MINIREVIEW

Genomic Approaches That Aid in the Identification of Transcription Factor Target Genes

ANTONIS KIRMIZIS AND PEGGY J. FARNHAM[1]

McArdle Laboratory for Cancer Research, University of Wisconsin Medical School, Madison, Wisconsin 53706

It is well-established that deregulation of the transcriptional activity of many different genes has been causatively linked to human diseases. In cancer, altered patterns of gene expression are often the result of the inappropriate expression of a specific transcriptional activator or repressor. Functional studies of cancer-specific transcription factors have relied upon the study of candidate target genes. More recently, gene expression profiling using DNA microarrays that contain tens of thousands of cDNAs corresponding to human mRNAs has allowed for a large-scale identification of genes that respond to increased or decreased levels of a particular transcription factor. However, such experiments do not distinguish direct versus indirect target genes. Coupling chromatin immunoprecipitation to microarrays that contain genomic regions (ChIP-chip) has provided investigators with the ability to identify, in a high-throughput manner, promoters directly bound by specific transcription factors. Clearly, knowledge gained from both types of arrays provides complementary information, allowing greater confidence that a transcription factor regulates a particular gene. In this review, we focus on Polycomb group (PcG) complexes as an example of transcriptional regulators that are implicated in various cellular processes but about which very little is known concerning their target gene specificity. We provide examples of how both expression arrays and ChIP-chip microarray-based assays can be used to identify target genes of a particular PcG complex and suggest improvements in the application of array technology for faster and more comprehensive identification of directly regulated target genes. Exp Biol Med 229:705–721, 2004

Key words: polycomb-group proteins; chromatin immunoprecipitation; ChIP-chip; gene expression; bioinformatics

[1] To whom correspondence should be addressed at UC Davis Genome Center, Genome and Biomedical Sciences Facility, 451 East Health Sciences Drive, Davis, CA 95616. E-mail: pjfarnham@ucdavis.edu

1535-3702/04/2298-0705 $15.00

Copyright © 2004 by the Society for Experimental Biology and Medicine

Introduction

Cancer arises when the delicate balance between cell proliferation and differentiation is lost. The predominance of proliferation over differentiation requires not only that genes that function to promote cell division be turned on, but also that genes that function to limit inappropriate proliferation by activating cell death pathways be turned off. Acquisition of these proliferation-promoting and death-inhibitory characteristics requires multiple genetic and epigenetic changes in the cell. A mutation that results in the inappropriate expression of a transcriptional activator or repressor is often an early event in the development of cancer. The increased expression of a transcriptional regulator can be achieved by multiple mechanisms. First, mutational loss of a negative regulator can lead to increased activity of a transcription factor without a change in the amount of the factor in the cell. For example, loss of the retinoblastoma protein leads to increased activity of the E2F family of transcriptional activators in multiple cancer types without change of E2F protein abundance (1). Genetic mutations can also lead to increased amounts of a transcriptional regulator, via either transcriptional or posttranscriptional mechanisms. For example, chromosomal rearrangements juxtapose a transcriptional enhancer next to the c-*myc* gene in certain lymphomas, resulting in increased c-Myc mRNA and protein (2). Alternatively, loss of the adenomatous polyposis coli protein leads to increased amounts of β-catenin protein, but not mRNA, caused by changes in stability of the β-catenin protein (3). Another mechanism by which both the levels and the activity of a transcription factor can be dramatically altered is through chromosomal rearrangements that fuse the factor to the N-terminal region of another protein. For example, the carboxy terminal region of the Ets family member FLI-1 is fused to the amino terminus of EWS; expression of the EWS-FLI fusion protein is regulated by the *ews* promoter region (4). Finally, epigenetic

changes can also modulate the expression pattern of a specific transcription factor. For example, hypermethylation of a CpG island within the promoter region of the transcription factor AP-2alpha results in loss of AP-2alpha expression (5).

Regardless of the mechanism by which it is achieved, altering the activity of a transcription factor can lead to major changes in the gene expression patterns of the abnormal cell when compared with its normal counterpart. Of course, many of the changes observed are due to a domino-like effect in which the altered expression of gene A results in the altered expression of gene B, which results in the altered expression of gene C, and so forth. An understanding of the biological consequences of altering a transcriptional activator or repressor in a specific cancer requires the cataloging of all such changes in gene expression. However, an understanding of the molecular mechanisms by which the changes are mediated requires that genes directly regulated by a factor be distinguished from those in the line of dominoes. In this review, we will use the Polycomb proteins as an example of transcription factors that display altered levels of expression in human cancers, describe how microarrays can be used to develop gene expression profiles and identify direct target genes, summarize the experiments performed to date by using these arrays to study Polycomb group (PcG) proteins, and then conclude with suggestions for future experimental approaches.

PcG Proteins

During embryonic development of vertebrates and invertebrates, normal anterior-posterior axis formation is controlled by the expression of the homeotic (*hox*) genes. Generally, *hox* genes are expressed within the posterior half of the embryo but are maintained inactive in the anterior portion. Deregulation of this spatially restricted pattern of *hox* expression can lead to transformation of embryonic body segments. For example, inappropriate expression of the *hox* gene *ultrabithorax* in the anterior compartment of a *Drosophila* embryo can transform thoracic segments into the more posterior abdominal segments (6). Because of the dire consequences of inappropriate expression of Hox proteins, *hox* genes are under tight transcriptional control during development. Transcription of *hox* genes can be influenced by both positive and negative regulators known respectively as Trithorax-group (Trx) proteins and PcG proteins. This review focuses on PcG transcriptional repressors.

The *polycomb* (*pc*) gene, discovered in *Drosophila melanogaster* by P.H. Lewis in 1947 (7), was the first gene shown to be involved in the control of *hox* gene expression. The gene was given the name *polycomb* to describe one of the phenotypes observed in male flies carrying mutations of that gene. Normally, male flies have a set of bristles on their first pair of legs, known as the sex comb, which assists them during mating. Mutations in *pc* caused the development of multiple sex combs (hence the term "polycomb") on all pairs of legs in adult male flies. This prominent phenotype of *pc* mutant flies (i.e., having all legs resemble front legs) is an example of a homeotic transformation that is caused by de-repression of the *hox* gene clusters Bithorax-complex and Antennapedia-complex in the anterior portion of mutant embryos (6). Since the discovery of the *pc* gene, 15 other genes have been identified in *Drosophila* that display similar homeotic transformation phenotypes when mutated, indicating their involvement in the regulation of *hox* expression (8–21). The discovery of the *Drosophila* PcG proteins prompted the identification of mammalian homologues. Unlike many families of transcription factors, the PcG proteins are not grouped on the basis of common domains in their protein structure. Rather, they are classified in the same group because they are all discovered on the basis of their ability to repress transcription of *hox* genes. A comprehensive list of mouse and human PcG homologues is shown in Table 1.

Studies of mammalian PcGs have linked these proteins to cellular processes in addition to the developmental control of *hox* gene expression. Mice lacking the PcG proteins B lymphoma Mo-MLV insertion region 1 (Bmi1), zinc finger protein 144 (also called Mel 18), chromobox homolog 2 (also called M33), mouse polyhomeotic homolog, or embryonic ectoderm development (EED) show severe hematopoietic abnormalities in addition to the skeletal transformations that are due to misexpression of *hox* genes (22–26). Additionally, PcG proteins have been implicated in the control of X-chromosome inactivation. Specifically, the absence of EED results in a failure to maintain X-chromosome inactivation (27). More-recent studies have shown that the PcG proteins EED and enhancer of zeste 2 (EZH2) are recruited to the imprinted X chromosome during initiation of X inactivation (28, 29). PcG proteins also play an important role during normal cell proliferation. Recent studies using Bmi1 null mice have demonstrated the requirement for this PcG protein in the control of cell proliferation of normal hematopoietic and neural stem cells as well as cerebellar precursor cells (30–33). In addition, Bracken and colleagues (34) have demonstrated that specific depletion of the PcG proteins EZH2 and EED with RNA interference (RNAi) results in reduced cell proliferation of normal diploid fibroblasts.

The discovery that PcGs affect cell proliferation suggests that deregulation of PcG expression might play an important role in tumorigenesis. Accordingly, deregulation of PcG proteins has been observed in several types of cancer. For example, EZH2 upregulation has been significantly correlated with the metastatic progression of prostate and breast cancers (35, 36). Another PcG protein, suppressor of zeste 12 (SUZ12), is often upregulated in tumors of the colon, breast, and liver (37). Additionally, the *SUZ12* gene is frequently translocated in endometrial stromal sarcomas and, as a result, is fused to a gene

Table 1. *Drosophila*, Mouse, and Human PcG Orthologs[a]

Drosophila	Mouse	Human	Complex	References
Pc (polycomb)	M33 Mpc2 Mpc3 Cbx7	HPC1/CBX2 HPC2/CBX4 HPC3/CBX8/CBX6 CBX7	PRC1	7, 39–41
Ph (polyhomeotic)	Rae-28/Mph1 Mph2 Phc3	HPH1/EDR1 HPH2/EDR2 HPH3/PHC3	PRC1	12, 39–41
Psc (posterior sex combs)	Bmi1 Mel-18	BMI1	PRC1	11, 39–41
Sce/dRing (sex combs extra)	Ring1a Ring1b	RING1A/RNF1 RING1B/RNF2	PRC1	19, 20, 39–41
Scm (sex combs on midleg)	Scmh1	SCML1/SCMH1 SCMH2	PRC1	11, 39–41
Sxc (super sex combs)	ND	ND	PRC1	7, 39
E(z) (enhancer of zeste)	Enx1/Ezh2 Enx2/Ezh1	EZH2 EZH1	PRC2/3/4	17, 42–45
Esc (extra sex combs)	Eed	EED	PRC2/3/4	8, 42–45
Su(z)12 (suppressor of zeste-12)	ND	SUZ12/JJAZ1	PRC2/3/4	16, 42–45
Pcl (polycomblike)	ND	PHF1	PRC2/3/4 (1-MDa complex)	9, 11, 46
Pho (pleiohomeotic)	Yy1	YY1	—	21, 64
E(Pc) (enhancer of polycomb)	Epc1/Epc2	EPC1/EPC2	—	5
Crm (cramped)	ND	CRAMP1L	—	13
Asx (additional sex combs)	ND	ASXL1 ASXL2	—	11
Su(z)2 (suppressor of zeste-2)	ND	ND	—	18
Mxc (multisex combs)	ND	ND	—	14

[a]Some *Drosophila* Polycomb group (PcG) proteins have multiple mammalian orthologs (i.e., Pc) whereas others have only one (i.e., Esc). Mammalian orthologs corresponding to some *Drosophila* PcG proteins are not yet known, indicated as "ND" (not determined). Alternative protein names are separated by a slash (i.e., HPC1/CBX2). A dash (—) indicates that the PcG protein has not been identified as a component of a known Polycomb Repressive Complex (PRC).

encoding a zinc finger protein (38). Finally, the PcG protein Bmi1 is overexpressed frequently in human medulloblastoma cell lines and primary tumors (32). These and many other studies clearly implicate PcG protein misregulation with cancer.

A growing body of evidence suggests that PcG proteins are important regulators of cell proliferation and development. However, a major limitation in our understanding of how they control these processes is the lack of known mammalian PcG target genes. The different phenotypes observed in PcG null mice and the implication of PcG proteins in various cellular processes suggest that these proteins regulate a broad spectrum of target genes. For example, the observed skeletal transformations in PcG null mice can be explained by the misregulation of developmental control genes such as the Hox family (22). However, the involvement of PcG proteins in the development of various human cancers might be due to altered expression of genes that control processes such as cell proliferation, differentiation, and apoptosis.

PcG-Mediated Transcriptional Regulation. Although the target genes responsible for PcG-mediated regulation of mammalian development and proliferation have not yet been identified, recent studies have led to the development of a model by which PcG proteins may regulate transcription. Early genetic studies of *Drosophila* predicted that PcG proteins exert their functions by forming multimeric complexes. For example, double and triple PcG mutant flies exhibit enhanced homeotic transformations when compared with single PcG mutant flies, suggesting functional interactions among the various PcG proteins (11). Recent biochemical studies have defined the composition of two such complexes that are present in both *Drosophila* and mammalian cells (Figure1). The first complex identified was named Polycomb Repressive Complex 1 (PRC1). The human complex includes the PcG proteins human polycomb homolog, human polyhomeotic homolog, BMI1, ring finger protein 1, and sex combs on midleg human homolog 1 (39–41). The second complex was more recently defined by four independent groups. This complex is referred to as Polycomb Repressive Complex 2 (PRC2, or EED-EZH2), and the human complex consists of five core subunits: the three PcG proteins EZH2, SUZ12, and EED, as well as the histone binding factors retinoblastoma associated proteins

p46 and p48 (RbAp46 and RbAp48) (42–45). In addition to the identified 600-kDa PRC2 complex, Tie *et al.* (46) isolated a variant of the PRC2 in *Drosophila* that was about 1 megadalton. This 1-megadalton complex contained the previously identified core subunits plus the PcG protein polycomblike and the histone deacetylase Rpd3 (46). The identification of these two PRC2s raises the question whether the 600-kDa PRC2 is a stable intermediate of the larger 1-megadalton complex or whether the two complexes form independently and have distinct biological functions.

As noted above, PcG proteins function as transcriptional repressors; thus, recent studies have focused on understanding the mechanisms by which the PcG-containing complexes turn off gene expression. All the groups that identified the 600-kDa PRC2 independently demonstrated that the purified complexes possess histone lysine methyltransferase (HKMT) activity, which is mediated by the SET [Su-(var)3-9;E(z);Trithorax] domain of EZH2. PRC2 was found to methylate lysine 27 (H3-K27) and, to a lesser extent, lysine 9 (H3-K9) of histone H3 *in vitro*. From these and previous observations, a model has been proposed that can explain how the PRCs use this histone-modifying activity to initiate and maintain transcriptional repression of their target genes. The PRC2 is thought to first catalyze the methylation of lysine 27 on histone H3. This posttranslational modification of histone H3 serves as a signal for specific binding of the chromodomain of Polycomb, thus mediating the recruitment of PRC1 (47). Binding of PRC1 is then proposed to block the recruitment of transcriptional activating factors, such as SWI/SNF (a Trx protein complex) and facilitate the establishment of a stable, repressive chromatin structure (39, 48). However, a recent report in *Drosophila* proposes an alternative model that describes a different role for PRC1. The authors of that study proposed that recruitment of PRC1 by the K27 methyl mark does not block access of activating factors to PRC target promoters, but rather the presence of the PRC1 prevents initiation of transcription by prebound factors, such as Tata box-binding protein and RNA polymerase II (RNAP II) (49). This second model is also supported by studies that demonstrate restriction enzyme accessibility and colocalization of PcG proteins with members of the general transcriptional machinery at repressed PcG target genes in *Drosophila* (50, 51). It is possible that both these mechanisms of regulation are used in mammalian cells with each mechanism controlling distinct sets of PcG target promoters. Recent work in our laboratory supports this hypothesis. Using chromatin immunoprecipitation (ChIP) assays, we have shown that certain human PcG target genes exhibit binding of PcG proteins in the absence of RNAP II, supporting the model that binding of PRCs to a promoter inhibits accessibility of activating factors. Interestingly, we have also found other promoters which bind both PcGs and RNAP II, supporting the model that PRCs can silence genes by inhibition of transcriptional initiation (52). However, we cannot exclude the possibility that the apparent colocalization of RNAP II and the PRC components is due to independent binding of PcGs and RNAP II to different alleles of the tested promoter. Also, it remains possible that in a population of cross-linked cells the genes exhibiting colocalization are active in some of the cells (i.e., bound by RNAP II but not PcGs) but inactive in others (i.e., bound by PcGs but not RNAP II). Future experiments using a sequential "double" ChIP procedure could distinguish between these possibilities. In such experiments, an initial immunoprecipitation (IP) would be performed with an antibody to one of the factors (e.g., RNAP II), and then the collected immunoprecipitates would be subjected to a second IP using an antibody to the second factor (e.g., a PcG protein).

It has recently been shown that one of the components of the PRC2, EED, exists in four different isoforms in human cells (53). Accordingly, Kuzmichev and colleagues (54) have demonstrated the existence of three PRC2-like complexes in human cells that each contain different isoforms of EED. All three complexes, which are now called PRC2,-3, and -4, contain the core subunits EZH2, SUZ12, RbAp46, and RbAp48. In addition to the core subunits, PRC2 contains the longest form of EED (EED 1), PRC3 contains the two shortest forms of EED (EED3 and EED4), and PRC4 contains the intermediate form of EED (EED2) plus the histone deacetylase sirtuin 1.[2] Intriguingly, the presence of the different EED isoforms results in different catalytic specificity of the HKMT-EZH2 *in vitro*. For example, PRC2 can methylate both H3-K27 and lysine 26 of histone H1 (H1-K26) on nucleosomal arrays. In contrast, PRC3 can methylate only H3-K27 whereas PRC4 methylates H1-K26 (Figure 1).

Identifying PRC Target Genes by a Candidate-Gene Approach. It is not known whether the complexes described above have redundant functions or whether they play different roles in distinct cellular processes. Distinguishing between these two possibilities requires the identification of target genes for each of the three complexes. It would seem that a simple approach would be to examine candidate target genes based on the presence of a consensus PcG element in a promoter region. However, none of the proteins purified in the different mammalian PRCs have been shown to be site-specific DNA binding proteins; therefore, simple sequence inspection cannot suffice to identify PcG target genes. How the mammalian PRCs are targeted to DNA remains unknown. In contrast, several DNA binding proteins have been implicated in targeting PcGs to the DNA in *Drosophila*. The identification of the *hox* genes as PcG target genes in *Drosophila* led to the discovery of cis-regulatory elements in the fly genome that are required for PcG-mediated repression. Genetic

[2] D. Reinberg, personal communication.

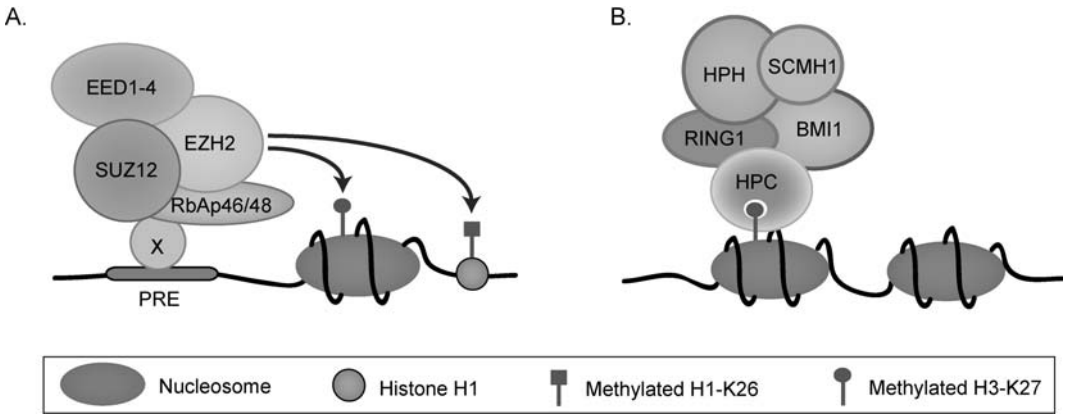

Figure 1. Polycomb Repressive Complexes (PRCs). (A) Mammalian core subunits of PRC2, -3, -4. These three complexes contain the proteins enhancer of zeste (EZH2), suppressor of zeste 12 (SUZ12), retinoblastoma associated p46 and p48 (RbAp46/48), and different isoforms of the embryonic ectoderm development (EED) protein. PRC2 contains the longest form of EED (EED2) and can methylate both lysine 26 of histone H1 (H1-K26) and lysine 27 of histone H3 (H3-K27). PRC3 contains the shortest forms of EED (EED3,4) and methylates H3-K27. PRC4 contains the intermediate form of EED (EED2) as well as the histone deacetylase sirtuin 1 (not shown) and methylates H1-K26. The histone lysine methyltransferase activity of the PRC2/3/4 complexes is mediated by the SET [Su-(var)3-9;E(z);Trithorax] domain of EZH2. In mammalian cells, it is not known whether the core components of PRC2/3/4 contact DNA directly or whether site-specific DNA binding proteins or noncoding RNAs (indicated by X) recruit these complexes to the DNA. (B) Mammalian core subunits of PRC1. This complex consists of the PcG proteins polycomb (HPC), B lymphoma Mo-MLV insertion region 1 (BMI1), polyhomeotic (HPH), ring finger protein 1 (RING1), and sex combs on midleg (SCMH1). PRC1 is recruited to the chromatin by interaction with methylated H3-K27 mediated by the chromodomain of HPC.

studies in combination with reporter assays defined the minimal DNA elements, termed PcG response elements (PREs), that regulate *hox* gene expression and mediate repression of reporter genes in *Drosophila* (55–57). Sequence alignments among the various PREs that control different *hox* genes reveal little similarity, and thus a strict consensus sequence has not been defined. However, one similarity among the different PREs is the presence of binding sites for three site-specific DNA binding proteins: the PcG protein pleiohomeotic (pho), GAGA, and zeste. In fact, Ringrose *et al.* (58) used the finding that these three sites occur frequently in PREs to develop a bioinformatic approach to identify other PREs in the *Drosophila* genome. Using this approach, the authors discovered 167 candidate PREs, many of which map close to genes that are involved in development and cell proliferation. Unfortunately, although it is clear that *Drosophila* PcG complexes use the pho, GAGA, or zeste sites in the PREs (59–63), none of these DNA binding factors or their mammalian orthologs copurify with the PRCs. The mammalian homologue of pho, known as YY1, was shown to physically interact with the WD-40 domains of EED in a yeast two-hybrid analysis. However, whether this factor facilitates recruitment of the PRCs to DNA *in vivo* (64) remains unclear.

Because the cis element corresponding to a mammalian PRE has not been identified, bioinformatic approaches have not been used to identify potential mammalian PRC target genes. In an attempt to identify mammalian PRC target genes, Jacobs and colleagues (65) demonstrated that Bmi1, a component of PRC1, cooperates with the oncogene c-Myc to repress the activity of the $p16^{INK4A}$ and $p19^{ARF}$ tumor suppressor genes, leading to transformation of lymphoid cells. Another suggestion that the $p16^{INK4A}/p19^{ARF}$ locus may be a putative mammalian PcG target came from a study where overexpression of the protein chromobox homolog 7 led to the downregulation of both the $p16^{INK4A}$ and the $p19^{ARF}$ gene, resulting in a longer lifespan than in normal cells (66). Other cell-cycle regulatory genes appear to be controlled by the modulation of the PRC components. Bracken *et al.* (34) proposed that *p53*, *cyclin D1*, *cyclin E1*, *cyclin A2*, and *cyclin B1* are potential PRC target genes because transformation of normal fibroblasts by depletion of EZH2 or EED correlates with altered expression of these genes. Additionally, Varambally and colleagues (36) used gene expression analysis to identify a large number of genes whose expression is reduced upon overexpression of EZH2 in prostate cancer cells. However, none of the above studies show recruitment of the PRCs to the promoters of the affected target genes. Therefore, whether the PcG complexes directly regulate these target genes or the target genes' expression is changed indirectly as a result of an altered cellular milieu remains unclear. As described below, the use of a high-throughput, microarray-based genomic approach that has been developed recently can address this issue.

Array-Based Approaches for the Identification of Target Genes

To determine which genes are directly regulated by the PRCs, it is necessary to identify the intersection of the set of genes whose activity is responsive to changes in levels or activity of the PRCs and the set of genes whose regulatory regions are bound by the components of the PRCs. Such an effort requires a two-step approach, but one may begin either with a gene expression array followed by a DNA binding assay or with a DNA binding array followed by a

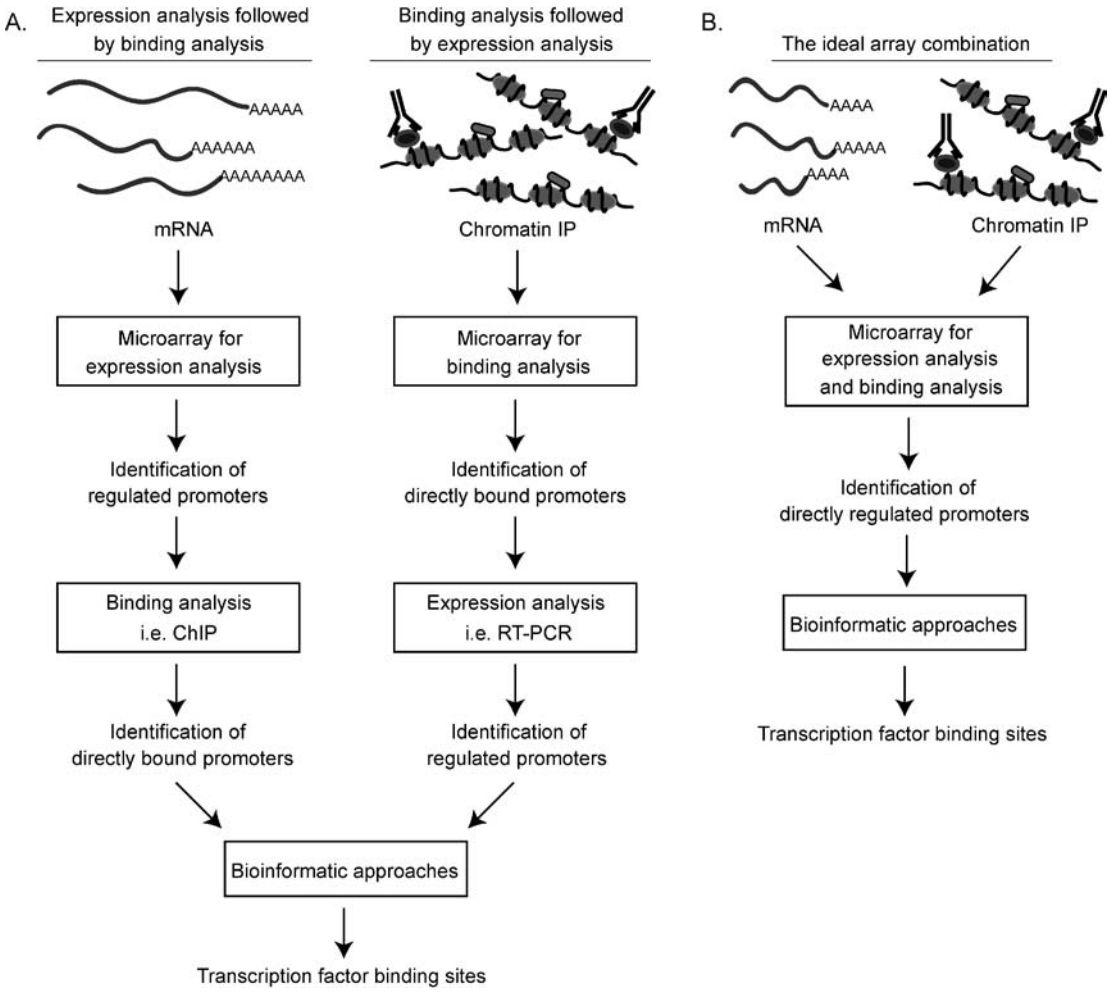

Figure 2. Microarray-based approaches used for the identification of transcription factor target genes. (A) Shown are two different methods that can be used to identify promoters that are directly bound and regulated by a specific transcription factor. The first method begins with a global gene expression analysis that can identify numerous genes whose activity is regulated by the transcription factor of interest. Then, binding analyses, such as chromatin immunoprecipitation (ChIP), are used to distinguish between directly and indirectly regulated promoters. The second method begins with a ChIP-chip analysis to identify numerous promoters that are directly bound by the transcription factor of interest, followed by gene expression approaches, such as RT-PCR, to identify functionally regulated promoters. Both methods lead to the identification of directly regulated target promoters whose primary sequence can be analyzed by computational programs to identify binding sites for the transcription factor in question. (B) Shown is an ideal approach for high-throughput identification of directly regulated target genes that circumvents many of the disadvantages of the two methods shown in (A). Such an approach requires the development of microarrays that could be used in both gene expression and ChIP-chip analysis.

gene expression assay. These two paths, each of which has distinct advantages and disadvantages, are described below (Figure 2).

Gene Expression Arrays Followed by ChIP. Using this approach, investigators first identify genes whose expression is regulated by the introduction or removal of a transcription factor and then perform follow-up studies to distinguish direct versus indirect targets. There are many ways to identify a set of genes whose expression changes upon over- or underexpression of a transcription factor. Such approaches include techniques based upon either subtractive hybridization, differential display, sequencing large numbers of clones, or hybridization to microarrays. Methods such as subtractive hybridization, differential display, and sequencing-based approaches such as serial analysis of gene expression (SAGE) have an advantage in that they are relatively unbiased and can allow the identification of mRNAs that are not in the current database. However, these methods are laborious, time consuming, and not quantitative, and it is likely that the rate of discovery of novel mRNAs will decrease quite rapidly in the near future because of the number of large-scale gene identification efforts in progress. Therefore, the advantages of these approaches will decrease, but the disadvantages will remain. In contrast, microarray-based techniques provide a rapid approach to the identification of genes responsive to a given factor. Currently available microarrays represent tens of thousands of mRNAs, and it is anticipated that future arrays will represent mRNAs of all genes. Therefore, this review will focus only on the use of

microarrays as the first step in a comprehensive identification of target genes.

Two different types of arrays can be used for the study of gene expression changes mediated by a transcription factor. The first type involves physically depositing (spotting) cDNAs or PCR fragments that were derived from mRNAs onto microscope slides. Such arrays were first produced in 1995 to study gene expression in *Arabidopsis thaliana* (67). Mammalian arrays containing cDNAs corresponding to about 1000 human genes were used in 1996 (68, 69), and 8600 human genes could be analyzed with arrays by 1999 (70). One of the first studies to analyze mRNAs from cells specifically lacking a transcription factor was the study of ATF-2 null mice (71). Innumerable studies using over- or underexpression of a particular transcription factor have been performed since then. Although a major step forward in gene expression analyses, the spotted arrays have several disadvantages. For example, the PCR fragments must be prepared, purified, quantitated, carefully catalogued, and stored. Each of these steps is expensive and subject to technical difficulties.

The second type of array used to analyze gene expression is composed of oligonucleotides that are synthesized directly on the solid phase surface based upon the sequence of known mRNAs (72). Because the oligonucleotides (which are commonly 20–25 nts in length but have been synthesized up to 60 nts) are synthesized directly on the array, many of the disadvantages associated with spotted arrays are eliminated. Initial arrays contained 65,000 probes that represented about 100 mammalian genes, but expanded sets of 4 arrays representing 6500 genes were soon created. An early example of the use of high density arrays to identify transcription factor target genes was the analysis of genes whose expression is altered after inducible expression of Wilms tumor 1 (73). Currently, commercially available arrays are used to examine gene expression changes for thousands of human and mouse genes. One commercially available array represents over 47,000 human transcripts corresponding to at least 14,500 well-characterized genes; the same company also produces a mouse array that represents about 39,000 transcripts corresponding to at least 14,000 genes (www.affymetrix.com).

Whether spotted or oligonucleotide arrays are used, it is necessary to collect mRNA from two samples that differ in the abundance of a factor of interest. One of the most common means investigators use to modulate a factor is through the introduction of a plasmid expressing a protein into a cell line and then the preparation of mRNA from the transfected cells, with mRNA preparation from the parental cell line serving as a control. This has proven to be a popular approach because it is technically easier to overexpress a protein than it is to remove a factor from a population of cells. However, overexpression has a distinct disadvantage when studying multisubunit complexes such as the PRCs. For example, if all components of a complex must be present in equal ratios, then overexpressing one component may have very little effect on expression of the target genes. A better approach is to reduce (or eliminate) one component of the complex, which would presumably lead to complex dissolution, and then to search for genes whose expression is increased or decreased (depending on whether the complex primarily activates or represses transcription). Traditionally, this has been performed by using mouse embryo fibroblasts from a knockout animal. Of course, loss of a transcription factor can be lethal, and in the past it has been difficult to study such factors. However, with the advent of RNAi technology, transient knockdowns of a factor in tissue culture cells can be achieved, but it is important to consider that this approach cannot overcome the problems associated with functional redundancy (i.e., multiple proteins, usually members of a family of transcription factors, may be able to regulate a common set of genes).

Once one prepares mRNA samples from the cells expressing normal versus altered levels of a transcription factor, the samples are labeled with fluorescent dyes and applied to microarrays. For some arrays, generally those consisting of spotted PCR fragments, two different dyes are used and the samples are applied to a single array. For oligonucleotide arrays, the samples are labeled with the same dye but applied to two separate arrays. In both cases, analysis programs are used to calculate a "fold difference" in expression levels of each analyzed mRNA in the two starting cell populations. Most mRNAs will not be changed by removal of the transcription factor and therefore will show fold differences close to 1. However, if studying normal versus "knockout" cells, levels of mRNAs whose expression is dependent upon the removed factor will decrease and levels of mRNAs whose expression is repressed by the removed factor will increase. Alternatively, if studying normal versus "overexpressing" cells, levels of mRNAs from promoters activated by the factor will increase and levels of mRNAs from promoters repressed by the factor will decrease.

In such studies, one can often end up with long lists of deregulated genes. The reason that a large number of genes are identified in such experiments is that the observed changes in mRNA may be due to direct and indirect effects of the removed factor. Clearly, removal of a factor can have major effects on multiple signaling pathways in a cell, with the deregulation of the direct target genes setting up cascades of effects on the expression of other genes. Investigators have tried to distinguish direct from indirect effects using, in the case of overexpression of a factor, approaches such as cycloheximide treatment or kinetic studies (74). However, it is possible to definitely prove that a gene is directly regulated by a factor only if one can demonstrate binding of that factor to a promoter or enhancer region of the gene in question. *In vitro* gel-shift studies have been used for such purposes; however, this type of *in vitro* experiment is no longer considered sufficient because multiple factors (e.g., different members of a family of transcription factors) can bind to the same sequence of DNA *in vitro*, especially when isolated

from other cellular proteins. Therefore, binding analyses should take into consideration the cellular milieu and the chromatin environment. Such analyses could be achieved by using the ChIP assay to determine if a candidate gene is directly regulated by a factor. Briefly, this assay involves the treatment of cells or tissue with formaldehyde, a procedure that was developed by Solomon and Varshavsky (75), to cross-link the factor to its genomic binding site. Protein-DNA cross-linking is followed by IP with an antibody specific for the factor of interest and then analysis by PCR with primers specific for a particular promoter region. With this assay, the promoters of the genes identified on the expression array can be analyzed to determine if they are directly or indirectly regulated by the factor.

Unfortunately, follow-up ChIP analysis of each of the perhaps hundreds of genes identified on an mRNA expression array would be very laborious. Also, it is often unclear which region of the promoter to analyze for direct binding. Although some factors tend to bind near the transcription start site, other factors (e.g., PRCs in *Drosophila*) bind to regions located at a great distance from the proximal promoter region. A recent study that focused on identifying target genes of human PRCs has taken an approach that reduces both of these concerns. Kirmizis *et al.* (52) first used siRNA to SUZ12 (a common component of PRC2/3/4), coupled with expression arrays, to identify a set of genes regulated directly and indirectly by SUZ12. The authors then prepared a custom oligonucleotide array consisting of 5 kb of promoter sequence from each gene that displayed significantly different expression levels. Using a ChIP-chip approach (described in more detail in the next section), they identified within the overall set of deregulated genes a set of genes bound by SUZ12. Although this subset of the genes could be conclusively classified as direct targets, it remains possible that other genes identified by the mRNA arrays are also direct targets but with the binding site located outside of the tested 5 kb region or with the antibody prevented from binding its epitope during the IP because of an unusual nucleoprotein conformation in that particular transcriptional complex.

In summary, the advantage of starting with a gene expression array is that the eventual list of identified genes will consist of those genes whose expression is regulated by the transcription factor in that particular tissue or cell type. The disadvantage is that many indirect targets will be identified, and therefore each gene must be checked as a direct versus indirect target with either individual ChIP assays or customized oligonucleotide arrays in a ChIP-chip assay. Another disadvantage is that it is not possible to know where the binding site for the factor occurs, relative to the transcription start site. Therefore, many direct targets may be mistakenly classified as indirect targets if the genomic region containing the binding site is not included in the follow-up ChIP experiments. The potential for false negatives is a serious problem for the study of mammalian PRCs because the binding site that recruits the complexes is still unknown; it is not yet possible to identify PRC binding sequences in the adjacent regions of the regulated genes and then include the identified region in follow-up analyses.

ChIP-Chip Followed by RNA Expression Analysis. A second general approach for identifying target genes of a given transcription factor is to begin with a high throughput analysis of a large number of promoter regions or a large span of genomic DNA to identify binding sites for a transcription factor. Most of the studies using this analysis rely on the technique of ChIP. However, one caveat of ChIP is that it provides information only about the binding activity of a transcription factor and does not link binding to a functional effect. For this reason, this second approach requires follow-up studies that can determine if the identified binding sites are functionally important in the regulation of a nearby gene.

The application of ChIP to the analysis of site-specific transcription factors has provided a major advance in the study of mammalian gene regulation. Although this technique has been used only in mammalian systems in the past decade (76, 77), it has now become the accepted method of linking a specific factor to the regulation of a specific gene. The success in adaptation of this technology to mammalian cells has now led to the subsequent modification of the assay from the one-gene-at-a-time approach to a more global screening of thousands of promoters. Although the ChIP-chip approach (i.e., ChIP followed by microarray analysis) was first used to study yeast transcription factors (78–84), several groups have now applied this technology to the study of mammalian factors. Several different types of microarrays have been used for the mammalian studies. One type, which consists of spotted PCR fragments corresponding to promoter regions, has been used to identify target genes of E2F, c-Myc, and hepatocyte nuclear factor (HNF) family members. For these studies, specific promoters were selected and small (less than 1 kb) regions of these promoters were created by PCR. These fragments were then spotted onto microscope slides. These studies began a few years ago with a modest number of promoter regions. For example, PCR fragments spanning from −700 to +200 of 1444 human genes were used to determine that ~9% of the promoters were bound by E2F4 (85). However, it may not be correct to assume that 9% of all promoters are regulated by E2F4, because the promoters chosen for the array were selected on the basis of their regulation during the cell cycle (86), a process known to be controlled, in part, by E2F family members. More recently, arrays containing thousands of promoters have been created. For example, Li *et al.* (87) used an array containing PCR products spanning from −650 to +250 of 4839 human genes to identify c-Myc binding sites in human Daudi cells and found that 15% of the tested promoters were occupied by c-Myc. Odom *et al.* (88) used arrays containing 13,000 human promoters (spanning from 700 bp upstream to 200 bp downstream of the transcription start sites) to identify binding sites for HNF1α (a homeodomain protein), HNF4α

(a nuclear receptor), and HNF6 (a member of the onecut family of transcription factors). The promoters chosen for analysis were those that are well characterized according to the National Center for Biotechnology Information annotation. The authors found that 1.6% and 0.8% of the promoters tested were bound by HNF1α in hepatocytes and pancreatic islets, respectively. Similarly, HNF6 bound to 1.7% and 1.4% of the promoters on the array when analyzed with hepatocytes or islets, respectively. In contrast, HNF4α bound to 11%–12% of the genes on the array in both tissues, suggesting that, like c-Myc and E2F, HNF4α may regulate a large percentage of mammalian genes. Unfortunately, these selected promoter arrays are not yet commercially available, and the cost and manpower associated with creating unique primers for tens of thousands of different promoters prohibit many labs from using this technology.

A slightly different approach in ChIP-chip assays has been to use libraries of CpG islands as a source of promoters. CpG islands are G+C-rich regions at least 200 bp long with an observed to expected ratio of CpG dinucleotides of at least 0.8. CpG islands are found in the promoters and first exons of an estimated 70% of human genes or at other regulatory regions in the genome (89). Arrays consisting of 8,000–12,000 CpG islands have been used to identify E2F, c-Myc, and SUZ12 target genes. Mao *et al.* (90) used a CpG island array and found that 12% of the clones were bound by c-Myc in human HL60 cells. The same CpG arrays have been used to identify CpG island clones bound by E2F1, E2F4, and E2F6 (91–93). Although the vast majority of the clones bound by E2F4 and E2F6 corresponded to CpG islands near promoter regions, many of the clones bound by E2F1 represented certain types of repeats. For example, sequences repeated on chromosomes 1 and 16 were specifically detected by the E2F1-immunoprecipitated DNA but not by the E2F4- or E2F6-precipitated DNA. Unfortunately, the different studies used different cell types, so it is not yet clear if the different types of identified sites are reflective of differences in the E2Fs or in the cell types used. For the E2F6 study, siRNA analysis was used to demonstrate that a subset of the identified genes were negatively regulated by E2F6 (93). One distinct advantage of the CpG island arrays is that they are now commercially available at a fairly low cost and therefore can be used by many different investigators (www.microarrays.ca).

A major disadvantage of both the "selected promoter" arrays and the CpG arrays is that they are not optimal for studying factors that regulate transcription by binding at a great distance from the start site of transcription. To overcome this problem, arrays consisting of PCR fragments of about 700 bp corresponding to 93% of the nonrepetitive regions of human chromosome 22 were created and used to identify nuclear factor-kappa B (NF-κB) and cAMP response element-binding protein (CREB) target genes (94, 95). The authors found that NF-κB bound to both noncoding and coding regions, primarily within 5 kb of the 5′ ends of genes and in introns. In the annotated region of chromosome 22, NF-κB bound to 15.5% of the loci, similar to the results obtained from the c-Myc and HNF4α ChIP-chip studies. Importantly, 90% of the identified NF-κB sites fell outside the 1-kb region upstream of a start site; using a selected promoter array would have missed these binding sites. Interestingly, NF-κB sites were also detected in unannotated regions of the genome, suggesting that yet-undiscovered genes may reside in these regions. The utility of a global genomic tiling approach to identify target genes was clearly demonstrated in this initial study. However, not only is this array not commercially available, it also suffers from the problem of having to create 21,024 unique PCR products to study this single chromosome. Clearly, expanding to the entire genome would require hundreds of thousands of PCR fragments and be very costly.

Perhaps the most promising type of array for whole genome profiling is a high-density oligonucleotide array. Such arrays have been used to identify thousands of binding sites for c-Myc, Sp1, and p53 on human chromosomes 21 and 22 (96). For these studies, tiled arrays containing on average one 25mer oligonucleotide spaced every 35 bp through the nonrepetitive regions of these two chromosomes were used in the ChIP-chip assays. The authors found 353 Sp1 sites, 756 c-Myc sites, and 48 p53 sites; extrapolation to the whole genome would suggest 25,000 Myc sites, 12,000 Sp1 sites, and 1,600 p53 sites (assuming that chromosome 21 and 22 contain an average number of genes and transcription factor binding sites as compared with the rest of the genome). The authors found that 43%, 24%, and 17% of the Sp1, c-Myc, and p53 sites, respectively, were located within 1 kb of CpG islands, indicating that only a fraction of sites would have been discovered by using CpG arrays. Interestingly, the authors found that 27%, 18%, and 0% of the Sp1, c-Myc, and p53 sites, respectively, were within 1 kb of a 5′ exon, suggesting that selected promoter arrays would have detected fewer binding sites than the CpG island arrays. Unfortunately, the authors did not attempt to determine which genes were regulated either positively or negatively by the binding of Sp1, Myc, or p53. A different array technology has recently been developed that allows the synthesis of custom high-density microarrays that can represent any genomic region of interest (97). These arrays have been used to identify PRC binding sites from a set of candidate target genes (52), as well as to identify E2F binding sites in 1% of the human genome.[3] Because these custom oligonucleotide arrays are produced by commercial sources, it is likely that they will soon be available to the scientific community. Scaling to the entire human genome will, of course, require many arrays and will most likely be quite expensive.

[3] Matthew Oberley and P.F., unpublished observations.

Although most studies have used variations of the ChIP-chip assay to identify target genes, two different approaches have also been described. One method uses a sequencing-based approach, and a second method is based on creation of a fusion between a transcription factor and a DNA-adenine methyltransferase (Dam). In the sequencing-based approach, the immunoprecipitated chromatin is not applied to an array. Rather, it is either directly cloned and then sequenced (77, 98) or turned into small tags similar to those used in SAGE analysis, concatamerized, cloned, and sequenced.[4] The sequencing-based approaches are not comprehensive and are very laborious, but they may identify targets that are not represented on selected promoter or CpG arrays. Another approach, termed DamID, circumvents the ChIP step entirely (99). In this approach, a DNA binding protein is fused to *Escherichia coli* Dam permitting methylation of DNA within 1.5–2 kb from the binding site of the DNA-bound fusion protein. Briefly, the fusion protein is introduced into cells, the cellular DNA is then extracted and digested with a restriction enzyme that cuts only at GATC (if the sequence is methylated), and then size fractionated. As a reference, the Dam protein (not fused to a DNA binding factor) is introduced into parallel cultures, the DNA extracted, digested, and size fractionated. The small DNA fragments produced by the Dam fusion protein versus the normal Dam protein are labeled with different fluorescent dyes and hybridized to a microarray. Initial experiments used cDNA-based microarrays, but more-recent studies have used arrays containing long contiguous regions of *Drosophila* genomic DNA. This technique has not yet been applied to mammalian cells and has the disadvantage in that an artificial protein must be expressed in cells, running the risk that non-physiological levels of the factor of interest may influence the number of binding sites identified. However, this technique might prove useful for identifying targets of factors that associate transiently with the chromatin and thus cannot be captured at the target locus by a cross-linking method.

The Ideal Array Combination. All the approaches described above (ChIP-chip with PCR fragments or oligonucleotide arrays, Sequence Tag Analysis of Genomic Enrichment [STAGE], or DamID arrays) provide relatively unbiased information concerning the location of binding sites for a particular transcription factor. However, they all suffer from a similar problem: it is not possible to know the precise function of each of the binding sites without additional experimentation. The genes closest to the identified binding sites must be checked individually for responsiveness to alterations in levels of the factor. However, some of the identified sites may be critical for regulation of the nearby gene in some, but not all, cells. Therefore, real targets that are regulated in a different cell type or under a different physiological condition may be inadvertently discarded with this approach. Despite these limitations, some studies have used these approaches to determine if regulation is mediated by a subset of identified binding sites. For example, some of the c-Myc target genes identified by ChIP-chip assays were analyzed for changes in gene expression by RT-PCR in experiments in which c-Myc levels were increased or decreased (90). Also, a subset of E2F6 sites identified by ChIP-chip assays were analyzed by RT-PCR after removal of E2F6 by using siRNA technology (93). However, it is clear that a complete follow-up analysis by RT-PCR or Northern blots is not possible if thousands of target genes have been identified in the binding site assays.

The ideal approach would be to create an array platform that could allow both the examination of RNA expression changes and the identification of DNA binding sites. The promoter arrays and the CpG island arrays correspond to the 5′ ends of genes and, as such, do not contain much of the transcribed regions of the genes. This makes it difficult to use these arrays for mRNA expression analysis. However, it is possible to produce 5′end–enriched cDNA populations for use with promoter or CpG arrays (100). Therefore, although not optimal, these arrays could be used to study changes in mRNA levels of the genes regulated by the CpG islands on the arrays. Clearly, a better approach would be to create arrays that tile through an entire genome at a resolution sufficient to identify a binding site. These arrays could be used to identify all the binding sites for a particular factor and to identify all RNAs (including protein-coding and noncoding RNAs) that respond to loss or over-expression of that factor. For example, Martone et al. (94) used a tiled genomic array platform, consisting of PCR fragments of about 700 bp in length, for both expression and DNA binding studies of NF-κB. Interestingly, they found that not all the promoters that are bound by NF-κB responded to changes in levels of NF-κB, suggesting either that some of the binding was nonfunctional or that these targets are regulated under different conditions or different cell types. It also remains possible that, due to inherent problems with microarray analysis, such arrays will not always provide a definitive set of target genes. Although the NF-κB study is a step in the right direction, only a small portion of the human genome (chromosome 22) was examined. Because of the size of the human genome, a comprehensive analysis using this approach would take dozens of arrays and would be quite expensive. An alternative approach, which would not be as comprehensive but which would be perhaps more generally useful, would be to create a one- or two-array set that represents 10-kb upstream of each gene plus a 1-kb portion of the 3′end of the coding region of each known gene. The probes representing the 10-kb region would mainly assist in the identification of binding sites in ChIP-chip experiments, whereas the 1-kb portion of the 3′end would primarily serve for determining RNA levels in gene expression experiments.

[4] V. Iyer, personal communication.

However, binding sites located outside the 10-kb regions would not be detected. Recent studies suggest that this is a true concern. For example, Cawley et al. (96) demonstrated that 36% of identified binding sites for the transcription factors Sp1, p53, and c-Myc were located within genes or downstream of the most 3′-end exon. As a compromise between a comprehensive genomic array and a promoter array, a "conserved region" array could be produced. Once more mammalian genome projects are completed and the comparative genomic approaches are improved, one could represent on arrays all the conserved mammalian genomic regions. This method would rely on the assumption that conserved regions represent functional domains of the genome where DNA-protein interactions and transcription would most likely occur. One problem with this method might be the need to use a considerable number of arrays to cover all the evolutionarily conserved regions as indicated by the fact that, at the nucleotide level, approximately 40% of the mouse genome is aligned to the human genome (101). However, unless the entire genome is represented on arrays, probably no other approach will provide a comprehensive identification of binding sites and examination of the transcriptome. We hope that future advances in the microarray technology will allow the fabrication of whole-genome arrays in both economical and practical ways.

Using Bioinfomatic Tools to Identify Transcription Factor Binding Sites

Several computational methods have recently been developed that use large data sets generated from microarray experiments to identify transcription factor binding sites and genomic regulatory elements. Here we describe examples of two general approaches, one that uses results from gene expression arrays and another that uses results from transcription factor binding analyses.

Over the past several years many computational programs have been developed that use global gene expression data to identify regulatory elements. In general, these computational programs use two different methodologies for identifying regulatory motifs. The first method is based on the ability to cluster genes according to their gene expression pattern (102). The underlying assumption is that genes classified in the same cluster are co-regulated and thus share similar regulatory motifs within their promoters. For example, Roth *et al.* (103) have used cDNA microarrays to identify genes that are involved in different cellular processes in yeast (i.e., galactose response). To identify regulatory elements that might play a role in the control of each cellular process, the authors first ranked the deregulated genes from each experimental system according to their changes in gene expression (i.e., from most upregulated to least upregulated). Then they selected the promoter sequences of the 10 genes with the highest changes in gene expression and used the application AlignACE to identify all the common DNA motifs in their promoter sequences. To validate the functionality of the identified motifs, the authors searched for these motifs in the promoters of other yeast genes and showed that additional genes containing the motifs were regulated similarly to the ones that were originally used for the identification of the motif.

Although clustering genes according to their expression and finding common motifs in their promoters is informative, it has limitations. This approach is based on the assumption that all co-regulated promoters share a common motif, and it does not take into account that some of the genes found in a given gene expression cluster might be a result of secondary gene expression perturbations and thus would not contain the same motif as the primary response genes. In addition, some promoters might contain the identified motif, but those genes are not regulated in a manner dictated by the identified motif because of context-dependent regulation at those promoters (i.e., control of expression is dependent on the synergy of adjacent motifs or transcription factors) (104). To avoid these limitations, the second method that uses gene expression data for the identification of regulatory motifs does not use clustering analysis. Rather, this second method initially uses computational programs to identify regulatory motifs occurring commonly in the promoter sequences of known genes, and then these motifs are correlated to collected gene expression data. The fitting of motifs to gene expression allows for the identification of the most relevant elements and also takes into account the combinatorial effects of these motifs on the control of gene expression (105, 106). However, this method is effective for discovering short and highly conserved motifs but is not reliable for identifying longer elements or motifs with degenerate sequences. To circumvent the disadvantages of both methods described above, Conlon *et al.* (107) have used a strategy, which they named "motif regressor," that combines both approaches. Using yeast that overexpress a particular transcription factor, the authors first cluster the genes according to their changes in gene expression. Then they use a motif-finding program (Motif Discovery scan [MDscan]) that allows the identification of all DNA elements that occur frequently in the promoters of the most highly responsive genes. After finding all the candidate elements, they correlate each sequence with the entire gene expression dataset to determine which motifs most likely affect transcription. Unlike the previous two approaches, this method provides higher specificity and sensitivity for finding relevant regulatory elements.

All the approaches described above were performed with yeast as a model system because the relatively small and simple yeast genome (which has a high gene density, small intergenic regions, and relatively few transcription factors) is amenable to bioinformatic analyses. Although application of bioinformatic approaches is much more difficult when studying higher eukaryotes, several groups have attempted to use computational programs to identify regulatory motifs in mammalian genomes. One example is a study that used a previous gene expression dataset (108) to

cluster all the human genes that are cell-cycle regulated (109). Using a computer program known as Promoter Integration in Microarray Analysis, Elkon *et al.* identified eight transcription factor binding sites that were over-represented in the promoter sequences of their clustered genes. Reassuringly, some of those sites corresponded to binding sites of transcription factors that are known to be involved in cell-cycle regulation (i.e., E2F).

In many of the experiments described above, different computational programs were used to identify regulatory motifs from gene expression results. Each method provides information about regulatory motifs that might control the expression of a set of genes under a specific experimental condition. However, unless the experimental condition entails either increased or decreased activity of a single transcriptional regulator, the promoter sequences of the majority of the deregulated genes identified in a microarray study will not contain a common transcription factor binding site. To circumvent this limitation, computational programs have recently been used in combination with location analyses to identify binding sites for specific transcription factors. The advantage of this approach is that all the genomic fragments that are enriched in binding analyses should contain a site that mediates the function of the transcription factor under examination. Using data from ChIP-chip experiments performed in yeast, Liu *et al.* (110) have shown that their computational method, MDscan, is able to identify known, as well as novel, consensus sites for a transcription factor. In the same study, the authors also tried three other algorithms—BioProspector, AlignACE, and CONSENSUS—in combination with the ChIP-chip dataset for the identification of the transcription factor binding sites, and they reported that those algorithms were much slower and less precise compared with MDscan. The MDscan program has also been applied to sequences derived from ChIP-chip experiments performed in human cells. Cawley and colleagues (96) identified consensus and degenerate binding sites of Sp1 in DNA fragments that were enriched with an antibody against the Sp1 transcription factor. However, in the same study, MDscan failed to discover binding sites in DNA sequences enriched by antibodies against two other known DNA binding transcription factors. This suggests that MDscan might be able to detect only mammalian binding sites, such as Sp1, which are most frequently found in core promoters. Another computational program was also used to identify binding sites in DNA sequences that were isolated by the DamID approach. Orian and colleagues (99) identified a large number of genomic loci bound by the Myc/Mad/Max family of transcription factors in *Drosophila* cells and then used the REDUCE algorithm to show a high correlation between the presence of the canonical E-box sequence (a Myc/Max/Mad binding site) and the identified transcription factor–bound regions.

Although, as described above, computational programs have been used successfully to identify binding sites from sequences enriched in location studies, in many situations they have failed to reveal the correct binding motifs. This might be because transcription factors can bind to non-consensus sequences. For example, previous ChIP-chip studies identified a large number of target promoters that did not contain a consensus site for the factor in question (87, 91, 94). Therefore, advanced bioinformatic approaches must be created that will allow the identification of degenerate binding sites. Such advancement may lie in the use of comparative genomics, also known as phylogenetic footprinting. This approach is based on the assumption that functionally important sequences are conserved through evolution and thus are maintained across several related species. One example of a study that integrates bioinformatics, phylogenetic footprinting, and experimental methods was performed by Kel *et al.* (111). The authors first identified putative binding sites for the E2F transcription factors within a large set of mammalian promoters by computer-based predictions and sequence conservation between mouse and human promoters, then they verified the binding of various E2F family members to those sites by performing ChIP assays in cultured cells. The E2F study did not begin with a set of promoters identified by ChIP-chip but instead selected the promoters by a consensus sequence. However, Kellis *et al.* (112) have used yeast ChIP-chip data and applied phylogenetic footprinting to genomic regions that are bound by transcription factors having known consensus binding sites. Surprisingly, only a few of the motifs were "discovered" by the comparative genomics approach. This result emphasizes the need for the development of improved computational methods that will aid in the identification of functional DNA motifs from sequences enriched in binding analysis.

Finally, primary DNA sequence is not the only determinant of where and when a factor will bind. DNA and histones can be modified, resulting in chromatin that contains epigenetic information that influences the binding of factors to specific genomic regions (113, 114). For example, over 25 posttranslational modifications of histone H3 have been identified that involve acetylation, methylation, and phosphorylation; these and many other modifications regulate recruitment of transcription factors and gene activity. No *in silico* approach has yet undertaken to include the epigenetic information along with the primary sequence information to determine algorithms for factor binding predictions.

Conclusions and Future Directions

This review describes two different array-based methods which, when used in combination, can identify a set of direct target genes of a specific transcription factor. Identification of a large number of transcription factor–bound loci permits the comparison of sequences for the development of a consensus binding site for transcription factors such as the PcG proteins. However, in addition to using an identified set of target genes in combination with bioinformatics approaches to develop a consensus "PRC

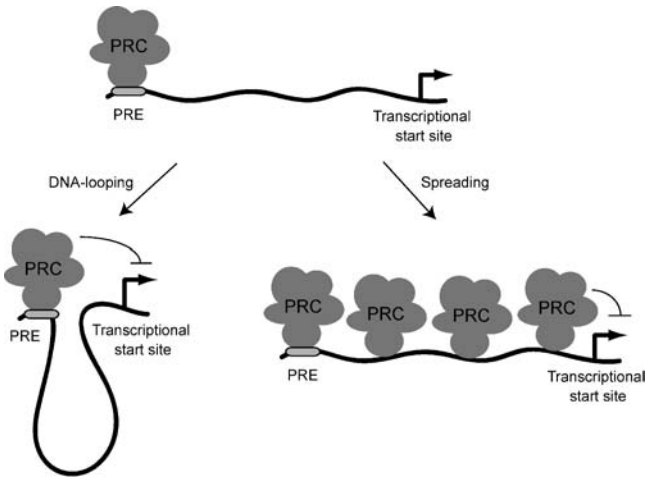

Figure 3. Model for two possible transcriptional mechanisms utilized by Polycomb Repressive Complexes (PRCs). The PcG complexes could use one of the two mechanisms indicated to regulate the transcriptional activity of promoters. The PRCs are first recruited at a distant Polycomb repressive element (PRE) and then associate with the core promoter either via a DNA-looping mechanism or by progressive spreading with adjacent lower-affinity binding sites. Once at the core promoter, the PRCs function to control the expression of the gene.

recruitment site," the array-based approaches can be used to address the following unanswered questions concerning the biological functions of PcG proteins and of the transcriptional mechanisms used by the PRCs to control gene expression.

How Can PcGs Activate Certain Genes and Repress Others? A greater knowledge of mammalian PRC target genes will allow the clarification of a dichotomy concerning PcG protein activity. Although PcG proteins and their complexes have been primarily studied in the context of their transcriptional silencing activities, several lines of evidence indicate that some of these proteins can also activate transcription in certain circumstances (34, 52, 115, 116). In fact, such PcG proteins are now classified as Enhancer of Polycomb and Trithorax proteins (117). Therefore, it will be of interest to determine how binding of PcG complexes to some target genes results in activation of gene expression and what mechanisms underline this activation (i.e., is histone methylation involved?). One approach to address this question would be to perform ChIP-chip assays with antibodies to components of the PRCs, differently modified histones, and components of the basal transcriptional machinery. The overlap of the array results could be used to determine if binding of PRCs correlates with active versus inactive chromatin and to develop hypotheses as to how recruitment of the PRC can lead to each type of chromatin state.

Do the PRCs Use the Same Mechanism to Imprint the X Chromosome As They Do to Silence Autosomal Target Genes? Evidence exists in support of the hypothesis that different mechanisms are involved in the regulation of genes on the X chromosome versus genes on the autosomes. For example, although PRC1 is needed for PRC2-mediated silencing of the *hox* genes in *Drosophila*, recent studies have demonstrated that PRC1 does not colocalize on the inactivated X chromosome with PRC2. Furthermore, PRC recruitment to the imprinted X chromosome is uniquely dependent on the Xist RNA, raising the possibility that a protein-RNA interaction mediates the recruitment of PRCs to the X chromosome (28, 29). Elucidation of the mechanisms by which PRCs mediate repression requires the identification of mammalian PRC target loci located on both autosomal and X-chromosomal regions. A ChIP-chip assay (with antibodies to the PRC components) with an X-chromosome–specific tiling array may show unique recruitment patterns.

How Do PRCs Communicate with the Core Promoter Region? Polycomb Regressive Complexes could use either one of the two modes of action depicted in Figure 3 to regulate their target genes. In the first model, the PRCs bind to a distant enhancer element and then, via a DNA looping mechanism, contact the core promoter via protein-protein interactions to regulate transcriptional activity. Alternatively, the PRCs could use an extensive spreading mechanism, which would entail binding of a PRC to a high-affinity binding site, followed by consecutive recruitment of additional PRCs to nearby low-affinity sites. Studies in *Drosophila* and recent preliminary evidence in human cells favor the DNA looping mechanism, even though strong evidence that would exclude the spreading mechanism is lacking (52, 118). The ability to develop special "tiling" arrays of target genes identified in ChIP-chip experiments could help distinguish between the two modes of action of PRCs.

In summary, several current genomic approaches can be used to identify a large set of PRC target genes, providing a better understanding of the function of the PRCs in both normal and diseased cells. Future studies that include the development of more-refined microarray platforms and the continued development of algorithms that take into account both primary sequence, as well as epigenetic information, should allow the derivation of likely transcription factor binding sites from sets of experimentally identified target genes.

We would like to thank Matthew J. Oberley for helpful discussions and insightful comments on the manuscript. We apologize to many colleagues for not citing their work because of space limitations. Work in our laboratory is supported by grants from the National Institute of Health (CA45240) and the Department of Defense (BC020760).

1. Nevins JR. The Rb/E2F pathway and cancer. Hum Mol Genet 10:699–703, 2001.
2. Boxer LM, Dang CV. Translocations involving c-myc and c-myc function. Oncogene 20:5595–5610, 2001.
3. Fodde R, Smits R, Clevers H. APC, signal transduction and genetic instability in colorectal cancer. Nat Rev Cancer 1:55–67, 2001.
4. May WA, Lessnick SL, Braun BS, Klemsz M, Lewis BC, Lunsford

LB, Hromas R, Denny CT. The Ewing's sarcoma EWS/FLI-1 fusion gene encodes a more potent transcriptional activator and is a more powerful transforming gene than FLI-1. Mol Cell Biol 13:7393–7398, 1993.

5. Douglas DB, Akiyama Y, Carraway H, Belinsky SA, Esteller M, Gabrielson E, Weitzman S, Williams T, Herman JG, Baylin SB. Hypermethylation of a small CpGuanine-rich region correlates with loss of activator protein-2alpha expression during progression of breast cancer. Cancer Res 64:1611–1620, 2004.
6. Lewis EB. A gene complex controlling segmentation in *Drosophila*. Nature 276:565–570, 1978.
7. Lindsley OL, Grell EH. Genetic variations of *Drosophila melanogaster*. Carnegie Inst of Wash Publ, Number 627:1968.
8. Struhl G. A gene product required for correct initiation of segmental determination in *Drosophila*. Nature 293:37–41, 1981.
9. Duncan IM. Polycomblike: a gene that appears to be required for the normal expression of the bithorax and antennapedia gene complexes of *Drosophila melanogaster*. Genetics 102:49–70, 1982.
10. Ingham PW. A gene that regulates the bithorax complex differentially in larval and adult cells of *Drosophila*. Cell 37:815–823, 1984.
11. Jurgens G. A group of genes controlling the spatial expression of the bithorax complex in *Drosophila*. Nature 316:153–155, 1985.
12. Dura JM, Brock HW, Santamaria P. Polyhomeotic: a gene of *Drosophila melanogaster* required for correct expression of segmental identity. Mol Gen Genet 198:213–220, 1985.
13. Yamamoto Y, Girard F, Bello B, Affolter M, Gehring WJ. The cramped gene of *Drosophila* is a member of the Polycomb-group, and interacts with mus209, the gene encoding Proliferating Cell Nuclear Antigen. Development 124:3385–3394, 1997.
14. Santamaria P, Randsholt NB. Characterization of a region of the X chromosome of *Drosophila* including multi sex combs (mxc), a Polycomb group gene which also functions as a tumour suppressor. Mol Gen Genet 246:282–290, 1995.
15. Stankunas K, Berger J, Ruse C, Sinclair DA, Randazzo F, Brock HW. The enhancer of polycomb gene of *Drosophila* encodes a chromatin protein conserved in yeast and mammals. Development 125:4055–4066, 1998.
16. Birve A, Sengupta AK, Beuchle D, Larsson J, Kennison JA, Rasmuson-Lestander A, Muller J. Su(z)12, a novel *Drosophila* Polycomb group gene that is conserved in vertebrates and plants. Development 128:3371–3379, 2001.
17. Persson K. Modification of the eye colour mutant zeste by suppressor, enhancer and minute genes in *Drosophila melanogaster*. Heretidas 82:111–119, 1976.
18. Adler PN, Charlton J, Brunk B. Genetic interactions of the suppressor 2 of zeste region genes. Dev Genet 10:249–260, 1989.
19. Breen TR, Duncan IM. Maternal expression of genes that regulate the bithorax complex of *Drosophila melanogaster*. Dev Biol 118:442–456, 1986.
20. Fritsch C, Beuchle D, Muller J. Molecular and genetic analysis of the Polycomb group gene Sex combs extra/Ring in *Drosophila*. Mech Dev 120:949–954, 2003.
21. Simon J, Chiang A, Bender W. Ten different Polycomb group genes are required for spatial control of the abdA and AbdB homeotic products. Development 114:493–505, 1992.
22. van der Lugt NM, Domen J, Linders K, van Roon M, Robanus-Maandag E, te Riele H, van der Valk M, Deschamps J, Sofroniew M, van Lohuizen M. Posterior transformation, neurological abnormalities, and severe hematopoietic defects in mice with a targeted deletion of the bmi-1 proto-oncogene. Genes Dev 8:757–769, 1994.
23. Akasaka Y, Kanno M, Balling R, Taniguchi M, Koseki H. A role for mel-18, a Polycomb group-related vertebrate gene, during the anteroposterior specification of the axial skeleton. Development 122:1513–1522, 1996.
24. Core N, Bel S, Gaunt SJ, Aurrand-Lions M, Pearce J, Fisher A, Djabali M. Altered cellular proliferation and mesoderm patterning in Polycomb-M33-deficient mice. Development 124:721–729, 1997.
25. Takihara Y, Tomotsune D, Shirai M, Katoh-Fukui Y, Nishii K, Motaleb MA, Nomura M, Tsuchiya R, Fujita Y, Shibata Y, Higashinakagawa T, Shimada K. Targeted disruption of the mouse homologue of the *Drosophila* polyhomeotic gene leads to altered anteroposterior patterning and neural crest defects. Development 124:3673–3682, 1997.
26. Tokimasa S, Ohta H, Sawada A, Matsuda Y, Kim JY, Nishiguchi S, Hara J, Takihara Y. Lack of the Polycomb-group gene rae28 causes maturation arrest at the early B-cell developmental stage. Exp Hematol 29:93–103, 2001.
27. Wang J, Mager J, Chen Y, Schneider E, Cross JC, Nagy A, Magnuson T. Imprinted X inactivation maintained by a mouse Polycomb group gene. Nat Genet 28:371–375, 2001.
28. Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y. Role of histone H3 lysine 27 methylation in X inactivation. Science 300:131–135, 2003.
29. Silva J, Mak W, Zvetkova I, Appanah R, Nesterova TB, Webster Z, Peters AH, Jenuwein T, Otte AP, Brockdorff N. Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. Dev Cell 4:481–495, 2003.
30. Molofsky AV, Pardal R, Iwashita T, Park IK, Clarke MF, Morrison SJ. Bmi-1 dependence distinguishes neural stem cell self-renewal from progenitor proliferation. Nature 425:962–967, 2003.
31. Lessard J, Sauvageau G. Bmi-1 determines the proliferative capacity of normal and leukaemic stem cells. Nature 423:255–260, 2003.
32. Leung C, Lingbeek M, Shakhova O, Liu J, Tanger E, Saremaslani P, van Lohuizen M, Marino S. Bmi1 is essential for cerebellar development and is overexpressed in human medulloblastomas. Nature 428:337–341, 2004.
33. Park IK, Qian D, Kiel M, Becker MW, Pihalja M, Weissman IL, Morrison SJ, Clarke MF. Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells. Nature 423:302–305, 2003.
34. Bracken AP, Pasini D, Capra M, Prosperini E, Colli E, Helin K. EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer. EMBO J 22:5323-5335, 2003.
35. Kleer CG, Cao Q, Varambally S, Shen R, Ota I, Tomlins SA, Ghosh D, Sewalt RG, Otte AP, Hayes DF, Sabel MS, Livant D, Weiss SJ, Rubin MA, Chinnaiyan AM. EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proc Natl Acad Sci U S A 100:11606–11611, 2003.
36. Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, Ghosh D, Pienta KJ, Sewalt RG, Otte AP, Rubin MA, Chinnaiyan AM. The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature 419:624–629, 2002.
37. Kirmizis A, Bartley SM, Farnham PJ. Identification of the Polycomb group protein SU(Z)12 as a potential molecular target for human cancer therapy. Mol Cancer Ther 2:113–121, 2003.
38. Koontz JI, Soreng AL, Nucci M, Kuo FC, Pauwels P, van den Berghe H, Dal Cin P, Fletcher JA, Sklar J. Frequent fusion of the JAZF1 and JJAZ1 genes in endometrial stromal tumors. Proc Natl Acad Sci U S A 98:6348–6353, 2001.
39. Shao Z, Raible F, Mollaaghababa R, Guyon JR, Wu CT, Bender W, Kingston RE. Stabilization of chromatin structure by PRC1, a Polycomb complex. Cell 98:37–46, 1999.
40. Saurin AJ, Shao Z, Erdjument-Bromage H, Tempst P, Kingston RE. A *Drosophila* Polycomb group complex includes Zeste and dTAFII proteins. Nature 412:655–660, 2001.
41. Levine SS, Weiss A, Erdjument-Bromage H, Shao Z, Tempst P, Kingston RE. The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. Mol Cell Biol 22:6070–6078, 2002.
42. Czermin B, Melfi R, McCabe D, Seitz V, Imhof A, Pirrotta V.

Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. Cell 111:185–196, 2002.

43. Muller J, Hart CM, Francis NJ, Vargas ML, Sengupta A, Wild B, Miller EL, O'Connor MB, Kingston RE, Simon JA. Histone methyltransferase activity of a *Drosophila* Polycomb group repressor complex. Cell 111:197–208, 2002.
44. Cao R, Wang L, Xia L, Erdjument-Bromage H, Tempst P, Jones RS, Zhang Y. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science 298:1039–1043, 2002.
45. Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P, Reinberg D. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev 16:2893–2905, 2002.
46. Tie F, Prasad-Sinha J, Birve A, Rasmuson-Lestander A, Harte PJ. A 1-megadalton ESC/E(Z) complex from *Drosophila* that contains polycomblike and RPD3. Mol Cell Biol 23:3352–3362, 2003.
47. Fischle W, Wang Y, Jacobs SA, Kim Y, Allis CD, Khorasanizadeh S. Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. Genes Dev 17:1870–1881, 2003.
48. Francis NJ, Saurin AJ, Shao Z, Kingston RE. Reconstitution of a functional core polycomb repressive complex. Mol Cell 8:545–556, 2001.
49. Dellino GI, Schwartz YB, Farkas G, McCabe D, Elgin SC, Pirrotta V. Polycomb silencing blocks transcription initiation. Mol Cell 13:887–893, 2004.
50. Breiling A, Turner BM, Bianchi ME, Orlando V. General transcription factors bind promoters repressed by Polycomb group proteins. Nature 412:651–655, 2001.
51. Fitzgerald DP, Bender W. Polycomb group repression reduces DNA accessibility. Mol Cell Biol 21:6585–6597, 2001.
52. Kirmizis A, Bartley SM, Kuzmichev A, Margueron R, Reinberg D, Green R, Farnham PJ. Silencing of human polycomb target genes is associated with methylation of histone H3 lysine 27. Genes Dev 18: 1592–1605, 2004.
53. Pasini D, Bracken AP, Helin K. Polycomb group proteins in cell cycle progression and cancer. Cell Cycle 3:22–26, 2004.
54. Kuzmichev A, Jenuwein T, Tempst P, Reinberg D. Different ezh2-containing complexes target methylation of histone h1 or nucleosomal histone h3. Mol Cell 14:183–193, 2004.
55. Muller J, Bienz M. Long range repression conferring boundaries of Ultrabithorax expression in the *Drosophila* embryo. EMBO J 10:3147–3155, 1991.
56. Simon J, Chiang A, Bender W, Shimell MJ, O'Connor M. Elements of the *Drosophila* bithorax complex that mediate repression by Polycomb group products. Dev Biol 158:131–144, 1993.
57. Chan CS, Rastelli L, Pirrotta V. A Polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression. EMBO J 13:2553–2564, 1994.
58. Ringrose L, Rehmsmeier M, Dura JM, Paro R. Genome-wide prediction of Polycomb/Trithorax response elements in *Drosophila melanogaster*. Dev Cell 5:759–771, 2003.
59. Tillib S, Petruk S, Sedkov Y, Kuzin A, Fujioka M, Goto T, Mazo A. Trithorax- and Polycomb-group response elements within an Ultrabithorax transcription maintenance unit consist of closely situated but separable sequences. Mol Cell Biol 19:5189–5202, 1999.
60. Fritsch C, Brown JL, Kassis JA, Muller J. The DNA-binding polycomb group protein pleiohomeotic mediates silencing of a *Drosophila* homeotic gene. Development 126:3905–3913, 1999.
61. Shimell MJ, Peterson AJ, Burr J, Simon JA, O'Connor MB. Functional analysis of repressor binding sites in the iab-2 regulatory region of the abdominal-A homeotic gene. Dev Biol 218:38–52, 2000.
62. Mishra RK, Mihaly J, Barges S, Spierer A, Karch F, Hagstrom K, Schweinsberg SE, Schedl P. The iab-7 polycomb response element maps to a nucleosome-free region of chromatin and requires both GAGA and pleiohomeotic for silencing activity. Mol Cell Biol 21:1311–1318, 2001.
63. Busturia A, Lloyd A, Bejarano F, Zavortink M, Xin H, Sakonju S. The MCP silencer of the *Drosophila* Abd-B gene requires both Pleiohomeotic and GAGA factor for the maintenance of repression. Development 128:2163–2173, 2001.
64. Satijn DP, Hamer KM, den Blaauwen J, Otte AP. The polycomb group protein EED interacts with YY1, and both proteins induce neural tissue in *Xenopus* embryos. Mol Cell Biol 21:1360–1369, 2001.
65. Jacobs JJ, Scheijen B, Voncken JW, Kieboom K, Berns A, van Lohuizen M. Bmi-1 collaborates with c-Myc in tumorigenesis by inhibiting c-Myc-induced apoptosis via INK4a/ARF. Genes Dev 13:2678–2690, 1999.
66. Gil J, Bernard D, Martinez D, Beach D. Polycomb CBX7 has a unifying role in cellular lifespan. Nat Cell Biol 6:67–72, 2004.
67. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470, 1995.
68. DeRisi J, Penland L, Brown PG, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JMG. Use of cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 14:457–460, 1996.
69. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A 93:10614–10619, 1996.
70. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson JJ, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO. The transcriptional program in the response of human fibroblasts to serum. Science 283:83–87, 1999.
71. Maekawa T, Bernier F, Sato M, Nomura S, Singh M, Inoue Y, Tokunaga T, Imai H, Yokoyama M, Reimold A, Glimcher LH, Ishii S. Mouse ATF-2 null mutants display features of a severe type of meconium aspiration syndrome. J Biol Chem 274:17813–17819, 1999.
72. Lockhart D, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675–1680, 1996.
73. Lee SB, Huang K, Palmer R, Truong VB, Herzlinger D, Kolquist KA, Wong J, Paulding C, Yoon SK, Gerald W, Oliner JD, Haber DA. The Wilms tumor suppressor WT1 encodes a transcriptional activator of amphiregulin. Cell 98:663–673, 1999.
74. Muller H, Bracken AP, Vernell R, Moroni MC, Christians F, Grassilli E, Prosperini E, Vigo E, Oliner JD, Helin K. E2Fs regulate the expression of genes involved in differentiation, development, proliferation, and apoptosis. Genes Dev 15:267–285, 2001.
75. Solomon MJ, Varshavsky A. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. Proc Natl Acad Sci U S A 82:6470–6474, 1985.
76. Boyd KE, Farnham PJ. Myc versus USF: Discrimination at the *cad* gene is determined by core promoter elements. Mol Cell Biol 17:2529–2537, 1997.
77. Grandori C, Mac J, Siebelt F, Ayer DE, Eisenman RN. Myc-Max heterodimers activate a DEAD box gene and interact with multiple E box-related sites *in vivo*. EMBO J 15:4344–4357, 1996.
78. 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. Genome-wide location and function of DNA binding proteins. Science 290:2306–2309, 2000.
79. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO.

Genomic binding sites of the yeast cell-cycle transcription factor SBF and MBF. Nature 409:533–538, 2001.

80. Lieb JD, Liu X, Botstein B, Brown PO. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet 28:327–334, 2001.
81. Wyrick JJ, Aparicio JG, Chen T, Barnett JD, Jennings EG, Young RA, Bell SP, Aparicio OM. Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. Science 294:2357–2360, 2001.
82. Ng HH, Robert F, Young RA, Struhl K. Genome-wide location and regulated recruitment of the RSC nucleosome-remodeling complex. Genes Dev 16:806–819, 2002.
83. Damelin M, Simon I, Moy TI, Wilson B, Komili S, Tempst P, Roth FP, Young RA, Cairns BR, Silver PA. The genome-wide localization of Rsc9, a component of the RSC chromatin-remodeling complex, changes in response to stress. Mol Cell 9:563–573, 2002.
84. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298:799–804, 2002.
85. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD. E2F integrates cell cycle progression with DNA repair, replication, and G2/M checkpoints. Genes Dev 16:245–256, 2002.
86. Ishida S, Huang E, Zuzan H, Spang R, Leone G, West M, Nevins JR. Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. Mol Cell Biol 21:4684–4699, 2001.
87. Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. Proc Natl Acad Sci U S A 100:8164–8169, 2003.
88. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA. Control of pancreas and liver gene expression by HNF transcription factors. Science 303:1378–1381, 2004.
89. Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. Nat Genet 29:412–417, 2001.
90. Mao DYL, Watson JD, Yan PS, Barsyte-Lovejoy D, Khosravi F, Wong WW-L, Farnham PJ, Huang TH-M, Penn LZ. Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. Curr Biol 13:882–886, 2003.
91. Weinmann AS, Yan PS, Oberley MJ, Huang TH-M, Farnham PJ. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev 16:235–244, 2002.
92. Wells J, Yan PS, Cechvala M, Huang T, Farnham PJ. Identification of novel pRb binding sites using CpG microarrays suggests that E2F recruits pRb to specific genomic sties during S phase. Oncogene 22:1445–1460, 2003.
93. Oberley MJ, Inman D, Farnham PJ. E2F6 negatively regulates BRCA1 in human cancer cells without methylation of histone H3 on lysine 9. J Biol Chem 278:42466–42476, 2003.
94. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. Distribution of NF-kappaB-binding sites across human chromosome 22. Proc Natl Acad Sci U S A 100:12247–12252, 2003.
95. Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M, Weissman S, Snyder M. CREB binds to multiple loci on human chromosome 22. Mol Cell Biol 24:3804–3814, 2004.
96. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K, Gingeras TR. 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.
97. Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, Pitas A, Richmond T, Gorski T, Berg JP, Ballin J, McCormick M, Norton J, Pollock T, Sumwalt T, Butcher L, Porter D, DeRosa T, Molla M, Hall C, Blattner F, Sussman MR, Wallace RL, Cerrina F, Green RD. Gene expression analysis using oligonucleotide arrays produced by masklass photolithography. Genome Res 12:1749–1755, 2002.
98. Weinmann AS, Bartley SM, Zhang MQ, Zhang T, Farnham PJ. The use of chromatin immunoprecipitation to clone novel E2F target promoters. Mol Cell Biol 21:6820–6832, 2001.
99. Orian A, van Steensel B, Delrow J, Bussemaker HJ, Li L, Sawado T, Williams E, Loo LW, Cowley SM, Yost C, Pierce S, Edgar BA, Parkhurst SM, Eisenman RN. Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. Genes Dev 17:1101–1114, 2003.
100. Shi H, Wei SH, Leu YW, Rahmatpanah F, Liu JC, Yan PS, Nephew KP, Huang TH. 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.
101. Waterson RH, Lindblad-Toh K, Birney E, Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562, 2002.
102. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863–14868, 1998.
103. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16:939–945, 1998.
104. Fry CJ, Farnham PJ. Context-dependent transcriptional regulation. J Biol Chem 274:29583–29586, 1999.
105. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nat Genet 27:167–174, 2001.
106. Keles S, van der Laan M, Eisen MB. Identification of regulatory elements using a feature selection method. Bioinformatics 25:1133–1136, 2002.
107. Conlon EM, Liu XS, Lieb JD, Liu JS. Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci U S A 100:3339–3344, 2003.
108. Whitfield ML, Zheng LX, Baldwin A, Ohta T, Hurt MM, Marzluff WF. Stem-loop binding protein, the protein that binds the 3′ end of histone mRNA, is cell cycle regulated by both translational and posttranslational mechanisms. Mol Cell Biol 20:4188–4198, 2000.
109. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Res 13:773–780, 2003.
110. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol 20:835–839, 2002.
111. Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ. Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. J Mol Biol 309:99–120, 2001.
112. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254, 2003.
113. Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC, Kouzarides T. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature 410:120–124, 2001.
114. Nielsen PR, Nietlispach D, Mott HR, Callaghan J, Bannister AJ, Kouzarides T, Murzin AG, Murzina N, Laue ED. Structure of the

HP1 chromodomain bound to histone H3 methylated at lysine 9. Nature 416:103–107, 2002.

115. LaJeunesse D, Shearn A. E(z): a polycomb group gene or a trithorax group gene? Development 122:2189–2197, 1996.
116. Gildea JJ, Lopez R, Shearn A. A screen for new trithorax group genes identified little imaginal discs, the *Drosophila melanogaster* homologue of human retinoblastoma binding protein 2. Genetics 156:645–663, 2000.
117. Brock HW, van Lohuizen M. The Polycomb group—no longer an exclusive club? Curr Opin Genet Dev 11:175–181, 2001.
118. Pirrotta V, Poux S, Melfi R, Pilyugin M. Assembly of Polycomb complexes and silencing mechanisms. Genetica 117:191–197, 2003.