Functional analysis and co-evolutionary model of chromatin and DNA methylation networks in embryonic stem cells

Enrique Carrillo de Santa Pau[1,*], Juliane Perner[2,*], David Juan[1,*], Simone Marsili[1], David Ochoa[3], Ho-Ryun Chung[4,&], Daniel Rico[1,&,$], Martin Vingron[2,&] and Alfonso Valencia[1,&,$]

[1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Center, CNIO. Melchor Fernandez Almagro, 3. 28029 Madrid, Spain.

[2]Computational Molecular Biology, Max Plank Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany.

[3]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus. Hinxton, Cambridge CB10 1SD, United Kingdom.

[4]Otto-Warburg-Laboratories Epigenomics, Max Plank Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany.

*Contributed equally.

[&]Jointly supervised research.

[$]Corresponding authors: drico@cnio.es, valencia@cnio.es

Abstract

We have analyzed publicly available epigenomic data of mouse embryonic stem cells (ESCs) combining diverse next-generation sequencing (NGS) studies (139 experiments from 30 datasets with a total of 77 epigenomic features) into a homogeneous dataset comprising various cytosine modifications (5mC, 5hmC and 5fC), histone marks and Chromatin related Proteins (CrPs). We applied a set of newly developed statistical analysis methods with the goal of understanding the associations between chromatin states, detecting co-occurrence of DNA-protein binding and epigenetic modification events, as well as detecting coevolution of core CrPs. The resulting

networks reveal the complex relations between cytosine modifications and protein complexes and their dependence on defined ESC chromatin contexts.

A detailed analysis allows us to detect proteins associated to particular chromatin states whose functions are related to the different cytosine modifications, i.e. RYBP with 5fC and 5hmC, NIPBL with 5hmC and OGT with 5hmC. Moreover, in a co-evolutionary analysis suggesting a central role of the Cohesin complex in the evolution of the epigenomic network, as well as strong co-evolutionary links between proteins that co-locate in the ESC epigenome with DNA methylation (MBD2 and CBX3) and hydroxymethylation (TET1 and KDM2A).

In summary, the new application of computational methodologies reveals the complex network of relations between cytosine modifications and epigenomic players that is essential in shaping the molecular state of ESCs.

Introduction

Cell identity depends on complex networks of chemical processes that modify the chromatin ("epigenomic remodelling") and lead to distinct epigenomic states. The current cell developmental state and its future possible fates are believed to be determined by the epigenomic landscape. Disruption of these landscapes is associated with disease and cellular transformation. In the case of embryonic stem cells (ESCs), the range of available cell differentiation options is very broad and changes in the epigenome are essential for lineage specification.

DNA methylation (5-methylcytosine; 5mC), certain histone marks and chromatin-related proteins (CrPs) critically contribute to the plastic state needed for induction and maintenance of pluripotency. Recently, novel cytosine modifications with potential regulatory roles such as 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC), and 5-carboxylcytosine (5-caC) have been described[1–6]. However, the physiological and disease role of these modifications is not yet well understood and the biological function is being elucidated[7].

ESCs constitute an ideal cellular model to study cytosine modifications, as they have very active

levels of TET1 that catalyze the oxidation of 5mC[8]. Several groups have reported genome-wide maps of cytosine modifications, histone marks and CrPs in mouse ESCs (for the full list of references, see Supplementary Table 1). Although many of the large-scale epigenomic datasets generated by different laboratories are currently available, the information is still heterogeneous and dispersed in different datasets. These datasets are normally processed with different bioinformatic protocols, making it even more challenging to analyze the relationships between the different epigenomic players and the functional implications.

Here, we combine diverse epigenomic datasets generated by next-generation sequencing (NGS, 139 experiments from 30 datasets with a total of 77 epigenomic features) to establish a homogeneous dataset between of different DNA methylation states, histone marks and CrPs using robust statistical approaches. We applied a set of newly developed analysis methods, including statistical analysis of networks of chromatin states, co-occurrence of DNA protein binding events and protein coevolution. Our results yield improved understanding of the functional role of different types of cytosine methylations marks within the molecular network of chromatin components and its evolutionary history and structure.

RESULTS

A chromatin atlas of mouse ESCs

We compiled a large collection of genome-wide data from 139 assays, profiled by ChIP-Seq (Chromatin Immunoprecipitation Sequencing), MeDIP (Methylated DNA immunoprecipitation) and GLIB (glucosylation, periodate oxidation and biotinylation) for mouse embryonic stem cells (mESCs, see Supplementary Table 1). This collection includes 3 types of cytosine modifications (5mC, 5hmC and 5fC), 13 histone marks (H2Aub1, H2AZ, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me2, H3K36me3, H3K79me2, H4K20me3) and 61 different Chromatin related Proteins (CrPs). CrPs include structural proteins, members of the machinery involved in cytosine and histone modifications, transcription factors (TFs, such as stemness-related TFs NANOG, OCT4 and SOX2), and four different post-translational modifications of RNA polymerase II (RNAPolII; S2P, S5P, S7P and 8WG16 - unmodified) with binding data available for ESCs. All data were downloaded from public databases and processed

in a systematic way using the same pipeline (see Methods). To visualize the full data set collected for the 77 epigenomic features a UCSC track hub has been generated and it is available at http://genome.ucsc.edu/cgi-bin/hgTracks?db=mm9&hubUrl=http://ubio.bioinfo.cnio.es/data/mESC_CNIO/mESC_CNIO_hub2/hub.txt

The genome of a given cell type can be classified in epigenetically-defined chromatin states[9]. Combinatorial methods such as ChromHMM[10] based on hidden markov models have become a de-facto standard for the functional classification of genomic elements, such as active/inactive/poised promoters, elongation, active/inactive enhancers, etc, based on epigenomic data. We applied this methodology using as input information 17 "core epigenomic features" including the 3 cytosine modifications, the 13 histone marks and the insulator protein CTCF - which has been previously shown to define a particular chromatin state *per se*[11]. In this framework, chromatin states are defined by the combinations of these "core epigenomic features" that would function as "platforms" for the interactions of other non-histone proteins[12].

An empirically defined model of 20-states was established following previous selection strategies[11,13] (see online Methods, Fig. 1A). We identified states related with transcription elongation (states 1-5), heterochromatin (6-10), enhancers (11-14), promoter activation (15-17) and repression (18-19), and the CTCF insulator (20), that are consistent with previous knowledge on the function of the features (Fig. 1A).

The good representation of several important chromatin related protein complexes in our ESC collection allows further exploration of their relationship with the 20 chromatin states. We identify clear associations between certain chromatin states and CrP complexes known to be functionally related to them (Supplementary Figures 1-5). For example, active promoter state 16 shows a clear enrichment of proteins of the Mediator complex, including RNAPolII (Fig. 1B and Supplementary Figures 2-3). States 18 and 19 characterized by repressive histone marks and related to promoter regions of lowly express genes (see Fig.1A), are associated with proteins of the Polycomb complex that play a key role in inactivation (Fig. 1C and Supplementary Figures 1 and 4). State 20 characterized by the CTCF insulator is additionally associated with cohesin proteins (Fig. 1D and Supplementary Figures 1 and 5), in agreement with previous analyses[12,14] and the recently proposed role of CTCF determining the localization of most of the cohesins in

the genome[15].

To disentangle the relationships between the chromatin states, we analyzed the association between the states based on the epigenomic features (histone marks, cytosine modifications and CrPs). The partial correlation analysis based on the frequency of occurrence of the features, provides a clean network of principal associations between states (Fig. 1E). This network contains 12 positive associations as well as 4 negatively correlated connections involving the empty heterochromatic state 7. Six out of the 12 positive links connect states with a similar functional role but with different levels of occupancy of the epigenomic features (Supplementary Figures 6-25). For example, we detect associations between stronger and weaker states of transcriptional elongation or promoter elongation (Fig. 1E).

The remaining links connect states with different but complementary roles, determined by a few key (or by some missing) epigenomic features. For instance, state 15 (promoter activation) shows positive association with states 1 and 2 (strong transcriptional elongation). These states present a similar profile of CrPs (Supplementary Figures 6, 7 and 20), as well as the strong enrichment in H3K79me2 in states 15 and 1 (Fig. 1E), a mark related with transcription initiation start just downstream to the promoter[16]. This suggests the involvement of states 1 and 15 in transcription initiation, while the higher enrichment in H3K4me2/3 for state 15 supports its classification as an active promoter. In contrast, state 2 show higher enrichment in the three cytosine modifications (Fig. 1E), consistent with the higher enrichment of the cytosine modification proteins TET1, OGT and the methyl-binding domain (MBD) proteins MBD1B and MBD2T (Supplementary Fig. 7).

Interestingly, the four enhancer states (11-14) were separated into two different subnetworks. States 11 and 12 are connected to the subnetwork implicated in transcriptional elongation (states 1-4); a link that might reflect the recently proposed role of elongation factors in posing enhancers in embryonic stem cells[17]. In contrast, enhancer states 13 and 14 (the most active enhancer states) form a different subnetwork together with the active promoter state 17. In fact, states 12 and 14 present very complementary profiles of the "core" genomic features. State 14 is highly enriched in H3K27ac, p300 and TFs (Supplementary Fig. 19), consistent to its role as active

enhancer. State 12 is highly associated to the cytosine modifications, H2AZ and cohesin (Supplementary Fig. 17) indicating that to these states may function as poised enhancers.

The five heterochromatic states are mostly unconnected in the network of chromatin states and show very different enrichments in the cytosine modifications. The state 6 enriched in H3K9me3 and the "empty state 7" are mutually exclusive (negative connection) and are not associated to any cytosine modification. 5mC is mainly enriched in the heterochromatic state 10, characterized by the absence of DNaseI signal and a clear enrichment in LaminB1, MBDs and regions with sequence repeats (Fig. 1A and Supplementary Fig. 15).

Finally, 5fC was most enriched in the unconnected state 8 (Fig. 1E), which is particularly enriched in simple repeats (Fig. 1A). DNA methylation (5mC) has been traditionally associated to retrovirus transcription repression in heterochromatin[18]. Our results suggest that the 5fC modification might be also important for this role. RYPB, usually associated to PRC1 in the Polycomb complex[19], is one of the two enriched CrPs in this quite empty state (together with MBD2T, see Supplementary Fig. 13). Interestingly, it has been recently described that RYBP participates in the repression of endogenous retrovirus in mouse stem cells[20]. Therefore, we hypothesize that RYBP-mediated retrovirus repression could be related to cytosine conversion processes that result into 5fC production or RYBP binding to 5fC rich regions.

By combining all available epigenomic data, we have characterized the 20 chromatin states of mouse ESCs. From the analyses above, we obtained a general overview of the relationships between chromatin states and the corresponding cytosine modifications, histone marks and CrPs, and a contrasted classification according to their functional role. To further investigate the organization of chromatin in ESCs, we inverted the problem to study the differential relations between cytosine modifications, histone marks and CrPs, in each one of the chromatin states.

Chromatin-State Dependent Co-Location Networks

To detect specific interactions between the components based on their binding co-localization it is necessary to eliminate indirect/transitive effects, i.e. co-localization that is introduced by other (observed) factors. To this end, we applied the method described in[21] (see Methods for details).

This methodology aims at unraveling interactions between factors, which cannot be "explained away" by the other observed factors and thus is more specific than an analysis of simple pair-wise correlations.

To obtain an interaction network associated to each state (context-dependent), we analyzed separately the genomic regions assigned to a certain chromatin state. The global network with the information of all the states (Fig. 2A) summarizes all direct interactions between cytosine modifications, histone marks and CrPs. In total, this "co-location network" connects 66 nodes by 149 edges, and as in the case of other biological networks it is dominated by a small number of hubs, i.e. MBD2T, G9A, SUZ12, RYBP and RNAPoIIIS2P (Supplementary Fig. 26). The networks specific of each one of the chromatin states have between 53 and 77 edges connecting between 51 and 62 nodes. These 20 chromatin specific networks, as well as the possible combinations of states, can be explored using EpiStemNet, an interactive viewer of the "co-location" network (http://dogcaesar.github.io/epistemnet).

About 13% of the interactions are present in all chromatin states (blue edges in Fig. 2A) and 25% of them are present in a single state (orange edges in figure 2), indicating a preferential association of CrPs to chromatin contexts. In fact, 85% of interactions are positive, supporting the classification in chromatin states and their proposed role as differential epigenomic binding platforms. Up to 36% (24 out of the 66) the direct protein-protein positive interactions in the "co-location" network as well as 11 out of the 62 indirect positive (positive-positive or negative-negative) associations mediated by a histone mark or a cytosine modification, correspond to previously characterized known and putative functional interactions obtained from STRING[22].

Moreover, the "co-location network" contains many interactions previously described in the literature, including the main epigenetic complexes, including known interactions among members of Polycomb, Cohesin and Mediator complexes (Fig. 2B). In addition, we identify interactions between members of complexes related with gene repression activity as the nucleosome remodeling deacetylase MI2/NuRD complex (MI2B/LSD1/MBD2/MBD3/HDAC1/HDAC2), the CoREST/Rest complex (Rest/CoREST/RYBP) or SETDB1 complex (SETDB1/H3K9me3). Another interesting group of

interactions are those related with different sites of phosphorylation of the RNA polymerase II, that are connected among them and with specific histone marks. S2P PolII is connected with H3K36m3 and H3K27ac, while unmodified PolII is connected with H3K4me3 reflecting the distribution of these modifications through coding regions or promoters, respectively[23]. In contrast, S5P RNAPolII is connected with H2Aub1 related with bivalent promoters and poised polymerase[23]. Remarkably, we also recover significant co-location signals of stemness-related complexes NANOG/OCT4/SOX2/TCF3 and SIN3A (SIN3A/TET1/OGT)[3,24].

Our analysis also recovers a number of interesting negative co-locations (mutually exclusive genomic features). About 68% of these negative associations (15 out of 22) involve either cytosine modifications or MBD2T, suggesting that cytosine modifications are probably important in defining mutually excluding epigenomic features within particular chromatin states. For instance, MBD2T and 5mC are positively connected to the repressive CBX3 protein, but they are negatively connected to a variety of activation histone marks, highlighting the repressive role of 5mC. In the next section we will explore this subnetwork in more detail.
As whole, our results show that our approach for disentangling unspecific co-location signals allows to retrieve a global picture of the main complexes operating in ESCs, as well as interesting logics operating within chromatin states.

Common and specific direct interactors of 5mC, 5hmC and 5fC

The co-location network can be used to shed light on the relation between methylation marks and the epigenetic machinery, since the biological roles for 5hmC and 5fC are not well understood and the literature remains somewhat controversial[1–6]. Figure 3 shows all the CrPs directly interacting with the three cytosine modifications in each one of the 20 chromatin states.

Overall, 5hmC and 5fC tend to be more connected via common interactors (NIPBL, RYBP and TET1-OGT), while 5mC appears to be more isolated and directly connected with 5hmC only via G9A histone methyltransferase in chromatin states 16 and 17 (active promoters, positive connection 5hmC-G9A) plus a negative connection between 5mC-G9A in states 1, 9 and 13. The position of G9A in the subnetwork indicates that it is indeed an essential protein in DNA methylation[25].

Little is known about cytosine modification readers[26,27] and only few of the cytosine modification interactions showed in the sub-network have been described previously, i.e. the positive interactions between TET1 with 5hmC[3,28] and p300 with 5fC[29]. KDM2B has been shown to bind preferentially to unmodified cytosine[26], which explains the anti-correlation to 5hmC in the Polycomb-related state 18. Interestingly, the complex association observed for 5hmC to the KDM proteins relates 5hmC with the demethylation of H3K4 and H3K36, respectively, affecting the gene expression status.

Moreover, we have identified a positive correlation (stronger in polycomb state) between MBD2T and 5mC[30] and a weakly negative interaction (mutual exclusion) between MBD2 and 5hmC across G9A in active promoter states (16 and 17). Although the interaction of MBD2 with 5mC has been extensively described[30,31], the recognition of 5hmC by MBD proteins remains controversial[30]. Our analysis suggests that MBD2 only interacts with 5mC - at least in ESCs.

Another interesting observation is the relation of 5hmC and 5fC with TET1. It is well known that TET proteins catalyze the oxidation of 5mC into 5hmC, 5fC and finally in 5caC[32]. We found that the interaction of TET1 with 5hmC is direct, while the interaction with 5fC is via OGT. The interaction between TET1 and OGT has been recently described in mESCs[24], where all OGT binding sites are co-occupied by TET1. However, no mechanistic roles for this complex have been described to date. Interestingly, the interaction of TET1 and OGT is found in particular chromatin states (2, 11, 12, 14, 15, 16, 18, 19) with different functional roles (elongation, enhancer, active transcription and repression/Polycomb; Supplementary Figures 6-25). Our analysis suggests that the interaction between OGT and TET1 could be important for TET1-mediated cytosine oxidation, or to recruit other proteins to the complex for specific cytosine modifications as 5fC.

The rest of the interactions in the co-location network based on the analysis of high-throughput experimental data, suggest specific roles of the cytosine modifications in CrPs complexes. Obviously, the verification of each one of them will require further detailed experimental work.

A co-evolution network of the chromatin-related proteins

We have described part of the functional structure of the epigenetic network of cytosine modifications, histone marks and CrPs in ESCs. Given its implication in cell differentiation, it is reasonable to assume that the ESC epigenetic network played an important role in metazoan evolution. To further characterize this network it is essential to get additional insight on the evolutionary relations between co-evolving protein components that form the functional core of the system. Detecting these evolutionary relations between the protein components of the epigenetic network will inform us about its basic structure and its adaptations to different biological scenarios.

To extract these evolutionary relations we implemented a co-evolutionary-based methodology particularly suitable for this scenario (for a general reference on co-evolution based methods see[33], specific details of the implementation are given in Methods). To collect the information required for the co-evolutionary study we retrieved 13,579 trees from eggNOG database[34] including over a million protein sequences from 88 metazoan species. For the co-evolutionary analysis we built a maximum-entropy model of pairwise interacting proteins[35,36] based on the similarities between the evolutionary trees of orthologs for the 58 mouse CrPs included in the "co-location network", while the whole set of 13,579 trees was used to evaluate the statistical significance of the results (see Methods). Similarly to the discussed methodology for the co-location network, this large-scale approach allows us to detect significant connections between functional and structural modules by dissecting direct protein-protein co-evolutionary relationships from the large "hairball" of indirect interactions[37].

The combination of co-evolutionary signal and complex membership yields a highly populated network (Fig. 4A). A third of the 34 statistically significant connections ($p < 0.05$) between the 38 connected nodes correspond to connections supported by the STRING database (seven) or described in the literature (three physical and two regulatory interactions; see Supplementary Table 2). The Cohesin complex acts as the central element of the network, with its proteins co-evolving with members of five different complexes (NuRD Mediator, RNA polymerase II-Activation/Repression , SOX2/OCT4/NANOG/TCF3 and TET1/OGT/SIN3A complexes).

The Cohesin complex is strongly coevolving with the NuRD complex, an evolutionary interdependence that points to their interaction with the SWI/SNF complex involved in cohesin loading[38]. In addition, the Cohesin complex co-evolves with the Mediator member NIPBL, a known loading factor of cohesin[39]. Taken together, these results point to the evolutionary importance of the process of Cohesin loading into the chromatin.

Interestingly, the functional relationship of five co-evolving protein pairs that are also connected by the co-location network has not been previously reported (see Supplementary Table 2). Among them, the SMC1A-HDAC1 protein pair, which show the strongest signal of co-evolution, is particularly interesting as it shows a negative co-localization with H3K79me2 and it puts in contact the HDAC (HDAC2) and cohesin (SMC3) complexes. Given that NIPBL is also recruiting HDAC1/3[40], our result suggests that Histone deacetylation and Cohesin functions are strongly associated.

We further detect a coevolution signal between RNA polymerase II-Activation/Repression and Cohesin as well as the RNA polymerase II-Mediator, which are involved in transcription regulation. Another highly connected complex is the NuRD complex that interacts with three other complexes and with several other proteins. In addition to Cohesin, NuRD co-evolves with SETDB1, a protein that interacts with MBD1 in the NuRD complex[41,42] and with SIN3A in the SIN3A/TET1/OGT and REST-CoREST complexes[43]. These and other associations in this network provide a compelling starting point for future work. In the next section, we are going to focus on the cytosine modification-related subnetwork previously defined and discussed in the co-location analysis.

Coevolution signals between proteins associated to 5mC, 5hmC and 5fC

In order, to further characterize the relationships among proteins co-localized with the different cytosine modifications in the epigenome of ESCs, we analyzed this co-evolutionary sub-network in more detail (Fig. 4B). There are four protein co-evolving protein pairs connecting seven proteins involved in co-location interactions with cytosine modifications (see Fig. 3 for a broader co-location context).

First, there is a strong co-evolution signal for the KDM2A/B pair of histone demethylases. KDM2A and KDM2B also show a strong co-location signal but they are differently associated to 5hmC (Fig. 3).

Second, NIPBL and BRG1 have also a strong co-evolutionary relationship. These two previously unrelated proteins have a negative co-location interaction via by 5fC (i.e. NIPBL co-locate with 5fC, while BRG1 is located in regions free from 5fC).

Finally, we detected two co-evolving pairs linking two proteins positively co-locating with 5mC (CBX3 and MBD2) to two proteins positively co-locating with 5hmC (KDM2A and TET1). KDM2A is known to be recruited by CBX5[44] (the closest paralog of CBX3) and CBX3 interacts with KDM2B[45,46], showing a very strong functional and evolutionary association between both protein families. This association is also relevant, because of the connection of the HDACs to cohesin discussed above.

The co-evolution between MBD2 and TET1 is particularly interesting, given the key role of TET1 in the oxidation of 5mC. MBD2 shows higher binding affinity to 5mC than to 5hmC[30] and MBDs have been suggested to modulate 5hmC levels inhibiting TET1 by their binding to 5mC[31]. The co-evolution detected between MBD2 (also co-evolving with MBD3, see Fig. 4A) and TET1 suggests an interdependence between the mechanisms maintaining 5mC and 5hmC in different epigenomic locations of ESCs. This interdependence can be also observed in the co-localization network, where TET1 shows a double negative connection to MBD2T through H3K4me2 and H3K4me1 (Fig. 2A). In turn, H3K4 demethylation is performed by the KDM2A protein[47] that co-localizes with 5hmC (and with H3K4me1). Therefore, the co-evolutionary analysis points to a strong interplay between histone demethylation and DNA demethylation processes, suggesting that MBD2, TET1 and KDM2A/B are key players in this process.

DISCUSSION

We have obtained a large epigenetic network composed of relations obtained by a co-location analysis of cytosine modifications, histone marks and proteins in mouse ESCs. The relations described in the network are specific of different chromatin states and potentially reflect the functional activities behind genome regulation. The core structure of this epigenetic network is formed by the intimate interplay between protein complexes, histone marks and the different cytosine modifications. The evolutionary analysis of the proteins implicated in the co-location network illustrates how the corresponding protein complexes co-evolved by working together in the regulation of the different cytosine modifications and their interplay with histone marks.

The analysis of the network of relations around the three best characterized DNA modifications (5mC, 5hmC and 5fC), allowed us for the first time to detect a set of proteins which genomic location is specifically dependent on the different cytosine modifications, as well a strong evolutionary co-dependence between some of them. In particular, our analyses reveal an evolutionary interaction between MBD2 and CBX3 proteins, linked to DNA methylation, with proteins TET1 and KDM2A, which genomic location is hydroxymethylation-dependent.

We are still in early stages of the exploration of the epigenetic network behind stemness and cell pluripotency. The epigenetic network includes many relations with proteins for which genome wide location data have still to be produced. The computational framework introduced here represents the basis for the exploration this vast space and provides the first integrative picture of the different players in epigenetic regulation. The future inclusion of experimental data on other states of cell differentiation will make possible to complete the picture and to follow the dynamics of the epigenetic network in different cell lineages.

Methods

ChIP-Seq, MeDIP and GLIB data processing

We downloaded sra files from Chromatin Immunoprecipitation Sequencing (ChIP-Seq),

Methylated DNA immunoprecipitation (MeDIP) and GLIB (glucosylation, periodate oxidation and biotinylation) technique experiments described in Supplementary Table 1. The MeDIP data for 5mC and the GLIB data for 5hmC were obtained from Pastor et al[2], as it has been previously shown that these datasets are less biased to antibody affinity in regions with repeats than other methods[7]. The sra files were transformed in fastq files with the sra-toolkit v2.1.12 and aligned to the reference mm9/NCBI37 genome with bwa[48] v0.5.9-r16 (Li et al, 2009) allowing 0-1 mismatches. Unique reads were converted to BED format.

Genome segmentation

The cytosine modifications, histone marks and CTCF were used to segment the genome in different chromatin states. A multivariate Hidden Markov Model (HMM) was used. This model uses two types of information, the frequency with which different chromatin mark combinations are found with each other and the frequency with which different chromatin states occur in spatial relationships of each other along the genome. To apply this method we used the ChromHmm software[10] (v1.03). The input data to generate the model were the ChIP-Seq, MeDIP and GLIB bed files containing the genomic coordinates and strand orientation of mapped sequences (see above). First ,the genome was divided in 200 bp non-overlapping intervals which we independently assigned if each of the marks was detected as present (1) or not (0) based on the count of tags mapping to the interval and on the basis of a Poisson background model[11] using a threshold of 10^{-4}. After binarization of each mark we trained the HMM model using a fixed number of randomly-initialized hidden states, varying from 20 up to 33 states. We focused on a 20-state model that provides enough resolution to resolve biologically meaningful chromatin patterns. We used this model to compute the probability that each location is in a given chromatin state, and then assigned each 200-bp interval to its most likely state (see Supplementary Tables 3 and 4). Only, intervals with a probability higher than 0.95 were considered in further analysis.

Segment enrichments

The percentage of genome overlap for each state and different annotation data was computed with ChromHmm software on the intervals selected (see above). The genomic, CpG islands, repeats and laminB1 annotations were downloaded from the UCSC Genome Browser website.

DNAseI and RNAseq were obtained from ENCODE E14 cell line (see Supplementary Table 1). The CAGE data were obtained from FANTOM[49] (see Supplementary Table 1). Data for 5hmC from ESC and NPC were obtained from GSE40810 (see Supplementary Table 1), fastq files processing and binarized calls were done as described previously. Interaction ChIA-PET data for ESCs were downloaded from the supplementary material of the original paper[50]. Enrichment fold changes for annotations and CrPs can be consulted in Supplementary Tables 5 and 6.

Simple Correspondence Analysis

The Simple Correspondence Analysis was carried out after calculate the enrichment of the different proteins included in the analysis (see Supplementary Table 7) in each state. For this purpose we binarized the ChIP-seq data as described above for histone modifications and calculated the enrichment in the selected intervals as described for gene annotations. From the fold change and the % of bins in each state we calculated for each protein the % of bins in each state in which is present, following this formula:

% bins state with the protein = Fold Change * % bins state all genome

Once we got the matrix with the % of protein in each state, we carried out the Simple Correspondence Analysis in R package with the CA library [51]. This library implement a multivariate statistical technique proposed that summarize a set of categorical data in a dimensional graphical space to reduce the dimensionality of a data matrix. The ca function provided us the percentages of explained inertias for all possible dimensions and values for the rows and columns dimensions (Supplementary Table 7). Additionally, principal coordinates, squared correlations and contributions for the points of the rows and columns were obtained. We generated a 3D plot with the plot3d.ca function in combination of the RGL library. We plotted the three first dimensions that let us recapitulate almost the 80% of the data variation (inertia) (Supplementary Fig. 1).

Partial correlation analysis of the chromatin states

Partial correlation analysis was performed using the matrix of the % of bins in each state where a histone mark, cytosine modification or CrPs have been detected as explained before. Here,

chromatin states were considered the variables and the genomic features the cases. We calculated the partial correlation matrix of the chromatin states using the GeneNet[52] R package (v1.2.10). In order to deal with the small variables-cases ratio (20 states and 78 genomic features) we used an analytically estimated shrinkage to the identity matrix, then partial correlation matrix was calculated for this corrected matrix. Statistical significance of every state-state partial correlation was determined using the two-sided Student's test implementation in the same package. Partial correlation linkages with a p-value < 0.05 were considered significant and represented in Fig. 2C.

Read counts and pre-processing for the co-location network inference

We used the ChromHMM segments as samples for the network inference. We filtered all bins that are unexpectedly large for each state (the upper 5% for each state) because they can lead to outliers in the data and it is hard to justify where the signal occurs within the region. The distributions of the bin width are shown in Supplementary Fig. 27 and the number of deleted bins per state are shown in Supplementary Fig. 28. For the resulting in 757,045 bins, we counted the overlapping ChIP-Seq reads using Rsamtools. Some of the ChIP-experiments had to be excluded from the network inference due to low number of reads per bin, low number of bins with signal or study dependent artifacts. These include SMAD1_GSE11431, MBD1A_GSE39610, MBD1B_GSE39610, MBD2A_GSE39610, MBD3A_GSE39610, MBD4_GSE39610, MECP2_GSE39610, CTCF_GSE11431, NANOG_GSE11431 and H3K27me3_GSE36114. Using hierarchical clustering with 1-cor(x,y) as distance measure, we find that most replicates or functionally related samples fall into the same branch (Supplementary Fig.29).

Next, replicates were merged by adding up the read counts in each bin. The resulting 71 samples were normalized against the corresponding input by using the same method as described in[21] In short, we estimated the median fold-change of the sample over the input and used this median to shrink the fold-change for each bin towards 1. Finally, the data was log-transformed adding a pseudocount of 1. The final correlation matrix is shown in Supplementary Fig. 30.

Co-location network inference

For each chromHMM state, we inferred an interaction network as previously described[21]. In short, for each state the samples were scaled to have mean 0 and a standard deviation of 1. An

Elastic net was trained in a 10-fold Cross-validation to predict each HM/CTCF/DNA methylation from the CrPs or to predict each CrP from all other CrPs;. Further, the SPCN was obtained using all the available samples. For visualization of the final networks, we selected the interactions between Histone marks/cytosine modifications and CrPs that obtain a high coefficient (w >= 2*sd(all_w)) in the Elastic Net prediction and have a non-zero partial correlation in the SPCN. All median coefficients of the Elastic Net as well as the R^2-values of the prediction over the 10-fold CV per state are given in Supplementary Tables 8 and 9. All partial correlation coefficients of the SPCN per state are given in Supplementary Table 10.

Coevolution-based analysis

We retrieved 13,579 protein trees of sequences at the metazoan level from eggNOG[34] v4.0. These trees include proteins from NS = 88 metazoan species that are either orthologs or paralogs duplicated after metazoan speciation split. Based on these trees, we extracted only-unique-orthologous protein trees for every mouse protein by inferring speciation and duplication nodes using a species-tree reconciliation approach[53] and a previously used pipeline to deal with tree inconsistencies[54]. When more than one ortholog was detected for a mouse protein in a species, the one with the overall shortest evolutionary distance (extracted from the tree) was selected. As a result, we obtained 13,579 only-unique-orthologous protein trees. From these, we extracted those including the mouse proteins with chip-seq data analysed in this study. So, we performed the main analysis on NP = 58 different protein trees including mouse CrPs. The whole population of trees was kept for performing a randomization test for assigning empiric statistical significances to our results.

We codified each protein tree as a vector containing the NS(NS – 1)/2 distances between all the pairs of species in the analysis, and formed a NS(NS – 1) X NP distance matrix containing these vectors as columns. Each row of this matrix represents a different instance of the measure of distances on the set of proteins, for a different pair of species. For each row of the matrix, distances were ranked and binned into five equally populated intervals {ss,s,m,l,ll} according to the four quintiles of the distribution: ss (very short distances), s (short distances), m (around median), l (large distances), ll (very large distances). An additional state NA was used for missing values in the distance matrix. Denoting with p,q two generic proteins and with a,b two

generic intervals, the single and pair frequencies $f_p(a)$ for protein p in bin a and $f_{p,q}(a,b)$ for the pair p,q respectively in bins a,b were computed as averages over the pairs of species for p,q in {1,2,...,NP} and a,b in {ss,s,m,l,ll,NA}.

The maximum-entropy distribution in the space of species-species distance bins {d} for fixed single and pair protein frequencies is given by:

$$P(\{d\}) = Z^{-1} \exp[\Sigma_p h(d_p) + \Sigma_{p,q} J_{p,q}(d_p,d_q)]$$

where Z is the partition function and the parameters h_p and $J_{p,q}$ have to be adjusted in order to match the empirical frequencies f_p and $f_{p,q}$. The parameters $J_{p,q}$ are of special interest here since they regulate the interactions between proteins in the model. For example, a strongly positive parameter $J_{p,q}(ss,ss)$ can be interpreted as the direct symmetrical interaction between the two proteins p and q, favoring the co-occurrence of short distances in the respective trees. The model parameters were determined by maximizing an l_2-regularized version of the (log) pseudo-likelihood[55] of the data:

$$\{\theta_k^*\} = \text{argmax}_\theta [l_{pseudo}(\{\theta_k\}) - \lambda \Sigma_k \theta_k^2]$$

where θ_k denotes a generic parameter of the model and $\lambda = 0.01$.

We determined a coevolutionary coupling $C_{p,q}$ for each pair of proteins p,q from the related set of couplings between bins, represented by the matrix $J_{p,q}(a,b)$ with a,b in {ss,s,m,l,ll}. Bin couplings involving missing values in the original set of distances (NA state) were not included in the definition of $C_{p,q}$. Following an established protocol for contact prediction in protein structural analysis[56], we double-centered the matrix $J_{p,q}$ and computed the Frobenius norm $F_{p,q} = [\Sigma_{a,b = ss,s,m,l,ll} J_{p,q}(a,b)^2]^{1/2}$.

Finally, we applied an average product correction[57] obtaining the coevolutionary coupling between proteins p and q, $C_{p,q} = F_{p,q} - F_p F_q / F$

Statistical significance:

In order to assign statistical significances to our co-evolutionary couplings, we randomly selected 10,000 groups of mouse proteins from the same size as our set of chromatin modifiers. We run the pipeline described above for every random set and retrieved the corresponding matrix

of coevolutionary couplings. P values were assigned based on the obtained random distribution and associations supported by p values < 0.05 were further considered. The matrix of coevolutionary couplings and corresponding p values are included in Supplementary Table 11.

URLs

UCSC Trackhub with chromatin states, cytosine modifications, histone marks and CrPs

http://genome.ucsc.edu/cgi-bin/hgTracks?db=mm9&hubUrl=http://ubio.bioinfo.cnio.es/data/mESC_CNIO/mESC_CNIO_hub2/hub.txt

EpiStemNet: chromatin state specific co-location networks in ESCs

http://dogcaesar.github.io/epistemnet

References

1. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473,** 398–402 (2011).
2. Pastor, W. A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473,** 394–397 (2011).
3. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473,** 343–348 (2011).
4. He, Y.-F. *et al.* Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science* **333,** 1303–1307 (2011).
5. Ito, S. *et al.* Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* **333,** 1300–1303 (2011).
6. Raiber, E.-A. *et al.* Genome-wide distribution of 5-formylcytosine in embryonic stem cells is

associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* **13,** R69 (2012).

7. Matarese, F., Carrillo-de Santa Pau, E. & Stunnenberg, H. G. 5-Hydroxymethylcytosine: a new kid on the epigenetic block? *Mol. Syst. Biol.* **7,** 562–562 (2011).
8. Tahiliani, M. *et al.* Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science* **324,** 930–935 (2009).
9. Chen, T. & Dent, S. Y. R. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat. Rev. Genet.* **15,** 93–106 (2014).
10. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).
11. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28,** 817–825 (2010).
12. Ernst, J. & Kellis, M. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res.* **23,** 1142–1154 (2013).
13. Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471,** 480–485 (2011).
14. Göke, J. *et al.* Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. *PLoS Comput. Biol.* **7,** e1002304 (2011).
15. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci.* **111,** 996–1001 (2014).
16. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

17. Lin, C., Garruss, A. S., Luo, Z., Guo, F. & Shilatifard, A. The RNA Pol II elongation factor Ell3 marks enhancers in ES cells and primes future gene activation. *Cell* **152,** 144–156 (2013).

18. Harbers, K., Schnieke, A., Stuhlmann, H., Jähner, D. & Jaenisch, R. DNA methylation and gene expression: endogenous retroviral genome becomes infectious after molecular cloning. *Proc. Natl. Acad. Sci. U. S. A.* **78,** 7609–7613 (1981).

19. Vidal, M. Role of polycomb proteins Ring1A and Ring1B in the epigenetic regulation of gene expression. *Int. J. Dev. Biol.* **53,** 355–370 (2009).

20. Hisada, K. *et al.* RYBP represses endogenous retroviruses and preimplantation- and germ line-specific genes in mouse embryonic stem cells. *Mol. Cell. Biol.* **32,** 1139–1149 (2012).

21. Juliane Perner, Julia Lasserre, Sarah Kinkley, Martin Vingron & Ho-Ryun Chung. Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling. *Under review.*

22. Jensen, L. J. *et al.* STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37,** D412–416 (2009).

23. Brookes, E. *et al.* Polycomb Associates Genome-wide with a Specific RNA Polymerase II Variant, and Regulates Metabolic Genes in ESCs. *Cell Stem Cell* **10,** 157–170 (2012).

24. Vella, P. *et al.* Tet Proteins Connect the O-Linked N-acetylglucosamine Transferase Ogt to Chromatin in Embryonic Stem Cells. *Mol. Cell* **49,** 645–656 (2013).

25. Leung, D. C. *et al.* Lysine methyltransferase G9a is required for de novo DNA methylation and the establishment, but not the maintenance, of proviral silencing. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 5718–5723 (2011).

26. Spruijt, C. G. *et al.* Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized

Derivatives. *Cell* **152,** 1146–1159 (2013).

27. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14,** R119 (2013).

28. Wu, H. *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473,** 389–393 (2011).

29. Song, C.-X. *et al.* Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. *Cell* **153,** 678–691 (2013).

30. Baubec, T., Ivánek, R., Lienert, F. & Schübeler, D. Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family. *Cell* **153,** 480–492 (2013).

31. Hashimoto, H. *et al.* Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res.* **40,** 4841–4849 (2012).

32. Bhutani, N., Burns, D. M. & Blau, H. M. DNA Demethylation Dynamics. *Cell* **146,** 866–872 (2011).

33. De Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14,** 249–261 (2013).

34. Powell, S. *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42,** D231–239 (2014).

35. Shannon, C. E. A mathematical theory of communications. *Bell Syst. Tech. J.* **27,** 379–423 (1948).

36. JAYNES, E. T. Information theory and statistical mechanics. **106,** 620–630 (1957).

37. Juan, D., Pazos, F. & Valencia, A. High-confidence prediction of global interactomes

based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 934–939 (2008).

38. Hakimi, M.-A. *et al.* A chromatin remodelling complex that loads cohesin onto human chromosomes. *Nature* **418,** 994–998 (2002).

39. Losada, A. Cohesin in cancer: chromosome segregation and beyond. *Nat. Rev. Cancer* **14,** 389–393 (2014).

40. Jahnke, P. *et al.* The Cohesin loading factor NIPBL recruits histone deacetylases to mediate local chromatin modifications. *Nucleic Acids Res.* **36,** 6450–6458 (2008).

41. Sarraf, S. A. & Stancheva, I. Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. *Mol. Cell* **15,** 595–605 (2004).

42. Lyst, M. J., Nan, X. & Stancheva, I. Regulation of MBD1-mediated transcriptional repression by SUMO and PIAS proteins. *EMBO J.* **25,** 5317–5328 (2006).

43. You, A., Tong, J. K., Grozinger, C. M. & Schreiber, S. L. CoREST is an integral component of the CoREST- human histone deacetylase complex. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 1454–1458 (2001).

44. Bartke, T. *et al.* Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143,** 470–484 (2010).

45. Gearhart, M. D., Corcoran, C. M., Wamstad, J. A. & Bardwell, V. J. Polycomb group and SCF ubiquitin ligases are found in a novel BCOR complex that is recruited to BCL6 targets. *Mol. Cell. Biol.* **26,** 6880–6889 (2006).

46. Sánchez, C. *et al.* Proteomics analysis of Ring1B/Rnf2 interactors identifies a novel complex with the Fbxl10/Jhdm1B histone demethylase and the Bcl6 interacting corepressor.

Mol. Cell. Proteomics MCP **6,** 820–834 (2007).

47. Gao, R. *et al.* Depletion of histone demethylase KDM2A inhibited cell proliferation of stem cells from apical papilla by de-repression of p15INK4B and p27Kip1. *Mol. Cell. Biochem.* **379,** 115–122 (2013).

48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

49. Fort, A. *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat. Genet.* **46,** 558–566 (2014).

50. Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature* **504,** 306–310 (2013).

51. Nenadic, O. & Greenacre, M. Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. *J. Stat. Softw.* **20,** 1–13 (2007).

52. Opgen-Rhein, R. & Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.* **1,** 37 (2007).

53. Ruan, J. *et al.* TreeFam: 2008 Update. *Nucleic Acids Res.* **36,** D735–D740 (2008).

54. Juan, D., Rico, D., Marques-Bonet, T., Fernández-Capetillo, O. & Valencia, A. Late-replicating CNVs as a source of new genes. *Biol. Open* **2,** 1402–1411 (2013).

55. Besag, J. EFFICIENCY OF PSEUDO-LIKELIHOOD ESTIMATION FOR SIMPLE GAUSSIAN FIELDS. *BIOMETRIKA* **64,** 616–618 (1977).

56. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E* **87,** 012707 (2013).

57. Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *BIOINFORMATICS* **24,** 333–340 (2008).

Figure legends

Figure 1. Chromatin states in ESCs

(a) Heatmaps with the emission probabilities of core epigenomic features (left) and fold change enrichments of genomic features (right) in the 20 chromatin states model.

(b-d) Enrichment of CrPs belonging to proteins tightly associated to Mediator (b), Polycomb (c) and Cohesin (d) complexes in particular chromatins states (16, 18 and 20) are indicated by red points. Members of every complex are highlighted in colored boxes. Boxplots indicate the distribution of each CrP in all 20 states. The solid horizontal line indicates the mean enrichment of all 58 CrPs and dashed lines indicate the corresponding standard deviation.

(e) Partial correlation analysis of chromatin states (see legend).

Figure 2. A network of interactions between chromatin modifiers, histone marks and cytosine methylations

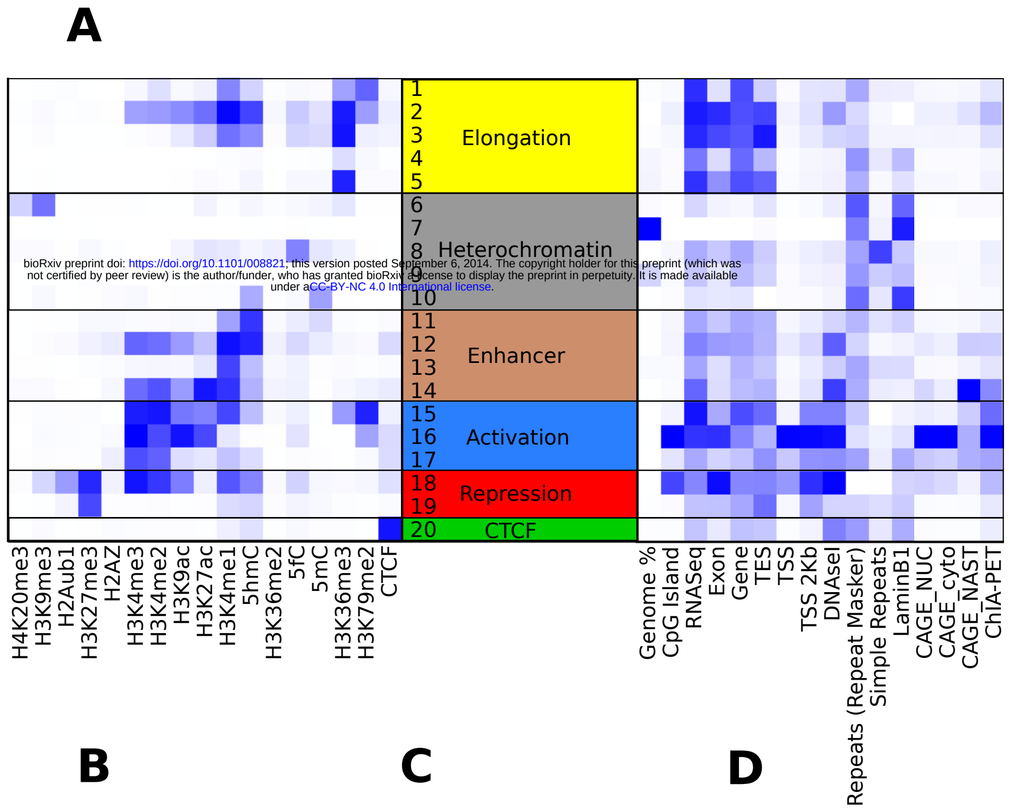
(a) Network of positive and negative interactions among epigenomic features. For visualization purposes, we select only the most reliable interactions within each state based on the coefficients of the Elastic Net and the SPCN (see methods). The thickness of the edges represents how well the participating proteins can be predicted in the Elastic Net measure by R^2. The summary figure only shows the maximal R^2 over all states. The color gradient on the edges indicates in how many states an interaction is observed (from orange=few over yellow=half to blue=all). Dashed lines indicate negative interactions (mutual exclusion). To see the resulting interaction networks per specific chromatin states, or combinations of chromatin states, visit EpiStemNet at http://dogcaesar.github.io/epistemnet

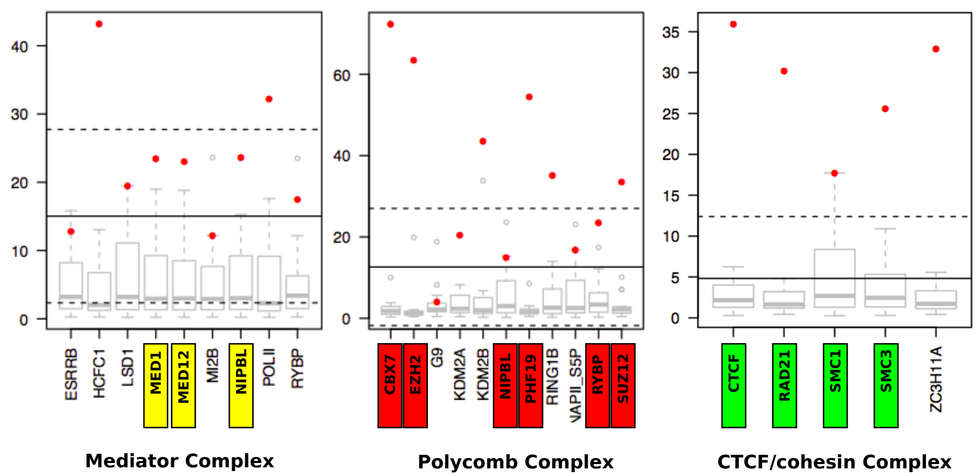
(b) Schematic representation of several complexes in the co-location network. For the sake of

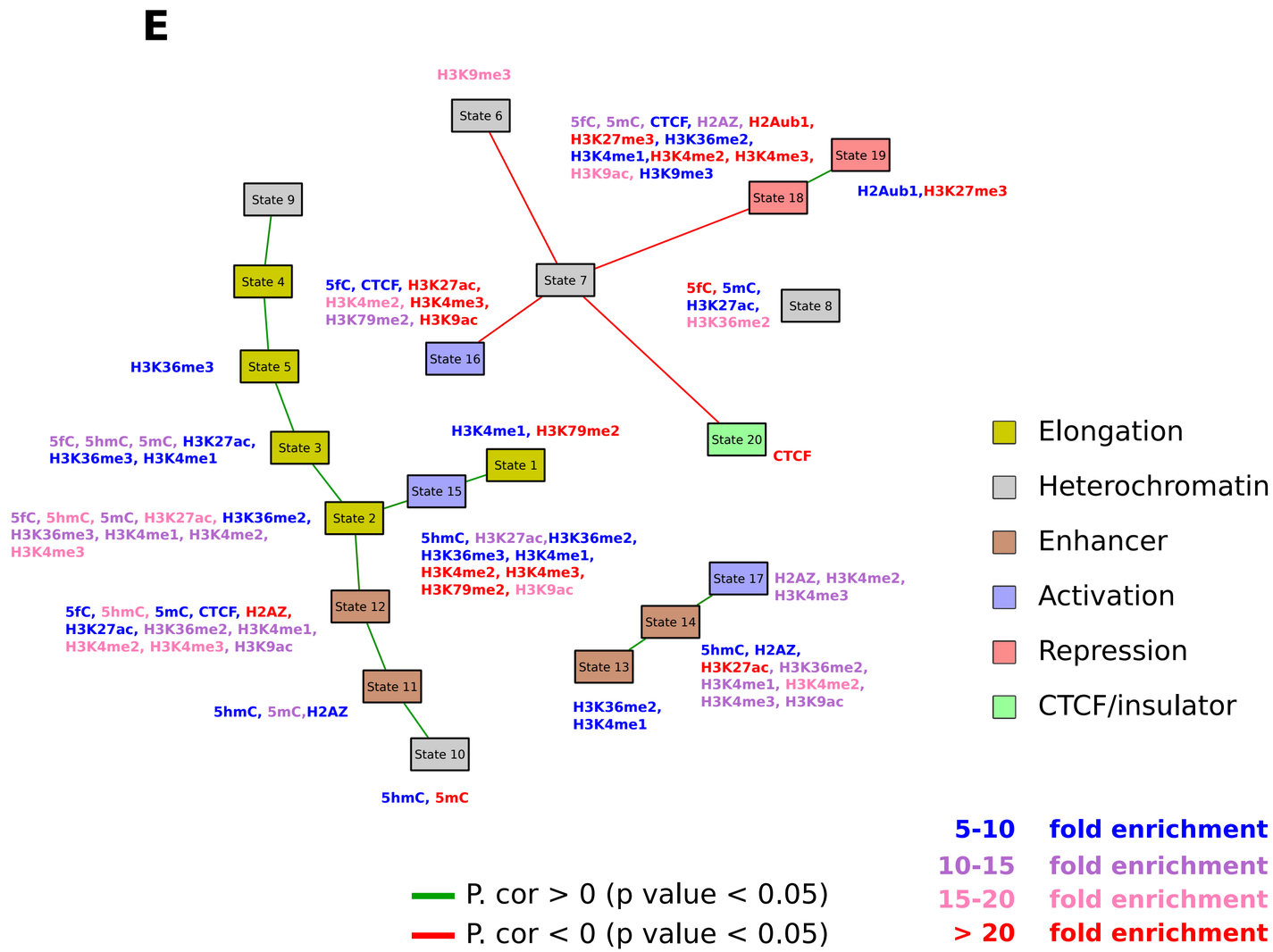
clarity, different representation is depicted for each complex, the region occupied by the complex is green shaded and nodes corresponding to the complex members are colored: NANOG/SOX2/OCT4/TCF3 (pink), Polycomb-PRC1/2 (red), Mediator (yellow), RNA polymerase II initiation/poised/elongation (orange), Rest/Co-Rest (magenta), TET1/OGT/SIN3A /OGT/SIN3A (light orange), NuRD (cyan), CTCF/Cohesin complex (green).

Figure 3. Co-location subnetwork of cytosine modifications and their direct interactors.
Subnetwork extracted from Figure 2 with the three cytosine modifications and their direct interactors. The connections relevant in each one of the 20 chromatin states are shown with different colors (see legend). Dashed lines indicate negative interactions.

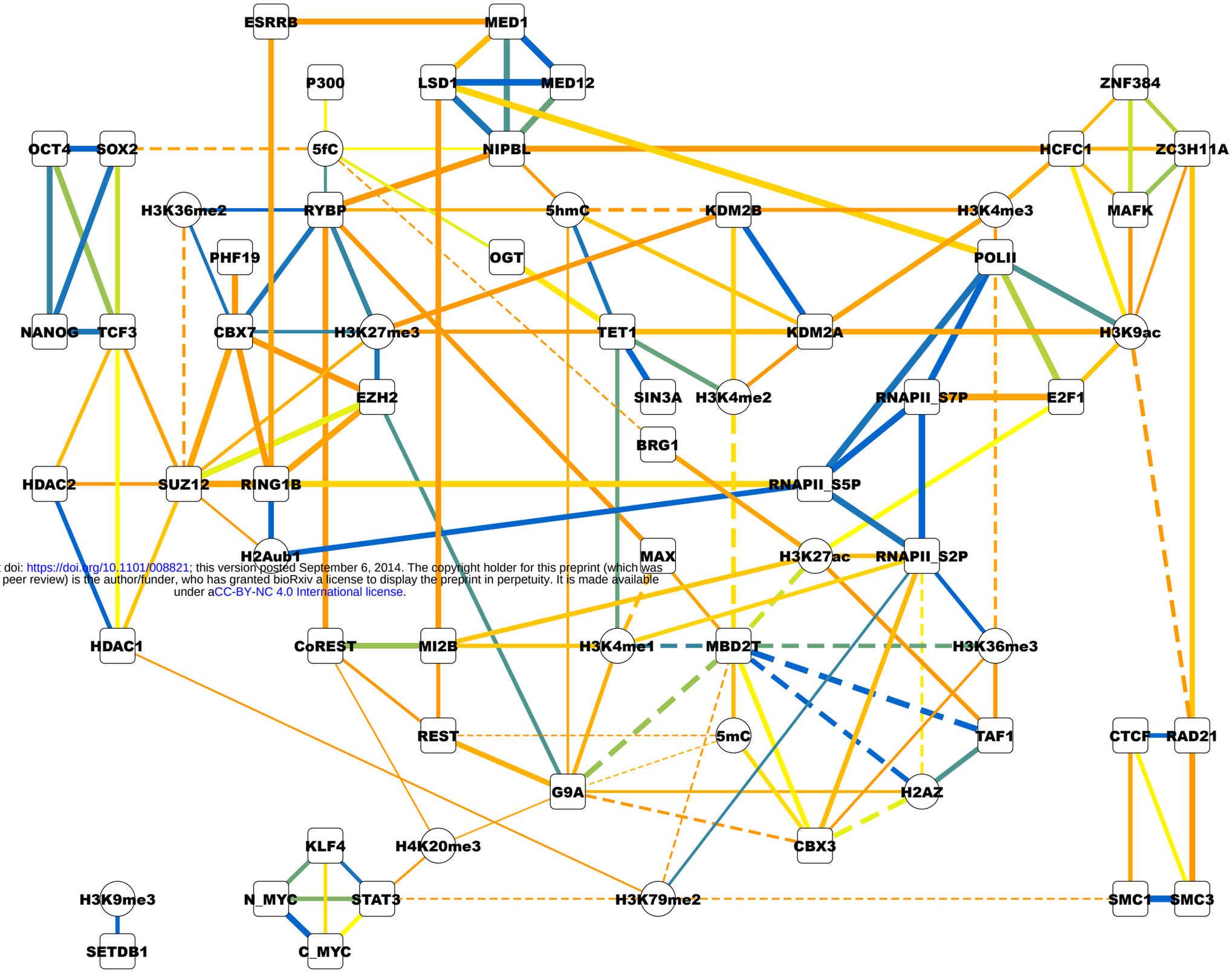
Figure 4. A co-evolutionary network of chromatin related proteins.
(a) Co-evolutionary coupled proteins are connected by black lines, with thicker lines for stronger couplings. Proteins belonging to different epigenetic complexes are represented in colored boxes: CTCF/Cohesin complex (green box), Polycomb-PRC1/2 (red), NANOG/SOX2/OCT4/TCF3 (pink), Mediator (yellow), TET1/OGT/SIN3A (light orange), NuRD (cyan), RNA polymerase II initiation/poised/elongation (orange), Setdb1 (grey).
(b) Subnetwork of co-evolving proteins (white nodes) that co-locate with cytosine modifications (green nodes). Red lines indicate pairs of co-evolving proteins. Black lines indicate co-location.



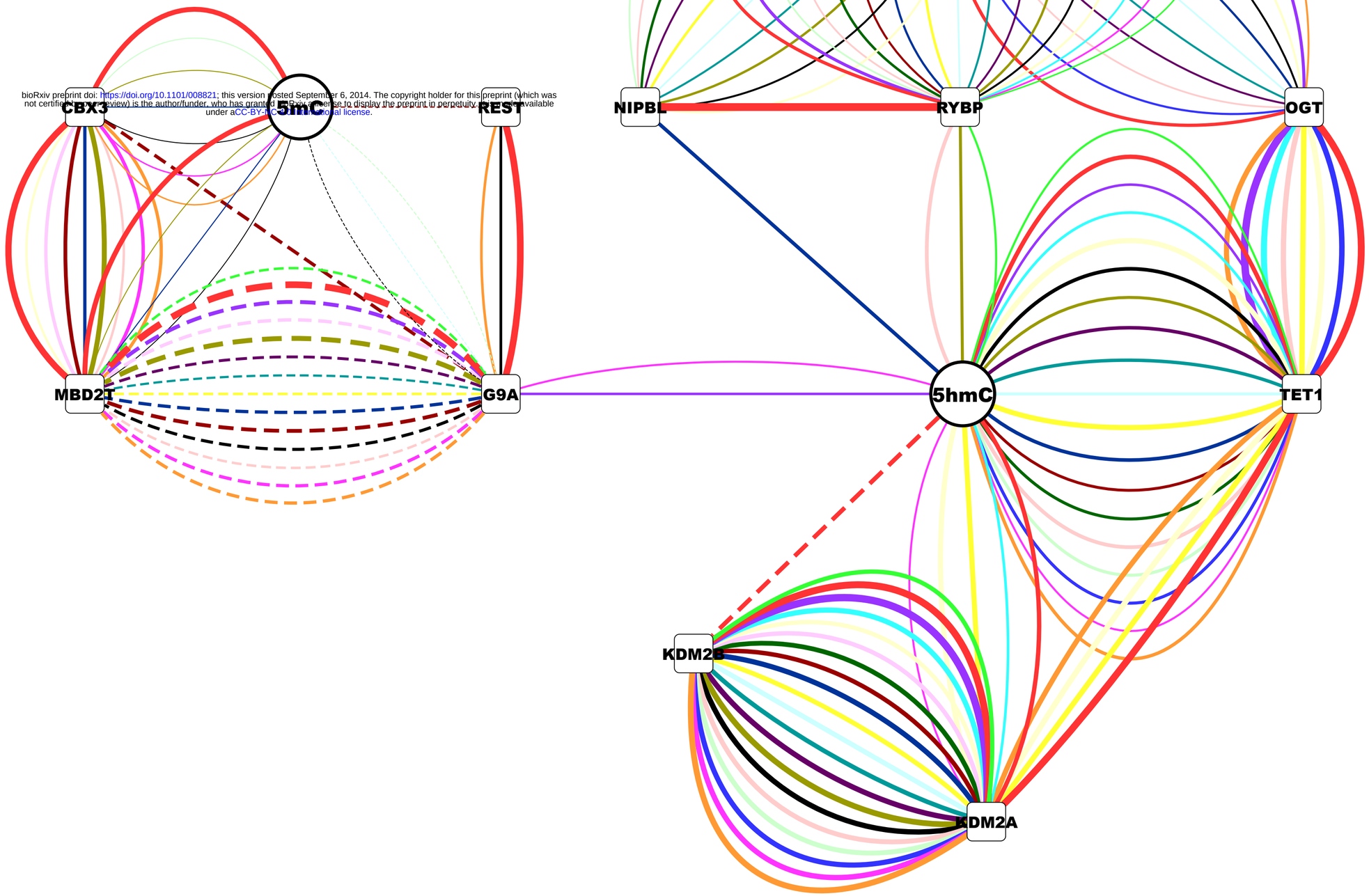


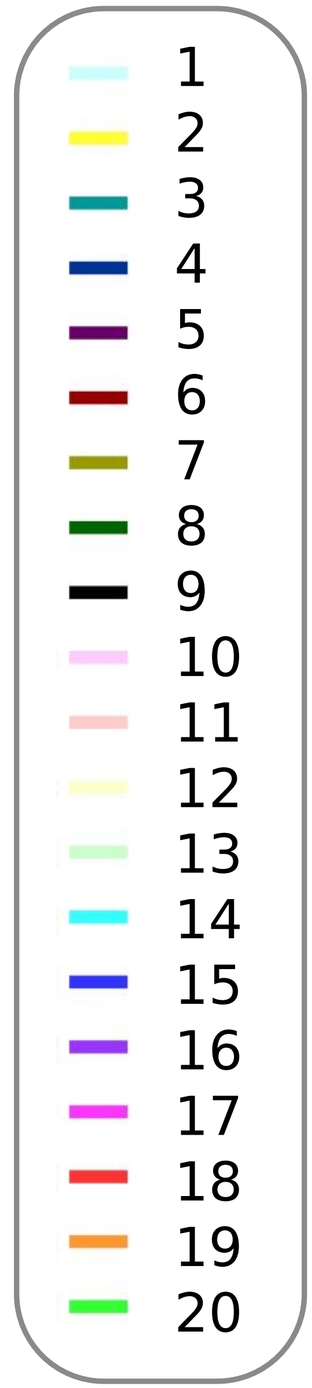


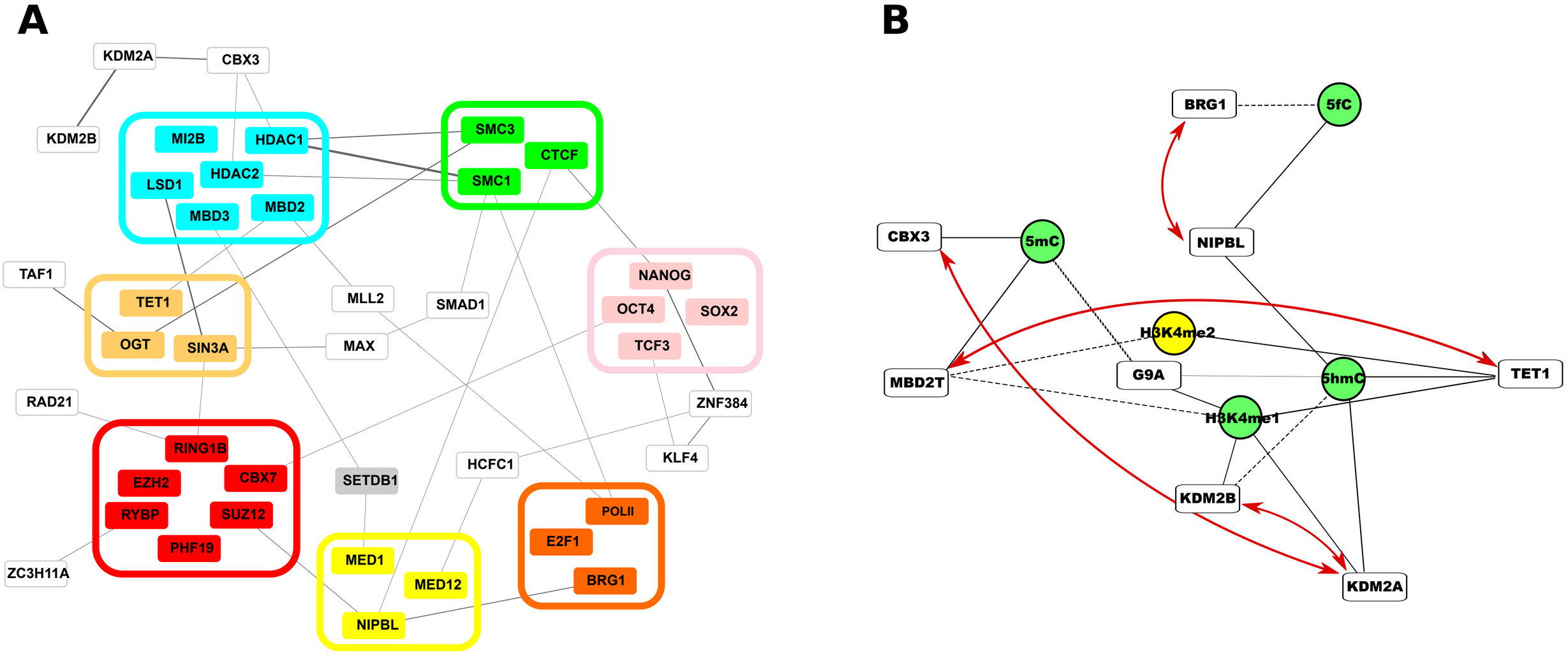
A



B






A
KDM2A
CBX3
KDM2B
MI2B
HDAC1
HDAC2
LSD1
MBD3
MBD2
SMC3
CTCF
SMC1
TAF1
TET1
OGT
SIN3A
MLL2
SMAD1
MAX
NANOG
OCT4
SOX2
TCF3
RAD21
RING1B
EZH2
CBX7
RYBP
SUZ12
PHF19
ZC3H11A
SETDB1
HCFC1
ZNF384
KLF4
POLII
E2F1
BRG1
MED1
MED12
NIPBL
B
BRG1
5fC
CBX3
5mC
NIPBL
H3K4me2
MBD2T
G9A
5hmC
TET1
H3K4me1
KDM2B
KDM2A