# Epigenetic regulation of the human genome: coherence between promoter activity and large-scale chromatin environment

Hanna Julienne, Azedine Zoufir, Benjamin Audit, Alain Arnéodo

# Epigenetic regulation of the human genome: coherence between promoter activity and large-scale chromatin environment

Hanna Julienne[a,b], Azedine Zoufir[a,b], Benjamin Audit[a,b] and Alain Arneodo[a,b]

[a]*Université de Lyon, F-69000 Lyon, France;* [b]*Laboratoire de Physique, CNRS UMR 5672, Ecole Normale Supérieure de Lyon, F-69007 Lyon, France*

Increasing knowledge of chromatin structure in various cell types raises the challenge of deciphering the contribution of epigenetic modifications to the regulation of nuclear functions in mammals. In a recent study, we have analysed the genome-wide distributions of thirteen epigenetic marks in the human cell line K562 at 100 kb resolution of Mean Replication Timing (MRT) data. Using classical clustering techniques, we have shown that the combinatorial complexity of these epigenetic data can be reduced to four predominant chromatin states that replicate at different periods of the S-phase. C1 is an early replicating transcriptionally active euchromatin state, C2 a mid-S repressive type of chromatin associated with Polycomb complexes, C3 a silent chromatin with lack of chromatin marks that replicates later than C2 but before C4, a HP1-associated heterochromatin state that replicates at the end of S-phase. These four chromatin states display remarkable similarities with those recently reported in fly, worm and plants at higher $\sim 1$ kb resolution of gene expression data. Here, we extend our integrative analysis of epigenetic data in the K562 human cell line to this smaller scale by focusing on gene promoters ($\pm 3$ kb around transcription start sites). We show that these promoters can similarly be classified into four main chromatin states: P1 regroups all the marks of transcriptionally active chromatin and corresponds to CpG rich promoters of highly expressed genes; P2 is notably associated with the histone modification H3K27me3 that is the mark of a polycomb repressed chromatin state; P3 corresponds to promoters that are not enriched for any available marks as the signature of a 'null' or 'black' silent heterochromatin state and P4 characterizes the few gene promoters that contain only the constitutive heterochromatin histone modification H3K9me3. When investigating the coherence between promoter activity (P1, P2, P3 or P4) and the large-scale chromatin environment (C1, C2, C3 or C4), we find that the higher the gene density in a considered 100 kb-window, the higher (resp. the lower) the probability of a P1 active promoter (resp. silent P2, P3 and P4 promoters) to be surrounded by an open euchromatin C1 (resp. facultative C2, black C3 or HP1-associated C4 heterochromatin) environment. From large to small scales, it is mainly C4 and to a lesser extent C3 heterochromatin environments both corresponding to gene poor regions, that strongly conditions promoters to belong to the inactive P3 and P4 classes. If C1 (resp. C2) environment surrounds a majority of corresponding active P1 (resp. P2) promoters, it also contains a non-negligible proportion of inactive P2 and P3 (resp. active P1 and inactive P3) promoters. When further investigating the large-scale organization of human genes with respect to 'master' replication origins that were shown to border megabase-sized U-shaped MRT domains, we reveal some significant enrichment of highly expressed P1 genes in a closed neighbourhood of these early initiation zones consistently with the gradient of chromatin states observed from C1 at U-domain borders followed by C2, C3 and C4 at U-domain centers. On the contrary to P2 promoters that are mainly found in the C2 environment at finite distance ($\sim$200–300 kb) from U-domain borders, the inactive P3 and P4 promoters are distributed rather homogeneously inside U-domains. The generalization of our study to different cell types including ES, somatic and cancer cells is likely to provide new insight on the global reorganization of replication domains during differentiation (or disease) in relation to coordinated changes in chromatin environment and gene expression.

**Keywords:** epigenetic modifications; chromatin states; mean replication timing; replication domains; promoter activity; clustering analysis

## Introduction

It is increasingly recognized that the dynamics of DNA folding and unfolding within the nucleus of eukaryotic cells plays a major role in the regulation of nuclear functions including gene expression and DNA replication (Belmont et al. 1999; Cook 1999, 2001; Cremer & Cremer 2001; Berezney 2002; Chakalova et al. 2005; Gilbert et al. 2005; Branco & Pombo 2007; Fraser & Bickmore 2007; Kouzarides 2007; Misteli 2007; Sexton et al. 2007; Gilbert 2010; Maric & Prioleau 2010; Arneodo et al. 2011; Zhou et al. 2011; Bickmore & van Steensel 2013). Eukaryotic chromatin can be viewed as a succession of superimposed organizational steps including the nucleosomal array, its condensation into the 30 nm chromatin fiber and the formation of chromatin loops, up to a full extent of condensation in metaphase chromosomes (van Holde 1988; Wolffe 1998;

Calladine & Drew 1999; Alberts et al. 2002; Felsenfeld & Groudine 2003). If specific chromatin configurations may be dictated by the DNA sequence itself (Satchwell et al. 1986; Ioshikhes et al. 1996; Widom 2001; Segal et al. 2006; St-Jean et al. 2008; Milani et al. 2009; Arneodo et al. 2011; Chevereau et al. 2011; Travers et al. 2012; Struhl & Segal 2013), the chromatin structure is subject to various epigenetic modifications in any given cell type, including DNA methylation, histone modifications, histone variant incorporation and DNA-binding proteins (Kouzarides 2007; Zhou et al. 2011; Zentner & Henikoff 2013). Recent technical advances in genomics and epigenomics including the combination of chromatin immunoprecipitation (ChIP) with massive parallel sequencing (ChIP-Seq) (Schones & Zhao 2008) have made available a wealth of genome-wide data in various eukaryotic organisms, from budding yeast, to plants, worm, fly and mammals (Bernstein et al. 2007; The ENCODE Project Consortium 2007; Rando & Chang 2009; Roudier et al. 2009; Gerstein et al. 2010; Kharchenko et al. 2010; The modENCODE Consortium 2010; Feng & Jacobsen 2011; The ENCODE Project Consortium 2011). Hopefully the efforts devoted to the study of these data sets will lead to significant progress in our understanding of the regulatory functions of the chromatin landscape in gene expression, genome maintenance, replication origin specification, cell differentiation and other key cellular processes (Hon et al. 2009; Wang et al. 2009; Ernst & Kellis 2010; Filion et al. 2010; Ernst et al. 2011; Liu et al. 2011; Roudier et al. 2011; Lee et al. 2012; Sexton et al. 2012). In the human genome of interest here, we have at our disposal for multivariate analysis in different cell types, chromosomal profiles of many epigenetic modifications (Bernstein et al. 2007; The ENCODE Project Consortium 2007, 2011), nucleosome positioning (Ozsolak et al. 2007; Schones et al. 2008; Valouev et al. 2011) and chromatin accessibility such as sensitivity to DNase I cleavage (Sabo et al. 2006; Boyle et al. 2008; The ENCODE Project Consortium 2011) that all characterize the primary chromatin structure. In addition, the recent development of the chromosome conformation capture (3C) technology (Dekker et al. 2002) and its high-throughput extensions (Dostie et al. 2006; Simonis et al. 2006; Zhao et al. 2006) including Hi-C (Lieberman-Aiden et al. 2009) and derivatives (Fullwood et al. 2009; Kalhor et al. 2012) has provided quantitative measurement of intra- and inter-chromosomal interaction maps (Dostie et al. 2006; Fullwood et al. 2009; Lieberman-Aiden et al. 2009; Dixon et al. 2012; Kalhor et al. 2012; Moindrot et al. 2012) from which very instructive informations can be extracted on the so-called tertiary (3D) chromatin structure and dynamics (Lieberman-Aiden et al. 2009; Dixon et al. 2012; Dostie & Bickmore 2012; Holwerda & de Laat 2012; Moindrot et al. 2012; Cavalli & Misteli 2013). Gene expression data obtained early with the RNA-Seq technique (Mortazavi et al. 2008; The ENCODE Project Consortium 2011) have been intensively used to address the question of the role of epigenetic modifications in transcription regulation and genome activity during development and differentiation or in response to the environment (Zhou et al. 2011). In contrast, progress in elucidating the chromatin-mediated control of replication origin usage and efficiency as well as of the maintenance of the spatio-temporal replication program in higher eukaryotes has been rather slow (Berezney et al. 2000; Bogan et al. 2000; Gilbert 2001, 2010; Méchali 2001, 2010; Bell & Dutta 2002; McNairn & Gilbert 2003; Aladjem 2007; Courbet et al. 2008; Hamlin et al. 2008). Only very recently nascent DNA strands synthesized at origins were purified by various methods to map replication origins genome-wide in mouse (Sequeira-Mendes et al. 2009; Cayrou et al. 2011) and human (Lucas et al. 2007; Cadoret et al. 2008; Karnani et al. 2010; Martin et al. 2011; Mesner et al. 2011; Valenzuela et al. 2011). The set of replication origins identified so far are strongly associated with annotated promoters and seem to be enriched in transcription factor binding sites (Cadoret et al. 2008; Karnani et al. 2010; Besnard et al. 2012) and in CpG islands (Cadoret et al. 2008; Sequeira-Mendes et al. 2009; Cayrou et al. 2011). But the correlation to transcription is not that obvious since a significant proportion of origins are found in regions void of DNase-I-hypersensitive sites (DHSs) and of histone marks found at active promoters (Cadoret et al. 2008; Maric & Prioleau 2010). Genome-wide profiling of Mean-Replication Timing (MRT) in mouse (Farkash-Amar et al. 2008; Hiratani et al. 2008, 2010) and human (Woodfine et al. 2004; Desprat et al. 2009; Chen et al. 2010; Hansen et al. 2010) in different cell lines have recently revealed a significant correlation with epigenetic modifications (Farkash-Amar & Simon 2010). Early replicating regions tend to be enriched in open chromatin marks, whereas late replicating zones likely correspond to constitutive heterochromatin (Hiratani et al. 2008; Ryba et al. 2010). Altogether these analyses of chromatin, gene expression and MRT data in mammals look very promising in the perspective of better understanding the role of epigenetic modifications in the co-regulation of transcription and replication.

Multivariate statistical analyses of epigenetic data sets in human have revealed that distinct epigenetic modifications often exist in a well-defined combinations corresponding to different genomic elements like promoters, enhancers, exons, repeat sequences and/or to distinct modes of regulation of gene expression such as actively transcribed, silenced and poised (Hon et al. 2009; Wang et al. 2009; Ernst & Kellis 2010; Ernst et al. 2011; Lee et al. 2012). In a recent work (Julienne et al. 2013), with the aim at quantifying the influence of epigenetic modifications on replication timing, we have used principal component analysis (PCA) and classical clustering method to analyse thirteen epigenetic mark maps in the K562 human cell line at the 100-kb-resolution of MRT data. This study reveals that the huge combinatorial epigenetic complexity can in fact be reduced to a rather small number of predominant chromatin states that interestingly share strong similarities with

the ones previously found in *Arabidopsis thaliana* (Roudier et al. 2011), *Caenorhabditis elegans* (Liu et al. 2011) and *Drosophila* (Filion et al. 2010; Sexton et al. 2012). These four main chromatin states were further shown to correlate with MRT, namely from early to late replicating, a transcriptionally active euchromatin state (C1) enriched in insulator binding protein CTCF, a polycomb repressed facultative heterochromatin state (C2), a silent heterochromatin state (C3) not enriched in any available marks and a HP1-associated heterochromatin state (C4). When mapping these chromatin states inside the megabase-sized U-domains (Baker, Audit et al. 2012; Audit et al. 2012, 2013), where the MRT is U-shaped and its derivative N-shaped like the nucleotide compositional asymmetry in the germline skew N-domains (Brodie of Brodie et al. 2005; Touchon et al. 2005; Audit et al 2007; Huvet et al. 2007; Audit et al. 2009; Baker et al. 2010; Chen et al. 2011; Baker, Chen et al. 2012; Baker, Julienne et al. 2012), we have shown that in these replication domains that cover about 50% of the human genome, the replication wave (Guilbaud et al. 2011) proceeds along a directional path through the four chromatin states, from the open euchromatin state C1 at U/N-domain borders successively followed by the three silent chromatin states C2, C3 and C4 at the U/N-domain centers (Julienne et al. 2013). The complete analysis, of the other half of the genome that is complementary to U-domains (Julienne et al. 2013) has confirmed the dichotomic picture proposed in early studies in mouse (Farkash-Amar et al. 2008; Hiratani et al. 2008, 2010) and human (Desprat et al. 2009; Ryba et al. 2010; Yaffe et al. 2010) genomes, where early and late replicating regions occur in separated compartments of open and closed chromatin, respectively. About 25% of the human genome is covered by megabase-sized GC-rich (C1 + C2) chromatin blocks that on average replicate early by multiple almost synchronous randomly positioned origins with almost equal proportions of forks coming from both directions which explains that the skew has not accumulated in these gene-rich regions devoided of N-domains (Brodie of Brodie et al. 2005; Touchon et al. 2005; Baker et al. 2010). The last 25% of the human genome corresponds to megabase-sized GC-poor domains of interspersed (C3 + C4) heterochromatin states or long C4 domains that on average replicate late by again multiple almost coordinated origins and that contain only a few genes (Julienne et al. 2013).

In this article, our goal is to extend our integrative analysis of epigenetic data in the K562 human cell line from the 100 kb scale of MRT data to a few kb scale characteristic of gene promoters as previously performed in plants (Roudier et al. 2011), worm (Liu et al. 2011) and fly (Filion et al. 2010; Sexton et al. 2012) Then by investigating the coherence between the chromatin states obtained at these two scales, we will be in a position to study to what extent the promoter activity does condition its large-scale chromatin environment and vice versa. The paper is organized as follows. The next section is devoted to materials and methods. In the third section, we perform a combinatorial analysis of chromatin marks in K562 and we describe the epigenetic content of the four prevalent chromatin states at gene promoters. In the fourth section, we study the coherence between promoter activity, as characterized by their 'small-scale' chromatin state, and the 'large-scale' chromatin environment (namely the C1, C2, C3 and C4 chromatin states found in Julienne et al. (2013)). In this comparative analysis we emphasize the expected as well as the unexpected importance of gene density on the observed relationship between these two scales characterizing transcription and replication data respectively. In the fifth section, we investigate the spatial distribution of these promoter chromatin states inside the three types of replication domains defined in our previous work (Julienne et al. 2013), namely the 50% of the human genome paved by MRT U-domains, the 25% covered by early replicating GC-rich (C1 + C2) chromatin blocks and the 25% covered by late replicating, GC-poor (C3 + C4 or long C4) heterochromatin blocks. We conclude, in the final section, by discussing some perspectives for further studies in different cell types, in other mammalian genomes in both health and disease.

## Materials and methods

### *Annotation and expression data*

Annotation and expression data were retrieved from the Genome Browser of the University of California Santa Cruz (UCSC). To construct our data set, we used RefSeq Genes track as human gene coordinates. Genes with alternative splicing were merged into one transcript by taking the union of exons. Hence the TSS was placed at beginning of the first exon. We obtained a table of 23,329 genes. We downloaded expression values from the release 2 of Caltech RNA-Seq track (ENCODE project at UCSC: http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeCaltechRnaSeq/). Expression for one transcript is given in reads per kilobase of exon model per million mapped reads (RPKM) (Mortazavi et al. 2008). RPKM is defined as:

$$R = \frac{10^9 C}{NL}, \tag{1}$$

where $C$ is the number of mappable reads that fall into gene exons (union of exons for genes with alternative splicing), $N$ is the total number of mappable reads in the experiment, and $L$ is the total length of the exons in base pairs. We associated 17,872 genes with a valid RPKM value in K562.

### *Histone marks, H2AZ, CTCF, RNAP II, Sin3A and CBX3 ChIP-Seq data*

For all ChIP-Seq data, we downloaded data in the ENCODE standard formats 'broadpeaks' and 'bigWig' (http://genome.ucsc.edu/ FAQ/FAQformat.html). Broadpeaks format is a table of significantly enriched genomic

intervals. BigWig format is a read count profile at high resolution of 25 bp. Most of the data correspond to release 3 (August 2012) of the Broad histone track. We downloaded the tables from:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/.

The CBX3 and Sin3A data correspond to release 3 (September 2012) of the HAIB TFBS track. Tables were downloaded from UCSC:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/.

For the K562 cell line, we downloaded the broadpeak tables for the following antibodies: CTCF, H3K27ac, H3K27me3, H3K36me3, H3K4me3, H3K9me3, RNAP ll, H2AZ, H3K79me2, H3K9me1, H4K20me1, CBX3, Sin3A.

### *Read density computation around promoters*

For each ChIP-Seq data, we filtered the high resolution profiles (BigWig format) by the significantly enriched intervals (Broadpeaks format). Then, for each gene with a valid expression value, the read density was computed as the number of reads that fall in a 6 kb window around the TSS divided by the window length. By doing so, we obtained a valid epigenetic value for 13 epigenetic marks around 17,724 promoters.

### *Rank transformation and Spearman correlation matrix*

All statistical computations were performed using the R software (http://www.r-project.org/).

In order to compute the Spearman correlation matrix, the read density around promoters was transformed with the R function *rank* with option $ties.method = max$. Then we computed the Pearson correlation matrix on the transformed dataset. To reorder the matrix in Figure 1, we computed the Spearman correlation distance *dSCor* as:

$$dSCor(X, Y) = 1 - SCor(X, Y), \tag{2}$$

where *SCor* is the Spearman correlation. Then, a dendrogram was computed using the R function *hclust* with option $method = average$ and with *dSCor* as dissimilarity.

### *Principal component analysis*

Principal component analysis was performed on the rank transformed dataset using the function *dudi.pca* from the R package *ade4* (see http://pbil.univ-lyon1.fr/ADE-4 and Chessel et al. (2004)) with the option $scale = TRUE$ (i.e. each variable is centered and normalized before the PCA computation). The first three components were retained which accounts for 74% of the dataset variance (see Figures 2(b) and (c)), and promoter states were defined in this 3D space.

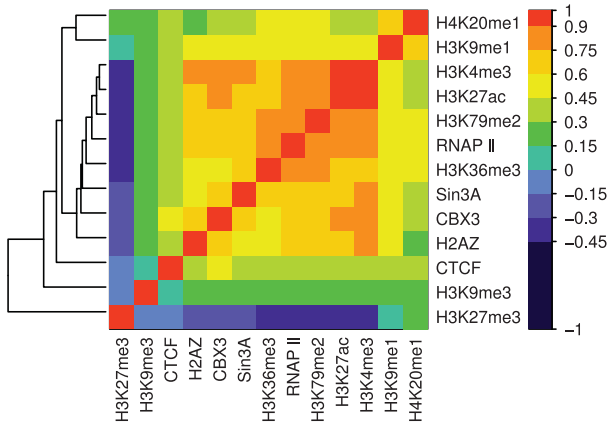


Figure 1. Spearman correlation matrix between epigenetic marks. For each pair of variables, we computed the Spearman correlation over 6 kb windows centered on human gene TSSs. Spearman correlation value is colour coded using the colour map shown on the right. Lines for the thirteen epigenetic marks were reorganized by a hierarchical clustering using Spearman correlation distances as illustrated by the dendrogram on the left of the heat map. This ordering implies that highly correlated epigenetic marks are close to each other.

### *Definition of promoter chromatin states*

Promoter chromatin states were defined as subdivisions of the 3D principal component space (Figure 3). Geometrical definitions of those subdivisions are given below:

$$P_4 = \{(x, y, z) \in \mathbb{R}^3 : x > 1.9, y > 0.5, z > 0.9\} \tag{3}$$

$$P_3 = \{(x, y, z) \in \mathbb{R}^3 : (x - 2.6)^2 + (y - 1)^2 < 1.4, (x, y, z) \notin P_4\} \tag{4}$$

$$P_2 = \left\{(x, y, z) \in \mathbb{R}^3 : y < \frac{4}{3}(x - 2), (x, y, z) \notin P_3 \cup P_4\right\} \tag{5}$$

$$P_1 = \left\{(x, y, z) \in \mathbb{R}^3 : y > \frac{4}{3}(x - 2), (x, y, z) \notin P_3 \cup P_4\right\} \tag{6}$$

where $x, y, z$ are the values along the first PC1, the second PC2 and the third PC3 principal components, respectively.

### *CpG o/e computation and GC content*

CpG observed/expected ratio (CpG o/e) was computed as $\frac{n_{CpG}}{L-l} \times \frac{L^2}{n_C n_G}$, where $n_C$, $n_G$ and $n_{CpG}$ are the numbers of C, G and dinucleotides CG, respectively, counted along the sequence, $L$ is the number of nonmasked nucleotides and $l$ is the number of masked nucleotide gaps plus one, i.e. $L - l$ is the number of dinucleotide sites. The CpG o/e was computed over the sequence after masking annotated CGIs.

### *100 kb resolution chromatin states*

Chromatin states for the myeloid cell line K562 were retrieved from a previous study by the authors (Julienne

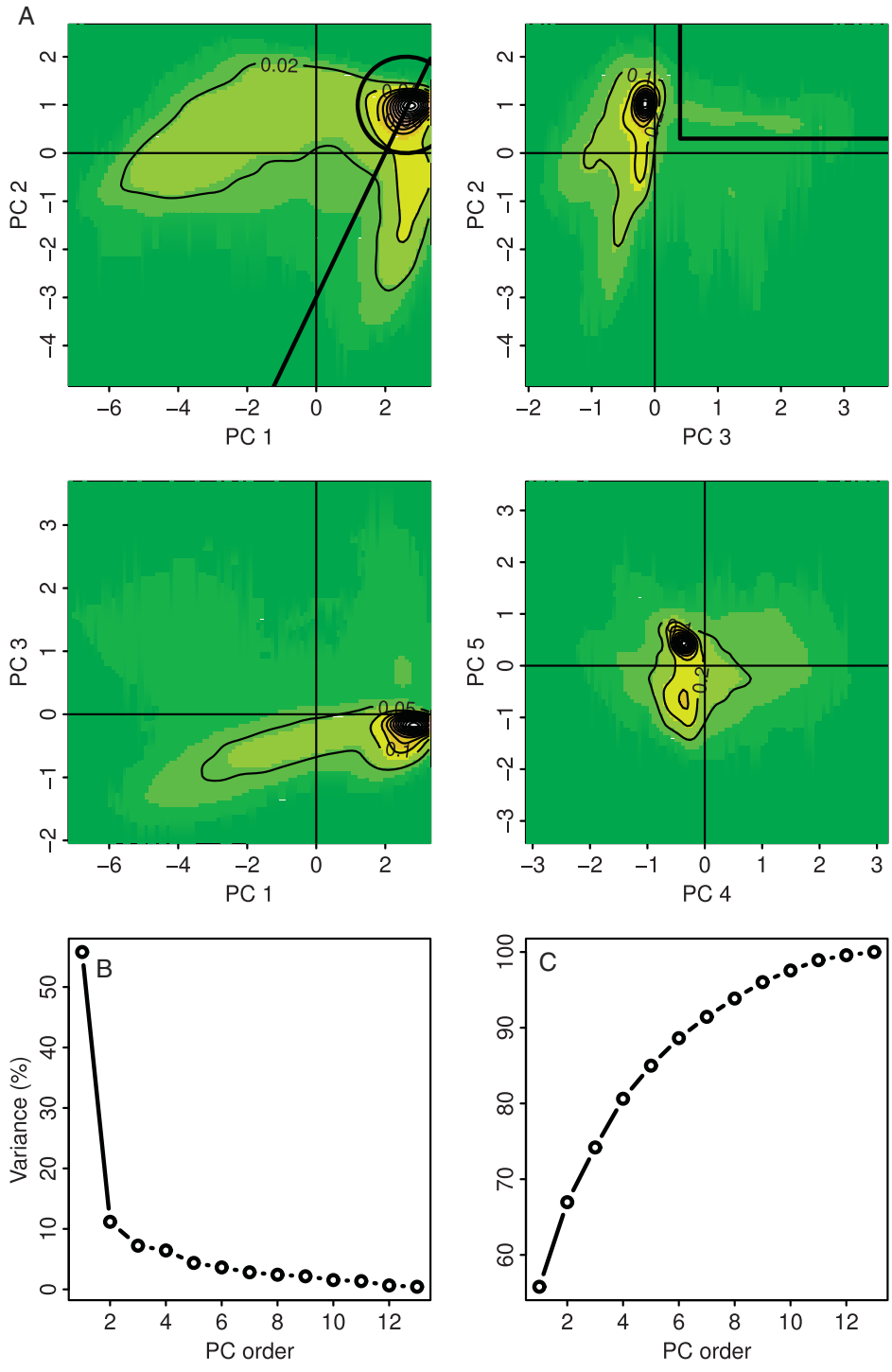


Figure 2. (a) Two-dimensional (2D) projections of the 6 kb promoter data points on the planes defined by (top left) the first (PC1) and second (PC2) principal components (top right) PC3 and PC2, (bottom left) PC1 and PC3 and (bottom right) PC4 and PC5. The density values are indicated by a colour code (white: high density, yellow: moderate density, green: low density) and a contour plot. Densities are computed with a kernel density estimator. The thick solid lines are the boundaries that separate promoter chromatin states P1, P2, P3 and P4 in the the 3D space (PC1, PC2, PC3) as defined in Equations (3) to (6). (b) Percentage of variance accounted for by the first thirteen principal components ordered according to their corresponding variance (eigenvalues). (c) Cumulative variance.

et al. 2013). Large scale chromatin states define an epigenetic segmentation of the human genome in four prevalent chromatin states C1, C2, C3 and C4, respectively, for 27,656 100 kb non-overlapping windows. The large scale chromatin state for a gene is the state of the 100 kb window its TSS is embedded in.

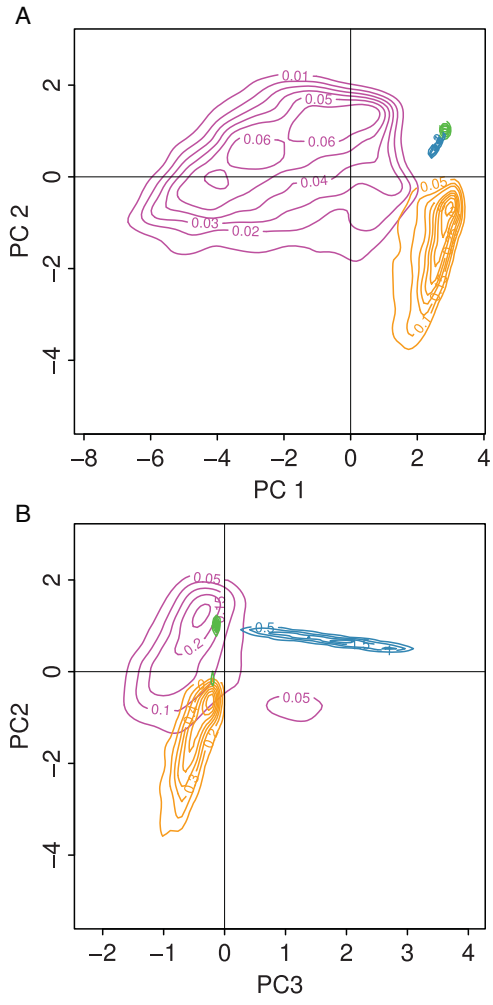


Figure 3. Contour plots of the densities of the four prevalent promoter groups P1 (pink): activated promoters, P2 (orange): Pc repressed promoters, P3 (green) unmarked promoters and P4 (blue): HP1 repressed promoters. The clustering in the 3D space generated by the first three principal components PC1, PC2 and PC3 is defined in Equations (3) to (6) and illustrated in Figure 2(a). (a) 2D-projection on the plane defined by PC1 and PC2; (b) 2D-projection on the plane defined by PC3 and PC2.

### *Promoter count definition*

Promoter count for a gene is the number of promoters that fall in a 100 kb window centered around its TSS. For each gene we compute five kinds of promoter count for:

- all genes. This gives an indication of the gene density around that gene;
- genes which belong to a promoter class giving four promoter counts.

### *Mean replication timing data and replication U-domain coordinates*

Timing profiles for the immature myeloid cell line K562 were obtained from the authors (Baker, Audit et al. 2012). The mean replication timing (MRT) is given for 27,656 100 kb non-overlapping windows in hg18 coordinates. We also retrieved the coordinates of the 876 U-domains in K562 from the authors (Baker, Audit et al. 2012).

## Combinatorial analysis of chromatin marks at human gene promoters

### *Fine-scale analysis of chromatin marks combinatorial complexity*

Mammalian promoter regions are well known to vary significantly in their positional relationships to genes (Kouzarides 2007; Zhou et al. 2011). The DNA sequence proximal to the transcriptional start site (TSS) of a gene is commonly regarded as a proxy region where the study of chromatin marks is likely to provide new insights into the regulatory state of promoters and genes. Here we investigate relationships between the genome-wide distributions of eight histone modifications, one histone variant and four DNA binding proteins in the myelogenous leukemia human cell line K562 around (±3 kb) the 17,872 gene TSS with a valid RPKM (Materials and methods). In Figure 1 is shown a heat map representing the Spearman correlation matrix between epigenetic marks after having reorganized rows and columns with a hierarchical clustering algorithm based on the Spearman correlation distance (Equation (2)). All the epigenetic marks that are known to be involved in transcription positive regulation, namely H4K20me1, H3K9me1, H3K4me3, H3K27ac, H3K79me2, RNAPII, H3K36me3, CBX3, H2AZ, together with the transcription factors CTCF and sin3A, form a block in the correlation matrix, meaning that they are all significantly correlated with each other (Ernst et al. 2011; Zhou et al. 2011). The maximum correlation is obtained between the two active promoter marks H3K4me3 and H3K27ac. Note also the preferential correlation between H4K20me1 and H3K9me1 consistent with previous observations of some enrichment of these marks in promoter or coding regions of active genes (Schotta et al. 2004; Talasz et al. 2005; Vakoc et al. 2006; Barski et al. 2007), with further evidence of significant colocalization (Sims et al. 2006). However there are mainly two lines that stand out from the block of active marks in the hierarchical clustering dendrogram in Figure 1. One of these lines corresponds to the polycomb (Pc) associated repressive chromatin marks H3K27me3 characteristics of the so-called facultative heterochromatin (Barski et al. 2007; Chandra et al. 2012). This is the only mark that anti-correlates with most of the active marks except H4K20me1. The other line corresponds to H3K9me3, commonly considered as a repressive chromatin mark associated with the heterochromatin protein 1 (HP1) known as a major actor in constitutive heterochromatin formation (Barski et al. 2007; Chandra et al. 2012). Surprisingly H3K9me3 is found to moderately correlate with all active marks. This confirms previous observations that this epigenetic modification may also be associated with transcriptional activation. When H3K9me3

is present in the promoter region in combination with all active marks, this may conduct in the anchoring of the $\gamma$ isoform of the HP1 protein (Minc et al. 2000; Li et al. 2002; Kellum 2003; Maison & Almouzni 2004), also called CBX3, which was recently shown to help the splicing of multi-exonic genes (Vakoc et al. 2005; Smallwood et al. 2012).

***Principal promoter chromatin states***

To objectively identify the prevalent combinatorial patterns of the thirteen chromatin marks at human gene promoters, we have performed a PCA (Chessel et al. 2004) to reduce the dimensionality of the data (Materials and methods). As shown in Figure 2, the first three principal components sum up 74% of the total data variance (Figures 2(b) and (c)). By projecting the 6 kb promoter loci on the (PC1, PC2), (PC3, PC2), (PC1, PC3) and (PC4, PC5) planes (Figure 2(a)), it is clear that most of the population is confined in the (PC1, PC2) plane. In this very dense plane, loci mainly lie along two straight lines with a very high density of loci at the intersection of these two lines. A rather wide diluted mode is observed parallel to the PC1 axis, whereas a more populated mode is concentrated along a line parallel to PC2. Furthermore, a simple inspection of the projections on the planes (PC3, PC2) and (PC1, PC3) in Figure 2(a) confirms that loci out of the (PC1, PC2) plane are rather scarce (less than 5% of the human gene promoters). This has led us to phemomenologically define four main promoter chromatin states in the 3D-space defined by Equations (3) to (6). When labeling each of these four promoter chromatin states with a colour, namely P1 (pink), P2 (orange), P3 (green) and P4 (blue), we obtain the density contour plots shown in Figure 3. Among the first three chromatin states that are confined in the (PC1, PC2) plane, P1 is by far the most populated state $N = 9643$ (54.4%) promoter loci as compared to P2 with $N = 3149$ (17.8%) and P3 with $N = 4252$ (24.0%). The fourth promoter chromatin state P4 is the only one that lies outside the (PC1, PC2) plane along a direction parallel to the PC3-axis (Equations (3) and Figure 3(b)). This state contains only $N = 679$ (3.8%) promoter loci, which is dramatically less than the P1, P2 and P3 populations. Since, as we will see in the next sections, P4 will turn out to be a relevant and epigenetically meaningful chromatin state, the fact that classical clustering algorithms similar to k-means would have missed this very poorly populated state (see Mackay (2003) for the limitations of these clustering methods) justifies, *a posteriori*, our phemomenological clustering in the four chromatin states defined by Equations (3) to (6).

*Remark* . When using the Clara clustering algorithm (Kaufman & Rousseeuw 1984) with the number of clusters fixed to four, we miss the chromatin state P4 that is then included in P3, whereas the most populated chromatin state P1 is split into two states. Indeed, Equations (4), (5) and (6) that respectively define the chromatin states P3, P2 and P1 mainly confined in the (PC1, PC2) plane, are inspired from the partitioning provided by the Clara algorithm. Let us point out that the results reported hereafter are robust to slight changes in the parameters in Equations (4) to (6).

***Epigenetic content of the four prevalent promoter chromatin states***

Visualization of the distributions of the thirteen epigenetic marks in each of the four promoter chromatin states in Figures 4 and 5, shows that most marks are not confined to a single promoter chromatin type. Rather, the four main promoter chromatin types are defined by a unique linear combination of these marks.

P1 (pink): active euchromatin state. More than 90% of the 6 kb promoter loci in P1 are associated (positive enrichment) with histone modifications H3K36me3, H3K4me3, H3K27ac and H3K79me2 (Figure 4), the hallmarks of transcriptionally active euchromatin (Barski et al. 2007; Kouzarides 2007; Zhou et al. 2011), as well as with RNA polymerase II (Figure 5) and to a slightly less extent with the RPD3-interacting protein SIN3A (Figure 5) as previously found in active euchromatin in *Drosophila* (Filion et al. 2010). P1 also regroups the majority of H3K9me1 marked promoter loci consistent with previous observation of higher H3K9me1 levels in the TSS surrounding of active promoters (Barski et al. 2007). Most of the promoter regions containing the histone variant H2AZ also belong to P1. The highly conserved histone variant H2AZ has been previously shown to affect nucleosome positioning *in vitro* and *in vivo* (Fan et al. 2002; Schones et al. 2008; Schones & Zhao 2008; Tolstorukov et al. 2009) and to be associated with chromatin activation *in vivo* (Barski et al. 2007; Schones et al. 2008) by contributing, via nucleosome sliding, to the phasing of a nucleosome free region at TSS (Schones & Zhao 2008; Talbert & Henikoff 2010; Vaillant et al. 2010; Arneodo et al. 2011).

P2 (orange): facultative heterochromatin state. P2 is notably associated with the histone modification H3K27me3 (Figure 4). This mark is well known to be recognized by the chromodomains of Pc proteins and to be implicated in gene silencing (Barski et al. 2007; Chandra et al. 2012).

P3 (green): silent 'unmarked' heterochromatin. Out of the four promoter chromatin states, P3 corresponds to promoter loci lacking a clear chromatin mark signature. As shown in Figures 4 and 5, most P3 promoters are not enriched for any available marks. P3 can indeed be compared to the 'null' or 'black' silent heterochromatin states previously found in *Drosophila* (Filion et al. 2010; Sexton et al. 2012) and *Arabidopsis* (Roudier et al. 2011) as covering a significant portion of the genome.

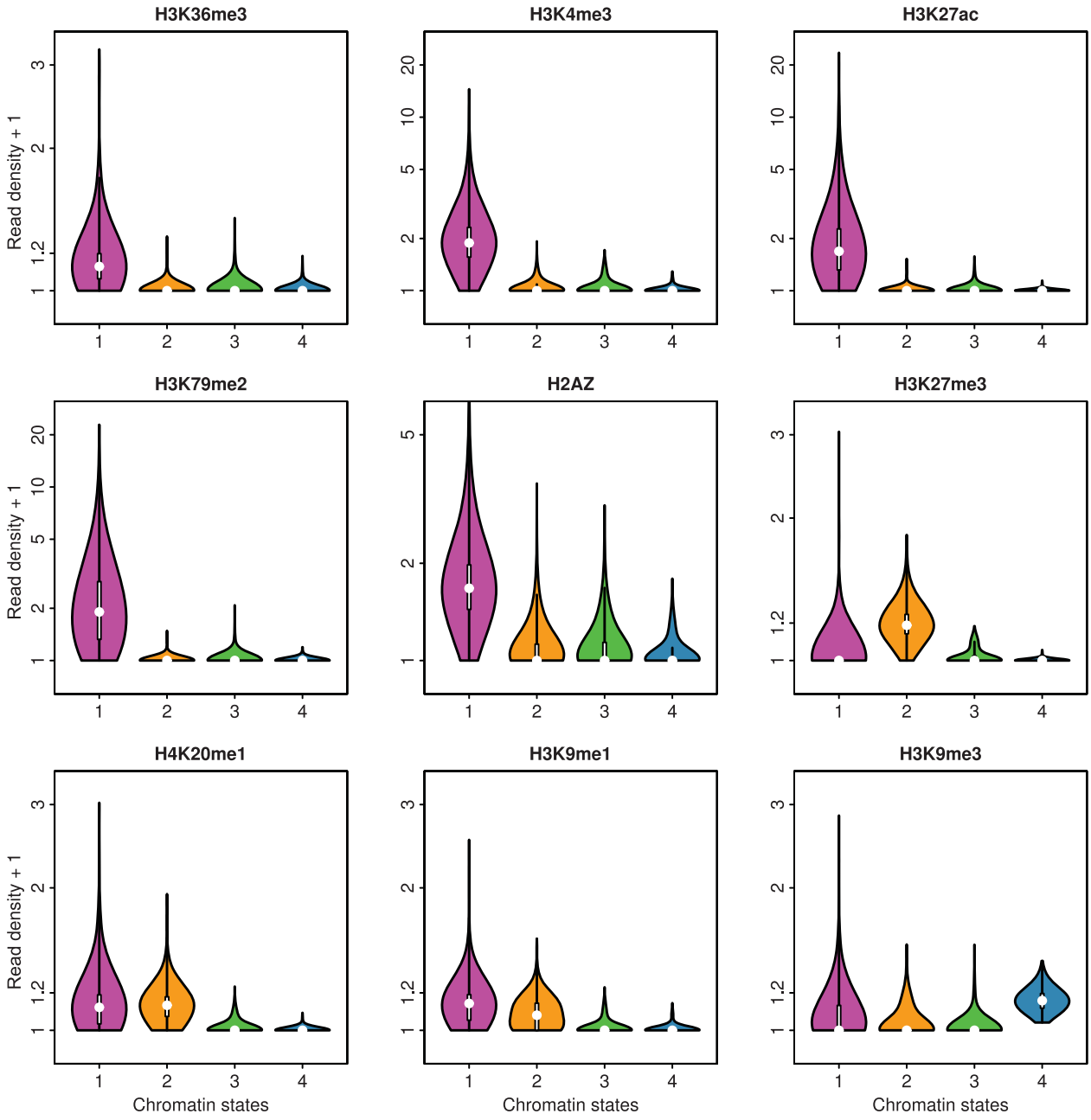


Figure 4. Repartition of histone marks in the four promoter chromatin states P1, P2, P3 and P4. Violin plots of the decimal logarithm of histone mark ChiP-Seq read density in 6 kb window around the TSS per promoter state. Violin plot combines a boxplot (in white) with a symmetric density plot (coloured area). The wider the coloured area is, the more points are associated with this value. Same colour coding as in Figure 3.

P4 (blue): HP1 associated heterochromatin state. P4 corresponds to the few (679) gene promoters containing the H3K9me3 mark and almost only that repressive mark (Figure 4) as the probable signature of its ability to anchor to the heterochromatin protein HP1 at the origin of establishment of heterochromatin (Barski et al. 2007; Chandra et al. 2012).

Methylation of H3K9 is well known to be implicated in heterochromatin formation and gene silencing (Kouzarides 2007; Zhou et al. 2011). The fact that H3K9me1 is found in P1 and to a less extent in P2 and not in P4 (Figure 4) confirms that this epigenetic modification, together with H4K20me1, may also be associated with transcriptional activation (Schotta et al. 2004; Talasz et al. 2005; Sims et al. 2006; Vakoc et al. 2006; Barski et al. 2007). Note that H3K9me3 is not exclusively found in P4 promoter regions; as seen in Figure 4, 42% of P1 promoters and 25% of the P2 promoters contain some H3K9me3 marks. As mentioned in the previous subsection,

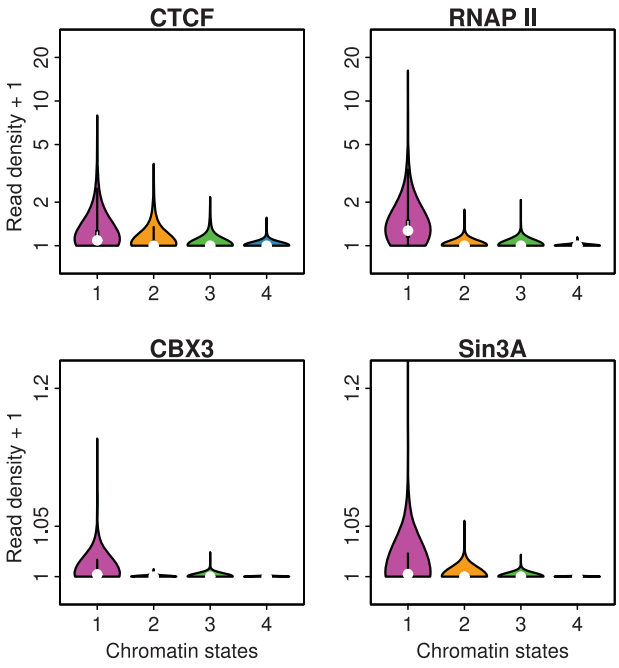


Figure 5. Repartition of transcription factors in the four promoter chromatin states P1, P2, P3 and P4. Violin plots of the decimal logarithm of transcription ChiP-Seq read density in 6 kb window around the TSS per promoter state. Same colour coding as in Figure 3.

when present in combination with all active marks, this mark may drive the anchoring of CBX3 (Figure 5) involved in gene splicing (Minc et al. 2000; Li et al. 2002; Kellum 2003; Maison & Almouzni 2004; Vakoc et al. 2005; Smallwood et al. 2012).

The insulator binding protein CTCF is known to establish chromatin boundaries to prevent the spreading of heterochromatin into transcriptionally active regions (Barski et al. 2007; Chandra et al. 2012). As shown in Figure 5, consistent with this picture, we get, in good agreement with previous observations in *Drosophila* (Filion et al. 2010; Sexton et al. 2012), that CTCF is found in P1 promoters and to a slight extent in P2 promoters. This can be understood by the fact that P1 and P2 genes lie together in gene rich, high GC megabase-sized domains of intermingled active euchromatin and facultative heterochromatin regions (see following sections).

To summarize, this simple classification into one active promoter chromatin state (P1) and three repressed promoter chromatin states (P2, P3 and P4) of human genes is strikingly similar to those recently reported in *Arabidopsis* (Roudier et al. 2011) and *Drosophila* (Filion et al. 2010; Sexton et al. 2012) suggesting the possible existence of some simple principles of epigenetic regulation of eukaryotic genomes.

### *A synthetic view of epigenetic regulation of gene activity*

#### *Gene expression*

As shown in Figure 6(a), when investigating gene expression data (Materials and methods), we find that a vast majority (8312, 88%) of expressed gene promoters with a $RPKM > 1$ [Equation (1)] are in the euchromatin state P1. As expected, most (2779, 89%) of the Pc repressed P2 promoters correspond to non-expressed genes. Interestingly, we find that the number of non-expressed genes in P1 (1250) is non-negligible and comparable to the one in P2 (2779). Most of the promoters in the heterochromatin states P3 (3124, 81%) and P4 (609, 91%) correspond to silent genes except for a minority of them.

#### *CpG-rich versus CpG-poor promoters*

Mammalian promoters can be classified according to their sequence content. Most promoters coincide with regions of high GC content and CpG ratio (or CpG islands) (Gardiner-Garden & Frommer 1987; Antequera & Bird

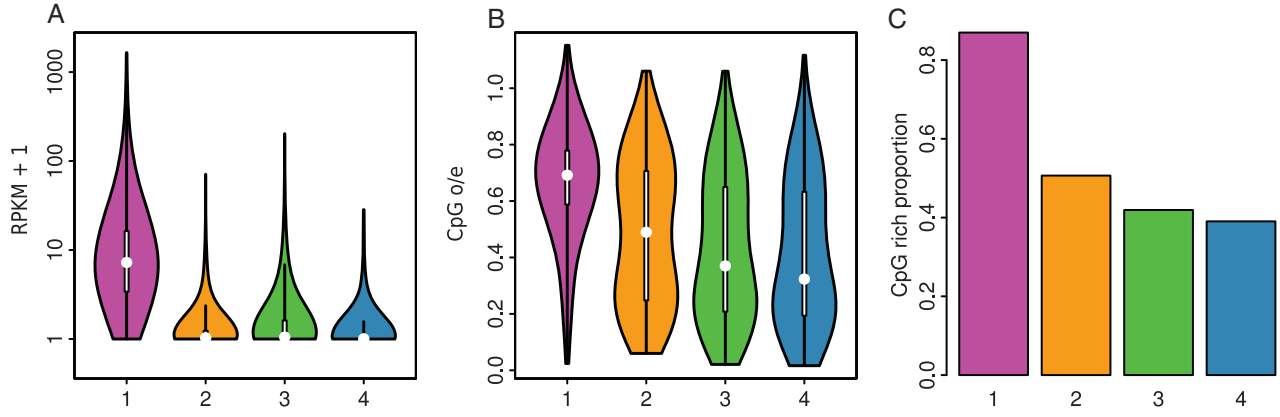


Figure 6. Expression level and CpG content in the four promoter chromatin states P1, P2, P3 and P4. (a) Violin plots of the decimal logarithm of RPKM expression score (see Materials and methods) in the four promoter states. (b) Violin plots of CpG o/e computed in the 6 kb windows around the TSS per promoter states. (c) Proportion of CpG rich genes per promoter state (a promoter is CpGrich if the CpGo/e around its TSS is above 0.48). Same colour coding as in Figure 3.

1993; Ponger et al. 2001; Antequera 2003; Suzuki & Bird 2008). As already noted by others (Saxonov et al. 2006; Tang & Epstein 2007; Weber et al. 2007; Mohn & Schüubeler 2009), the distribution of CpG enrichment is bimodal which is also the case in other mammalian genomes, including the mouse genome. As proposed in a previous work (Zaghloul et al. 2012), we can use a threshold value $r^*$ (0.48 in Figures 6(b) and (c)) so that promoters with a CpG enrichment $> 0.48$ are considered CpG rich and with CpG enrichment $< 0.48$ CpG poor (Materials and methods). These two classes of promoters have different regulations and present different characteristics. Whereas CpG-poor genes have a specific initiation site, usually a TATA-box, CpG-rich genes have a broad initiation site (Carninci et al. 2006). Besides, CpG-rich promoters evolve more rapidly than CpG-poor ones. A hypothesis on the origin of these two gene categories was proposed in Mohn and Schübeler (2009) but not investigated further: these two categories could have a different evolutionary history, with CpG-rich genes being the oldest ones, present before the global methylation appeared on vertebrate genomes (Gardiner-Garden & Frommer 1987; Antequera 2003) and CpG-poor being more recent. As shown by the violin plot of CpGo/e in Figure 6(b), gene promoter loci in P1 are significantly enriched in CpG as compared to P2, P3 and P4 promoter loci. We clearly find a significant shift of the CpG pdf to smaller values when going from P1 ($\overline{\text{CpG } o/e} = 0.69$)) to P2 ($\overline{\text{CpG } o/e} = 0.49$), P3 ($\overline{\text{CpG } o/e} = 0.37$)) and P4 ($\overline{\text{CpG } o/e} = 0.32$)). Thus relative to the genome average 0.57, the P1 promoter loci are clearly CpG-rich. In terms of promoter states previously defined, 87% of P1 promoter loci belong to the CpG-rich class as compared to 51% of P2, 42% of P3 and 39% of P4 promoter loci. Thus, a non-negligible proportion of gene promoter loci in the repressed heterochromatin states P2, P3 and P4 are CpG-rich but mostly non-expressed in K562 human cell line.

## Interplay between promoter activity and large-scale chromatin environment

### *Distribution of promoter states in the four prevalent large-scale chromatin states*

In our previous work (Julienne et al. 2013), we identified four main large-scale chromatin states C1, C2, C3 and C4 that were respectively found in 6572 (23.8%), 5312 (19.2%), 6603 (23.9%) and 6758 (24.4%) loci among the 27,656 100 kb loci with a defined MRT. Note that we removed from the analysis 2411 (8.7%) loci that were not properly classified in any of these chromatin states. To address the question of the gene content of these four chromatin states, we used a data set of 17,724 genes whose promoters have a valid epigenetic value for the considered 13 epigenetic marks. Some of these genes (1832) were not taken into account in our analysis because their promoters did not belong to any C1, C2, C3 or C4 100 kb loci.

Table 1. Density of promoters per Mbp in the four large scale chromatin states C1, C2, C3 and C4, for the four epigenetic promoter states, P1, P2, P3 and P4.

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| P1 | 11.8 | 0.53 | 0.1 | 0.06 |
| P2 | 1.29 | 2.98 | 0.19 | 0.01 |
| P3 | 1.6 | 1.18 | 2.03 | 0.29 |
| P4 | 0.13 | 0.05 | 0.02 | 0.75 |

Table 2. Number of promoters P1, P2, P3 and P4 in large scale chromatin states C1, C2, C3 and C4.

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| P1 | 8797 | 304 | 57 | 41 |
| P2 | 961 | 1721 | 113 | 7 |
| P3 | 1193 | 682 | 1191 | 191 |
| P4 | 99 | 26 | 14 | 495 |

The mean density of the 15,892 genes that belong to one of the four large-scale chromatin states is 6.25 promoters per Mb. As reported in Tables 1 and 2, the early replicating active euchromatin state C1 is highly enriched in gene promoters (14.82 promoters/Mb) and harbours 69.5% of gene promoters even though it represents about 25% of the total genome coverage by the four large-scale chromatin states. The mid S facultative heterochromatin state C2 also contains a non-negligible percentage (17.2%) of gene promoters that indeed corresponds to a modest density 4.74 promoters/Mb. The late replicating unmarked and HP1-associated heterochromatin states C3 and C4 are genuinely gene poor with very low gene densities 2.34 promoter/Mb and 1.11 promoter/Mb for a total of 8.6% and 4.7% of gene promoters, respectively. Let us point out that the mean gene length increases gradually from C1 (42.5 kb), to C2 (59.4 kb), C3 (83.5 kb) and C4 (133.1 kb), which explains why the gene coverage decreases less abruptly than the promoter density, with C1 mainly genic (62.9%), C2 modestly genic (49.8%) and C3 (39.5%) and C4 (29.3%) mostly intergenic.

As reported in Table 3, when comparing the data in Table 2 and the expected promoter number if the probability

Table 3. Observed/expected ratio of a promoter Pi to be in a large-scale chromatin state Cj. The expected number is given by $\frac{n_{Pi} * n_{Cj}}{N}$ where $n_{Pi}$ is the number of promoters in $P_i$, $n_{Cj}$ the number in $C_j$ and N the total number of promoters.

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| P1 | 1.38 | 0.19 | 0.07 | 0.1 |
| P2 | 0.49 | 3.57 | 0.47 | 0.05 |
| P3 | 0.53 | 1.22 | 4.23 | 1.27 |
| P4 | 0.22 | 0.24 | 0.26 | 16.91 |

of belonging to any promoter state Pi were independent from the probability of being in the chromatin state Cj, we find observed/expected ratio values significant greater than 1 for the four (Pi/Ci) associations as the signature of an increasing dependency from (P1/C1) (1.38), to (P2/C2) (3.57), (P3/C3) (4.23) and (P4/C4) (16.91). In contrast, the observed/expected ratio values obtained for the $(Pi,Cj)_{i\neq j}$ associations are all smaller than 1 as an indication of some anti-correlation except for (P3, C2) (1.22) and (P3, C4) (1.27) which shows that unmarked P3 promoters are more abundant than expected in both the facultative C2 and C4 heterochromatin states.

### *Conditional analysis of promoter activity and large-scale chromatin environment*

In Table 4, we have expressed the results reported in Table 2 in terms of the probability of a promoter to be classified in the promoter state Pi knowing that it is embedded in the large-scale chromatin state Cj. The large scale unmarked C3 and HP1-associated C4 states likely corresponding to nuclear lamina pericentric heterochromatin (Barski et al. 2007; Chandra et al. 2012; Zullo et al. 2012) only contain silent genes with P3 and P4 promoters (∼90%). If large-scale transcriptional activity in C1 euchromatin state is recovered in a large majority (∼80%) of genes with P1 promoters, it does not exclude the presence of inactive genes with P2 (9%) and P3 (11%) promoters. Large-scale facultative heterochromatin state C2 is not very predictive of promoter states since besides a majority of Pc repressed P2 gene promoters (63%) it also contains a significant and non-negligible proportion of silent unmarked P3 (25%) and of active P1 (11%) promoters.

Reciprocally, when revisiting the results in Table 2 in terms of the probability of a promoter in a given promoter state Pi to be in large-scale chromatin environment Cj, we find in Table 5 that with very high probability (96%) P1 promoters have an active euchromatin C1 environment. This contrasts with the Pc repressed P2 promoters that in the majority (61%) belong to the corresponding large-scale facultative heterochromatin C2, but with a significant proportion of them (35%) being contained in an active C1 environment. The unmarked P3 promoters are rather evenly distributed in C1 (37%), C2 (21%) and C3 (37%). Let us point out that the poorly populated P4 promoter state is consistently found in the majority (78%) in the corresponding constitutive heterochromatin state C4 but also in the gene rich euchromatin state C1 (16%) where 1/3 (resp. 2/3) of them are expressed (resp. silent) genes.

Table 4. Transition matrix from large-scale chromatin states to promoter states. Probability of being classified in the promoter state Pi knowing that the promoter is embedded in the large scale chromatin state Cj.

| | from C1 | from C2 | from C3 | from C4 |
|---|---|---|---|---|
| to P1 | 0.79 | 0.11 | 0.04 | 0.06 |
| to P2 | 0.09 | 0.63 | 0.08 | 0.01 |
| to P3 | 0.11 | 0.25 | 0.87 | 0.26 |
| to P4 | 0.01 | 0.01 | 0.01 | 0.67 |

Table 5. Transition matrix from promoter states to large scale chromatin states. Probability that a promoter in the class Pi to be embedded in the large scale chromatin state Cj.

| | to C1 | to C2 | to C3 | to C4 |
|---|---|---|---|---|
| from P1 | 0.96 | 0.03 | 0.01 | 0 |
| from P2 | 0.35 | 0.61 | 0.04 | 0 |
| from P3 | 0.37 | 0.21 | 0.37 | 0.05 |
| from P4 | 0.16 | 0.04 | 0.02 | 0.78 |

Further understanding of these results can be obtained when taking into account gene density. As shown in Figure 7(a), when classifying promoters according to gene promoter number in their 100 kb neighbourhood, we see that the proportion of active P1 promoter increases when increasing the local promoter count at the expense of the proportions of inactive P2, P3 and P4 promoters. Even more spectacular, similar tendencies are observed in Figure 7(b) when considering now the relative proportions of consistent pairing (Pi, Ci) of a promoter Pi embedded in the corresponding large-scale chromatin environment Ci, when increasing the local density of promoters of the same state Pi. As expected the proportion of transcriptionally active pairing (P1, C1) increases when the 100 kb windows surrounding a P1 promoter contains more and more P1 promoters. Naively we would have expected the same increase in the probability of an inactive promoter P2, P3 or P4 to be embedded in the corresponding heterochromatin environment C2, C3 or C4, respectively, when enriching its 100 kb neighbourhood in promoters belonging to the same promoter state. However, this is only true for HP1-associated promoters. This observation is consistent with P4 promoters being mostly in a separated nuclear compartment (Table 5). For promoter states P2 and P3, the pairing (Pi,Ci) doesn't increase with promoter density. Indeed, as shown in Figure 8 (upper left panel), this is only true if this neighbourhood contains no P1 promoter. As soon as one or more P1 promoters belong to the neighbourhood of a P2 or P3 promoter, then the probability for this promoter to be embedded in the gene rich euchromatin state C1 increases (Figure 8, other panels), which explains the observed behaviour of the proportions of inactive pairing (P2, C2) and (P3, C3) in Figure 7(b). Conversely to the P1 promoter, the presence of one P4 promoter doesn't imply a C4 environment suggesting that a P4 promoter is not sufficient to drive the association with the pericentric compartment (data not shown). Altogether these results confirm that gene density is a key parameter underlying the coherence between promoter activity and a large-scale chromatin environment.

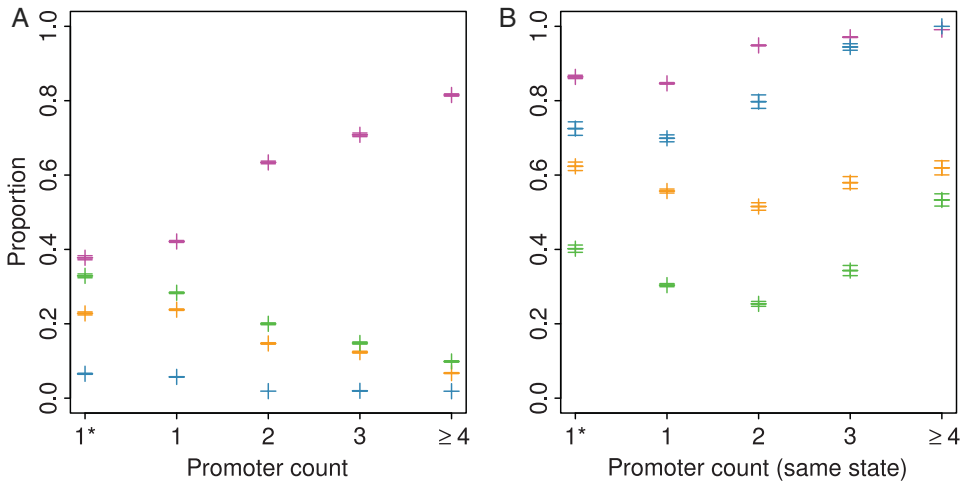


Figure 7. Effect of local promoter density on large-scale chromatin state. Promoter count for a gene is the number of promoters that fall in a 100 kb window centered around its TSS. The more the promoter count is high, the more gene rich is the surrounding region. A promoter count of 1* means that the gene is isolated and that its length is smaller than 50 kb (so that the surrounding of this gene is mostly intergenic). Promoter count (same state) is the promoter count taking into account only genes with the same promoter state as the considered gene. (a) Proportions of promoter states P1, P2, P3 and P4 with respect to promoter count. (b) Proportion of promoters with a large chromatin state corresponding to their promoter state (e.g. P1 in C1 etc.) with respect to promoter count (same state). Same colour coding as in Figure 3.

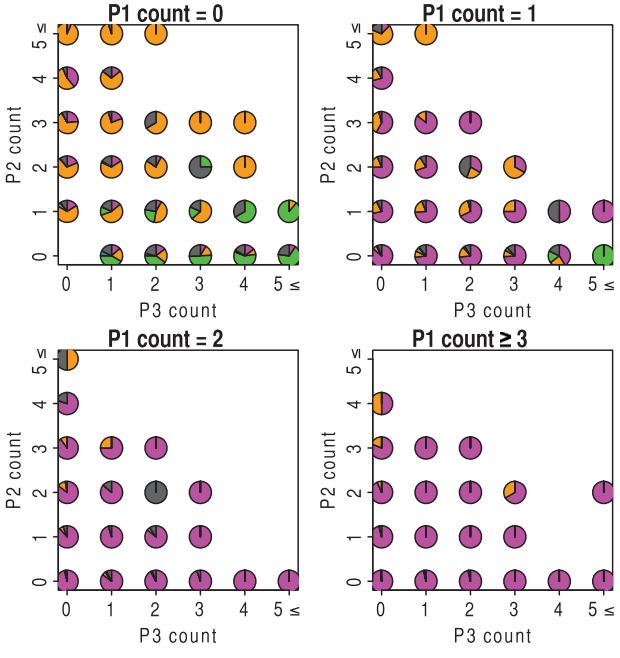


Figure 8. Large-scale chromatin states with respect to promoter counts. Pi count of a given gene is the number of Pi promoters that fall in a 100 kb window centered around its TSS. Each panel corresponds to a different active P1 promoter count. For each possible value of the three considered promoter counts (P1, P2, P3), we calculated the proportions of large-scale chromatin states C1 (pink), C2 (orange), C3 (green) and C4 (blue); these proportions are represented by a pie chart. Because the P4 promoter state is poorly populated (Table 2), we have fixed P4 count = 0.

## Repartition of promoter chromatin states along human chromosomes

### *Distribution of promoter chromatin states inside replication timing U-domains*

When first concentrating on the gene distribution inside the 876 replication timing U-domains previously identified in K562 cells (Baker, Audit et al. 2012), we reveal a remarkable organization of the four prevalent promoter chromatin states. This is particularly patent in Figure 9(a)–(d) where the 876 U-domains were centered and ordered vertically from the smallest (top) to the largest (bottom) and only gene promoters are represented. By simple visual inspection, we recognize in Figure 9(a) the edges of the U-domains from the local enrichment of active P1 promoters that are mainly confined in a closed (∼150 kb) C1 neighbourhood of the 'master' replication origins that border these replication domains (Julienne et al. 2013). Note that this result is quite consistent with the previous observation (Zaghloul et al. 2012) that CpG-rich gene promoters that are likely to be active in the germ line and do present an important transcription-associated nucleotide compositional asymmetry (Green et al. 2003; Touchon et al. 2003, 2004; Baker et al. 2010), also lie preferentially nearby the edges of replication skew N-domains. In Figure 9(b), the Pc repressed P2 promoters are mostly found at finite distance (∼200–300 kb) from U-domain borders whose centers are significantly devoided of P2 promoters. In small U-domains (< 1.2 Mb), P2 promoters mainly occupy their centers that are replicated in mid-S phase. In contrast

unmarked P3 promoters do not seem to have any preferential positioning inside U-domains where they look rather homogeneously distributed as shown in Figure 9(c). Despite their small number, inactive HP1-associated P4 promoters are mostly found in the central region of large (> 1 Mb) U-domains in Figure 9(d); they consistently lie in a late replicating heterochromatin C4 environment (Julienne et al. 2013). As confirmed on the corresponding mean occupation profiles in Figure 9(e), this remarkable organization of gene promoters inside U-domains is consistent with the gradient of chromatin states observed across these replication domains, from C1 at U-domain borders followed by C2, C3 and C4 at centers (Julienne et al. 2013). Note that as shown in Figures 9(f) and (g), a similar organization is found for CpG-rich and CpG-poor promoters, respectively, except that CpG-poor P1 promoters are about one order of magnitude less numerous than CpG-rich P1 promoters.

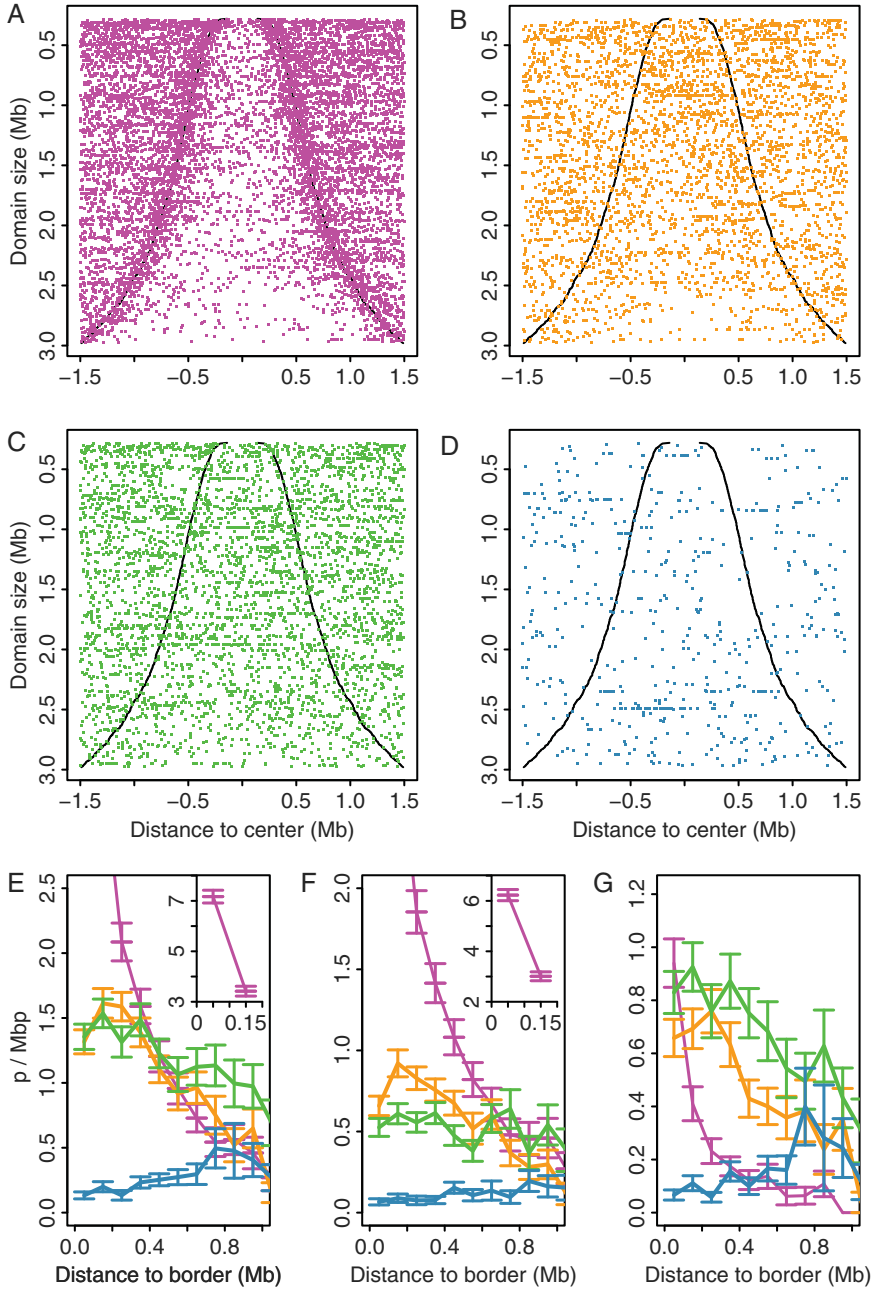


Figure 9. Distribution of promoter states inside replication U-domains. (a) The 876 K562 U-domains were centered and ordered vertically from the smallest (top) to the largest (bottom). All active P1 promoters are represented by a dot (pink). (b) Same as (a) for Pc repressed P2 promoters (orange). (c) Same as (a) for the unmarked promoters P3 (green). (d) Same as (a) for HP1-repressed promoters P4. (e) Mean promoter density with respect to the distance to the closest U-domain border. Error bars represent standard deviation. Same colour coding as in (a)–(d). (f) Same as (e) for CpG rich genes. (g) Same as (e) for CpG poor genes.

Table 6. Distribution of promoter chromatin states P1, P2, P3 and P4 inside replication U-domains, (C1 + C2) blocks and (C3 + C4) blocks (Julienne et al. 2013).

| | U-domains | C1 + C2 | C3 + C4 |
|---|---|---|---|
| Total length (Mb) | 1293.9 | 750.6 | 745.5 |
| Mean length (kb) | 1431.3 | 561.8 | 723.1 |
| promoter number | | | |
| P1 | 3029 | 6224 | 197 |
| P2 | 1550 | 1449 | 103 |
| P3 | 1656 | 1218 | 826 |
| P4 | 306 | 70 | 285 |
| density of promoters per Mb | | | |
| P1 | 2.3 | 8.29 | 0.26 |
| P2 | 1.20 | 1.93 | 0.14 |
| P3 | 1.28 | 1.62 | 1.1 |
| P4 | 0.24 | 0.09 | 0.38 |

### *Distribution of promoter chromatin states outside replication U-domains*

Replication timing U-domains actually cover about 50% of the human genome. In our previous study (Julienne et al. 2013), we have shown that the other half of the human genome is more in agreement with the dichotomic picture proposed in early studies of the mouse (Farkash-Amar et al. 2008; Hiratani et al. 2008, 2010) and human (Desprat et al. 2009; Ryba et al. 2010; Yaffe et al. 2010) genomes, where early and late replicating regions occur in separated compartments of open and close chromatin, respectively.

- High GC, gene rich (C1 + C2) blocks: About 25% of the human genome (Table 6) are covered by megabase-sized GC-rich (C1 + C2) chromatin blocks that on average replicate early by multiple almost synchronous origins (e.g. the region from 151.5 to 155.8 Mb of human chromosome 1 in Figure 10(a)). As reported in Table 6, these regions are gene rich with a high density of P1 promoters (6.85 promoters/Mb) and a significant density of P2 promoters (2.15 promoters/Mb) that replicate slightly earlier than the mid-S phase P2 promoters found in replication timing U-domains. Some unmarked P3 promoters (1.41 promoters/Mb) also belong to these (C1 + C2) blocks and correspond to the sub-class of genes with P3 promoters that are expressed in K562. Only a few P4 promoters (0.09 promoters/Mb) are found in these early replicating (C1 + C2) block regions.
- Low GC, gene poor (C3 + C4) blocks: The last 25% of the human genome correspond to megabase-sized GC-poor domains of interspersed (C3 + C4) heterochromatin states or of long C4 domains that on average replicate late by again multiple almost coordinated origins (e.g. the region from 185 to 190 Mb of human chromosome 1 in Figure 10(b)). As reported in Table 6, these regions are gene deserts with, relatively to their genome mean densities, almost no P1 (0.17 promoters/Mb) and P2 (0.10 promoters/Mb) promoters, and in contrast contain most of the P4 promoters (0.43 promoters/Mb) as well as a significant proportion of P3 promoters.

As reported in Figure 6, P1 and P2 promoters are in the large majority CpG rich, which further indicates that C1 + C2 blocks are enriched in CpG-rich gene promoters consistent with previous observations that CpG-rich genes tend to be seated in high GC isochores (Julienne et al. 2013).

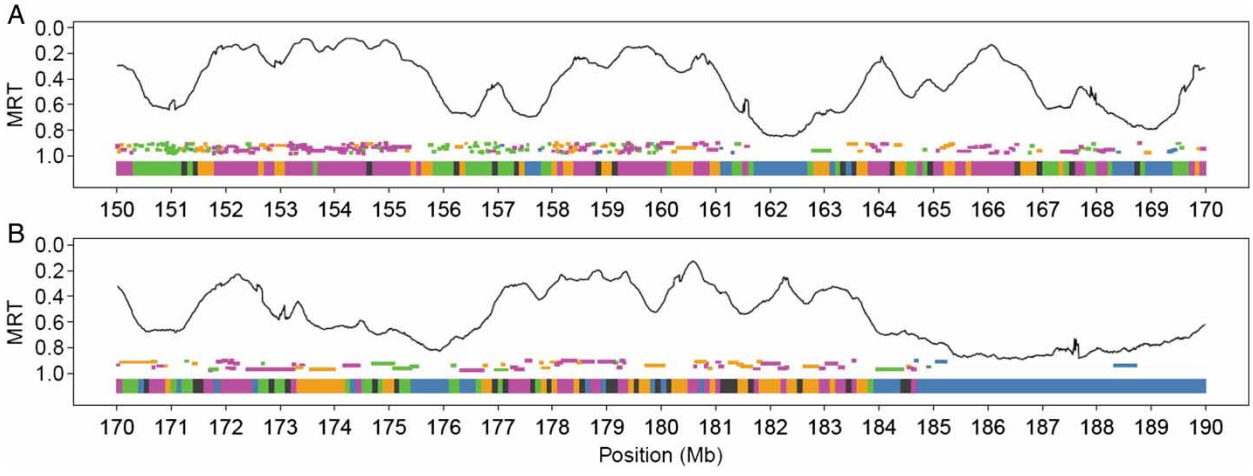


Figure 10. Distribution of promoter states along the MRT profile. (a) K562 MRT profile along a 20 Mb long fragment of human chromosome 1. Below the MRT profile, gene positions are indicated by a segment. The segment colour indicates the promoter state. Same colour coding as in Figure 9. At the bottom of the plot, the chromatin state of each 100 kb window is represented with the following coding: active euchromatin state C1 (pink), Pc repressed facultative heterochromatin C2 (orange), silent unmarked heterochromatin state C3 (green) and HP1-associated heterochromatin state C4 (blue) (Julienne et al. 2013. (b) Same as (a) for the following 20 Mb fragment of the human chromosome 1.

In contrast, C3 + C4 blocks, as the low GC isochores counterpart, contain only a few genes mostly inactive and with a CpG-poor promoter.

## Conclusion/perspectives

In summary, the integrative analysis of epigenetic mark maps in the myelogenous leukemia human cell line K562 has shown that, at the gene promoter scale (±3 kb around TSS), the combinatorial complexity of these epigenetic data can be reduced to four prevalent promoter chromatin states that display remarkable similarties with those found in different cell types in *Drosophila* (Sexton et al. 2012) and *Arabidopsis* (Roudier et al. 2011): P1 regroups all the marks of transcriptionally active chromatin and corresponds to CpG-rich promoters of highly expressed genes; P2 is notably associated with the histone modification H3K27me3 that is the mark of Pc repressed falcultative heterochromatin; P3 corresponds to promoters that are not enriched in any marks as the signature of silent heterochromatin; and P4 characterizes the few gene promoters that contain only the HP1-associated histone modification H3K9me3. When analysing the coherence between promoter activity (P1, P2, P3 ad P4) and the corresponding large-scale (100 kb) chromatin states (C1, C2, C3 and C4) that were shown to replicate at different periods of the S-phase (Julienne et al. 2013), we confirm gene density as a central parameter underlying the interplay between transcription and replication. Among the striking results obtained about the large-scale chromatin environment from the local knowledge of a gene-promoter activity is the fact that a P1 promoter is almost surely surrounded by an early replicating, gene-rich, transcriptionally active euchromatin state C1. Reciprocally, it is the spreading of the late replicating, gene-poor, HP1-associated heterochromatin large-scale state C4 that almost surely governs the local inactivity of the few unmarked P3 and constitutively silent P4 promoters. When further investigating the spatial distribution of the P1, P2, P3 and P4 promoters along human chromosomes, our study reveals a remarkable gene organization in relation with the MRT. In 50% of the human genome that are covered by megabase-sized replication U-domains (Baker, Audit et al. 2012; Julienne et al. 2013), a significant enrichment of highly expressed P1 genes is observed in a closed neighbourhood of the early C1 initiation zones that border these domains. P2 promoters are mainly found in the mid-S C2 environment at finite distance (∼200–300 kb) from U-domain borders. Inactive P3 and P4 promoters are distributed more homogeneously inside U-domains with a majority of the poorly populated P4 promoter set in the C4 central region of large U-domains likely associated with pericentric nuclear heterochromatin. Thus, in these U-domains where the replication wave starting at bordering 'master' replication origins, keeps accelerating thanks to the firing of secondary origins (Guilbaud et al. 2011), some gradient of gene promoter activity is also observed as the possible consequence of some epigenetic co-regulation of replication and transcription. This intimate relationship between gene activity and MRT is also observed in the other half of the human genome with mainly P1 and P2 promoters in megabase-sized GC-rich and highly genic (C1 + C2) chromatin blocks that replicate early in the S-phase, and P3 and P4 promoters in late replicating, gene-poor and GC-poor megabase-sized (C3 + C4) blocks (Julienne et al. 2013).

Extending this study to different cell types including ES, somatic and cancer cells looks very promising. Previous comparative analyses of replication timing profiles during development have revealed important dynamical changes leading to cell type specific patterns of replication (Hiratani et al. 2008, 2010; Ryba et al. 2011). Importantly, these specific replication timing patterns are conserved between human and mouse syntenic regions of related cell types despite the length of evolutionary divergence (Ryba et al. 2010). Thus MRT profiles likely capture the epigenetic differences between cell types, even when they are closely related, and should be considered as a *bona fide* epigenetic mark (McNairn & Gilbert 2003; Hiratani & Gilbert 2009). By performing our integrative analysis at low (100 kb) and high (6 kb) resolutions in parallel, we should be in position to investigate the global reorganization of replication domains during differentiation in relation to coordinated changes in chromatin state and gene expression. A number of studies have also demonstrated a clear association between the replication program and cancer genome rearrangement events (Letessier et al. 2011; De & Michor 2011a, 2011b; Ryba et al. 2012). In particular, MRT was shown to capture important epigenetic modifications involved in genomic misregulation and chromosomal instability during tumoral progression prior to rearrangement events (Ryba et al. 2012). Extending the present study to cancer cell lines with well defined temporally ordered steps of tumoral progression will provide new knowledge that hopefully will turn out very helpful for cancer diagnosis, prognosis and cancer treatment. This work is under progress.

## Acknowledgements

We are very grateful to Y. d'Aubenton-Carafa, A. Baker, JC. Cadoret, E. Cascales, CL. Chen, L. Duret, A. Goldar, O. Hyrien, F. Picard, MN. Prioleau, C. Thermes and C. Vaillant for helpful discussions.

## Funding

This work was supported by ANR (REFOPOL, ANR 10 BLAN 1615).

## References

Aladjem, MI. 2007. Replication in context: dynamic regulation of DNA replication patterns in metazoans. Nat Rev Genet. 8:588–600.

Alberts B, Bray D, Lewis J, Raff M, Roberts, K., Watson JD. 2002. Molecular biology of the cell. 4th ed. New York: Garland Publishing.

Antequera, F. 2003. Structure, function and evolution of CpG island promoters. Cell Mol Life Sci. 60:1647–1658.

Antequera F, Bird A. 1993. Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci USA. 90:11995–11999.

Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton-Carafa Y, Thermes C. 2011. Multi-scale coding of genomic information: From DNA sequence to genome structure and function. Phys Rep. 498:45–188.

Audit B, Baker A, Chen CL, Rappailles A, Guilbaud G, Julienne H, Goldar A, d'Aubenton Carafa Y, Hyrien O, Thermes C, Arneodo A. 2013. Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. Nat Protoc. 8:98–110.

Audit B, Nicolay S, Huvet M, Touchon M, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2007. DNA replication timing data corroborate in silico human replication origin predictions. Phys Rev Lett. 99:248102.

Audit B, Zaghloul L, Baker A, Arneodo A, Chen CL, d'Aubenton-Carafa Y, Thermes C. 2012. Megabase replication domains along the human genome: relation to chromatin structure and genome organisation. Subcell Biochem. 61: 57–80.

Audit B, Zaghloul L, Vaillant C, Chevereau G, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2009. Open chromatin encoded in DNA sequence is the signature of "master" replication origins in human cells. Nucleic Acids Res. 37:6064–6075.

Baker A, Audit B, Chen CL, Moindrot B, Leleu A, Guilbaud G, Rappailles A, Vaillant C, Goldar A, Mongelard F, et al. 2012. Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. PLoS Comput Biol. 8:e1002443.

Baker A, Chen CL, Julienne H, Audit B, d'Aubenton Carafa Y, Thermes C, Arneodo A. 2012. Linking the DNA strand asymmetry to the spatio-temporal replication program: II. Accounting for neighbor-dependent substitution rates. Eur Phys J E. 35:123.

Baker A, Julienne H, Chen CL, Audit B, d'Aubenton Carafa Y, Thermes C, Arneodo A. 2012. Linking the DNA strand asymmetry to the spatio-temporal replication program. I. About the role of the replication fork polarity in genome evolution. Eur Phys J E. 35:92.

Baker A, Nicolay S, Zaghloul L, d'Aubenton-Carafa Y, Thermes C, Audit B, Arneodo A. 2010. Wavelet-based method to disentangle transcription- and replication-associated strand asymmetries in mammalian genomes. Appl Comput Harmon Anal. 28:150–170.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. Cell. 129:823–837.

Bell SP, Dutta A. 2002. DNA replication in eukaryotic cells. Annu Rev Biochem. 71:333–374.

Belmont AS, Dietzel S, Nye AC, Strukov YG, Tumbar T. 1999. Large-scale chromatin structure and function. Curr Opin Cell Biol. 11:307–311.

Berezney R. 2002. Regulating the mammalian genome: the role of nuclear architecture. Adv Enzyme Regul. 42:39–52.

Berezney R, Dubey DD, Huberman JA. 2000. Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. Chromosoma. 108:471–484.

Bernstein BE, Meissner A, Lander ES. 2007. The mammalian epigenome. Cell. 128:669–681.

Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, Marin JM, Lemaitre JM. 2012. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat Struct Mol Biol. 19:837–844.

Bickmore WA, van Steensel B. 2013. Genome architecture: domain organization of interphase chromosomes. Cell. 152:1270–1284.

Bogan JA, Natale DA, Depamphilis ML. 2000. Initiation of eukaryotic DNA replication: conservative or liberal? J Cell Physiol. 184:139–150.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. Cell. 132:311–322.

Branco MR, Pombo A. 2007. Chromosome organization: new facts, new models. Trends Cell Biol. 17:127–134.

Brodie of Brodie EB, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2005. From DNA sequence analysis to modeling replication in the human genome. Phys Rev Lett. 94:248103.

Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, Prioleau MN. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc Natl Acad Sci. USA. 105:15837–15842.

Calladine CR, Drew HR. (1999). Understanding DNA. San Diego: Academic Press.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CAM, Taylor MS, Engström PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 38:626–635.

Cavalli G, Misteli T. 2013. Functional implications of genome topology. Nat Struct Mol Biol. 20:290–299.

Cayrou C, Coulombe P, Vigneron A, Stanojcic S, Ganier O, Peiffer I, Rivals E, Puy A, Laurent-Chabalier S, Desprat R, Méchali M. 2011. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res. 21:1438–1449.

Chakalova L, Debrand E, Mitchell JA, Osborne CS, Fraser P. 2005. Replication and transcription: shaping the landscape of the genome. Nat Rev Genet. 6:669–677.

Chandra T, Kirschner K, Thuret JY, Pope BD, Ryba T, Newman S, Ahmed K, Samarajiwa SA, Salama R, Carroll T, et al. 2012. Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. Mol Cell. 47:203–214.

Chen CL, Duquenne L, Audit B, Guilbaud G, Rappailles A, Baker A, Huvet M, d'Aubenton Carafa Y, Hyrien O, Arneodo A, Thermes C. 2011. Replication-associated mutational asymmetry in the human genome. Mol Biol Evol. 28:2327–2337.

Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, d'Aubenton-Carafa Y, Arneodo A, Hyrien O, Thermes C. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res. 20:447–457.

Chessel D, Dufour A, Thioulouse J. 2004. The ade4 package -I- One-table methods. R News. 4:5–10.

Chevereau G, Arneodo A, Vaillant C. 2011. Influence of the genomic sequence on the primary structure of chromatin. Frontiers Life Sci. 5:29–68.

Cook PR. 1999. The organization of replication and transcription. Science. 284:1790–1795.

Cook PR. 2001. Principles of nuclear structure and functions. New York: Wiley.

Courbet S, Gay S, Arnoult N, Wronka G, Anglana M, Brison O, Debatisse M. 2008. Replication fork movement sets chromatin loop size and origin choice in mammalian cells. Nature. 455:557–560.

Cremer T, Cremer C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet. 2:292–301.

De S, Michor F. 2011a. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat Biotechnol. 29:1103–1108.

De S, Michor F. 2011b. DNA secondary structures and epigenetic determinants of cancer genome evolution. Nat Struct Mol Biol. 18:950–955.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. Science. 295:1306–1311.

Desprat R, Thierry-Mieg D, Lailler N, Lajugie J, Schildkraut C, Thierry-Mieg J, Bouhassira EE. 2009. Predictable dynamic program of timing of DNA replication in human cells. Genome Res. 19:2288–2299.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 485:376–380.

Dostie J, Bickmore WA. 2012. Chromosome organization in the nucleus charting new territory across the Hi-Cs. Curr Opin Genet Dev. 22:125–131.

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 16:1299–1309.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 28:817–825.

Ernst J, Kheradpour P, Mikkelsen T, Shoresh N, Ward L, Epstein C, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 473:43–49.

Fan JY, Gordon F, Luger K, Hansen JC, Tremethick DJ. 2002. The essential histone variant H2A.Z regulates the equilibrium between different chromatin conformational states. Nat Struct Biol. 9:172–176.

Farkash-Amar S, Lipson D, Polten A, Goren A, Helmstetter C, Yakhini Z, Simon I. 2008. Global organization of replication time zones of the mouse genome. Genome Res. 18:1562–1570.

Farkash-Amar S, Simon I. 2010. Genome-wide analysis of the replication program in mammals. Chromosome Res. 18:115–125.

Felsenfeld G, Groudine M. 2003. Controlling the double helix. Nature. 421:448–453.

Feng S, Jacobsen SE. 2011. Epigenetic modifications in plants: an evolutionary perspective. Curr Opin Plant Biol. 14: 179–186.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. Cell. 143:212–224.

Fraser P, Bickmore W. 2007. Nuclear organization of the genome and the potential for gene regulation. Nature. 447:413–417.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 462:58–64.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. J Mol Biol. 196:261–282.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science. 330:1775–1787.

Gilbert DM. 2001. Making sense of eukaryotic DNA replication origins. Science. 294:96–100.

Gilbert DM. 2010. Evaluating genome-scale approaches to eukaryotic DNA replication. Nat Rev Genet. 11:673–684.

Gilbert N, Gilchrist S, Bickmore WA. 2005. Chromatin organization in the mammalian nucleus. Int Rev Cytol. 242:283–336.

Green P, Ewing B, Miller W, Thomas PJ, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet. 33:514–517.

Guilbaud G, Rappailles A, Baker A, Chen CL, Arneodo A, Goldar A, d'Aubenton-Carafa Y, Thermes C, Audit B, Hyrien O. 2011. Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. PLoS Comput Biol. 7:e1002322.

Hamlin JL, Mesner LD, Lar O, Torres R, Chodaparambil SV, Wang L. 2008. A revisionist replicon model for higher eukaryotic genomes. J Cell Biochem. 105:321–329.

Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, Stamatoyannopoulos JA. 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci. USA. 107:139–144.

Hiratani I, Gilbert D. 2009. Replication timing as an epigenetic mark. Epigenetics. 4:93–97.

Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, Papp B, Fussner E, Bazett-Jones DP, Plath K, Dalton S, et al. 2010. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res. 20:155–169.

Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, Lyou Y, Townes TM, Schubeler D, Gilbert DM. 2008. Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol. 6:e245.

Holwerda S, de Laat W. 2012. Chromatin loops, gene positioning, and gene expression. Front Genet. 3:217.

Hon G, Wang W, Ren B. 2009. Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput Biol. 5:e1000566.

Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, Thermes C. 2007. Human gene organization driven by the coordination of replication and transcription. Genome Res. 17:1278–1285.

Ioshikhes I, Bolshoy A, Derenshteyn K, Borodovsky M, Trifonov EN. 1996. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J Mol Biol. 262:129–139.

Julienne H, Zoufir A, Audit B, Arneodo A. Forthcoming 2013. Human genome replication proceeds though four chromatin states. PloS Comput Biol. 9:e1003233.

Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. 2012. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol. 30:90–98.

Karnani N, Taylor CM, Malhotra A, Dutta A. 2010. Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. Mol Biol Cell. 21:393–404.

Kaufman L, Rousseeuw PJ. 1984. Finding groups in data: An introduction to cluster analysis. New York: John Wiley & Sons.

Kellum R. 2003. HP1 complexes and heterochromatin assembly. Curr. Top. Microbiol. Immunol. 274:53–77.

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. 2010. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. Nature. 471:480–485.

Kouzarides T. 2007. Chromatin modifications and their function. Cell. 128:693–705.

Lee BK, Bhinge AA, Battenhouse A, McDaniell RM, Liu Z, Song L, Ni Y, Birney E, Lieb JD, Furey TS, et al. 2012. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. Genome Res. 22:9–24.

Letessier A, Millot GA, Koundrioukoff S, Lachagès AM, Vogt N, Hansen RS, Malfoy B, Brison O, Debatisse M. 2011. Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature. 470:120–123.

Li Y, Kirschmann DA, Wallrath LL. 2002. Does heterochromatin protein 1 always follow code? Proc Natl Acad Sci USA. 99 Suppl. 4:16462–16469.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 326:289–293.

Liu T, Rechtsteiner A, Egelhofer TA, Vielle A, Latorre I, Cheung MS, Ercan S, Ikegami K, Jensen M, Kolasinska-Zwierz P, et al. 2011. Broad chromosomal domains of histone modification patterns in *C.elegans*. Genome Res. 21:227–236.

Lucas I, Palakodeti A, Jiang Y, Young DJ, Jiang N, Fernald AA, Le Beau MM. 2007. High-throughput mapping of origins of replication in human cells. EMBO Rep. 8:770–777.

Mackay D. 2003. Information theory, inference, and learning algorithms. Cambridge (UK): Cambridge University Press.

Maison C, Almouzni G. 2004. HP1 and the dynamics of heterochromatin maintenance. Nat Rev Mol Cell Biol. 5:296–304.

Maric C, Prioleau MN. 2010. Interplay between DNA replication and gene expression: a harmonious coexistence. Curr Opin Cell Biol. 22:277–283.

Martin MM, Ryan M, Kim R, Zakas AL, Fu H, Lin CM, Reinhold WC, Davis SR, Bilke S, Liu H, et al. 2011. Genome-wide depletion of replication initiation events in highly transcribed regions. Genome Res. 21:1822–1832.

McNairn AJ, Gilbert DM. 2003. Epigenomic replication: linking epigenetics to DNA replication. Bioessays. 25:647–656.

Méchali M. 2001. DNA replication origins: from sequence specificity to epigenetics. Nat Rev Genet. 2:640–645.

Méchali M. 2010. Eukaryotic DNA replication origins: many choices for appropriate answers. Nat Rev Mol Cell Biol. 11:728–738.

Mesner LD, Valsakumar V, Karnani N, Dutta A, Hamlin JL, Bekiranov S. 2011. Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription. Genome Res. 21:377–389.

Milani P, Chevereau G, Vaillant C, Audit B, Haftek-Terreau Z, Marilley M, Bouvet P, Argoul F, Arneodo A. 2009. Nucleosome positioning by genomic excluding-energy barriers. Proc Natl Acad Sci USA. 106:22257–22262.

Minc E, Courvalin J, Buendia B. 2000. HP1gamma associates with euchromatin and heterochromatin in mammalian nuclei and chromosomes. Cytogenet. Cell Genet. 90:279–284.

Misteli T. 2007. Beyond the sequence: cellular organization of genome function. Cell. 128:787–800.

Mohn F, Schübeler D. 2009. Genetics and epigenetics: stability and plasticity during cellular differentiation. Trends Genet. 25:129–136.

Moindrot B, Audit B, Klous P, Baker A, Thermes C, de Laat W, Bouvet P, Mongelard F, Arneodo A. 2012. 3D chromatin conformation correlates with replication timing and is conserved in resting cells. Nucleic Acids Res. 40:9470–9481.

Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 5:621–628.

Ozsolak F, Song JS, Liu XS, Fisher DE. 2007. High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol. 25:244–248.

Ponger L, Duret L, Mouchiroud D. 2001. Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res. 11:1854–1860.

Rando OJ, Chang HY. 2009. Genome-wide views of chromatin structure. Annu Rev Biochem. 78:245–271.

Roudier F, Ahmed I, Bérard C, Sarazin A, Mary-Huard T, Cortijo S, Bouyer D, Caillieux E, Duvernois-Berthet E, Al-Shikhley L, et al. 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J. 30:1928–1938.

Roudier F, Teixeira FK, Colot V. 2009. Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. Trends Genet. 25:511–517.

Ryba T, Battaglia D, Chang BH, Shirley JW, Buckley Q, Pope BD, Devidas M, Druker BJ, Gilbert DM. 2012. Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. Genome Res. 22:1833–1844.

Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM. 2010. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. 20:761–770.

Ryba T, Hiratani I, Sasaki T, Battaglia D, Kulik M, Zhang J, Dalton S, Gilbert DM. 2011. Replication timing: a fingerprint for cell identity and pluripotency. PLoS Comput Biol. 7:e1002225.

Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods. 3:511–518.

Satchwell SC, Drew HR, Travers AA. 1986. Sequence periodicities in chicken nucleosome core DNA. J Mol Biol. 191:659–675.

Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci. USA. 103:1412–1417.

Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell. 132:887–898.

Schones DE, Zhao K. 2008. Genome-wide approaches to studying chromatin modifications. Nat Rev Genet. 9: 179–191.

Schotta G, Lachner M, Sarma K, Ebert A, Sengupta R, Reuter G, Reinberg D, Jenuwein T. 2004. A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes Dev. 18:1251–1262.

Segal E, Fondufe-Mittendorf Y, Chen L, Thaström A, Field Y, Moore IK. 2006. A genomic code for nucleosome positioning. Nature. 442:772–778.

Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gomez M. 2009. Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS Genet. 5:e1000446.

Sexton T, Schober H, Fraser P, Gasser SM. 2007. Gene regulation through nuclear organization. Nat Struct Mol Biol. 14:1049–1055.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the *Drosophila* genome. Cell. 148:458–472.

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 38:1348–1354.

Sims JK, Houston SI, Magazinnik T, Rice JC. 2006. A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin. J Biol Chem. 281:12760–12766.

Smallwood A, Hon GC, Jin F, Henry RE, Espinosa JM, Ren B. 2012. CBX3 regulates efficient RNA processing genome-wide. Genome Res. 22:1426–1436.

St-Jean P, Vaillant C, Audit B, Arneodo A. 2008. Spontaneous emergence of sequence-dependent rosettelike folding of chromatin fiber. Phys Rev. E 77:061923.

Struhl K, Segal E. 2013. Determinants of nucleosome positioning. Nat Struct Mol Biol. 20:267–273.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 9:465–476.

Talasz H, Lindner HH, Sarg B, Helliger W. 2005. Histone H4-lysine 20 monomethylation is increased in promoter and coding regions of active genes and correlates with hyperacetylation. J Biol Chem. 280:38814–38822.

Talbert PB, Henikoff S. 2010. Histone variants–ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol. 11:264–275.

Tang CSM, Epstein RJ. 2007. A structural split in the human genome. PLoS One. 2:e603.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 447:799–816.

The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9:e1001046.

The modENCODE Consortium. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science. 330:1787–1797.

Tolstorukov MY, Kharchenko PV, Goldman JA, Kingston RE, Park PJ. 2009. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. Genome Res. 19:967–977.

Touchon M, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2004. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. Nucleic Acids Res. 32:4969–4978.

Touchon M, Nicolay S, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. FEBS Lett. 555:579–582.

Touchon M, Nicolay S, Audit B, Brodie of Brodie EB, d'Aubenton-Carafa Y, Arneodo A, Thermes C. 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc Natl Acad Sci. USA. 102:9836–9841.

Travers AA, Vaillant C, Arneodo A, Muskhelishvili G. 2012. DNA structure, nucleosome placement and chromatin remodelling: a perspective. Biochem Soc Trans. 40:335–340.

Vaillant C, Palmeira L, Chevereau G, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2010. A novel strategy of transcription regulation by intragenic nucleosome ordering. Genome Res. 20:59–67.

Vakoc CR, Mandat SA, Olenchock BA, Blobel GA. 2005. Histone H3 lysine 9 methylation and HP1$\gamma$ are associated with transcription elongation through mammalian chromatin. Mol Cell. 19:381–391.

Vakoc CR, Sachdeva MM, Wang H, Blobel GA. 2006. Profile of histone lysine methylation across transcribed mammalian chromatin. Mol Cell Biol. 26:9185–9195.

Valenzuela MS, Chen Y, Davis S, Yang F, Walker RL, Bilke S, Lueders J, Martin MM, Aladjem MI, Massion PP, Meltzer PS. 2011. Preferential localization of human origins of DNA replication at the 5'-ends of expressed genes and at evolutionarily conserved DNA sequences. PLoS One. 6:e17308.

Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. Nature. 474:516–520.

van Holde KE. 1988. Chromatin. New York: Springer-Verlag.

Wang Z, Schones DE, Zhao K. 2009. Characterization of human epigenomes. Curr Opin Genet Dev. 19:127–134.

Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet. 39:457–466.

Widom J. 2001. Role of DNA sequence in nucleosome stability and dynamics. Q Rev Biophys. 34:269–324.

Wolffe AP. 1998. Chromatin structure and function. 3rd ed. London: Academic Press.

Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, Debernardi S, Mott R, Dunham I, Carter NP. 2004. Replication timing of the human genome. Hum Mol Genet. 13:191–202.

Yaffe E, Farkash-Amar S, Polten A, Yakhini Z, Tanay A, Simon I. 2010. Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. PLoS Genet. 6:e1001011.

Zaghloul L, Baker A, Audit B, Arneodo A. 2012. Gene organization inside replication domains in mammalian genomes. C. R. Mécanique. 340:745–757.

Zentner GE, Henikoff S. 2013. Regulation of nucleosome dynamics by histone modifications. Nat Struct Mol Biol. 20: 259–266.

Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, et al. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 38:1341–1347.

Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet. 12:7–18.

Zullo JM, Demarco IA, Pique-Regi R, Gaffney DJ, Epstein CB, Spooner CJ, Luperchio TR, Bernstein BE, Pritchard JK, Reddy KL, Singh H. 2012. DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. Cell. 149:1474–1487.