Epigenomic co-localization and co-evolution reveal a key role for 5hmC as a communication hub in the chromatin network of ESCs

Network approaches to decipher the epigenetic communication of embryonic stem cells

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

[1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Center - CNIO, Melchor Fernandez Almagro 3, 28029 Madrid, Spain.
[2]Computational Molecular Biology, Max Planck 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 Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany.

*These authors contributed equally to this work.
[$]Corresponding authors: valencia@cnio.es, drico@cnio.es

Abstract

Epigenetic communication through histone and cytosine modifications is essential for gene regulation and in defining cell identity. Among the possible cytosine modifications, 5-hydroxymethylcytosine (5hmC) has been related with the pluripotent status of ESCs, although its precise functional role remains unclear. To fully understand the functional role of epigenetic modifications, it is necessary to analyze the whole chromatin network. Here, we propose a framework that is based on a communication model in which histone and cytosine modifications are considered epigenetic signals, while chromatin-associated proteins (CrPs) can act as emitters or receivers of these signals. We inferred the epigenetic communication network of mouse ESCs from genome-wide location data (77 different epigenomic features) combined with extensive manual annotation of epigenetic emitters and receivers based on the literature. Notably, 5hmC represents the most central hub of this network, connecting DNA demethylation to most of the nucleosome remodeling complexes and to several key transcription factors of pluripotency. An evolutionary analysis of the network revealed that most co-evolving CrP pairs are connected by 5hmC. Further analysis of the genomic regions marked with 5hmC and bound by specific interactors (ESRRB, LSD1, TET1 and OGT) shows that each interaction points to different aspects of chromatin remodeling, cell stemness, differentiation and metabolism. Taken together, our results highlight the essential role of cytosine modifications in the epigenetic communication of ESCs.

Introduction

Intracellular and intercellular communication between proteins and/or other elements in the cell is essential for homeostasis and to respond to stimuli. Communication may originate through multiple sources and it can be propagated through different compartments, including the cell membrane, the cytoplasm, the nuclear envelope or chromatin. Indeed, a cell's identity is defined by complex communication networks, involving chemical processes that ultimately modify the DNA, histones and other chromatin proteins ("epigenomic remodeling").

It has been proposed that multiple histone modifications confer stability, robustness and adaptability to the chromatin signaling network (Schreiber & Bernstein, 2002). In fact, it is now clear that the combination of different histone marks defines the epigenomic scaffolds that affect the binding and function of other epigenetic elements (e.g., different protein complexes). The increasing interest in characterizing the epigenomic network of many biological systems has led to an impressive accumulation of genome-wide experimental data from distinct cell types. This accumulation of experimental data has meant that the first chromatin signaling co-localization networks of histone marks and chromatin remodelers could be inferred in the fly (van Bemmel *et al*, 2013) and at promoters in human (Perner *et al*, 2014). However, we are still far from understanding the epigenomic "syntax" and how different chromatin components communicate with each other to control biological processes. In addition, a variety of cytosine modifications with possible regulatory roles have emerged as potentially important pieces of this 'chromatin puzzle', such as 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC: Ficz *et al*, 2011; Pastor *et al*, 2011; Williams *et al*, 2011; He *et al*, 2011; Ito *et al*, 2011; Raiber *et al*, 2012). However, the biological function and the role of these modifications in epigenetic signaling is not yet clear (Pfeifer *et al*, 2013; Liyanage *et al*, 2014; Moen *et al*, 2015). Moreover, we still do not understanding how these and the other elements involved in epigenomic communication shape the functional landscape of mammalian genomes.

It has been proposed that evolution can be used to discern the basis of meaningful communication in animals (Smith & Harper, 2003). The continuous adaptation of

living organisms to different scenarios requires a fine-tuning of molecular communication. As a consequence, the conservation of communication pathways is often challenged by ever-changing selection pressures. Recent research pointed to protein co-evolution as a source of change in systems where the ability to interact with the environment and adapt are essential for fitness (de Juan *et al*, 2013). In addition, communication frequently occurs among mutualistic and symbiotic species, as the evolution of communicative strategies requires co-adaptation between signal production/emission and signal reception/interpretation (Smith & Harper, 2003; Scott-Phillips, 2008). At a molecular level, long-standing protein co-evolution can be reliably detected through directly correlated evolutionary histories. In fact, co-evolutionary analysis has successfully identified molecular interactions at different levels of detail (de Juan *et al*, 2013). As a consequence, protein co-evolution signatures would be expected to clearly reflect essential communicative interactions that have been frequently challenged by fluctuating evolutionary pressures.

Here, we establish a new framework to rationalize and study epigenomic communication. This framework combines network-based analyses and an evolutionary characterization of the interactions of chromatin components derived from high-throughput data and literature mining. In particular, we followed a systems biology approach to investigate the functional interdependence between chromatin components in mouse embryonic stem cells (ESCs), whereby changes to their epigenome control a very broad range of alternative cell differentiation options and they are essential for lineage specification. We constructed the epigenetic signaling network of ESCs as a combination of high-quality genomic co-localization networks of 77 different epigenomic features: cytosine modifications, histone marks and chromatin-related proteins (CrPs) extracted from a total of 139 ChIP-seq experiments. We labeled histone marks and cytosine modifications as *signals* and we classified the proteins that co-localize with them as their *emitters* (writers or erasers) or *receivers* (readers) based on information in the literature **(Figure 1)**. To our knowledge the resulting communication network is the most complete global model of epigenetic signaling currently available and therefore, we propose it to be a valuable tool to understand such processes in ESCs.

By analyzing this network, we found 5hmC to be a key node that mediates communication between different regions of the network. In addition, our co-evolutionary **analysis of this network identified 5hmC as a central node that connects most co-evolving CrPs**. Exploration of 5hmC-centered communication revealed that specific co-localization of 5hmC with the TET1, OGT, ESRRB and LSD1 produces alternative partner-specific activity, such as chromatin remodeling, cell stemness and differentiation, and energy metabolism. Thus, we propose that 5hmC acts as a central signal in ESCs for the self-regulation of epigenetic communication.

Results

Inference of the chromatin signaling network in mouse ESCs

We built an epigenetic signaling network in mESCs through a two-step process. First, we inferred the network connectivity based-on co-localization in the genome-wide distribution of chromatin components. In this analysis, we included 139 ChIP-Seq, MEDIP and GLIB assays for 77 epigenetic features (3 cytosine modifications, 13 histone marks and 61 CrPs: **Supplementary Table 1)**. Accordingly, we employed a method described recently (Perner *et al*, 2014) that reveals direct co-dependence between factors that cannot be "explained" by any other (indirect) factor observed. Thus, we detected only relevant interactions in different functional chromatin domains (see **Methods** for details).

Second, we annotated the direction of the interactions in the network (as shown in **Figure 1**), for which we relied on classifying a CrP as an emitter based-on previously reported experimental evidence of its specific ability to write or erase an epigenetic signal (either a histone mark or a cytosine modification, **Suppl. Table 2**). This evidence can be roughly summarized within two possible scenarios: (1) Protein A is a known *writer* or *eraser* of signal B; (2) Alterations to the genome-wide distribution of protein A (e.g., through its knock-out) affect the distribution of signal B in the genome. In the absence of any such evidence, proteins were defined as receivers of the interacting signal.

This epigenetic communication network (**Figure 2A**) recovered 236 connections between 68 nodes, the latter represented by cytosine modifications, histone marks or CrPs. The network contains 192 positive interactions (simultaneous interactions, 81.4%) and 44 negative (mutually exclusive interactions, 18.6%). A web interactive browser of the global co-localization network enables users to explore the interactions among these chromatin components in more detail (see http://dogcaesar.github.io/epistemnet).

Our approach detected 115 direct CrP-CrP interactions that are mostly due to protein complexes given that these components coincide at chromatin. These include complexes such as Polycomb (RYBP/CBX7/PHF19/SUZ12/EZH2), Cohesin (RAD21/SMC1/SMC3), Mediator (MED1/MED12/NIPBL), the nucleosome remodeling deacetylase MI2/NuRD complex (MI2B/LSD1/HDAC1/HDAC2) and CoREST/Rest (Rest/CoREST/RYBP: **Figure 2A**).

In order to understand the epigenetic interaction network and its activity as a communication system, we focused our analyses on directional interactions: *emitter-signal* and *signal-receiver* associations. Based on the experimental information extracted from the literature, we established “communication arrows" from “emitter-CrPs” to their signals and from the signals to their epigenetic "receiver-CrPs". In general, we could establish 124 (52.5%) directional interactions involving an epigenetic emitter and a signal (56 edges), or a receiver and a signal (68 edges), and as a consequence, we identified 8 emitter-CrPs, 17 receiver-CrPs and 18 CrP nodes that can act simultaneously as emitters and receivers of different signals.

The hubs of a network are highly connected nodes that facilitate the networking of multiple components. Directional edges allowed us to distinguish between two types of hubs: in-hubs (nodes with a large number of incoming arrows) and out-hubs (with a large number of outgoing arrows). Not surprisingly, the main in-hub was RNA polymerase II with S2 phosphorylation of the C-terminal (RNAPII_S2P). Indeed, 9 out of 16 signals in the network pointed to this form of RNAPII, which is involved in transcriptional elongation and splicing (**Suppl. Figure 1**). Here, the strong in-hub

nature of RNAPII_S2P in the network coincided with the many different signals that independently contribute to transcription and expression in the genome.

By contrast, the two main out-hubs in the network revealed a different aspect of epigenetic regulation. The main hubs that accumulated connections with receivers were H3K79me2 (12) and 5hmC (10: **Figure 2B, Suppl. Figure 2**). H3K79me2 is involved in transcription initiation and elongation, as well as promoter and enhancer activity, suggesting that it is a key signal for different aspects of transcriptional regulation. Interestingly, two groups of transcription factors (TFs) were connected to H3K79me2: one composed of TCF3, OCT4, SOX2 and NANOG; and another that contains CMYC, NMYC, STAT3, KLF4, TCFCP2L1 and E2F1. Conversely, 5hmC is particularly interesting as it is thought to be a key element in different processes even though its role in gene regulation remains controversial (Pfeifer *et al*, 2013; Liyanage *et al*, 2014). Whereas initially related to gene activation (Song *et al*, 2011), others claimed that 5hmC associates with weakly expressing poised promoters (Pastor *et al*, 2011; Williams *et al*, 2011), while both roles were elsewhere claimed to be possible depending on the context (Wu *et al*, 2011). In addition, 5hmC was shown to play a major role in enhancer activation (Stroud *et al*, 2011; Szulwach *et al*, 2011) or silencing (Choi *et al*, 2014). This apparent controversy could be explained by the role of 5hmC as a central node of the communication network. Indeed, 5hmC was the node that is traversed by the highest number of paths between nodes (**Suppl. Figures 3 and 4**), which implied that this node concentrates the information flow of the mESC network.

Co-evolution among chromatin components

Cell stemness evolved very early in metazoan evolution and it is a critical phenomenon that enhances the viability of multicellular animals (Hemmrich *et al*, 2012). Thus, it can be assumed that CrP-mediated communication in stem cells has also been essential for metazoan evolution. As co-evolution consistently reflects important functional interactions among conserved proteins (de Juan *et al*, 2013), we studied the signatures of protein co-evolution within the context of the epigenetic communication network in stem cells. We focused our analysis on the CrPs in the network for which there is sufficient sequence and phylogenetic information in order

to perform a reliable analysis of co-evolution (de Juan *et al*, 2013). We extracted evolutionary trees for 59 orthologous CrPs in our epigenetic communication network and calculated their degree of co-evolution. To disentangle the direct and uninformative indirect evolutionary correlations, we developed a method that recovers protein evolutionary partners based on a maximum-entropy model of pairwise interacting proteins (see Methods for specific details of the implementation).

Using this approach, we retrieved 34 significant co-evolutionary interactions among 54 CrPs (see **Supplementary table 3**). A total of 27 co-evolved relationships were identified based on the direct functional protein-protein interactions evident in prior experimental data: external sources, indirect evidence in the literature and/or from our communication network (see **Supplementary Table 3**). These co-evolutionary associations reflected the evolutionary relevance of different epigenetic communication pathways that might be at play in essential, evolutionary maintained cell types like ESCs.

We identified epigenetic signals that connect CrPs related by co-evolution (i.e.: those connecting co-evolving pairs) and we considered the historically influential signals as those that were best connected in a co-evolutionary filtered network. This co-evolutionary filtered network was obtained by maintaining the pairs of CrPs that both co-evolve and that are included in a protein/signal/protein triplet (see **Figure 3**). Co-evolving CrP pairs are not evenly distributed in the epigenetic communication network but rather, we found a statistically significant correspondence between signal-mediated communication and co-evolution for H3K4me2, H3K4me3 and 5hmC (p-value < 0.05, see Methods). Of these, 5hmC mediates communication between four different co-evolving pairs and seven different CrPs (**Figure 3**), clearly standing out as the epigenetic signal connecting more co-evolving CrPs. Notably, the three positively co-occurring emitters of 5hmC (TET1, OGT and LSD1) co-evolved with three different receivers (MBD2, TAF1 and SIN3A). Thus, from the combination of the 5hmC interactors (see **Fig. 2C**), three specific emitter/signal/receiver triplets with coordinated evolution were identified: LSD1-5hmC-SIN3A, TET1-5hmC-MBD2 and OGT-5hmC-TAF1. In other words, co-evolutionary signals reflected very important interactions due to the multiple

241 connections that are possible in the network. In addition, we detected co-evolution
242 between the 5fC-emitter BRG1 and the 5fC-receiver NIPBL.
243
244 The case of MBD2 and TET1 is particularly interesting given the biological activities
245 of these proteins. One of the key functions of TET1 is the oxidation of 5mC, while
246 MBD2 is a methyl-binding domain protein (MBD) that shows higher binding affinity
247 to 5mC than to 5hmC (Baubec *et al*, 2013). In addition, MBDs are thought to
248 modulate 5hmC levels, inhibiting TET1 by their binding to 5mC (Hashimoto *et al*,
249 2012). The co-evolution of MBD2 and TET1 suggests certain dependence between
250 the mechanisms that maintain 5mC and 5hmC at different epigenomic locations in
251 ESCs.
252
253 The well-known TET1 interactors OGT and SIN3A each co-evolved with a different
254 CrP: TAF1 and LSD1, respectively. OGT co-occurs with 5hmC while TAF1 binding
255 is significantly enriched in 5hmC depleted regions. Similarly, LSD1 positively
256 interacts with 5hmC while its co-evolving partner SIN3A was found in a pattern that
257 is mutually exclusive to 5hmC. As in the case of TET1 and MBD2, these results
258 suggest the remarkable influence of 5hmC on the differential binding of CrPs to
259 distinct genomic regions in the ESC epigenome during metazoan evolution.
260
261 Accordingly, these results confirmed our working hypothesis that chromatin proteins
262 interconnected via epigenetic signals have evolved in a concerted manner.
263 Interestingly, our results also suggest that 5hmC is a communication hub as it
264 connects processes that have been coordinated during metazoan evolution.
265

266 **Functional modularization of the network reveals protein complexes and**
267 **star-shaped structures**

268 Having shown that 5hmC and H3K79me2 are the most influential signals in the ESC
269 epigenetic communication network, and that 5hmC mediates the communication
270 between CrPs that have co-evolved in Metazoa, recent research has shown that the
271 genomic localization of certain combinations of core epigenetic features allows
272 different chromatin states associated with functional processes to be reliably
273 identified (Filion *et al*, 2010; Ernst & Kellis, 2010). Here, we examined how the

positive interactions in the network are distributed in relation to these different functional contexts. In particular, we focused on the modules of co-localizing chromatin components with similar peak frequencies that were associated with the diverse chromatin states in ESCs (see Methods and **Supplementary Fig. 5**).

We found 15 groups of interactions that yielded sub-networks associated with distinctive functional chromatin profiles (**Figure 4**). These **chrom**atin context-specific **net**works (*chromnets*) were made up of CrPs and epigenetic signals that tended to co-exist in the different chromatin states at a similar frequency in ESCs. We found that most chromnets could be classified into two groups: protein complexes and communication chromnets. Specific examples of protein complexes chromnets were Polycomb (CBX7/PHF19/SUZ12/EZH2) in chromnet-5, Cohesin (RAD21/SMC1/SMC3) in chromnet-10 or Mediator (MED1/MED12/NIPBL) in chromnet-11 (**Figure 4A and Supp. Figs 6-20**). These chromnets had high clustering coefficients and a high proportion of CrP-CrP interactions, and their frequency in different chromatin states was coherent with their known function. For example, chromnet-5 (Polycomb) was strongly enriched in the two chromatin states enriched in H3K27me3 (**Supp. Figure 10**).

We also noted the presence of star-like chromnets with very low clustering coefficients. These star-like chromnets are mostly generated by emitter/signal and signal/receiver interactions, suggesting that these are communication modules that connect different protein complexes. For example, chromnet-3 contains two central connectors (5fC and RYBP) connecting Polycomb, Mediator and TET1-SIN3A complexes, and this chromnet is enriched in active transcription states and regulatory elements (**Suppl. Figure 8**).

Interestingly, chromnet-2 was a star-like module centered on 5hmC (the most central hub in the network) and it contained all its positively co-localizing interactors: LSD1, RYBP, ESRRB, KDM2A, TET1, OGT, G9A, and MBD2T (**Figure 4B**). In addition, 5hmC indirectly connects to H3K4me1 via TET1, and with 5mC via MBD2T. This chromnet was clearly enriched in regulatory elements.

In summary, we have decomposed the communication network into communication chromnets, functional modules of interactions with similar frequencies in the different chromatin contexts. The components, structure and genomic distribution of these chromnets provided information about their functional role. In particular, we detected several star-like chromnets that are important to distribute epigenetic information to different regions of the communication network. The wide range of functional chromatin states that were enriched in these chromnets further supports their potential role in mediating communication between distinct processes.

Independent co-localization of 5hmC with ESRRB, LSD1, OGT and TET1 was associated with different biological activities

Having identified 5hmC as an important communication signal in extant mouse ESCs and during metazoan evolution, we also found that the eight positive interactions (co-existence) between 5hmC and CrPs are part of a star-like chromnet with similar enrichment associated with chromatin states. We further characterized the genomic regions where 5hmC co-localized independently with the stemness factor ESRRB and with the three independent emitters of 5hmC, LSD1, OGT and TET1, which were also identified in our co-evolutionary analysis (see above).

Remarkably, we found 6,307 genomic regions where 5hmC co-localized with its receiver ESRRB in the absence of TET1, and with the rest of its interactors (**Figure 5A**). ESRRB is a transcription factor that is essential for the maintenance of ESCs (Papp & Plath, 2012; Zwaka, 2012), yet to our knowledge the binding of ESRRB to DNA has not been previously associated with the presence of 5hmC. However, the ESRRB gene locus is known to be strongly enriched in 5hmC in ESCs (Doege *et al*, 2012), suggesting that 5hmC and ESRRB form a regulatory loop. **Gene ontology analysis** carried out with the genes closest to these specific regions (McLean *et al*, 2010) identified stem cell maintenance, MAPK and Notch cell signaling cascades as the most enriched functions **(Figure 5E)**, highlighting the importance of ESRRB for stemness maintenance. Surprisingly, the expression of the ESRRB gene is not ESC-specific but rather it is expressed ubiquitously in most differentiated cell types (Zwaka, 2012). Thus, its specific role in stemness probably requires ESC-specific interactions with other components of the communication network and our results

suggested that 5hmC might be the key signal connecting ESRRB function with stemness.

LSD1 is a H3K4- and H3K9-demethylase that can act as either a transcriptional co-activator or co-repressor (Wang *et al*, 2007). To our knowledge, this was the first time 5hmC and LSD1 were found to coincide in the epigenome of ESCs (**Figure 5B**). Interestingly, it is well known that there is a functional co-dependence between histone demethylation and DNA methylation (Vaissière *et al*, 2008; Ikegami *et al*, 2009). Indeed, we consider LSD1 is an emitter of 5hmC because there is a global loss of DNA methylation in the LSD1 knockout (Wang *et al*, 2009, 1). Remarkably, we found that the 9,714 5hmC-**LSD1 specific** regions are significantly enriched with specific terms associated with histone acetylation and DNA modification (**Figure 5E**), strengthening the dependent relationship between histone and DNA modifications. Indeed, LSD1 not only functions as a histone demethylase by itself but also, in association with 5hmC it can regulate the expression of proteins that modify both histone acetylation and DNA methylation. These results suggest the presence of a second regulatory loop involving 5hmC.

TET1 and OGT are two of the best known emitters of 5hmC (**Figure 5C-D**), with TET1 a DNA demethylase that catalyzes the conversion of 5mC to 5hmC and OGT a regulator of TET1 (Vella *et al*, 2013; Balasubramani & Rao, 2013). In fact, the role of OGT in DNA demethylation was associated to its co-localization with TET1. However, OGT is a N-acetylglucosaminyltransferase that can also bind to different TFs independently of TET1 (Bond & Hanover, 2015). Notably, we observed different functional enrichment of the 5hmC-TET1 and 5hmC-OGT regions (**Figure 5E**). While the 27,721 5hmC-TET1 regions were enriched in stem cell maintenance and morphogenesis, highlighting the role of both 5hmC and TET1 in stemness, the 1,017 5hmC-OGT regions were related with the metabolism of glycerophospholipids and carbohydrates. Interestingly, OGT is known to bind phosphatidylinositol-3,4,5-trisphosphate, regulating insulin responses and gluconeogenesis through glycosylation of different proteins (Yang *et al*, 2008). Our results suggest that the alternative role of OGT in gene regulation is also associated to 5hmC (but not to TET1). As the presence of 5hmC requires the action of TET1, our results suggest that OGT might remain in certain locations after TET1 removal, probably associated to the presence of specific

TFs in order to regulate the metabolism of glycerophospholipids and carbohydrates. In this scenario, OGT would act as an emitter regulating 5hmC production and as a receiver by acting with other proteins in the presence of 5hmC to regulate gene expression.

In summary, the analysis of specific genomic regions revealed that different processes and functions could be regulated and may be interconnected via 5hmC interactions with other proteins. These processes include functions as relevant as epigenetic self-regulation, cell signaling, maintaining stemness, morphogenesis and metabolism.

Discussion

ESCs constitute an ideal model to explore the epigenomic communication that directly influences the phenotype of cells. Cytosine modifications, certain histone marks and CrPs contribute to the plasticity required for the induction and maintenance of pluripotency. Thus, the abundant epigenomic data from mouse ESCs has enabled us to investigate how the different chromatin components communicate with each other within a complex network. Using high-throughput genome-wide data and information from the literature, we reconstructed the epigenetic communication network of ESCs. In addition to the rigorously established co-incidence and mutual exclusion, we also annotated the directions of the CrP interactions mediated by epigenetic signals (cytosine modifications and histone marks) based on information extracted manually from the literature. This information allows CrPs to be classified as emitters or receivers of these more basic epigenetic signals.

Our results provide a framework for future studies of the chromatin network in ESCs and other cell types, and we highlight the importance of using information taken from the literature. This biological knowledge allowed us to understand the network of co-localization patterns from high-throughput data, permitting us to obtain the first global picture of the information flow that could take place in the ESC epigenome. For example, we identified the hubs that receive more independent signals - in-hubs - and those that emit signals to a larger number of receivers - out-hubs. Not surprisingly, active RNA polymerase II was identified as the main in-hub of the

406 network, which shows that our approach is able to recover biologically meaningful,
407 data, as many components of the network regulate transcription.
408
409 More surprisingly, our analysis revealed that 5hmC is the main out-hub and the most
410 central node in this network. 5hmC interacts with a total of sixteen CrPs: five emitters
411 of 5hmC and eleven receivers. Of these sixteen interactions, half are positive (co-
412 occurrence) and half are negative (mutual exclusion). The large number of CrPs that
413 preferentially bind to chromatin in the presence or absence of 5hmC indicates that this
414 cytosine modification is an influential signal for chromatin communication in ESCs.
415
416 The elements that drive epigenetic communication constitute an intricate and dynamic
417 network that produces responses that range from stable programs defining cell-
418 identity to fast cellular responses. In this context, the fine-tuning of epigenetic
419 communication pathways is likely to have been a key aspect in the evolution of
420 multicellular organisms, such as metazoans. Co-evolutionary analyses highlight the
421 conservation and co-ordinated changes in interactions, and this is a particularly
422 adequate approach to reveal strong functional links in the context of complex and
423 dynamic protein interactions. Co-evolution can occur between proteins that interact
424 directly or that participate in the same communication processes – for example, via
425 chromatin interactions mediated by histone marks or cytosine modifications.
426
427 Remarkably, the majority of the co-evolutionary associations related to epigenetic
428 communication are triplets formed by an emitter, a signal and a receiver.
429 Unexpectedly, four different co-evolutionary associations were found between
430 proteins interacting with 5hmC: SIN3A with LSD1, TET1 with MBD2, MBD2 with
431 MLL2, and OGT with TAF1. Strikingly, all three co-occurring 5hmC emitters (TET1,
432 OGT and LSD1) co-evolve with three different 5hmC receivers, forming different
433 emitter-5hmC-receiver triplets. Interestingly, these associations do not reflect direct
434 physical interactions of the protein pairs but rather, complementary roles in the
435 control of cytosine modifications and gene regulation. Thus, we speculate that the
436 balance between 5mC, 5hmC and other cytosine modifications has been very
437 important in fine-tuning epigenomic communication during the evolution of
438 metazoans.

Identifying modules in networks helps better understand their distinct components (Mitra *et al*, 2013). Here, we followed a simple approach to identify functional sub-networks of chromatin communication, or chromnets, clustering positive interactions in function of their relative frequency in different chromatin states. This analysis revealed the functional structure of the communication network and we were able to automatically recover known protein-complexes, such as Polycomb and Mediator. By contrast, we found that 5hmC and 5fC establish two different star-shaped chromnets, suggesting that they might be involved in communication between distinct epigenetic components and processes in distinct locations of the ESC epigenome.

While further experiments will be needed to reveal the functional roles of the different independent interactions of 5hmC, our results generate some interesting hypotheses about the possible independent functions played by 5hmC in ESCs. We propose that the stem-specific role of ESRRB in ESCs could be linked to its co-occurrence with 5hmC, as this cytosine modification is less common in most differentiated cell types (Zwaka, 2012). Our results also show that LSD1-5hmC might be specifically involved in the regulation of histone modifications and DNA methylation, while the TET1-5hmC interaction is associated with stem cell maintenance and morphology. Furthermore, our data suggest a TET1-independent interaction between 5hmC and OGT might participate in the regulation of energy metabolism, and an interaction between 5hmC and LSD1 regulates histones and DNA methylation.

The combination of genome-wide location data, prior knowledge from the literature and protein co-evolution highlights conserved functional relationships between 5hmC-interacting CrPs that have been dynamically coordinated during evolution. Based on our co-evolution analysis, we hypothesize that the different cytosine modifications in different regions of the genome might have been important during metazoan evolution. Our results suggest that the interaction of 5hmC with specific emitters is involved in regulating different specific and critical functions.

In conclusion, network architecture conveys relevant contextual information that cannot be easily obtained from analyses that focus on only a few epigenetic features.

The computational framework introduced here represents the basis to explore this vast space and it provides the first integrated picture of the different elements involved in epigenetic regulation. Accordingly, this analysis enables us to attain an integrated vision of epigenetic communication in ESCs that highlighted the relevance of 5hmC as a central signal. Notwithstanding, we are still at the early stages of exploring the epigenetic network. Thus, the future inclusion of experimental data regarding genome-wide localization profiles for many additional proteins, as well as that related to other states of cell differentiation, will make it possible to draw-up a more complete picture and to define the dynamics of the epigenetic network in different cell lineages.

Materials and Methods

ChIP-Seq, MeDIP and GLIB data processing

We downloaded sra files from 139 Chromatin Immunoprecipitation Sequencing (ChIP-Seq), Methylated DNA immunoprecipitation (MeDIP) and GLIB (glucosylation, periodate oxidation and biotinylation) experiments described in 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, elements of the machinery involved in cytosine and histone modification, transcription factors (TFs, such as the 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. The MeDIP data for 5mC and the GLIB data for 5hmC were taken from Pastor et al (Pastor *et al*, 2011) as it has been previously shown that these datasets are less biased to antibody affinity in regions with repeats than other methods (Matarese *et al*, 2011). The sra files were transformed into fastq files with the sra-toolkit (v2.1.12) and aligned to the reference mm9/NCBI37 genome with bwa v0.5.9-r16 (Li & Durbin, 2009) allowing 0-1 mismatches. Unique reads were converted to BED format.

Genome segmentation

The input information used to segment the genome into different chromatin states was that derived from 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* (Ernst & Kellis, 2010). A multivariate Hidden Markov Model (HMM) was employed that uses two types of information: the frequency with which different combinations of chromatin marks are found with each other, and the frequency with which different chromatin states are spatially related in the genome. To apply this method we used the ChromHmm software (Ernst & Kellis, 2012: 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 the mapped sequences (see above). First, the genome was divided in 200 bp non-overlapping segments which were independently assigned a value of 1 or 0 in function of the presence or absence of histone marks, respectively, based on the count of the tags mapping to the segment and on a Poisson background model (Ernst & Kellis, 2012) using a threshold of 10^{-4}. After establishing a binary distribution for each mark, we trained the HMM model using a fixed number of randomly-initialized hidden states that varied from 20 to 33 states. We focused on a 20-state model that provided sufficient resolution to resolve biologically meaningful chromatin patterns according to previous selection strategies (Ernst & Kellis, 2012; Kharchenko *et al*, 2011 - see Suppl. Figure 21). We used this model to compute the probability that each location is in a given chromatin state and we then assigned each 200 bp segment to its most likely state (see **Suppl. Tables 4 and 5**). Only, intervals with a probability higher than 0.95 were considered for further analysis.

We identified states related with enhancer (states 1-3), transcription elongation (states 4-5), heterochromatin (6-10), enhancers (11-14), promoter activation (15-17), Polycomb (18-19) and the CTCF insulator (20), that are consistent with prior knowledge regarding the function of these features (**Suppl. Figures 21, 22** and **Suppl. Table 6**).

Segment enrichment

The proportion of overlap for each state and annotation in the genome was computed with ChromHmm software on the selected segments (see above). The genomic annotations, CpG islands, repeats and laminB1 annotations were downloaded from the UCSC Genome Browser, and DNAseI and RNAseq were obtained from ENCODE E14 cell line (see **Suppl. Table 1**). The processed CAGE data (Fort *et al*, 2014) and ChIA-PET data (Zhang *et al*, 2013) for ESCs were downloaded from the supplementary material of the original papers (see **Suppl. Table 1**). The enrichment of the annotations and CrPs can be consulted in **Suppl. Tables 7 and 8**.

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

We used the ChromHMM segments with a probability higher than 0.95 as samples for the network inference. We filtered all bins for each state that were unexpectedly large (the upper 1% for each state) because they might produce outliers in the data and it is hard to justify where the signal occurs within the region (**Suppl. Figure 23**). We counted the overlapping ChIP-Seq reads for the resulting segments using Rsamtools, although some of the ChIP-experiments had to be excluded from the network inference due to the low number of reads per bin, or the low number of bins with signal or study dependent artifacts, including: CTCF_GSE11431, NANOG_GSE11431, LAMIN1B and H3K27me3_GSE36114, SMAD1_GSE11431, MBD1A_GSE39610, MBD1B_GSE39610, MBD2A_GSE39610, MBD3A_GSE39610, MBD4_GSE39610, and MECP2_GSE39610 (as MBD2A was not used, the MBD2 co-localization data corresponds to MBD2T). Using hierarchical clustering with 1-cor(x,y) as a distance measure, we find that most replicates or functionally related samples fall into the same branch (**Suppl. Figures 24 and 25**).

Next, the replicates were merged by adding up the read counts in each segment. The resulting 71 samples were normalized against the corresponding input using the same method described in (Perner *et al*, 2014). In short, we estimated the median fold-change of the sample over the input and used this median to shrink the change in each segment towards 1. Finally, the data was log-transformed adding a pseudocount of 1.

Co-location network inference

To detect specific interactions between the components based on their co-localization it is necessary to eliminate indirect/transitive effects, i.e.: co-localization that might be

introduced by other factors. To this end, we applied the method described in (Perner *et al*, 2014) that aims to unravel the interactions between factors that cannot be "explained" by the other observed factors and thus, this is a more specific approach than an analysis of simple pair-wise correlations.

We inferred an interaction network for each chromHMM state. Briefly, for each state the samples were scaled to have a mean of 0 and a standard deviation of 1. An Elastic Net was trained in a 10-fold Cross-validation to predict the HM/CTCF/DNA methylation of the CrPs or to predict each CrP from all other CrPs. Furthermore, the sparse partial correlation network (SPCN) was obtained using all the samples available. To visualize the final networks, we selected the interactions between Histone marks/cytosine modifications and CrPs that obtained a high coefficient (w >= 2*sd(all_w)) in the Elastic Net prediction and that have a non-zero partial correlation coefficient 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 **Suppl. Tables 9 and 10**. All partial correlation coefficients of the SPCN per state are given in **Suppl. Table 11**.

The global network with the information of all the states (**Figure 2A**) summarizes all the direct interactions between cytosine modifications, histone marks and CrPs. The global network, as well as the chromnets, can be explored using EpiStemNet, an interactive viewer of the "co-location" network (http://dogcaesar.github.io/epistemnet).

Co-evolution-based analysis

We retrieved 46,041 protein trees of sequences at the metazoan level from eggnog v4.0 (Powell *et al*, 2014), including over a million protein sequences. These trees include proteins from NS = 88 metazoan species that are either orthologs or paralogs that were duplicated after the metazoan speciation split. Based on these trees, we extracted only-unique-orthologous protein trees for each mouse protein by inferring speciation and duplication nodes using a species-tree reconciliation approach (Nenadic & Greenacre, 2007) and a previously developed pipeline to deal with tree inconsistencies (Juan *et al*, 2013). When more than one ortholog was detected for a mouse protein in a species, the one selected was that extracted from the tree with the shortest overall evolutionary distance. As a result, we obtained 13,579 only-unique-

orthologous protein trees. From these, we extracted those that included the mouse proteins for which ChIP-seq data was analyzed in this study, before performing the main analysis on NP = 58 different protein trees that include mouse CrPs. The whole population of trees was kept to perform a randomization test in order to assign empirical statistical significance to our results.

We encoded each protein tree as a vector containing the NS(NS – 1)/2 distances between all the pairs of species in the analysis, and we formed a NS(NS – 1) X NP distance matrix containing these vectors as columns. Each row in this matrix represents a different instance of the distances in the set of proteins for a different pair of species. For each row of the matrix, the 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 the median), l (large distances), ll (very large distances). An additional state, NA, was used for any missing values in the distance matrix. Denoting two generic proteins as p,q and two generic intervals as a,b, 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 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 the 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 (Besag, 1977) 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 co-evolutionary 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 (Ekeberg *et al*, 2013), 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 (Dunn *et al*, 2008) obtaining the co-evolutionary coupling between proteins p and q, $C_{p,q} = F_{p,q} - F_pF_q/F$.

In order to assign statistical significance 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 ran the pipeline described above for every random set and retrieved the corresponding matrix of co-evolutionary couplings. P-values were assigned based on the random distribution obtained and associations supported by p values < 0.05 were further considered. The matrix of co-evolutionary couplings and corresponding p values are included in **Suppl. Table 3**.

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 (Weigt *et al*, 2009; de Juan *et al*, 2013).

Identification of epigenetic signals with a statistically significant co-evolutionary effect

For each epigenetic signal (histone mark/cytosine modification), we identified all the pairs of CrPs that satisfy the following two conditions: 1) the proteins in the pair are co-evolutionary coupled (see above); and 2) each of the proteins in the pair directly interacts with the epigenetic signal. We then used the number of unique CrPs in the resulting set of pairs (Co-evolutionary Filtered Centrality, CFC) as a measure of the influence of the signal on co-evolution between the CrPs in the epigenetic signaling network. The analysis resulted in 7 signals with a CFC greater than zero: H3K4me1 (CFC=2), H3K4me2 (4), H3K4me3 (5), H3K9ac (4), H3K27ac (2), 5fC (2), 5hmC

(9). 5hmC is clearly the signal with the strongest effect on co-evolution, with a CFC=9 almost double that of the second ranking signal (H3K4me3).

The statistically significance of each CFC was evaluated by computing a p-value that corresponded to the probability of obtaining a CFC greater or equal to that observed in a network model with randomly-generated edges among the CrPs in the co-evolutionary analysis. This procedure identified three signals with a significant CFC (p-value < 0.05): 5hmC (p-value approx 0.04), H3K4me2 (0.01), H3K4me3 (0.02).

Functional Modularization of the Co-localization Network

The co-localization network was decomposed into local networks of positive interactions. First, we calculated the frequency of each positive interaction using ChromHMM peaks, considering that an interaction is present if both interactors are 'present' in the same 200 bp genomic window. We calculated this frequency for each of the 20 chromatin states, such that we have a vector of 20 frequencies for each positive interaction. In order to reduce state-specific biases, the frequencies of the interactions were standardized separately for every state. These vectors were clustered by hierarchical clustering (Pearson correlation, average linkage) and the largest statistically supported clusters (p-value < 0.05, n=10,000) according to Pvclust (Suzuki & Shimodaira, 2006) were defined as chromnets (see **Supp. Figure 5**).

Gene Ontology enrichment analysis

Gene Ontology enrichment analyses were carried out with GREAT v3.0.0 (McLean *et al*, 2010). We used the independent segments of 5hmC co-localization with its emitters and receivers as the input to predict biological functions of the associations analyzing the closest genes. The genomic regions were associated to genes with a minimum distance of 5Kb upstream and 1Kb downstream, with the whole genome as the background. The False Discovery Rate (FDR) considered was 0.05 (see Supplementary Table 12).

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

Acknowledgements

Author contributions

Conflict of interest

Funding

References

Balasubramani A & Rao A (2013) O-GlcNAcylation and 5-Methylcytosine Oxidation: An Unexpected Association between OGT and TETs. *Mol. Cell* **49:** 618–619

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

Van Bemmel JG, Filion GJ, Rosado A, Talhout W, de Haas M, van Welsem T, van Leeuwen F & van Steensel B (2013) A network model of the molecular organization of chromatin in Drosophila. *Mol. Cell* **49:** 759–771

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

Bond MR & Hanover JA (2015) A little sugar goes a long way: the cell biology of O-GlcNAc. *J. Cell Biol.* **208:** 869–880

Choi I, Kim R, Lim H-W, Kaestner KH & Won K-J (2014) 5-hydroxymethylcytosine represses the activity of enhancers in embryonic stem cells: a new epigenetic signature for gene regulation. *BMC Genomics* **15:** 670

Doege CA, Inoue K, Yamashita T, Rhee DB, Travis S, Fujita R, Guarnieri P, Bhagat

731 G, Vanti WB, Shih A, Levine RL, Nik S, Chen EI & Abeliovich A (2012) Early-
732 stage epigenetic modification during somatic cell reprogramming by Parp1
733 and Tet2. *Nature* **488:** 652–655
734 Dunn SD, Wahl LM & Gloor GB (2008) Mutual information without the influence
735 of phylogeny or entropy dramatically improves residue contact prediction.
736 *BIOINFORMATICS* **24:** 333–340
737 Ekeberg M, Lovkvist C, Lan Y, Weigt M & Aurell E (2013) Improved contact
738 prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys.*
739 *Rev. E* **87:** 012707
740 Ernst J & Kellis M (2010) Discovery and characterization of chromatin states for
741 systematic annotation of the human genome. *Nat. Biotechnol.* **28:** 817–825
742 Ernst J & Kellis M (2012) ChromHMM: automating chromatin-state discovery
743 and characterization. *Nat. Methods* **9:** 215–216
744 Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, Marques CJ,
745 Andrews S & Reik W (2011) Dynamic regulation of 5-hydroxymethylcytosine
746 in mouse ES cells and during differentiation. *Nature* **473:** 398–402
747 Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman
748 W, de Castro IJ, Kerkhoven RM, Bussemaker HJ & van Steensel B (2010)
749 Systematic protein location mapping reveals five principal chromatin types
750 in Drosophila cells. *Cell* **143:** 212–224
751 Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, Bonetti A,
752 Voineagu I, Bertin N, Kratz A, Noro Y, Wong C-H, de Hoon M, Andersson R,
753 Sandelin A, Suzuki H, Wei C-L, Koseki H, Hasegawa Y, Forrest ARR, et al
754 (2014) Deep transcriptome profiling of mammalian stem cells supports a
755 regulatory role for retrotransposons in pluripotency maintenance. *Nat.*
756 *Genet.* **46:** 558–566
757 Hashimoto H, Liu Y, Upadhyay AK, Chang Y, Howerton SB, Vertino PM, Zhang X &
758 Cheng X (2012) Recognition and potential mechanisms for replication and
759 erasure of cytosine hydroxymethylation. *Nucleic Acids Res.* **40:** 4841–4849
760 Hemmrich G, Khalturin K, Boehm A-M, Puchert M, Anton-Erxleben F, Wittlieb J,
761 Klostermeier UC, Rosenstiel P, Oberg H-H, Domazet-Lošo T, Sugimoto T,
762 Niwa H & Bosch TCG (2012) Molecular Signatures of the Three Stem Cell
763 Lineages in Hydra and the Emergence of Stem Cell Function at the Base of
764 Multicellularity. *Mol. Biol. Evol.* **29:** 3267–3280
765 He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L, Sun Y, Li X,
766 Dai Q, Song C-X, Zhang K, He C & Xu G-L (2011) Tet-Mediated Formation of 5-
767 Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science* **333:**
768 1303–1307
769 Ikegami K, Ohgane J, Tanaka S, Yagi S & Shiota K (2009) Interplay between DNA
770 methylation, histone modification and chromatin remodeling in stem cells
771 and during development. *Int. J. Dev. Biol.* **53:** 203–214
772 Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C & Zhang Y (2011) Tet
773 Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-
774 Carboxylcytosine. *Science* **333:** 1300–1303
775 De Juan D, Pazos F & Valencia A (2013) Emerging methods in protein co-
776 evolution. *Nat. Rev. Genet.* **14:** 249–261
777 Juan D, Rico D, Marques-Bonet T, Fernández-Capetillo O & Valencia A (2013)
778 Late-replicating CNVs as a source of new genes. *Biol. Open* **2:** 1402–1411
779 Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo

PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TK, et al (2011) Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471:** 480–485

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

Liyanage VRB, Jarmasz JS, Murugeshan N, Del Bigio MR, Rastegar M & Davie JR (2014) DNA Modifications: Function and Applications in Normal and Disease States. *Biology* **3:** 670–723

Matarese F, Carrillo-de Santa Pau E & Stunnenberg HG (2011) 5-Hydroxymethylcytosine: a new kid on the epigenetic block? *Mol. Syst. Biol.* **7:** 562–562

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM & Bejerano G (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28:** 495–501

Mitra K, Carvunis A-R, Ramesh SK & Ideker T (2013) Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* **14:** 719–732

Moen EL, Mariani CJ, Zullow H, Jeff-Eke M, Litwin E, Nikitas JN & Godley LA (2015) New themes in the biological functions of 5-methylcytosine and 5-hydroxymethylcytosine. *Immunol. Rev.* **263:** 36–49

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

Papp B & Plath K (2012) Pluripotency re-centered around Esrrb. *EMBO J.* **31:** 4255–4257

Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, Ko M, McLoughlin EM, Brudno Y, Mahapatra S, Kapranov P, Tahiliani M, Daley GQ, Liu XS, Ecker JR, Milos PM, Agarwal S & Rao A (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473:** 394–397

Perner J, Lasserre J, Kinkley S, Vingron M & Chung H-R (2014) Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling. *Nucleic Acids Res.* **42:** 13689–13695

Pfeifer GP, Kadam S & Jin S-G (2013) 5-hydroxymethylcytosine and its potential roles in development and cancer. *Epigenetics Chromatin* **6:** 10

Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, Jensen LJ, von Mering C & Bork P (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42:** D231–239

Raiber E-A, Beraldi D, Ficz G, Burgess HE, Branco MR, Murat P, Oxley D, Booth MJ, Reik W & Balasubramanian S (2012) Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. *Genome Biol.* **13:** R69

Schreiber SL & Bernstein BE (2002) Signaling Network Model of Chromatin. *Cell* **111:** 771–778

Scott-Phillips TC (2008) Defining biological communication. *J. Evol. Biol.* **21:** 387–395

Smith T late JM & Harper D (2003) Animal Signals Oxford Series in Ecology and Evolution

Song C-X, Szulwach KE, Fu Y, Dai Q, Yi C, Li X, Li Y, Chen C-H, Zhang W, Jian X,

Wang J, Zhang L, Looney TJ, Zhang B, Godley LA, Hicks LM, Lahn BT, Jin P & He C (2011) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29:** 68–72

Stroud H, Feng S, Morey Kinney S, Pradhan S & Jacobsen SE (2011) 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.* **12:** R54

Suzuki R & Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinforma. Oxf. Engl.* **22:** 1540–1542

Szulwach KE, Li X, Li Y, Song C-X, Han JW, Kim S, Namburi S, Hermetz K, Kim JJ, Rudd MK, Yoon Y-S, Ren B, He C & Jin P (2011) Integrating 5-Hydroxymethylcytosine into the Epigenomic Landscape of Human Embryonic Stem Cells. *PLoS Genet* **7:** e1002154

Vaissière T, Sawan C & Herceg Z (2008) Epigenetic interplay between histone modifications and DNA methylation in gene silencing. *Mutat. Res.* **659:** 40–48

Vella P, Scelfo A, Jammula S, Chiacchiera F, Williams K, Cuomo A, Roberto A, Christensen J, Bonaldi T, Helin K & Pasini D (2013) Tet Proteins Connect the O-Linked N-acetylglucosamine Transferase Ogt to Chromatin in Embryonic Stem Cells. *Mol. Cell* **49:** 645–656

Wang J, Hevi S, Kurash JK, Lei H, Gay F, Bajko J, Su H, Sun W, Chang H, Xu G, Gaudet F, Li E & Chen T (2009) The lysine demethylase LSD1 (KDM1) is required for maintenance of global DNA methylation. *Nat. Genet.* **41:** 125–129

Wang J, Scully K, Zhu X, Cai L, Zhang J, Prefontaine GG, Krones A, Ohgi KA, Zhu P, Garcia-Bassets I, Liu F, Taylor H, Lozach J, Jayes FL, Korach KS, Glass CK, Fu X-D & Rosenfeld MG (2007) Opposing LSD1 complexes function in developmental gene activation and repression programmes. *Nature* **446:** 882–887

Weigt M, White RA, Szurmant H, Hoch JA & Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **106:** 67–72

Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PAC, Rappsilber J & Helin K (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473:** 343–348

Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, Sun YE & Zhang Y (2011) Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev.* **25:** 679–684

Yang X, Ongusaha PP, Miles PD, Havstad JC, Zhang F, So WV, Kudlow JE, Michell RH, Olefsky JM, Field SJ & Evans RM (2008) Phosphoinositide signalling links O-GlcNAc transferase to insulin resistance. *Nature* **451:** 964–969

Ye T, Krebs AR, Choukrallah M-A, Keime C, Plewniak F, Davidson I & Tora L (2011) seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.* **39:** e35

Zhang Y, Wong C-H, Birnbaum RY, Li G, Favaro R, Ngan CY, Lim J, Tai E, Poh HM, Wong E, Mulawadi FH, Sung W-K, Nicolis S, Ahituv N, Ruan Y & Wei C-L (2013) Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature* **504:** 306–310

Zwaka TP (2012) Pluripotency network in embryonic stem cells: maybe Leibniz was right all along. *Cell Stem Cell* **11:** 441–442

Figure legends

Figure 1. A framework to study communication among chromatin components

Our network approach is based on a classification of epigenomic features into three component classes, where histone and cytosine modifications are always considered to be signals and the chromatin-related proteins (CrPs) can be either co-occurring (or mutually exclusive) emitters (writers/erasers) or receivers (readers) of those epigenetic signals.

Figure 2. Chromatin communication network in ESCs

A full chromatin communication network in which the edges represent positive or negative interactions that indicate genomic co-localization or mutual exclusion, respectively. Arrows associated with the directional edges represent communication flux for emitter-signal or signal-receiver pairs retrieved from the literature. The colors indicate membership of known protein complexes. B Emitters and receivers of the H3K79me2 hub signal. C Emitters and receivers of the 5hmC hub signal.

Figure 3. Co-evolution of CrPs

Coupling analysis of the phylogenetic histories of CrPs revealed significant co-evolution between emitters and receivers of 5hmC and 5fC. Co-evolving pairs are indicated by thick colored dashed lines. The grey lines indicate co-localization or mutual exclusion in the chromatin communication network (see Figure 2 for more details).

Figure 4. Chromnets recover known protein complexes and star-shaped structures

The chromnets are sub-networks of interactions with similar co-occurrence across the chromatin states and they have different topologies. Each bar plot indicates the overall enrichment of the chromatin states in each chromnet along (see B for details of the chromatin states). B Star-like 5hmC sub-network and the overall enrichment of chromatin states.

Figure 5. 5hmC genomic regions have different functional enrichment depending on the co-localizing partner

A-D Read densities over a 10Kb windows centered on the 5hmC-ESRRB (A), 5hmC-LSD1 (B), 5hmC-TET1 (C) and 5hmC-OGT (D) peaks. We calculated the read density of 5hmC, ESRRB, LSD1, TET1 (N- and C-terminal ChIP-seqs) and OGT in 10Kb windows centered on the genomic bins (200 bp), where 5hmC co-localizes exclusively with each specific partner (i.e.: the rest of the 5hmC interactors are not present). The read density plots were obtained with the SeqMINER platform v1.3.3e (Ye *et al*, 2011). The average density of the reads in 50 bp bins was plotted from the center of the 5hmC independent genomic regions to +/-5000 bp. E Gene Ontology enrichment analysis of peaks in A-D using GREAT.

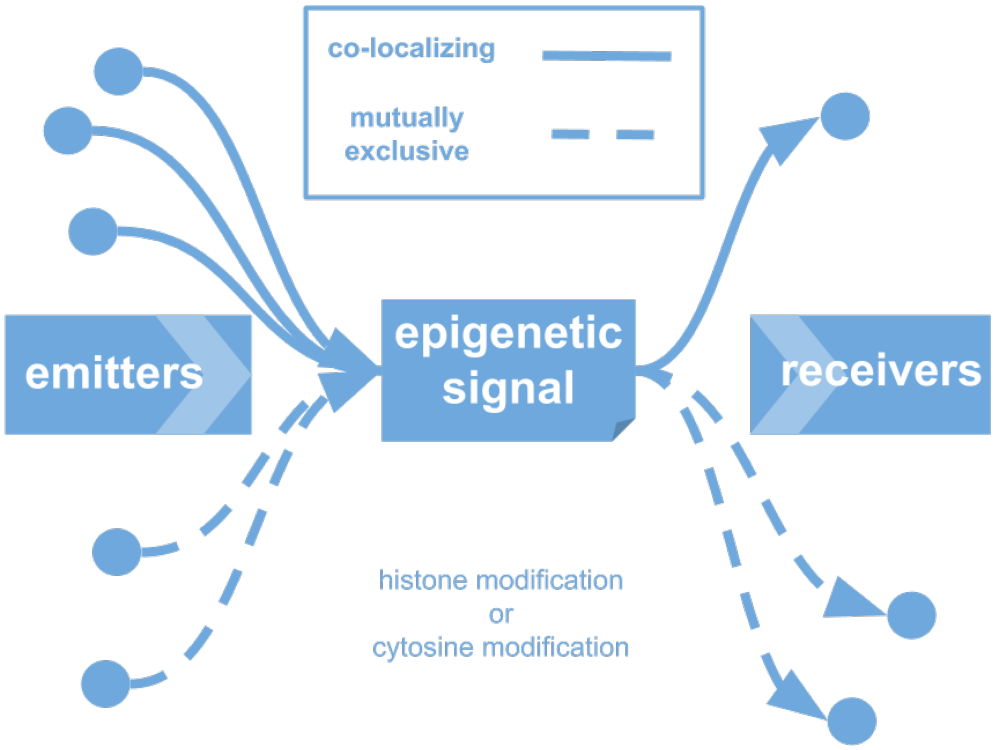


Figure 1

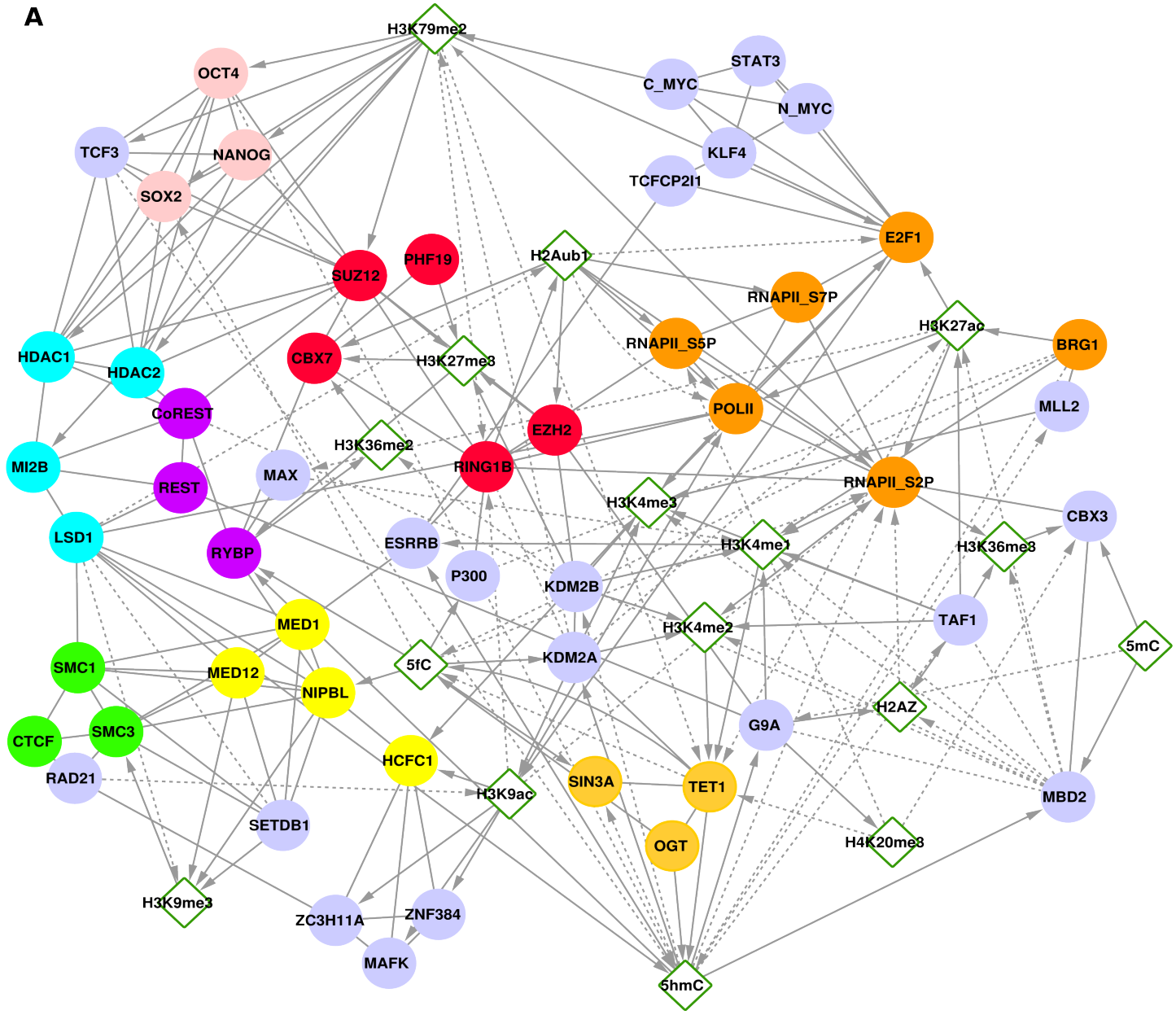


Figure 2

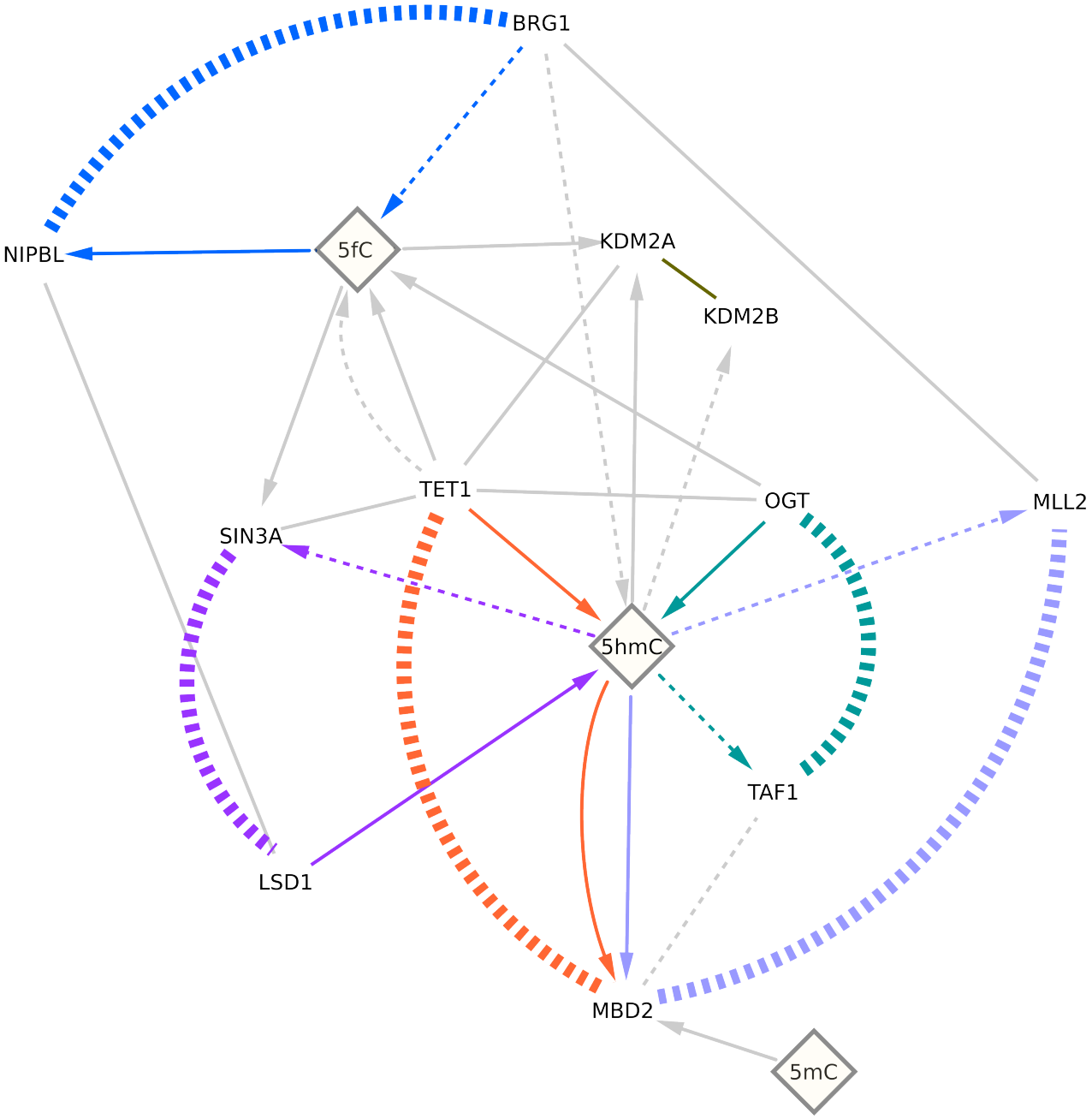
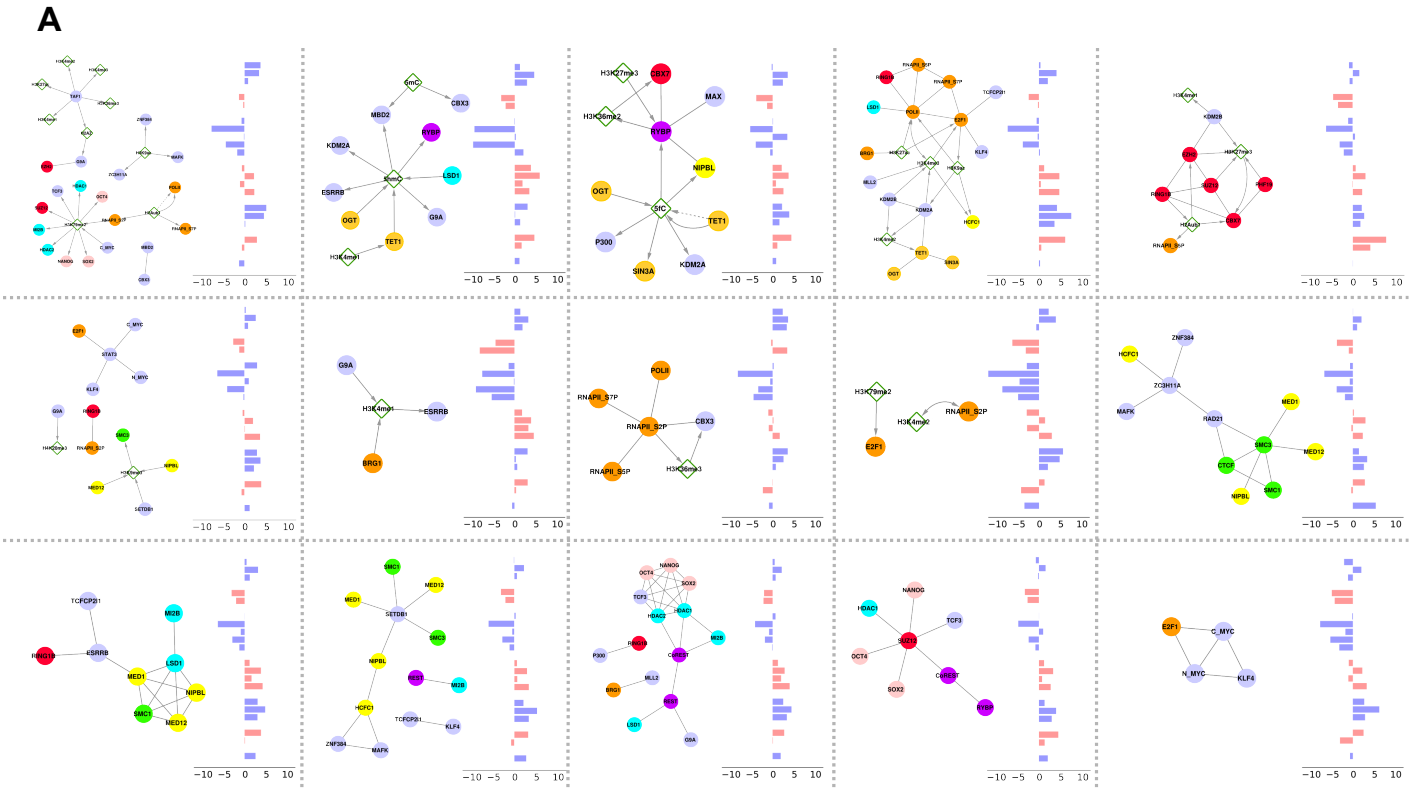


Figure 3



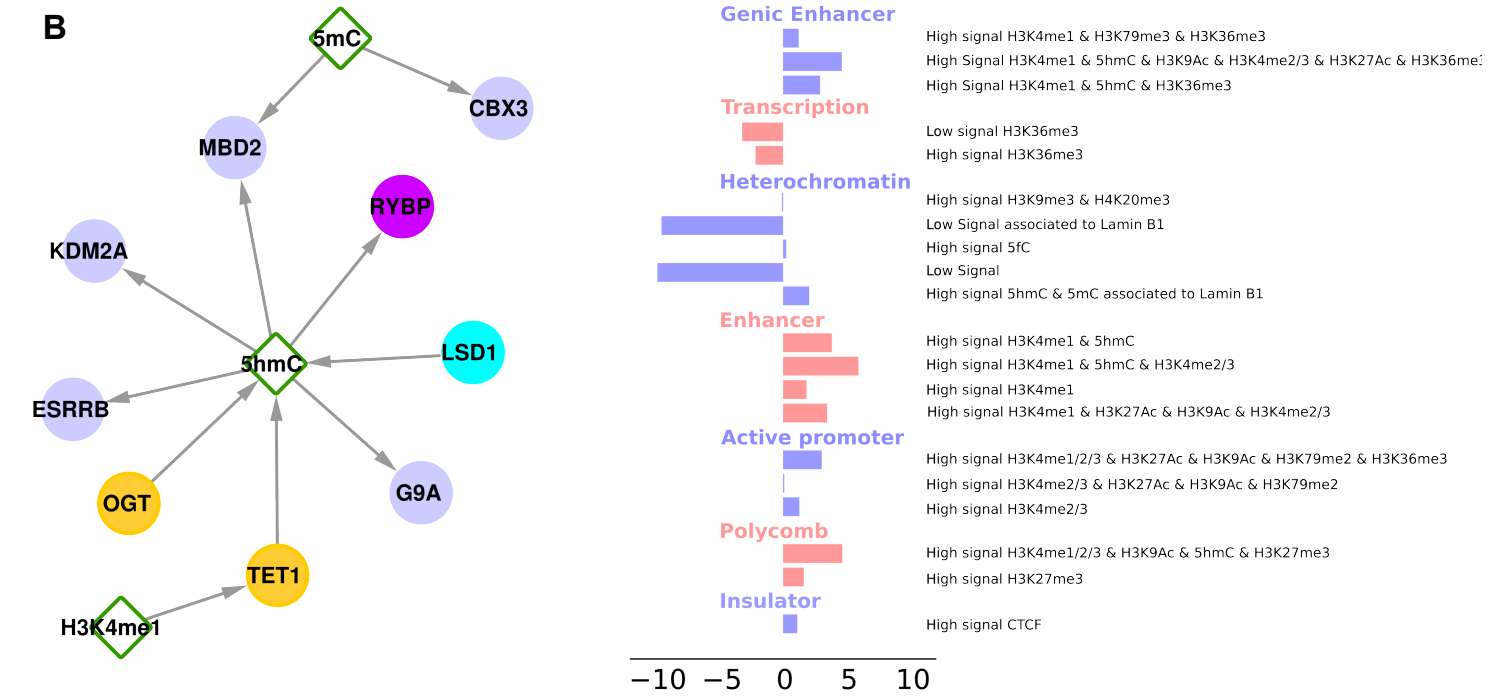
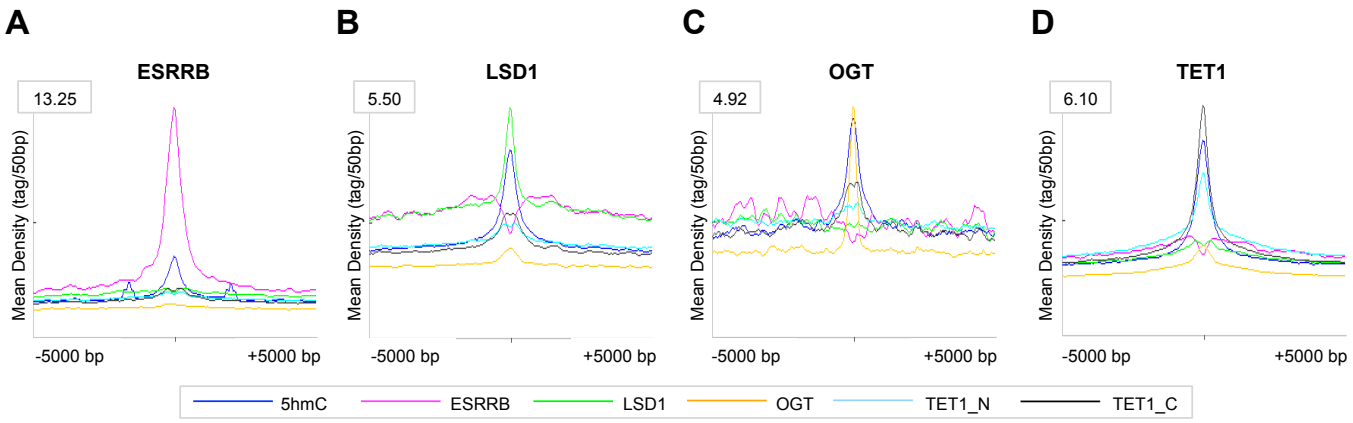


Figure 4



E

GO term	ESRRB	LSD1	TET1	OGT
Maturation of SSU-rRNA	■			
Regative regulation of MAPK cascade	■			
Plasma membrane organization	■			
Intrinsic apoptotic signaling pathway	■			
Cell-cell junction organization	■			
Epithelial cell morphogenesis	■			
Somatic stem cell maintenance	■		■	
Regulation of Notch signaling pathway	■		■	
Mesoderm morphogenesis	■		■	
Histone acetylation	■	■		
Regulation of gene expression, epigenetic	■	■		
DNA methylation or demethylation		■		
Histone H3-H4 acetylation		■		
Placenta development		■		
Blastocyst formation		■		
Proteasomal protein catabolic process		■		
Regulation of translation		■		
Regulation of stem cell maintenance			■	
Regulation of TGF-beta receptor signaling			■	
Apoptotic process involved in morphogenesis			■	
Cell-cell signaling involved in cell fate commitment			■	
Developmental induction			■	
Mesoderm development			■	
Body and head morphogenesis			■	■
Carbohydrate biosynthetic process				■
Glycerophospholipid metabolic process				■
Regulation of lipid storage				■
Maintenance of location in cell				■
Modulation by virus of host morphology or physiology				■
Regulation of organelle organization				■
Regulation of chromatin organization				■
Regulation of histone modification				■
Regulation of cytokinesis				■

Figure 5