Defining an epigenetic code

Bryan M. Turner

The nucleosome surface is decorated with an array of enzyme-catalysed modifications on histone tails. These modifications have well-defined roles in a variety of ongoing chromatin functions, often by acting as receptors for non-histone proteins, but their longer-term effects are less clear. Here, an attempt is made to define how histone modifications operate as part of a predictive and heritable epigenetic code that specifies patterns of gene expression through differentiation and development.

The nucleosome is the fundamental unit of chromatin structure in all eukaryotes. It comprises a core of eight histones (two H2A, H2B, H3 and H4 histones) around which 147 base pairs of DNA are wrapped in 1.75 superhelical turns[1]. Given the intimate association between histones and DNA, it is not surprising that histones influence almost every aspect of DNA function. In some cases they are influential just by their presence — for example by hiding or revealing transcription factor binding sites or influencing polymerase progression. In other cases their effects can be more subtle and can depend on chemical modification of specific histone amino acids. The amino-terminal tails of all eight core histones protrude through the DNA and are exposed on the nucleosome surface, where they are subject to an enormous range of enzyme-catalysed modifications of specific amino-acid side chains[2,3], include acetylation of lysines, methylation of lysines and arginines, and phosphorylation of serines and threonines. The modifications decorate the nucleosome surface with an array of chemical information. Some years ago it was suggested that specific modifications may act as signalling receptors on the nucleosome surface that would be recognised by specific binding proteins, which would then, in turn, exert an effect on chromatin structure and function[4]. Work in many laboratories over the past ten years or so has confirmed this suggestion, both by identifying proteins that bind selectively to modified histones and by linking this binding to functional outcomes[2,3].

Bryan M. Turner is in the Institute of Biomedical Research, University of Birmingham Medical School, Birmingham B15 2TT, UK.
e-mail: b.m.turner@bham.ac.uk

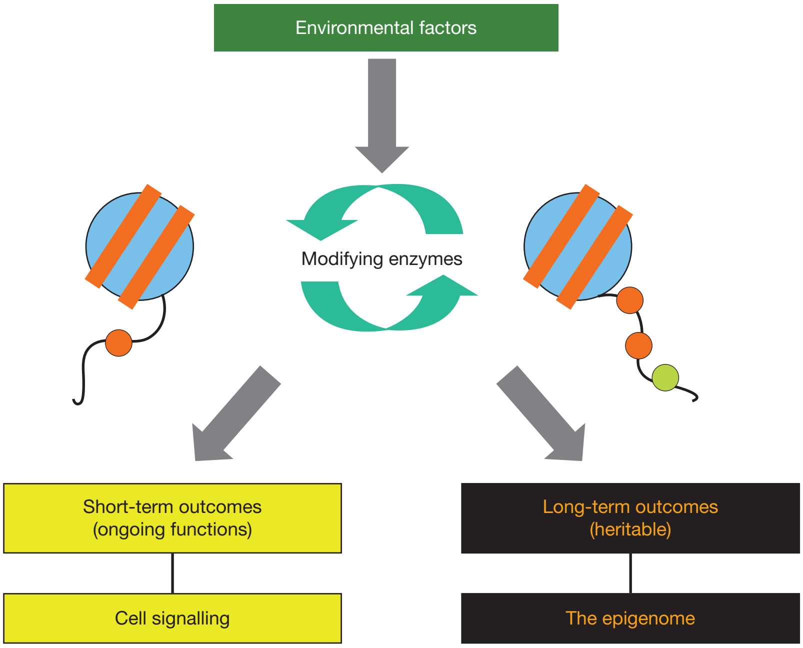

Figure 1 Histone modifications can generate both short-term and long-term outcomes. Histone tail modifications are put in place by modifying and demodifying enzymes, whose activities can be modulated by environmental and intrinsic signals. Modifications may function in both short-term, ongoing processes (such as transcription, DNA replication and repair) and in more long-term functions (as determinants of chromatin conformation, for example, heterochromatin formation, or as heritable markers that both predict and are necessary for, future changes in transcription). Short-term modifications are transient and show rapidly fluctuating levels. Long-term, heritable modifications need not necessarily be static: in theory they could still show enzyme-catalysed turnover, but the steady-state level must be relatively consistent.

Histone modifications have both short-term and long-term functional effects

Histone modifications are functionally linked to a variety of processes that are continuously occurring within the cell — for example, transcriptionally active promoters show an overall increase in acetylation of core histones and a more selective increase in methylation at particular lysines and arginines[3]. Patterns of histone modification associated with ongoing transcription can change rapidly and cyclically in response to external stimulation[5,6]. In this context, histone modifications can be considered the endpoints, on chromatin, of cellular signalling pathways and a mechanism through which the genome can respond to

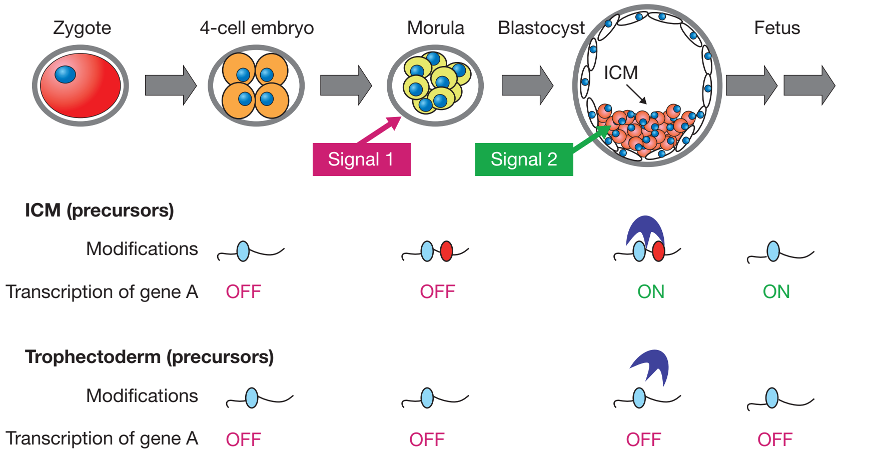

Figure 2 A hypothetical illustration of how an epigenetic code may work in the pre-implantation embryo. In response to an environmental or intrinsic signal at the morula stage, a histone modification (red oval) occurs on nucleosomes packaging a regulatory region of gene A, in those cells destined to form the inner cell mass (ICM) of the blastocyst. Together with a pre-existing modification (blue oval) this generates a distinctive epigenetic 'sign'. There is no immediate change in transcription of gene A, but the sign is passed on to successive cell generations, in the absence of the initial signal. In the ICM, receipt of a second signal initiates transcription of gene A through attachment of a binding protein to the sign (histone modifications) put in place at the morula stage. Once transcription of gene A is underway, it can be stabilised (maintained) by other 'memory' mechanisms[25] and the original sign need not be retained.

environmental stimuli[7]. To allow such responses, it is likely that the modifications themselves are rapidly turning over[5].

Histones can also exert longer-term effects on genomic function, largely by defining and maintaining chromatin structures throughout the cell cycle, or from one cell generation to the next (Fig. 1). To date, the long-term effects of histone modifications have mainly been examined through studies of heterochromatin. Constitutive heterochromatin is gene poor, rich in certain families of DNA repeats, largely transcriptionally silent and is marked by a characteristic array of histone modifications, including enrichment in H3 trimethylated at Lys 9 (H3K9me3), H4 trimethylated at Lys 20 (H4K20me3) and depletion in overall histone acetylation. H3K9me3 acts as a binding site (receptor) for the heterochromatin-forming protein HP1. Crucially, HP1 binding can be diminished or abolished by modification of residues adjacent to Lys 9, namely phosphorylation of Ser 10 and/or acetylation of Lys 14 (refs 8, 9), illustrating how modifications can interact to modulate functional effects. Modifications can distinguish large domains in heterochromatin and euchromatin, sometimes in conjunction with histone variants, and may have an 'indexing' function involved in large scale genome organization[10,11]. Facultative heterochromatin (such as the inactive X chromosome, Xi, in female mammals), is also marked by a characteristic array of histone modifications (for example, loss of H3K4 methylation, increased H3K27 methylation and increased levels of the variant macroH2A), together with selective DNA methylation and coating in *cis* with the non-coding *Xist* RNA[12]. Importantly, the silent state of Xi, once established, is extraordinarily stable and is faithfully passed on from one cell generation to the next throughout the life of the organism, exemplifying the heritability of chromatin states.

Defining a code: insights from semiotics

The increasing variety and interactive properties of histone modifications has led to the use of the terms 'histone code'[2] and 'epigenetic code'[3] — partly as a convenient shorthand to encompass the phenomenon of interacting histone modifications and partly as a means of indicating the important concept that specific combinations of modification may be linked to defined functional effects. Although this may seem reasonable, it is timely to ask whether the concept of a 'code' can appropriately or usefully be applied to protein–protein interactions that are often integral components of ongoing chromatin processes, such as transcription. I will argue that the use of the term in this context is not only inappropriate, but potentially misleading: an epigenetic code becomes a valuable hypothesis primarily in the context of possible long-term, heritable effects of histone modification, and here it is potentially very powerful indeed. However, the potential will only be realised if we decide precisely what we mean by, and therefore can expect from, such a code.

Semiotics is the study of signs and symbols and their use or meaning, and although primarily a tool for linguistic analysis, it can be applied to biological questions[13]. A semiotic system consists of just three components: a sign, its meaning (that is, the outcome it produces) and the code by which the sign can make its meaning known. In addition, in the systems discussed here, one or more adaptors are required to make a physical link between the sign and its outcome, that is, to 'read' the code[13]. To take a simple example, a red traffic light (the sign) will generally bring traffic to a standstill (the outcome). However, this will only happen because we have established a code that prescribes a red light as meaning stop and because the driver's brain (the adaptor) can read that code and carry out the appropriate actions. This simple system (summarized in Table 1) has two important properties that are relevant to our discussion and that exemplify two fundamental rules of semiotics: first, the code is arbitrary — we could just as well adopt a green light as the stop signal and, provided the Highway Code (or its equivalent outside the UK) was changed accordingly, the outcome would be the same; second, there is no direct, causal relationship between the sign and its outcome. For example, a pile of rocks in the road is certainly effective at stopping traffic, but does not constitute a semiotic system; it causes the outcome by itself without the need for a code or adaptors.

The genetic code meets both fundamental requirements of a semiotic system. Triplet codons on DNA and mRNA give rise to specified amino acids in proteins according to rules specified by the code and interpreted by transfer RNAs (tRNAs) bearing complementary anticodons (Table 1). There is no direct link between, for example, ACT in the DNA helix and serine in the protein. ACT is an arbitrary sign and its outcome (meaning) could be altered by generating a mutant tRNA carrying the appropriate anticodon. It is also noteworthy that DNA sequences that lead to nucleosome positioning, despite their undoubted importance, do not constitute a code that conforms to semiotic rules[14]. Such sequences, through their characteristic bending properties, directly favour nucleosome assembly — no code is needed, just biochemistry.

Table 1 Three semiotic systems

Sign	Number of signs	Code	Adaptor	Meaning	Number of meanings
Red/green traffic light	2	Highway code	Driver's brain	Traffic stops/goes	2
Triplet codon (chosen from 4 bases in DNA)	64 (4^3)	Genetic code	tRNA etc.	Amino acid in protein	20 (plus stop)
Combination of histone/DNA modifications	?	Epigenetic code	Modification-dependent binding proteins	Transcription starts or stops at a specified future time and place	?

A semiotic system consists of a sign, its meaning and the code used to interpret the sign. For the three examples given, an adaptor is needed to put the meaning into effect. It is proposed that the meanings specified by an epigenetic code are expression of defined genes, or sets of genes, at defined stages of development ('time' in the table) and in specific cell types ('place' in the table).

The genetic code as a paradigm

How can we identifying an epigenetic code that conforms to semiotic rules? I suggest that if we are serious in our search, we should aspire to define the epigenetic code with a clarity that comes close to that achieved for the genetic code more than 40 years ago. Francis Crick's minimalistic 1963 definition is a good starting point: "The genetic code describes the way in which a sequence of 20 or more things is determined by a sequence of four things of a different type"[15]. When this was written, both the chemical nature of the 'things' (nucleotides and amino acids) was known with certainty, as were the numbers involved, but the manner in which the code operated was still being worked out by technically demanding experiments whose interpretation was often controversial. There were competing hypotheses and the (putative) triplet codons specifying each amino acid had yet to be defined[16]. However, among the continuing uncertainty, Crick's precise definition of the problem, which makes no operational assumptions, served to guide the planning of experiments and their interpretation.

Perhaps the most crucial question at this stage is to decide what the epigenetic code is actually coding for. A heritable code linking chromatin modifications (the signs) with functionally relevant epigenetic outcomes (the meanings) must, in some sense, be a determinant of gene expression, but the signs involved should not, indeed cannot, be directly linked to ongoing transcription. Experiments that link specific patterns of histone acetylation to transcriptional activity, although valuable guides to transcriptional mechanisms, fail the first test of a semiotic system in that the modification is very likely directly involved in the outcome[17,18]. Acetylation is particularly problematic in that the neutralization of the positive charge of lysine residues inevitably reduces their DNA binding and may contribute directly to chromatin opening and indirectly to transcriptional upregulation. A code is not a useful or necessary concept in interpreting these experiments.

A much more promising coding problem is set when we consider the processes by which patterns of gene expression are put in place during differentiation and development. During development of multicellular organisms, cells progress through a series of stages during which they become increasingly specialised. Each specialization step is in response to intrinsic and extrinsic signals and involves characteristic changes in expression of key genes. Epigenetic mechanisms, almost by definition, underpin these expression changes, and it is here that an epigenetic code is potentially most valuable. I suggest that the code comprises combinations of chromatin modifications that allow the transcriptional status of specific genes to be switched (from on to off, or off to on), at a defined stage of development or differentiation. A key property is that the modifications are put in place before transcription of the target gene begins; they are neither contingent on, nor contemporary with, transcription. This is a truly predictive epigenetic code (Table 1). Using Crick's pared-down description of the genetic code as a template, I suggest the following definition: the epigenetic code describes the way in which the potential for expression of genes in a particular cell type is specified by chromatin modifications put in place at an earlier stage of differentiation.

Specifying signs and meanings

With regard to the epigenetic coding elements (signs), a reasonably proposition is that these signs comprise combinations of histone modifications, probably in association with DNA methylation, when available. As methylated CpGs can be recognised by a well-characterized family of meCpG-binding proteins, and given the central importance of DNA methylation in heritable epigenetic phenomena such as imprinting and X-inactivation, the involvement of DNA methylation in setting and reading an epigenetic code is likely if not inevitable, in those organisms in which it is present[19,20]. In contrast with the sequence-based genetic code, the number of modifications that constitute the coding units of the epigenetic code need not be fixed. The code must be combinatorial, but the number of modifications involved, or how they are configured, has yet to be established. In some cases, two or more specifically modified histone tails may be involved, and possibly involve more than one nucleosome[3,21].

The number of meanings (outcomes) specified by such a predictive code is likely to be large, but not unmanageably so. A typical mammal is made up of several hundred different cell types, most generated by progression down branching pathways of differentiation. The code would define expression of key genes in specified cell types and at specified stages of development. Guessing at possible numbers of outcomes that the code may specify is not useful, but what is certain is that the number will be smaller in a model organism such as *Caenorhabditis elegans*, with a limited number of cell types and developmental pathways, than it will be in mammals. What is also certain is that histone modifications, in combination, offer enough variety (coding potential) to deal with any likely number of outcomes.

Chromatin context and molecular mechanisms

The operation of predictive signs based on combinations of histone modifications is likely to be dependent on the chromatin context in which they find themselves, and this is likely to change through development. Factors such as chromatin condensation, nucleosome density, position within the nucleus or coating with non-coding RNAs can all alter the character of chromatin regions so as to exert a strong, and possibly overriding influence on gene expression[22,23,24]. Such mechanisms can set a context over large chromatin domains, or even whole chromosomes, and are themselves likely to be mediated by changes in histone modification. It is interesting to note that context matters even for the genetic code, for example, most AGT triplets in the genome do not result in a serine residue in a protein, either because they are in non-transcribed regions or they are not in a valid reading frame.

At this stage we do not know whether all, or just a selected subset, of the multitude of possible histone modifications are involved in a predictive and heritable epigenetic code. Some modifications may be exclusively concerned with ongoing processes (transcription, replication and repair), whereas others may set the context (that is, overall regional chromatin structure) within which the code operates. Yet others may function in 'cellular memory' — the mechanisms by which patterns of gene expression, once in place, are maintained from one cell generation to the next[25]. Given that histone modifications exerting these different functional effects may all occur together in the same chromatin domain, or even on the same nucleosome, distinguishing the different functional outcomes of each one presents a major experimental challenge.

Where to look for an epigenetic code

An epigenetic code, as defined here, will be detected by studying patterns of histone modification at key regions of defined genes before their expression during differentiation or development, and by showing that these patterns are predictive of, and necessary for, such expression. A possible, but entirely hypothetical, scenario is illustrated in Fig. 2. The proposed outcome is expression of gene A in cells of the inner cell mass (ICM) of the early blastocyst. This outcome is specified by a defined combination of histone modifications put in place at an earlier stage of development. In the model, the predictive mark is set at the morula stage in response to a specific environmental or intrinsic signal. It consists of a new modification (shown in red) alongside a pre-existing one (blue), which together make the predictive sign. The new modification is placed in those cells destined to form the ICM, but not in those destined to form the trophectoderm. The mechanism underpinning this selective response is not specified, but individual cells at the morula stage have lost the totipotency of the zygote and early cleavage stages, and are becoming restricted in their developmental options[26]. Differential responses to a common signal can reasonably be considered a consequence of these developmental changes. Crucially, the modification induces no change in transcription of gene A at the morula stage, but is stably transmitted through successive cell cycles as the embryo develops. At the blastocyst stage, in the ICM, the presence of the predictive mark allows gene A to respond to a second (common) environmental or intrinsic signal by enhanced transcription, possibly mediated by a specific binding protein. Such a combination of histone modifications predicting a future functional outcome and mediated by an adaptor molecule (binding protein) is a true code and broadly consistent with semiotic rules.

The early embryo is used for illustration in Fig. 2, but any differentiation pathway where changes in gene expression are well established provides a suitable model. A recent study of the mouse *λ5-VpreB1* locus, activated early in B-lymphocyte development, has identified a small (≈2 kb) intergenic region that is marked by high levels of H3 and H4 acetylation and H3K4 dimethylation in pluripotent ES cells, where the gene is not expressed[27,28]. These modifications spread across the *λ5-VpreB1* locus as cells progress towards the B-cell lineage (where the gene is switched on) and disappear from the gene altogether in non-lymphoid cells.

Homeotic genes are key regulators of development in multicellular organisms, and if it could be shown that histone modifications are both predictive of, and necessary for, their expression at precisely defined developmental stages, this would go some way towards confirming the likely existence of an epigenetic code. Some progress has already been made: for example, the promoter and exon 1 of the murine homeobox gene, *Hoxb9,* are marked by high levels of H3K9 acetylation and H3K4 trimethylation (both modifications associated with active genes) in undifferentiated mouse ES cells in which the gene is silent[29]. There is no increase in these modifications when the *Hoxb9* gene is switched on after 10 days of ES-cell differentiation (that is, they are not responsive to altered transcription). However, expression is preceded by changes in chromatin conformation and intranuclear positioning of the locus[29]. Recently, evidence has been presented to show that H3 trimethylated at Lys 4 is necessary for assembly of the nucleosome remodelling factor (NURF) chromatin-remodelling complex on the *Hoxc8* locus[30]. This modification functions by binding the NURF subunit, BPTF (bromodomain and PHD finger transcription factor), and depletion of BPTF causes abnormalities in the location of *Hoxc8* expression in *Xenopus* tadpoles[30]. Whether these changes are involved directly in triggering gene expression, or in setting a chromatin context, remains to be established.

Conclusion

Epigenetic mechanisms are responsible for putting in place and maintaining the patterns of gene expression that specify the many different cell types required to make a higher eukaryote. The mechanisms involved must meet two seemingly conflicting requirements: they must be sensitive to the intrinsic and environmental cues that specify when patterns of expression must switch during development, but be stable enough to allow transmission of established patterns across cell generations, even in the presence of variable, possibly hostile, environments. An epigenetic code based primarily on enzyme-catalysed histone modifications can meet both these requirements.

I have suggested that an essential prerequisite for defining an epigenetic code, or even deciding whether or not it exists, is to distinguish the different functional roles of histone modifications. Those that have a long-term, heritable (potentially coding) function must be distinguished from those concerned with short-term signalling processes, and from those that may function in determining the conformation, or intranuclear positioning, of chromatin domains (here, referred to as chromatin context). This presents a major experimental challenge.

A final trip down to the dusty journals in the basement of our library provided an appropriate ending, again from Francis Crick and referring to the early days of the genetic code: "These theories may not be correct but they … enable us to tighten up our logic and make us scrutinize the experimental evidence to some purpose…. In the final analysis it is the quality of the experimental work that will be decisive"[16]. □

ACKNOWLEDGMENTS
I am grateful to M. Barbieri for drawing my attention to the world of semiotics, and to A. Ferguson-Smith and colleagues in the Chromatin and Gene Expression Group for their thoughts, criticisms and insights.

1. Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389,** 251–260 (1997).
2. Margueron, R., Trojer, P. & Reinberg, D. The key to development: interpreting the histone code? *Curr. Opin. Genet. Dev.* **15,** 163–176 (2005).
3. Nightingale, K.P., O'Neill, L.P. & Turner, B.M. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr. Opin. Genet. Dev.* **16,** 125–136 (2006).
4. Turner, B.M., Birley, A.J. & Lavender, J. Histone H4 isoforms acetylated at specific lysine residues define individual chromosomes and chromatin domains in *Drosophila* polytene nuclei. *Cell* **69,** 375–384 (1992).
5. Hazzalin, C.A. & Mahadevan, L.C. Dynamic acetylation of all lysine 4-methylated histone H3 in the mouse nucleus: analysis at c-fos and c-jun. *PLoS Biol.* **3,** e393 (2005).

6. Metivier, R. *et al.* Estrogen receptor-α directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell* **115,** 751–763 (2003).
7. Schreiber, S. L. & Bernstein, B. E. Signaling network model of chromatin. *Cell* **111,** 771–778 (2002).
8. Fischle, W. *et al.* Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature* **438,** 1116–1122 (2005).
9. Mateescu, B., England, P., Halgand, F., Yaniv, M. & Muchardt, C. Tethering of HP1 proteins to chromatin is relieved by phosphoacetylation of histone H3. *EMBO Rep.* **5,** 490–496 (2004).
10. Hake, S. B. & Allis, C. D. Histone H3 variants and their potential role in indexing mammalian genomes: the "H3 barcode hypothesis". *Proc. Natl Acad. Sci. USA* **103,** 6428–6435 (2006).
11. Peters, A. H. *et al.* Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. *Mol. Cell* **12,** 1577–1589 (2003).
12. Heard, E. Delving into the diversity of facultative heterochromatin: the epigenetics of the inactive X chromosome. *Curr. Opin. Genet. Dev.* **15,** 482–489 (2005).
13. Barbieri, M. *The Organic Codes; An Introduction to Semantic Biology* (Cambridge University Press, Cambridge, 2003).
14. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442,** 772–778 (2006).
15. Crick, F. H. On the genetic code. *Science* **139,** 461–464 (1963).
16. Crick, F. H. The recent excitement in the coding problem. *Progress in Nucleic Acid Research* **1,** 163–217 (1963).
17. Dion, M. F., Altschuler, S. J., Wu, L. F. & Rando, O. J. Genomic characterization reveals a simple histone H4 acetylation code. *Proc. Natl Acad. Sci. USA* **102,** 5501–5506 (2005).
18. Liu, C. L. *et al.* Single-nucleosome mapping of histone modifications in *S. cerevisiae. PLoS Biol.* **3**, e328 (2005).
19. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16,** 6–21 (2002).
20. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genet* **33**, 245–254 (2003).
21. Sun, Z. W. & Allis, C. D. Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature* **418,** 104–108 (2002).
22. Wilson, C. B. & Merkenschlager, M. Chromatin structure and gene regulation in T cell development and function. *Curr. Opin. Immunol.* **18,** 143–151 (2006).
23. Gilbert, N., Gilchrist, S. & Bickmore, W. A. Chromatin organization in the mammalian nucleus. *Int. Rev. Cytol.* **242,** 283–336 (2005).
24. Sproul, D., Gilbert, N. & Bickmore, W. A. The role of chromatin structure in regulating the expression of clustered genes. *Nature Rev. Genet.* **6,** 775–781 (2005).
25. Ringrose, L. & Paro, R. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.* **38,** 413–443 (2004).
26. Ralston, A. & Rossant, J. Genetic regulation of stem cell origins in the mouse embryo. *Clin. Genet.* **68,** 106–112 (2005).
27. Szutorisz, H. *et al.* Formation of an active tissue-specific chromatin domain initiated by epigenetic marking at the embryonic stem cell stage. *Mol. Cell Biol.* **25,** 1804–1820 (2005).
28. Szutorisz, H. & Dillon, N. The epigenetic basis for embryonic stem cell pluripotency. *Bioessays* **27,** 1286–1293 (2005).
29. Chambeyron, S. & Bickmore, W. A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev.* **18,** 1119–1130 (2004).
30. Wysocka, J. *et al.* A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442,** 86–90 (2006).