An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment

This article has had an erroneous label (H3K27me3) removed from Figure 4C since Advance Online Publication.

Magnus D Lynch1, Andrew JH Smith1,2, Marco De Gobbi1, Maria Flenley1, Jim R Hughes3, Douglas Vernimmen1, Helena Ayyub1, Jacqueline A Sharpe1, Jacqueline A Sloane-Stanley1, Linda Sutherland3, Stephen Meek3, Tom Burdon3, Richard J Gibbons1, David Garrick1,4 and Douglas R Higgs1,4,*

1MRC Haematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK, 2Institute for Stem Cell Research, University of Edinburgh, Edinburgh, UK and 3The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Scotland, UK

The role of DNA sequence in determining chromatin state is incompletely understood. We have previously demonstrated that large chromosomal segments from human cells recapitulate their native chromatin state in mouse cells, but the relative contribution of local sequences versus their genomic context remains unknown. In this study, we compare orthologous chromosomal regions for which the human locus establishes prominent sites of Polycomb complex recruitment in pluripotent stem cells, whereas the corresponding mouse locus does not. Using recombination-mediated cassette exchange at the mouse locus, we establish the primacy of local sequences in the encoding of chromatin state. We show that the signal for chromatin bivalency is redundantly encoded across a bivalent domain and that this reflects competition between Polycomb complex recruitment and transcriptional activation. Furthermore, our results suggest that a high density of unmethylated CpG dinucleotides is sufficient for vertebrate Polycomb recruitment. This model is supported by analysis of DNA methyltransferase-deficient embryonic stem cells.

Introduction

The mechanism by which gene expression and the associated chromatin states are encoded in primary DNA sequence is a fundamental question in molecular biology. Orthologous chromosomal regions from closely related species can exhibit differing patterns of transcription factor binding, histone modifications and transcriptional output in the same cell type. Many factors could play a role in this process including changes in DNA sequence, positioning within the nucleus, alterations in the epigenetic machinery and in the levels or modifications of the transcription factors.

Transfer of large chromosomal segments into the genome of another species has provided evidence for the primacy of DNA sequence in the encoding of gene expression and chromatin states. A comparison of transcription factor binding and H3K4me3 modification in hepatocytes from an aneuploid mouse carrying human chromosome 21 found that the human chromosome adopted the human rather than the mouse pattern of chromatin modifications (Wilson et al., 2008). Furthermore, we have previously reported that, following the replacement of ~87 kilobases (kb) of the mouse α globin locus with a corresponding 120-kb region from the human genome, the human sequence adopts its native chromatin state in erythroid cells (Wallace et al., 2007). Thus, the information required for species-specific regulatory differences appear to be encoded in cis-acting DNA sequences.

An important outstanding question is whether this code is local or global (Coller and Kruglyak, 2008). Interspecies differences in the chromatin landscape could either reflect alterations restricted to the site of chromatin modifications or alternatively a synergistic interaction between regulatory elements distributed throughout the locus. This question is particularly pertinent to the establishment of bivalent chromatin domains, which have been defined as promoters marked by both the active H3K4me3 modification (mediated by Trithorax Group (TrxG) proteins) and the repressive H3K27me3 modification (mediated by Polycomb Group (PcG) proteins) (Azuara et al., 2006; Bernstein et al., 2006). Sites of H3K4 methylation are conserved between orthologous locations in human and mouse genomes; however, the underlying DNA sequences are often no more conserved than background (Bernstein et al., 2005). This implies either that the DNA elements directing H3K4 methylation represent only a small fraction of the underlying sequence or that they are influenced by distal flanking sequences. Sequences responsible for vertebrate PcG recruitment and modification by H3K27me3 are also incompletely characterized (Margueron and Reinberg, 2011). In support of a local model, CpG islands (CGIs; Ku et al., 2008; Mendenhall et al., 2010) and transcription factor binding sites (Barna et al., 2002; Caretti et al., 2004; Kim et al., 2005) are key regions.
The \(\alpha\) globin genes provide a useful model for investigating the role of primary DNA sequences in the templating of chromatin states. In pluripotent embryonic stem (ES) cells, the human \(\alpha\) globin locus contains prominent sites of PCg recruitment and chromatin bivalency whereas the corresponding mouse locus does not (Garrick et al., 2008). We have undertaken a comparative analysis of these loci to investigate the sequences encoding the bivalent chromatin state. We first confirm, by comparing these two loci within the same nucleus in a humanized mouse model, that cis-acting sequences are responsible for differential recruitment of PCg proteins. Next, to determine which sequences are responsible, we have used recombinase-mediated cassette exchange (RMCE) to insert various fragments of the human locus into the orthologous position in the mouse locus. We find that a 4-kb region of human sequence establishes a novel bivalent chromatin domain. Analysis of non-overlapping fragments shows that chromatin state is redundantly encoded across this 4-kb region. Using this model we provide evidence that, consistent with a recent report (Mendenhall et al., 2010), chromatin bivalency reflects competition between PCg recruitment and transcriptional activation at CGIs. These analyses highlight a correlation between density of unmethylated CpG dinucleotides and PCg recruitment and a causative relationship is supported by the finding of multiple sites of de-novo Polycomb repressive complex 2 (PRC2) recruitment at CpG-rich regions that are methylated in wild-type ES cells and lose methylation in Dnmt3a/b-/- ES cells.

**Results**

Local sequences are sufficient to encode chromatin bivalency

The \(\alpha\) globin genes are similarly arranged in the human and mouse genomes (Figure 1A); however, the human \(\alpha\) globin locus is associated with prominent sites of PCg recruitment and chromatin bivalency in pluripotent cells whereas the corresponding mouse locus is not (Garrick et al., 2008; Figure 1A; Supplementary Figure S1A–D). To confirm that cis-acting sequences are responsible for these differences in chromatin state, we analysed ES cells from a mouse model in which the entire mouse \(\alpha\) globin locus is replaced with a syntenic region (~120 kb) from the human locus (Wallace et al., 2007; Figure 1B). Since only one mouse chromosome is modified, species-specific real-time qPCR probes can be used to compare the chromatin profiles at the mouse and human loci within the same nucleus. There is a clear difference in chromatin state between the human and the mouse \(\alpha\) globin genes: Cbx7, a component of the Polycomb repressive complex 1 (PRC1), Ezh2, a component of the PRC2 and the H3K4me3 histone modification are templated specifically to the human but not to the mouse locus (Figure 1C–E). Similarly, although the level of H3K27me3 is slightly above background levels at the mouse locus, the level at the human gene is considerably greater (Figure 1F). Thus, these differences in chromatin state between mouse and human \(\alpha\) globin genes must be determined by cis-acting sequences rather than trans-acting factors which differ between human and mouse.

Next, we determined which sequences encode chromatin bivalency. Using RMCE (Figure 2A), we introduced test fragments from the human \(\alpha\) globin cluster into the orthologous region of the mouse locus, which contains duplicated copies of the mouse \(\alpha\) and \(\beta\) genes in two homology blocks. We replaced the downstream homology block with an RMCE cassette by homologous recombination. Subsequently, using RMCE we introduced a 4-kb fragment of the \(\alpha\) globin locus, including the human \(HBA2\) gene and flanking sequences, into this cassette. A linked Hprt selective marker was excised using Flp recombinase so that the human fragments were flanked only by frt and lox511 sites in the mouse locus (Figure 2A). We also introduced DNA fragments encoding FERD3L, another short gene associated with a bivalent chromatin state in human ES cells (Supplementary Figure S1E) and as a negative control HBB, which does not recruit PCg or the H3K4me3 mark in human ES cells (Supplementary Figure S1F).

Both \(HBA2\) and FERD3L recruit PRC2 (Ezh2) and are modified by H3K27me3 at this ectopic location. The \(HBA2\) gene was also modified by H3K4me3 consistent with a bivalently modified chromatin domain. The levels of H3K4me3 modification observed were lower for FERD3L but still above background (Figure 2B and C). There was no recruitment of PRC2 or modification of chromatin with H3K27me3 or H3K4me3 for the negative control HBB fragment (Figure 2D). Thus, all of the sequences required for chromatin bivalency appear to be encoded locally in these short DNA fragments. ChIP with an antibody to unmodified histone H3 confirms that these differences do not reflect alterations in histone occupancy (Supplementary Figure S1G–I).

Chromatin state is redundantly encoded throughout a bivalent domain

Domains of chromatin bivalency may extend over many kilobases (Ku et al., 2008). This could either reflect recruitment to an initial site followed by spreading along the chromosome.

---

**Figure 1** Differences in chromatin state at human and mouse \(\alpha\) globin loci in humanized mouse ES cells. (A) The human \(\alpha\) globin cluster is located close to the telomere (16p13.3), whereas the mouse cluster lies at an interstitial chromosomal position (11qA4). The positions of \(\alpha\) globin genes (\(HBA1/Hba1\) and \(HBA2/Hba2\)) and other globin genes within the loci are indicated. The \(\alpha\) globin genes in both species are arranged in duplicated homology blocks. In the human locus, these blocks contain only the \(\alpha\) globin genes whereas in the mouse, the \(\beta\) globin genes are also present. Published ChIP-seq data illustrate differences in the epigenetic status at human versus mouse \(\alpha\) globin genes in human and mouse ES cells. To facilitate analysis of duplicated homology blocks, reads were remapped with Bowtie permitting up to two copies of a sequence in the genome. Read count is normalized to reads mapped per 10 million. (B) For one copy of mouse chromosome 11, the illustrated region was replaced with the 120-kb orthologous region of human chromosome 16 including the \(\alpha\) globin genes (Wallace et al., 2007) permitting comparison of human and mouse globin loci within the same nucleus of pluripotent cells. (C–F) ChIP was performed with antibodies to (C) Ezh2, (D) Cbx7, (E) H3K4me3 and (F) H3K27me3 in these transgenic mouse ES cells bearing one wild-type allele and one humanized allele. Enrichment was quantified as percentage of input DNA and is plotted according to the position of species-specific qPCR amplicons relative to the human or mouse \(\alpha\) globin gene. Positive and negative control points are shown on the right side of each panel (from left to right: 5’ mouse \(\beta\) globin gene, mouse \(\beta\) globin exon1, mouse \(\beta\) actin, mouse Gata6). Data are the result of at least two biological replicates ± s.d.
or recruitment by multiple redundant sequence elements throughout the domain. Initially, we assessed the relative contributions of the promoter and gene body sequences to chromatin state. To address this, the HBA2 4-kb test fragment was divided into three subfragments, which were separately integrated into the same genomic locus using RMCE (Figure 3). Fragment (I) is the original 4-kb fragment. Fragment (II) and Fragment (III) extend from the start of Fragment (I) to the transcriptional start site (TSS) and exon 2, respectively. Fragment (IV) extends from the TSS to the end of the original 4-kb fragment. The nucleotide composition of these various fragments and the HBB and FERD3L fragments is illustrated (Supplementary Figure S2A–F). For each fragment, three independently derived cell lines were analysed.

---

**Figure 3**

**A**

Human α globin locus

Chr16

Mouse α globin locus

Chr11

**B**

Mouse chromosome 11

Human sequence (from Chr16p13.3)

Mouse chromosome 11

Globin genes

**C**

Ezh2

% Input

0 200 400 600 800 1000 1200 1400 1600 1800 2000

0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2

Human or mouse α globin

**D**

CBX7

% Input

0 0.5 1 1.5 2 2.5 3 3.5

0 500 1000 1500 2000

Human or mouse α globin

1.8 kb

**E**

H3K4me3

% Input

0 5 10 15 20 25

0 500 1000 1500 2000

Human or mouse α globin

**F**

H3K27me3

% Input

0 5 10 15 20 25

0 500 1000 1500 2000

Human or mouse α globin

1.8 kb

---

**Supplementary Figures S2A–F**

Human or mouse α globin

HUMAN-specific qPCR amplicons

MOUSE-specific qPCR amplicons

Positive and negative control amplicons
Figure 2 Establishment of a novel bivalent chromatin domain in the mouse α globin locus. (A) An RMCE cassette was targeted to the wild-type mouse α globin locus deleting the 3′ homology block containing the Hba2 and 3′ theta genes. The position of qPCR primers located within the mouse α globin locus upstream of the RMCE exchange site (1–2) and in exon 1 of the mouse α globin gene (3) are indicated. (B–D) DNA fragments encoding the human (A) HBA2, (B) FERD3L and (C) HBB genes were separately integrated into this locus using the RMCE system. Following Flp-mediated excision of the Hprt selective marker gene, ChIP was performed with antibodies to Ezh2, H3K4me3 and H3K27me3 and enrichment quantified with species-specific qPCR probes. The positions of qPCR probes within the tested fragments are indicated (numbered 4–14). Other numbered probes are control points in the mouse genome as follows: (15) 5′ β globin; (16) β globin exon 1; (17) 3′ β globin; (18) β actin promoter; (19) HoxB7; (20) HoxC5 and (21) Gata6 promoter. Test probes are shown in green along with mouse genomic negative control points (red) and genomic positive control points (grey). Other points are shown as unfilled bars. Enrichment was quantified as percentage of input DNA. Data are shown for three separately derived and analysed cell lines ± s.d.
Figure 3 Chromatin state is redundantly encoded in the 4-kb HBA2 fragment. (A) The indicated subfragments were separately integrated into the mouse α globin locus using the RMCE system. The positions of tested subfragments relative to the original 4-kb HBA2 fragment are shown. (B) Following Flp-mediated excision of the Hprt selective marker gene, ChIP was performed with antibodies to Ezh2, Cbx7, H3K4me3 and H3K27me3 and enrichment was quantified with species-specific qPCR probes. qPCR probes 1–3 are located within the endogenous mouse α globin locus as indicated in Figure 1. Other probe positions are identical to Figure 2. Enrichment was quantified as percentage of input DNA. Data are shown for three separately derived and analysed cell lines ± s.d. (C) ChIP-seq was performed with an antibody to Ezh2 for a cell line with Fragment (IV) inserted into the mouse α globin locus (orange bar). Reads were mapped to the transgenic locus (middle track). For comparison, ChIP-seq data are shown for Ezh2 at the wild-type mouse α globin locus (upper track; data from Ku et al). An input track is also displayed (lower track). ChIP-seq enrichment is normalized to mapped reads per 10 million.
Remarkably, all three of the test fragments became modified by H3K4me3 and H3K27me3 and recruited PRC1 and PRC2 complexes, albeit to varying degrees (Figure 3B). Therefore, it appears that the signal for chromatin bivalency is redundantly encoded throughout Fragment (I).

Surprisingly, the greatest magnitude of PcG recruitment and H3K27me3 modification was observed not for the largest Fragment (I) but for a smaller Fragment (IV) which lacks the gene promoter. This fragment also had the lowest level of H3K4me3 modification. To compare the degree of PcG recruitment at the newly inserted human fragments with the flanking mouse locus, we performed ChIP-seq in transgenic ES cells containing Fragment (IV) using an antibody to Ezh2 (Figure 3C, lower track). The corresponding region in the unmodified genomic locus is illustrated for comparison (Figure 3C, upper track). This shows that the bivalent chromatin state is templated by sequences within the fragment rather than acquiring modification that has spread from flanking sequences. We also note that there is not marked spreading of the Ezh2 signal into adjacent chromatin.

**Chromatin bivalency reflects competition between PcG recruitment and transcriptional activation**

Since PcG recruitment is greater when the promoter sequences are deleted it seemed possible that chromatin bivalency reflects a competition between the recruitment of PcG and activating proteins at the promoter. If so, the presence of additional activating sequences should shift the balance towards an active epigenetic state. To test this, we generated another transgenic cell line in which an 200-bp fragment encoding an MC1 promoter, which is constitutively active in mouse ES cells (Thomas and Capecchi, 1987), was inserted upstream of the 4-kb HBA2 fragment (Figure 4A), and the resulting construct was integrated by RMCE and compared with cell lines carrying Fragment (I) or Fragment (IV).
Quantification of spliced HBA2 cDNA in these three cell lines revealed a very low level of expression when the promoter had been deleted, higher expression for the HBA2 gene with intact promoter and higher still for the fragment containing the MC1 promoter (Figure 4B). It should be noted that the level of expression observed even for this fragment is several orders of magnitude lower than for HBA2 in an erythroid cell line. Next, we compared active and repressive histone modifications and PRC2 recruitment for all three of these RMCE-modified cell lines. There is an inverse relationship between H3K27me3 and H3K4me3 (Figure 4C) with the highest level of PRC2 recruitment and lowest level of H3K4me3 seen when the activating sequences associated with the promoter were deleted. Complete clearing of PRC2 and H3K27me3 occurred in the presence of the MC1 promoter and an intermediate level of H3K4me3 and H3K27me3 was observed for the wild-type construct.

**CGI erosion during mammalian evolution is associated with loss of chromatin bivalency**

A clear difference in primary sequence between the human and mouse γ globin genes is the presence of prominent CGIs in the human but not in the mouse locus; this reflects erosion of the CGI in the mouse lineage compared with a common mammalian ancestor (Antequera and Bird, 1993). To investigate whether this association between CGI erosion and loss of PcG recruitment is a general phenomenon, we identified 2088 peaks of H3K27me3 in human ES cells that were associated with elevated CpG density (>6% in a 500-bp window). These were compared with their corresponding genomic intervals (using the UCSC liftOver tool) in mouse and rat ES cells.

Peaks are ranked according to CpG density (in a 500-bp window) in the mouse genome (Figure 5A). Each line displays a peak of H3K27me3 in human ES cells and the chromatin state of corresponding genomic regions in mouse ES cells. CpG density is generally conserved between corresponding regions in the human and rodent genomes and this is associated with conservation of H3K27me3 recruitment in ES cells. However, numerous examples were identified for which CGI erosion between human and mouse is associated with diminution of the H3K27me3 histone modification (Figure 5A, below dashed line). This is confirmed by a pileup analysis comparing enrichment of H3K27me3 at mouse regions associated with conserved (>3%; green) and eroded (<3%; blue) CpG density (Figure 5B). The correlation is also apparent when the data are presented as a scatter plot (Supplementary Figure S3). To confirm that this is not a species-specific phenomenon, a map of H3K27me3 was generated for rat ES cells and an analogous pileup analysis performed with similar findings (Figure 5C), although it is noted that the ratio of signal to background for this data set is inferior to the human and mouse data sets. Finally, three loci are illustrated for which CGI erosion in the mouse compared with human genome is associated with loss of the bivalent chromatin state in ES cells (Figure 5D–F).

**De-novo PRC2 recruitment to CpG-rich sequences in DNA methyltransferase-deficient ES cells**

Most sites of PcG recruitment in ES cells are associated with an elevated density of CpG dinucleotides (Ku et al., 2008). CGIs associated with actively expressed promoters are marked exclusively by H3K4me3 (Mikkelsen et al., 2007) and results presented here and elsewhere (Mendes-Moura et al., 2010) suggest that the binding of activating factors reduces PcG recruitment. However, this does not explain why some CGIs recruit neither PcG nor the H3K4me3 modification. We have found that DNA methylation can explain a proportion of these CGIs. An example is the CGI present in body of the human RHBDF1 gene. In human ES cells (Figure 6A), the promoter is marked by H3K4me3 but the intragenic CGI is not modified by either H3K4me3 or H3K27me3. As the gene is silenced during erythropoiesis, the promoter recruits H3K27me3 but the intragenic CGI remains unmarked. We have previously reported that this intragenic CGI is methylated in multiple adult somatic tissues including erythroid cell lines (Vyas et al., 1992). This was confirmed for human ES cells by inspection of published genome-wide bisulphite sequencing data (Figure 6A; Lister et al., 2009). A genome-wide comparison of DNA methylation and the H3K4me3 and H3K27me3 histone modifications confirms that most methylated CGIs are unmethylated by both histone modifications in human ES cells (Figure 6B–E). Conversely, inspection of chromatin state at unmethylated (<5% methylated) CGIs (Supplementary Figure S4A) reveals most unmethylated CGIs to be modified by either the H3K27me3 or H3K4me3 mark in human ES cells; moreover, there is a positive correlation with CGI size so that essentially all unmethylated CGIs >1 kb in size are modified by one or other of these marks (Supplementary Figure S4B–E).

The anticorrelation between DNA methylation and PcG recruitment in ES cells could reflect either (i) inhibition of PcG recruitment by DNA methylation, (ii) inhibition of DNA methylation by H3K4me3 and/or H3K27me3 or (iii) a confounding factor, such as the restriction of DNA methylation to non-promoter CGIs in human ES cells. To distinguish these models, we performed genome-wide ChIP-seq for H3K27me3 in mouse ES cells deficient for Dnmt3a and Dnmt3b methyltransferases and in wild-type mouse ES cells. The general pattern of H3K27me3 recruitment at positive (Figure 7A; Supplementary Figure S5A) and negative (Figure 7B; Supplementary Figure S5B) control regions is conserved between wild-type and knockout cell lines, although there appears to be somewhat greater spreading of the signal in knockout cells. However, numerous _de novo_ sites of H3K27me3 recruitment were observed at CpG-rich regions in the knockout cells (Figure 7C–J; Supplementary Figure S5C and D). A number of these sites were confirmed by qPCR (Supplementary Figure S5E). We hypothesized that DNA associated with these regions is methylated in wild-type ES cells and loses methylation in Dnmt3a/b-/- ES cells and this was confirmed by bisulphite sequencing (Supplementary Figure S6). These results demonstrate that loss of DNA methylation leads to _de novo_ sites of PcG recruitment at CpG-rich sequences, suggesting that PcG recruitment is the default state for such sequences in the absence of binding of transcriptional activators.

**Discussion**

Studies in which large chromosomal segments are introduced into the cells of another species (Wallace et al., 2007; Wilson et al., 2008) suggest that differences in chromatin state and
Figure 5 CpG island erosion during mammalian evolution is associated with loss of chromatin bivalency. (A) In all, 2088 peaks of H3K27me3 enrichment associated with a CpG density of 6% or greater (500 bp window) were identified in human ES cells (Ernst et al., 2011). Data for each peak are plotted on a single line with flanking regions of 5 kb (X axis). The corresponding genomic intervals in the mouse genome were identified (UCSC liftOver tool). CpG density in a 500-bp window (percent) and H3K27me3 read density in a 500-bp window (percentage of maximal enrichment) for mouse ES cells are plotted adjacent to the corresponding human peaks (Ku et al., 2008). Peaks are displayed in order of maximum CpG density in the mouse genomic interval. CpG erosion in the mouse genome is associated with a diminution of H3K27me3 recruitment in mouse ES cells. The locations of selected genes are indicated. (B) Pileup analysis of the same data comparing mouse genomic regions associated with a maximum CpG density (500 bp window) of > 3% (green) to those with < 3% (blue). (C) A map of H3K27me3 was generated for rat ES cells and an analogous pileup analysis was performed. (D, E) Selected examples of CpG island erosion between human and mouse genomes at the (D) MYO1G, (E) CLEC4G and (F) MYF6 genes associated with loss of chromatin bivalency. ChIP-seq data are displayed as mapped reads per 10 million.
transcriptional output largely reflect differences in primary DNA sequence. Consequently, comparative genomics provides a powerful strategy for deciphering this regulatory code (Waterston et al., 2002). Sequence conservation across multiple organisms is predictive of functional elements (Pennacchio and Rubin, 2001; Hughes et al., 2005). However, not all functionally important sequences exhibit a high degree of conservation (Ruvinsky and Ruvkun, 2003; Bernstein et al., 2005), and at present the analysis of conserved sequence cannot predict which functional elements are required for a given effect. To date, the function of a particular sequence is usually addressed by transgenic experiments. However, the analysis of expression and chromatin state in such experiments is often confounded by copy number variation and position effects. Here, we have analysed elements required for the recruitment of a specific bivalent chromatin signature (H3K27me3 and H3K4me3) using RMCE, which allows the analysis of all tested fragments at single copy in a defined chromosomal location. This provides a powerful system for identifying the precise cis elements involved in chromatin templating and how far the effect may spread beyond the primary elements.

Bivalent domains are thought to result from the interplay between repressive (PcG/H3K27me3) and activating (TrxG/ H3K4me3) pathways and play an important role in marking the promoters of key developmental regulators in multipotent cells (Boyer et al., 2006; Lee et al., 2006). The identity of cis-acting sequences responsible for PcG recruitment and the establishment of chromatin bivalency in mammals is controversial. In support of local sequences, a 1.8-kb region between human HOXD11 and HOXD12 was demonstrated to recruit PRC1 and PRC2 (Woo et al., 2010) and CpG-rich sequences were reported to recruit PcG components to a transgenic BAC (Mendenhall et al., 2010). Conversely, there is evidence that genomic context plays an important role in PcG recruitment. A 3-kb element located adjacent to the mouse MafB recruited only PRC1 and not PRC2 components when assayed in mammalian cells (Sing et al., 2009). In addition, a number of reports have identified non-coding RNAs encoded in cis or in trans that are required for the recruitment of PcG (Rinn et al., 2007; Pandey et al., 2008; Zhao et al., 2008; Yap et al., 2010).

In this study, we were initially struck by the observation that human locus encoding α globin (HBA) contains prominent sites of PcG recruitment and chromatin bivalency in pluripotent cells whereas the orthologous mouse locus does not. We first confirmed that this is due to cis sequence differences by comparing these loci within the same nucleus. Furthermore, we found that a relatively small 4 kb DNA fragment containing the human HBA2 gene was sufficient to recreate a novel site of chromatin bivalency when inserted into the corresponding position in the mouse locus. This observation is not specific to the α globin genes since another 4 kb region containing the human FERD3L gene also established chromatin bivalency. Thus, for these examples, chromatin bivalency is encoded by local sequences. Nevertheless, some domains of chromatin bivalency extend over many kilobases (Ku et al., 2008). Here, by analysing the HBA2 gene, we have shown that this could be explained by redundant recruitment to multiple sequence elements as opposed to recruitment of chromatin modifying enzymes to a single site and subsequent spreading. Redundant encoding of a chromatin state may confer robustness to gene regulation in the face of single base substitutions.
Figure 7  De-novo genomic sites of H3K27me3 methylation in Dnmt3a/b−/− ES cells. Genome-wide maps of H3K27me3 were generated for Dnmt3a/b−/− and wild-type ES (E14-TG2a) cell lines. For each locus, the profile of H3K27me3 (reads/10 million) is displayed for wild-type and mutant cells and compared with CpG density (CpGs/500 bp). (A) Positive control region (Hox A locus). (B) Negative control region (Hbb locus). (C–J) Examples of de-novo PRC2 recruitment at intragenic CpG islands that are methylated in wild-type cells and become demethylated in mutant cells. The locations of bisulphite amplicons displayed in Supplementary Figure S6 are indicated.
Our results suggest that the bivalent chromatin state reflects a competitive equilibrium between the recruitment of PcG to CpG-rich sequences and gene activation associated with recruitment of TrxG complexes that contain H3K4 methyltransferases. Deletion of promoter sequences from the HBA2 gene increases PcG recruitment and H3K27me3 modification relative to the wild-type sequence, with corresponding diminution of H3K4 methylation. Conversely, the addition of a constitutively active promoter increases the level of H3K4me3 in association with a reduction in PcG recruitment. These findings are consistent with a study of randomly integrated transgenic BACs in mouse ES cells which found that the deletion of activating motifs from a housekeeping CGI promoter led to PcG recruitment (Mendenhall et al., 2010). A limitation of that study is that the site of genomic integration and copy number is not controlled; this is important since the presence of a transgene in multiple copies is sufficient to initiate PcG silencing in Drosophila (Pal-Bhadra et al., 1997).

Our results are also informative regarding the sequences responsible for recruitment of H3K4 methyltransferases to bivalent domains. Genome-wide studies have established that most sites of PcG recruitment in ES cells are also associated with at least a low level of the H3K4me3 modification (Mikkelsen et al., 2007) and it has been suggested that this ‘pre-marks’ the gene for activation. Recruitment of the H3K4me3 modification, whereas a higher level of transcription is associated with clearing of PcG proteins.

The nature of the genomic signals for vertebrate PcG recruitment is a central question in the field of epigenetics. On the other hand, there is a correlation between the magnitude of the H3K4me3 modification and the transcriptional activity of the gene (Thomson et al., 2008). Our results suggest that the bivalent chromatin state reflects a competitive equilibrium between PcG recruitment and H3K4me3 modification relative to the wild-type sequence, with corresponding diminution of H3K4 methylation. Conversely, the addition of a constitutively active promoter increases the level of H3K4me3 in association with a reduction in PcG recruitment. These findings are consistent with a study of randomly integrated transgenic BACs in mouse ES cells which found that the deletion of activating motifs from a housekeeping CGI promoter led to PcG recruitment (Mendenhall et al., 2010). A limitation of that study is that the site of genomic integration and copy number is not controlled; this is important since the presence of a transgene in multiple copies is sufficient to initiate PcG silencing in Drosophila (Pal-Bhadra et al., 1997).

Recombinase-mediated cassette exchange

Test sequences were cloned into the AseI site of a plasmid containing an RMCE donor cassette (loxP/Hprt−AS/loxP/MC1neo/loxS1I) created by modification of a previously described chromosome engineering cassette (Wallace et al., 2007) and flanked by homology arms designed to delete the 3′/5′ homology block in the mouse β globin locus. E14-TG2a IV mouse ES cells, which are hypoxanthine phosphoribosyltransferase deficient (HPRT−), were cultured and gene targeting by homologous recombination was performed as previously described (Wallace et al., 2007). Correctly targeted clones were identified by Southern blot with HindIII and BglII digests. Further details of the constructs used are available on request.

Chromatin immunoprecipitation

Chromatin immunoprecipitation was performed with the Millipore ChIP Assay Kit (Millipore, 17-295). Briefly, ES cells were cross-linked with 1% formaldehyde in PBS for 10 min at 37 °C. Chromatin was then isolated and sonicated to an average size of 500–1000 bp using a Diagenode Bioruptor. Chromatin fragments were immunoprecipitated with antibodies to H3K4me3 (Millipore, 05-745R), H3K27me3 (Millipore, 07-449), Ezh2 (Abnova, pAB08649) or Cbx7 (Santa Cruz, P-15 sc70232). Immunoprecipitated DNA was either analysed by real-time qPCR or prepared for ChIP sequencing according to standard Illumina protocols. Enrichment was quantified by real-time qPCR as a percentage of input DNA with Taqman probes specific to the human β globin locus. E14-TG2a IV mouse ES cells, which are hypoxanthine phosphoribosyltransferase deficient (HPRT−), were cultured and gene targeting by homologous recombination was performed as previously described (Wallace et al., 2007). Correctly targeted clones were identified by Southern blot with HindIII and BglII digests. Further details of the constructs used are available on request.
DNA and RNA analysis
RNA was prepared with TRI reagent (Sigma) and quantified relative to mouse Gapdh with RT-qPCR primers specific to the spliced human $\alpha$ globin transcript (Anguita et al., 2004). Bisulfite conversion of genomic DNA was performed with the EZ DNA Methylation-Gold kit (Zymo Research, D5005) and methylation was quantified by cloning into pGEM-T Easy (Promega) and sequencing.

Bioinformatic analysis
To investigate the relationship between CGI erosion and chromatin bivalency, publicly available ChiP-seq data sets for H3K27me3 in human (Ernst et al., 2011) and mouse (Mikkelsen et al., 2007) ES cells were analysed with custom Python scripts. Peaks in human ES cells were identified with a sliding window of 500 bp and moving increment of 50 bp. Peaks separated by 1 kb or less were merged and peaks associated with enrichment on an input DNA track (Ernst et al., 2011) were eliminated. For peaks associated with a CpG density of 6% or greater in a 500-bp window, the corresponding mouse genomic regions were identified using the UCSC liftOver tool. Density of H3K27me3 and CpG dinucleotides was plotted for these genomic regions in human and mouse ES cells. Finally, a pileup analysis was performed to compare the genomic regions in these genomic regions in human and mouse ES cells. Finally, a pileup analysis was performed for H3K27me3 in rat ES cells.

To quantify histone modifications and DNA methylation at CGIs in human ES cells, publicly available data sets for H3K27me3 (Ernst et al., 2011), H3K4me3 (Ernst et al., 2011) and high-coverage bisulfite sequencing (Lister et al., 2009) were analysed. For each annotated CGI (UCSC definition, hg18), CpG methylation was quantified as a fraction by dividing the total number of methylated cytosines by the total number of unmethylated cytosines in sequencing reads that mapped to that CGI. The density of histone modifications was quantified by taking the maximum read density in a sliding 500 bp window at any position within the CGI and flanking 1 kb regions (to account for nucleosome depletion at a subset of CGIs). Sex chromosomes and CGIs to which reads could not be mapped (UCSC mappability track) were excluded from the analysis.

ChiP-sequencing data from this study have been deposited with the GEO database (accession number GSE27580).

Supplementary data
Supplementary data are available at The EMBO Journal Online (http://www.embojournal.org).

Acknowledgements
We thank S Butler for assistance with tissue culture, the Computational Biology Research Group, Oxford University for bioinformatic support and T Milne for critical reading of the manuscript. This work was supported by the Medical Research Council and the Oxford Biomedical Research Centre. MDL was the recipient of a clinical research training fellowship (MRC).

Author contributions: MDL, DRH, DG, RG and AJS designed experiments. MDL, MF, HA and AJS performed experiments. LS, SM and TB derived and expanded rat ES cells. AJS, JRH, DV and MDG provided reagents and scientific inputs.

Conflict of interest
The authors declare that they have no conflict of interest.

References

Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331: 76–79

DNA methylation
This process involves the transfer of a methyl group from S-adenosylmethionine to cytosine residues in DNA. In mammals, this process is catalyzed by DNA methyltransferases, which add a methyl group to the 5-position of cytosine residues within CpG dinucleotides.

CpG islands
CpG islands are regions of DNA that are rich in CpG dinucleotides. They are typically defined as regions with a high frequency of CpG dinucleotides, with at least one CpG dinucleotide every 200 bp.

Bisulfite sequencing
Bisulfite sequencing is a technique used to determine the methylation status of DNA at specific loci. It involves treating DNA with bisulfite, which converts unmethylated cytosines to uracil, while methylated cytosines are left unchanged. The resulting DNA is then amplified and sequenced, allowing the methylation status of specific loci to be determined.

ChIP-sequencing
Chromatin immunoprecipitation (ChIP) followed by sequencing (ChIP-sequencing) is a technique used to determine the binding sites of proteins on the genome. It involves cross-linking proteins to DNA, shearing the DNA into small fragments, and then using antibodies to pull down the DNA-protein complexes. The resulting DNA is then amplified and sequenced, allowing the binding sites of specific proteins to be determined.

Functional assessment
Functional assessment is the process of determining the biological function of a gene, protein, or other molecular entity. This can be achieved through a variety of experimental approaches, such as knockdown experiments, overexpression experiments, or analysis of gene expression in different cell types or conditions.

Multispecies conservation
Multispecies conservation refers to the preservation of genetic diversity across multiple species. This can be achieved through a variety of approaches, such as captive breeding programs, habitat conservation, or genetic management plans.

Polycomb complexes
Polycomb complexes are multicomponent protein complexes that play a role in epigenetic regulation, specifically in the maintenance of gene silencing and the establishment of stable epigenetic states.

ChIP-sequencing data from this study have been deposited with the GEO database (accession number GSE27580).