

Genetics and epigenetics: stability and plasticity during cellular differentiation

Fabio Mohn and Dirk Schübeler

Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, 4058 Basel, Switzerland

Stem cells and multipotent progenitor cells face the challenge of balancing the stability and plasticity of their developmental states. Their self-renewal requires the maintenance of a defined gene-expression program, which must be stably adjusted towards a new fate upon differentiation. Recent data imply that epigenetic mechanisms can confer robustness to steady state gene expression but can also direct the terminal fate of lineage-restricted multipotent progenitor cells. Here, we review the latest models for how changes in chromatin and DNA methylation are regulated during cellular differentiation. We further propose that targets of epigenetic repression share common features in the sequences of their regulatory regions, thereby suggesting a co-evolution of epigenetic pathways and classes of *cis*-acting elements.

Epigenetic mechanisms and chromatin modulate gene expression

The tight control of gene expression programs at a given developmental stage is crucial to govern cell function and identity. The balance of stability versus plasticity in transcriptional programs presents an inherent regulatory challenge for developing organisms [1]. This difficulty is most obvious in mammalian embryonic stem (ES) cells, which have the potential to develop into every cell type of the adult organism. At the same time, these cells can be readily maintained in a pluripotent state *ex vivo* under defined culture conditions but can also be induced rapidly to differentiate.

The most important mediators for turning on or off expression of particular genes are DNA-sequence-specific transcription factors. Over the past few years, a large body of evidence indicated that chromatin-based regulatory mechanisms, in addition to transcription factors, could have important roles in establishing and maintaining transcriptional programs. This layer of control comprises post-translational modifications of DNA-bound histones, DNA methylation and chromatin remodeling [2,3]. All these pathways are currently referred to as being epigenetic, which, by stringent definition, involves a sequence-independent inheritance pattern during cell division in the absence of the initial trigger [4]. Currently, however, the mode of propagation is only known for DNA methylation [5], whereas several models have been proposed for post-translational histone modifications, including the involvement of positive-feedback loops [2,6,7]. Such models are

compatible with the phenotypes observed in relevant knockout models and predict a self-perpetuation of modifications after the deposition of new nucleosomes, which is supported by protein interaction data [2,6,8]. However, it is inherently difficult to clearly distinguish between sequence-independent self-propagation of chromatin states and a re-establishment after cell division mediated by sequence-dependent recruitment of proteins or RNA, which in turn modify chromatin. To circumvent a discussion of the use of the term epigenetics, we will hereafter refer to a recent definition, which states that lasting chromatin changes can be termed epigenetic irrespective of proof of inheritance [9] (this nevertheless emphasizes the need for further investigation into if and how chromatin modifications can be inherited).

In eukaryotes, the packaging of DNA into nucleosomes provides a basic layer of repression because it reduces DNA access [10–12]. In a simplified view, any additional change in chromatin structure could thus further restrict access for DNA-binding factors or relieve repression and, therefore, potentially reside upstream of sequence-based regulation. These epigenetic modifications are thought to modulate DNA accessibility for transcription factors and the transcription machinery itself. Although the exact molecular mechanisms and their interplay with epigenetic modifications in controlling accessibility remain incompletely understood, it is hypothesized that they operate by directly blocking transcription-factor binding and/or by establishing higher-order chromatin structures, which would be either permissive or restrictive for transcription. The best studied example for the latter is histone H4 Lys16 (H4K16) acetylation, which can prevent higher-order structure formation, thus, indicating that H4K16 acetylation is directly involved in mediating an accessible chromatin state [13]. Conversely, it seems that Polycomb target genes localize to structures termed Polycomb bodies and adopt a higher-order conformation, which excludes active chromatin and is thought to enhance repression [14,15].

Here, we review recent insight into epigenome changes during cellular differentiation and their potential impact on gene regulation and further developmental potential. In particular, we discuss evidence that CpG-rich and CpG-poor promoters are differentially regulated by epigenetic pathways, which is compatible with a model of genome partitioning through chromatin modifications in vertebrates.

Genome partitioning by chromatin

Over evolutionary time, genome sizes increased along with organismal complexity. For example, the

Corresponding authors: Mohn, F. (fabio.mohn@fmi.ch); Schübeler, D. (dirk@fmi.ch)

roundworm *Caenorhabditis elegans* genome ($\sim 10^8$ base pairs [bp]) is 30 times smaller than the human genome ($\sim 3 \times 10^9$ bp). However, both contain approximately the same number of genes ($\sim 20\,000$), showing that increased organismal complexity and genome size is not paralleled by a rise in gene number (the so-called C-value enigma [16]). In humans, and vertebrates in general, genome expansion results largely from the accumulation of repetitive and transposable elements, which is suggested to be a consequence of obligatory sexual reproduction [17]. As a result, only a small portion of vertebrate genomes encode proteins or regulatory RNA [18]. The transcriptional machinery therefore faces the challenge of locating *cis*-regulatory regions in a 'sea' of seemingly non-functional DNA sequence. One might expect that the complexity of sequence motifs that are recognized by transcription factors has similarly increased to enable specific binding to defined genomic sites. Surprisingly, however, eukaryotic transcription-factor-recognition motifs tend to be as short (6–8 bp) as those in prokaryotes and, in many cases, their binding sites are degenerate [19]. To illustrate the problem, consider that any 6-mer recognition motif occurs by chance every 4096 bp. If we assume a random sequence distribution, this would predict $>781\,200$ binding sites in the human genome, which, for a given transcription factor, needs to be multiplied by the number of degenerate motifs it can recognize. *In vivo*, only a subset of these millions of sites is occupied, raising the question of how specificity is generated. One prominent way is cooperative binding of multiple factors [20]. In addition, the large genomes of higher eukaryotes require further structuring to direct transcription factors to appropriate targets and to reduce random binding, which in turn would dilute the pool of available factors and potentially lead to inappropriate gene regulation [21]. Epigenetic pathways that modify chromatin and DNA also coevolved along with increasing genome size. They are thus *bona fide* candidates to function in a potential partitioning of the genome into 'accessible' genic and regulatory compartments and 'inaccessible' repeat-containing regions. Indeed, previous work indicated that two major evolutionary steps, the origin of eukaryotes and the origin of vertebrates, were only possible owing to the parallel evolution of new mechanisms to control 'transcriptional noise' as an otherwise unavoidable by-product of increasing genome complexity [22]. The prokaryote-to-eukaryote transition was paralleled by the appearance of nucleosomes, which, compared with naked DNA, reduce the chance of aberrant transcription initiation [10,11]. The invertebrate-to-vertebrate step was accompanied by the advent of genome-wide DNA methylation, a modification that enables efficient transcriptional repression [23,24]. A second major repressive epigenetic pathway that coevolved in multicellular organisms along with increasing genome size and organismal complexity is mediated by Polycomb group (PcG) proteins. PcG proteins underwent marked expansion over evolutionary time [25], which is in line with the concept that increasing genome size requires additional repressive mechanisms to enhance specificity of transcription initiation and suppression of 'transcriptional noise'.

DNA methylation shapes mammalian promoter structure and stabilizes pluripotency shut-down during differentiation

DNA methylation is an efficient epigenetic repression pathway, which, in vertebrates, occurs only at cytosines in the context of CpG dinucleotides. It is catalyzed by three DNA methyltransferases (Dnmts), which are all essential [5]. *Dnmt1*- or *Dnmt3b*-deficient mouse embryos die by embryonic day 10.5 and *Dnmt3a*-deficient mice are born occasionally but suffer serious malformations and die within weeks [5]. Species that undergo widespread DNA methylation in their genome have lost CpG dinucleotides over evolutionary time. This is a direct consequence of DNA methylation because it results from increased C-to-T transitions that occur after deamination of methylated cytosines [26–28]. This loss, however, is non-uniform because certain regions are 'CpG-rich' and display the expected frequency of CpGs. They are referred to as CpG islands [29,30] and represent a large fraction of *cis*-regulatory sequence because $\sim 60\%$ of all mammalian gene promoters are CpG-rich [26,31,32]. In addition, several studies indicate that many non-promoter CpG islands probably serve an important regulatory function as distal regulators such as insulators and enhancers [33]. The localized depletion of CpGs results in a characteristic bimodal distribution of CpGs across vertebrate genomes (Figure 1). In invertebrates, DNA methylation is only present in some species, in which it occurs in mosaic patterns that are not genome-wide; consequently, no depletion or resulting bimodal CpG distribution is observed [34,35] (Figure 1b).

Recent genome-wide surveys revealed that DNA methylation at CpG-rich sequences is very low in stem cells [36–38]. During cellular differentiation, hypermethylation can occur at CpG island promoters and at CpG-rich sequences outside of promoter regions [37,38]. Remarkably, almost no demethylation is detected, indicating that DNA-methylation-mediated epigenetic repression increases during lineage-specification. Direct comparisons of differentiated cell types support this conclusion [28,39–41]. Notably, sequence-based detection methods such as microarrays and high-throughput sequencing, which have been employed for these studies, cannot comprehensively measure repetitive DNA, leaving the dynamics of DNA methylation at non-unique sequences an open question.

Many of the identified targets of differentiation-coupled *de novo* DNA methylation are promoters of stem-cell- and germline-specific genes [28,38,39]. One interpretation of this selectivity is that DNA methylation might stably repress the pluripotency program and prevent its aberrant reactivation and de-differentiation under physiological conditions. Experimental support for this model comes from a recent report showing that reprogramming of somatic cells into 'induced pluripotent' stem (iPS) cells is greatly enhanced upon treatment with the DNA methyltransferase inhibitor 5-aza-cytidine [42]. Collectively, the genetic and molecular data are compatible with a role for DNA methylation in the shut-down of pluripotency and, eventually, cellular specification. Nevertheless, formal proof for this model is still missing because the requirement of DNA methylation for development could reflect its

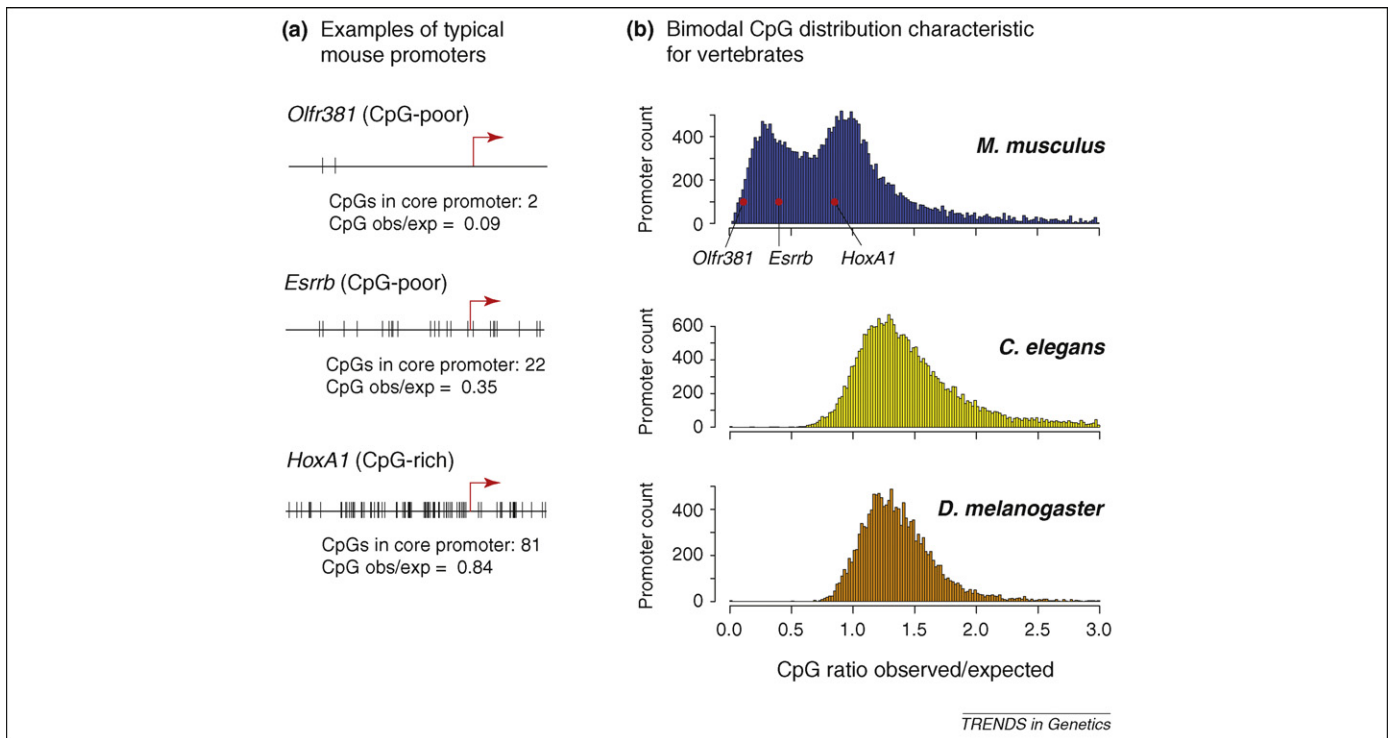


Figure 1. CpG distribution is markedly different in the genomes of vertebrates and invertebrates. Invertebrate genomes show homogenous distribution and no depletion of CpG dinucleotides. Vertebrate genomes, by contrast, are globally depleted for CpG dinucleotides, except at CpG-islands, which often mark regulatory elements such as enhancers and promoters. **(a)** Representative examples of mouse promoters with up to 40-fold different abundance of CpG dinucleotides. CpG-poor: olfactory receptor 381 (*Olf381*) and estrogen-related receptor β (*Esrrb*); CpG-rich: *HoxA1*. In each case, a 1.3-kb window around the transcription start site (red arrow) is shown. Each vertical line represents an individual CpG dinucleotide. **(b)** CpG dinucleotide distribution at promoter sequences in a vertebrate (*Mus musculus*) and two invertebrate (*C. elegans* and *D. melanogaster*) genomes. For each genome, we calculated the CpG content of annotated transcription start sites and calculated the relative abundance as the ratio of observed CpGs versus the number of CpGs expected based on sequence composition assuming equal abundance of all bases in a 1.3-kb window. The resulting ratio is plotted as a histogram for all promoters. Vertebrates (here, mouse) show a bimodal distribution of promoter CpGs with two major peaks corresponding to CpG-poor (left peak) and CpG-rich (right peak) promoters. Invertebrates, illustrated by *D. melanogaster* and *C. elegans*, do not show such a distribution; instead, the major peak of CpG content is close to 1, indicating that invertebrates contain the expected number of CpGs in their promoters.

function in repeat inactivation [24] and maintenance of differentiated states rather than their establishment.

Profound differences in the regulation of CpG-rich versus CpG-poor promoters

In vertebrates, CpG-rich and CpG-poor promoters differ not only in sequence and DNA methylation but also in the spatial precision of transcriptional initiation and in their chromatin states. Massively parallel Cap-analysis gene expression (CAGE)-tag sequencing revealed that transcription initiates at defined nucleotide positions in CpG-poor promoters, which mostly rely on the initiation factor TATA-box-binding protein (TBP). At CpG-rich promoters, transcriptional start sites (TSS) are loosely defined and initiation can occur in a region spanning 10–60 bp [32,43]. This finding seems to contrast with the general precision of initiation reported for *in vitro* systems, however, these biochemical assays were performed almost exclusively with TBP-dependent CpG-poor promoters. The underlying difference(s) in TSS definition and regulation remains elusive, but a hint comes from genome-wide profiles of histone modifications (see later).

H3K4 methylation at CpG islands

Several recent studies show a differential distribution of active chromatin marks, such as methylation of Lys4 of histone H3 (H3K4me), at CpG-rich versus CpG-poor pro-

moters. H3K4me creates a chromatin signal that is recognized by a large number of multiprotein complexes and which has been implicated in many gene activation pathways [44,45]. CpG-poor promoters harbor di- or tri-methylated H3K4 only when the gene is actively transcribed. This is reminiscent of the situation in invertebrates in which promoter H3K4 methylation reflects the active state [46,47]. By contrast, CpG-rich promoters behave differently: they are methylated at H3K4 constitutively and independently of the transcriptional activity of the corresponding gene [28,38,48,49]. Furthermore, a study in T cells showed that CpG-rich promoters are hyperacetylated at H3 in a transcription-independent manner [50]. Interestingly, low levels of RNA polymerase (Pol) II levels can be detected at many of these inactive CpG islands [49], raising the question of whether active histone modifications are a cause or consequence of this low and inefficient Pol II recruitment. It is conceivable that these promoters, albeit inactive, reside in an 'open', transcriptionally permissive environment, which leads to occasional Pol II binding but does not enable productive elongation [51]. This state, however, requires the absence of DNA methylation [28,36–38], implying that unmethylated CpG-rich elements are recognized by *trans*-acting factors that mediate the unique chromatin state of CpG islands (Figure 2). Such factors could include chromatin modifying enzymes (e.g. mixed-lineage leukemia [MLL] H3K4 methyltransfer-

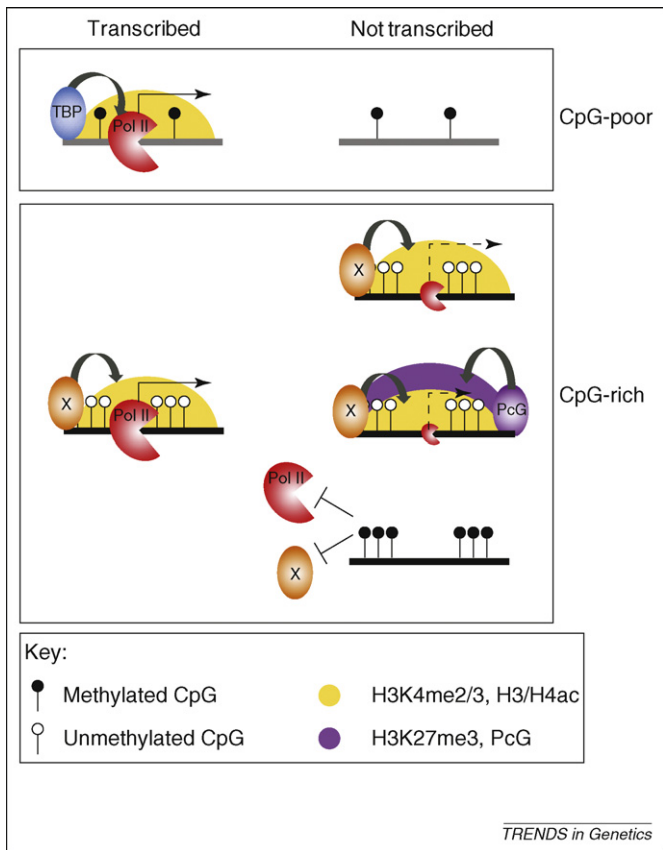


Figure 2. CpG-poor and CpG-rich promoters are differentially regulated, which is reflected in their chromatin configuration in the transcriptionally inactive state. CpG-poor promoters, which are often regulated via TBP (blue)-dependent pathways, only carry H3K4me and H3 and H4 acetylation (H3/H4ac) marks (yellow) when actively transcribed by Pol II (red). CpG-rich promoters are mostly DNA unmethylated (white lollipops), decorated by H3K4me and H3/H4ac and display low levels of Pol II even when inactive. This state could be mediated by proteins that recognize unmethylated CpG motifs, such as CXXC domain proteins (X, orange). When repressed by PcG proteins (purple), the active environment persists, indicating transient repression. Upon DNA methylation (black lollipops) of CpG-rich promoters, active histone modifications and Pol II can no longer be detected, indicating more stable silencing.

ases) containing CXXC domains, which bind preferentially to unmethylated CpGs [52]. Moreover, this ‘active’ chromatin environment at CpG-rich sequences could protect regulatory elements against DNA methylation and a resulting loss of ‘accessibility’. Direct evidence for H3K4 methylation protection against DNA methylation comes from a structural analysis of Dnmt3L, a germline-specific co-factor essential for *de novo* methylation of imprinting control regions [53,54]. Dnmt3L in complex with the *de novo* methyltransferase Dnmt3a can only bind nucleosomes that are unmodified at H3K4, whereas H3K4 methylation blocks this interaction and prevents DNA methylation. Hence, it is tempting to speculate that an H3K4-dependent pathway also operates during stem-cell-differentiation-coupled *de novo* methylation. By definition, *de novo* methylation can only occur at unmethylated sequences, which, at large, are the CpG-rich sequences in the genome [21,28,36–38,40]. The fact that H3K4 methylation and DNA methylation are mutually exclusive [28] predicts that *de novo* methylation of CpG-rich sequences coincides with a loss of H3K4 methylation, which is indeed the case during stem-cell differentiation [37,38].

Only a small fraction of CpG-rich sequences is *de novo* DNA methylated during cellular differentiation, hence, the majority of CpG-rich regulatory regions in mammalian genomes are accessible at any developmental stage. By contrast, the non-regulatory CpG-poor part of the genome would reside in a less accessible chromatin environment, which in turn might reduce the binding of transcription factors to randomly occurring sites.

In support of the concept that DNA sequence features are sufficient to establish an accessible chromatin state, a recent study in mosaic mice carrying human DNA indicated that homologous transcription factors can establish tissue-specific transcription and H3K4 methylation in closely related species [55].

Lessons from Polycomb

PcG-mediated gene repression regulates gene transcription during development. Originally discovered in *Drosophila melanogaster* as a system that controls *Hox* gene expression for correct body patterning [56,57], Polycomb has a broad regulatory potential in mouse and human ES cells because it targets many developmental transcription factors [58,59]. PcG-mediated repression entails histone H3 Lys27 trimethylation (H3K27me3), which is set by the Polycomb repressive complex 2 (PRC2) [56,57]. Polycomb repression can be overcome upon gene activation by specific stimuli once pluripotent cells are induced to differentiate, whereas non-induced Polycomb targets maintain H3K27me3 and PRC2 occupancy [58]. These data led to the model that Polycomb targets are specified early in development and continue to be repressed unless activated. However, this view has been challenged by a series of recent studies that reveal additional, non-stem-cell-specified Polycomb targets in various primary and transformed mammalian cell types [38,60–66]. For example, in neuronal progenitors, genes required for further developmental fates comprise differentiation-specific PcG targets [38]. Together, these findings support a model in which Polycomb repression could act not only in pluripotent stem cells to ensure proper lineage choice, but also in progenitor cells to guide their further developmental potential by ensuring proper regulation of subtype-specific genes. These findings further exemplify that stem cells are not necessarily unique with regard to the epigenetic mechanisms they employ but, rather, are unique in the genomic targets of these pathways. Cell-type-specific recruitment is probably mediated by transcription factors. For example, many targets of Oct4, a key transcription factor in the control of stem-cell pluripotency, carry the repressive H3K27me3 mark, pointing to a mechanism by which PcG proteins can be recruited to cell-type-specific targets [58]. Moreover, in cancer cells, PcG proteins are recruited by the transcription factor Snail1 to silence E-cadherin expression, which often correlates with increasing invasiveness and malignancy of the cancer owing to detachment from the extracellular matrix upon E-cadherin loss [67].

Several recent reports showed that H3K27me3 in mammalian cells is largely confined to CpG-rich sequences [33,38,62,68]. This is unexpected because the majority of tissue-specific genes, the *bona fide* targets of Polycomb

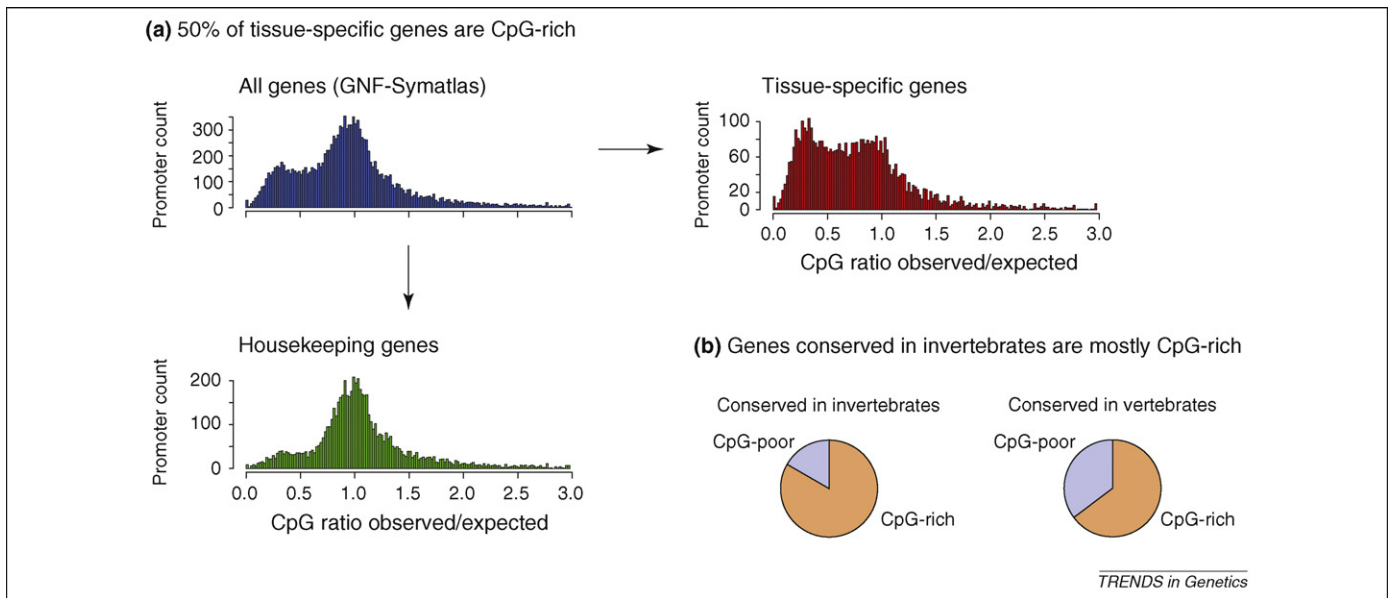


Figure 3. Genes can be grouped as housekeeping or tissue-specific based on a broad or restricted activity in different tissues. Most housekeeping genes are under the control of CpG island promoters. Nevertheless, many CpG island promoters control tissue-specific genes, which tend to regulate ancestral genes. **(a)** Gene expression data for 61 mouse tissues were retrieved from SymAtlas (<http://symatlas.gnf.org>) [87] and CpG content was determined for all promoters in the dataset ($n = 13\,729$) as described in Figure 1. A histogram is shown of the CpG ratios for all genes, including both tissue-specific and broadly expressed housekeeping genes. All genes display a similar bimodal distribution (top left) as shown in Figure 1 [69]. As expected, ubiquitously expressed ‘housekeeping’ genes (expression value >200 in >50 tissues out of 61; $n = 5481$) are mostly CpG-rich. However, tissue-specific genes (expression value >200 in <11 tissues out of 61; $n = 5294$) show almost equal numbers of CpG-rich and -poor promoters. These CpG-rich tissue-specific genes are preferably regulated by DNA methylation and Polycomb, whereas CpG-poor promoters seem not to require active repression. **(b)** Interestingly, ‘old’ genes (i.e. conserved from invertebrates) are mostly CpG-rich, whereas ‘newer’ genes, which arose in the vertebrate lineage, are more often controlled by CpG-poor promoters.

regulation, are under the control of CpG-poor promoters [69] (Figure 3). This surprising preference towards CpG-rich elements indicates a sequence contribution to the yet-to-be-determined targeting mechanism of the Polycomb machinery (see later). It furthermore provides a DNA-sequence-based explanation for the existence of so-called ‘bivalent’ chromatin domains, which harbor both H3K4 methylation and H3K27 methylation. As outlined earlier, CpG-rich promoters are ubiquitously marked by H3K4me. Thus, any CpG-rich sequence targeted by Polycomb will consequently form a bivalent domain carrying both the ‘activating’ H3K4 and the ‘repressive’ H3K27 methylation mark [68]. This model is compatible with the finding that bivalent domains are not a unique feature of pluripotent cells but, instead, are also present in differentiated cell types and can even form *de novo* during cellular differentiation [38,61–64]. In agreement with a CpG island explanation for bivalent chromatin, both marks are mutually exclusive in *D. melanogaster* [70], the genome of which does not contain global DNA methylation or CpG islands.

It remains to be tested if some aspects of the regulatory functions of Polycomb are unique to mammalian (and probably vertebrate) promoters, which display bivalency. One possibility is that, in addition to the function in invertebrates, Polycomb repression could be used to reduce partially intrinsic transcriptional noise at selected CpG island promoters because these reside in an ‘accessible’ chromatin state with detectable levels of Pol II even when inactive (Figure 2 and Box 1).

Control of CpG-poor promoters

In vertebrates, CpG-poor promoters are largely DNA methylated independently of activity state. When they

are not active, they are neither H3K4 methylated nor bound by Pol II [38,49,62]. Thus, they are reminiscent of inactive promoters in lower eukaryotes such as yeast, in which nucleosomal packaging and the resulting decreased sequence accessibility seem sufficient to mediate a stable off-state [10]. Similarly to the general situation in yeast, and unlike CpG-rich promoters, CpG-poor promoters contain very precise start sites that are set by sequence-specific activators (see earlier). Interestingly, Orford and colleagues reported a set of CpG-poor promoters in hematopoietic progenitor cells (erythroid-myeloid-progenitors [EML]), which show H3K4 dimethylation (H3K4me₂) at lineage-specific targets in the absence of transcription [71]. These EML-specific H3K4me₂-marked promoters control several hematopoietic genes, which are expressed in either erythroid or myeloid lineages. Upon stimulation of EML cells towards erythroid differentiation, H3K4me₂ positive erythroid genes are activated and retain H3K4me₂, whereas myeloid-specific genes remain silent but lose the H3K4me₂ mark. Hence, in this system, H3K4 methylation seems to mark lineage-specific genes in progenitor cells for later activation. In the absence of transcription, H3K4 methylation might lead to a more accessible chromatin environment at these CpG-poor promoters and it remains to be tested whether they are also occupied by Pol II or Polycomb, similarly to the situation at CpG-rich promoters.

Despite this example, ‘open’ chromatin is usually absent at inactive CpG-poor promoters. This finding might explain why they tend to rely less frequently on additional epigenetic repression, illustrated by the marked bias of Polycomb binding for CpG-rich promoters [33,68]. At the same time, H3K9 methylation and heterochromatin

Box 1. Is lineage choice driven by transcriptional oscillations from CpG-rich promoters?

Most evidence in mammalian systems and derived models indicates that tight control of gene 'on' and 'off' states defines cell identity. In many prokaryotic systems, however, stochastic gene expression and oscillations of transcription levels seem to function as triggers or switches in decision making [81]. Such events also occur in mammalian cells, for example, in the stochastic expression of odorant receptors in the mouse sensory neurons [82]. A recent report showed that clonal populations of mouse hematopoietic progenitor cells sorted for either high or low expression of the stem-cell marker Sca-1 can reconstitute the full parental spectrum of Sca-1 expression within a few days after isolation [83]. Notably, not only is Sca-1 expression different, but the entire expression program is markedly altered between the sorted populations and reverts back to the average of the unsorted pool of hematopoietic progenitors within a few days. Cells expressing low Sca-1 levels are prone to erythroid differentiation, whereas cells expressing high Sca-1 levels are more prone to differentiate into a myeloid lineage, indicating that these apparently random fluctuations have functional implications [83]. Interestingly, recent data showed that ES cells are also more heterogeneous than previously thought and seem to oscillate between different states. For example, Nanog, a transcription factor important for maintenance of pluripotency, and Stella, a protein specifically expressed in pre-implantation embryos and the germ line, show fluctuating expression in ES cells, which has an impact on the differentiation potential of the cells [84,85]. On a related note, caudal type homeobox 2 (Cdx2) and Eomes, two key transcription factors of the trophoblast lineage, show stochastic expression in the early blastocyst, which eventually is sufficient to promote expression of the Elf5 transcription factor and induction of trophoblast differentiation [86]. However, in the inner cell mass (ICM) and in ES cells, the *Elf5* promoter is DNA methylated and the sporadic activity of Cdx2 and Eomes is not sufficient to activate *Elf5*. Consequently, the ICM remains pluripotent and will give rise to all cells of the embryo but will not contribute to trophectoderm [86]. These examples illustrate how stochastic gene expression, in combination with epigenetic pathways, can contribute to cell-fate decision switches. Moreover, it is tempting to speculate that these random fluctuations are facilitated by promoter structure and, thus, mostly occur at CpG-rich promoters, which seem to be less stringently controlled than CpG-poor promoters. Along the same line, it will be interesting to determine if other genes show similar behavior and how this relates to decision making *in vivo*.

protein 1 (HP1), which binds H3K9 methylated nucleosomes, seem to confer repression to some CpG-poor promoters [62,72]. These H3K9 targets are often members of large gene families such as olfactory receptors and Krüppel-associated box (KRAB)-zinc finger genes, which are clustered in the genome. Their repetitive nature might explain the recruitment of H3K9 and HP1, which have a crucial role in silencing centromeric and intergenic repeats at constitutive heterochromatin [8]. Remarkably, H3K9 methylation does not occur at CpG-rich promoters [62], indicating that H3K9 and H3K27 methylation function at separate targets that differ in sequence composition.

Promoter structure and epigenetic regulation

Transient and more stable forms of epigenetic repression are targeted to particular promoter classes based upon CpG content. The observed bias of tissue-specific DNA methylation and Polycomb repression toward CpG islands seems surprising because CpG islands are thought to mostly regulate housekeeping genes [43,69]. This, however, is an oversimplification. Several thousand CpG-rich promoters control genes that are expressed in a tissue-

specific manner (Figure 3). These promoters reside in accessible chromatin, as indicated by H3K4 methylation, and low, but detectable, Pol II levels (Figure 2). Their chromatin state might reflect a form of genome partitioning that maintains regulatory regions accessible and free of DNA methylation. As a consequence, this open structure might require active repression via chromatin and DNA methylation to stabilize gene 'off' states. The transcriptionally permissive environment at CpG-rich promoters could furthermore result in stochastic transcription from inactive or oscillatory expression at active promoters, which has been implicated in contributing to cell fate decisions (Box 1). Moreover, recent studies found divergent short sense and antisense transcripts at the transcription start sites of active genes, which might reflect less stringent control of transcription initiation owing to the accessible environment at CpG-rich promoters [73–76].

How did these remarkably different promoter classes evolve? To maintain high concentrations of CpG over evolutionary time, CpG islands must be unmethylated in the germline. At first glance, tissue-specific genes controlled by CpG islands tend to have arisen in invertebrates before the advent of global DNA methylation. For example, nearly all *Hox* genes are under the control of CpG island promoters. Conversely, vertebrate-specific classes of genes, such as those encoding immunoglobulins or olfactory receptors, tend to be under the control of CpG-poor promoters (Figure 3). It thus seems plausible that promoter classes primarily reflect the evolutionary history of the genes they control rather than their precise biological function. Further work is required to better understand the evolution of *cis*-regulatory regions; however, CpG-rich and CpG-poor promoters seem to be regulated differently, not only at the level of DNA methylation but also in their selective use of the Polycomb system.

Concluding remarks and future perspectives

Stem-cell-based differentiation models have proven informative for studying the cellular changes that accompany the loss of pluripotency and terminal differentiation. Because these systems can be genetically modified and generate pure populations of primary, non-transformed cells with defined developmental potential and function, they will continue to provide important information. Indeed, in this context, the application of modern genomics and proteomics tools can be used to identify regulatory principles that can be tested *in vivo*. Such approaches will also guide our understanding of epigenetic gene regulation; however, limitations of these systems should not be ignored. Rapidly dividing stem cells are often compared with post-mitotic differentiated cells bearing the risk that 'stem-cell-specific' characteristics are actually general features of mitotically dividing cells. In addition, current *in vitro* stem-cell systems do not enable the study of asymmetric cell division, a key feature of stem-cell maintenance *in vivo*, nor the testing of potential asymmetric segregation of epigenetic information on chromatin.

Equally importantly, most current assays that map sites of epigenetic modifications, such as chromatin immunoprecipitation (ChIP) coupled to microarrays or massive parallel sequencing, cannot account for heterogeneity

within pools of cells. New developments that enable ChIP to be performed on few cells [77] or high-throughput sequencing of bisulfite-converted DNA [37,78] will increase resolution and enable the identification of differences in populations of cells as in rare primary cells, for example, in the early embryo. Another important question for the future concerns the role of non-coding RNAs in differentiation processes; indeed, these RNAs can target sites of chromatin changes in several organisms, including vertebrates, and help to regulate DNA methylation in plants [79].

In this review, we have highlighted the interplay between promoter sequence features and chromatin and DNA modifications. A logical next step will be to identify the exact contribution of a given DNA sequence motif to chromatin-regulatory processes. Such information might provide a key to understanding targeting pathways and the propagation of epigenetic states and potentially link them to extracellular signaling pathways that are crucial for stem-cell function *in vivo* [80].

Acknowledgements

We thank Marc Bühler, Antoine Peters, Susan Gasser and members of the Schubeler laboratory for critical reading of the manuscript and helpful comments. Furthermore, we apologize to colleagues whose work could not be cited owing to space limitations. Research in the laboratory of D.S. is supported by the Novartis Research Foundation and the European Union (LSHG-CT-2004-503433 and LSHG-CT-2006-037415). F.M. is supported by a pre-doctoral fellowship from the Boehringer Ingelheim Fonds.

References

- Reik, W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447, 425–432
- Bernstein, B.E. *et al.* (2007) The mammalian epigenome. *Cell* 128, 669–681
- Cairns, B.R. (2005) Chromatin remodeling complexes: strength in diversity, precision through specialization. *Curr. Opin. Genet. Dev.* 15, 185–190
- Ptashne, M. (2007) On the use of the word ‘epigenetic’. *Curr. Biol.* 17, R233–R236
- Goll, M.G. and Bestor, T.H. (2005) Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.* 74, 481–514
- Groth, A. *et al.* (2007) Chromatin challenges during DNA replication and repair. *Cell* 128, 721–733
- Hansen, K.H. *et al.* (2008) A model for transmission of the H3K27me3 epigenetic mark. *Nat. Cell Biol.* 10, 1291–1300
- Peters, A.H. and Schubeler, D. (2005) Methylation of histones: playing memory with DNA. *Curr. Opin. Cell Biol.* 17, 230–238
- Bird, A. (2007) Perceptions of epigenetics. *Nature* 447, 396–398
- Struhl, K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* 98, 1–4
- Workman, J.L. and Kingston, R.E. (1998) Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.* 67, 545–579
- Lam, F.H. *et al.* (2008) Chromatin decouples promoter threshold from dynamic range. *Nature* 453, 246–250
- Shogren-Knaak, M. *et al.* (2006) Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* 311, 844–847
- Lanzuolo, C. *et al.* (2007) Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nat. Cell Biol.* 9, 1167–1174
- Terranova, R. *et al.* (2008) Polycomb group proteins Ezh2 and Rnf2 direct genomic contraction and imprinted repression in early mouse embryos. *Dev. Cell* 15, 668–679
- Gregory, T.R. (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev. Camb. Philos. Soc.* 76, 65–101
- Bestor, T.H. (2003) Cytosine methylation mediates sexual conflict. *Trends Genet.* 19, 185–190
- Waterston, R.H. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562
- Wray, G.A. *et al.* (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20, 1377–1419
- Bilu, Y. and Barkai, N. (2005) The design of transcription-factor binding sites is affected by combinatorial regulation. *Genome Biol.* 6, R103
- Rollins, R.A. *et al.* (2006) Large-scale structure of genomic methylation patterns. *Genome Res.* 16, 157–163
- Bird, A.P. (1995) Gene number, noise reduction and biological complexity. *Trends Genet.* 11, 94–100
- Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.* 16, 6–21
- Walsh, C.P. *et al.* (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* 20, 116–117
- Whitcomb, S.J. *et al.* (2007) Polycomb Group proteins: an evolutionary perspective. *Trends Genet.* 23, 494–502
- Antequera, F. and Bird, A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. U. S. A.* 90, 11995–11999
- Bestor, T.H. and Coxon, A. (1993) Cytosine methylation: the pros and cons of DNA methylation. *Curr. Biol.* 3, 384–386
- Weber, M. *et al.* (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* 39, 457–466
- Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261–282
- Takai, D. and Jones, P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. U. S. A.* 99, 3740–3745
- Ioshikhes, I.P. and Zhang, M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nat. Genet.* 26, 61–63
- Bajic, V.B. *et al.* (2006) Mice and men: their promoter properties. *PLoS Genet.* 2, e54
- Tanay, A. *et al.* (2007) Hyperconserved CpG domains underlie Polycomb-binding sites. *Proc. Natl. Acad. Sci. U. S. A.* 104, 5521–5526
- Glass, J.L. *et al.* (2007) CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res.* 35, 6798–6807
- Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Natl. Rev.* 9, 465–476
- Fouse, S.D. *et al.* (2008) Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell* 2, 160–169
- Meissner, A. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454, 766–770
- Mohn, F. *et al.* (2008) Lineage-specific polycomb targets and *de novo* DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell* 30, 755–766
- Farthing, C.R. *et al.* (2008) Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. *PLoS Genet.* 4, e1000116
- Illingworth, R. *et al.* (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* 6, e22
- Eckhardt, F. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* 38, 1378–1385
- Mikkelsen, T.S. *et al.* (2008) Dissecting direct reprogramming through integrative genomic analysis. *Nature* 454, 49–55
- Carninci, P. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38, 626–635
- Sims, R.J., 3rd and Reinberg, D. (2008) Is there a code embedded in proteins that is based on post-translational modifications? *Nat. Rev. Mol. Cell Biol.* 9, 815–820
- Taverna, S.D. *et al.* (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.* 14, 1025–1040
- Pokholok, D.K. *et al.* (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* 122, 517–527
- Schubeler, D. *et al.* (2004) The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* 18, 1263–1271

- 48 Barrera, L.O. *et al.* (2008) Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res.* 18, 46–59
- 49 Guenther, M.G. *et al.* (2007) A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130, 77–88
- 50 Roh, T.Y. *et al.* (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* 19, 542–552
- 51 Lorincz, M.C. and Schubeler, D. (2007) RNA polymerase II: just stopping by. *Cell* 130, 16–18
- 52 Voo, K.S. *et al.* (2000) Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol. Cell. Biol.* 20, 2108–2121
- 53 Jia, D. *et al.* (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for *de novo* DNA methylation. *Nature* 449, 248–251
- 54 Ooi, S.K. *et al.* (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to *de novo* methylation of DNA. *Nature* 448, 714–717
- 55 Wilson, M.D. *et al.* (2008) Species-specific transcription in mice carrying human chromosome 21. *Science* 322, 434–438
- 56 Schwartz, Y.B. and Pirrotta, V. (2007) Polycomb silencing mechanisms and the management of genomic programmes. *Natl. Rev.* 8, 9–22
- 57 Schuettengruber, B. *et al.* (2007) Genome regulation by polycomb and trithorax proteins. *Cell* 128, 735–745
- 58 Boyer, L.A. *et al.* (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 441, 349–353
- 59 Lee, T.I. *et al.* (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125, 301–313
- 60 Barski, A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837
- 61 Bracken, A.P. *et al.* (2006) Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev.* 20, 1123–1136
- 62 Mikkelsen, T.S. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560
- 63 Pan, G. *et al.* (2007) Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* 1, 299–312
- 64 Pasini, D. *et al.* (2007) The polycomb group protein Suz12 is required for embryonic stem cell differentiation. *Mol. Cell. Biol.* 27, 3769–3779
- 65 Roh, T.Y. *et al.* (2006) The genomic landscape of histone modifications in human T cells. *Proc. Natl. Acad. Sci. U. S. A.* 103, 15782–15787
- 66 Squazzo, S.L. *et al.* (2006) Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res.* 16, 890–900
- 67 Herranz, N. *et al.* (2008) Polycomb complex 2 is required for E-cadherin repression by the Snail1 transcription factor. *Mol. Cell. Biol.* 28, 4772–4781
- 68 Bernstein, B.E. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125, 315–326
- 69 Schug, J. *et al.* (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* 6, R33
- 70 Tolhuis, B. *et al.* (2006) Genome-wide profiling of PRC1 and PRC2 polycomb chromatin binding in *Drosophila melanogaster*. *Nat. Genet.* 38, 694–699
- 71 Orford, K. *et al.* (2008) Differential H3K4 methylation identifies developmentally poised hematopoietic genes. *Dev. Cell* 14, 798–809
- 72 Vogel, M.J. *et al.* (2006) Human heterochromatin proteins form large domains containing KRAB-ZNF genes. *Genome Res.* 16, 1493–1504
- 73 Core, L.J. *et al.* (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322, 1845–1848
- 74 He, Y. *et al.* (2008) The antisense transcriptomes of human cells. *Science* 322, 1855–1857
- 75 Preker, P. *et al.* (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 322, 1851–1854
- 76 Seila, A.C. *et al.* (2008) Divergent transcription from active promoters. *Science* 322, 1849–1851
- 77 O'Neill, L.P. *et al.* (2006) Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. *Nat. Genet.* 38, 835–841
- 78 Cokus, S.J. *et al.* (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452, 215–219
- 79 Zaratiegui, M. *et al.* (2007) Noncoding RNAs and gene silencing. *Cell* 128, 763–776
- 80 Moore, K.A. and Lemischka, I.R. (2006) Stem cells and their niches. *Science* 311, 1880–1885
- 81 Raj, A. and van Oudenaarden, A. (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135, 216–226
- 82 Vassar, R. *et al.* (1993) Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell* 74, 309–318
- 83 Chang, H.H. *et al.* (2008) Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* 453, 544–547
- 84 Chambers, I. *et al.* (2007) Nanog safeguards pluripotency and mediates germline development. *Nature* 450, 1230–1234
- 85 Hayashi, K. *et al.* (2008) Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell* 3, 391–401
- 86 Ng, R.K. *et al.* (2008) Epigenetic restriction of embryonic cell lineage fate by methylation of Elf5. *Nat. Cell Biol.* 10, 1280–1290
- 87 Su, A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067