The Plant Cell, Vol. 26: 2351–2366, June 2014, www.plantcell.org

LARGE-SCALE BIOLOGY ARTICLE

The Functional Topography of the *Arabidopsis* Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States[C][W]

Joana Sequeira-Mendes,[a] Irene Aragüez,[a] Ramón Peiró,[b] Raul Mendez-Giraldez,[b,1] Xiaoyu Zhang,[c,2] Steven E. Jacobsen,[c] Ugo Bastolla,[b] and Crisanto Gutierrez[a,3]

[a] Department of Genome Dynamics and Function, Centro de Biologia Molecular Severo Ochoa, CSIC-UAM, Cantoblanco, 28049 Madrid, Spain

[b] Bioinformatics Unit, Centro de Biologia Molecular Severo Ochoa, CSIC-UAM, Cantoblanco, 28049 Madrid, Spain

[c] Department of Molecular, Cellular, and Developmental Biology, University of California, Los Angeles, California 90095

Chromatin is of major relevance for gene expression, cell division, and differentiation. Here, we determined the landscape of *Arabidopsis thaliana* chromatin states using 16 features, including DNA sequence, CG methylation, histone variants, and modifications. The combinatorial complexity of chromatin can be reduced to nine states that describe chromatin with high resolution and robustness. Each chromatin state has a strong propensity to associate with a subset of other states defining a discrete number of chromatin motifs. These topographical relationships revealed that an intergenic state, characterized by H3K27me3 and slightly enriched in activation marks, physically separates the canonical Polycomb chromatin and two heterochromatin states from the rest of the euchromatin domains. Genomic elements are distinguished by specific chromatin states: four states span genes from transcriptional start sites (TSS) to termination sites and two contain regulatory regions upstream of TSS. Polycomb regions and the rest of the euchromatin can be connected by two major chromatin paths. Sequential chromatin immunoprecipitation experiments demonstrated the occurrence of H3K27me3 and H3K4me3 in the same chromatin fiber, within a two to three nucleosome size range. Our data provide insight into the *Arabidopsis* genome topography and the establishment of gene expression patterns, specification of DNA replication origins, and definition of chromatin domains.

INTRODUCTION

The genetic information is packed into chromatin consisting of DNA and all associated proteins that contribute to its structure and function. The nucleosome is the structural unit of chromatin and is made of 147 bp of DNA wrapped around a histone protein octamer core formed by two molecules of each histone H2A, H2B, H3, and H4. At first sight, this may seem a static and repetitive structure. However, this is far from reality, since at least three major sources of variations exist. One is DNA modifications, primarily cytosine methylation (Law and Jacobsen, 2010). Another is the plethora of posttranslational modifications of histones that most frequently include acetylation, methylation, phosphorylation, and ubiquitination, among others (Berger, 2007; Kouzarides, 2007). Finally, individual histone molecules can be replaced within the nucleosome by histone variants such as H2A.Z and H3.3 with the aid of various histone chaperones and remodeling complexes (Filipescu et al., 2013; Skene and Henikoff, 2013). Altogether, these variations provide a very high combinatorial diversity at individual genomic loci (Berger, 2007; Kouzarides, 2007). Additionally, nucleosome positioning can also vary and nonhistone proteins that function as readers, erasers, and writers of histone marks increase the local complexity across the genome. This large diversity of chromatin composition has significant consequences, for example, in transcription (Berger, 2007; Lee et al., 2010) and genome replication (Dorn and Cook, 2011).

The first effort to identify chromatin types was performed in *Drosophila melanogaster* cells using genomic information of 53 chromatin proteins (Filion et al., 2010). This allowed the identification of five major chromatin states, namely, heterochromatin, Polycomb, repressed, and two types of active chromatin regions. A more recent study based on 18 histone modifications in *Drosophila* cultured cells identified nine chromatin states (Kharchenko et al., 2011), whose functional significance was investigated by integrating chromosome organization with data of DNase I hypersensitivity, RNA transcripts, and nonhistone protein binding.

[1] Current address: Department of Biophysics and Biochemistry Genetic Medicine, University of North Carolina, 120 Mason Farm Road, Chapel Hill, NC 27599.

[2] Current address: Department of Plant Biology, University of Georgia, Athens, GA 30602-7271.

[3] Address correspondence to cgutierrez@cbm.csic.es.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Crisanto Gutierrez (cgutierrez@cbm.csic.es).

[C] Some figures in this article are displayed in color online but in black and white in the print edition.

[W] Online version contains Web-only data.

www.plantcell.org/cgi/doi/10.1105/tpc.114.124578

A similar approach was performed in the model plant *Arabidopsis thaliana* using information derived from histone marks across tiling arrays of chromosome 4 (Roudier et al., 2011). In this case, four major chromatin states (heterochromatin, Polycomb, active genes, and intergenic regions) were reported. It was found that these chromatin domains are frequently small due to the compact nature of the *Arabidopsis* genome and appear interspersed with each other. Indeed, the *Arabidopsis* genome is particularly attractive for genomic studies since it is relatively small (~125 Mb), fully sequenced, and well annotated. Furthermore, genome-wide epigenomic maps of a large collection of histone marks, CG methylation, and histone variants have been reported. The availability of this full genomic information prompted us to ask whether, in addition to the classical active and repressed states, other predominant combinations of marks could be identified in the *Arabidopsis* genome. Here, we used genome-wide data of 11 histone modifications, CG methylation, nucleosome occupancy, and three histone variants, together with DNA sequence features to carry out a high-resolution study of chromatin signatures defining chromatin states throughout the entire genome. In this article, we expand the current view of the *Arabidopsis* epigenome organization, reporting nine chromatin states identified through a maximum likelihood probabilistic model that optimally describes chromatin features and their positional order. The model parameters were carefully optimized, and we tested that the resulting states are robust with respect to changes in the parameters of the analysis. Although we cannot exclude that a larger number of states describes finer details of the genome organization, the nine states that we report provide a robust description that is a good compromise between economy and accuracy. Interestingly, the topographical relationships between chromatin domains indicate preferential association of certain chromatin states. Our data provide a ground for a better understanding of the linear organization of the genome and the relevance and/or preference that certain signatures of genomic elements may have for either establishing gene expression patterns or specifying DNA replication origins. Furthermore, they are a useful resource to identify potentially relevant structural and functional elements, or combinations of them, in the *Arabidopsis* genome.

RESULTS AND DISCUSSION

High-Resolution Identification of Chromatin States

Our computational procedure started from the published profiles of nine histone modification marks (H3K9me2, H3K27me1, H4K5ac, H3K4me1, H2Bub, H3K36me3, H3K4me2, H3K4me3, and H3K27me3), three histone variants (H2A.Z, H3.1, and H3.3), the nucleosome density (total H3 histone content), the genomic G+C content, and CG methylated residues (Supplemental Data Set 1) (Bernatavichute et al., 2008; Zilberman et al., 2008; Zhang et al., 2009; Costas et al., 2011; Roudier et al., 2011; Stroud et al., 2012). In addition, we generated genome-wide chromatin immunoprecipitation (ChIP)-chip data for H3K9ac and H3K14ac. This rendered a total of 16 genome-wide chromatin and DNA sequence features that we have combined for the analysis.

These genomic profiles were homogenized by projecting them onto windows of fixed size w that cover the whole genome and gently smoothed to take care of windows where no data are present (see Methods). The histone modifications were normalized by the local nucleosome content and all data were subjected to principal component analysis (PCA) to reduce their dimensionality. We then clustered from two to six principal components (PCs) describing the chromatin marks of the genomic fragments, fitting the sequence of components across the genome to a Hidden Markov Model with K hidden states, similar to the method adopted by Kharchenko et al. (2011). The parameters of the analysis, window size w, smoothing parameter s, and number of PCs were optimized to maximize the minimum separation between states (see Figure 1A and Methods). The optimal parameters are almost independent of K in the range $K \leq 10$, and we chose the values $w = 150$ bp that coincide approximately with the size of DNA wrapped around the nucleosome core and produce 794,305 genomic fragments, $s = 0.1$, and PC = 4, one more than the number of principal components contributing to the variance more than average.

The first principal component (PC1) can be interpreted largely as gene expression potential, since it separates histone modifications associated with heterochromatin, which contribute negatively, from marks associated with euchromatin, which yield a positive contribution. The Polycomb mark H3K27me3, which is known to be a repressive mark, enters PC1 with a positive sign. The second (PC2) weights H3K27me3, which gives the dominant contribution, together with H2A.Z and H3K27me1, while all other marks give negative contributions. PC3 weights the presence of heterochromatin marks and the activation marks H3K4me3 and H3K36me3. PC4 separates from the rest a different combination of open chromatin marks, namely, H2A.Z, H3K9ac, H3K14ac, and H3K4me3, together with the histone components H3 and H3.1. The clusters that define distinct chromatin states were sorted according to the first PC, so that globally chromatin state 1 contains the largest set of histone modifications associated with open/active chromatin and the state with the largest index corresponds to the silenced pericentromeric heterochromatin.

To choose a convenient number of chromatin states, we examined the interstate similarity defined as the normalized scalar product of chromatin signatures q, whose value is one for completely identical sets of distinctive features (see Methods). With $K = 9$ states, the maximum similarity is q = 0.80, which increases to q = 0.86 for 10 states, suggesting that the additional state only provides minor distinctions with respect to the nine states (rendering essentially the same biological information in a more economic way). These reasons led us to conclude that nine states render a solid and coherent biological interpretation of the *Arabidopsis* genome, without excluding that new chromatin information could help in the future to refine the current knowledge.

Distinctive Properties of *Arabidopsis* Chromatin States

Chromatin state 1 (red; Figure 1B) is characterized by high amounts of H3K4me2 and H3K4me3, H3 acetylation, H3K36me3, and H2Bub, typically associated with transcribed regions and

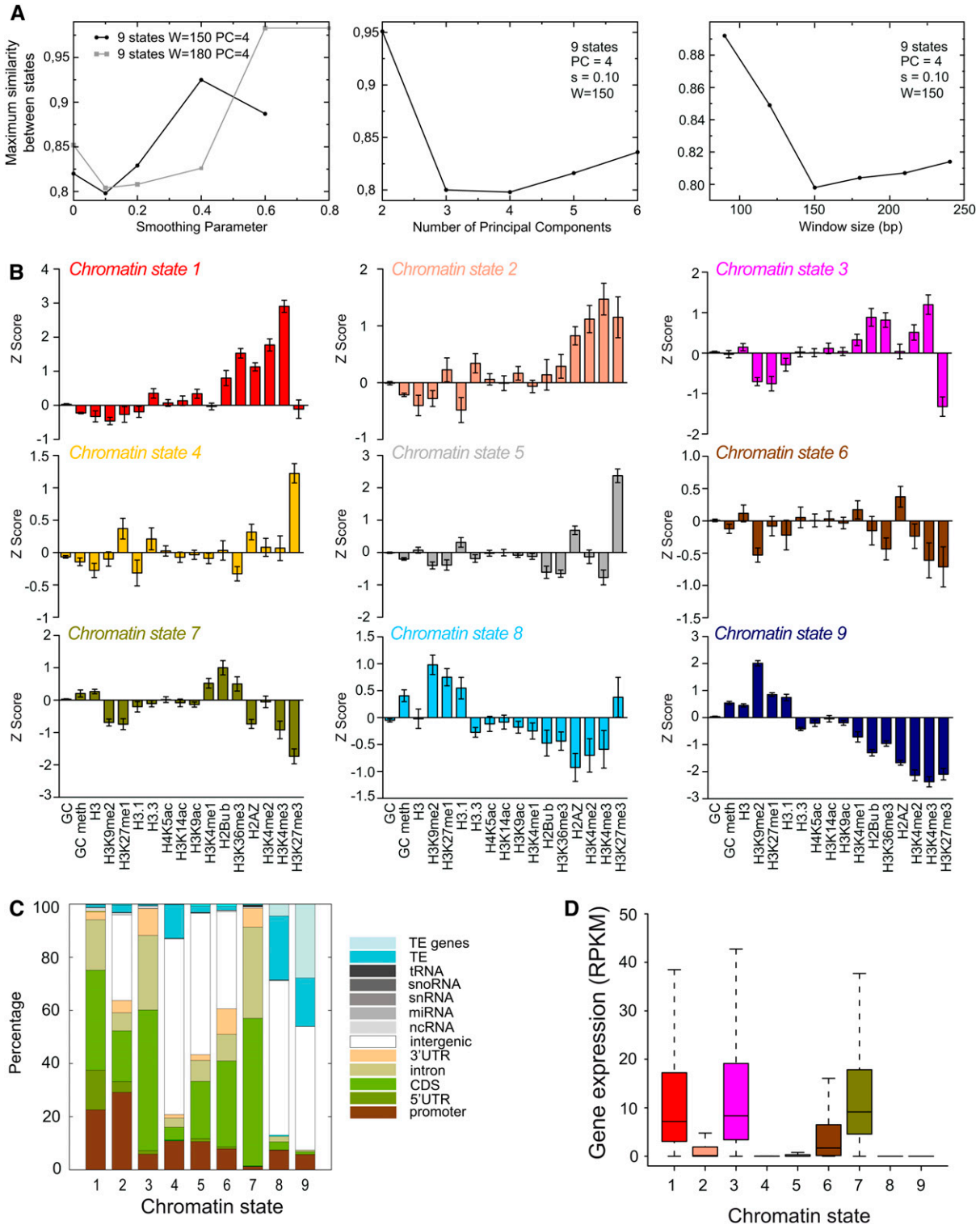


Figure 1. Genome-Wide Annotation of the *Arabidopsis* Chromatin Defined by Specific Signatures.

(A) Optimization of parameters. The smoothing parameter (left panel), number of PCs (middle panel), and window size (right panel) were optimized based on the minimum similarity between clusters. See Methods for further details.

(B) Prevalent chromatin states, as a result of a PCA, as described in the text, were defined by a combinatorial pattern of genomic features. They are characterized by a unique combination of values (positive and negative z-score indicate values above or below the average in the genome, respectively)

transcription start sites (TSSs; see also below) and by a relatively low nucleosome density, enriched in H3.3 and H2A.Z.

A similar set of active marks but also including high levels of the repressive modification H3K27me3 defines the chromatin state 2 (salmon; Figure 1B). This state presents lower levels of H3K36me3, H2Bub, H3ac, and nucleosome density, the latter possibly due to its higher than average AT richness.

Chromatin state 3 (magenta, Figure 1B) is defined by high levels of histone H3K4me1, H2Bub, H3K36me3, and H3K4me2/3, is highly depleted in Polycomb marks, and represents a transcription elongation signature.

While chromatin states 1 and 3 are very similar in histone modification marks, they differ in their enrichment in certain genomic elements. Thus, 37.5% of the genome with chromatin state 1 colocalizes with promoter and 5′ untranslated regions, whereas only ~7% of state 3 chromatin associates with these elements (Figure 1C). Interestingly, the promoter+5′UTR (red; state 1) and transcriptionally elongating (magenta; state 3) chromatin states correspond qualitatively with similar states of the *Drosophila* genome (Filion et al., 2010; Kharchenko et al., 2011). Conversely, association studies with genomic elements revealed that 29.1% of state 2 overlaps with promoters and 32.3% with intergenic regions (Figure 1C).

State 4 (gold, Figure 1B) is similar to state 2, maintaining H3.3, H2A.Z, and high levels of H3K27me3 but with reduced levels of marks typical of active transcription. In fact, chromatin state 4 largely overlaps with noncoding intergenic regions (66.2%). However, in this state, the overlap with putative promoters increases (from 10.9 to 19.4%) when we consider as promoters regions of up to 1000 bp upstream of the TSS instead of 650 bp (Figure 1C; Supplemental Figure 3), suggesting that this chromatin in state 4 could correspond to the most upstream region of promoters. Therefore, whereas states 1 and 3 tend to be present at the 5′ half of genic regions, states 2 and 4 seem to be more characteristic of intergenic regions containing proximal (state 2) and distal (state 4) promoter elements and perhaps regulatory regions.

State 5 (gray, Figure 1B) corresponds to the typical Polycomb-regulated chromatin and is defined by a lower than average amount of all marks analyzed except for high levels of H3K27me3 and moderate H2A.Z, within a nucleosome context enriched in H3.1. It is worth noting that H3K27me3 is also present to a significant amount in chromatin states 2 and 4, as has been observed in the *Drosophila* genome (Kharchenko et al., 2011). Chromatin state 5 colocalizes primarily with intergenic regions (63.9%; Figure 1C) and to a lesser extent with genic regions (32.7%; Figure 1C). This is fully consistent with the differences between H3K27me3 targets in plant and animal cells (Zhang et al., 2007). This chromatin state also emerged in previous studies in *Arabidopsis* (Zhang et al., 2007; Roudier et al., 2011) and in *Drosophila* as the Polycomb chromatin (Filion et al., 2010; Kharchenko et al., 2011).

Chromatin state 6 (brown, Figure 1B) is typically an intragenic state (52.7%), and it is characterized by a slight enrichment in H2A.Z, a higher than average nucleosome density, and H3K4me1, typical of gene bodies (Figure 1B).

Likewise, state 7 (green, Figure 1B) is also intragenic and has H3K4me1, H2Bub, and H3K36me3 as the most prominent marks. Strikingly, this state appears almost exclusively related to intragenic regions (97.2%), with 55.6 and 34.3% colocalizing with coding sequences and introns, respectively (Figure 1C), roughly similar to the *Drosophila* chromatin state located preferentially at introns (Kharchenko et al., 2011). It is interesting to note that, although most chromatin states colocalize to different extents with genes, state 7 is associated with transcription units longer than average (Supplemental Figure 3). This state 7, together with state 4, is difficult to assimilate to any of the chromatin states identified in *Drosophila* (Filion et al., 2010; Kharchenko et al., 2011).

Heterochromatin is defined by enrichment in H3.1, CG methylation, H3K9me2, and H3K27me1, as already known (Bernatavichute et al., 2008; Cokus et al., 2008; Jacob et al., 2010; Stroud et al., 2012). Interestingly, we can identify two distinct types of heterochromatic regions distinguished by their C+G content: GC-rich heterochromatin (navy, state 9; Figure 1B) and the less frequent AT-rich heterochromatin (sky blue, state 8; Figure 1B). The Polycomb and heterochromatin states largely coincide with two states identified previously (Roudier et al., 2011). We found that chromatin state 8 colocalizes preferentially with intergenic regions (58.2%) and transposon elements (TEs; 28.6%), of which 24.1% correspond to TEs and only 4.5% to TE genes (TEs are larger genomic elements that may contain one or more TE genes associated with them). However, the more GC-rich chromatin state 9, typically corresponding to heterochromatic pericentromeric regions, is mostly located at intergenic regions (46.5%) and transposable elements (46.0%), with a large proportion of both TEs (18.2%) and TE genes (27.8%) (Figure 1C).

A previous study integrating epigenomic maps in *Arabidopsis* chromosome 4 described four main chromatin states, namely, active, repressed, silent, and intergenic domains (Roudier et al., 2011). In that report, a large amount of the genome was considered as globally active (31%) or undefined (28%) chromatin. The high resolution of our analysis allowed a significant

Figure 1. (continued).

of each of the chromatin features considered. Error bars represent the SE of the mean. This number is estimated as the total number of windows divided by the correlation length of the mark considered.

(C) Relationship between genomic elements and chromatin states. The overlap (in base pairs) between the indicated genomic elements and each chromatin state was computed and expressed as a percentage. A promoter region of 0.65 kb was considered. Note that TEs are large genomic elements that may have one or more TE genes associated with them. Here, the class TE refers to genomic regions that contain TEs but do not overlap with TE genes.

(D) Relationship between gene expression level and chromatin states. RNA sequence reads normalized per kilobase and million reads (RPKM) obtained in whole seedlings (15 d old; see Methods) were computed for each chromatin state. Note the agreement with data presented in **(B)**.

advancement to this view, based on the identification of specific signatures within open chromatin: four states (1, 3, 6, and 7) primarily related with promoter and distinct regions within gene bodies and two additional states (2 and 4) mainly containing intergenic regions. Furthermore, analysis of gene expression levels (Kurihara et al., 2012) revealed that in addition to the two heterochromatic states (8 and 9), chromatin states 2, 4, and 5 are those containing the lowest amount of RNA transcripts (Figure 1D), consistent with their preferential location associated with intergenic regions and/or genes enriched for H3K27me3.

The GC content may affect nucleosome density and CG methylation, and there is a strong GC bias depending on the nature of genomic elements. Thus, a potential risk of our analysis is that including the GC content may bias toward differentiating between genetic elements rather than bona fide chromatin states. Therefore, we repeated the PCA without including the GC content and found essentially the same nine chromatin states (Supplemental Figure 4) but the likelihood was slightly less, indicating that the GC content is a valuable data set contributing positively to discriminate the different chromatin states.

A summary of representative signatures defining each of the nine chromatin states is shown in Figure 2. The full list of genomic coordinates defining each chromatin state is provided in Supplemental Data Set 2. These coordinate files can be used in genome browsers for convenient visualization of chromatin states (see also below).

Identification of Genomic Regions Sharing H3K27me3 and H3K4me3 Histone Modifications

Most of the identified states are characterized by previously described combinations of marks, such as H3K4me2, H3K4me3, H3 acetylation, or H3K36me3 in the most active states and H3K9me2 associated with H3K27me1 in heterochromatin (Deal and Henikoff, 2011; Feng and Jacobsen, 2011; Henikoff and Shilatifard, 2011). Surprisingly, our analysis identified genomic regions with somewhat unexpected combinations of histone modifications (e.g., H3K4me3 and H3K27me3). This observation could be the result of two different, though not mutually exclusive, scenarios: regions carrying one of the marks in one subpopulation of cells/tissues and the other mark in another and/or both marks occurring concomitantly in the same chromatin fiber and defining a state at the cellular level.

To experimentally address these possibilities, we performed a sequential ChIP (re-ChIP) in which chromatin was immunoprecipitated first using anti-H3K27me3 antibody and second using anti-H3K4me3 antibody. This strategy allows the purification of the chromatin molecules that carry both modifications simultaneously. We focused on state 2 chromatin since it is characterized by relatively high amounts of both H3K27me3 and H3K4me3 and by having a higher gene density (Figure 3A). We also included another region lacking these histone modifications, as a negative control (state 6; Figure 3A).

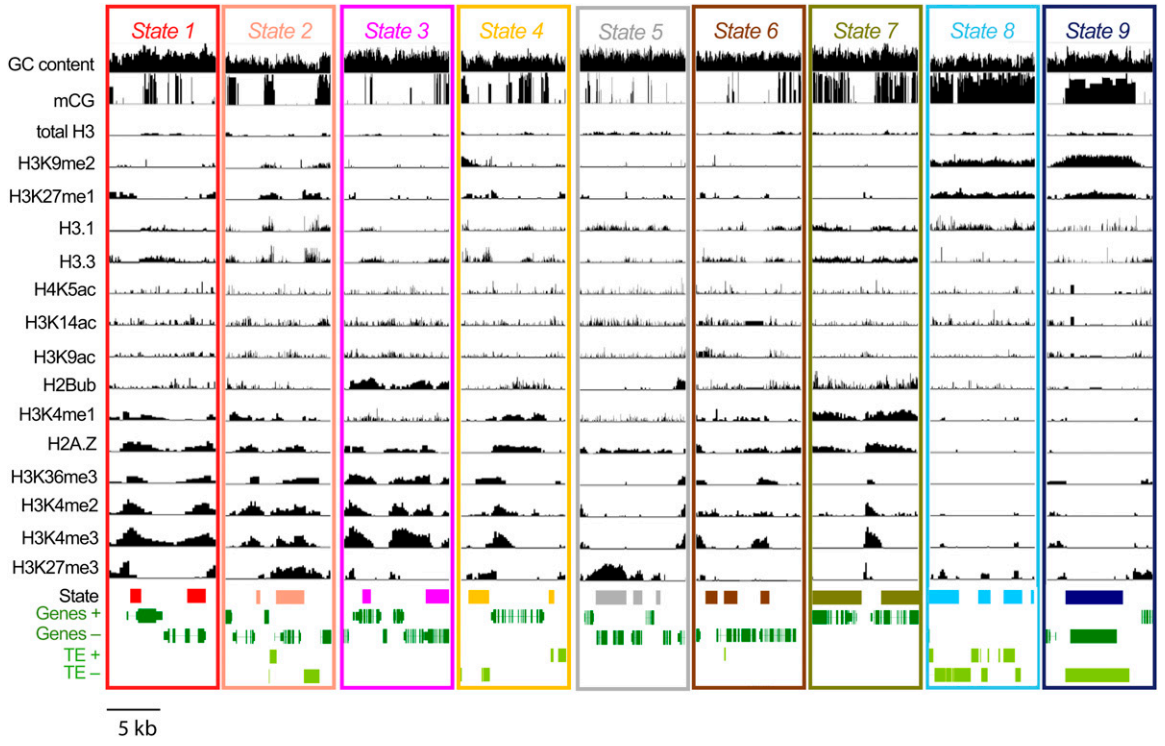


Figure 2. Representative Genomic *loci* with Distinct Features Defining Each Chromatin State.

Integrated Genome Browser (Nicol et al., 2009) views illustrating genomic regions containing each color-coded chromatin state characterized by their combinatorial profiles of chromatin and DNA sequence features. Note that each panel contains one to four domains of each type (blank regions between them are occupied by other chromatin states that have been omitted in the figure for simplicity).

Each of the regions analyzed was significantly enriched relative to a control region in the first ChIP with anti-H3K27me3 (Figure 3B) and additionally enriched in the second ChIP with anti-H3K4me3 (Figure 3C; see Methods for details). These results show that in all cases analyzed a large fraction of chromatin that contains H3K27me3 also holds H3K4me3 in the same molecule.

To rule out that the enrichments obtained after the second immunoprecipitation step were due to carried-over first antibody (anti-H3K27me3) present in the first eluate, a control with no antibody was performed during the second ChIP. The background ratios obtained in this control clearly confirmed that the observed enrichments in the second ChIP were produced by the second antibody (anti-H3K4me3) and not by traces of H3K27me3 antibody from the first ChIP (Figure 3C, gray bars).

The sequential ChIP was performed also in the reverse order of immunoprecipitation (first using anti-H3K4me3 antibody and anti-H3K27me3 as the second antibody). Again, the regions analyzed were significantly enriched for the simultaneous presence of both modifications (Figures 3C and 3D). Notably, region Chr1-16,578 kb (genomic region a in Figure 3A) that has lower amounts of H3K4me3 compared with the other regions analyzed, both in the epigenomic map data and in the primary ChIP data (Figure 3A), also holds lower values of enrichment in the re-ChIP experiments. Nevertheless, it is still enriched relative to the control region (and also for regions enriched only in H3K27me3; data not shown). These data ascertain the validity and sensitivity of the experimental procedure. Altogether, our results clearly identify a true epigenetic state at the cellular level defined by the coexistence of H3K27me3 and H3K4me3 in the same chromosome fiber, at least in a subpopulation of cells/tissues in the seedling. Based on the size of sheared DNA for the ChIP experiments (200 to 600 bp), our results indicate that these modifications could coexist, if not in the same nucleosome particle, in adjacent nucleosomes. A large fraction of these regions might also present alternative states with either H3K27me3 or H3K4me3, corresponding to the complete repression or activation of the associated genes in the different tissues and/or cell differentiation stages. Our sequential ChIP results directly point to the occurrence of bivalent regions similar to those described in mammalian pluripotent and primary cells (Bernstein et al., 2006; Roh et al., 2006). Overlap between H3K4me3 and H3K27me3 marks was observed in a number of genes in a study comparing chromatin profiles and gene expression in two different cell types of the *Arabidopsis* root, although their simultaneous presence in the same chromatin fiber was not assessed (Deal and Henikoff, 2010). A global analysis of chromatin regions containing simultaneously H3K4me3 and H3K27me3 is under way to understand their role during development.

Domain Organization of the *Arabidopsis* Genome

About half of the *Arabidopsis* genome (50.1% of the bins covering the entire genome) contains marks associated with expressed genes and proximal upstream promoter regions (states 1, 2, 3, 6, and 7), whereas 21.6 and 13.5% of the genome corresponds to heterochromatin (states 8 and 9) and Polycomb-regulated regions (state 5), respectively. The remaining 14.8% corresponds to chromatin state 4, primarily associated with intergenic regions (Figure 4A), which appears as a "hub" state that characterizes the transitions between the three big groups of chromatin states described above.

Neighbor fragments of 150 bp belonging to the same chromatin state were grouped together to define chromatin domains with more biological relevance. Interestingly, the relative amount of active chromatin domains is high, a situation particularly evident for chromatin states typical of the promoter and 5′-end of genes (states 1, 2, and 3; 47.2%). However, the opposite occurs for the heterochromatin domains (states 8 and 9; 9.4%; Figures 4A and 4B), as active domains tend to be smaller than the large inactive chromatin domains (Figures 4A and 4B). The largest number of domains is found for states 2 and 4 (Figures 4A and 4B).

Domain sizes are roughly exponentially distributed, with the exception of the GC-rich heterochromatin (state 9) that is broadly distributed (Supplemental Figure 5). The typical size of each chromatin domain is similar in states 1 through 6 (except classic Polycomb state 5), ranging from ~0.7 kb (states 1, 2, and 3) to 1.0 to 1.3 kb (states 4 and 6), whereas the state with the largest domains is the typical pericentromeric heterochromatin (state 9). Polycomb-associated chromatin (state 4) and the intragenic state 7 exhibit intermediate domain sizes (Figure 4C; Supplemental Figure 5), in agreement with data indicating that, unlike in animal cells where the H3K27me3 targets occupy ~20- to 50-kb-long genomic regions, the H3K27me3 domains in *Arabidopsis* are preferentially restricted to smaller regions (Zhang et al., 2007). The typical GC-rich heterochromatin (state 9) is organized in large domains. If we join the two heterochromatin states 8 and 9, the size distribution of the combined state becomes almost a power law, indicating the absence of a characteristic scale (Supplemental Figure 6).

Relationship of Chromatin States to Genes

To further determine possible relationships between each chromatin domain and transcriptional organization of the genome, we aligned the genomic states with TSS, transcriptional termination site (TTS), and gene bodies, including TE and various noncoding RNAs. We found a striking coincidence of active chromatin state 1 with TSS, whereas the intragenic states 3, 7, and 6 were enriched inside gene bodies, ~1.0, ~1.9, and ~2.3 kb downstream of TSS, respectively (Figure 5A). State 6 is the dominant state near TTS (Figure 5A); conversely, the TTSs peak at the center of state 6 domains (Supplemental Figure 7). States 2 and 4, which have a higher A+T content and H3K27me3 show a different pattern since they peak at ~0.4 and ~1 kb upstream of the TSS, respectively (Figure 5A), suggesting that they preferentially contain gene regulatory elements.

Interestingly, the genes that overlap with state 7 are on average much longer than genes that do not contain it (Supplemental Figure 3), so that this relatively rare state characterizes long genes, whereas genes that overlap with the AT-rich states 2, 4, and 5 are the shortest ones. This is suggestive of a transcription-dependent generation of chromatin marks, where gradients of distinct histone modifications along the genomic units could define predominant chromatin signatures.

A complementary way to visualize this organization consists of representing the genomic marks that define the nine states

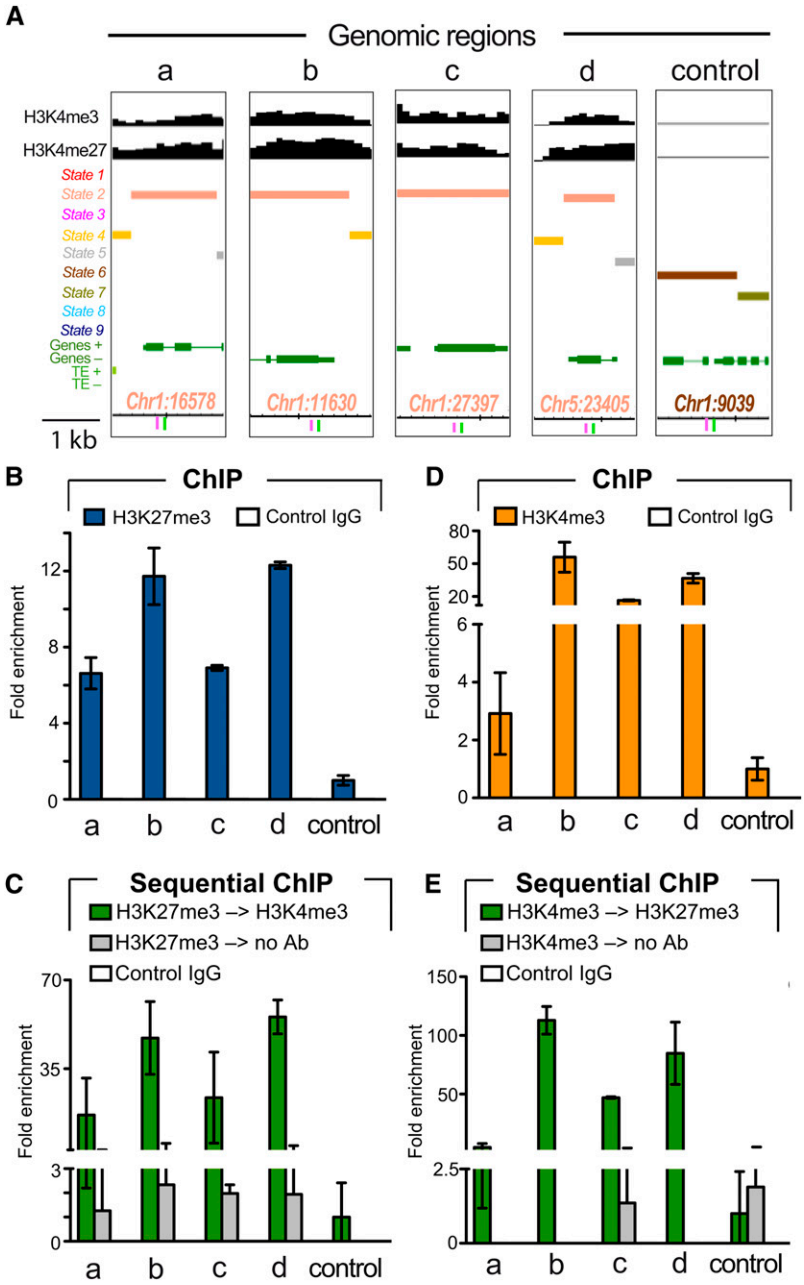


Figure 3. ChIP and Sequential ChIP Analyses.

(A) Browser views of genomic *loci* containing unexpected combinations of H3K4me3 and H3K27me3 observed in chromatin state 2 (a to d). The rightmost panel represents a randomly chosen control region with none of the mentioned histone modifications. At the bottom of each panel are depicted the positions of the primers used for the quantitative PCR analysis.

(B) Real-time PCR enrichment ratios of the indicated sites for H3K27me3 modification relative to the control region, detected by ChIP.

(C) Real-time PCR ratios reflecting the fold enrichment of the analyzed regions after sequential chromatin immunoprecipitations with H3K27me3 antibody and subsequently H3K4me3 antibody. Controls for the ChIP specificity (Control IgG) and for the second ChIP (H3K27me3 → no Ab) are presented.

(D) Real-time PCR enrichment ratios of the indicated sites for H3K4me3 modification relative to the control region, detected by ChIP.

(E) Real-time PCR ratios reflecting the fold enrichment of the analyzed regions after inverted order of sequential chromatin immunoprecipitations (first H3K4me3 antibody and second H3K27me3 antibody). Error bars in **(B)** to **(E)** represent the SD of the duplicates in one representative experiment. Controls for the ChIP specificity (Control IgG) and for the second ChIP (H3K4me3 → no Ab) are also presented.

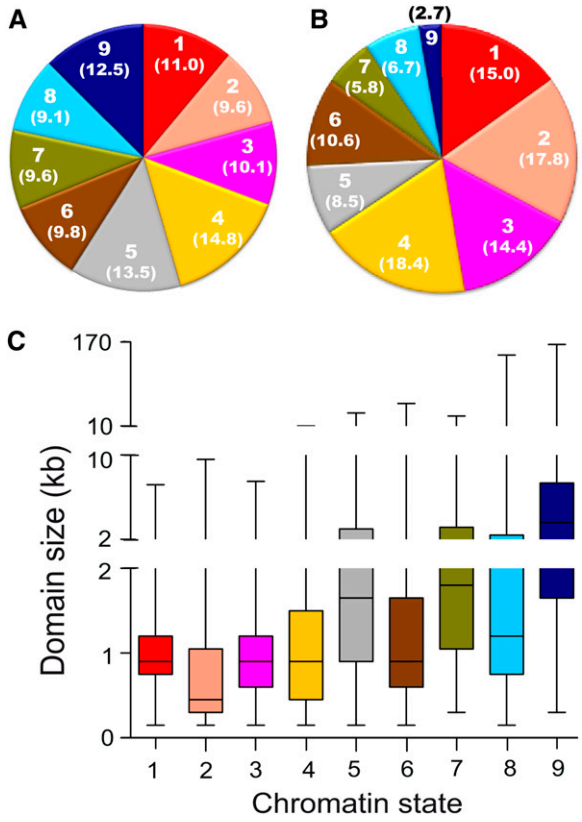


Figure 4. Genome-Wide Annotation of the *Arabidopsis* Chromatin Defined by Specific Signatures.

(A) Fraction of the genome (indicated as a percentage in parenthesis) occupied by each of the nine chromatin states.
(B) Fraction of chromatin domains (indicated as a percentage in parenthesis) occupied by each of the nine chromatin states.
(C) Size distribution of chromatin domains. See text for details. Box and whiskers show the minimum to the maximum of all data in each domain. The bar within the box depicts the median. The sample size for each domain appears in Supplemental Figure 5.

with respect to the TSS and TTS (Figure 5B) or with respect to the center of the domains of each of the nine states (Supplemental Figure 7). Thus, the marks typical of active transcription, H3K4me2, H3K4me3, H3K36me3, and the histone variant H2A.Z peak slightly upstream of the TSS, and slightly upstream of the center of the state 1 domains, and H3K4me1 is high at the center of state 7 domain, where the nucleosome content is also much above average (Supplemental Figure 7). A sharp peak of H3K27me3 and H2Bub characterizes the center of state 2, immediately upstream of the TSS, whereas the center of the hub state 4 and canonical Polycomb state 5 are characterized by H3K27me1 and H3K27me3, respectively.

The highly accessible and highly inaccessible regions identified by DNase I sensitivity (Shu et al., 2012) provide another interesting characterization of the genomic states. As expected, the promoter- and TSS-associated state 1 is highly accessible (Figure 6). However, remarkably, state 3, preferentially located just downstream of promoters, is even more accessible. The intragenic chromatin (states 6 and 7) with an average accessibility follows promoter chromatin (state 1) in this rank. On the contrary, the heterochromatin states are the most inaccessible, with state 9 being more inaccessible than state 8 (Figure 6). This observation is in agreement with the higher AT content of state 8, which is associated with a reduced nucleosome density and an enrichment in TE genes, suggesting that the transcriptional potential of these elements might contribute to a milder inaccessibility when compared with the classic GC-rich heterochromatin (state 9).

Genome-Wide Topographical Relationships between Chromatin States

A simple inspection of our data strongly suggests that chromatin states associate with each other in a nonrandom manner, as clearly visualized in Figure 7A. To gain quantitative insight into the large-scale organization of chromatin domains, we analyzed the spatial relationship between nearby domains by computing the propensities of each pair of chromatin states to be adjacent along the genome (see Methods). A positive propensity indicates that the pairs of domains co-occur more frequently than expected at random. Strikingly, we found that each of the nine chromatin states have very strong propensities to associate with only a subset of other states (Figure 7B). Hence, the typical domain of actively transcribed genes containing the TSS (state 1) tends to associate exclusively with chromatin states involving the 5′ half of genes and proximal promoters (states 2 and 3, respectively), whereas intragenic chromatin state 3 is preferentially flanked by the intragenic states 6 and 7 but not by chromatin state 4. Furthermore, Polycomb-associated chromatin (state 5) is commonly in contact with chromatin state 4 (also enriched in H3K27me3) but not with any other domain associated with active chromatin. At the same time, H3K27me3-enriched states 4 and 5 are the preferred chromatin domains to be in contact with the AT-rich heterochromatin (state 8) found largely interspersed in the euchromatic regions of chromosome arms that, in turn, is the chromatin state exclusively found to be present at the transition to the pericentromeric heterochromatin regions (state 9). From this analysis, chromatin state 4 appears as a communication hub, being the only state preferentially associated with the three main types of chromatin (Figure 7C): genic (through direct contact with state 2 and state 6), Polycomb repressed (state 5), and heterochromatin (through state 8).

The topographical association of different chromatin states can be graphically summarized as a network diagram in which the associated pairs are represented (Figure 7C). As expected, we found that the propensity between two chromatin states is related to the similarity of their average histone modifications (see Methods and Supplemental Figure 8).

We observed that the propensity network is a more stringent representation, since not all similar pairs showed a positive propensity to associate. It is clear that the two extreme chromatin states in terms of chromatin features, i.e., fully active and fully repressed (states 1 and 9), are the most physically distant in

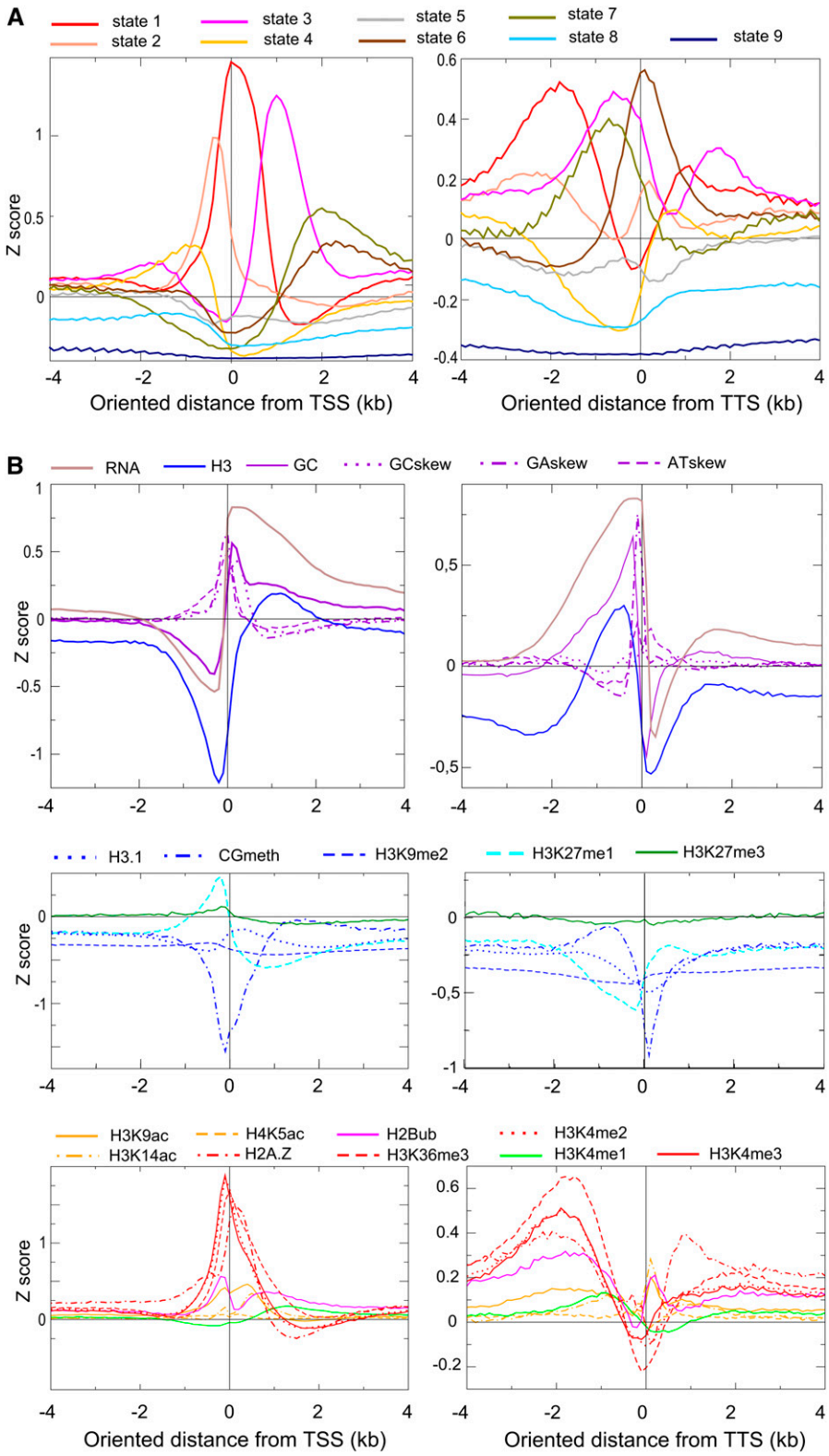


Figure 5. Localization of Chromatin States and Features Relative to Genomic Elements.

(A) The distribution of each chromatin state was determined around the TSSs (left panel) and the TTSs (right panel) of *Arabidopsis* genes. Colocalization analysis was performed taking into account the orientation of the transcription unit.

(B) Estimation of the relative enrichment of histone marks and DNA sequence features around the TSS and TTS.

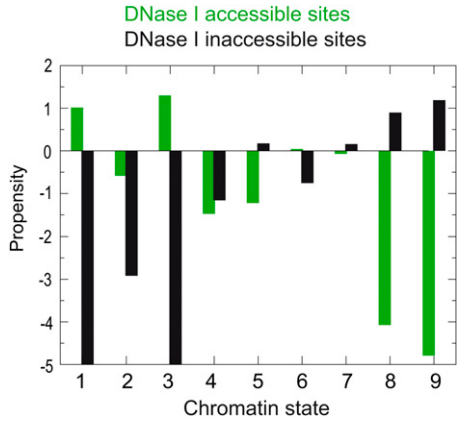


Figure 6. Accessibility of Chromatin States.

Propensity of colocalization of each chromatin state with DNase I accessible and inaccessible chromatin fractions, as described (Shu et al., 2012), compared with genome average probability/to random.
[See online article for color version of this figure.]

the linear scale of the genome. Thus, reaching active chromatin (state 1) from repressed heterochromatin (state 9) needs an almost mandatory path through AT-rich heterochromatin (state 8). Then, AT-rich heterochromatin (state 8) neighbors Polycomb chromatin (state 5) and intergenic region chromatin (state 4). Consistent with this observation, patches of H3K27me3-enriched (state 5) chromatin are frequently flanked by TEs in two *Arabidopsis* accessions (Dong et al., 2012). Interestingly, the Polycomb-associated chromatin and, subsequently, heterochromatin domains can be reached from state 1 through two main alternative paths with different chromatin features. One is through the relatively GC-rich chromatin states 3 and 6 (with or without state 7 in between) and another through the relatively AT-rich chromatin states 2 and 4, which also contain mid levels of H3K27me3, H3.1, and H2A.Z (and H3K4me2 in domain 2) (Figure 7C).

Consistent with previous analyses, the histone modifications used in this study show the expected correlations (Supplemental Figures 9). Furthermore, analysis of the distribution of different histone marks over the different domains revealed that there are three major types of marks (Supplemental Figure 7): (1) marks with a well-defined peak at the domain center as illustrated by the marks characteristic of active chromatin (state 1); (2) marks with a defined peak displaced from the oriented domain center as seen in histone modifications of chromatin present in promoters and the 5′-end of transcribed genes (states 2 and 3), where the displaced peak likely reflects the presence of another chromatin state nearby, e.g., state 1; and (3) marks distributed uniformly across the domain such as H3K9me2 in heterochromatin (states 8 and 9). These relationships of different histone marks may be of functional relevance, in agreement with previous observations (Deal and Henikoff, 2011; Henikoff and Shilatifard, 2011).

The relationships among various chromatin states are also evident when plotting each chromatin domain relative to the midpoint of all domains (Figure 8A). An example showing the topography of chromatin domains and transitions discussed above is shown in Figure 8B. Interestingly, recent experiments aimed at defining the chromosomal architecture of *Arabidopsis* nuclei using circular chromosome conformation capture (4C) revealed that the linear organization of the genome is largely translated into the genome-wide interactome (Grob et al., 2013). This is evident for interactions between euchromatic and heterochromatic regions. It would be relevant in the future to evaluate such 4C interactions between the various chromatin states identified in our study.

Genomic Motifs

Finally, we extended this analysis by computing all combinations of chromatin domains ranging from three to nine elements. Figure 9A represents the most numerous combinations for each number of elements, selecting only those combinations that occur more frequently than expected based on the lower number of elements. These frequent motifs can be attributed to three major chromatin meta-states: (1) Polycomb-repressed chromatin is mainly represented by the sequence of states 4-5-4-5 and its extensions; (2) typical heterochromatin is frequently formed by repeated tracts of chromatin states 8 and 9; and (3) the euchromatin, the more common and likely functionally relevant motif in TSS chromatin (state 1) flanked by states 2 and 3. This motif is by far the most frequent combination of three chromatin states, and it is frequently flanked by chromatin of distal regulatory regions (state 4) on its 5′ side, in the direction of transcription, and by chromatin states 6 or 7 on the 3′ side (Figure 9A; see also Figure 8B), leading to the consensus domain sequence 5-4-2-1-3-(7)-6-4-2, whose fragments and variations characterize the open chromatin. Two other important motifs of three elements in euchromatin are 3-7-6 (Figure 9B), which characterizes long transcriptional units (note that state 7 almost only occurs flanked by states 3 and 6) and 1-3-6, which characterizes medium-sized genes. When we computed the motifs associated with transcribed sequences, we found three main classes of transcriptional units based on the specific association of chromatin domains and gene size: (1) those that are contained within a unique domain, mostly of a bivalent (state 2) or repressed type (states 5, 8, and 9), and with a tendency to be associated with short genes; (2) those containing the sequence 1-3-7-(6), preferentially associated with the longest genes; and (3) those containing the sequence 1-3-6, with genes of intermediate size (Figure 9B).

Conclusions

The arsenal of molecular features that characterize the chromatin of eukaryotic organisms is large, such as local nucleosome content, histone variants, histone modifications, and DNA methylation, and one could expect that the epigenomic landscape should be very rich in combinatorial properties. Strikingly, however, current studies have unveiled a relatively simple linear organization of the epigenomic landscape (Filion et al., 2010; Kharchenko et al., 2011; Roudier et al., 2011; Ernst and Kellis, 2013). We speculate that this redundancy may serve to increase the functional robustness of the epigenome.

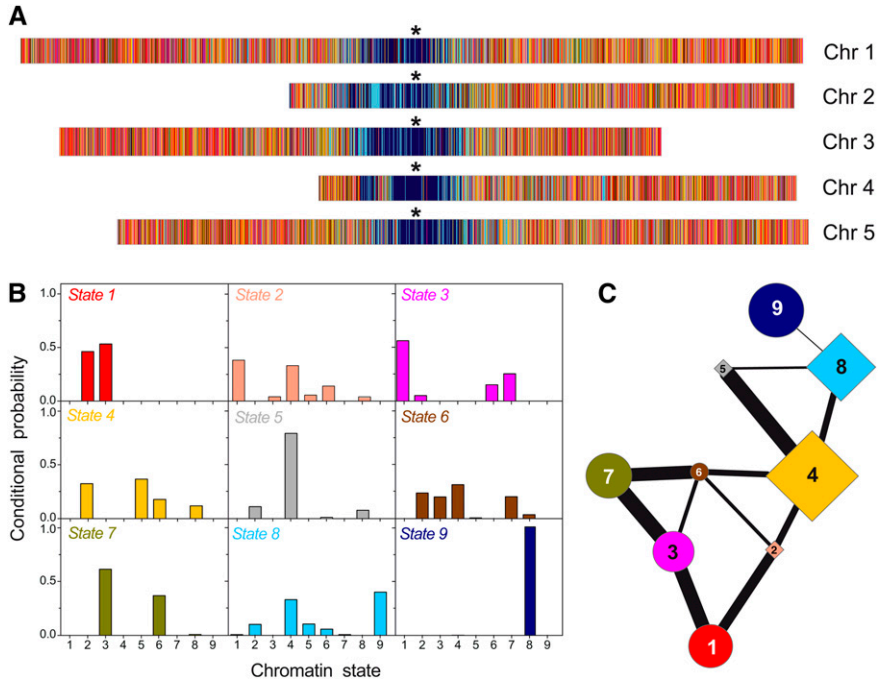


Figure 7. Transition Properties between Chromatin Domains.

(A) A karyotype view of *Arabidopsis* showing the relative location of color-coded chromatin domains identified in the five chromosomes. Asterisks indicate the position of centromeres.

(B) Frequency of transition between chromatin states. Bar graphs for each chromatin state show the conditional probability of a given state having another as a neighbor. See text for further details.

(C) Network propensity diagram of the frequency of transition between the nine chromatin states. Diamonds and circles represent AT-rich and GC-rich states, respectively. Symbol size represents the deviation in GC content with respect to the average genomic content. Circles are states with GC content larger than average, and diamonds are states with low GC content. The thickness of lines connecting chromatin states is proportional to the propensity of transition between two given states.

Given the combinatorial complexity of the epigenetic landscape, a more interpretable view of chromatin organization requires reducing the dimensionality of the epigenomic data, which we have achieved through PCA. Here, we present an objective estimate of the number of chromatin states by determining the optimum number of principal components with a criterion of maximum intercluster separation. As a result, we obtained a high-resolution topography of the *Arabidopsis* genome with the following major findings. (1) Four different chromatin states enriched in genes can be identified each with its own functional role: chromatin around TSS and enriched in histone modifications associated with active genes (state 1); the most accessible chromatin that tends to colocalize with the start of coding sequences (state 3); chromatin colocalizing with TTS (state 6); and the facultative chromatin state, mainly associated with long genes and intronic regions and with enrichment in H3K4me1, H2Bub, and H3K36me3 marks (state 7), in agreement with its localization over gene bodies. (2) Two types of heterochromatin states: the well-defined GC-rich heterochromatin (state 9) and a previously unnoticed AT-rich heterochromatin (state 8) interspersed within the typical heterochromatin, less inaccessible, less rich in histone modifications associated with repressed chromatin and with a much smaller typical domain size. These less inactivated areas belonging to state 8 can have an important role in facilitating the access of DNA-interacting proteins to the heterochromatin. (3) Three chromatin states enriched in the distinctive Polycomb mark H3K27me3: the classical Polycomb chromatin (state 5), depleted in all other marks; a H3K27me3-containing chromatin that has some histone modifications of active chromatin (state 4); and another H3K27me3-containing chromatin, located just upstream of the TSS and enriched in proximal promoter regions (state 2), which is the second most active state according to the first principal component. When further analyzing this unexpected coexistence of active/repressive marks (e.g., H3K4me3 and H3K27me3), using state 2 as an example, we observe that a subpopulation of cells in the young seedling indeed harbors both modifications in the same chromatin fiber. These states, in particular states 2 and 4, have an AT content larger than the average of the genome, they are prevalently intergenic, and they tend to be organized along the sequence of states 5-4-2. This mosaic structure of interspersed active and inactive regions seems to be an important characteristic of the linear organization of the genome. (4) Chromatin states show differential

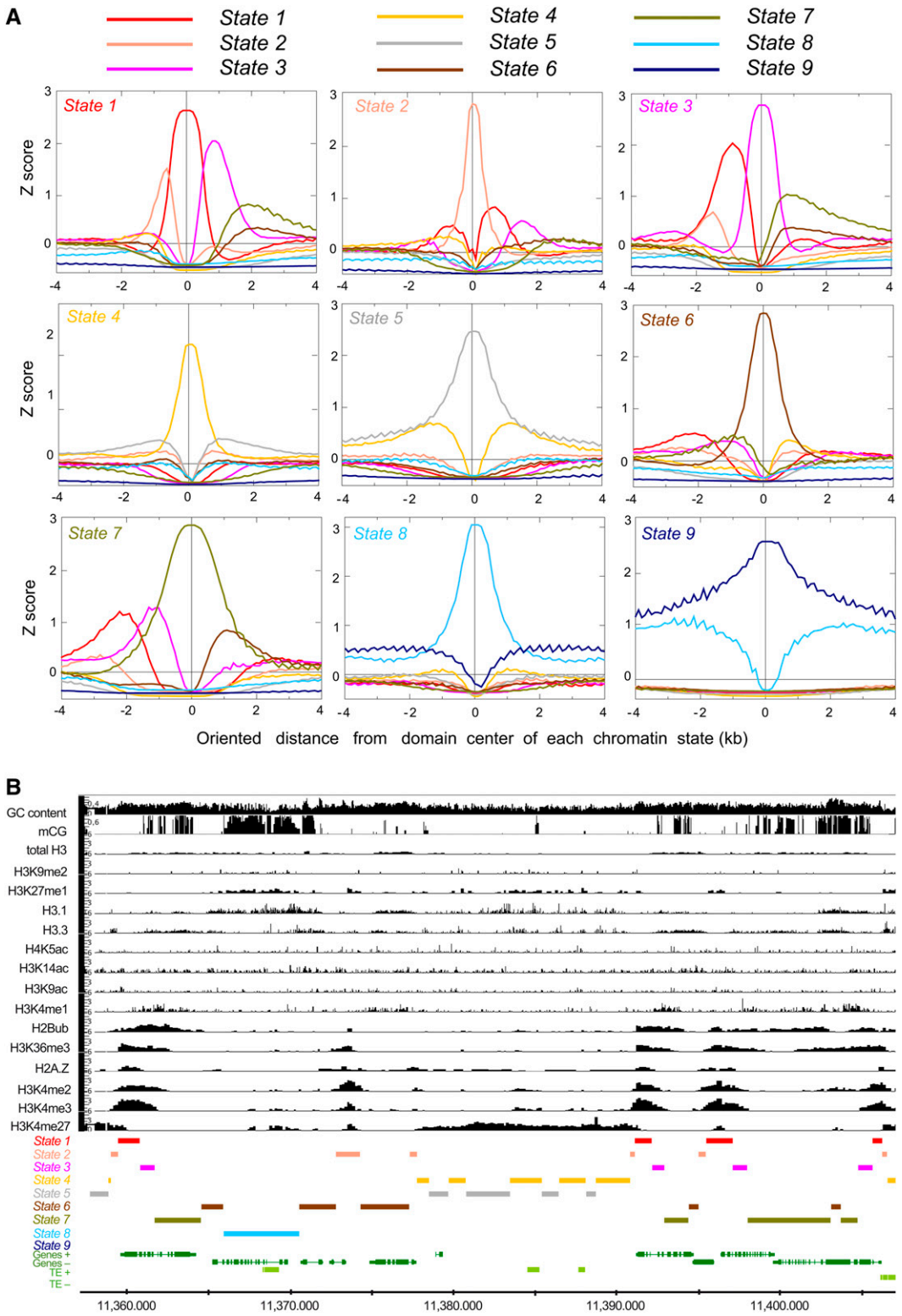


Figure 8. Local Relationships of Chromatin States.

(A) The distribution of each chromatin state determined around the center of the domain taking into account the transcription-based orientation of each domain.
(B) Genome browser view of the color-coded chromatin domain annotation over an ~50-kb region of the *Arabidopsis* Chromosome 1 (coordinates correspond to the TAIR10 version). Representative chromatin states in gene-dense and gene-poor regions, the chromatin state transition propensities, and the chromatin features of specific loci are highlighted in this epigenomic landscape.

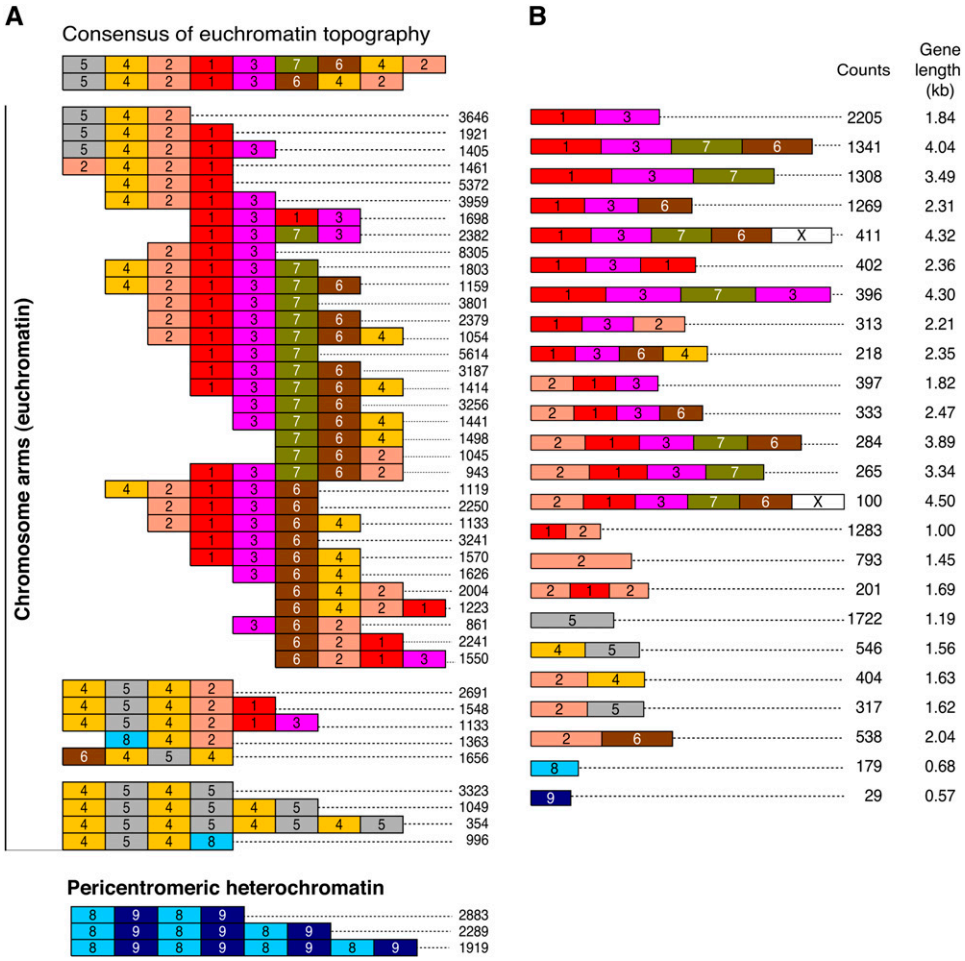


Figure 9. Domain Combinations.

(A) Most frequent combinations of chromatin states according to propensities between neighbors (total number in the genome at the right side of the panel).

(B) Relationship between chromatin domain combination and gene length (in kilobases; indicated at the right side of the panel). The size of each combination was made proportional to the gene length. Note that combinations containing the pair 7-6 tend to be associated with longer genes.

preferences to associate with each other. Chromatin state 1 is flanked only by state 3 and state 2, and almost always according to the sequence 2-1-3 (along the direction of transcription). This is one of the most important examples of a reduced number of chromatin motifs that simplify even further the linear organization of the chromatin of *Arabidopsis*. The motif 2-1-3 most likely has a functional role in facilitating transcription, with the AT-rich and nucleosome-poor chromatin (state 2) upstream of chromatin where the TSS is preferentially located and containing histone modifications associated with active genes (state 1), followed by the less active but more accessible chromatin (state 3), probably due to the presence of labile nucleosomes at the 5′-end of genes. Other important chromatin motifs in the euchromatin are (1) 1-3-7-(6), which characterizes long genes where the relatively rare chromatin marked by H3K4me1, H2Bub, and H3K36me3 (state 7) probably facilitates the processivity of transcription, and (2) 1-3-6, which characterizes medium-sized genes. Motifs that associate with Polycomb-repressed chromatin consist of domain sequences characterized by the alternation of more and less active domains, such as 4-5-4, 5-4-2, and 4-5-4-5. Finally, the heterochromatin is characterized by long stretches of the domain combination 8-9-8-9… Strikingly, state 9 cannot be contiguous to almost any other state except the AT-rich heterochromatin state 8, 9-8-4 (or to a lesser extent 9-8-5) being the typical sequence through which the heterochromatin communicates with Polycomb-repressed chromatin.

Together, our data could serve the basis to anticipate expression patterns of genes of interest based on the associated combination of chromatin domains. It will be interesting to study the evolutionary relationships of the peculiar simplicity and redundancy of chromatin and genome organization, as well as their functional relevance.

METHODS

Plant Material

For the sequential ChIP experiment, *Arabidopsis thaliana* seedlings (Columbia-0 ecotype) were grown in Murashige and Skoog medium supplemented with 1% (w/v) sucrose and 1% (w/v) agar in a 16-h:8-h light/dark regime at 22°C. For the ChIP-chip experiments, plants were grown on soil under 24-h light.

ChIP and Microarray (chip)

The H3K9ac and H3K14ac ChIP analyses were performed using the aboveground tissues of 3-week-old seedlings and the Affymetrix *Arabidopsis* Tiling 1.0R arrays as described (Costas et al., 2011). Antibodies used for H3K9ac, H3K14ac, and H4K5ac are AB4441 (Abcam; lot number 511238), 07-353 (Millipore; lot number DAM1462567), and 06-759 (Millipore; lot number DAM1549961), respectively. The H3 ChIP control was performed using an antibody from Abcam (ab1791). Data sets are deposited at the National Center for Biotechnology Information Gene Expression Omnibus under accession code GSE54489.

Sequential ChIP (Re-ChIP)

Sequential ChIP experiments were performed using 10-d-old seedlings. The plantlets were cross-linked with 1% formaldehyde by vacuum infiltration and quenched with 0.125 M glycine. After grinding, nuclei were isolated in extraction buffer (10 mM Tris-HCl pH 8.0, 0.25 M sucrose, 10 mM $MgCl_2$, 1% Triton X-100, and protease inhibitors). Nuclei were pelleted by centrifugation, resuspended in lysis buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, and 1% SDS) and disrupted by sonication in a Bioruptor (Diagenode), yielding genomic fragments of 200 to 600 bp. For each ChIP/re-ChIP assay, 8 to 15 µg of DNA/protein complexes was immunoprecipitated with α-H3K27me3 (Upstate 07-449; Abcam ab6002) and α-H3K4me3 (Abcam ab8580) using the Re-ChIP-IT kit (Active Motif) and following the manufacturer's instructions. After the de-cross-linking step, DNA was extracted with phenol:chloroform:isoamyl alcohol (25:24:1), ethanol precipitated, and resuspended in TE. Quantitative real-time PCR was performed in an ABI Prism 7900HT detection system (Applied Biosystems) using GoTaq pPCR Master Mix (Promega). The sequence of primers used in this analysis is provided in Supplemental Table 1.

Genomic Profiles

Experimental details and the source of material of published epigenomic data sets are provided in Supplemental Data Set 1. All these data sets were converted into TAIR9 compatible coordinates. The profiles of each epigenetic feature were standardized and normalized in windows of size w ranging from 90 to 360 bp. In short, each profile was averaged in each window. Since some windows did not contain any data, the resulting profile was smoothed over five adjacent windows centered on the target window with weighting coefficients of 1 for the target window, s for neighboring windows, and s^2 for the next neighbors if data were present, or otherwise 0. We determined nearly optimal values of the window size w and smoothing parameter s as explained below, finding that the optimal values are w = 150 bp and s = 0.1. This analysis shows that smoothing improves the likelihood values, as expected, since it takes care of missing data without modifying too many windows for which data are available, but it does not change the properties of the states qualitatively.

Importantly, histone modification marks H2Bub, H3K4me2, H3K4me3, H3K27me1, H3K27me3, and H3K36me3 were normalized by the local H3 content and the H3.1 and H3.3 contents by their geometric mean, so that one of these marks became redundant and was not used in the computation. Finally, each profile was shifted by subtracting its mean over the whole genome. In this way, we obtained a standardized value of each epigenomic mark at each window of the *Arabidopsis* genome (794,305 windows for the optimal window size w = 150 bp).

The RNA-seq data alignment sets used in Figure 1, corresponding to 2-week-old wild-type seedlings of *Arabidopsis* ecotype Columbia-0, were downloaded in SAM format from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). SAM files were converted into PILEUP format using SAMTOOLS (http://samtools.sourceforge.net). Finally, expression levels (reads normalized per kilobase and million reads) were calculated from PILEUP format files using a Perl script.

Clustering of Principal Components

The dimensionality of the 16 epigenomic and genomic variables was reduced by a PCA. For the optimal window size, only the first three PCs contribute to the variance more than the average component, which suggests that the dimensionality reduction should improve the results, as we directly verified (see next section). The first n PCs were clustered by fitting their sequence to a Hidden Markov Model (HMM) with K hidden states described as in the previous case, and also accounting for the transition probabilities from each state to the other ones. The $K(1+n+n(n+1)/2+(K-1)/2)$ parameters of the HMM and the assignation of each genomic window to one of the K states were computed iteratively through a standard algorithm that works iterating two steps: First, the weights of each window in each hidden state that maximizes the a posteriori probability is computed with the Viterbi algorithm and then weights are used to determine parameters for each state. When the algorithm converges, each window is assigned to the hidden states with maximum posterior probability. The computation was performed $K-1$ times with different initial parameters derived from the $K-1$ clusters identified at the previous step, randomly splitting each of the $K-1$ clusters to generate different sets of initial parameters. We verified that the similarity between the final states obtained for different initial conditions is very large, more than 90% for more than $K = 4$ states and often almost 100% similarity. For each number n of PCs, the clustering procedure was performed for K ranging from 1 to 16. We measured the Bayes Information Criterion (BIC) and Akaike Information Criterion (AIC) scores, which score the log of the likelihood L of the observed components, given the HMM model and the parameters, penalizing the number of parameters N_{para}, which increases with the number of states, according to the AIC ($AIC = -2\log(L) + 2N_{para}$) and the more stringent BIC ($BIC = -2\log(L) + N_{para}\log(N_{windows})$), where $N_{windows}$ is the number of windows considered in the computation. We also measured the maximum and average intercluster similarity. We found that both the BIC and the AIC scores improve for larger numbers of clusters, but the clusters become more and more similar so that distinguishing them provides less and less information. The algorithm that carries out these computations was programmed by us.

Parameter Optimization and Robustness of the Analysis

Choosing optimal parameters is probably the most important part of any computational analysis. In this case, there are four main parameters to optimize: (1) the size w of the windows in which histone modification data are averaged; (2) the smoothing parameter s; (3) the number n of PCs used in the clustering procedure; and (4) the number of clusters K, the most important parameter as it may have a biological significance if a clear optimum emerges from the analysis. Due to the intensiveness of the calculations, it is not possible to explore parameter space exhaustively. For every set of tested parameters, we clustered our data by fitting the sequence of the properties of each cluster with the HMM and for each number of clusters we determined the maximum value of the intercluster similarity measured as the cosine between the vectors defined by the

mean marks in the state, Sim(a,b)=sum$_k$ M*ak**M$_{bj}$/sqrt(sum$_j$ M$_{aj}^2$*sum$_j$ M$_{bj}^2$), where M$_{aj}$ is the mean value of the *j*-th mark in state *a*. This measure was used as an objective function, since minimizing it optimizes the distinction between different states. We made an initial guess of parameters and plotted the objective function versus the parameters of the analysis, finding candidate optimal parameters $w = 150$, $s = 0.10$, and $n = 4$. Subsequently, we verified that these parameters are at a local optimum, measuring the objective function varying one parameter at a time, as shown in Figure 1A. The results shown in the text were obtained with these optimized parameters.

To determine the optimal number of states, we observed that the maximum interstate similarity is q = 0.76 for eight states, q = 0.80 for nine states, and it grows to q = 0.86 for 10 states, suggesting that the additional state is very similar to the ones that have been already found. Moreover, we examined the properties of the new state, finding that it only provides minor distinctions with respect to the nine states, rendering essentially the same biological information in a more economic way.

The robustness of the procedure is shown by the average values of the properties of the nine states for window size $w = 120$ (Supplemental Figure 10A) or $w = 180$ (Supplemental Figure 10C). These profiles are similar to each other and to the standard window $w = 150$. We also found that not performing the smoothing ($s = 0$; Supplemental Figure 10B) yielded qualitatively similar profiles, although it worsens the intercluster separation. Finally, if the GC content (which strictly speaking is not an epigenetic property) is not included, the analysis is worse, but the result does not change qualitatively, in particular we can still distinguish the two heterochromatin states 8 and 9 with different accessibility (Supplemental Figure 4; Figure 6).

Network of Chromatin Domains

The propensity of domains of states *a* and *b* being adjacent along the genome was computed as Prop(a,b)=log[f(a,b)]-log[f(a)f(b)], where f(a,b) is the frequency of adjacent pairs of type a,b and f(n) is the frequency of type n. Positive propensity characterizes pairs that co-occur more frequently than expected at random. The propensity between two states is related to the similarities of the average marks of each state, Sim(a,b) defined above. This procedure was extended to the computation of all sequences of states of length *l*, with *l* ranging from 3 to 9. Since a given chromatin domain sequence could not be distinguished from the complementary one, for instance 2-1-3 and 3-1-2, complementary sequences were grouped together. Sequences were ranked according to their number in the genome. To assess whether a sequence of length *l* was significantly overrepresented, its probability was computed based on the two sequences of length *l*-1 contained in it and the transition probabilities. Only sequences that are overrepresented more than a given threshold (0.3 logarithmic units) were considered for Figure 9.

Quantification of Chromatin States at Various Genomic Elements

The TAIR10 annotation file containing coordinates of each genomic element of the *Arabidopsis* genome, including transposable elements, was downloaded from the TAIR website (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/TAIR10_gff3/TAIR10_GFF3_genes_transposons.gff). The file containing coordinates of chromatin states was compared with the annotation file to compute the length (in base pairs) of each genomic element corresponding to every chromatin state. In addition to the defined genomic elements, a "promoter region" was defined as either the 0.65- or 1-kb region immediately upstream of the TSS of each gene. In the case of overlaps between a "transposable element" and a "transposable element gene," the overlapping fragment was assigned to the "transposable element gene" class. Raw base pair counts of each genomic element in each chromatin state were normalized with the cumulative base pair length of each chromatin state.

Relationship between Chromatin States, Gene Size, and DNase I Accessibility

For each transcriptional unit, its overlap with genomic domains was computed by: (1) counting the number of units that overlap with one domain of each given type and (2) calculating the mean length of these overlapping genes. For accessibility analysis, each accessible and inaccessible region's data sets (Shu et al., 2012) was overlapped with domains of each given type and the overlapping length divided by the total length of domains of this type was calculated.

Accession Numbers

Gene Expression Omnibus data sets mentioned in this study are under the following accession numbers: GSM852792, GSM852793, and GSM852794.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Values of Each of the Genomic Features Used in This Study for Each of the Principal Components Considered.

Supplemental Figure 2. Relationship between Genomic Elements and Chromatin States.

Supplemental Figure 3. Calculation of the Number of Transcripts per Domain (Black Bars) and the Average Size of Genes (in kb; Green Bars) Associated with Each Chromatin State.

Supplemental Figure 4. Evaluation of PCA without Considering the GC Content.

Supplemental Figure 5. Domain Size Distributions for the Different Chromatin States.

Supplemental Figure 6. Distribution of the Domain Size of the Heterochromatin States 8 and 9 Considered Individually or Combined.

Supplemental Figure 7. Estimation of the Relative Enrichment of Histone Marks and DNA Sequence Features in the 9 Chromatin States.

Supplemental Figure 8. Network Similarity Diagram of the Frequency of Transition between the 9 Chromatin States.

Supplemental Figure 9. Network Correlation Diagram of the Frequency of Transition between Different Chromatin Features (Histone Marks and GC Content) Used in This Study.

Supplemental Figure 10. Robustness of the Nine Chromatin States Obtained with Different Parameters.

Supplemental Table 1. Oligonucleotide Pairs Used in the Sequential ChIP-PCR Experiments.

Supplemental Data Set 1. Information and Accession Data of the Epigenomic Profiles Used in This Study.

Supplemental Data Set 2. Assignment of the *Arabidopsis* Genome to 9 Chromatin States.

ACKNOWLEDGMENTS

We thank Y. Yu for his participation in the ChIP-chip H3 acetylation experiments, E. Martinez-Salas for comments, and members of the C.G. laboratory for helpful discussions. J.S.-M. was the recipient of a Juan de la Cierva contract from MINECO. This research was supported by Grants BFU2009-9783, BFU2012-34821, and CSD2007-0057 (MICINN) to C.G.,

BFU2012-40011 to U.B., and GM60398 to S.E.J. S.E.J. is an Investigator of the Howard Hughes Medical Institute. The Centro de Biologia Molecular Severo Ochoa received an institutional grant from Fundación Ramon Areces.

AUTHOR CONTRIBUTIONS

C.G. conceived the study and designed the approach with the input of J.S.-M. and I.A. U.B. conceived and developed the computational approach and carried out the corresponding analysis and the study of motifs. R.P., R.M.G., and U.B. carried out the bioinformatics analysis. U.B., C.G., J.S.-M., I.A., and R.P. analyzed the data. X.Z. and S.E.J. generated H3K9ac and H3K14ac data sets. J.S.-M. carried out and analyzed the re-ChIP experiments. C.G. and U.B. wrote the article with the participation, direct input, and approval of all authors.

Received February 21, 2014; revised May 16, 2014; accepted May 27, 2014; published June 16, 2014.

REFERENCES

Berger, S.L. (2007). The complex language of chromatin regulation during transcription. Nature **447:** 407–412.

Bernatavichute, Y.V., Zhang, X., Cokus, S., Pellegrini, M., and Jacobsen, S.E. (2008). Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in *Arabidopsis thaliana*. PLoS ONE **3:** e3156.

Bernstein, B.E., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell **125:** 315–326.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature **452:** 215–219.

Costas, C., de la Paz Sanchez, M., Stroud, H., Yu, Y., Oliveros, J.C., Feng, S., Benguria, A., López-Vidriero, I., Zhang, X., Solano, R., Jacobsen, S.E., and Gutierrez, C. (2011). Genome-wide mapping of *Arabidopsis thaliana* origins of DNA replication and their associated epigenetic marks. Nat. Struct. Mol. Biol. **18:** 395–400.

Deal, R.B., and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev. Cell **18:** 1030–1040.

Deal, R.B., and Henikoff, S. (2011). Histone variants and modifications in plant gene regulation. Curr. Opin. Plant Biol. **14:** 116–122.

Dong, X., Reimer, J., Göbel, U., Engelhorn, J., He, F., Schoof, H., and Turck, F. (2012). Natural variation of H3K27me3 distribution between two Arabidopsis accessions and its association with flanking transposable elements. Genome Biol. **13:** R117.

Dorn, E.S., and Cook, J.G. (2011). Nucleosomes in the neighborhood: new roles for chromatin modifications in replication origin control. Epigenetics **6:** 552–559.

Ernst, J., and Kellis, M. (2013). Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. Genome Res. **23:** 1142–1154.

Feng, S., and Jacobsen, S.E. (2011). Epigenetic modifications in plants: an evolutionary perspective. Curr. Opin. Plant Biol. **14:** 179–186.

Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J., and van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell **143:** 212–224.

Filipescu, D., Szenker, E., and Almouzni, G. (2013). Developmental roles of histone H3 variants and their chaperones. Trends Genet. **29:** 630–640.

Grob, S., Schmid, M.W., Luedtke, N.W., Wicker, T., and Grossniklaus, U. (2013). Characterization of chromosomal architecture in Arabidopsis by chromosome conformation capture. Genome Biol. **14:** R129.

Henikoff, S., and Shilatifard, A. (2011). Histone modification: cause or cog? Trends Genet. **27:** 389–396.

Jacob, Y., Stroud, H., Leblanc, C., Feng, S., Zhuo, L., Caro, E., Hassel, C., Gutierrez, C., Michaels, S.D., and Jacobsen, S.E. (2010). Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. Nature **466:** 987–991.

Kharchenko, P.V., et al. (2011). 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.

Kurihara, Y., Schmitz, R.J., Nery, J.R., Schultz, M.D., Okubo-Kurihara, E., and Ecker, J.R. (2012). Surveillance of 3′ noncoding transcripts requires FIERY and XRN3. G3 (Bethesda) **2:** 487–498.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. **11:** 204–220.

Lee, J.S., Smith, E., and Shilatifard, A. (2010). The language of histone crosstalk. Cell **142:** 682–685.

Nicol, J.W., Helt, G.A., Blanchard, S.G., Jr., Raja, A., and Loraine, A.E. (2009). The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics **25:** 2730–2731.

Roh, T.Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. Proc. Natl. Acad. Sci. USA **103:** 15782–15787.

Roudier, F., et al. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J. **30:** 1928–1938.

Shu, H., Wildhaber, T., Siretskiy, A., Gruissem, W., and Hennig, L. (2012). Distinct modes of DNA accessibility in plant chromatin. Nat. Commun. **3:** 1281.

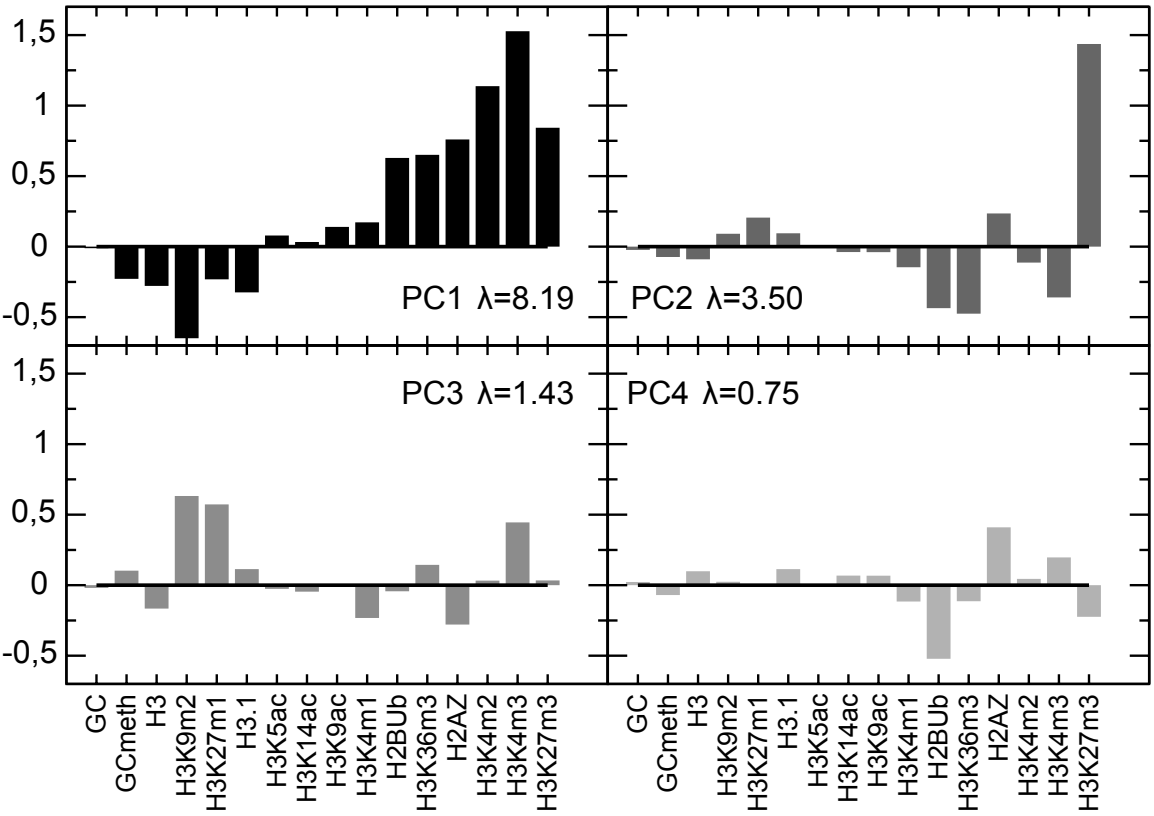
Skene, P.J., and Henikoff, S. (2013). Histone variants in pluripotency and disease. Development **140:** 2513–2524.

Stroud, H., Otero, S., Desvoyes, B., Ramírez-Parra, E., Jacobsen, S.E., and Gutierrez, C. (2012). Genome-wide analysis of histone H3.1 and H3.3 variants in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **109:** 5370–5375.

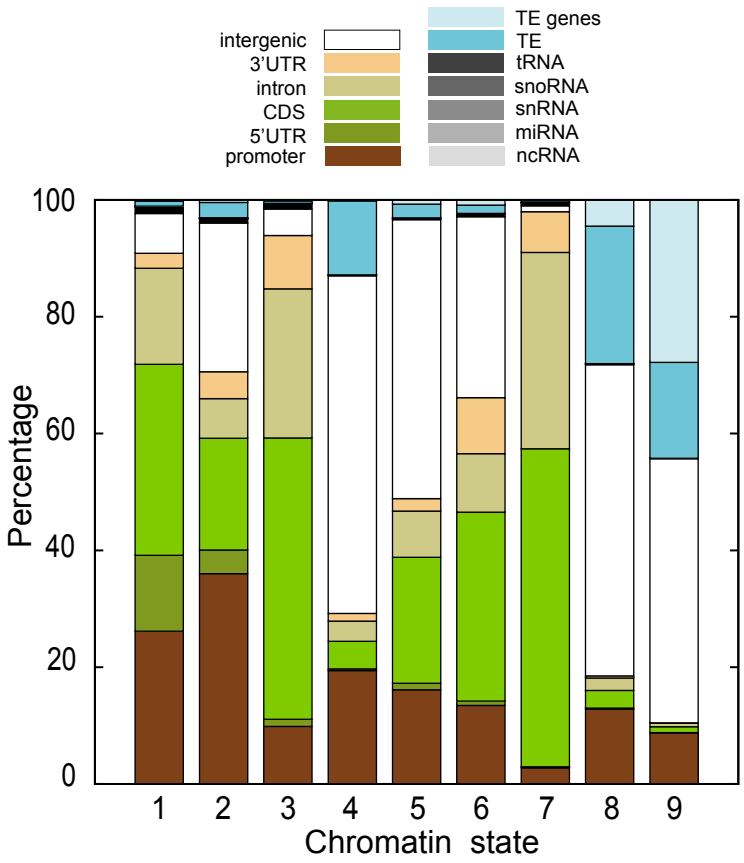
Zhang, X., Bernatavichute, Y.V., Cokus, S., Pellegrini, M., and Jacobsen, S.E. (2009). Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. Genome Biol. **10:** R62.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y.V., Pellegrini, M., Goodrich, J., and Jacobsen, S.E. (2007). Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. PLoS Biol. **5:** e129.

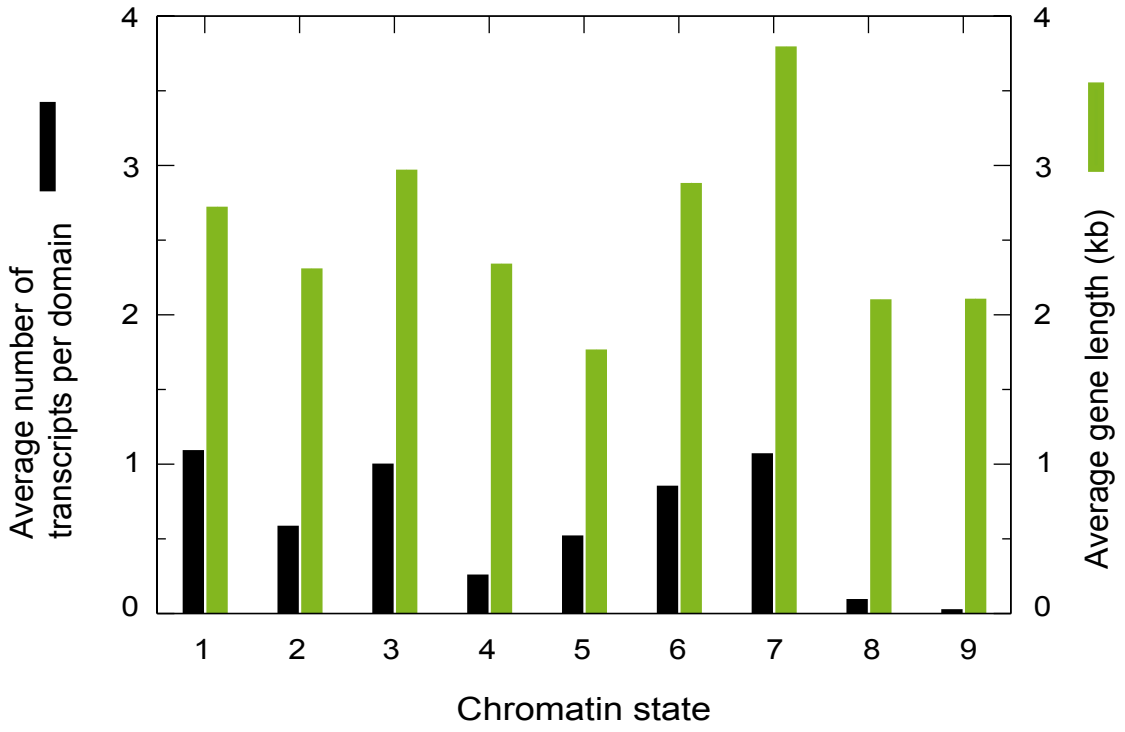
Zilberman, D., Coleman-Derr, D., Ballinger, T., and Henikoff, S. (2008). Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature **456:** 125–129.



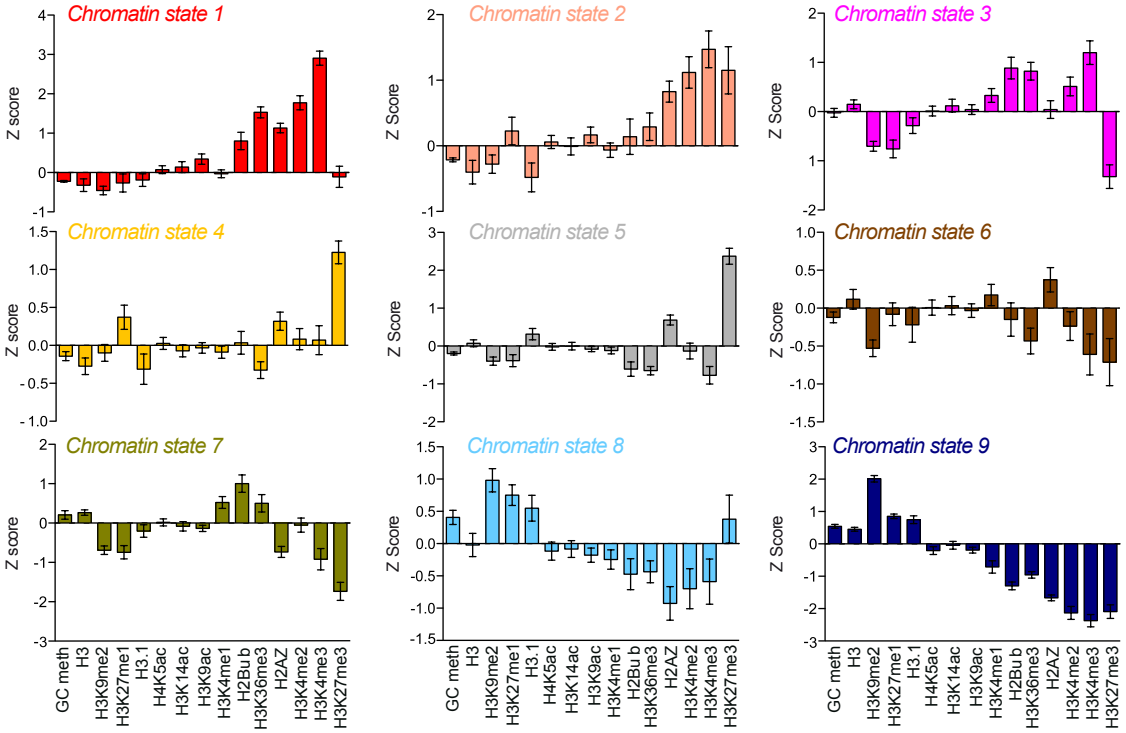
Supplemental Figure 1. Relative values of each of the genomic features used in this study for each of the principal components considerd. See Methods for details.



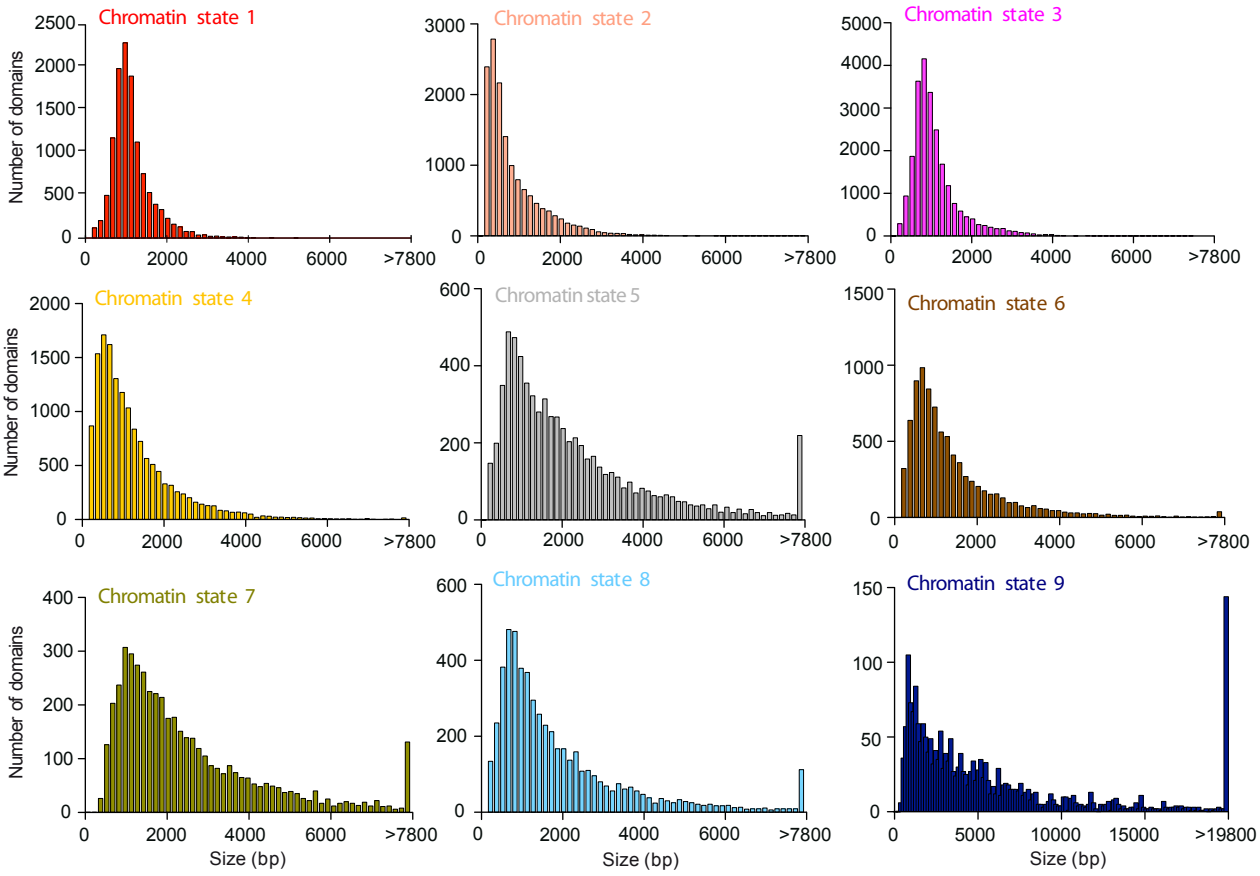
Supplemental Figure 2. Relationship between genomic elements and chromatin states. The overlap (in bp) between the indicated genomic elements and each chromatin state was computed and expressed as a percentage. A promoter region of 1 kb was considered.



Supplemental Figure 3. Calculation of the number of transcripts per domain (black bars and the average size of genes (in kb; green bars) associated with each chromatin state. See Methods for details and Supplemental Figure 5 for domain size summary.

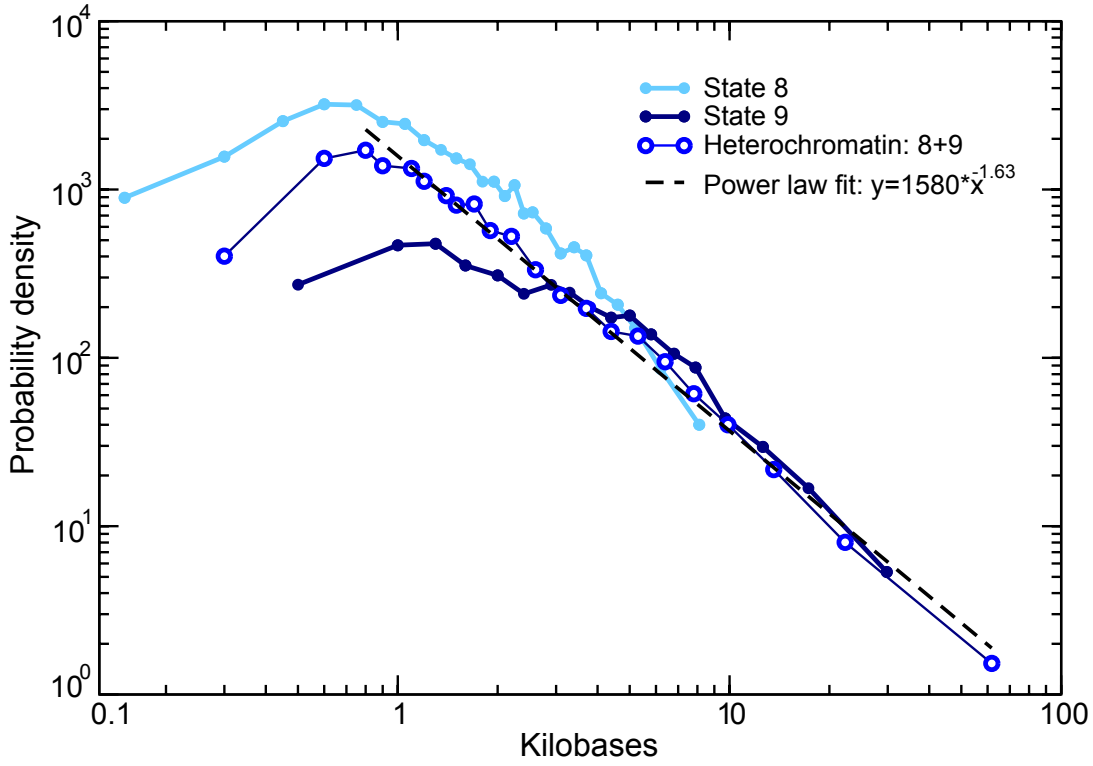


Supplemental Figure 4. Evaluation of PCA without considering the GC content. The robustness of the nine chromatin states obtained if the GC content is not used for the clustering is evaluated here by plotting the average values for each epigenetic property. The parameters w=150, s=0.10, n=4 were used. Error bars represent the standard error of the mean. This number is estimated as the total number of windows divided by the correlation length of the mark considered.

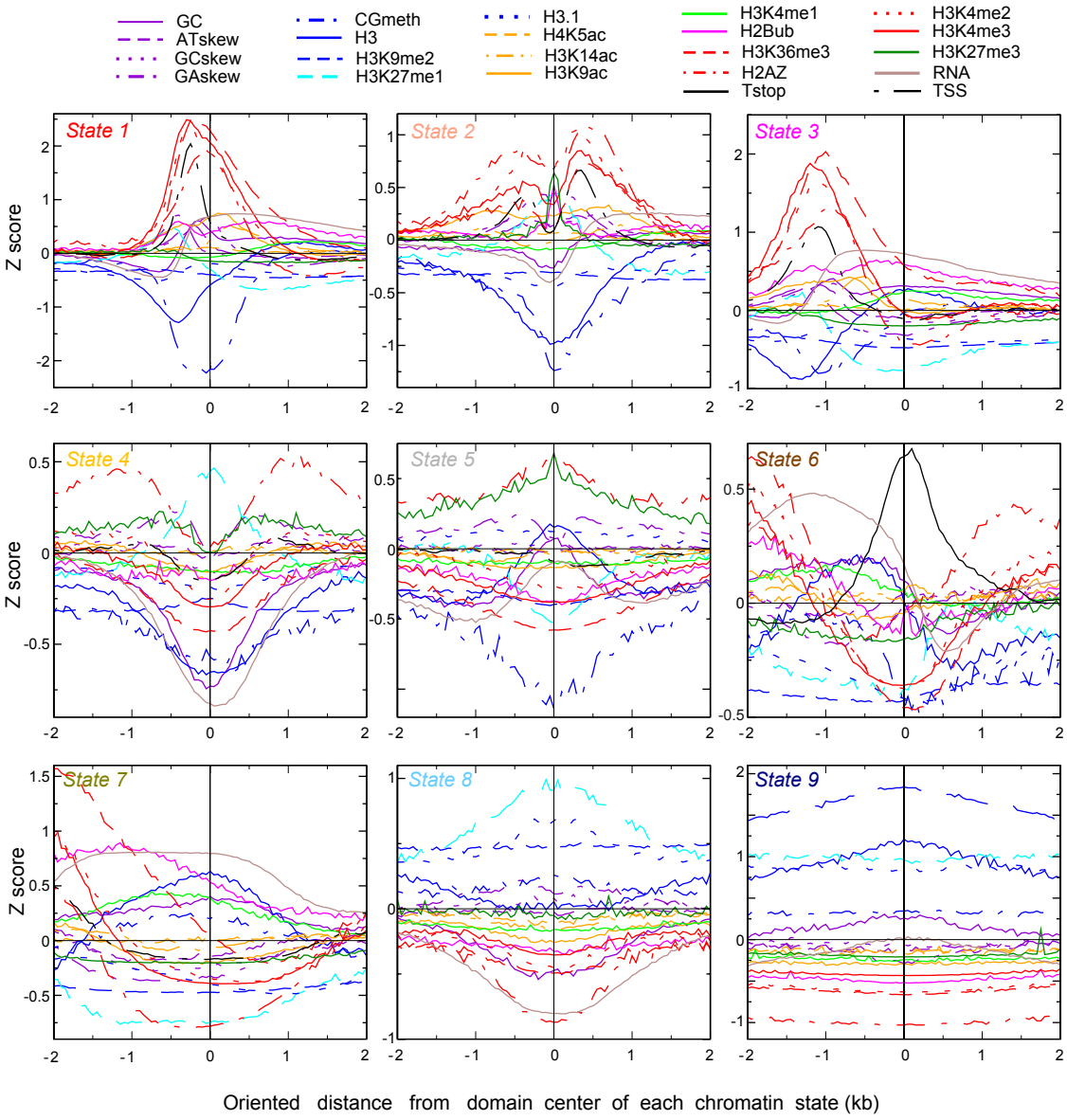


States	n	Mean Size (kb)	Median Size (kb)	Typical Size (kb)
1	12220	1.079	0.90	0.687
2	14502	0.785	0.45	0.725
3	11751	1.023	0.90	0.721
4	14950	1.179	0.90	1.060
5	6943	2.299	1.65	2.196
6	8603	1.368	0.90	1.279
7	4715	2.497	1.80	2.039
8	5470	1.965	1.20	1.845
9	2191	6.702	3.60	4.185

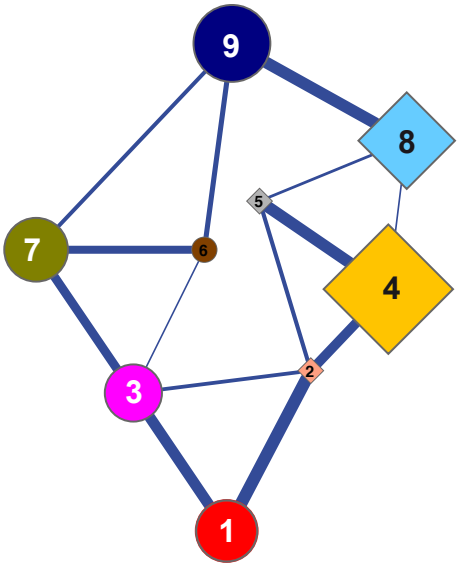
Supplemental Figure 5. Domain size distributions for the different chromatin states. Information of number of domains, mean, median and typical size (i.e. the parameter l of the exponential distribution of domain size, $e^{-s/\lambda}/l$) (in kb) is given at the bottom.



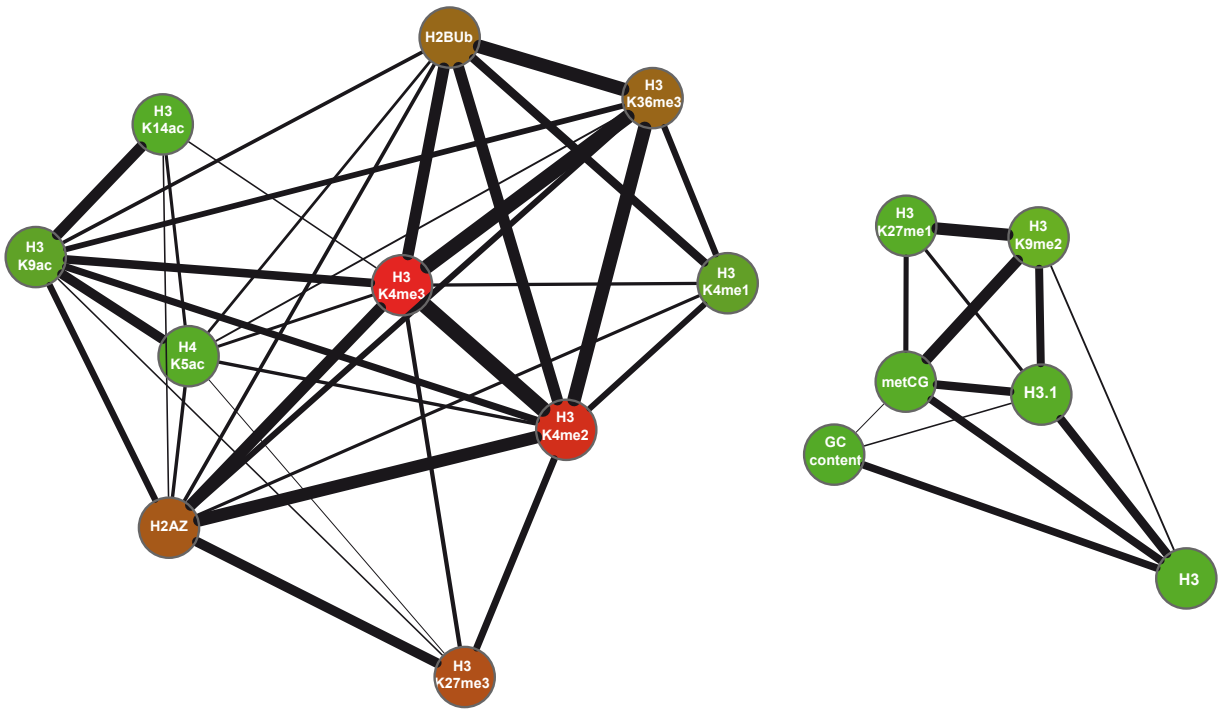
Supplemental Figure 6. Distribution of the domain size of the heterochromatin states 8 and 9 considered individually or combined. While state 8 domains (AT-rich heterochromatin) show an exponential distribution, and state 9 domains are depleted of short domains, the heterochromatin state obtained by joining states 8 and 9 presents a power law distribution with exponent -1.63 over one and half decade (from 2 to 70 kbp), which suggests that heterochromatin is approximately scale-free.



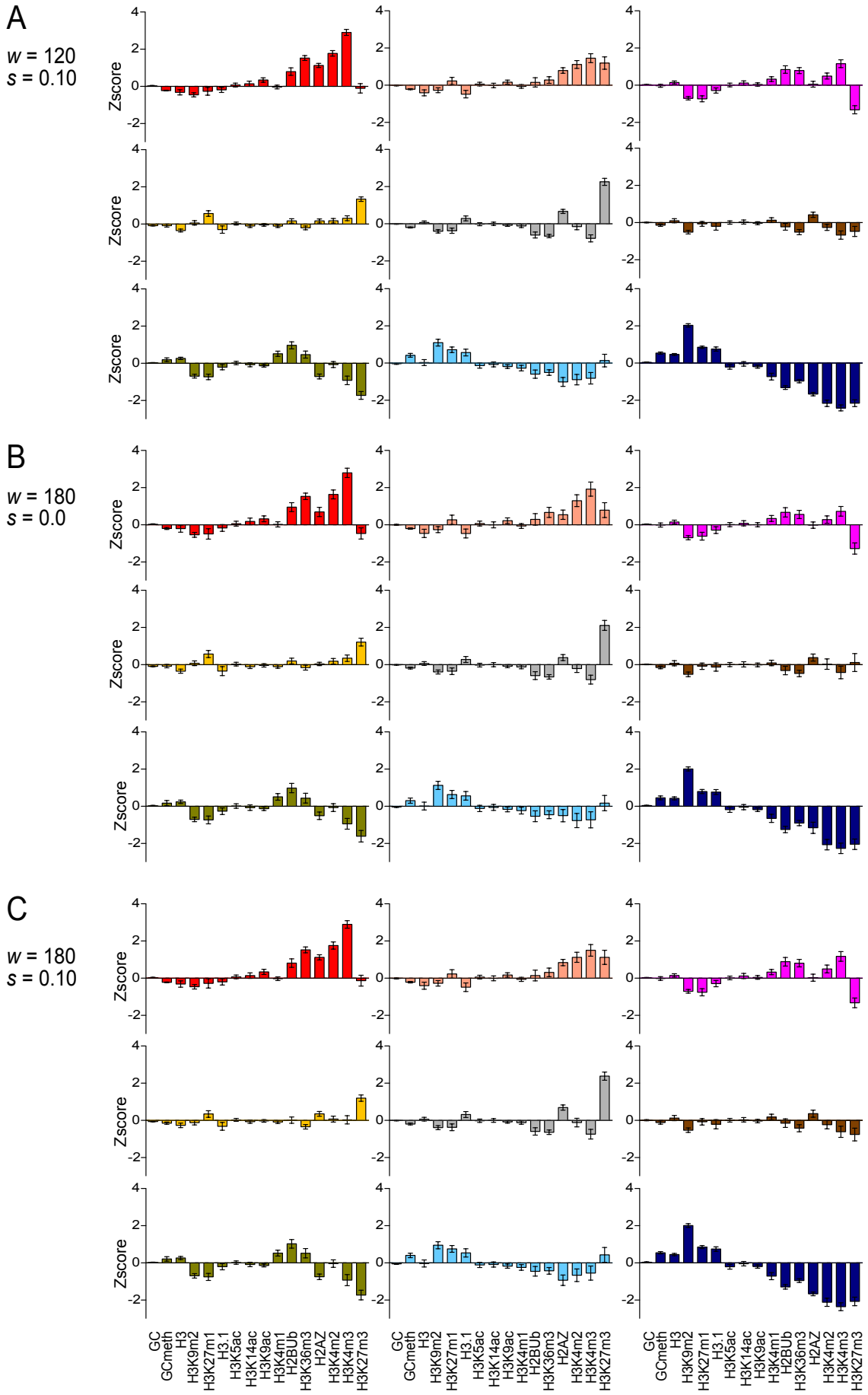
Supplemental Figure 7. Estimation of the relative enrichment of histone marks and DNA sequence features in the 9 chromatin states. The distribution of each chromatin and DNA feature was determined around the center of the domain taking into account the orientation of transcription.



Supplemental Figure 8. Network similarity diagram of the frequency of transition between the 9 chromatin states. Diamonds and circles represent AT-rich and GC-rich states, respectively. Symbol size represents the deviation in GC content with respect to the average genomic content. Circles are states with GC content larger than average, and diamonds are states with low GC content. The thickness of the lines connecting chromatin states is proportional to the similarity degree between two given states.



Supplemental Figure 9. Network correlation diagram of the frequency of transition between different chromatin features (histone marks and GC content) used in this study. Chromatin features are connected by lines, the thickness of which is proportional to the positive correlation between them. Colors of marks depend on their contribution to PC1 (red, high contribution; brown, mid contribution; green, low contribution). Note that chromatin features separate into two groups largely corresponding to euchromatin and heterochromatin.



Supplemental Figure 10. Robustness of the nine chromatin states obtained with different parameters.
The average values of the genomic and epigenomic marks of each of the nine states are shown for parameters w=120, s=0.10 (A), w=180, s=0.0 (B), w=180, s=0.10 (C). Similarity with the optimized parameters w=150, s=0.10 shows that the results are robust even to large variations in parameters.

Supplemental Table 1. Oligonucleotides used in the sequential ChIP (Re-ChIP)-qPCR

Region a (Chr1-16,578 kb)		
	1F	ccggctctaaaacaccaaaa
	1R	gggtcgggtaagaaagaagc
Region b (Chr1-11,630 kb)		
	1F	tcttctctgccatgtcgatg
	1R	catctgtggaaaccgactga
Region c (Chr1-27,397 kb)		
	1F	tctcgaagcaaaggtggatt
	1R	cccttggctgagatgagaag
Region d (Chr5- kb)		
	1F	caacggttcttcatccgatt
	1R	ctgctcgaaatggctctacc
Region control (Chr1-9,039 kb)		
	1F	tgctcgtcccatttcctatc
	1R	ggcatagtgattttgccaca