeling enzymes such as Esa1p and RSC have illuminated important aspects of the regulatory programs these proteins are involved in [Wyrick and Young, 2002].

An exciting recent development have been pilot studies applying "ChIP on a chip" to transcription factors in the human genome. M. Snyder and colleagues analyzed binding of the transcription factor GATA-1 to a 40 kb stretch of the β-globin locus tiled onto a custom microarray, and discovered that there is "...no correlation between the number of [GATA-1] sites [per 1 kb stretch] and observed enrichment [of that stretch in the output of the chromatin immunoprecipitation]" [Horak et al., 2002]. These data emphasize how little we understand DNA binding in vivo by Zn finger proteins [Urnov, 2002]—the major class of transcription factors in all metazoa.

Two research groups reported "ChIP on a chip" analysis of binding by members of the winged-helix-related E2F transcription factor family. The laboratories of R. Young and B. Dynlacht used a custom microarray that contained core promoter sequences for ~1,200 human genes earlier identified as being cell-cycle regulated, and made the unexpected discovery that E2F family genes are bound to promoters of genes involved in the DNA damage pathway, as well as the G_2 and mitotic checkpoints [Ren et al., 2002]. A related finding was made by P. Farnham, T. Huang and colleagues, who used a microarray with ~8,000 computationally identified CpG islands to connect E2F with genes involved in DNA repair and recombination [Weinmann et al., 2002].

The success of these experiments highlights the urgency of addressing the third problem in mammalian transcriptional network

analysis—lack of information about the location of regulatory DNA elements in the human genome. A back-of-the envelope estimate—one core promoter and, on average, two additional regulatory elements such as an enhancer or an insulator per gene—places the total number of these stretches in the human genome at ~100,000 (this estimate does not include *cis*-acting DNA elements involved in non-transcriptional genomic processes). As of August, 2002, less than 5% of such stretches have been experimentally defined, and published computational predictions exist for a number of others, but await experimental validation. The last section describes the challenges, and possible solutions, to a comprehensive identification of regulatory DNA in the human genome.

THE CHROMATIN EPIGENOME AS A WINDOW ONTO THE ELUSIVE REGULATORY DNA

A familiar parable describes a late-night passerby who sees a man crawling on his knees in a circle of light cast by the sole streetlight on the block—he lost his keys someplace else, he explains, and is looking for them in that particular spot *because that is where the light is*. With some notable exceptions, the majority of computational and experimental effort in human transcription biology focuses on core gene promoters [Werner, 2001]—the few 100 bp surrounding the transcription start site of a gene. The reason for such emphasis is the core promoter is the sole stretch of regulatory DNA that can be identified relatively quickly by aligning the genomic sequence with the full-length cDNA (see, for example, the useful eukaryotic promoter database at http://www.epd.isb-sib.ch/). There is universal agreement that in emphatic contrast to bacteria and budding yeast, genes in all metazoa are regulated by complex interactions of multiple *cis*-regulatory domains—well-studied example include five such elements spread over a 15-kb LCR in the human β-globin locus [Bulger et al., 2002], and the near-astonishing complexity and size of regulatory domains of such *Drosophila* homeotic gene clusters BX-C and ANT-C [Lyko and Paro, 1999].

At present, however, "the light" reveals only the core promoter. A small number of human genes (<5%) have experimentally identified nonpromoter regulatory elements. Efforts are being made to use cross-species conservation of regulatory elements to identify such motifs on a genome-wide scale, but face major obstacles [Pennacchio and Rubin, 2001]. Such analysis yields a number of false-negatives, i.e., experimentally defined regulatory DNA stretches that are not conserved between species even as closely related as human and mouse: examples include some of the SCL enhancers, and regulatory elements in the α-globin locus. Cross-species computational comparison also produces false-positives—conserved stretches the regulatory relevance of which is entirely unclear until verified by experimentation.

It is puzzling that an experimental effort to identify all regulatory DNA in the human genome has not yet been completed– this delay is, perhaps, to some extent due to an existing bias against the "historic, labor-intensive, wet-laboratory methods" [Pennacchio and Rubin, 2001] of identifying such elements. In actual fact, given the established limitations of computational methods [Pennacchio and Rubin, 2001], experimental approaches, while indeed labor-intensive, are the only way to *comprehensively* identify all regulatory DNA in a mammalian genome, and a conceptual framework for such identification has existed since the late 1970s.

The human genome undergoes a ~15,000 compaction upon its assembly into the nucleus. Rather than form a structurally uniform compact array, chromatin—the union of genomic DNA, histone proteins, and a wide variety of associated nonhistone regulators—assumes a wide variety of distinct functional states, each associated with, and caused by, a particular form of chromatin remodeling and modification [Wolffe, 1998; Elgin and Workman, 2000; Jenuwein and Allis, 2001; Narlikar et al., 2002]. The nucleus, therefore, contains information additional to that encoded in the primary DNA sequence of the genome—this extra information, the epigenome, results from the sum total of functional states assumed by each gene, and represents the program being executed by the genome in a given cell type. In the late 1970s, several groups of researchers—W. Scott and D. Wigmore, A. Varshavsky and colleagues, C. Wu and S. Elgin, and S. Nedospasov and G. Georgiev—discovered an important general property of the chromatin-based epigenome: the existence of active regulatory DNA stretches in an unusual chromatin conformation that was termed "nuclease

hypersensitive site" [Elgin, 1988; Gross and Garrard, 1988].

The first demonstration that the promoter of an endogenous chromosomal gene in a eukaryote is found in a "DNase I hypersensitive site" in the context of chromatin was made by C. Wu in 1980, i.e., 22 years ago, and has since been expanded to >95% of all promoters, and other *cis*-regulatory DNA elements that have been examined experimentally [Elgin, 1988; Gross and Garrard, 1988]. It appears that the binding of certain transcription factors to their target sites within the context of unperturbed chromatin is accompanied by the recruitment of large chromatin remodeling and modification complexes, such as the ATPase-containing complex SWI/SNF, or the histone acetyltransferase-containing complex SAGA [Narlikar et al., 2002]. This recruitment leads to a dramatic and localized perturbation of histone-DNA contacts of 1-2 nucleosomes immediately adjacent to the transcription factor binding site, and manifests itself as a marked increase in accessibility of that DNA stretch to nucleases, such as DNase I and restriction enzymes.

A laboratory technician can use simple procedures—treatment of cell nuclei with nucleases, DNA extraction, and Southern blotting—to map all the DNase I hypersensitive sites in any 50 kb stretch of a mammalian genome within 1 week, thereby yielding a comprehensive experimental map of regulatory DNA in that locus. It is important to appreciate a fundamental difference between this technique and all other existing approaches: 24 years of scholarship in transcription biology have yielded the firm observation that a sequence found in a DNase I hypersensitive site must be functioning to regulate some chromosomal process. This approach, therefore, does not generate a prediction that must then be tested experimentally, but rather an experimental *fait accompli*.

Since the problem of experiment-based identification of regulatory DNA was solved in 1978, immediately following the release of the human genome sequence in the summer of 2000 (see "Author's note"), a collection of methods to map nuclease hypersensitive sites in a massively parallel fashion was developed (Fig. 2). Its application to human tissue culture cells yielded the observation that a considerable (>60%) fraction of active regulatory DNA stretches in a given cell type are not found in core gene promoter regions, and that a large number of such elements are not CpG-rich. The feasibility of nuclease-based profiling of the chromatin remodeling-based epigenome in a massively parallel fashion solves a major problem—comprehensive experimental identification of nonpromoter regulatory DNA—but also illuminates several new ones.

Published data on "ChIP on a chip" for GATA-1 [Horak et al., 2002] illustrate how challenging genome-wide mapping of protein binding sites is in the human genome—it remains to be seen if use of chromatin to parse the genomic sequence and reveal bona fide regulatory DNA alleviates this predicament. A very considerable number of nonpromoter regulatory DNA elements in the human genome await analysis by computational and biochemical methods—integration of data from genome-wide expression profiling with massively parallel profiling of the chromatin-remodeling-based epigenome is expected to be of major use in this regard (Fig. 2).

CONCLUSION

The complexity of regulatory DNA in the human genome offers a perspective on what we may expect to be lurking in the Nietzschean "muddy water" [Ptashne and Gann, 2002] of mammalian genome control. Mechanisms of gene regulation in bacteria and such unicellular eukaryotes as budding yeast provide an invaluable conceptual foundation for study of gene control in metazoa. In the words of François Jacob, one of the founding fathers of molecular biology, "...the same principles [of transcriptional control] operating in bacteria are also operating in higher organisms with added complexity" [Ptashne and Gann, 2002]. Two billion years of evolutionary time separate a given *E. coli* from the particular *H. sapiens* whose intestines it lives in—is the "added complexity" a simple addition, like extra blades on a pocket knife? Or is overall evolutionary analogy of phenomenon—such as, for example, between lactose induction of the gene for β-galactosidase and of the malic enzyme gene by thyroid hormone—obscuring a fundamental difference in mechanism?

Precedent exists for the latter scenario. For example, both bacteria and mammalian cells can move towards a source of some chemoattractant, such as a nutrient. On a macroscopic

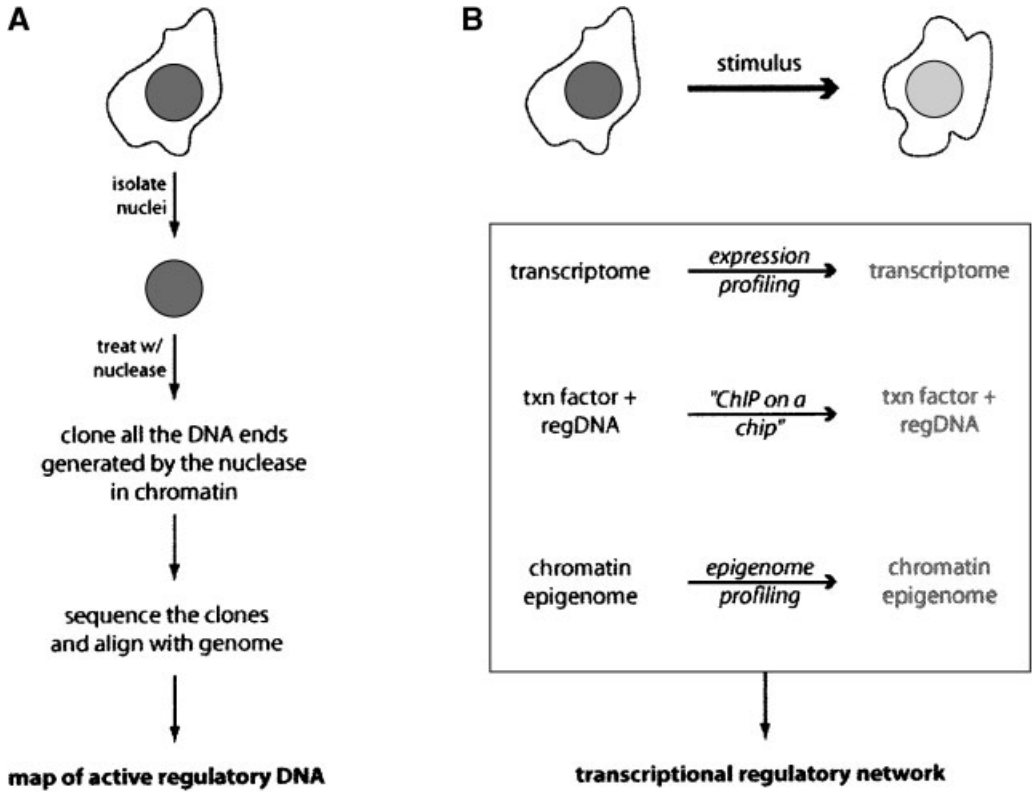

Fig. 2. **A**: Massively parallel identification of active regulatory DNA stretches by mapping nuclease hypersensitive sites in chromatin (F. Urnov and A.P. Wolffe, manuscript in preparation). Nuclei are isolated from cells and treated with a nuclease, such as DNase I or a restriction enzyme. All the ends generated by the nuclease in chromatin are then cloned into a library, and the library is sequenced. The clones isolated identify DNase I hypersensitive sites, i.e., active regulatory DNA. **B**: Chromatin-based mapping of transcriptional regulatory networks during a cell phenotype change driven by a stimulus (e.g., oncogenic transformation of breast epithelium by estrogen). Expression profiling yields the change in the transcriptome, which is difficult to interpret. Regulatory DNA identified as in (A) can be used for a ''ChIP on a chip'' to identify all direct genome targets of stimulus sensor (e.g., estrogen receptor). Use of a microarray for massively parallel profiling of regulatory DNA state changes driven by stimulus—when compared to changes in transcriptome—can identify genes poised for stimulus response, and illuminate the transcriptional regulatory network.

level, this process of positive chemotaxis is identical between prokarya and eukarya. The make up of the underlying mechanism—i.e., the process whereby the cell translates output from a measurement of attractant concentration into physical motion towards the source of that attractant—is fundamentally different between bacteria and mammalian cells. The difference manifests itself as an inevitable biochemical divergence—the proteins used by bacteria and by humans for this process are entirely different. Much more importantly, the larger size and lower mobility of mammalian cells necessitated the evolution of an entirely novel way of translating attractant gradient into motion—briefly, bacteria perform a "biased random walk" towards an attractant, while mammalian cells do not [Weiner, 2002].

Function of the mammalian genome involves a number of processes and phenomena that are not found in genomes of unicellular organisms such as *E. coli* or *S. cerevisiae*. In all taxa studied, gene control involves the interactions of DNA with regulatory proteins—a key function of these regulators is to determine where in the genome transcription will occur, and how frequently it will do so [Ptashne and Gann, 2002]. In eukarya, however, the primary DNA sequence is not the major determinant of where regulatory proteins will bind [Ren et al., 2000; Horak et al., 2002], these regulatory proteins shuttle on and off DNA on the scale of seconds in an ATP-dependent mechanism [Wolffe and Hansen, 2001], a single gene is controlled by a large number of DNA elements spread over many thousands of base pairs and bound by

multiple regulators [Bulger et al., 2002], and chromatin remodeling and modification complexes orders of magnitude larger than the regulators that target them are required for proper gene regulation [Narlikar et al., 2002]. Most importantly, the function of these complexes is *not* to eliminate chromatin and thereby allow gene control to occur on a stage of naked DNA, although such "chromatin elimination" can be achieved in artificial constructs [Lomvardas and Thanos, 2002]—chromatin and the complexes that regulate its structure are an intergral part of the regulatory mechanism, as we have known for more than 10 years since the pioneering work of T. Archer and G. Hager on the MMTV LTR, and of F. Winston's laboratory—on yeast histones and SWI/SNF.

The range of acute and epigenetically stable modes of gene expression mammalian genomes are capable of producing is remarkable—all these modes are built on a foundation that bacteria lack: chromatin. Paraphrasing Hamlet's comment to Horatio, our present philosophy has not even dreamt of the complexity of chromatin-based regulatory mechanisms that remain hidden in the mammalian cell nucleus.

ACKNOWLEDGMENTS

This article is dedicated to the memory of Alan Wolffe, who in the summer of 2000 summed up a discussion of future research plans with the following emphatic statement: "Now that the genome has been sequenced, there is only one thing left to do—study the epigenome!"

I apologize to the authors of the >200 primary research articles that were not cited in this minireview because of editorial restrictions on the number of references allowed. The reader is urged to consult other reviews for primary references, in particular, the article by J. Wyrick and R. Young on transcriptional network analysis, by R. Eisenman, on myc, by J. Zhang and M. Lazar, on the thyroid hormone receptor, by L. Pennachio and E. Rubin, on computational approaches to identification of mammalian regulatory DNA, and Alan's book *Chromatin Structure and Function*, along with more recent reviews by T. Jenuwein and D. Allis, and G. Narlikar and R. Kingston, on chromatin in eukaryotic gene regulation. I thank Orion Weiner for explanations of principles of chemotaxis, and Carl Pabo for a critical reading of the manuscript.

REFERENCES

Berns K, Hijmans EM, Koh E, Daley GQ, Bernards R. 2000. A genetic screen to identify genes that rescue the slow growth phenotype of c-myc null fibroblasts. Oncogene 19:3330–3334.

Birrell GW, Brown JA, Wu HI, Giaever G, Chu AM, Davis RW, Brown JM. 2002. Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents. Proc Natl Acad Sci USA 99:8778–8783.

Bulger M, Sawado T, Schubeler D, Groudine M. 2002. ChIPs of the beta-globin locus: Unraveling gene regulation within an active domain. Curr Opin Genet Dev 12: 170–177.

Bulyk ML, Huang X, Choo Y, Church GM. 2001. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci USA 98:7158–7163.

Bush A, Mateyak M, Dugan K, Obaya A, Adachi S, Sedivy J, Cole M. 1998. c-myc null cells misregulate cad and gadd45 but not other proposed c-Myc targets. Genes Dev 12:3797–3802.

Dang CV. 1999. c-Myc target genes involved in cell growth, apoptosis, and metabolism. Mol Cell Biol 19:1–11.

Eisenman RN. 2001. Deconstructing myc. Genes Dev 15: 2023–2030.

Elgin SC. 1988. The formation and function of DNase I hypersensitive sites in the process of gene activation. J Biol Chem 263:19259–19262.

Elgin SCR, Workman JL. 2000. Chromatin structure and gene expression. In: Hames BD, Glover DM, editors. Frontiers in molecular biology. Oxford; Oxford University Press.

Frank SR, Schroeder M, Fernandez P, Taubert S, Amati B. 2001. Binding of c-Myc to chromatin mediates mitogen-induced acetylation of histone H4 and gene activation. Genes Dev 15:2069–2082.

Ghosh S, Karin M. 2002. Missing pieces in the NF-kappaB puzzle. Cell 109(Suppl):S81–S96.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M. 2002. Functional profiling of the Saccharomyces cerevisiae genome. Nature 418:387–391.

Gregory PD. 2001. Transcription and chromatin converge: Lessons from yeast genetics. Curr Opin Genet Dev 11: 142–147.

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

Horak CE, Mahajan MC, Luscombe NM, Gerstein M, Weissman SM, Snyder M. 2002. GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. Proc Natl Acad Sci USA 99:2924–2929.

Jenuwein T, Allis CD. 2001. Translating the histone code. Science 293:1074–1080.

Judelson C, Privalsky ML. 1996. DNA recognition by normal and oncogenic thyroid hormone receptors. Unexpected diversity in half-site specificity controlled by non-zinc-finger determinants. J Biol Chem 271:10800–10805.

Lomvardas S, Thanos D. 2002. Modifying gene expression programs by altering core promoter chromatin architecture. Cell 110:261–271.

Lyko F, Paro R. 1999. Chromosomal elements conferring epigenetic inheritance. Bioessays 21:824–832.

Narlikar GJ, Fan HY, Kingston RE. 2002. Cooperation between complexes that regulate chromatin structure and transcription. Cell 108:475–487.

Nikiforov MA, Chandriani S, O'Connell B, Petrenko O, Kotenko I, Beavis A, Sedivy JM, Cole MD. 2002. A functional screen for myc-responsive genes reveals serine hydroxymethyltransferase, a major source of the one-carbon unit for cell metabolism. Mol Cell Biol 22:5793–5800.

Pennacchio LA, Rubin EM. 2001. Genomic strategies to identify mammalian regulatory sequences. Nat Rev Genet 2:100–109.

Ptashne M, Gann A. 2002. Genes and signals. Cold Spring Harbor, NY; Cold Spring Harbor Laboratory Press.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. 2000. Genome-wide location and function of DNA binding proteins. Science 290:2306–2309.

Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD. 2002. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev 16:245–256.

Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N, Bucher P. 2002. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol 20:831–835.

Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nat Genet 14:450–456.

Smith AG. 2001. Embryo-derived stem cells: of mice and men. Annu Rev Cell Dev Biol 17:435–462.

Thanos D, Maniatis T. 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell 83:1091–1100.

Urnov FD. 2002. A feel for the template: Zinc finger protein transcription factors and chromatin. Biochem Cell Biol 80:321–333.

Urnov FD, Wolffe AP. 2001. A necessary good: Nuclear hormone receptors and their chromatin templates. Mol Endocrinol 15:1–16.

Weiner OD. 2002. Regulation of cell polarity during eukaryotic chemotaxis: The chemotactic compass. Curr Opin Cell Biol 14:196–202.

Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ. 2002. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev 16:235–244.

Werner T. 2001. The promoter connection. Nat Genet 29:105–106.

Wolffe AP. 1998. Chromatin structure and function. San Diego, CA: Academic Press.

Wolffe AP, Hansen JC. 2001. Nuclear visions: Functional flexibility from structural instability. Cell 104:631–634.

Wyrick JJ, Young RA. 2002. Deciphering gene expression regulatory networks. Curr Opin Genet Dev 12:130–136.

Zhang J, Lazar MA. 2000. The mechanism of action of thyroid hormones. Annu Rev Physiol 62:439–466.