Epigenetic regulation of the transcriptome

Beatrice Borsari

TESI DOCTORAL UPF / 2020

THESIS SUPERVISOR

Roderic Guigó Serra

Department of Experimental and Health Sciences
UNIVERSITAT POMPEU FABRA (UPF)

Bioinformatics and Genomics Programme
CENTRE FOR GENOMIC REGULATION (CRG)

A mamma e papà, e a Diego

Abstract

We have monitored the transcriptome and the epigenome of human pre-B cells transdifferentiating into macrophages. Analysis of these data provides a general framework to understand the relationship between gene expression and chromatin. We have observed widespread uncoupling of gene expression and epigenetic features during transdifferentiation, with several genes characterized by unvaried chromatin state throughout the process, irrespective of changes in gene expression. Nevertheless, we report a strong association between transcription and chromatin marking of promoter regions at the time of initial gene activation. We have also analyzed the genomic location of distal regulatory elements in developmental and adult samples, and found that tissue-specific enhancer signatures in the human genome tend to accumulate within introns, while those shared among tissues are more frequently intergenic. By focusing on intronic segments, we have additionally uncovered both constraint and variation in the timing of splicing, with a subset of introns that switch from co-transcriptional to post-transcriptional splicing across distinct cell types.

Resumen

Hemos monitorizado el transcriptoma y el epigenoma de células pre-B durante su transdiferenciación en macrófagos. El análisis de estos datos proporciona un marco general para comprender la relación entre la expresión génica y la cromatina. Observamos un desacoplamiento generalizado entre la expresión génica y las marcas epigenéticas durante la transdiferenciación, con multitud de genes caracterizados por un único estado de la cromatina, independientemente de los cambios en su expresión. No obstante, encontramos una fuerte asociación entre la transcripción y las marcas de la cromatina en los promotores durante la activación inicial de los genes. También hemos estudiado la localización genómica de elementos reguladores distales (*enhancers*), en muestras obtenidas tanto de tejidos embrionarios como adultos, encontrando que los *enhancers* específicos de tejido tienden a estar situados en los intrones, mientras que aquellos compartidos entre tejidos son, a menudo, intergénicos. Por último, centrándonos en el estudio de los intrones, hemos identificado tanto conservación como variabilidad en la temporalidad del *splicing*, con un subconjunto de intrones que cambian de *splicing* co-transcripcional a post-transcripcional en distintos tipos celulares.

Preface

In the post-genomic era, efforts are focused on characterizing all functional elements of the genome, with the ultimate goal of understanding how this information translates into specific features and functions of the cell. RNA biogenesis and processing play a fundamental role in the flow of information from DNA to proteins, and large scale transcriptomic experiments are nowadays a common framework to investigate general principles of gene expression regulation. In this context, the level of expression of a gene represents a direct measure of its transcriptional activity. Nevertheless, these measurements do not inform as to the molecular processes preceding and/or causing gene activation, nor about the mechanisms required to cease transcription.

Chromatin is the complex of DNA and histone proteins found in the nucleus of eukaryotic cells, and its impact on gene expression has been subject to extensive research, accelerated in the last fifteen years by the development of genome-wide sequencing technologies. By means of these methods, it is possible to investigate which precise biochemical activities, besides transcription, occur at a genomic locus at a specific time or under a given circumstance. Over the years, these analyses have enabled deciphering molecular events and epigenetic patterns associated with the activation and silencing of genes, a brief description of which is provided in the introductory section.

Although mostly correlative, these observations have led to the formulation of the histone code hypothesis, according to which the combinatorial nature of histone marks at individual genomic loci dictates certain biochemical outputs, and orchestrates transcriptional programs that are finely tuned in time and space. Importantly, the widely accepted belief of a causal role of histone marks on gene expression builds upon studies that are often conducted in steady-state conditions, and therefore do not explore the interplay between expression and chromatin marks over time. To address this shortcoming, we have monitored the transcriptome and the epigenome during the transdifferentiation of human pre-B cells into macrophages, generating RNA-seq profiles and ChIP-seq maps of nine histone modifications over a period of seven days (Chapter 1). Analysis of these data reveals that human genes display a limited number of combinations of histone marks, which behave in a coordinated manner over time.

We have observed widespread uncoupling of gene expression and histone modifications during transdifferentiation, with several genes characterized by unvaried chromatin state throughout the process, irrespective of changes in gene expression. Notwithstanding, we report positive association between transcription and chromatin modifications at the time of initial gene activation, a stage that is actually characterized by a precise order of events bringing about transcription initiation and histone marks' deposition.

While the focus of Chapter 1 is on the temporal relationship between transcription and chromatin marking at promoter regions, the work described in Chapter 2 addresses the role of distal regulatory elements in orchestrating general and tissue-specific gene expression patterns. By leveraging the ENCODE registry of candidate *cis*-regulatory elements (cCREs), we have identified sets of common and tissue-specific enhancer-like signatures in the human genome. We report that distal regulatory activity shared among human tissues is more frequently located in intergenic regions, while tissue-specific enhancers are more abundant within introns. Remarkably, we have observed that intronic regulatory elements are associated with genes involved in tissue-specific functions and homeostasis. Furthermore, the enrichment in tissue-specific intronic regulatory elements appears to correlate with the degree of specialization of the tissue. In fact, more differentiated tissues present higher rates of intronic enhancers, while the lowest rate is observed in embryonic stem cells.

Although often assayed in whole cell fractions, transcriptomic profiles can significantly vary between the nucleus and the cytosol, mainly as a consequence of RNA processing events. For this reason, in Chapter 3 we evaluate differences in splicing completion between nuclear and cytosolic transcripts, considering such differences as a proxy for the moment at which splicing of a given intron occurs (i.e. co-transcriptionally or post-transcriptionally). We have performed this exercise across a panel of 13 human cell lines, and uncovered that the proportion of introns undergoing post-transcriptional splicing dramatically varies across cellular conditions. Specifically, we have identified groups of introns with constrained timing of splicing (constitutively either co- or post-transcriptional) across cell lines, but also a subset of introns that switch from co-transcriptional to post-transcriptional splicing and more often belong to protein-coding genes. We have analyzed patterns of chromatin features and RNA-binding proteins, and found that specific components of the spliceo-

some machinery are preferentially bound to introns retained for a longer time within the transcript.

List of publications during the thesis:

1. **Borsari B.**, Peña Castillo L. and Guigó R. Variation and constraint in the timing of splicing. *In preparation*.
2. **Borsari B.**, Abad A., Klein C.K., Nurtdinov R., Esteban A., Palumbo E., Ruiz-Romero M., Sanz M., Correa B.R., Johnson R., Pérez-Lluch, S. and Guigó R. (2020). Dynamics of gene expression and chromatin marking during cell state transition. *Submitted*.
3. **Borsari B.***, Villegas-Mirón P.*, Laayouni H., Segarra-Casas A., Bertranpetit J., Guigó R. and Acosta S. (2020). Intronic enhancers regulate the expression of genes involved in tissue-specific functions and homeostasis. *Submitted*.
4. Garrido-Martín D., **Borsari B.**, Calvo M., Reverter F. and Guigó R. (2020). Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat Commun.*, in press.
5. The ENCODE Project Consortium., Moore, J.E., Purcaro, M.J., Pratt H.E., Epstein C.B., Shoresh N., Adrian J., Kawli T., Davis C.A., Dobin A., Kaul R., Halow J., et al., (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature 583*(7818), 699-710.
6. The ENCODE Project Consortium, Snyder M.P., Gingeras T.R., Moore J.E., Weng Z., Gerstein M.B., Ren B., Hardison R.C., Stamatoyannopoulos J.A., Graveley B.R., Feingold E.A., Pazin M.J., et al., (2020). Perspectives on ENCODE. *Nature 583*(7818), 693-698.
7. Ramilowski J.A., Yip C.W., Agrawal S., Chang J. C., Ciani Y., Kulakovskiy I.V., Mendez M., Ooi J.L.C., Ouyang J.F., Parkinson N., Petri A., et al. (2020). Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res. 30*(7), 1060-1072.
8. Stik G., Vidal E., Barrero M., Cuartero S., Vila-Casadesús M., Mendieta-Esteban J., Tian V.T., Choi J., Berenguer C., Abad

A., **Borsari B.**, le Dily F., et al. (2020). CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response. *Nat Genet. 52*(7), 655-661.

9. Pervouchine D., Popov Y., Berry A., **Borsari B.**, Frankish A. and Guigó R. (2019) Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. *NAR 47*(10), 5293-5306.

Contents

INTRODUCTION

Back to the origins: what is epigenetics?

Chromatin identifies the complex of DNA, histone and non-histone proteins that constitutes the chromosomes found in the nucleus of eukaryotic cells. The word chromatin – which means "stainable material" – was coined in 1882 by Walther Flemming, who discovered chromosomes by studying the distribution of the genetic material during cell division (Flemming Walther, 1882). In fact, because of the low resolution of the microscopes of the time, researchers used to stain fixed cells to enhance the contrast of their contents.

Although the genetic information encoded in the DNA sequence is largely identical in every cell of an eukaryotic system, cells in different tissues and organs exhibit distinct gene expression patterns, and ultimately perform distinct functions (Felsenfeld, 2014). Most importantly, these features can be clonally inherited, suggesting that chromatin changes other than those affecting the DNA sequence (i.e. genetic variation) can shape cell memory and lead to heritable phenotypic traits.

The branch of biology that studies such changes is called epigenetics. Its origins trace back to the early decades of the 20th century, and its history is tightly linked with the study of evolution and development. In 1904, Theodor Boveri’s work on sea urchins demonstrated that a specific number of chromosomes is required for normal development (Boveri, 1904). Roughly in the same years, independent work by Walter Sutton with lubber grasshoppers related, for the first time, the segregation of chromosomes to Mendel’s laws of heredity (Sutton, 1902, 1903). These discoveries set the basis of the chromosome theory of inheritance, which postulates that chromosomes are the carriers of genetic material, and thus of hereditary information (Wilson, 1925). However, experimental evidence of this theory did not come until 1911, when Thomas Hunt Morgan unequivocally demonstrated that genes are located on chromosomes (Morgan, 1911). Although skeptical about both the Mendelian and the chromosome theories of inheritance, Morgan reconsidered his posi-

tions when he discovered that eye color in the fruit fly expresses a sex-linked trait, and that the mechanistic bases of this association reside in a gene located on the X chromosome. This was followed by a rapid expansion of genetic linkage studies, which culminated in 1913 with the development of the first gene map in *Drosophila* by Alfred Sturtevant (Sturtevant, 1913).

The breakthrough by Morgan certainly inaugurated classical experimental genetics, but also opened to further, more intriguing questions. The fundamental question – resumed years later by Conrad Waddington (Waddington, 1953) – focused on how a single fertilized egg could give rise to a mature organism formed by cells of heterogeneous phenotypes. Researchers began to wonder which specific molecules within the chromosomes and which chemical reactions account for the transfer of genetic information from one cell to the other, and thus ensure the correct execution of developmental programs. In 1930, Hermann Muller provided the first evidence that genes do not function as independent portions of chromosomes, and that altering their location within the genome, but not their composition, can lead to heritable phenotypic changes. In fact, X ray-induced chromosome rearrangements in the fruit fly could result in a mottled-eyes progeny (Muller, 1930). The phenomenon, later on described as position-effect variegation (PEV; reviewed in Elgin and Reuter, 2013), was caused by the juxtaposition of the white gene with a heterochromatin region (Hannah, 1951). Based on these results, it became evident that several mechanisms other than alterations of individual genes (point mutations) may impact the inheritance of genetic information.

The discovery of deoxyribonucleic acid (DNA) as the "transforming principle" that accounts for inheritance of specific characteristics (Avery et al., 1944) relighted the debate on developmental processes. At the time, karyotype studies suggested that all somatic cells present the same set of chromosomes, but it was unclear whether that corresponded to the genetic makeup of the zygote. Briggs and King demonstrated in 1952 that transplantation of the nucleus of an early embryonic cell into an enucleated egg of *Rana pipiens* enables the egg to develop into a normal embryo (Briggs and King, 1952). However, the ultimate proof in this sense was provided in 1970 by Laskey and Gurdon, who obtained tadpoles by transplanting nuclei of somatic cells into enucleated eggs of *Xenopus* (Laskey and Gurdon, 1970). This was the most persuasive evidence that the signals orchestrating developmental and differentiation programs do not af-

fect the germline DNA sequence, but actually lay on top of it (*epi*-genetics), mostly in the form of modifications to the bases and to the proteins complexed with the DNA, as well as genome rearrangements (Felsenfeld, 2014).

Epigenetic paradigms in development and reprogramming

Gene activation and silencing during development and differentiation lay at the heart of epigenetic studies. There is evidence that a number of mechanisms, including chemical modifications of DNA bases and histone proteins, incorporation of histone variants and changes in chromatin accessibility and conformation, contribute to the establishment of specific gene expression patterns in the developing embryo, and to their stable propagation through cell division (reviewed in Cantone and Fisher, 2013). How these mechanisms concretely contribute to cellular identity and lineage fidelity is still largely unknown. Nevertheless, at least two global perturbations of the epigenetic landscape seem to be required for correct development in mammals. The first wave occurs at fertilization, when gametes fuse to generate the first cell of the embryo, the zygote. At this stage, independent remodeling (mostly in the form of DNA demethylation and histone hyperacetylation) of the paternal and maternal genomes resets the epigenetic landscape to a ground state characteristic of cellular totipotency. This competence is progressively lost during cleavage divisions. A second *in vivo* "reprogramming" occurs when primordial germ cells (the gamete precursors) develop and migrate in the embryo, and it primes an epigenetic status required to reestablish totipotency in the next generation. This stage is characterized by a switch in chromatin repressive marking, with loss of H3K9me2 and increase of H3K27me3 (Hajkova et al., 2008, Seki et al., 2005).

Several studies have suggested that histone modifications and remodelers play a fundamental role both in early development and throughout differentiation. For instance, altered levels of H3R26me2 in blastomeres are associated with impaired lineage fate decisions (Torres-Padilla et al., 2007), and the lack of specific DNA and histone methyltransferases affects embryonic stem cells (ESCs) differentiation (Fisher and Fisher, 2011). Perhaps the most important feature of early developing cells is their flexibility in lineage commitment, and

Polycomb repressor complexes (PRCs), in particular the H3K27me3-depositing PRC2, seem to be involved in maintaining a functional "bivalency" of their epigenetic landscape (Azuara et al., 2006, Bernstein et al., 2006, Voigt et al., 2012). This means that several genes – including key developmental regulators – are kept in a "poised" condition, and can rapidly switch from a silent to a transcriptionally active chromatin state.

Epigenetic phenomena are also at the basis of the reprogramming of somatic cells towards pluripotency (also reviewed in Cantone and Fisher, 2013). In these experimental models, the forced expression of a set of key transcription factors (Oct4, Sox2, Klf4 and c-Myc, or OSKM factors) alters the gene expression program of the differentiated cell. The mechanisms by which these transcription factors (TFs) reactivate the pluripotent status of the cell remain largely unexplained. Nonetheless, because of the low rate of successful cell reprogramming observed so far, stochastic epigenetic events are likely to play an important yet undisclosed role in this process (Buganim et al., 2012, Hanna et al., 2009). In line with this, a number of studies have shown that cellular plasticity and reprogramming success can be improved by perturbing the epigenetic landscape of the somatic cell, for instance by increasing the levels of histone modifications typically associated with active gene expression, such as H3K9ac and H3K4me3 (Mali et al., 2010, Mikkelsen et al., 2008, Sridharan et al., 2009). On the other hand, the forced deposition of the repressive mark H3K9me3 by SUV39H1 (but not SUV39H2) impairs reprogramming (Burton et al., 2020).

Epigenetic modes of transcriptional regulation

One of the critical questions in epigenetics is how genes are switched on and off appropriately in time and space. Below, we explore how distinct epigenetic mechanisms are related to different aspects of gene expression regulation. We acknowledge that some long-standing questions in the field remain the contribution of transcription factors’ binding and histone post-translational modifications to transcriptional programs, as well as which features permit identifying different types of regulatory genomic elements, such as promoters and enhancers.

Nucleosome positioning and the accessible genome

The DNA inside the nucleus of an eukaryotic cell is wrapped around proteins called histones. The basic structural units of this packaging – the nucleosomes – have a non-uniform distribution across the genome, and portions of DNA that lie free of histones are more prone to physically interact with nuclear macromolecules (Klemm et al., 2019). These interactions play a key role in the expression of genes. As a consequence, active genes and enhancers (small segments of DNA that recruit TFs and regulate gene expression at distal loci) tend to reside in a more relaxed context of open chromatin, originally referred to as *euchromatin*. On the other hand, compact, *heterochromatic* regions typically host transcriptionally silent genes, or perform more structural functions, such as centromeres and telomeres (Xhemalce B et al., 2011).

The degree of accessibility of the DNA sequence to *trans* factors depends on both the density of the associated histone proteins (nucleosome occupancy) and their fractional residence time (nucleosome turnover). While contributing to a gradient of chromatin accessibility that varies across different regulatory elements, these two features are not always inversely correlated. For instance, nucleosome occupancy is usually low at both structural insulators and actively transcribed Transcription Start Sites (TSSs), but the former display low nucleosome turnover rates while the latter are associated with unstable nucleosomes. Similarly, active and inactive enhancers have comparable nucleosome occupancy rates but different nucleosome turnover (Klemm et al., 2019).

The accessible genome corresponds to a small fraction of the total DNA sequence (2-3%) (Klemm et al., 2019), and can change in response to external stimuli and developmental cues. As a matter of fact, profiles of single-cell genome accessibility can recapitulate cell type-specificity (Cusanovich et al., 2018, Ma et al., 2020). Therefore, the landscape of chromatin accessibility of a given cell ultimately reflects its capacity to read and execute specific portions of the genetic information, and this regulatory potential contributes in turn to the establishment and maintenance of cellular identity.

Moreover, the fraction of accessible DNA sequence remarkably coincides with more than 90% of the regions bound by TFs (Thurman et al., 2012). In fact, TFs perform a central function in chromatin ac-

cessibility, especially by directing nucleosome remodeling to specific loci (pioneer TFs), since other components of the remodelling machinery do not typically present DNA sequence specificity (Klemm et al., 2019). The internucleosomal DNA is also bound by RNA Polymerases, chromatin modifiers and architectural proteins such as insulators. Collectively, these elements compete with histones and other chromatin-binding proteins to regulate nucleosome positioning and modulate access to the DNA sequence.

Context-specific binding of transcription factors to regulatory elements

TFs’ binding to the genome is a key step in gene expression regulation. For instance, the broadly conserved transcriptional programs of vertebrate tissues are known to be governed by analogously conserved sets of tissue-specific TFs, which contribute to the establishment of characteristic functions throughout development and to ensure tissue homeostasis in adulthood (Villar et al., 2014). Nevertheless, recent studies have highlighted a number of features, including pervasive transcription, stochastic gene expression and widespread binding of TFs, which uncover a rather "leaky" nature of transcriptional regulation, in contrast to the tightly regulated essence of developmental programs (Spitz and Furlong, 2012). The reason behind this likely resides in the integrated yet intricate nature of events required for the expression of a given gene, which should be studied as a whole and not as individual phenomena (Spitz and Furlong, 2012).

A number of features take part in these events. Besides promoter regions, several other promoter-proximal and distal *cis*-regulatory elements contribute to transcriptional regulation in a coordinated manner. In this context, the role of enhancers and their associated TFs in controlling the expression of target genes is paramount yet poorly understood. What most likely hinders the reconstruction of gene regulatory networks is the combinatorial nature of TFs’ binding to enhancers, which highly varies in space and time. Enhancers typically contain clusters of distinct TF binding sites, and combinatorial binding of multiple TFs at the same region has been linked to precise transcriptional patterns in overlapping spatial domains of the developing embryo (Halfon et al., 2000, Lettice et al., 2012, Sandmann et al., 2007, Small et al., 1992, Yuh et al., 1994). Moreover, increasing evidence suggests that this binding is highly context-dependent. In

fact, the same TF can occupy different sets of enhancers in different times or conditions (Jakobsen et al., 2007, Sandmann et al., 2007, 2006), promoting recruitment of additional TFs (Mullen et al., 2011, Trompouki et al., 2011, Zeitlinger et al., 2003). Furthermore, a gene can be controlled by multiple enhancers (Osterwalder et al., 2018), as well as the same enhancer can target multiple genes (Gasperini et al., 2019), all in a context-dependent manner.

Another level of complexity is given by the fact that enhancers are not genomic regions that are simply turned on or off. For instance, the activity status of an enhancer can change during development, based on the TFs that are bound to it and the chromatin remodelling they promote, as described for the macrophage-specific *c-fms* locus, progressively silenced during B-lymphopoiesis (Krysinska et al., 2007, Tagoh et al., 2004). As anticipated, pioneer TFs can recognize DNA motifs, likely within nucleosomal DNA or at the exit of the nucleosome, that are inaccessible to other factors, and promote histone displacement by recruiting chromatin remodelers. Moreover, although not sufficient to form an activating complex, the binding of a pioneer TF to an enhancer can have a priming function, by triggering a cascade of events necessary for subsequent recruitment of additional TFs, which are sometimes required at later stages in development (Spitz and Furlong, 2012). Therefore, the function of pioneer TFs, like their binding, is highly context-dependent: while in some cases they have a more transient role limited to early nucleosome remodeling and histone modification (Hoogenkamp et al., 2009, Liber et al., 2010, Xu et al., 2009), in some other cases they exert multi-lineage priming of enhancers (Mercer et al., 2011). Because of this dual function, several pioneer TFs – such as MYOD1 (de la Serna et al., 2005), FOXA1 (Lupien et al., 2008), PU.1 (Ghisletti et al., 2010), PAX5 (McManus et al., 2011), and C/EBPβ (Siersbæk et al., 2011) – act at the top of gene regulatory networks. It has also been suggested that TF binding may prevent, within enhancer regions, methylation of cytosine (Xu et al., 2007), an epigenetic feature typically associated with genome silencing in mammals (for more details see section "DNA methylation and the silent genome").

The manifold roles of histone post-translational modifications

Although the establishment and maintenance of distinct cellular characteristics is tightly linked with the expression and binding of specific TFs, other chromatin features are also associated with cellular identity. One example is represented by ESCs, whose pluripotent capacity is associated with the action of the core OSKM factors, but which are also characterized by more condensed chromatin compared to differentiated cell types (Park et al., 2004). Among other mechanisms, post-translational modifications (PTMs) of histones – including phosphorylation, acetylation, methylation and ubiquitylation – affect the compaction of chromatin, promoting a more or less tight interaction of the DNA sequence with histone proteins. Moreover, these chemical tags can function as binding sites for other proteins, and have been implicated in the regulation of transcription initiation and elongation (see below). These modifications are typically reversible, and several enzymes are involved in their deposition and removal, including histone acetyltransferases (HATs) and deacetylases (HDACs), lysine methyltransferases (KMTs) and demethylases (KDMs), as well as ubiquitylation enzymes (E1-3) and deubiquitylases (DUBs). These enzymes typically form multisubunit complexes, which chemically modify specific residues either within the globular domains of core histones or in their amino-terminal tails.

Multiple functions have been attributed to histone modifications, from a synergistic interplay with TFs to transcriptional regulation and cell memory (Chen and Dent, 2014). As described above, the binding of pioneer TFs to closed chromatin can initiate nucleosome remodeling and trigger activation of regulatory regions. In some cases, the presence of an epigenetic pre-patterning, represented by certain combinations of histone modifications, seems to either positively or negatively correlate with pioneer factor binding. Different types of epigenetic pre-patterning have been observed in distinct processes, including differentiation and disease. For instance, in breast cancer cells FOXA1 binds to open, DNA-unmethylated regions which are enriched in H3K4me1 and H3K4me2 (Zaret and Carroll, 2011), while failure to reprogramming of human somatic cells is associated with an enrichment of H3K9me3 which impedes binding of the OSKM factors (Soufi et al., 2012). These and similar observations are suggestive of a synergistic interplay between TFs and histone modifications. Actually, some histone modifications have been reported as

predictive of TFs binding sites (Gerstein et al., 2010). In line with this, several histone-modifying enzymes belong to co-regulator complexes, which cooperate with TFs towards transcriptional regulation (Chen and Dent, 2014). As an example, during neural differentiation the TF SMAD3 recruits at distal regulatory loci the chromatin modifiers JMJD3 and CHD8, which contribute to full enhancer activation (Fueyo et al., 2018).

The fact that the disruption of the histone modification machinery leads to altered phenotypes and disease (Bungard et al., 2010, Cao and Zhang, 2004, Smith et al., 2011) has prompted the need to understand how these epigenetic marks contribute to eukaryotic gene regulation and chromosome packaging. In trying to make sense of this complexity and relate it to gene regulation, the notion of a histone code (Strahl and Allis, 2000) has become a popular paradigm in the last two decades. According to this hypothesis, specific combinations of histone marks over the genome dictate precise downstream molecular events, such as transcription initiation and elongation, splicing or silencing. Similarly, these combinations would allow the identification of gene-proximal and -distal regulatory elements at a genome-wide scale. Remarkably, the definition of "code" implies a key for a direct conversion between certain inputs (combinations of histone marks) and outputs (transcriptional events) (Henikoff and Shilatifard, 2011). Nevertheless, accumulating evidence points to the lack of an overarching conversion factor, whereby certain histone marks or combinations of them are systematically linked to a particular molecular phenotype. Instead, what has emerged in recent times is a picture of correlative patterns between certain histone marks and transcriptional outcomes, which do not necessarily translate into a relationship of causality. Further details on the relationship between gene expression and some well-studied histone modifications is provided in the section "The complex interplay between gene expression and histone modifications".

Although transcriptomic measurements are often taken as the main determinant of cell type, the identity of a cell is intimately linked to its epigenetic landscape, as predicted by early epigenetic studies. For instance, cellular reprogramming is a slow and difficult process to achieve, and large domains of H3K9me3 (a modification typically associated with heterochromatin) observed in differentiated cells are considered major barriers for their reprogramming into induced pluripotent stem cells (iPSCs) (Papp and Plath, 2013). This is

suggestive that histone modifications may play a role in the memory that the cell preserves of its lineage commitment. In further support of this, genome-wide patterns of histone modifications are largely maintained across cell divisions, and this underscores the fact that they are stably inherited despite the disruption of chromatin caused by DNA replication. Some mechanistic studies have shed light on this process. Upon eviction from DNA, (old) parental histones are kept in close proximity (Gruss1 et al., 1993, Madamba et al., 2017), and shortly after the fork passage, they appear to be re-deposited together with newly synthesized unmarked histones in a 1:1 ratio (Alabert et al., 2015, Alabert and Groth, 2012, Almouzni and Cedar, 2016, Annunziato, 2015). Recently, it has been proposed that the recycled histones preserve the parental histone modifications, and induce feedforward stimulation of modification on the adjacent, newly incorporated histones (Reverón-Gómez et al., 2018). Nevertheless, the mechanisms behind inheritance of histone modifications remain largely unexplored, and their study may disclose important principles in the establishment and maintenance of cellular identity.

DNA methylation and the silent genome

Although less common than histone modifications, methylation of cytosine DNA bases is an epigenetic feature typically associated with transcriptional silencing. The vast majority of the DNA sequence in mammals is non-coding, yet it comprises several regions with latent transcriptional potential, such as pericentromeric repeats and parasitic transposable elements. DNA hypermethylation at these regions, together with H3K9me3 deposition, has been linked to their constitutively silent transcriptional status, which is required for the stability and maintenance of the genome (in the case of transposons) as well as for proper chromosome alignment and centromeric assembly (in the case of pericentromeres) (Smith and Meissner, 2013). This epigenetic mark is characterized by rather stable patterns across tissues and throughout life, with only a few exceptions. During preimplantation development, active and passive waves of demethylation affect the paternal and maternal genomes, respectively (Inoue and Zhang, 2011, Mayer et al., 2000, Oswald et al., 2000, Santos et al., 2002, Smith et al., 2012, Wossidlo et al., 2010) with the exception of imprinted loci (Lane et al., 2003, Olek and Walter, 1997, Tremblay et al., 1997), and this global unmethylated state is not reverted until early embryonic progression. Although most CpGs in

mammalian genomes remain methylated during development, CpG islands located within promoters of several genes (either housekeeping or developmentally regulated) are constitutively hypomethylated (Smith and Meissner, 2013). Classical studies (Brandeis et al., 1994, Macleod et al., 1994) showed that TF binding negatively regulates the methylated state of promoter CpG islands, which can progressively acquire heritable methylation if depleted of known TF binding sites. Moreover, it has been suggested that histone marks and variants typically associated with transcription may be constrained at promoter loci by surrounding DNA methylation patterns (Conerly et al., 2010, Yang et al., 2012), and that spurious DNA methylation at these regions is contrasted by H3K27me3 in normal conditions (Bartke et al., 2010, Brinkman et al., 2012), but often altered in cancer and age-related diseases (Jones, 2012).

3D genome organization and chromatin hubs

The hierarchical folding of chromatin in the 3D nuclear space comprises a number of different structural entities, from low-level nucleosome-nucleosome interactions and long-range chromatin loops, to higher-level topologically associating domains (TADs) and megabase-scale compartments. The spatial positioning of genes within this complex structure correlates with their chromatin status and transcriptional activity: while heterochromatic and gene-poor regions tend to localize close to the nuclear envelope, chromosomal gene-rich segments typically locate inside the euchromatic internal part of the nucleus (Bonev and Cavalli, 2016). Nevertheless, it is unclear what are the consequences of perturbing the genome architecture, and if they are related to the transcriptional status of the involved loci. It has been shown that forcing a loop between one promoter and its distal regulatory element can trigger gene expression (Deng et al., 2012, 2014), but also that the induced relocation of genes towards the nuclear periphery or the loss of anchoring of heterochromatic regions to the nuclear lamina does not affect the transcriptional outcome of the locus (Gonzalez-Sandoval et al., 2015, Shachar et al., 2015, Therizols et al., 2014). Indeed, long-range interactions between promoters and enhancers can occur at earlier stages than gene activation (Ghavi-Helm et al., 2014), suggesting that they may set a permissive condition for gene expression without strictly accounting for RNA Polymerase recruitment and transcription initiation (Bonev and Cavalli, 2016). Because of this, the study of 3D

genome organization is lately being integrated in the context of chromatin hubs, which dictate coordinated responses to stimuli by bringing together distal regulatory regions and their target genes. The identification of chromatin hubs allows analysis of the chromatin and transcriptional status of loci close in the 3D space. In line with this, recently it has been shown that genetic variation in the population is associated with changes in chromatin modification and accessibility, as well as transcription, not only locally but also at distal ($>$ 50 Kb) but physically interacting loci (Grubert et al., 2015, Waszak et al., 2015).

Enhancer and promoter regulatory activity

The epigenetic mechanisms described so far ultimately affect the activity of specific portions of the genome, which regulate local and/or distal transcriptional outputs, and have been implicated in the response to external and developmental cues, cell and tissue homeostasis, and disease (Andersson and Sandelin, 2020). Promoters and enhancers, which initiate and amplify transcription, respectively, have been extensively studied in the last decades. Although historical definitions of these elements are dichotomic, their genome-wide identification and characterization have highlighted similar properties and functions, suggesting that conventional definitions of promoters and enhancers should be revisited.

The primary event in the expression of a gene is the selection of the RNA Polymerase II TSS (i.e. the first transcribed nucleotide of a transcript), which is assisted by the assembly of the pre-initiation complex, a multi-subunit complex comprising, besides the RNA Pol II, general transcription factors and other co-activators elements, including the Mediator complex. The core promoter, a region typically surrounding $\pm$ 50 bp the TSS, contains sufficient information to select the TSS, in the form of either specific elements such as the TATA box and the initiator element (INR), or of more flexible and degenerate sequences. Nevertheless, the rate of RNA Pol II recruitment, initiation and elongation can be influenced by other signals, such as local and distal binding of TFs and patterns of epigenetic marks. The distal input can be up to 1 Mb away, and as described in the previous section, is brought in the close proximity of the promoter by long-range interaction loops and other higher-order chromatin folding structures. These regions, called enhancers, were originally discovered as se-

quences capable of increasing the expression of a reporter gene, independently of their distance or orientation with respect to the core promoter.

Genome-wide maps across a large number of mammalian cells have highlighted key features of transcription initiation sites, including clusters of alternative promoters (Batut et al., 2013, Carninci et al., 2006, Landry et al., 2003, Zavolan et al., 2002) (whose choice may impact the final protein product, Valen et al., 2009), pausing of RNA Pol II downstream the TSS (a step that precedes active elongation over the gene body, see next section) (Core and Adelman, 2019, Core et al., 2008), and divergent antisense transcription (Core et al., 2008, Preker et al., 2008, Seila et al., 2008). Interestingly, this last feature does not appear to be exclusive of promoters, since bidirectional transcriptional products have also been observed at enhancer regions (eRNAS) (De Santa et al., 2010, Kim et al., 2010). Nevertheless, the amount of transcriptional activity at enhancers is lower compared to promoters (Andersson et al., 2014b, Core et al., 2014), and leads to RNAs that are often degraded in the nucleus (Andersson et al., 2014a). In addition to their transcriptional outputs, enhancers and promoters present distinct sequence and chromatin properties. The former are depleted of CpG islands, which instead overlap around 50% of gene promoters (Andersson et al., 2014a). Patterns of histone acetylation and methylation are also considered features that can help distinguish these two classes of elements, and even predict their activity (Creyghton et al., 2010, Heintzman et al., 2007, Robertson et al., 2008). For instance, marking by H3K27ac, combined with high H3K4me1 and low H3K4me3 signals, is considered a signal of active enhancers (for more details see section "The complex interplay between gene expression and histone modifications"). However, given the association of certain histone modifications with transcriptional activity, these differences may be reconducted to the local transcriptional output of regulatory elements, rather than to their ability to modulate distal transcription (Andersson and Sandelin, 2020). Moreover, long-range chromatin interactions between pairs of promoters have also been observed (Javierre et al., 2016, Mifsud et al., 2015, Schoenfelder et al., 2015), and several promoters have reported enhancer activity *in vitro* and *in vivo* (Dao et al., 2017, Diao et al., 2017, Engreitz et al., 2016).

Given the numerous epigenetic and transcriptional similarities between promoters and enhancers, an updated model of regulatory ele-

ments has recently been proposed (Andersson and Sandelin, 2020), in which these two types of activity are not mutually exclusive, but can rather coexist on the same locus. According to this model, a regulatory element is a segment of the genome depleted of nucleosomes and often bound by TFs, which can present both enhancer and promoter activity in different combinations, and even in a cell type- or tissue-specific manner. In this view, enhancers may also be regions of the genome that elicit distal regulation in a distance and/or orientation-specific way.

Promoter-proximal escape of RNA Polymerase II

Promoter-proximal pausing of RNA Polymerase II is a widespread phenomenon in metazoans, and represents an important regulatory step in the complex series of events leading to the expression of a gene. Early experiments with mammalian cells in the 1970s and 1980s showed that the initiation of transcription did not necessarily result in the production of a full-length transcript (Fraser et al., 1978, Gariglio et al., 1981). These results were consistent with studies of the heat shock (Hsp) genes in the fruit fly, in which a transcriptionally engaged Polymerase, associated with a short nascent RNA, was accumulating downstream of the promoter (Rasmussen and Lis, 1993, Rougvie and Lis, 1988). Roughly in the same years, an enrichment of Pol II downstream of the TSS was also observed for key regulatory mammalian genes, such as MYC and FOS (Krumm et al., 1992, Plet et al., 1995, Strobl and Eick, 1992). These observations motivated the study of promoter-proximal pausing and release of Pol II as an important regulatory step in the transcriptional cycle, contrary to traditional models of gene regulation in *S. cerevisiae* that mostly focused on the recruitment and formation of the pre-initiation complex.

Analyses of nascent RNA and Pol II occupancy have highlighted, consistently across the fruit fly, mouse and human species, that approximately 30% of the genes display Pol II promoter-proximal pausing (Core et al., 2008, Larschan et al., 2011, Min et al., 2011). Nevertheless, there is great variation in the types of genes that are paused in different cell types and conditions (Min et al., 2011, Nechaev and Adelman, 2008). Importantly, less than 1% of these genes are transcriptionally silent, suggesting that Pol II pausing does not mediate expression inactivation, but should be rather considered as a mechanism for fine-tuning the expression of transcriptionally active genes,

perhaps in response to specific signals.

Early mechanistic studies showed that, before entering a phase of productive RNA synthesis, Pol II elongates inefficiently within the first 100 nucleotides from the TSS, and its block at this level is mainly mediated by the association of two pause-inducing factors, DSIF and NELF (Kephart et al., 1992, Marshall and Price, 1992, Wada et al., 1998, Yamaguchi et al., 1999). Other factors such as GDOWN1 and TFIIF have also been involved in the stability of the paused Polymerase and the lifetime of the early elongation complex (Cheng et al., 2012). Importantly, the release of the paused Polymerase requires the phosphorylation of specific serine residues on the Pol II carboxy-terminal domain (CTD), which serves as a binding platform for chromatin modifiers and RNA-processing factors. In this sense, there is a growing body of evidence implicating histone modifications, in particular H3K4me3 and H3K9ac, in the regulation of pause release and elongation of Pol II (for more detail see section "The complex interplay between gene expression and histone modifications").

A number of putative functions have been attributed to this molecular phenomenon. Again, studies of the Hsp genes in the fruit fly have shown that a Polymerase paused at the promoter is associated with nucleosome clearance and binding of transcription factors and members of the transcription machinery (Adelman and Lis, 2012). These observations are supported by a study in which depletion of NELF leads to a significant decrease of paused Pol II at promoters and concomitant increase of nucleosome occupancy (Gilchrist et al., 2010, 2008). Pausing of the Polymerase has also been proposed as a mechanism to favour rapid gene activation in response to stimuli, since it maintains a context of open chromatin that can be rapidly accessed by co-activators (Adelman and Lis, 2012). However, this hypothesis remains rather speculative, since analyses of signal transduction networks in the fruit fly and mouse have highlighted that Pol II pausing is more frequently observed at promoters of constitutively expressed genes encoding receptors, kinases and TFs (Gilchrist et al., 2012). Although the enrichment of Pol II is typically higher at promoters compared to gene bodies (Gilchrist et al., 2010, Rahl et al., 2010), the Polymerase can slow down its elongation rate and even stop during active production of RNAs. These pauses of the Polymerase are thought to coordinate RNA processing events, such as 5' capping, splicing, and 3' cleavage and poly-adenylation. For instance, pausing of Pol II over exons is thought to promote splic-

ing (Carrillo Oesterreich et al., 2010), and lower elongation rates at exons with weak splice sites have been associated with their inclusion (De La Mata et al., 2003). Moreover, accumulation of Pol II at the 3’ end has been proposed to couple transcription termination and 3’ end cleavage (Core et al., 2008, Gromak et al., 2006, Proudfoot, 2011).

Chromatin and splicing regulation

The nascent RNA of an actively transcribed gene is tethered to its template DNA sequence by the elongating RNA Polymerase, and as introduced in the previous section, a number of processing events are coupled with transcription (i.e. they occur co-transcriptionally) (Neugebauer, 2019). Genome-wide studies have reported widespread co-transcriptional splicing across different species (budding yeast: 75%; fruit fly: 83%; human: 74%-85%), although a lower rate is observed in mouse liver (45%) (Ameur et al., 2011, Carrillo Oesterreich et al., 2010, Khodor et al., 2012, Tilgner et al., 2012). Nevertheless, not all introns are removed in a transcription-coupled manner, in particular those that are alternatively spliced (Ameur et al., 2011, Tilgner et al., 2012) and/or reside at the 3’ end of the transcript (Schmidt et al., 2011, Tilgner et al., 2012). Moreover, 20% of the activated spliceosomes are not tethered to chromatin (Girard et al., 2012), and intron retention plays a regulatory function in a number of cell types (Boutz et al., 2015, Braunschweig et al., 2014, Pandya-Jones et al., 2013, Pimentel et al., 2016, Ullrich and Guigó, 2020, Wong et al., 2013).

Since the discovery of the coupling between transcription and splicing, and that the latter is affected by Pol II elongation rates, a number of chromatin-related features, including nucleosome positioning, histone modifications and chromatin remodelers, have been investigated for their role in splicing. More than one decade ago, a number of studies reported the striking observations that nucleosome occupancy is higher in exons compared to introns independently of the transcriptional status of the gene, and that this feature is conserved across species (Andersson et al., 2009, Hon et al., 2009, Schwartz et al., 2009, Spies et al., 2009, Tilgner et al., 2009, Wilhelm et al., 2011). In line with this, computational analyses have shown that exons are enriched in high-affinity nucleosome sequences (Schwartz et al., 2009), and display higher GC content compared to introns, be-

sides being relatively short (roughly 150 bp) (Zhu et al., 2009). More specifically, nucleosomes do not localize at splice sites, but rather at the centre of exons, especially of those that have weak splicing potential (Tilgner et al., 2009), suggesting that they may act as barriers to the elongating Pol II, allowing more time for the recognition of splice sites by the splicing machinery.

While the nucleosome pattern over exons is not transcription-dependent, marking by H3K36me3 was found preferentially over expressed genes. In contrast to the epigenetic patterns that regulate TSS switching, which are common across the majority of genes, the relationship between histone marks and splicing is rather gene-specific, with the exception of H3K36me3, whose levels moderately correlate with exon inclusion rates at a genome-wide scale (Podlaha et al., 2014). Interestingly, promoter-like chromatin features have been reported in a small but well-defined subset (approximately 4%) of exons subject to tight splicing regulation across human cell types, which do not correspond to TSSs but lie close to and contact them via chromatin looping. The cell type-specific inclusion levels of these exons correlate with marking of H3K9ac, H3K27ac and H3K4me3 (Curado et al., 2015).

Are histone marks instructive for gene expression?

In what follows, we revise the role of some well-studied histone lysine modifications in the regulation of gene expression. Besides the six marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K36me3 and H3K9me3) endorsed by the reference epigenome criteria (International Human Epigenome Consortium, IHEC - `http://ihec-epigenomes.org/research/reference-epigenome-standards/`), we also focus on H3K4me2, H3K9ac and H4K20me1. In the past, in fact, distinct combinations of these histone marks have been associated with some of the gene regulatory mechanisms described above. Notwithstanding, different and sometimes contradictory observations divide the field as to whether histone marks are the cause or consequence of transcription.

H3K4me3

One of the first studies investigating the genomic distribution of different degrees of Histone 3 Lysine 4 methylation (H3K4me) is described in Santos-Rosa et al., 2002. By analyzing the methylase activity of Set1 and the effects of its depletion in *S. cerevisiae*, Santos-Rosa and colleagues found that the tri- (but not the di-) methylation of H3K4 is associated with Set1-dependent transcription. Because Set1 is the only methyltransferase in budding yeast, and given that H3K4me2 is often observed in the absence of H3K4me3, they concluded that specific mechanisms may prevent, at discrete genomic locations, the global transition from di- to tri-methylated K4. For the first time, the degree of methylation of a lysine residue was considered an indicator of transcriptional activity.

Since then, the genomic distribution of H3K4 methylation has been profiled in a range of organisms (Barski et al., 2007, Bernstein et al., 2005, Chih et al., 2005, Guenther et al., 2007, Li et al., 2008, Ram et al., 2011, Zhang et al., 2009). Despite the numerous differences in gene length and architecture, the most conserved feature across species is the H3K4me3 peak around the TSS of actively transcribed genes (Barski et al., 2007, Mikkelsen et al., 2007). Concomitant to these genome-wide analyses, a number of experimental reports began to suggest that the enrichment of H3K4me3 at promoter regions may trigger transcription initiation. In fact, DNA-binding of BPTF, a member of the NURF chromatin-remodeling complex, is stabilized by H3K4me3, and is associated with the control of developmental *H0XC8* gene expression via binding of the SNF2L ATPase (Wysocka et al., 2006). Moreover, a number of protein folds present in chromatin remodeling and histone modifying factors can recognize H3K4me3 (Ruthenburg et al., 2007). The best known example is perhaps TAF3 – a component of the general transcription factor TFIID – which can bind H3K4me3 proximal to TSSs through its PHD finger domain and mediate the recruitment of RNA Polymerase II (Lauberth et al., 2013, Vermeulen et al., 2007).

The study of methyltransferase complexes also definitely contributed to shed light on the controversial role of H3K4me3 in gene expression. In mammals, four different complexes account for the deposition of either bulk (Set1A-B/COMPASS complexes) or specific – mainly at homeotic genes (MLL1-2/COMPASS-like complexes) – H3K4me3. While recruitment of Set1/COMPASS to chromatin de-

pends on H2B monoubiquitination and occurs after the establishment of the basal transcriptional machinery, this is not the case for the MLL1-2/COMPASS complexes (reviewed in Smith et al., 2011). For these reasons, H3K4me3 has often been assigned an instructive role for the transcription of genes, further supported by the overall positive correlation between the genes' expression levels and their associated H3K4me3 promoter signals. It is important to highlight, however, that these correlation analyses are typically conducted in steady-state conditions, and can be strongly affected by the constrained nature of both transcriptional and epigenetic programs across tissues and species (Pervouchine et al., 2015).

In the last years, a number of time-series studies have explored the dynamic association between gene expression and histone marks over time. Many of them have actually reported a degree of correlation between gene expression and H3K4me3 substantially lower than what previously described in steady-state conditions, and have also questioned the causal role of H3K4me3 in triggering transcription initiation. In yeast, H3K4me3 deposition is delayed with respect to up-regulation of gene expression during meiosis, and constitutively high levels of H3K4me3 are found at double-strand breaks hotspots, independently of the local absolute level of gene expression (Borde et al., 2009). This observation for the first time uncoupled the distribution of H3K4me3 from transcriptional regulation, proposing instead a role in the initiation of meiotic recombination. Delayed or stable H3K4me3 signal at TSSs are also observed during the yeast metabolic cycle, compared to the massive transcriptional and lysine acetylation oscillations (Kuang et al., 2014).

These considerations rapidly expanded to other model organisms. In the fruit fly, gene activation during pre-midblastula transition occurs in the absence of H3K4me3 (Chen et al., 2013), and approximately 34% of the genes expressed during *Drosophila* development lack H3K4me3 at their TSSs (Nègre et al., 2011). In particular, global absence of H3K4me3 and other canonical active chromatin marks involves developmentally regulated genes (Pérez-Lluch et al., 2015). Studies of circadian, developmental and differentiation programs in mammals have also reported uncoupled patterns of gene expression and H3K4me3 over time. During the circadian cycle in the mouse liver, for instance, H3K4me3 peaks long after RNA Pol II recruitment and transcription initiation (Koike et al., 2012, Le Martelot et al., 2012), and there is a pervasiveness of circadian rhythms in

RNA Pol II binding and histone marking that is independent of RNA cycling (Koike et al., 2012) and the other way around (Le Martelot et al., 2012). Furthermore, genes that share similar expression profiles throughout cardiac differentiation in mouse can present distinct H3K4me3 temporal patterns, which often recapitulate gene-specific functions (Wamstad et al., 2012).

In combination with time-resolved studies, several experimental assays have also tried to uncover a general mechanism by which H3K4me3 is instructive for transcription activation. It is difficult, though, to genetically ablate one histone modification without conflating effects (Howe et al., 2017). Deletion of yeast *SET1*, for instance, interferes with other non-histone substrates, such as the kinetochore protein Dam1 (Zhang et al., 2005). Moreover, this and other strategies – such as the substitution of H3K4 with a non-methylatable residue – can result in the removal of all three methylated states (mono-, di- and tri-) of H3K4, besides H3K4 acetylation, further complicating the interpretation of causal/consequential effects (Howe et al., 2017). For these reasons, several studies have instead aimed to perturbe the integrity of methyltransferase complexes. One example is the deletion of *Spp1*, a component of the Set1 complex in yeast. Spp1 is thought to stabilize the binding of Set1 by recognizing H3K4me2/3 (Kirmizis et al., 2007, Shi et al., 2007). Its ablation specifically reduces gene-specific H3K4me3 but not H3K4me1/2 (Morillon et al., 2005, Schneider et al., 2005). Nevertheless, levels of steady-state or dynamically changing mRNAs and transcription are not affected (Howe et al., 2017, Lenstra et al., 2011, Margaritis et al., 2012).

Similar approaches have been applied in the fruit fly. Although lethal at the late L3 - early pupa stage (Beltran et al., 2003), removal of *Ash2* – a member of the trxG required for deposition of H3K4me3 (Beltran et al., 2007, Steward et al., 2006) – affects, in the wing imaginal discs, the expression of a minor subset of genes, which are unexpectedly either up- or down-regulated (Pérez-Lluch et al., 2011). Furthermore, substituting K4 with a non-methylatable residue does not impair transcriptional regulation and activation of major developmental pathways, which can occur in the absence of H3K4 methylation (Hödl and Basler, 2012).

The effect of the removal of H3K4me3 on gene expression has also been investigated in mammals, leading to similar observations. De-

pletion of *Cfp1*, a subunit of mammalian SET1A/B complexes, causes the loss of H3K4me3 from the promoters of actively transcribed genes and from CpG islands in mouse ESCs (Clouaire et al., 2012), which are viable but unable to differentiate (Carlone et al., 2005). This does not affect the transcription of the associated genes but, intriguingly, it is linked with a leakage of H3K4me3 to inappropriate chromatin compartments, as well as increased local gene expression (Clouaire et al., 2012). While this observation holds for most of the genome-wide H3K4me3 signal, which depends on the Set1/COMPASS complex, it cannot be extended to marking of trimethyl H3K4 at bivalent genes in ESCs, which is performed by MLL2. It has been previously proposed that MLL recruitment could precede transcription activation and even promote it (Smith et al., 2011). Nevertheless, the expression of only a few MLL2-target genes depends on its H3K4 methylation activity (Hu et al., 2017), and the rescue of gene expression upon MLL2-KO is not accompanied by an increase of H3K4me3 (Douillet et al., 2020), demonstrating that transcription at MLL2-targeted loci can also proceed without H3K4me3.

H3K9ac

One of the initial findings regarding acetylation of histone H3 was its depletion from telomeric regions compared to actively transcribed genes (Bernstein et al., 2002, Braunstein et al., 1993). In line with this, it was originally reported that acetylation neutralizes the positive charge of lysine residues, thus reducing binding of the negatively-charged DNA to histone proteins and favouring an open chromatin conformation (Simpson, 1978). Acetyl-lysine is also one of the most dynamic modifications (Barth and Imhof, 2010, Weinert et al., 2018, Zheng et al., 2013).

The genome-wide signal of H3K9ac typically coincides with transcriptionally active promoters, and positively correlates with other acetylation signals like H3K14ac and H3K27ac, as well as CpG content and H3K4me3 (Birney et al., 2007, Dunham et al., 2012, Heintzman et al., 2007, Karmodiya et al., 2012, Wang et al., 2008). In particular, the genome-wide co-occurring patterns of H3K4me3 and H3K9ac have prompted further mechanistic studies, which have shown that in a cell-free system, the presence of H3K4me3 on nucleosomes promotes acetylation on the neighbouring Lys 9 residue by recruitment of the Spt-Ada-Gcn5 acetyltransferase (SAGA) complex

(Foulds et al., 2013). Nevertheless, in HeLa cells, acetylation of Lys 9 has been reported to instruct the recruitment of the Super Elongation Complex (SEC), a function that H3K9ac can perform independently of the presence of H3K4me3 (Gates et al., 2017). In this context, H3K9ac mediates the release of stalling RNA Pol II from the promoter of a subset of genes, with H3K4me3 at promoters, instead, likely blocking the binding of specific elongation factors until deposition of the acetylation signal. Indeed, this is consistent with an independent observation in the fruit fly, according to which ablation of *Ash2* (a component of the trithorax group – trxG – required for deposition of H3K4me3) is associated with a global reduction of the paused form of RNA Pol II (Pérez-Lluch et al., 2011). Therefore, although often enriched at overlapping genomic sites, H3K9ac and H3K4me3 seem to play distinct roles in the process of gene activation. In further support of this, acetylation patterns tightly oscillate along with transcriptional changes during the yeast metabolic process, with peaks of H3K9ac appearing a few minutes before transcription, in contrast to H3K4me3, which is more stable or often delayed (Kuang et al., 2014). Similarly, H3K9ac precedes nascent transcription throughout the circadian cycle in the mouse liver (Koike et al., 2012).

Nonetheless, as in the case of H3K4me3, the role of H3K9ac during gene expression activation is far from being fully understood, since a number of studies have reported binding of functional histone deacetylases (HDACs) at the promoter of active genes (Kim et al., 2013, Wang et al., 2002, 2009), as well as a positive role of HDACs in transcription elongation (Greer et al., 2015). Furthermore, as in the case of H3K4me3 and several other histone modifications (see below), roles of H3K9ac other than in transcription regulation have been proposed, which are closer to the primordial definition of epigenetics. For instance, levels of H3K9ac decrease during ESCs differentiation (Krejčí et al., 2009), and correlate with their pluripotency and reprogramming capacity. In support of this, HDAC inhibitors can rescue the stem cell potential of a less potent murine ESC line, with a minimal impact on gene expression (Hezroni et al., 2011), and again in mouse ESCs, H3K9ac can mark developmentally regulated bivalent promoters together with H3K27me3, H3K4me3 and H3K14ac, and can also discriminate, like H3K27ac, active enhancers from inactive ones (Karmodiya et al., 2012).

H3K27ac

One of the first genome-wide surveys of acetyl-histone modifications in human cells also reported the enrichment of H3K27ac at promoters of transcriptionally active genes, including several Polycomb targets (Wang et al., 2008). Soon afterwards, the antagonistic and mutually exclusive behavior of H3K27ac and H3K27me3 at promoter regions was documented in the fruit fly, together with the notion that deposition of H3K27ac is performed by the acetyltransferase complex CBP/p300 but requires TRX (Tie et al., 2009).

The discovery that binding of p300 is predictive of enhancer activity in fetal mouse tissues (Visel et al., 2009) inspired the dissection of H3K27ac signal at distal regulatory regions. By taking into account the expression of proximal genes and preferential enhancer-RNA (eRNA) production, levels of H3K27ac were found, indeed, to be a discriminating factor of active *vs* poised enhancer regions in both murine (Creyghton et al., 2010) and human (Rada-Iglesias et al., 2011) ESCs (see below). Furthermore, a counteracting signal of H3K27me3 at poised enhancers revealed an unappreciated role of tri-methyl H3K27 at distal regions (Rada-Iglesias et al., 2011). Nevertheless, as extensively described above, these correlative studies are not sufficient to drive any statement of causality, and need to be supported by specific experimental assays.

In this regard, disruption of the CBP/p300 bromodomain has proved to suppress the deposition of H3K27ac selectively at distal enhancer regions in human cells. Although this decrease in H3K27ac signal is accompanied by a reduction in transcription at enhancer regions and enhancer-proximal genes, gene expression levels remain globally unaltered (Raisner et al., 2018). Likewise, replacement of Lys 27 by a non-modifiable residue does not affect chromatin accessibility, gene transcription, and self-renewal in mouse ESCs, proving that H3K27ac is dispensable for enhancer activity at least in these cells (Zhang et al., 2020). This is further supported by the fact that, in the absence of H3K27ac, acetylation of H4K16 (Taylor et al., 2013), as well as of the globular domain residues H3K64 and H3K122 (Pradeepa et al., 2016), can be associated with active enhancers.

Although, in the last years, the focus has been predominantly on the relationship between H3K27ac and distal regulatory activity, promoters of actively transcribed genes also display peaks of Lys 27 acety-

lation, as aforementioned. It is unclear, however, whether the signal of H3K27ac at enhancer regions is in some way related to the one at the corresponding promoter targets, since disruption of the CBP/p300 bromodomain selectively affects distal H3K27ac signals (Raisner et al., 2018). Overall, this suggests a different regulation of Lys 27 acetylation between promoter and enhancer regions, pointed out also by other studies. During prometaphase, for instance, human cell lines lose most of the genomic H3K27ac signal, which is recovered in anaphase/telophase but markedly more at promoter regions, and in concomitance with the reactivation of gene expression (Kang et al., 2020). In light of this, the function of H3K27ac signal at promoter regions still remains elusive. Recently, YEATS2 has been proposed as a histone H3K27ac reader that promotes the GCN5/PCAF-mediated deposition of H3K9ac at promoters (Mi et al., 2017). While mechanistically linking the marking of H3K27ac and H3K9ac, these results demonstrate the importance of genome-wide integrative analyses of multiple histone modifications for a better understanding of molecular events.

H3K4me1

Although the identification of active distal regulatory regions has been lately driven by the H3K27ac signal, initial efforts to map enhancers genome-wide focused on the localization of the histone mark H3K4me1 (Birney et al., 2007, Ghisletti et al., 2010, Heintzman et al., 2009, Kim et al., 2010). Genomic loci enriched in mono-methyl H3K4, as well as DNaseI hypersensitive sites, show a stronger cell type-specific distribution across the genome, compared to the more stable chromatin states at promoters and insulators, and overall correlate with cell type-specific gene expression programs (Heintzman et al., 2009, Xi et al., 2007). Although these regions typically display regulatory activity when tested in reporter assays (Heintzman et al., 2009), it was later demonstrated that H3K4me1 can mark both active and poised or predetermined enhancers, while H3K27ac is selectively enriched on the active ones (Creyghton et al., 2010, Rada-Iglesias et al., 2011).

Interestingly, the priming of lymphoid and myeloid specific enhancers by H3K4me1 occurs before lineage commitment (Mercer et al., 2011), and contributes to an epigenetic landscape that orchestrates transcriptional events in the differentiated cell. However, under spe-

cific stimuli, selection of *de novo* enhancers can happen at distal regions that do not display occupancy of TFs or the canonical regulatory chromatin state in basal conditions, and is associated with cellular plastic changes (Kaikkonen et al., 2013, Ostuni et al., 2013). Of note, deposition of mono- and di-methyl H3K4 occurs after the production of eRNAs at these newly active enhancers, and persists upon washout of the stimulus (Ostuni et al., 2013) and in the absence of enhancer transcription (Kaikkonen et al., 2013). It has been proposed that this marking at *de novo* enhancers provides an epigenomic memory of the initial stimulation, and that the deposition of H3K4me1 and H3K4me2 at pre-existing enhancer regions may also depend, originally, on transcription (Kaikkonen et al., 2013).

H3K4me1, like other modifications, can serve as a docking site for chromatin remodelers and other reader proteins to execute downstream molecular functions. Recently, multiple components of the transcriptional regulatory machinery, including the BAF complex, were reported to bind H3K4me1 at putative enhancer regions (Local et al., 2018). MLL3 and MLL4 account for most of the H3K4me1 signal at distal regulatory regions, and depletion of these enzymes results in highly reduced H3K4me1 levels, as well as decreased binding of BAF and other H3K4me1-associated proteins to enhancers (Local et al., 2018). Nevertheless, as pointed out in Rada-Iglesias, 2018, this study does not directly test whether the presence of H3K4me1 marking has itself a causative role on the ability of enhancers to control the expression of their target genes. By separating the catalytic activity-dependent and -independent functions of Mll3/4, it has been shown that H3K4me1 at enhancers is largely dispensable for eRNA production and expression of the target genes in murine ESCs (Dorighi et al., 2017), consistent with previous time-resolved studies (Kaikkonen et al., 2013). Similar observations have also been reported during fruit fly (Rickels et al., 2017) and mouse (Cao et al., 2018) development. In contrast, loss of Mll3/4 proteins leads to compromised enhancer activity and downregulation of target genes, suggesting that the long-range coactivator function of Mll3/4 does not reside in its methyltransferase activity (Dorighi et al., 2017).

As for H3K27ac, the presence of H3K4me1 at promoters has been investigated in a comparative minor way. The first genome-wide maps of H3K4 methylation across different species reported that the flanking regions of actively transcribed TSSs – which usually display a narrow H3K4me3 peak – are typically marked by H3K4me1 and

H3K4me2 (Barski et al., 2007, Ernst and Kellis, 2010, Ernst et al., 2011). It is known that the enzymatic activity of histone H3K4 methyltransferase is characterized by a high degree of specificity not only towards the substrate residue but also towards the degree of methylation (mono-, di- or tri-methyl state) (Cenik and Shilatifard, 2020). Nevertheless, while H3K4me1 marking does not necessarily convert into higher methyl states, it is not clear whether it serves as a platform for deposition of H3K4me2/3 at promoters, or these latter marks can instead be deposited on an unmethylated H3K4 substrate. During cardiac differentiation, cardiovascular genes gain H3K4me1 marking at promoter regions prior to transcriptional activation and deposition of H3K4me3 (Wamstad et al., 2012), and during myogenesis, muscle genes that become activated lose MLL3 occupancy and their promoters switch from H3K4me1 to H3K4me3 marking by recruitment of a distinct COMPASS complex (Cheng et al., 2014). Nonetheless, this gradient of increasing methyl states is not observed at enhancer regions. In this regard, it has been proposed that the different H3K4me3-to-H3K4me1 signal ratio observed between enhancers and promoters correlates with their different recruitment rate of RNA Pol II (Andersson et al., 2015, Core et al., 2014).

H3K4me2

As previously mentioned, a preferential association of active coding regions with the tri- but not the di-methylated form of H3K4 was originally reported in yeast (Santos-Rosa et al., 2002). The persistence of H3K4me2 at coding regions for a long time after gene expression down-regulation has been interpreted as a molecular signature of previous transcription (Ng et al., 2003). Although less frequently profiled than the mono- and tri-methylation of H3K4, the genomic distribution of H3K4me2 has been well characterized in a number of reprogramming studies (Koche et al., 2011, Sardina et al., 2018, Shao et al., 2008), and proposed as an epigenetic feature of both promoter and enhancer activity (Koche et al., 2011). In fact, the early stages in the reprogramming of murine fibroblasts – upon ectopic expression of OSKM factors – are characterized by a burst of H3K4me2 deposition, which is not accompanied by global changes neither in the repressive H3K27me3 mark, nor in gene expression and in the transcriptionally-related H3K36me3. Importantly, this result has been interpreted as an early shift in cell identity which is not reflected at the transcriptome level (Koche et al., 2011). These H3K4me2-gaining loci often corre-

spond to the promoters of pluripotency and developmentally regulated genes, such as *Sall4*, *Lin28*, and *Fgf4*, which become activated later on during the formation of iPSCs. Interestingly, *de novo* gain of H3K4me2 at promoter regions is not accompanied by other histone H3K4 methylation signals (me1/3). In contrast to the general gain of promoter marking by H3K4me2, distal intragenic loci (enhancers) either gain or lose H3K4me2, suggesting distinct regulation of this epigenetic mark at promoters and enhancers. An increase of H3K4me2 is also observed, concomitantly to active DNA demethylation and prior to chromatin opening, at roughly 3% of candidate regulatory loci during reprogramming of mouse pre-B cells (Sardina et al., 2018), and deposition of H3K4me2 together with H3K27ac has also been reported at enhancers that become activated during neural differentiation (Fueyo et al., 2018). During T-cell differentiation, H3K4me2 co-localizes with binding of master regulators PU.1 and GATA-3, and H3K4me2 marking at promoters in the absence of histone 3 acetylation characterizes either newly silenced genes, or genes primed for activation in subsequent stages of differentiation (Zhang et al., 2012). This is indeed similar to what has been reported for H3K4me1 during cardiac differentiation (Wamstad et al., 2012), and the opposite to what has been described for H3K4me3 in a number of studies (see above). Collectively, there are dispersed findings suggesting that deposition of H3K4me1 and H3K4me2 at promoters often anticipates gene activation, while marking in H3K4me3 follows it. In line with this, it has been proposed that in yeast, transitions from H3K4me1 to H3K4me2 to H3K4me3 at promoter regions occur in a stepwise manner over multiple rounds of transcription, according to a "time" model that postulates that the H3K4me level is a function of the time Set1 is tethered on a nucleosome (Soares et al., 2017). In this sense, the three marks may be considered as being one the prerequisite for the other, instead of unrelated epigenetic features.

H3K36me3

Differently from the histone modifications analyzed so far, which show a preferential location at the 5' end of the genes, H3K36me3 is typically found at the gene 3' end (Bannister et al., 2005), with an enrichment in expressed exons compared to introns (Kolasinska-Zwierz et al., 2009), and higher marking in intron-containing compared to intron-less genes (De Almeida et al., 2011). Because of this, in the last decade several reports have suggested a role of H3K36me3

in interfacing the elongating RNA Polymerase and regulating RNA processing events. For instance, it was proposed that H3K36me3 acts as a binding site for the chromatin reader MRG15, which in turn recruits the splicing regulator PTB and favours inclusion of alternatively spliced exons, such as exon IIIc in gene *FGFR2* (Luco et al., 2010). Consistently, overexpression and depletion of *SETD2*, the main effector of H3K36me3 marking, lead to higher and lower inclusion rates, respectively, of these exons (Luco et al., 2010). Similar observations have also been made for Psip1/p52, a protein that recognizes H3K36me3 and mediates recruitment of Srsf1 and other splicing factors (Pradeepa et al., 2012). An alternative hypothesis is that H3K36me3 marking over genes is a consequence rather than an instruction for splicing. In support of this, de Almeida and colleagues have demonstrated that splicing inhibition leads to a loss of Setd2 recruitment and H3K36me3 deposition with no effect on transcription, and that forcing the inclusion of alternative exons has the opposite effect, i.e. it enhances Setd2 binding and H3K36me3 deposition (De Almeida et al., 2011). Other experiments have also reported a role of this modification in X-chromosome dosage compensation (by recruiting the fruit fly MSL complex, Larschan et al., 2007), DNA damage response (reviewed in Sun et al., 2020), and 3D chromosome organization (Evans et al., 2016, Smith et al., 2013, Ulianov et al., 2016).

While postulating a role of H3K36me3 in splicing regulation, the studies above do not address the role of the histone modification by itself, but rather the function of the depositing enzyme. The generation of H3K36R *Drosophila* mutants has shown that H3K36me3 is actually dispensable for alternative splice site choice and efficient removal of canonical introns, and its depletion is associated with an increase of H4 acetylation (Meers et al., 2017), as previously reported in budding yeast (Carrozza et al., 2005, Keogh et al., 2005). A widespread dysregulation of the transcriptome is observed in the absence of H3K36me3, with genes least expressed in the wild-type showing the largest increases in expression, and the other way around. Interestingly, differences in nuclear *vs* PolyA+ RNA-seq expression of these genes upon depletion of H3K36me3 were linked to defects in 3' end formation and polyadenylation (but not to PolyA site choice), suggesting a post-transcriptional role of this mark in transcript fitness (Meers et al., 2017).

H4K20me1

The role of H4K20me1 in transcriptional regulation is unclear and to some extent contradictory. This is mainly due to the lack of studies investigating this histone modification in depth (only 140 Pubmed publications contain the word "H4K20me1" in their title and/or abstract, in contrast to the over 2,000 results that can be retrieved for H3K4me3 – as of December 2020), but species- and context-specific functions should also be taken into account (Beck et al., 2012). Biochemical and non genome-wide experiments have suggested a role of this modification in transcription repression. Loss of PR-Set7, which exclusively accounts for H4K20 mono-methylation, causes a cell cycle arrest in G2/M phase and a general decrease in chromosome compaction (Houston et al., 2008, Oda et al., 2009). Besides, H4K20me1 is required for higher degrees of H4K20 methylation (Schotta et al., 2008) (with H4K20me3 considered a hallmark of silenced heterochromatic regions – Jørgensen et al., 2013), and is recognized by the Polycomb group protein L3MBTL1 (Trojer and Reinberg, 2007). In the fruit fly, H4K20me1 is found in condensed regions on polytene chromosomes, where it antagonizes the acetylation of H4K16 in vitro (Nishioka et al., 2002), while mouse cells show an enrichment of H4K20me1 on the inactive X chromosome (Kohlmaier et al., 2004).

On the other hand, analyses at specific loci in both mouse and human genomes highlighted positive correlation between H4K20me1 marking and gene expression (Talasz et al., 2005, Vakoc et al., 2006, Wakabayashi et al., 2009), which have been confirmed by genome-wide studies (Barski et al., 2007, Cui et al., 2009). H4K20me1 is typically enriched over the gene body of actively transcribed genes, and it is highly correlated with gene expression (Wang et al., 2008). Besides, it has been reported that H3K27me3 and Polycomb factors, contrary to their canonical repressive role in transcriptional regulation, are associated with active expression of a subset of genes marked by H4K20me1 in *Drosophila* wing discs (Lv et al., 2016).

Besides its role in mitotic condensation and gene expression, H4K20me1 seems to be required for DNA repair and genome integrity (Beck et al., 2012). Indeed, very recent evidence shows that reduction of H4K20me1 induced either by inhibition of PR-SET7, or overexpression of H4K20M, blocks murine embryos at the one- or two-cell stage, respectively, and causes the accumulation of DNA

double-strand breaks, uncovering a key role of H4K20me1 in preimplantation development (Shikata et al., 2020).

H3K27me3

Enriched over the inactive X chromosome of placental mammals (Erhardt et al., 2003, Plath et al., 2003, Silva et al., 2003), H3K27me3 is considered an epigenetic feature of transcriptionally silent loci, and it is deposited by the Polycomb Repressive Complex 2 (PRC2). PRC2 works in conjunction with the PRC1 complex, which recognizes H3K27me3 and promotes ubiquitination of H2AK119 (Cao et al., 2005, Wang et al., 2004). On the other hand, H2AK119ub can recruit PRC2, favouring the propagation of H3K27me3 over nearby nucleosomes (Blackledge et al., 2014, 2020, Cooper et al., 2014, Tamburri et al., 2020), but PRC2 methyltransferase activity is also stimulated by a feed-forward mechanism via binding of H3K27me3 (Margueron et al., 2009, Poepsel et al., 2018). The role of Polycomb (PcG) proteins was first described by developmental studies in *Drosophila*, which showed that the balanced, counteracting activity of Trithorax/COMPASS (which account for canonically active H3K4 methylation) and PcG complexes is required for proper expression of homeotic genes, thus ensuring correct development of anatomical structures (reviewed in Piunti and Shilatifard, 2016). In line with this antagonistic behavior of Trx and PcG, promoters of key developmental TFs were found to be marked by both H3K4me3 and H3K27me3 in mouse ESCs (Bernstein et al., 2006). This bivalency, i.e. the presence of both active and repressive modifications at the same locus, is thought to poise silent genes for later activation, and progressively disappears during differentiation, with active and repressive genes losing H3K27me3 and H3K4me3, respectively. Similar results also apply to human ESCs (Pan et al., 2007, Zhao et al., 2007), with bivalent promoters often displaying enrichment of paused Pol II downstream the TSS (Ferrai et al., 2017, Levine, 2011). In both murine and human ESCs, sequential ChIP experiments have shown that this bivalent nature is not due to a heterogeneous population of cells, carrying either H3K4me3 or H3K27me3, but that these two marks actually coexist within the same nucleosome (Bernstein et al., 2006, Pan et al., 2007, Voigt et al., 2012). On the contrary, there is no evidence in the fruit fly of a real co-existence of H3K4me3 and H3K27me3 (Voigt et al., 2013), suggesting a species-specific role of this epigenetic feature. Moreover, H3K27me3 marking has been described,

together with H3K4me1, at poised enhancers, likely contrasting the deposition of H3K27ac (Rada-Iglesias et al., 2011, Zentner et al., 2011). Bivalent domains are not an exclusive feature of ESCs, since they have been described also in neural progenitor cells, mouse embryonic fibroblasts, CD14+ T cells and hematopoietic stem cells (reviewed in Blanco et al., 2020).

Nevertheless, bivalency appears to be restricted to the promoters of key developmental TFs. For instance, while cardiac TFs and members of key signaling pathways (such as TGFβ family, Wnt and Notch) show reciprocal regulation of H3K4me3 and H3K27me3 during cardiac differentiation, cardiac-specific genes, such as those encoding cardiomyocyte contractile proteins, are not marked by H3K27me3 at any time, despite being tightly regulated during differentiation (Paige et al., 2012). Nor is it the case that H3K27me3 is always associated with a poised or silent state, since in this model genes involved in mesodermal differentiation are actively transcribed and constitutively H3K27me3-marked. In line with this, positive regulation of gene expression by PcG has also been reported in other studies (Jacob et al., 2008, Pasini et al., 2007, Schaaf et al., 2013), and has been recently associated with H4K20me1 marking (Lv et al., 2016, see above).

As for other histone marks, histone gene replacement experiments have provided insight into the mechanistic role of H3K27me3, but differently from other marks, substitution of Lys 27 with a non-methylatable residue phenocopies the homeotic dysregulation observed upon mutation of its catalyzing enzyme PRC2, suggesting that H3K27me3 is indeed required for correct development in both the fruit fly and in mammals (Lavarone et al., 2019, Leatham-Jensen et al., 2019, McKay et al., 2015, Pengelly et al., 2013).

H3K9me3

In contrast to H3K27me3, which typically marks developmentally- and cell type-specific silent loci, H3K9me3 is a feature of constitutively heterochromatic regions, such as transposable elements (TE), centromeres and telomeres, and its association with gene repression (via recruitment of HP1 proteins) was first documented by the study of position-effect variegation (PEV, reviewed in Elgin and Reuter, 2013). KRAB-containing zinc finger proteins and short RNAs coupled to Ago members are responsible of targeting H3K9 methyltrans-

ferases to specific genomic loci, and depletion of both guiding and effector elements leads to expression of TE as well as normally silenced host genes (Ninova et al., 2019). H3K9me3 deposition depends on H3K9me2, and its accumulation at promoters of cell cycle genes impairs proliferation of neural progenitors (Pappa et al., 2019). H3K9me3 has also been observed at bivalent promoters in trophoblast, extraembryonic endoderm stem cells and cultured ESCs (Voigt et al., 2013).

Nevertheless, as for H3K27me3, marking by H3K9me3 is not always coupled with transcriptional repression. Enrichment of H3K9me3 towards the 3’ end is observed in heterochromatin-residing and active genes in the fruit fly, with loss of H3K9me3 followed by down-regulation of these genes (Ninova et al., 2020, Riddle et al., 2011). H3K9me3 also marks, in combination with H3K36me3, the 3’ exons of zinc finger genes (Blahnik et al., 2011). In contrast, enrichment of H3K9me3 at the TSS of actively transcribed genes was early reported only in cancer cells (Wiencke et al., 2008). A very recent study has re-evaluated the role of this modification at promoter regions, as well as its context- and enzymatic-dependent deposition during development (Burton et al., 2020). In the earliest stages of mouse development, SUV39H2-driven deposition of H3K9me3 at the TSS was actually found compatible with active gene expression, especially in the case of genes *de novo* marked in the paternal genome at the zygotic stage (Burton et al., 2020). On the other hand, embryos with forced expression of *Suv39h1wt* (but not of *Suv39h1mut*, nor *Suv39h2*) do not down-regulate two-cell stage-specific genes, accumulate precocious constitutive heterochromatin and fail to reach the blastocyst stage. Overall, this suggests a distinct role of early and late appearance of heterochromatic H3K9me3 during development, and highlights an unappreciated non-repressive function of this histone mark at promoter regions.

CHAPTER 1

A general framework to understand the relationship between expression and histone marks

To investigate the relationship between gene expression and histone marks from a temporal perspective, we have monitored the transcriptome and the epigenome of human pre-B cells transdifferentiating into macrophages. Analysis of these data reveals that the large associations between gene expression and chromatin marking previously reported in steady-state conditions are partially artifactual, mostly because of the constrained nature of the transcriptome and the epigenome. Furthermore, modeling of the transdifferentiation process shows that only a limited number of chromatin states actually mark the genes, in contrast to the highly combinatorial behavior of histone marks previously described. Genes tend to remain in the same chromatin state throughout the entire transdifferentiation process, even those that undergo substantial changes in gene expression. We have also observed chromatin changes that are not necessarily accompanied by changes in gene expression, suggesting that the contribution of epigenetic modifications to cell fate transitions cannot be fully recapituted by transcriptomic profiles. We report, however, a strong association between chromatin marking and expression at the time of initial gene activation. We have been able to determine the precise order of histone marks' deposition at that time, and found that only H3K4me1 and H3K4me2 appear to be deposited prior to gene activation. Further changes in gene expression, comparable or even stronger than those at initial gene activation, are instead mostly uncoupled from chromatin changes.

Borsari B., Abad A., Klein C.K., Nurtdinov R., Esteban A., Palumbo E., Ruiz-Romero M., Sanz M., Correa B.R., Johnson R., Pérez-Lluch, S. and Guigó R. (2020). Dynamics of gene expression and chromatin marking during cell state transition.

Submitted. Available on *bioRxiv*: `https://doi.org/10.1101/2020.11.20.391524`

Dynamics of gene expression and chromatin marking during cell state transition

Beatrice Borsari[1], Amaya Abad[1], Cecilia C. Klein[1,2,3], Ramil Nurtdinov[1], Alexandre Esteban[1,4], Emilio Palumbo[1], Marina Ruiz-Romero[1], María Sanz[1,5], Bruna R. Correa[1], Rory Johnson[1,6], Sílvia Pérez-Lluch[1,*] and Roderic Guigó[1,7,*]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona 08003, Catalonia, Spain
[2]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia and Institut de Biomedicina (IBUB), Universitat de Barcelona, Barcelona 08028, Catalonia, Spain
[3]Present address: Clarivate, Barcelona 08025, Catalonia, Spain
[4]Present address: “la Caixa” Foundation, Department of Research and Innovation, Barcelona 08028, Catalonia, Spain
[5]Present address: Universidad Camilo José Cela (UCJC), Madrid 28692, Spain
[6]Present address: Department of Medical Oncology, Inselspital, University Hospital and University of Bern, Bern 3010, Switzerland. School of Biology & Environmental Science, University College Dublin (UCD), Dublin 4, Ireland.
[7]Universitat Pompeu Fabra (UPF), Barcelona 08003, Catalonia, Spain
[*]Correspondence: silvia.perez@crg.cat (S.P-L.), roderic.guigo@crg.cat (R.G.)

Abstract

The role of histone modifications in the regulation of gene expression remains controversial. This is partially due to the lack of deep transcriptomics and epigenomics data on a temporal scale, so that hypotheses of causality can be properly assessed. Here, we have generated such data at twelve time points during the transdifferentiation of pre-B cells into macrophages. Modeling of the transdifferentiation process highlights selective combinations of histone marks which, over time, behave in a coordinated manner, defining the major chromatin states in which genes can be found. Changes in gene expression and histone marks strongly correlate only at the time of initial gene activation. At this time we observe a precise order of events, with gene expression activation preceded by H3K4me1 and H3K4me2, and followed by other canonically active marks. Subsequent changes in gene expression, comparable or even stronger, are instead uncoupled from chromatin changes.

Introduction

Chromatin is the complex of DNA, histone and non-histone proteins that constitutes the chromosomes found in the nucleus of eukaryotic cells. Post-translational modifications (PTMs) of histone proteins, together with other epigenetic features, can alter the overall chromatin structure and are thought to play a critical role in the regulation of all DNA-based processes. In particular, interest has grown in understanding the relationship between chromatin and transcriptional regulation.

Several histone marks have been associated with either active or silent gene expression. For instance, high levels of H3K27ac and H3K4me1 are considered a feature of active transcriptional enhancers[1], whereas active promoters are typically marked by H3K4me3[2,3]. Conversely, features of constitutive and facultative heterochromatin correlate with levels of H3K9me3 and H3K27me3, respectively[4,5]. This is strongly suggestive of an association between gene expression and chromatin modifications. According to the histone code hypothesis[6], distinct combinations of histone modifications over regulatory regions — associated with specific arrangement of transcription factors — confer to each gene a unique temporal and spatial transcriptional program. Based on this hypothesis, methods to predict gene expression from combinations of different histone marks have been developed with great accuracy, even when the predictions are obtained in a cell type other than the one in which the model is inferred[7,8].

The majority of these predictions are conducted in steady-state conditions, and therefore do not track the association between gene expression and histone marks over time. Studies along time, however, are essential to decipher the mechanisms behind transcriptional control and maintenance, since an appropriate balance of stability and dynamics in epigenetic features seems to be required for accurate gene expression[9]. Interestingly, a number of studies in different species and biological models have highlighted a degree of correlation between gene expression and chromatin marks over time substantially lower than what previously described in steady-state conditions. For instance, during fruit fly development, around 34% of the expressed genes lack H3K4me3 at their promoters[10], while transcription can occur in the absence of most active marks[11,12]. It has also been reported that, upon stimulation, changes in gene expression are not always accompanied by changes in histone modifications[13], and that chromatin marks do not represent linear measures of transcriptional activity[14,15]. Overall, it has been suggested that the contribution of chromatin to gene expression depends largely on the promoter architecture of genes[16].

Time-series studies have also striven to elucidate the temporal ordering in which transcription factor (TF) binding, deposition of histone marks and RNA Polymerase recruitment occur at both enhancer and promoter regions. For instance, it has been reported that enhancers required for hematopoietic differentiation are already primed with H3K4me1 in multipotent progenitors[17]. However, *de novo* enhancers' transcription seems to precede local deposition of H3K4me1 and H3K4me2 marks[18]. Furthermore, deposition of H3K4me1 is dispensable for either enhancer or promoter transcription, and does not affect the maintenance of transcriptional programs[19,20].

Nevertheless, most time-series studies so far have monitored a few histone modifications in a limited number of time-points. To address these limitations, here we have generated gene expression profiles and

maps of nine histone modifications at twelve time-points along a controlled cellular differentiation process: the induced transdifferentiation of human BLaER1 cells into macrophages[21]. BLaER1 is a human B-cell precursor leukemia cell line, stably transfected with a construct containing cEBPα fused with the estrogen hormone receptor binding domain[21]. These cells are able to transdifferentiate into functional macrophages at a high efficiency rate upon induction with beta-estradiol, which induces the internalization of the transcription factor into the nucleus, promoting massive transcriptomic changes. We believe that the data that we have generated constitutes an unprecedented resource in the field of time-series transcriptional and chromatin studies.

Analysis of these data reveals that the large steady-state associations between gene expression and chromatin marking previously reported are partially artifactual, and mainly arise from the constrained nature of the transcriptome and the epigenome. We found that only a limited number of combinations of histone modifications are actually marking the genes, defining the major genic chromatin states in the human genome. Genes tend to remain in the same state throughout the entire transdifferentiation process, even those that change expression substantially. We have also observed substantial chromatin changes that are not necessarily accompanied by changes in gene expression, suggesting that epigenetic modifications contribute to cell state in a manner that cannot be fully recapituted by gene expression. We did find, however, a strong association between chromatin marking and expression at the time of initial gene activation. We have been able to determine the precise order of histone modifications at that time, and found that only H3K4me1 and H3K4me2 appear to be deposited prior to gene activation. Further changes in gene expression, comparable or even stronger than those at gene activation, seem to be mostly uncoupled from changes in histone modifications.

Results

A rich resource for time-series analysis of chromatin and gene expression dynamics

To investigate the temporal interplay between transcriptional activity and chromatin marking during the transdifferentiation of BlaER1 cells into macrophages[21], we monitored this process at 12 time-points, from 0 to 168 hours post-induction (p.i.) (Figure 1a). Reciprocal regulation of B-cell and macrophage antigens CD19 and Mac-1, respectively, was assessed by flow cytometry throughout the process (Supplementary Figure 1a). For each time-point we characterized, in two biological replicates, the whole cell RNA-seq gene expression profiles and the ChIP-seq maps of nine histone post-translational modifications. Besides the six marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K36me3 and H3K9me3) endorsed by the reference epigenome criteria (International Human Epigenome Consortium, `http://ihec-epigenomes.org/research/reference-epigenome-standards/`), we have profiled H3K4me2, H3K9ac and H4K20me1 (Figure 1b). This has allowed us to characterize the interchange between different degrees of lysine four methylation over time, but also to compare acetylation patterns on distinct lysine residues, and to explore the alternation of broad marks over actively transcribed gene bodies. In addition, we have

generated, for each time-point, ChIP-seq profiles of the transcription factor cEBPα, RNA-seq data from the cytosol and the nucleus, as well as riboprofiling and proteomics maps (Correa *et al.*, in preparation).

To avoid any bias due to differences in the transdifferentiation process between experiments, a crucial component of our experimental design is that the RNA and the chromatin to perform immunoprecipitations with all histone marks were obtained from the same pool of cells in each biological replicate (see Methods). To efficiently and reproducibly analyze the wealth of data generated in a controlled environment, we developed *ChIP-nf* (`https://github.com/guigolab/chip-nf`), a pipeline implemented in NextFlow[22] (see Methods).

Gene expression recapitulates transdifferentiation more accurately than chromatin

To characterize gene expression and histone modifications' profiles during the pre-B cell transdifferentiation process, we selected 12,248 genes — out of 19,831 protein-coding genes annotated in Gencode[23] version 24 — that were either expressed in at least one time-point ($\geq$ 5 TPM, 10,696 genes), or silent all along the process (0 TPM in all time-points, 1,552 genes) (Supplementary Figure 1b). Within expressed genes, we identified 8,030 genes characterized by significant changes in their expression profiles over time (differentially expressed, DE; Supplementary Figure 1b; see Methods). Half of these genes are down-regulated during the process, 25% are up-regulated, and for the remaining 25% we observed transient increases (peaking) and decreases (bending) in expression. 2,666 expressed genes do not display changes in expression over time (stably expressed).

For every gene in these sets, we also computed the level of each histone modification at a specific time-point, either over the gene body in the case of H3K36me3 and H4K20me1, or at promoter regions (± 2 Kb with respect to the transcription start site) for the remaining marks (Supplementary Figure 1c, see Methods). Roughly all expressed genes are marked by the canonical active histone modifications, whereas the proportion of silent genes showing peaks of these marks is low, except for H3K4me1 and H3K4me2 (Supplementary Table 1). Unexpectedly, marks typically associated with silent transcription (H3K9me3 and H3K27me3) are not abundant in either expressed or silent genes.

To visually summarize the gene expression and individual histone modification profiles during transdifferentiation, we performed Principal Component Analysis (PCA), in which we plotted the 12 time-points based on these profiles (Figure 1c). Even though the PCA was performed jointly on gene expression and all chromatin marks — which show different patterns of variation —, the first two principal components (PC1 and PC2) still capture about one fifth of the total variance of the data. Whereas gene expression is able to recapitulate the process in the space of the first two principal components, the chromatin marks are less resolutive, with H3K27ac, H3K9ac and H4K20me1 showing the clearest trends. The trajectory of gene expression in the PCA space suggests that the process occurs in two different transcriptional phases, with PC1 explaining the main differences between pre-B cells and macrophages, and PC2 representing early transcriptional changes within the first 24 hours of transdifferentiation. Instead, for several chromatin marks we observed parabolic trajectories, with PC2 mainly separating the intermediate stages of transdif-

Figure 1

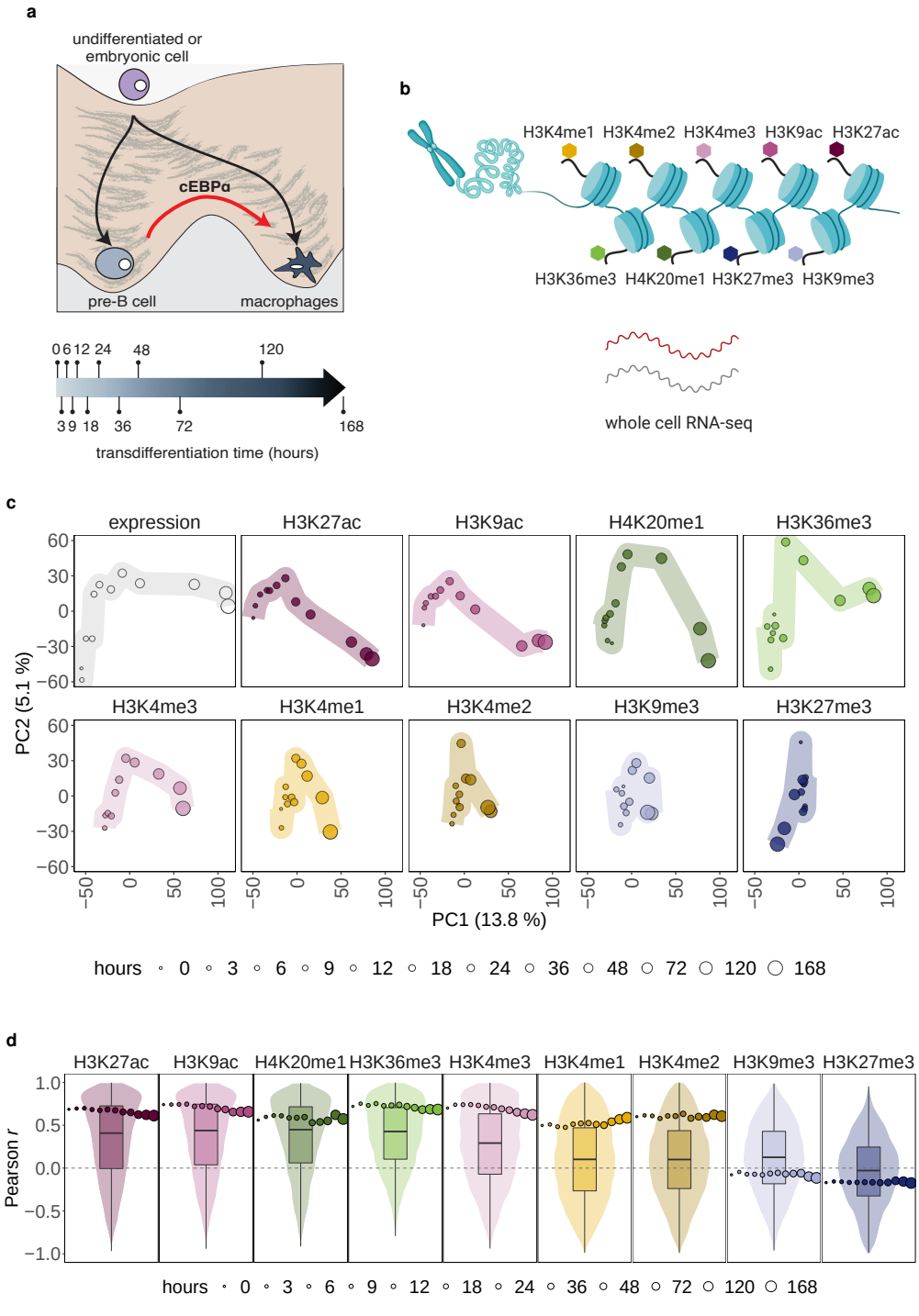

a
undifferentiated or
embryonic cell
cEBPα
pre-B cell
macrophages
0 6 12 24 48 120
3 9 18 36 72 168
transdifferentiation time (hours)
b
H3K4me1 H3K4me2 H3K4me3 H3K9ac H3K27ac
H3K36me3 H4K20me1 H3K27me3 H3K9me3
whole cell RNA-seq
c
expression
H3K27ac
H3K9ac
H4K20me1
H3K36me3
H3K4me3
H3K4me1
H3K4me2
H3K9me3
H3K27me3
PC2 (5.1 %)
PC1 (13.8 %)
hours 0 3 6 9 12 18 24 36 48 72 120 168
d
H3K27ac
H3K9ac
H4K20me1
H3K36me3
H3K4me3
H3K4me1
H3K4me2
H3K9me3
H3K27me3
Pearson r
hours 0 3 6 9 12 18 24 36 48 72 120 168

Figure 1: Global behaviour and relationship between chromatin and expression during transdifferentiation — See also Supplementary Figures 1-2,6; Supplementary Tables 1-2. **a**: The transdifferentiation of human pre-B cells into macrophages lasts a period of seven days, which we monitored at twelve time-points. **b**: We have performed ChIP-seq of nine histone modifications and RNA-seq in whole-cell fraction, at twelve time-points along the process of transdifferentiation. All experiments were performed in two biological replicates. **c**: Trajectories of transdifferentiation derived from a Principal Component Analysis performed jointly on time-series gene expression and chromatin marks' profiles. **d**: Correlations between levels of gene expression and histone marks. For a given mark and for each of the twelve time-points, we computed the steady-state Pearson *r* value between the vector of expression levels and the vector of chromatin signals corresponding to the 12,248 genes. These twelve correlation values are represented by single dots, the size of the dot being proportional to the hours of the corresponding time-point. The median Pearson *r* values for each mark are: H3K27ac: 0.67; H3K9ac: 0.72; H4K20me1: 0.59; H3K36me3: 0.72; H3K4me3: 0.70; H3K4me1: 0.51; H3K4me2: 0.61; H3K9me3: -0.07; H3K27me3: -0.17. In the case of time-course correlations, we obtained a Pearson *r* value for each expressed gene, and the distributions for all genes are represented by violin and box plots. Median Pearson *r* values across genes for each mark are: H3K27ac: 0.41; H3K9ac: 0.44; H4K20me1: 0.45; H3K36me3: 0.43; H3K4me3: 0.29; H3K4me1: 0.10; H3K4me2: 0.10; H3K9me3: 0.13; H3K27me3: -0.03.

ferentiation from the differentiated cell types. Genes contributing to PC1 are mostly up- or down-regulated (Supplementary Figure 1d), and display significant enrichment in Gene Ontology terms associated with immune response and cell motility (Supplementary Table 2). Instead, PC2-contributing genes perform functions related to nucleic acids metabolism and protein modification (Supplementary Table 2), and comprise a large proportion of genes either displaying no changes in gene expression, or presenting transient increases or decreases (Supplementary Figure 1d). Taken all together, these results suggest that while there are major changes in gene expression and chromatin leading from one differentiated cell type to another (PC1), there are also changes that may be involved in a transient de-differentiation from pre-B cells into an intermediate state, and in the re-differentiation into macrophages (PC2), with a distinct contribution of expression and chromatin marks.

The association between chromatin marking and gene expression is overestimated by correlations computed in steady-state conditions

We computed, at each time-point, the steady-state correlation between levels of expression and histone modifications across the set of 12,248 genes (Figure 1d). As previously observed, we found a strong positive correlation for most active marks (median Pearson *r* value across time-points between 0.51 and 0.72), and a (weak) negative correlation for the repressive marks H3K9me3 and H3K27me3 (-0.07 and -0.17, respectively). However, when computing, for individual genes, the correlation between expression and chromatin profiles through time (time-course correlations), the values are substantially lower for active marks (median Pearson *r* ranging between 0.10 and 0.45), and higher for repressive marks (0.13 and -0.03 for H3K9me3 and H3K27me3, respectively; Figure 1d). Remarkably, for H3K9me3 the time-course correlation with expression is positive, in contrast to what has been previously described[24], and that we also measured in steady states.

It appears, therefore, that correlations measured in steady-state conditions artificially inflate the true degree of association between gene expression and chromatin modifications. This can be dramatically seen by randomizing the real temporal association between gene expression and chromatin marks. Within each gene’s time-series profile, we permuted histone modification levels among time-points, while keeping the actual gene expression values (see Methods; for an example with H3K4me3, compare upper and lower panels in Supplementary Figure 2a). As expected, the average time-course correlation is zero for all marks (Supplementary Figure 2b). However, the steady-state correlations are unexpectedly large for canonically active marks upon randomization, despite the fact that any meaningful association between gene expression and chromatin marks has been eliminated (Supplementary Figures 2a lower panel and 2b). This is likely due to a considerable fraction of genes displaying stable expression and chromatin profiles over time, which are either relatively highly expressed and marked (housekeeping genes)[25], or silent and not marked. Indeed, after removing the genes with silent or stable expression profiles over time, the steady-state correlations (Supplementary Figure 2c) are lower compared to those computed on the entire set of genes (Figure 1d), and become more similar to the time-course correlations.

Genes are characterized by a limited number of major chromatin states, which are more stable than expression

Next, we investigated the dynamics of chromatin marking during transdifferentiation. Towards that end, we summarized the chromatin state of each gene at each time-point, by building a multivariate Hidden Markov Model (HMM) on the signal of the nine histone marks along the twelve transdifferentiation points. More specifically, we produced a segmentation of the transdifferentiation time by assigning a given chromatin state to each gene at each time-point. This is in contrast to previous uses of HMMs in the field, where the segmentation is produced along the genome sequence by assigning a given chromatin state to every genome interval[26–30]. We explored configurations with up to twenty different states, and found that five states are a good compromise between optimizing the likelihood of the model and the number of states capturing the epigenetic status of genes (Supplementary Figure 3a and Figure 2a, see Methods). These five states correspond to the major combinations of histone modifications in which genes can be found (major chromatin states): a) absence of marking (with the exception, in some cases, of moderate marking by H3K9me3), b) low marking (mono and di-methylation of H3K4), c) bivalent marking (mostly marking by H3K4me1, H3K4me2 and/or H3K27me3), d) canonical active marking (all canonical active marks) and e) strong canonical active marking in the presence of H3K9me3 signal. These states (from a to e) correspond to increasing marking by canonically active histone modifications, with the exception of the bivalent marking state (c), which is also characterized by high H3K27me3 signal. These results suggest that only a limited number of combinations of marks can co-occur in a given gene at a given time-point. They also suggest that marking by H3K4me1 and H3K4me2 appears to be a precondition for marking by any other active histone modification, since for none of the configurations that we have explored, we have found states in which there is marking by an active histone modification without H3K4me1 and H3K4me2. The most frequent states among expressed genes are active and strong active marking (d and e, respectively), while the most frequent state among silent genes is absence of marking (a) (Supplementary Figure 3b).

Hierarchical clustering of genes based on the sequence of the five states along the twelve time-points revealed a limited number of temporal chromatin state profiles (Figures 2b-c). Most of the genes remain in the same chromatin state during transdifferentiation (constant state profiles), irrespective of whether they are stably (79%) or differentially expressed (70%) during transdifferentiation (Figure 2d, left panels). Thus, during the process, most changes in gene expression are not accompanied by chromatin changes.

Of the remaining genes, the vast majority (90%) go over just one-state transition during transdifferentiation. When considering DE genes, these transitions are generally associated with the expected transcriptional changes (Figure 2c). Transitions from weaker to stronger active chromatin marking are accompanied by increases in gene expression (Figure 2c, upper side; Figure 2d, middle panels), while transitions from stronger to weaker active chromatin states are accompanied by decreases in gene expression (Figure 2c, lower side; Figure 2d, right panels). However, while transitions from active to strong active marking states (and vice versa) are more numerous, the corresponding fold changes in gene expression are lower, compared to transitions from low marking to active marking states (and vice versa). We observed activating

Figure 2

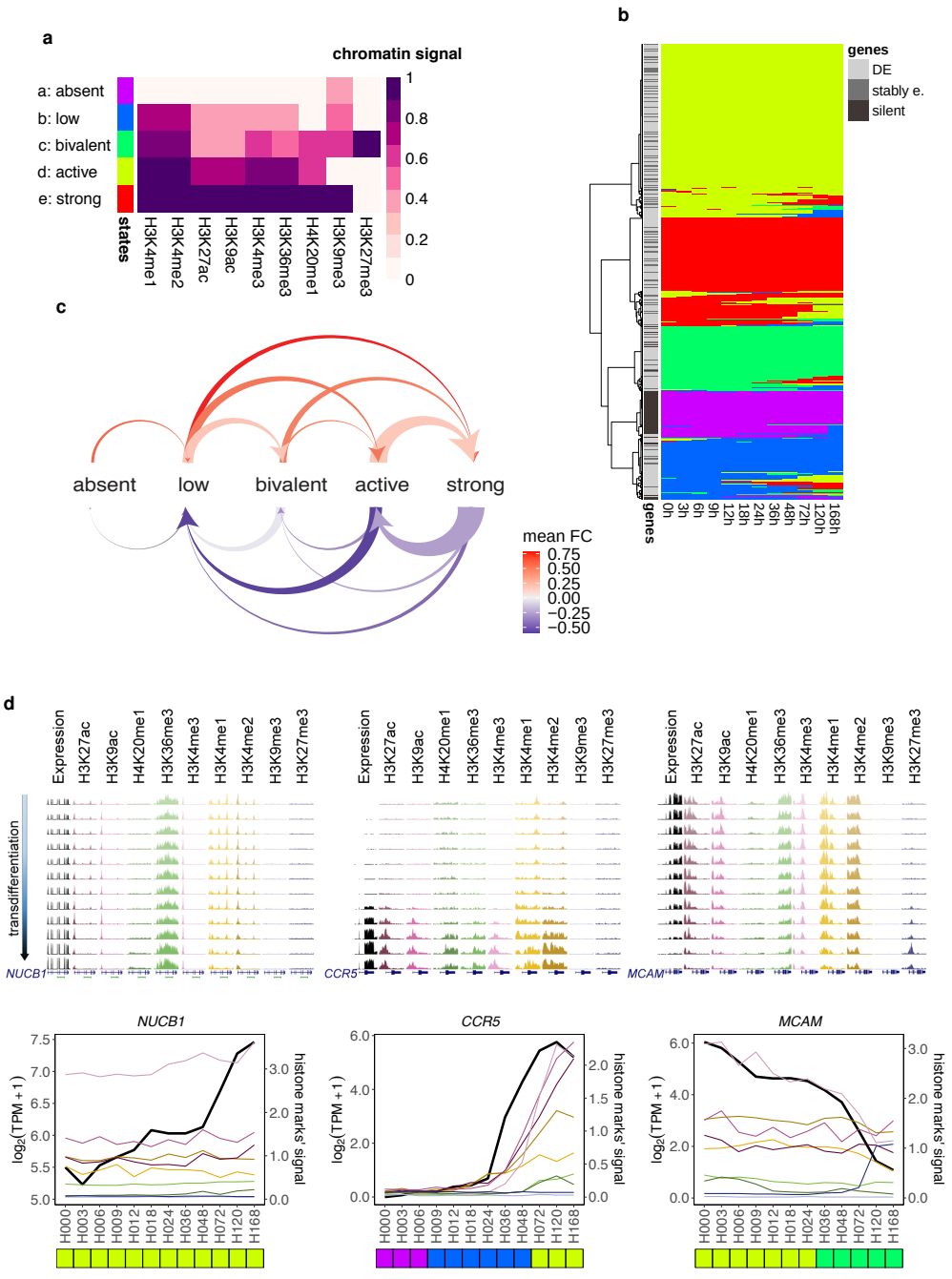

a
chromatin signal
a: absent
b: low
c: bivalent
d: active
e: strong
states
H3K4me1
H3K4me2
H3K27ac
H3K9ac
H3K4me3
H3K36me3
H4K20me1
H3K9me3
H3K27me3
1
0.8
0.6
0.4
0.2
0
b
genes
DE
stably e.
silent
0h
3h
6h
9h
12h
18h
24h
36h
48h
72h
120h
168h
c
absent
low
bivalent
active
strong
mean FC
0.75
0.50
0.25
0.00
−0.25
−0.50
d
Expression
transdifferentiation
NUCB1
CCR5
MCAM
log2(TPM + 1)
histone marks' signal

Figure 2: Genes are characterized by a limited number of major chromatin states, which are more stable than expression — See also Supplementary Figure 3. **a**: A five-state multivariate HMM. Each state is defined by a combination of histone marks. We report the histone marks' signals corresponding to each state. The states are sorted by increasing level of marking averaged over the nine histone modifications, with a and e states characterized by the lowest and highest average level of marking, respectively. **b**: Heatmap representing the hierarchical clustering of the HMM profiles built along the transdifferentiation process for the 12,248 genes. **c**: Arc diagram representing the types of state transitions observed in the HMM-sequence profiles of DE genes. The size of the arrow base is proportional to the number of genes reporting a given transition. Only transitions involving $\geq$ 10 genes are shown. We tested, for the sets of genes reporting each type of transition, the significance in gene expression fold-change (FC) (Wilcoxon Rank-Sum paired test, two-sided). The color of the arrow represents the average FC among genes experiencing a given transition. Transitions characterized by no significant changes in expression FC (Benjamini-Hochberg FDR $\geq$ 0.05) are represented by gray arrows. Upper panel: transitions from weaker to stronger active chromatin marking. Lower panel: transitions from stronger to weaker active chromatin marking. **d**: Examples showing different HMM states along transdifferentiation. For each gene, expression and chromatin tracks from one biological replicate are displayed, as well as normalized line plots averaging the signal from the two replicates. Profiles of HMM states for the three genes are shown at the bottom. Left panels: example of an up-regulated gene (*NUCB1*) with a constant HMM state profile along transdifferentiation. Middle panels: example of an up-regulated gene (*CCR5*) transitioning first from absence of marking state (a) to low marking state (b), and from this to active marking state (d). Right panels: example of a down-regulated gene (*MCAM*) transitioning from active marking state (d) to bivalent marking state (c).

transitions from the absent state mainly to the low marking state, further supporting the fact that marking by H3K4me1 and H3K4me2 is a prerequisite for the deposition of any other active histone modifications. On the other hand, we did not observe transitions from the strong active marking state to absence of marking, suggesting that the erasing of chromatin marks is not as an efficient process as its deposition.

Analysis of individual histone marks confirmed the HMM results. We determined whether the marks' signals are stable or variable over time, analogously to what was done for gene expression profiles. The majority of genes present, indeed, stable chromatin profiles during transdifferentiation, even when focusing only on the differentially expressed ones (Supplementary Table 3, left side; Figure 3a). Lysine acetylation (H3K27ac and H3K9ac) is the most dynamic signal (Supplementary Table 3, left side). Still, around 35% of DE genes show no changes in histone acetylation, despite being marked. Unexpectedly, only 8.5% of DE genes show changes in H3K27me3 throughout the process, although roughly half of them are down-regulated. Conversely, for a smaller number of silent and stably expressed genes we observed significant variations in their chromatin profiles over time (Supplementary Table 4, Figures 3b-c), comparable or even larger than for DE genes (Supplementary Figure 3c), although no changes could be detected in their expression profiles. We observed, in general, that differentially marked genes display clearer transdifferentiation trajectories compared to genes that are stably marked (Supplementary Figure 3d), further supporting that the contribution of gene expression and chromatin marks to cell state is not fully overlapping.

Chromatin marking is associated with expression specifically at the time of gene activation

The limited number of chromatin HMM states indicates a coordinated behaviour of histone modifications. To investigate this behaviour at the resolution of individual marks and how it relates to gene expression, we first determined the type of association between each mark and expression along transdifferentiation, for each of the 8,030 genes that are differentially expressed (labels: unmarked, stably marked, positively correlated, uncorrelated and negatively correlated; see Figure 4a, Supplementary Table 3 and Methods). Then, we clustered the combinations of marks and types of association, and found that, in general, in a given gene, most marks show indeed the same type of association with expression (Figure 4b). When clustering the genes based on these combinations, we found essentially three major groups (Figure 4c, Supplementary Figure 4a). The first and largest cluster includes 4,995 DE genes (62%), presenting either stable or uncorrelated profiles for the majority of active marks, and absence of marking for H3K27me3 and H3K9me3 (Figures 5a-b, upper panels). The second cluster includes 2,993 DE genes (37%), showing the canonical positive correlation between expression and most active modifications. A large proportion of these genes lack repressive marks, but a few of them (9%) exhibit the expected negative correlation with H3K27me3 (Figures 5a-b, middle panels). Finally, the third and smallest cluster includes 102 genes (1%) characterized by an overall absence of both active and repressive marking, with the exception of H3K4me1 and H3K4me2 (Figures 5a-b, lower panels).

Especially in the case of up-regulated genes, these clusters mostly reflect the level of gene activation when transdifferentiation starts (Figure 5c, Supplementary Figures 4b-c). Genes in cluster 1 are already

Figure 3

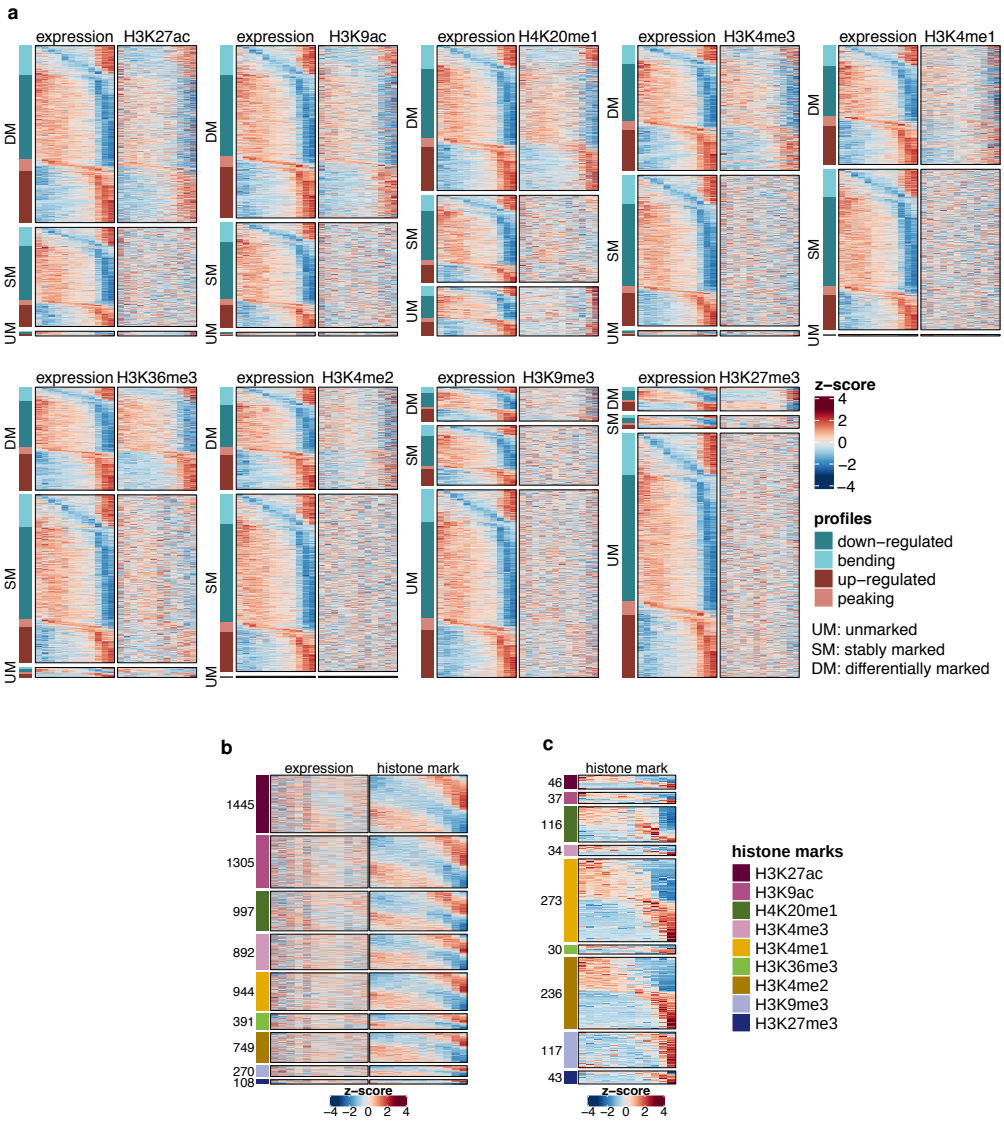

a
expression H3K27ac
expression H3K9ac
expression H4K20me1
expression H3K4me3
expression H3K4me1
expression H3K36me3
expression H3K4me2
expression H3K9me3
expression H3K27me3
DM
SM
UM
z−score
4
2
0
−2
−4
profiles
down−regulated
bending
up−regulated
peaking
UM: unmarked
SM: stably marked
DM: differentially marked
b
expression
histone mark
1445
1305
997
892
944
391
749
270
108
z−score
−4 −2 0 2 4
c
histone mark
46
37
116
34
273
30
236
117
43
z−score
−4 −2 0 2 4
histone marks
H3K27ac
H3K9ac
H4K20me1
H3K4me3
H3K4me1
H3K36me3
H3K4me2
H3K9me3
H3K27me3

Figure 3: Uncoupling of expression and chromatin marks throughout transdifferentiation — See also Supplementary Figure 3, Supplementary Tables 3-4. **a**: Expression and chromatin profiles across the 12 time-points (columns) for the set of 8,030 DE genes, distinguishing between differentially marked (DM), stably marked (SM) and unmarked (UM) genes (rows). The profiles consist of row-normalized *z*-scores, computed independently for expression and chromatin marks. **b**: Expression and chromatin profiles over the 12 time-points (columns) for the set of stably expressed genes that are differentially marked for a given histone modification along transdifferentiation. The profiles consist of row-normalized *z*-scores, computed independently for expression and chromatin marks. The largest numbers of significantly variable profiles are observed for H3K27ac and H3K9ac. **c**: analogous representation to Figure 3b for silent genes. In this case, H3K4me1 and H3K4me2 are the most variable marks throughout the process.

activated at the beginning of transdifferentiation, genes in cluster 2 are in early stages of activation or are activated early during transdifferentiation, while genes in cluster 3 are activated late during the process. The functions of the genes in these clusters are consistent with their level of activation at the beginning of transdifferentiation (Supplementary Figures 4d-e). In particular, genes in cluster 3 are associated with macrophage-specific functions, and we have found them lowly expressed and lowly marked in other cell types but CD14+ monocytes (Supplementary Figures 4f-g). Down-regulation of gene expression, on the other hand, appears to be largely uncoupled from chromatin changes, since most genes decreasing expression belong to cluster 1 (Supplementary Figure 4h).

Gene expression changes anticipate changes in most active marks for up-regulated genes

The results above are suggestive that the association between gene expression and histone modifications occurs preferentially in a limited window of time during the initial stage of gene activation. Thus, to investigate the relationship between expression and chromatin marking precisely at this stage, we focused on the set of 257 up-regulated genes that are not expressed at 0 hours p.i., and that are, therefore, specifically activated during transdifferentiation. The vast majority of these genes (230, 89%) belong to cluster 2, that is, they are indeed characterized by positive correlation between gene expression and active chromatin marks. They are mostly associated with low and bivalent marking HMM states and, in 25% of the cases, transition into stronger marking states towards the end of transdifferentiation (Supplementary Figure 5a, upper panel).

To investigate the temporal relationship between gene activation and chromatin marking, for each up-regulated gene and histone mark we rescaled the expression and chromatin time-series profiles to the same range (0-100%), and identified the first time-point at which the expression level and the chromatin signal reach at least 25%, 50%, 75% and 100% (Supplementary Figure 5b). In this way, we determined whether active chromatin marking anticipates, co-occurs with, or follows gene expression. In contrast to the prevalent view, we did not find that most active marks anticipate activation of gene expression. At the first stage of up-regulation (25%), only marking by H3K4me1, H3K4me2 and H3K27ac anticipates more often than follows activation of gene expression (Figures 6a-b), whereas for the other marks most changes follow expression up-regulation. These differences are progressively lost towards the end of the process (Figure 6a, Supplementary Figure 5c).

To further decipher the precise order in which active chromatin signals are established over time, we computed, for a given mark, the fraction of genes whose changes either anticipate (Figure 6c, upper panel) or co-occur with (Supplementary Figure 5d, upper panel) changes in each of the other six marks. When considering 25% of up-regulation, we observed that, in general, no marks anticipate H3K4me1, indicating that it is the first mark to increase, followed by H3K4me2 and H3K27ac (Figure 6c, upper panel). This is consistent with the HMM analysis, which suggested that marking by H3K4me1 and H3K4me2 is a prerequisite for marking by other histone modifications (Figure 2a). Changes in H3K4me1, H3K4me2 and H3K27ac most frequently precede increases in H3K9ac and H3K4me3. In all the comparisons, H3K36me3

Figure 4

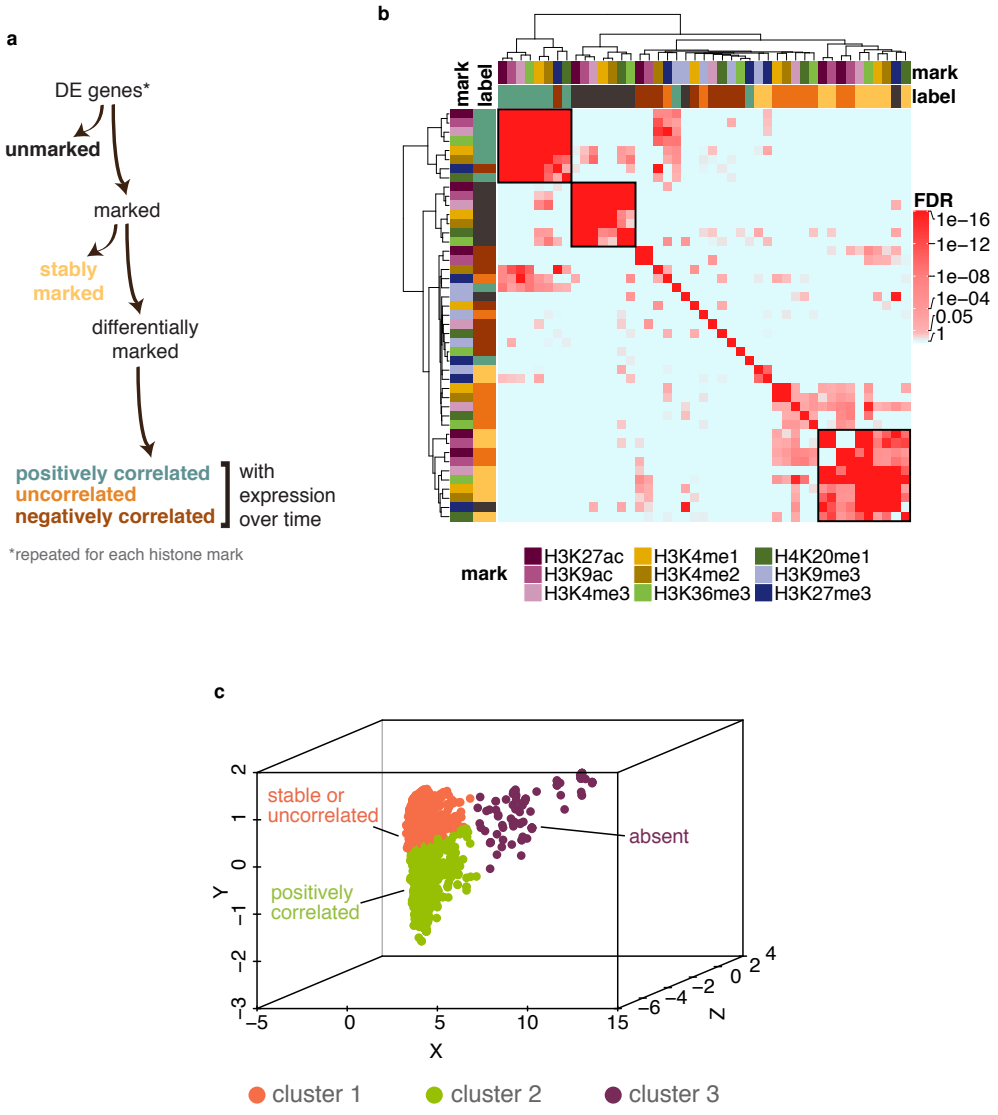

a
DE genes*
unmarked
marked
stably marked
differentially marked
positively correlated
uncorrelated
negatively correlated
with expression over time
*repeated for each histone mark
b
mark
label
FDR
1e−16
1e−12
1e−08
1e−04
0.05
1
mark
H3K27ac
H3K9ac
H3K4me3
H3K4me1
H3K4me2
H3K36me3
H4K20me1
H3K9me3
H3K27me3
c
stable or uncorrelated
positively correlated
absent
X
Y
Z
cluster 1
cluster 2
cluster 3

Figure 4: Chromatin marks show a coordinated behavior along transdifferentiation — See also Supplementary Figure 4, Supplementary Table 3. **a**: Decision-tree approach to label each of the 8,030 DE genes based on their chromatin marking status and its relationship with the expression profile over time. The approach is applied independently for each of the nine histone marks. The first branch distinguishes between unmarked (absence of peaks across all twelve time-points) and marked (presence of peaks in at least one time-point) genes. Within the set of marked genes, it further distinguishes between stably and differentially marked genes, i.e. genes characterized by absence and presence, respectively, of significant (maSigPro Benjamini-Hochberg FDR < 0.05) changes in chromatin signal along the process. Differentially marked genes are further classified into genes with positive, null or negative time-course correlation with expression. **b**: We assessed the overlap between sets of genes corresponding to the decision-tree labels across different histone marks (hypergeometric test). Hierarchical clustering of the FDR values identifies three main clusters: a) genes showing expression profiles positively correlated with H3K27ac, H3K9ac, H3K4me3, H3K36me3, H3K4me1, H3K4me2, H4K20me1, and negatively correlated with H3K27me3; b) genes unmarked for H3K27ac, H3K9ac, H3K4me3, H3K4me1, H3K4me2, H4K20me1 and H3K36me3; c) genes with stable or uncorrelated profiles for H3K27ac and H3K9ac, stable profiles for H3K4me3, H3K36me3, H3K4me1, H3K4me2, H4K20me1, and unmarked for H3K27me3. The color code for the labels is analogous to Figure 4a. **c**: Similar results are obtained with Cluster Correspondence Analysis, a method that combines dimension reduction and cluster analysis for categorical data. Three-dimensional representation of the genes (analysis objects), grouped into three clusters (color-coded) based on the combinations of histone marks and labels they display.

Figure 5

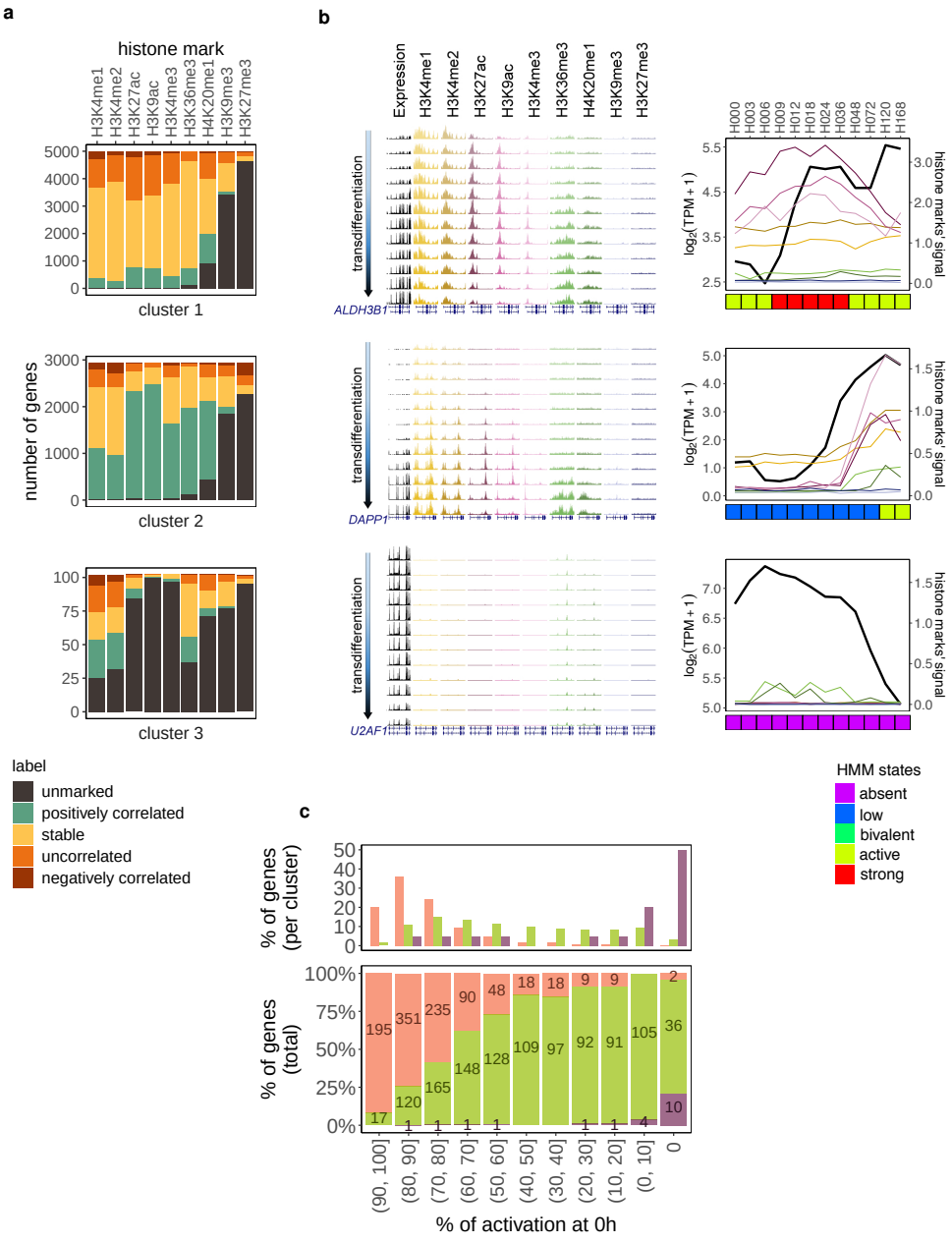

a
b
c
histone mark
H3K4me1
H3K4me2
H3K27ac
H3K9ac
H3K4me3
H3K36me3
H4K20me1
H3K9me3
H3K27me3
cluster 1
cluster 2
cluster 3
number of genes
Expression
transdifferentiation
ALDH3B1
DAPP1
U2AF1
H000
H003
H006
H009
H012
H018
H024
H036
H048
H072
H120
H168
log2(TPM + 1)
histone marks' signal
label
unmarked
positively correlated
stable
uncorrelated
negatively correlated
HMM states
absent
low
bivalent
active
strong
% of genes (per cluster)
% of genes (total)
% of activation at 0h
(90, 100]
(80, 90]
(70, 80]
(60, 70]
(50, 60]
(40, 50]
(30, 40]
(20, 30]
(10, 20]
(0, 10]
0

Figure 5: Chromatin marking is associated with expression specifically at the time of gene activation — See also Supplementary Figure 4, Supplementary Tables 5-6. **a**: Percent stacked bar plot representing, for each of the three clusters, the proportion of unmarked, stably marked, positively correlated, uncorrelated, and negatively correlated genes identified with respect to each histone mark. **b**: Examples of genes belonging to each cluster. For each gene, expression and chromatin tracks from one biological replicate are displayed, as well as normalized line plots averaging the signal from the two replicates. Profiles of HMM states for the three genes are shown at the bottom. Upper panels: example of an up-regulated gene (*ALDH3B1*) showing stable and uncorrelated profiles for active marking and absence of H3K9me3 and H3K27me3 along transdifferentiation. Middle panels: example of an up-regulated gene (*DAPP1*) showing positively correlated profiles for active marking and absence of H3K9me3 and H3K27me3 along transdifferentiation. Lower panels: example of a down-regulated gene (*U2AF1*) showing absence of marking along transdifferentiation. **c**: Percent stacked bar plot reporting the proportion of up-regulated genes in clusters 1-3 characterized by decreasing degrees of gene expression activation (bins of 10% decrement) at time-point 0h p.i. The degree of gene expression activation is defined as the ratio between the gene’s expression level at 0h and its maximum expression level along transdifferentiation.

and H4K20me1 follow the other marks (Figure 6c, upper panel). As observed for gene expression, this precise order of marks' deposition is progressively lost along transdifferentiation (Figure 6c upper panel, Supplementary Figure 5d upper panel). Overall, this suggests that the deposition of active chromatin modifications follows a precise order at the time of initial gene activation (H3K4me1 $>$ H3K4me2 $>$ H3K27ac $>$ expression $>$ H3K9ac $>$ H3K4me3 $>$ H3K36me3 $>$ H4K20me1; Figure 6d, left panel).

We performed a similar analysis with the set of 629 up-regulated genes that are already substantially expressed at 0 hours p.i. ($>$ 25 TPM). These genes belong mostly to cluster 1 (389, 62%), that is, their expression profiles are uncoupled from changes in chromatin marking, and they actually remain in active chromatin states during transdifferentiation (Supplementary Figure 5a lower panel). For these genes we did not find preservation in the pattern of chromatin deposition with respect to expression (Supplementary Figure 5e), nor in the deposition of the marks (Figure 6c lower panel; Figure 6d right panel; Supplementary Figure 5d lower panel).

A model to explain the coupling between transcription and chromatin marking over time

Altogether, our results show that the canonical association between histone modifications and gene expression mainly occurs in a limited window of time preceding and following initial gene activation. We specifically propose a model (Figure 7a) in which the activation of gene expression is anticipated by deposition of H3K4me1, H3K4me2 and, less frequently, of H3K27ac at promoter regions. The deposition of other marks typically enriched either at promoters (H3K9ac, H3K4me3) or over the gene body (H3K36me3, H4K20me1) is concomitant to or, more often, follows (and may be induced by) gene activation. After this initial stage of gene activation, further changes in gene expression, comparable or even stronger, appear to be mostly uncoupled from changes in histone modifications (Figure 7b, compare left and right panels).

This model explains our observations well. The patterns of association between chromatin marking and gene expression (as defined in Figure 4a) for genes in different degrees of activation when transdifferentiation starts (0h p.i.) reflect how this association changes as gene activation proceeds (Figure 7c). Up-regulated genes that are silent when transdifferentiation starts (mostly in cluster 3) lack almost all “activating” histone modifications, possibly with the exception of H3K4me1 and H3K4me2 (i.). Up-regulated genes in cluster 2 that are lowly or not activated at 0h show mostly correlated patterns of expression and chromatin marking. In these genes, most marks, with the exception of H3K4me1, H3K4me2 and H3K27ac, follow rather than anticipate expression (ii., see also Figure 7b, left panel). As we consider genes with increasing degrees of activation at 0h (and thus, in increasingly advanced states of activation), the fraction of genes with correlated patterns of expression and chromatin marking decreases, while the fraction of genes with stable or uncorrelated chromatin profiles (iii. and iv.) proportionally increases. The temporal order of activation of marks observed in early activation stages is also gradually lost. Finally, for genes in cluster 1 (v.), which are already highly active when transdifferentiation starts, changes in gene expression, even if substantial, are mostly uncoupled from chromatin marking, showing uncorrelated or stable profiles (see also Figure 7b, right panel).

Figure 6

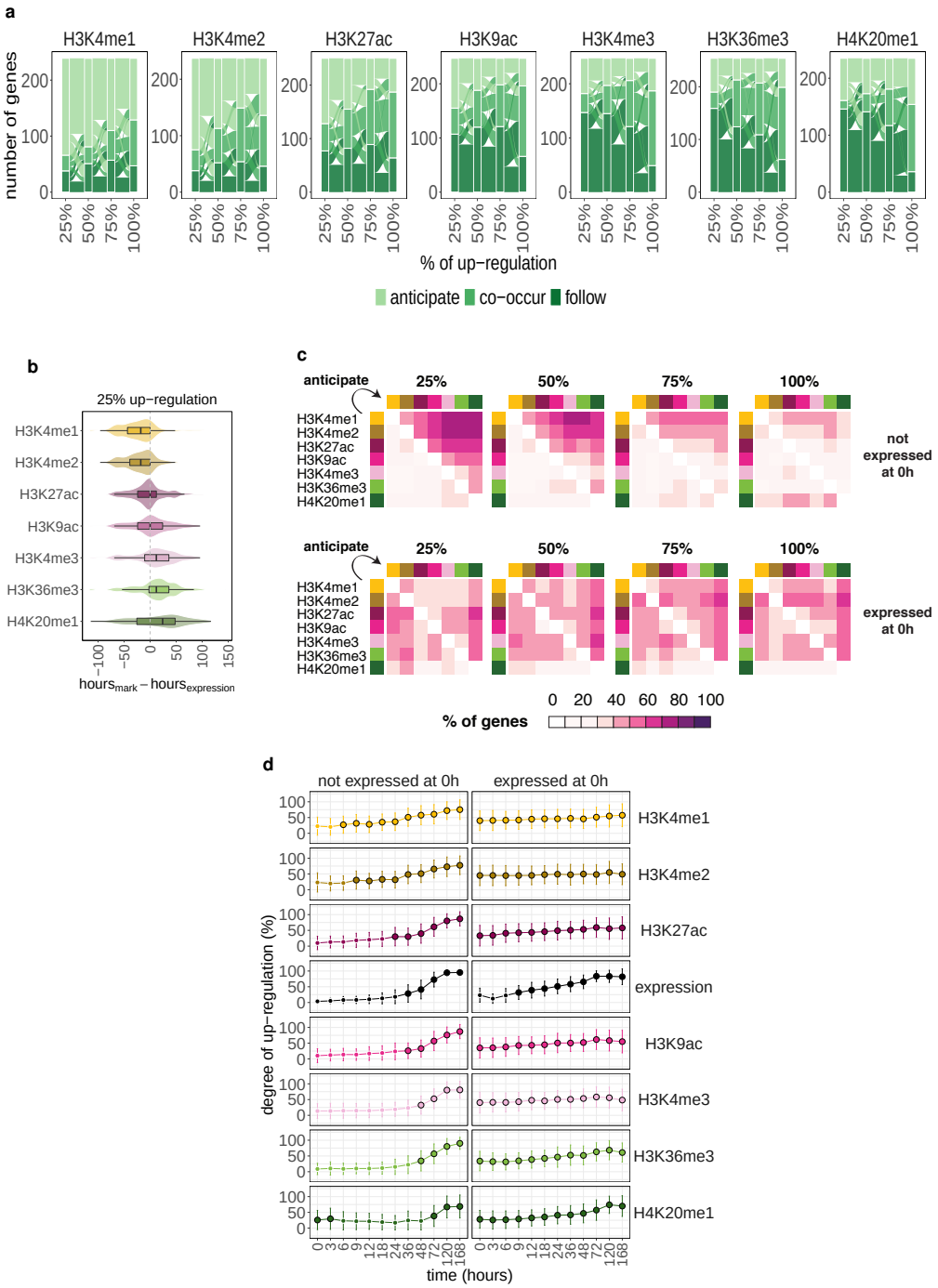

a
H3K4me1
H3K4me2
H3K27ac
H3K9ac
H3K4me3
H3K36me3
H4K20me1
number of genes
% of up−regulation
anticipate co−occur follow
b
25% up−regulation
hours_mark − hours_expression
c
anticipate
not expressed at 0h
expressed at 0h
% of genes
d
not expressed at 0h
expressed at 0h
degree of up−regulation (%)
expression
time (hours)

Figure 6: Gene expression changes anticipate changes in most active marks for up-regulated genes. — See also Supplementary Figure 5. **a**: Alluvial plot describing, for each of the seven canonical active histone marks, the number of genes, out of 257 genes activated during transdifferentiation (i.e. up-regulated genes not expressed ($<$ 1 TPM) at 0 hours p.i.), for which the up-regulation in a given mark's signal anticipates (light green), co-occurs with (green) or follows (dark green) gene expression up-regulation. For more details see Supplementary Figure 5b. The flow lines indicate the number of genes exchanged among the three groups across increasing degrees of up-regulation. **b**: Lag (hours) between 25% up-regulation in histone marks' signal and expression level for the 257 selected up-regulated genes. Negative lags correspond to changes in chromatin marks anticipating changes in gene expression; positive lags correspond to changes in chromatin marks following changes in gene expression. **c**: Upper panel: Heatmaps reporting the proportion (%) of genes activated during transdifferentiation whose changes in the chromatin mark on row *i* anticipate changes in the chromatin mark on column *j*. Like in the previous analyses, we considered four subsequent degrees of up-regulation (25%, 50%, 75% and 100%). e.g. the fraction reported in cell [row 1, column 2] of the first heatmap (25%), corresponds to the percentage of genes for which the 25% up-regulation in H3K4me1 signal (yellow - row 1) anticipates the 25% up-regulation in H3K4me2 signal (ochre - column 2). Lower panel: analogous to upper panel for the 629 up-regulated genes already expressed ($>$ 25 TPM) at 0h p.i. For this latter set of genes there is not a precise order of increase in chromatin marks. **d**: Mean and standard deviation of time-series expression and chromatin profiles for the 257 (left panel) and 629 (right panel) up-regulated genes that are not expressed and highly expressed, respectively, at 0 hours p.i. The expression and histone marks' time-series profiles of each gene were re-scaled to a 0-100% range prior to the analysis. We highlight in black the time-points at which the mean value is $\geq$ 25%.

Figure 7

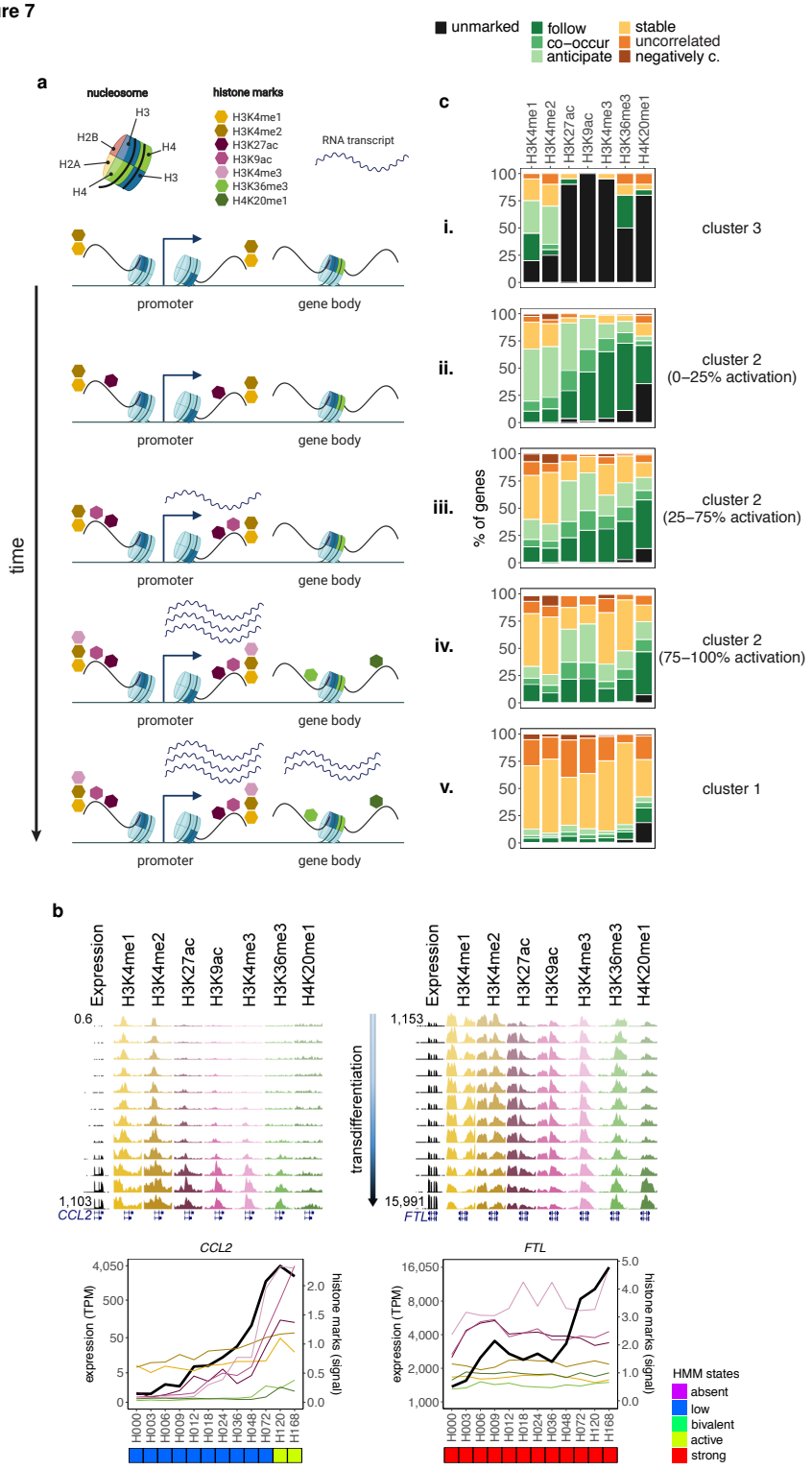

a
nucleosome
histone marks
H3
H2B
H4
H2A
H3
H4
H3K4me1
H3K4me2
H3K27ac
H3K9ac
H3K4me3
H3K36me3
H4K20me1
RNA transcript
time
promoter
gene body
b
Expression
H3K4me1
H3K4me2
H3K27ac
H3K9ac
H3K4me3
H3K36me3
H4K20me1
0.6
1,103
CCL2
transdifferentiation
1,153
15,991
FTL
CCL2
FTL
expression (TPM)
histone marks (signal)
4,050
500
50
5
0
16,050
8,000
4,000
2,000
1,000
2.0
1.5
1.0
0.5
0.0
5.0
4.0
3.0
2.0
1.0
0.0
H000
H003
H006
H009
H012
H018
H024
H036
H048
H072
H120
H168
HMM states
absent
low
bivalent
active
strong
c
unmarked
follow
co−occur
anticipate
stable
uncorrelated
negatively c.
% of genes
100
75
50
25
0
i.
cluster 3
ii.
cluster 2
(0–25% activation)
iii.
cluster 2
(25–75% activation)
iv.
cluster 2
(75–100% activation)
v.
cluster 1

Figure 7: A model to explain the coupling between transcription and chromatin marking over time. a: According to our model, chromatin marking correlates with expression specifically during the first stage of gene activation, and the deposition of histone marks follows a specific order. Further changes in gene expression that happen later in time are mostly uncoupled from chromatin marking. **b**: Examples of up-regulated genes inactive (*CCL2*) and highly active (*FTL*) at the beginning of transdifferentiation. For each gene, expression and chromatin tracks from one biological replicate are displayed, as well as normalized line plots averaging the signal from the two replicates. Profiles of HMM states for the two genes are shown at the bottom. Left panels: for *CCL2*, most active histone modifications follow gene activation, with the exception of H3K4me1 and H3K4me2, which anticipate it. Right panels: for *FTL*, most active histone modifications remain stable along transdifferentiation, even though its absolute increase in expression is much higher than that of *CCL2*. **c**: Percentage (%) of unmarked, stably marked, positively correlated, uncorrelated and negatively correlated profiles within cluster 3, cluster 2 (0-25%, 25-75%, 75-100% activation level), and cluster 1 up-regulated genes. Positively correlated genes are further separated into genes whose histone mark’s up-regulation anticipates, co-occurs with or follows gene expression up-regulation.

Discussion

Epigenetics was initially defined as "the branch of biology that studies the causal interactions between genes and their products which bring the phenotype into being"[31]. In a more contemporary definition, "an epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence"[32]. The epigenetic mechanisms leading to the development of an individual or to the differentiation of a cell lineage from the unique genotype of the organism have been largely studied during decades. Although initial references to the mechanisms by which epigenetics promotes cell memory and leads cell fate did not relate to its ability to regulate gene expression, a causative role for epigenetic modifications in controlling transcription has been later pointed out (see 33,34 for reviews about different aspects related to epigenetics and its role in regulating gene expression), and it has even been shown that some epigenetic features, such as histone modifications, are accurate predictors of gene expression[7,8,35] and the other way around[36].

However, the causal/consequential relationship between chromatin modifications and gene expression represents a long-standing discussion[37], and a number of reports have challenged the causal role that has been broadly attributed to chromatin modifications[11,19,38,39]. Still, and despite the efforts dedicated to this problem and the vast literature produced, the actual relationship between histone modifications and the regulation of gene expression remains unsolved.

This is partially due to the few available studies in which gene expression and histone modifications have been both consistently monitored through time in a given dynamic system. Differentiation models are suitable to study the relationship between gene expression and chromatin marking, as they provide a dynamic system that allows to decipher the order of the events. In this work, we have used the transdifferentiation of BLaER1 cells (pre-B cells) into macrophages, a model that has proven to be highly efficient[21], and we have generated high-quality data on the transcriptome and the epigenome in twelve time-points along the seven days the transdifferentiation process lasts. Our analysis of these data has uncovered some fundamental features of chromatin organization in human genes and of the relationship between gene expression and histone modifications.

Our analyses have also contributed to a better understanding of the molecular events underlying transdifferentiation of pre-B cells into macrophages. Despite the fact that, to our knowledge, there is no retrodifferentiation during the process[21,40], the joint PCA of gene expression and chromatin marks suggests that BLaER1 cells undergo an intermediate state (Figure 1c). This intermediate state is characterized by chromatin changes not accompanied by changes in gene expression (Supplementary Figure 6), and vice versa by changes in gene expression not associated with chromatin changes (Supplementary Figure 6a). Although it is often assumed that the transcriptome is the main determinant of cell state, these results suggest that epigenetic modifications contribute to cell state in a manner that cannot be fully recapituted by gene expression. Thus, neither the epigenome nor the transcriptome can be fully predictive of one another.

Consistently, we found that the association between gene expression and chromatin modifications is overall weaker than reflected by the correlations reported so far, which have been mostly computed in a

particular steady-state cellular condition (Figure 1d). These artifactually strong correlations result from the largely constrained nature of the human epigenome and transcriptome. In particular, a large fraction of genes in the human genome (likely more than 50%[25]) are either invariably silent and not marked, or expressed and marked across most cellular states. Genes with stable epigenomes and transcriptomes drive the correlations to large values when computed in a particular cell condition, and explain why models relating gene expression to histone modifications inferred in a particular cell type have high predictive power in other cell types[7,8,35,36], even though there is no true causality involved in the relationship between chromatin and expression. The steady-state correlations represent an example of the Sympson's paradox[41], by which the data can show different or even opposite behavior if subgroups within the dataset are considered.

HMMs have been widely used to summarize patterns of combinations of multiple histone modifications into a limited number of chromatin states. However, in most cases so far, they have been used to segment the genome sequence[26–30]. Here, instead, we used them, we believe for the first time, to segment time along a dynamic differentiation process. The HMM segmentation reveals that, even though the number of possible histone combinations is very large (if nine histones are considered, 2^9 = 512 combinations are possible), most genes are actually found in one among only about five major states (Figure 2a). This challenges to some extent the notion of a histone code[6]. Further supporting the limited number of genic chromatin states, we found that marks act in a coordinated manner, meaning that genes showing a stable profile for one histone modification tend also to present stable profiles of the other marks, and that genes showing absence of one active mark tend to be void of all positive modifications (Figures 4b-c, Supplementary Figure 4a). Most genes remain in the same chromatin state during transdifferentiation, irrespective of whether they are or not differentially expressed, explaining the low correlation between gene expression and chromatin marks throughout time. Analysis of individual histone modifications further uncovered two unexpected findings regarding the chromatin marks typically associated with gene silencing. First, we observed that, although roughly 4,000 genes are down-regulated during the process, only 10% of them present H3K27me3 marking in at least one time-point, indicating that the majority of genes that are silenced along transdifferentiation do not depend on Polycomb repression. Second, we saw that H3K9me3 marking at transcription start sites is associated more frequently with active transcription than gene silencing (see Supplementary Table 3, right side), contrary to what has been previously reported[24]. Actually, H3K9me3 at the transcription start site has been previously related to active expression in malignant cells[42]. Furthermore, these analyses also allowed us to identify a number of silent or stably expressed genes along transdifferentiation that show changes in chromatin marking (Figures 3b-c).

While there is a general lack of coupling between gene expression and chromatin marking, there is a temporal relationship between gene expression and the different histone modifications at the time of gene activation. We propose a model (Figure 7a) in which activation of gene expression is anticipated by deposition of H3K4me1, H3K4me2, while deposition of other marks is concomitant or, more often, follows gene activation, being the gene body marks the last ones to be incorporated. The order of chromatin marking in our model is in agreement with the observed deposition of histone modifications upon induction

of gene expression in human melanoma cells[43], and with the notion that the methylation of some histone residues depends on the transcription machinery[39]. While we observed that certain modifications, such as H3K4me1/2 and H3K27ac tend to anticipate gene expression, this does not necessarily mean that they are the cause of transcription initiation. Actually, we have also observed particular cases in which these marks are deposited post-activation (for an example see Figure 5b, middle panels). After the initial stage of gene activation, further changes in gene expression, even if substantial, appear to be mostly uncoupled from changes in histone modifications (Figure 7b). It is tempting to speculate that after the initial burst of transcription, histone residues are saturated with modifications, and that therefore, any further up-regulation of gene expression cannot possibly be accompanied by increased levels of histone modifications.

We do have identified a small set of genes that are expressed in the absence of any histone modification, with the exception of H3K4me1 and H3K4me2 (Figures 4c, 5a-b lower panels). A few of these are activated later during the transdifferentiation process, and therefore we lack the temporal resolution to detect post-activation marking. Still, many of these genes are down-regulated or stably expressed, and are unmarked even at the beginning of transdifferentiation (for an example see Figure 5b, lower panels). Gene activation without histone modifications has been previously observed for developmentally regulated genes in the fruit fly[12].

Here we have focused specifically on the dynamics of chromatin modifications during up-regulation. Our results suggest that down-regulation appears to be largely uncoupled from chromatin changes (Supplementary Figure 4h). However, while RNA sequencing-inferred expression levels can be used to approximately identify the time at which a gene is initially activated, differences in RNA stability may confound the identification of the time-point at which a gene is fully inactivated. Indeed, RNAs can be detected long after gene inactivation, for a time likely to be specific to each individual gene. Therefore, the data that we have generated does not have the appropriate resolution to discard that this lack of coupling during down-regulation is partially caused by the difficulty in precisely identifying the time-point at which genes stop being expressed.

The multi-omics data that we have generated during the pre-B cell transdifferentiation into macrophages has allowed us to address with unprecedented resolution some fundamental questions regarding the dynamics of chromatin marking and gene expression during cellular differentiation, and have contributed to shed light on some long-standing questions in the field. Further mining of this data resource will certainly contribute to a deeper understanding of the epigenetic layer of gene regulation.

Methods

RESOURCE AVAILABILITY

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The code generated during this study is available at https://github.com/bborsari/Borsari_et_al_transdifferentiation_chromatin. A complete list of scripts used for each analysis described in the section *Method details* can be found at https://github.com/bborsari/Borsari_et_al_transdifferentiation_chromatin/blob/master/bin/table.scripts.tsv. When not specified in the text, the code used for a given analysis is included in the corresponding figure's script.

RNA-seq and ChIP-seq raw and processed data from this study have been submitted to ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) under accession numbers E-MTAB-9790 and E-MTAB-9825, respectively.

Processed data in GRCh38/hg38 assembly from this study is available for visualization at the UCSC Genome Browser[44] (http://genome.ucsc.edu/). The track data hub is available at https://public-docs.crg.es/rguigo/Data/bborsari/hubs/ERC_human_hub/hub.txt.

A web page has also been implemented to gather all information regarding the Chromatin and Transcriptomics Dynamics Project (http://rnamaps.crg.eu/). The web page provides information about all experiments and replicates performed during the project, as well as access to the data in ArrayExpress and the UCSC Genome Browser.

ENCODE data is freely available on the ENCODE portal (https://www.encodeproject.org/). Experiments and files accession IDs for RNA-seq and ChIP-seq data are reported in Supplementary Tables 5 and 6, respectively.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Transdifferentiation of BLAER1 cells to macrophages

For the transdifferentiation process we made use of the Burkitt lymphoma cell line BlaER1, as described in 21. Induction of transdifferentiation (treatment with 100 μM β-estradiol and growth in the presence of 10 nM Il-3 and 10 nM CSF-1) has been described in 45 and 46. The process was monitored at 12 time-points (as described in 21): 0, 3, 6, 9, 12, 18, 24, 36, 48, 72, 120 and 168 hours post-induction (p.i.; Figure 1a).

METHOD DETAILS

RNA-seq library preparation and sequencing

Two independent biological replicates for each time-point were performed. Briefly, cells were lysed with QiAzol (Qiagen, The Netherlands). Chloroform was added to each sample, and RNA contained in the aqueous solution was isolated and purified by using RNeasy mini kit columns (Qiagen, The Netherlands). Poly A+ libraries were prepared with 1 μg of total RNA and using TruSeq Stranded mRNA Library Prep Kit (Illumina, USA) according to the manufacturer's protocol. Libraries were analyzed using Agilent DNA 1000 chips to determine the quantity and size distribution, and sequenced paired-end 75-bp on an Illumina

HiSeq 2000.

ChIP-seq library preparation and sequencing

ChIP-seq experiments of nine histone marks (H3K4me1: Abcam ab8895; H3K4me2 : Millipore 07-030; H3K4me3: Abcam ab8580; H3K9ac: Abcam ab4441; H3K27ac: Diagenode C15410192; H3K36me3: Abcam ab9050; H4K20me1: Abcam ab9051; H3K9me3: Abcam ab8898; H3K27me3: Millipore 07-449) were performed in two independent biological replicates for each time-point. Cells were crosslinked with formaldehyde 1% (Sigma) for 10’ at room temperature. The reaction was stopped by adding glycine to 0.25 M final concentration for 10’ at room temperature. Fixed cells were resuspended in 100 µL of lysis buffer (SDS 1%, EDTA 10 mM, TrisCl 50 mM and protease inhibitors). The lysate was sonicated for 25’ using Covaris S2 system in TC12 tubes (Duty cycle 20%, Intensity 8, cycles/burst 200, water level 15). The cleared supernatant was used immediately in ChIP experiments or stored at -80 °C. 5 µg of sonicated chromatin were diluted in 900 µL RIPA buffer — H3K4me3, H3K9ac, H4K20me1, H3K27me3 and H3K27ac (140 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% Na deoxycholate, protease inhibitors) —, RIPA 2X — H3K4me1, H3K4me2 and H3K9me3 (280 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 2% Triton X-100, 0.2% SDS, 0.2% Na deoxycholate, protease inhibitors) —, or RIPA 1X 1% triton — H3K36me3 (280 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 1% Triton X-100, 0.2% SDS, 0.2% Na deoxycholate, protease inhibitors). For H3K4me3, H3K36me3, H3K9ac and H3K27me3 ChIPs, chromatin and antibodies were incubated overnight, rotating at 4 °C with 0.125-5 µg of specific antibody and samples were then incubated for 2 hours rotating at 4 °C with Dynabeads protein A for immunoprecipitation (Invitrogen) to recover the bound material. For H3K4me1, H3K4me2, H3K9me3, H4K20me1 and H3K27ac ChIPs, antibodies were coated to protein A magnetic beads for 2 hours at 4 °C prior to overnight incubation with chromatin. In all cases, beads were washed for 10’ three times in 1 mL of the corresponding immunoprecipitation buffer without protease inhibitors, then washed once in 1 mL LiCl buffer (0.25 M LiCl, 0.5% NP-40, 0.5% sodium deoxycholate, 1 mM Na-EDTA, 10 mM Tris-HCl, pH 8.0), and finally washed twice in 1 mL of TE buffer (1 mM Na-EDTA, 10 mM Tris-HCl, pH 8.0). ChIPped material was incubated with DNase-free RNase at 50 µg/mL for 30’ at 37 °C. Chromatin was reverse-crosslinked by adding SDS (0.5% final concentration) and Proteinase K (500 µg/mL final concentration) and incubated overnight at 65 °C. ChIPped chromatin was then purified with Qiaquick PCR purification columns (Qiagen) following the manufacturer’s instructions. ChIP libraries were prepared with 1-5 ng of DNA and using NebNext Ultra DNA library prep kit for Illumina (New England Biolabs) according to the manufacturer’s protocol. Libraries were analyzed using Agilent DNA High Sensitivity chips to determine the quantity and size distribution, and sequenced single-read 50-bp on an Illumina HiSeq 2000.

In total, 264 samples were sequenced (24 by RNA-seq, 216 by ChIP-seq, 24 by ChIP input).

RNA-seq data processing and analysis

Data was processed using the *grape-nf* (`https://github.com/guigolab/grape-nf`) Nextflow[22] pipeline. RNA-seq reads were aligned to the human genome (assembly GRCh38, Gencode annotation version 24) using the STAR[47] software version 2.4.0j . We allowed a maximum number of mismatches equal to 4% of the read length. Only alignments for reads mapping to ten or fewer loci were reported. Quantification of genes and transcripts was done with RSEM[48] version 1.2.21. TPM calculation was performed after removing mitochondrial genes.

From the set of 19,831 protein-coding genes (Gencode v24), we selected 10,696 expressed genes with a maximum expression during transdifferentiation $\geq$ 5 TPM in both replicates, and 1,552 silent genes (0 TPM in all time-points and replicates). Based on this set of 12,248 genes, we quantile-normalized the expression matrices (log_2-transformed TPM, pseudocount of 1) across replicates and time-points using the R package preprocessCore[49] (script: `quantile.normalization.R`), and obtained the mean expression levels between replicates (script: `matrix_matrix_mean.R`).

To detect significant gene expression changes along transdifferentiation, we used the R package maSigPro[50] with replicates handled internally. Function `p.vector()` was run with default parameters: `Q = 0.05`, `MT.adjust = "BH"`, `min.obs = 20` (script: `maSigPro.wrapper.R`). We defined as stably expressed those genes reporting a maSigPro FDR value $\geq$ 0.05 (n = 2,666).

As concerns the identification of up-regulated, down-regulated, peaking and bending genes, we performed a two-step classification across the 8,030 genes with significantly variable gene expression profiles. Briefly, we first focused on profiles with at least two-fold change (in log_2 scale this change corresponds to 1) and identified monotonic up-regulations and down-regulations; peaking profiles were defined as monotonic increases followed by monotonic decreases, bending profiles as the opposite (script: `classification.log2.pl`). All other significantly variable genes with fold-change $<$ 2 were assigned to one of these four groups following hierarchical clustering (distance measure: *euclidean*; clustering method: *complete*; script: `classification.2.R`).

ChIP-seq data processing and analysis

Data was processed using the *ChIP-nf* (`https://github.com/guigolab/chip-nf`) Nextflow[22] pipeline. ChIP-seq reads were aligned to the human genome assembly (GRCh38) using the GEM[51] mapping software, allowing up to two mismatches. Only alignments for reads mapping to ten or fewer loci were reported. Duplicated reads were removed using Picard (`http://broadinstitute.github.io/picard/`). Pile-up signal from bigWig files was obtained running MACS2[52] on individual replicates. No shifting model was built. Instead, fragment length was set to 250 bp and was used to extend each read towards the 3' end (using the `--extsize` option). Pile-up signal was normalized by scaling larger samples to smaller samples (using the default for the `--scale-to` option) and adjusting signal per million reads (enabling the `--SPMR` option). Peak calling was performed using Zerone[53] with replicates handled internally, and passed the filter for all pairs of replicates (`advice: accept discretization`).

To check library complexity, we computed the fraction of non-redundant mapped reads[54] (recommended threshold: NRF $\geq$ 0.8) for each ChIP-seq experiment, and found a minimum NRF value of 0.92. Additionally, to evaluate the global ChIP enrichment, we computed the fraction of reads in peaks[54] (recommended threshold: FRiP $\geq$ 0.01), and found a minimum FRiP value of 0.05.

The intersection / overlap analyses described below were performed with the function `intersectBed` of BEDTools[55] software v2.27.1.

To select the genomic location enriched, on average, in a specific histone mark (region of interest), we focused on an up-stream and down-stream 5 Kb region (±5 Kb) with respect to the first annotated Transcription Start Site (TSS) of the gene, and retrieved 6,063 protein-coding genes that did not overlap any other gene body ±5 Kb. For each histone modification we then selected, among the 6,063 genes, those with peaks in the ±5 Kb promoter region in all the 12 time-points, and computed, using the function `aggregate` from the bwtool[56] software (script: `bwtool.aggregate.ChIPseq.sh`), the mean pile-up signal for each experiment. Based on this analysis, we decided to select as regions of interest i) the gene body for H3K36me3 and H4K20me1, ii) ±2 Kb with respect to the TSS for all other marks (Supplementary Figure 1c). A comprehensive catalogue of all non-redundant (same ensembl gene ID and start coordinate) TSSs annotated for the selected 12,248 in Gencode v24 was obtained with the script `non.redundant.TSS.sh`.

To compare expression and chromatin profiles over time, we quantified, for each of the nine histone marks, the amount of pile-up signal associated with a gene at each time-point (script: `get.matrix.chipseq.sh`). Briefly, if a peak was present in the region of interest of a gene at a specific time-point, we considered the mean pile-up signal in the intersection between the peak and the region of interest, otherwise we computed the mean pile-up value in the entire region of interest. In the presence of multiple peaks and/or multiple regions of interest (e.g. in case of multiple TSSs annotated for the same gene), we considered the highest of all observed values. Matrices of histone marks' signals for the selected 12,248 protein-coding genes were quantile-normalized across replicates and time-points using the R package preprocessCore[49] as done for gene expression. For all down-stream analyses, we used the mean signal between replicates.

Principal Component Analysis of expression and chromatin data

For this type of analysis we made use of the transposed expression and chromatin For this type of analysis we made use of the transposed expression and chromatin matrices generated as described in sections *RNA-seq data processing and analysis* and *ChIP-seq data processing and analysis*, respectively. Therefore, genes (columns) and time-points (rows) were used as variables and observations, respectively. We centered and scaled each of the ten transposed matrices independently, obtaining *z*-score profiles for each time-point monitored at expression and histone marks' level. For the joint Principal Component Analysis (PCA) reported in Figure 1c across expression and the nine histone marks, we included as variables the subset of 10,658 genes with non-missing (NA) *z*-score profiles in all ten matrices. As a consequence,

1,590 genes were excluded from this analysis, 98% of them being the silent genes (1,552). For the PCAs reported in Supplementary Figure 3d, we considered for each histone modification the corresponding sets of DE genes that are either stably or differentially marked.

Analysis of the degree of correlation between expression levels and chromatin signals

Steady-state correlations between gene expression levels and each histone mark's signals were computed at individual time-points considering the entire set of 12,248 selected protein-coding genes. In this case, Pearson *r* measured the degree of correlation between the vector of 12,248 expression levels and the vector of 12,248 mark signals at a given time-point (Figure 1d, dots). Time-course correlations were measured, instead, at the level of individual expressed genes. Silent genes were not considered for this analysis, because of the zero standard deviation in their time-series expression profile (i.e. 0 TPM in all time-points). Thus, for each gene and histone mark we obtained the Pearson *r* correlation coefficient between the vector of 12 expression levels (i.e. the expression levels measured at the 12 time-points) and the vector of 12 mark signals. The distributions of Pearson *r* correlation coefficients for the set of (differentially + stably) expressed genes are depicted with box plots and violin plots in Figure 1d. Randomized steady-state and time-course correlation coefficients were computed as described above following a 1,000-permutations scheme on each histone mark's matrix. Briefly, while we kept the original expression matrix, the columns (time-points) of the matrix corresponding to a given mark's signal were permuted without repetition 1,000 times (for an example, see Supplementary Figure 2a, lower panel). In the case of steady-state correlations we report, for each expression time-point, the Pearson *r* averaged over 1,000 rounds of permutation of chromatin time-points (Supplementary Figure 2b, dots). In the case of correlations computed across time-points (time-course), we computed, for each gene, the Pearson *r* averaged over the 1,000 rounds of permutations. The distributions of the resulting coefficients across the set of expressed genes are depicted in Supplementary Figure 2b (box plots and violin plots). Correlations were computed with the R function `cor()`. Permutations without replacement of the chromatin time-points were performed consistently across histone marks with the R function `sample()`, by setting an independent seed for each round of permutations. The correlation values reported in Supplementary Figure 2c are an analogous exercise to Figure 1d on the set of 8,030 differentially expressed genes.

Multivariate Hidden Markov Model analysis

A multivariate Hidden Markov Model (HMM) was fitted to the entire ChIP-seq dataset to approximate the set of underlying chromatin states reported by the 12,248 selected protein-coding genes along the transdifferentiation process. Specifically, we provided as input a matrix of dimensions 146,976 rows $\times$ 9 columns, which collected for each gene and time-point (12,248 genes, 12 time-points) the signal of each of the 9 histone marks after quantile normalization (for a description of these calculations see previous section *ChIP-seq data processing and analysis*). The collective behavior of the nine histone marks along the twelve time-points was modelled as an independent time-series for each gene, using Gaussian distributions. The

model then reprocessed each gene's data to estimate the chromatin state of each gene at each time-point, and provide a time series of chromatin states for each gene. HMM was performed using the R package depmixS4[57], in particular functions `depmix()`, `fit()` and `posterior()` (script: `HMM.wrapper.marks.R`). We repeated the analysis for increasing numbers of states (between 2 and 20), and recorded the log likelihood of each model (the 20-states model reached the maximum number of iterations in EM without convergence). We found that somewhere between five and eight states approximate the elbow point of the log likelihood curve (Supplementary Figure 3a), and observed that the combinations of histone marks represented by five states were consistent with manual inspection of pile-up histone marks profiles in the UCSC genome browser. We thus set for five states. The response parameters of the nine histone marks corresponding to each of these states are reported in Figure 2a. In this case, the *Intercept* values of each histone mark across the five states were re-scaled to a range 0-1 to enable the comparison among different states and marks. HMM sequence hierarchical clustering across the 12,248 genes was performed with the TraMineR[58] and pheatmap (`https://github.com/raivokolde/pheatmap`) R packages (clustering distance: *euclidean*, clustering method: *Ward.D2*). The arc diagram representation in Figure 2c was obtained with the R package arcdiagram (`https://github.com/gastonstat/arcdiagram`).

Decision-tree labelling

In the Methods section *ChIP-seq data processing and analysis* we introduced the distinction between genes with and without peaks of a given mark at a given point in the region of interest (gene body for H3K36me3 and H4K20me1; TSS ±2 Kb for all other marks). Following this first assessment, we classified as unmarked those genes that were consistently unmarked throughout the whole process of transdifferentiation, i.e. with no peaks called at any time-point in the region of interest. Conversely, marked genes reported peak calls of a given mark in the region of interest in at least one time-point (Figure 4a).

Within the set of marked genes, we defined as stably marked (SM) those that did not report significant changes detected by maSigPro[50] over time (FDR ≥ 0.05). On the contrary, differentially marked (DM) genes reported significant changes in a given mark's profile over time (FDR < 0.05). To ensure a multiple testing correction procedure consistent among the nine marks and also with respect to gene expression, maSigPro was run, as described for gene expression (default parameters, replicates handled internally), on the initial set of 12,248 genes, which also included unmarked genes.

The next branch of classification (Figure 4a) was applied only to the set of differentially marked genes that are also differentially expressed. To ensure consistent results among histone marks, the following multiple testing correction procedures were always applied to the set of 8,030 DE genes. For each DE gene, we computed at each time-point the breadth of a given mark's signal, defined as the fraction of the gene's size (from the first annotated region of interest until the last annotated Transcription Termination Site, TTS) covered by peaks of the mark. We refer to this vector of length 12 as the mark's coverage vector. We next considered i) Pearson *r* correlation coefficient between the time-series expression levels and mark's signals; ii) Pearson *r* correlation coefficient between the time-series expression levels and mark's coverage

values; iii) statistical significance of the Needleman-Wunch (NW) dynamic time warping alignment between the time-series expression levels and mark's signals (following Benjamini-Hochberg multiple testing correction; script: `p-adjust.R`). We used as input for the NW alignments (scripts: `NW.alignment.path.R`, `NW.bidirectional.matches.py`) the *z*-score profiles of expression and mark obtained after applying polynomial regression (`degree = 2`) on the original matrices (scripts: `loess.polynomial.regression.R`, `NW.generate.input.matrix.sh`). This procedure was applied to remove the noise due to occasional fluctuations in signal over time. A permutation *p* value for each gene was computed (script: `NW.pvalue.permutation.test.py`), based on a 100,000-permutations scheme (script: `NW.alignment.permutations.R`). To classify a gene as positively correlated, we required at least two of the following conditions: i) Pearson *r* correlation coefficient between the time-series expression levels and mark's signals $\geq$ 0.60 and FDR $<$ 0.05; ii) Pearson *r* correlation coefficient between the time-series expression levels and mark's coverage values $\geq$ 0.60 and FDR $<$ 0.05; iii) NW alignment between the time-series expression levels and mark's signals with FDR $<$ 0.05. For negatively correlated genes, we required at least two of the following conditions: i) Pearson *r* correlation coefficient between the time-series expression levels and mark's signals $\leq$ -0.60 and FDR $<$ 0.05; ii) Pearson *r* correlation coefficient between the time-series expression levels and mark's coverage values $\leq$ -0.60 and FDR $<$ 0.05; iii) NW alignment between the time-series expression levels and mark's signals with FDR $\geq$ 0.05. Genes that did not meet these requirements were classified as uncorrelated. The same decision-tree classification was performed independently for each of the nine histone marks, to ensure comparable results among all modifications (script: `define.6.groups.R`).

Clustering analysis

We considered all 45 combinations between the 9 histone marks and the 5 decision-tree labels described in the previous section. For instance, one combination may be "stably marked + H3K4me3", and another combination may be "positively correlated + H3K27ac". To test the co-occurrence of this pair of combinations, we retrieved the set of DE genes that are labelled "stably marked" for H3K4me3, and the set of DE genes that are labelled "positively correlated" for H3K27ac. The significant overlap between these two sets of genes was tested by the hypergeometric distribution (R function `phyper()`). We repeated this procedure for all possible pairs of combinations. We next clustered the *p* values obtained after applying the Benjamini-Hochberg False Discovery Rate (FDR) multiple testing correction. Hierarchical clustering was performed with the ComplexHeatmap[59] R package (clustering distance = *Manhattan*, clustering method = *Ward.D2*). Cluster correspondence analysis[60] of the 45 categorical variables (combinations of histone marks and decision-tree labels) across the 8,030 selected genes was performed with the R package clustrd[61]. To select the optimal number of clusters and dimensions, we first run the function `tuneclus()` with the following parameters: `nclusrange = 3:10`, `ndimrange = 2:9`, `method = "clusCA"`, `nstart = 100`, `seed = 1234`. This indicated that the optimal number of dimensions and clusters was two and three, respectively. We then obtained the three clusters of genes running the function `clusmca` with the

following parameters: `nclus = 3`, `ndim = 2`, `method = "clusCA"`, `nstart = 100`, `smartStart = NULL`, `gamma = TRUE`, `seed = 1234`. We obtained the same clusters of genes when running the function clusmca with the following parameters: `nclus = 3`, `ndim = 3`, `method = "MCAk"`, `alphak = 0.5`, `nstart = 100`, `smartStart = NULL`, `gamma = TRUE`, `seed = 1234`). This allowed us to explore the clustering of genes also in the third dimension (Figure 4c, Supplementary Figure 4a).

Gene Ontology enrichment analysis

We used the R package GOstats[62] to identify Gene Ontology (GO) terms related to biological processes (BP) and cellular compartments (CC). We set a *p* value threshold of 0.01 to identify significantly enriched terms. For the GO enrichment analysis on the genes contributing to Principal Components (PC) 1 and 2 (described in Results, section *Gene expression recapitulates transdifferentiation more precisely than chromatin*; Figure 1c, Supplementary Table 2), we used the function `get_pca_var()` from the R package factoextra (`https://CRAN.R-project.org/package=factoextra`) to extract the 10% genes (n = 1,066) with the highest contribution to each of the two first principal components. The union of these two sets of genes was used as background for the GO enrichment analysis. We used REVIGO[63] (`http://revigo.irb.hr/`) to summarize the lists of enriched GO terms. For the GO enrichment analysis on the up-regulated genes that belong to the three chromatin clusters (described in Results, section *Chromatin marking is associated with expression specifically at the time of gene activation*), we provided as background the set of 2,103 up-regulated genes. In this case, we used REVIGO and the R package ggplot2[64] to compute and visualize, respectively, maps of the identified GO terms based on their frequency, $-log_{10}$ *p* value, uniqueness and dispensability. Only children terms with dispensability < 0.5 are shown.

Analysis of ENCODE RNA-seq and ChIP-seq data

To investigate differences in gene expression levels and chromatin marking among the three clusters of DE genes in other biological models, we obtained RNA-seq data and ChIP-seq data for histone marks generated by the ENCODE Project[65,66] (`https://www.encodeproject.org/`). Besides B cells and CD14-positive monocytes, which are biologically more similar to pre-B cells and macrophages, respectively, we selected five cancer cell lines (K562, HepG2, GM12878, MCF-7, A549) that are comprehensively characterized by ENCODE ChIP-seq data for the nine histone marks that we have profiled in our study. To assess differences in gene expression levels between the three clusters of DE genes, we obtained gene expression quantifications (with respect to Gencode v24) from polyA+ RNA-seq experiments (accession date: 10/06/2019). We computed, for each gene, the average TPM values between two biological replicates. The list of experiments and datasets' accession IDs used for this analysis is reported in Supplementary Table 5.

To assess differences in chromatin marking, we obtained ChIP-seq data available for the nine histone marks profiled in our study. (Assay title: Histone ChIP-seq; Genome assembly: GRCh38; Output type: replicated peaks or stable peaks; Accession date: 10/06/2019). The list of experiments and datasets'

accession IDs used for this analysis is available in Supplementary Table 6. In all cases, we excluded experiments associated with AUDIT errors. In case of multiple experiments on the same target and cell type, the experiment associated with the lowest number of AUDIT terms was selected. The scripts used to retrieve and filter the ENCODE experiments are: `download.metadata.sh`, `parse.metadata.audit.categories.py`, `retrieve.encode.identifiers.sh`, `parse.list.identifiers.sh`.

For each experiment and cell type, we computed the proportion of genes with at least one peak called over the gene body (H3K36me3, H4K20me1) or in the promoter region (TSS ±2 Kb for all other marks; script: `intersect.peaks.regions.sh`). In the presence of multiple TSSs annotated for the same gene, multiple regions were considered. This is consistent with the analyses described in section *ChIP-seq data processing and analysis*.

Analysis of temporal dynamics

For this analysis we first identified, within the set of 2,103 up-regulated genes, 257 with expression at 0 hours p.i. $<$ 1 TPM. These genes were, therefore, specifically activated during transdifferentiation. Expression and chromatin profiles of each of the considered genes were re-scaled to range 0-100 (script: `rescale.R`): in this way, the minimum and maximum expression level or chromatin signal over the 12 time-points were set to 0% and 100% of up-regulation, respectively. We next considered, for each gene, pairs of consecutive time-points along transdifferentiation (e.g. 0h and 3h; 3h and 6h; 6h and 9h; etc.), and recorded the first time-point at which the expression / chromatin profile crossed ($\geq$) 25%, 50%, 75% and 100% degree of up-regulation (Supplementary Figure 5b). This "crossing" step implies that, in a pair of consecutive time-points, the signal corresponding to the first time-point is, for instance, $<$ 25%, and the signal corresponding to the second time-point is, for instance, $\geq$ 25%. This assessment is performed for each of the four degrees of up-regulation. To ensure monotonic increases consistently across all histone marks, we excluded genes for which this "crossing" step could not be observed for all four degrees of up-regulation in a given mark's time-series profile. This explains the different numbers of genes, among marks, reported in Figure 6a and Supplementary Figure 5e. For a given gene and for each of the four degrees of up-regulation, the recorded time-points (tp) for expression and chromatin profiles were compared, and a label was assigned depending on whether the up-regulation of chromatin signal anticipated ($tp_{mark} < tp_{expression}$), co-occurred ($tp_{mark} = tp_{expression}$) or followed ($tp_{mark} > tp_{expression}$) the up-regulation of gene expression. We analogously compared the up-regulation between pairs of histone marks (Figure 6c, Supplementary Figure 5d). In this case, we analyzed whether the up-regulation of histone mark's signal on row i anticipated ($tp_i < tp_j$) or co-occurred with ($tp_i = tp_j$) the up-regulation of histone mark's signal on column j. To assess whether the specific order of up-regulation in expression levels and chromatin signals depended on the initial level of expression of the genes, these analyses were repeated starting on a set of 629 up-regulated genes with expression at 0 hours p.i. $>$ 25 TPM.

QUANTIFICATION AND STATISTICAL ANALYSIS

Details regarding statistical tests, significance assessment, dispersion and precision measures are reported both in the section *Method details* and in the figures' legends. All statistical analyses were performed using the R language for statistical computation and graphics[67](`http://www.R-project.org/`). In all cases, the multiple testing correction procedure was performed by applying the Benjamini-Hochberg[68] False Discovery Rate (FDR). Wilcoxon rank-sum tests were performed with the `wilcox.test()` R function in a two-sided manner.

When not specified, plots were made using the R package ggplot2[64]. All box plots depict the first and third quartiles as the lower and upper bounds of the box, with a band inside the box showing the median value and whiskers representing 1.5x the interquartile range. All scripts used in the analyses are publicly available (see the *Data and Code Availability* statement).

Acknowledgments

We thank Thomas Graf and Francesca Rapino for donating BLaER1 cells and for helpful discussions. We thank Sebastian Ullrich, Carme Arnan and Vasilis Ntasis for helpful discussion about the data. We thank Montserrat Corominas, Guillaume Filion and Luciano Di Croce for insightful suggestions. We thank Diego Garrido-Martín, Manuel Muñoz and Javier Martín-Vallejo for statistical advice. We also thank the Genomics, the Flow Cytometry and the Bioinformatics Core Units of the CRG (Barcelona, Spain). We thank Romina Garrido for administrative support. We thank the ENCODE Consortium, in particular Thomas Gingeras', Bradley Bernstein's, John Stamatoyannopoulos' and Peggy Farhnam's laboratories, for data production. This work was performed under the financial support of the European Community under the FP7 program (ERC-2011-AdG-294653-RNA-MAPS). B.B. is supported by the fellowship 2017FI_B00722 from the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement (Generalitat de Catalunya) and the European Social Fund (ESF). C.C.K. is supported by the CERCA Programme / Generalitat de Catalunya and FEDER under project VEIS-001-P-001647. B.R.C. is supported by the Ministerio de Ciencia, Innovación y Universidades de España under grant FJCI-2017-34353. We also acknowledge Agencia Estatal de Investigación (AEI) and FEDER under project PGC2018-094017-B-I00. All authors acknowledge the support of the Ministerio de Ciencia, Innovación y Universidades de España to the EMBL partnership, the Centro de Excelencia Severo Ochoa, and the CERCA Programme / Generalitat de Catalunya. Figures 1b and 7a were created with `https://biorender.com`.

Author Contributions

R.G. and R.J. conceived the project. B.B., S.P-L. and R.G. designed the study. B.B. performed the computational analyses. A.A. performed the ChIP-seq experiments. A.E. and M.S. performed the RNA-seq experiments. C.C.K. and E.P. contributed to data quality check and processing. C.C.K., R.N., M.R-R. and

B.R.C. contributed tools and ideas to perform experiments and computational analyses. B.B., S.P-L. and R.G. wrote the manuscript with the contribution of all authors.

Competing Interests

The authors declare no competing interest.

References

[1] Hon, G., Wang, W. & Ren, B. Discovery and Annotation of Functional Chromatin Signatures in the Human Genome. *PLoS Computational Biology* **5**, e1000566 (2009).

[2] Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).

[3] Schneider, R. *et al.* Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nature Cell Biology* **6**, 73–77 (2004).

[4] Trojer, P. & Reinberg, D. Facultative Heterochromatin: Is There a Distinctive Molecular Signature? *Molecular Cell* **28**, 1–13 (2007).

[5] Hansen, K. H. *et al.* A model for transmission of the H3K27me3 epigenetic mark. *Nature Cell Biology* **10**, 1291–1300 (2008).

[6] Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).

[7] Karlić, R., Chung, H. R., Lasserre, J., Vlahoviček, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2926–2931 (2010).

[8] Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology* **13**, R53 (2012).

[9] Greer, E. L. & Shi, Y. Histone methylation: A dynamic mark in health, disease and inheritance. *Nature Reviews Genetics* **13**, 343–357 (2012).

[10] Nègre, N. *et al.* A cis-regulatory map of the Drosophila genome. *Nature* **471**, 527–531 (2011).

[11] Hödl, M. & Basler, K. Transcription in the absence of histone H3.2 and H3K4 methylation. *Current Biology* **22**, 2253–2257 (2012).

[12] Pérez-Lluch, S. *et al.* Absence of canonical marks of active chromatin in developmentally regulated genes. *Nature Genetics* **47**, 1158–1167 (2015).

[13] Vandenbon, A., Kumagai, Y., Lin, M., Suzuki, Y. & Nakai, K. Waves of chromatin modifications in mouse dendritic cells in response to LPS stimulation. *Genome Biology* **19**, 138 (2018).

[14] Le Martelot, G. *et al.* Genome-Wide RNA Polymerase II Profiles and RNA Accumulation Reveal Kinetics of Transcription and Associated Epigenetic Changes During Diurnal Cycles. *PLoS Biology* **10**, e1001442 (2012).

[15] Wang, S. *et al.* A dynamic and integrated epigenetic program at distal regions orchestrates transcriptional responses to VEGFA. *Genome Research* **29**, 193–207 (2019).

[16] Rach, E. A. *et al.* Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level. *PLoS Genetics* **7**, e1001274 (2011).

[17] Mercer, E. M. *et al.* Multilineage Priming of Enhancer Repertoires Precedes Commitment to the B and Myeloid Cell Lineages in Hematopoietic Progenitors. *Immunity* **35**, 413–425 (2011).

[18] Kaikkonen, M. U. *et al.* Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular Cell* **51**, 310–325 (2013).

[19] Dorighi, K. M. *et al.* Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Molecular Cell* **66**, 568–576 (2017).

[20] Cao, K. *et al.* An Mll4/COMPASS-Lsd1 epigenetic axis governs enhancer function and pluripotency transition in embryonic stem cells. *Science Advances* **4**, eaap8747 (2018).

[21] Rapino, F. *et al.* C/EBPα Induces Highly Efficient Macrophage Transdifferentiation of B Lymphoma and Leukemia Cell Lines and Impairs Their Tumorigenicity. *Cell Reports* **3**, 1153–1163 (2013).

[22] DI Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**, 316–319 (2017).

[23] Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766–D773 (2019).

[24] Ninova, M., Tóth, K. F. & Aravin, A. A. The control of gene expression and cell identity by H3K9 trimethylation. *Development* **146**, dev.181180 (2019).

[25] Pervouchine, D. D. *et al.* Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nature Communications* **6**, 1–11 (2015).

[26] Ernst, J. & Kellis, M. ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215–216 (2012).

[27] Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9**, 473–476 (2012).

[28] Song, J. & Chen, K. C. Spectacle: Fast chromatin state annotation using spectral learning. *Genome Biology* **16**, 33 (2015).

[29] Zhang, Y., An, L., Yue, F. & Hardison, R. C. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Research* **44**, 6721–6731 (2016).

[30] Zhang, Y. & Hardison, R. C. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Research* **45**, 9823–9836 (2017).

[31] Waddington, C. H. The epigenotype. *Endeavour* **1**, 18–20 (1942).

[32] Berger, S. L., Kouzarides, T., Shiekhattar, R. & Shilatifard, A. An operational definition of epigenetics. *Genes and Development* **23**, 781–783 (2009).

[33] Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* **21**, 381–395 (2011).

[34] Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).

[35] Sekhon, A., Singh, R. & Qi, Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics* **34**, i891–i900 (2018).

[36] Yin, Q., Wu, M., Liu, Q., Lv, H. & Jiang, R. DeepHistone: A deep learning approach to predicting histone modifications. *BMC Genomics* **20**, 193 (2019).

[37] Henikoff, S. & Shilatifard, A. Histone modification: Cause or cog? *Trends in Genetics* **27**, 389–396 (2011).

[38] Rickels, R. *et al.* Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nature Genetics* **49**, 1647–1653 (2017).

[39] Krogan, N. J. *et al.* The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: Linking transcriptional elongation to histone methylation. *Molecular Cell* **11**, 721–729 (2003).

[40] Di Tullio, A., Vu Manh, T. P., Schubert, A., Månsson, R. & Graf, T. CCAAT/enhancer binding protein α (C/EBPα)-induced transdifferentiation of pre-B cells into macrophages involves no overt retrodifferentiation. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 17016–17021 (2011).

[41] Simpson, E. H. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **13**, 238–241 (1951).

[42] Wiencke, J. K., Zheng, S., Morrison, Z. & Yeh, R. F. Differentially expressed genes are marked by histone 3 lysine 9 trimethylation in human cancer cells. *Oncogene* **27**, 2412–2421 (2008).

[43] Rybtsova, N. *et al.* Transcription-coupled deposition of histone modifications during MHC class II gene activation. *Nucleic Acids Research* **35**, 3431–3441 (2007).

[44] Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids Research* **45**, D626–D634 (2017).

[45] Bussmann, L. H. *et al.* A Robust and Highly Efficient Immune Cell Reprogramming System. *Cell Stem Cell* **5**, 554–566 (2009).

[46] Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise reprogramming of B cells into macrophages. *Cell* **117**, 663–676 (2004).

[47] Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

[48] Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

[49] Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

[50] Nueda, M. J., Tarazona, S. & Conesa, A. Next maSigPro: Updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* **30**, 2598–2602 (2014).

[51] Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: Fast, accurate and versatile alignment by filtration. *Nature Methods* **9**, 1185–1188 (2012).

[52] Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

[53] Cuscó, P. & Filion, G. J. Zerone: A ChIP-seq discretizer for multiple replicates with built-in quality control. *Bioinformatics* **32**, 2896–2902 (2016).

[54] Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22**, 1813–1831 (2012).

[55] Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

[56] Pohl, A. & Beato, M. bwtool: A tool for bigWig files. *Bioinformatics* **30**, 1618–1619 (2014).

[57] Visser, I. & Speekenbrink, M. depmixS4: An R package for hidden markov models. *Journal of Statistical Software* **36**, 1–21 (2010).

[58] Gabadinho, A., Ritschard, G., Müller, N. S. & Studer, M. Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software* **40**, 1–37 (2011).

[59] Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

[60] van de Velden, M., D’Enza, A. I. & Palumbo, F. Cluster Correspondence Analysis. *Psychometrika* **82**, 158–185 (2017).

[61] Markos, A., D’Enza, A. I. & van de Velden, M. Beyond tandem analysis: Joint dimension reduction and clustering in R. *Journal of Statistical Software* **91**, 1–24 (2019).

[62] Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007).

[63] Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE* **6**, e21800 (2011).

[64] Wickham H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, New York City, New York, 2009).

[65] Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research* **46**, D794–D801 (2018).

[66] Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

[67] Team R. C. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2017).

[68] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

Dynamics of gene expression and chromatin marking during cell state transition

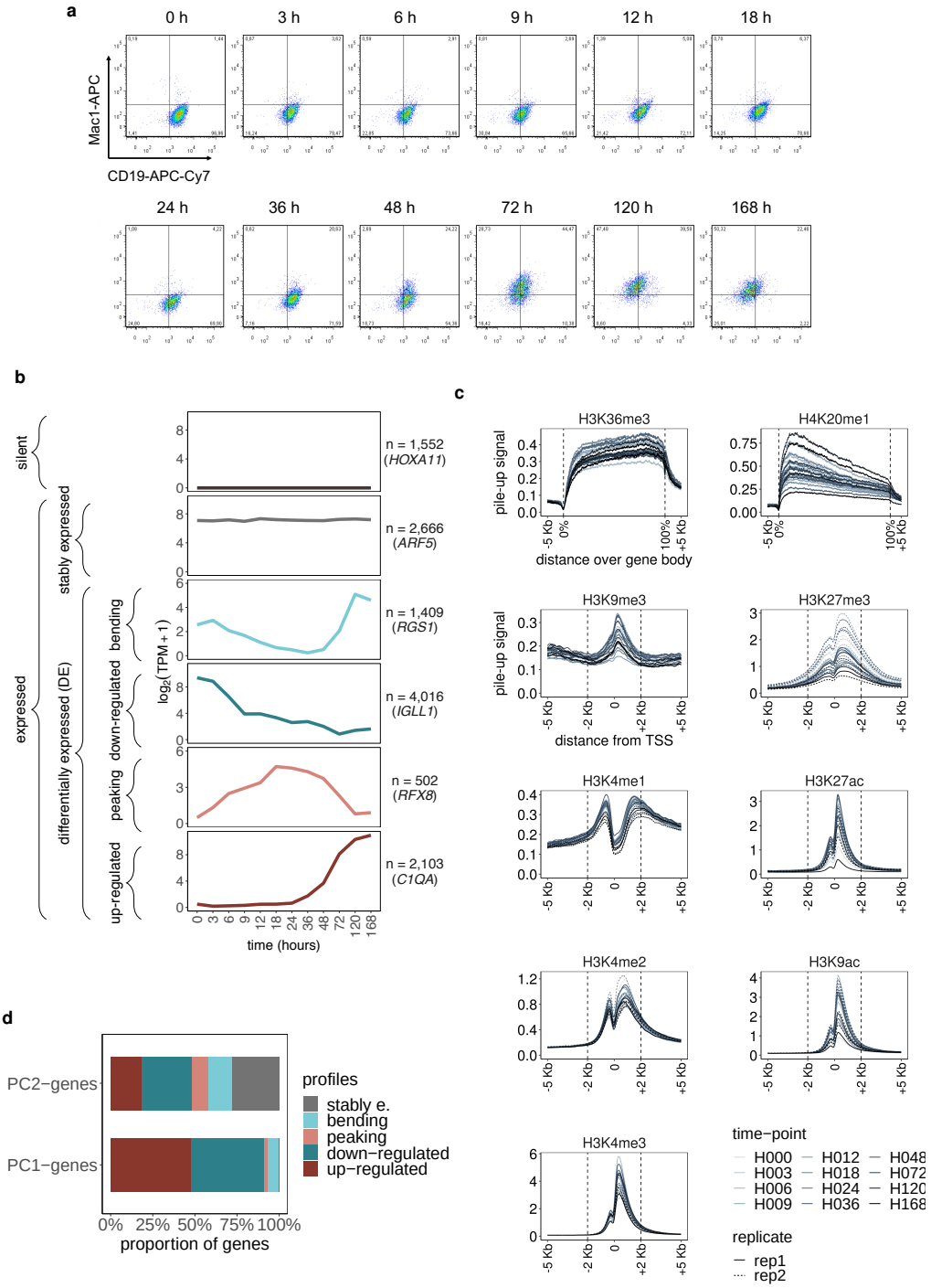
Beatrice Borsari, Amaya Abad, Cecilia C. Klein, Ramil Nurtdinov, Alexandre Esteban, Emilio Palumbo, Marina Ruiz-Romero, María Sanz, Bruna R. Correa, Rory Johnson, Sílvia Pérez-Lluch and Roderic Guigó

Supplementary Information

Supplementary Figures 1-6

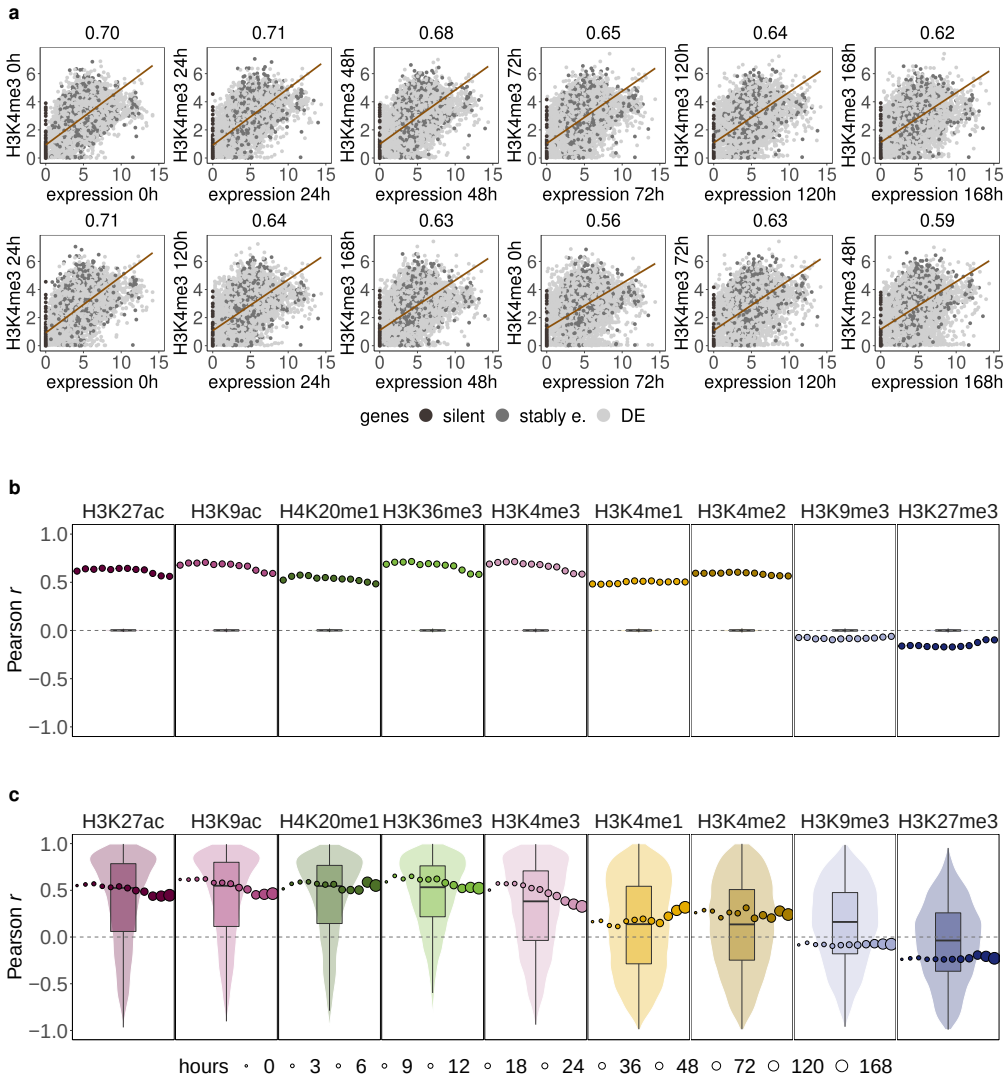
Supplementary Tables 1, 3-6

Supplementary Figure 1



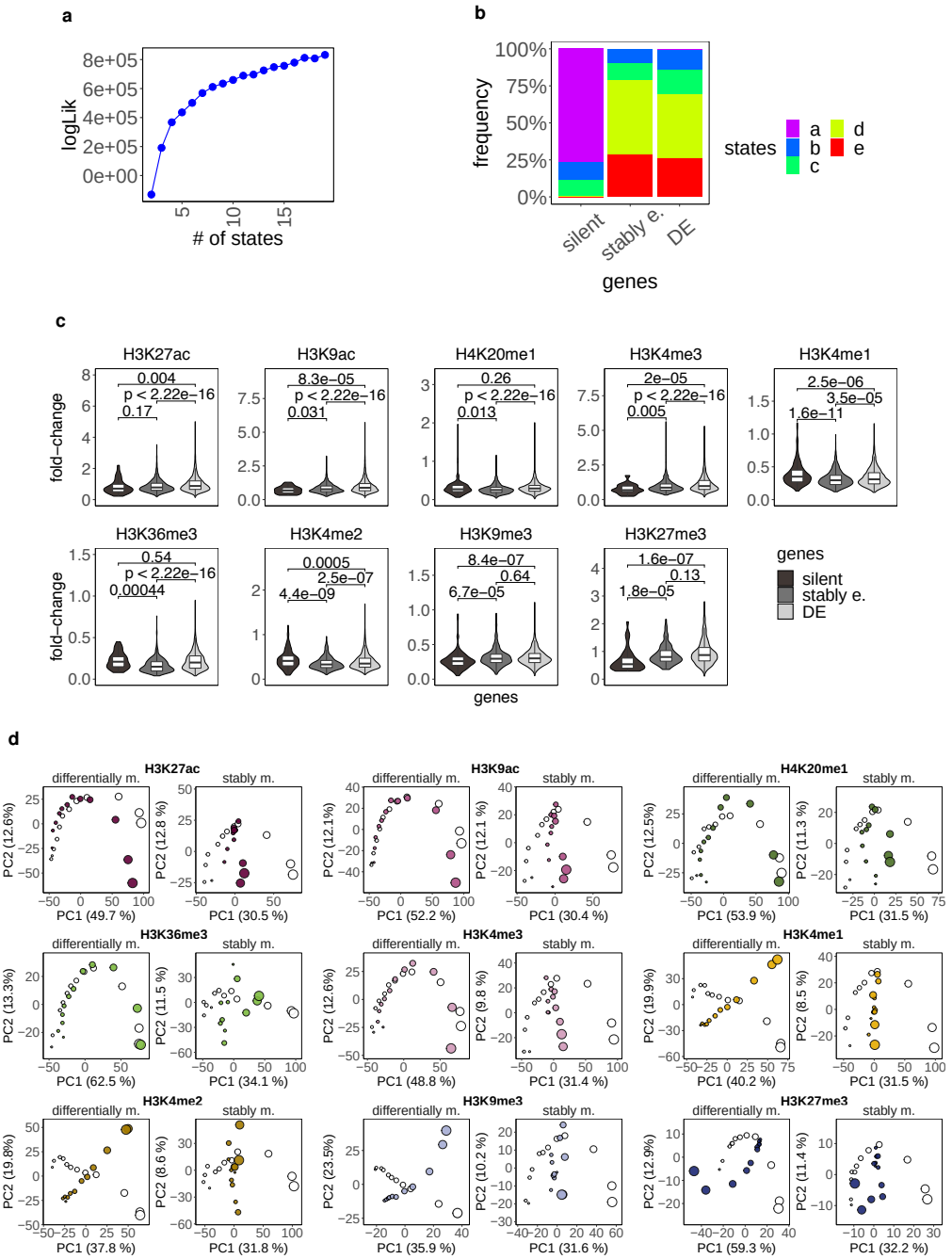
Supplementary Figure 1: Characterization of gene expression and histone modifications' profiles during transdifferentiation — See also Figure 1, Supplementary Tables 1-2. **a**: Flow-cytometry plots assessing expression of CD19 and Mac-1 antigens at the 12 time-points monitored during transdifferentiation. **b**: Classification of time-series expression profiles. We selected a set of 12,248 protein-coding genes, which comprises 1,552 not expressed genes (0 TPM in all time-points and biological replicates) and 10,696 expressed genes ($\geq$ 5 TPM in at least one time-point, and in both biological replicates). Within the set of expressed genes, we distinguished between genes with a stable expression profile throughout transdifferentiation (stably expressed; maSigPro FDR $\geq$ 0.05; n = 2,666), and genes showing significant changes in gene expression over time (differentially expressed or DE; maSigPro FDR $<$ 0.05; n = 8,030). DE genes were further characterized into bending (1,409), down-regulated (4,016), peaking (502) and up-regulated (2,103) genes. Examples of genes belonging to the six types of expression profiles are provided. Gene expression values are reported in $\log_2$ (TPM + 1). **c**: Average pile-up signal over the gene body $\pm$ 5 Kb (H3K36me3 and H4K20me1), or promoter regions $\pm$ 5 Kb from the Transcription Start Site (TSS; all other marks), computed at each of the 12 time-points. The vertical dashed lines mark the selected region of $\pm$ 2 Kb around the TSS. **d**: Proportion of genes contributing to the first two principal components (PC1 and PC2) of the joint PCA on expression and chromatin marks (Figure 1c), that are classified as bending, down-regulated, peaking, up-regulated or stably expressed.

Supplementary Figure 2



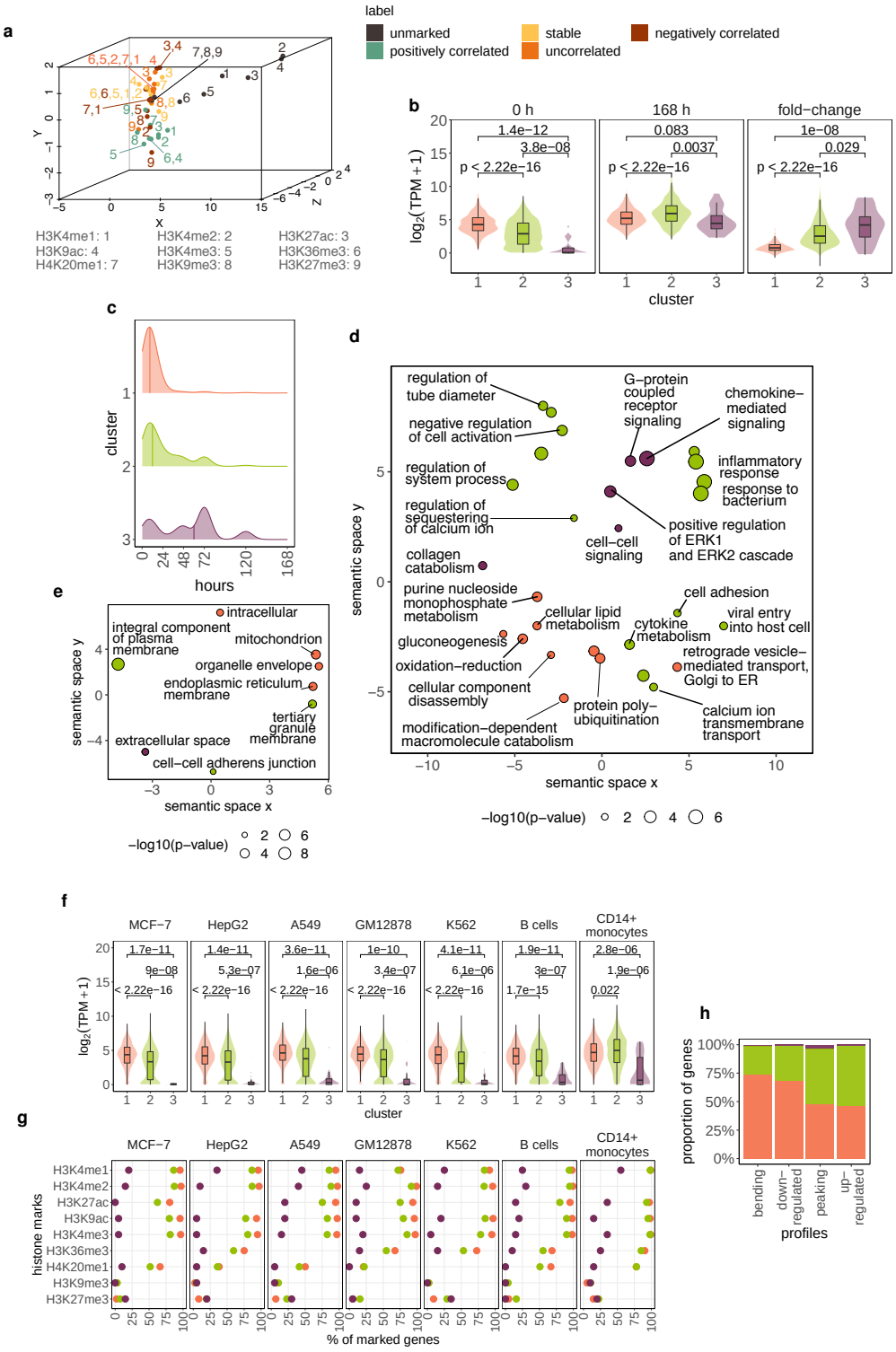
Supplementary Figure 2: The correlation between chromatin marking and gene expression over time is lower than the one reported in steady-state conditions. — See also Figure 1. **a**: Steady-state correlations between expression levels (x-axis) and H3K4me3 signals (y-axis) computed on the set of 12,248 genes (silent genes: dark gray; stably expressed genes: gray; DE genes: light gray). Upper panel: Pearson *r* between expression levels and H3K4me3 signals at paired time-points (0, 24, 48, 72, 120 and 168 hours). The magnitude of the correlation is reported on the top of each scatterplot. The linear regression line is depicted in brown. Lower panel: analogous representation after randomly shuffling the H3K4me3 signals among time-points. As a result, we computed the expression *vs.* chromatin correlation between unpaired time-points (0h - 24h; 24h - 120h; 48h - 168h; 72h - 0h; 120h - 72h; 168h - 48h). Steady-state correlations computed on the whole set of genes are large despite the randomization of the data. **b**: Steady-states (dots) and time-course (violin and box plots) correlation values between expression levels and chromatin signals (analogous to Figure 1d) computed after randomly permuting the genes' signals of a given mark among time-points. In all cases we report the Pearson *r* values averaged over 1,000 permutations. For steady-states correlations, the median Pearson *r* values across time-points are: H3K27ac: 0.63; H3K9ac: 0.68; H4K20me1: 0.54; H3K36me3: 0.69; H3K4me3: 0.69; H3K4me1: 0.50; H3K4me2: 0.59; H3K9me3: -0.08; H3K27me3: -0.16. For time-course correlations, the median Pearson *r* values across genes are 0 ($|r| < 0.001$) for all marks. **c**: Steady-states (dots) and time-course (violin and box plots) correlation values between expression levels and chromatin signals (analogous to Figure 1d), computed after removing stably expressed and silent genes (i.e. only on the set of differentially expressed genes). The median steady-state Pearson *r* values for each mark are: H3K27ac: 0.53; H3K9ac: 0.58; H4K20me1: 0.56; H3K36me3: 0.60; H3K4me3: 0.51; H3K4me1: 0.17; H3K4me2: 0.26; H3K9me3: -0.08; H3K27me3: -0.23. The median time-course Pearson *r* values for each mark are: H3K27ac: 0.53; H3K9ac: 0.55; H4K20me1: 0.55; H3K36me3: 0.53; H3K4me3: 0.38; H3K4me1: 0.14; H3K4me2: 0.14; H3K9me3: 0.16; H3K27me3: -0.04. Silent genes contribute substantially to the steady state correlations, and partially contribute to the differences observed in Figure 1d between steady-state and time-course correlations, since the latter cannot be computed for silent genes (see Methods).

Supplementary Figure 3



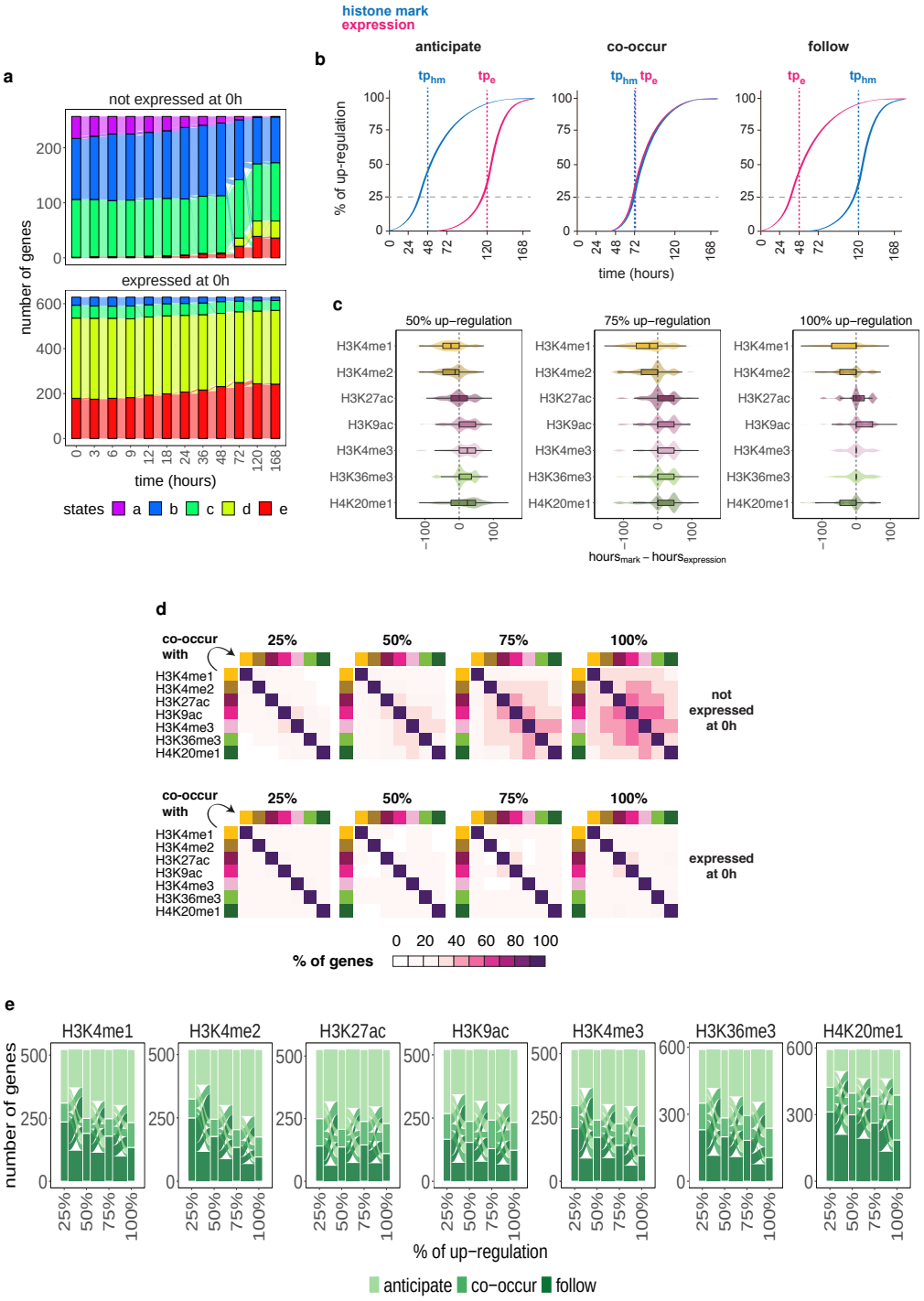
Supplementary Figure 3: Changes in chromatin marking over time can be uncoupled from changes in gene expression — See also Figures 2-3, Supplementary Tables 3-4. **a**: Log likelihood values for HMM models with increasing number of states (between 2 and 20). **b**: Frequency of the five states observed along the twelve time-points of transdifferentiation in the HMM-sequence profiles of the sets of silent, stably expressed and differentially expressed (DE) genes. **c**: Distributions of genes' fold-change (FC: difference between maximum and minimum signals along transdifferentiation) for each histone mark. Differences in FC among sets of silent, stably expressed and DE genes were statistically assessed with Wilcoxon Rank-Sum test (two-sided). The magnitude of chromatin changes observed in stably expressed and silent genes is, in some cases, comparable to (H4K20me1 and H3K36me3 for silent; H3K27me3 and H3K9me3 for stably expressed), or even larger (H3K4me1 and H3K4me2 for silent) than the one observed for DE genes. **d**: Trajectories of transdifferentiation derived from a Principal Component Analysis performed jointly on expression and each histone mark's time-series profiles of DE genes, distinguishing between differentially marked (left) and stably marked (right) genes. Across all histone marks, transdifferentiation trends are clearer using the former set of genes, suggesting that the different resolution of PCA trends initially observed (Figure 1c) may be explained by the different amount of changes observed, over time, across histone marks. Unexpectedly, H3K4me1-, H3K4me2- and H3K9me3-differentially marked genes show a contrasting profile for expression and chromatin modifications along PC2, but different to the pattern observed for H3K27me3.

Supplementary Figure 4



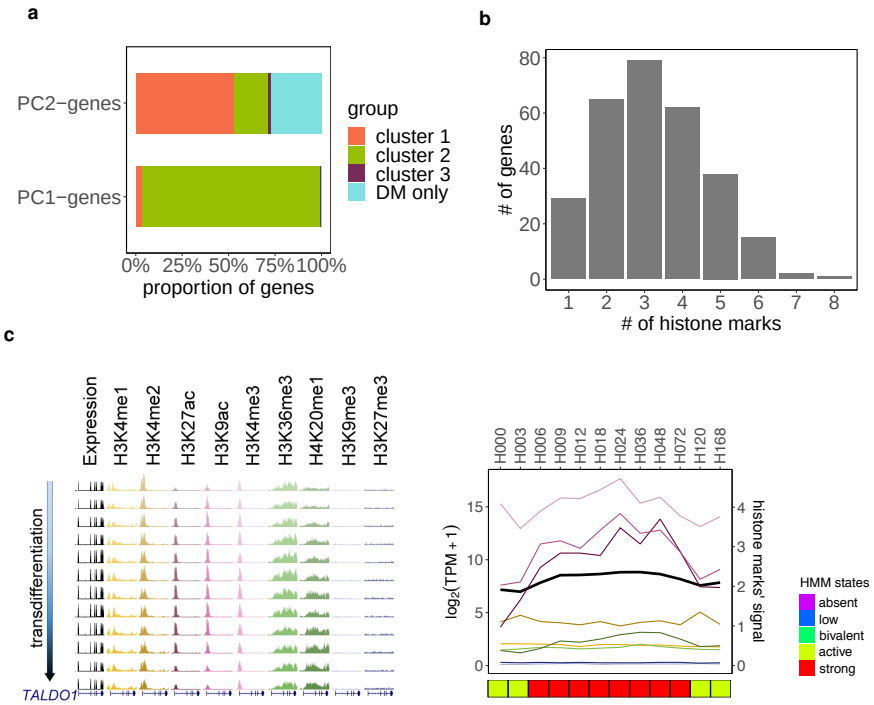
Supplementary Figure 4: Genes in different stages of activation are associated with specific chromatin and gene expression patterns, and perform distinct functions — See also Figures 4-5, Supplementary Tables 5-6. **a**: Three-dimensional representation of the combinations of labels and histone marks (analysis attributes). The color code for the labels is analogous to Figure 4a. Histone marks are represented by numbers. **b**: Distributions of gene expression levels at 0 and 168 hours p.i., and fold-change (FC) in gene expression (168h - 0h) for up-regulated genes that belong to clusters 1-3. Differences in gene expression levels among clusters were assessed with the Wilcoxon Rank-Sum test (two-sided). **c**: Density plot reporting the time-point at which the time-series expression profiles of up-regulated genes in clusters 1-3 reach a degree of up-regulation $\geq$ 25%. For this analysis, the time-series expression profile of each gene was re-scaled to a 0-100% range. **d**: Multidimensional scaling-based representation of the semantic dissimilarities between non-redundant Gene Ontology Biological Process terms enriched among up-regulated genes in clusters 1-3. Each circle represents a term, with the size and the color of the circle denoting the $-log_{10}$ *p* value and the cluster of the term, respectively. GO terms that lie close to each other are semantically more similar. **e**: Analogous representation to Supplementary Figure 4d for cellular compartments. Cluster 1 genes are associated with metabolic functions mostly performed in intracellular compartments, suggesting a more housekeeping nature of these up-regulated genes. Cluster 2 genes perform functions related to the inflammatory response and to the cell membrane and projections, and are thus more likely to be involved in the transition from pre-B cells to macrophages. Cluster 3 genes are associated with macrophage-specific functions. **f**: Analysis of ENCODE RNA-seq data available for five cancer cell lines (MCF7, HepG2, A549, GM12878, K562), and two primary cell types (B cells and CD14+ monocytes) that are biologically similar to the cell types present at the beginning (pre-B) and at the end (macrophages), respectively, of our transdifferentiation model. Distributions of gene expression levels for up-regulated genes that belong to clusters 1-3. Differences in gene expression levels among clusters were assessed using Wilcoxon Rank-Sum test (two-sided). **g**: Analysis of ENCODE ChIP-seq data for the nine histone marks we have monitored along transdifferentiation in five cancer cell lines (MCF7, HepG2, A549, GM12878, K562) and two primary cell types (B cells and CD14+ monocytes). Proportions (%) of marked genes at gene body (H3K36me3, H4K20me1) and promoter regions (all other marks) among up-regulated genes in clusters 1-3. **h**: Percent stacked bar plot depicting the proportion of bending, down-regulated, peaking and up-regulated genes that belong to the three clusters.

Supplementary Figure 5



Supplementary Figure 5: The up-regulation of chromatin marks and gene expression follows a precise order only during the initial stage of gene activation — See also Figure 6. **a**: Alluvial plot describing the HMM time-series profiles for the 257 (upper panel) and 629 (lower panel) up-regulated genes that are not expressed ($<$ 1 TPM) and expressed ($>$ 25 TPM), respectively, at 0 hours p.i. **b**: Graphical representation of cases in which the up-regulation of chromatin signal anticipates (left), co-occurs with (middle), or follows (right) the up-regulation of gene expression. The expression and histone marks' time-series profiles of each gene were re-scaled to a 0-100% range prior to the analysis. We considered four degrees of up-regulation (25%, 50%, 75% and 100%) and computed, for each gene and histone mark, the time-point at which the expression and chromatin re-scaled values reach each of the four degrees of up-regulation. Here we depict a representation for the degree of up-regulation of 25%. **c**: Lag (hours) between 50%, 75% and 100% up-regulation in histone marks' signal and expression level for the 257 up-regulated genes not expressed at 0 hours p.i. Negative lags correspond to changes in chromatin marks anticipating changes in gene expression; positive lags correspond to changes in chromatin marks following changes in gene expression. **d**: Analogous representation to Figure 6c for co-occurring changes between pairs of histone marks in genes that are either silent (upper panel) or expressed (lower panel) at 0 hours. For genes specifically activated during transdifferentiation (upper panel), the amount of co-occurring changes increases towards the end of the up-regulation process. **e**: Analogous representation to Figure 6a for the 629 up-regulated genes expressed at 0 hours p.i.

Supplementary Figure 6



Supplementary Figure 6: Chromatin marking cannot be fully recapitulated by gene expression — See also Figure 1, Supplementary Figure 1. **a**: Proportion of genes contributing to the first two principal components (PC1 and PC2) of the joint PCA on expression and chromatin marks in Figure 1c, that belong to the three clusters of DE genes (clusters 1-3) or that are stably expressed and differentially marked (“DM only”). While genes contributing to the transition from pre-B cells to macrophages (pc1-contributing genes, Figure 1c, Supplementary Figure 1d) show the canonical correlation with chromatin changes (cluster 2), a considerable fraction of genes involved in the intermediate stages of transdifferentiation (pc2-contributing genes) display expression and chromatin changes uncoupled from one another (cluster 1, or stably expressed and differentially marked - “DM only”). This further supports the hypothesis that chromatin changes are involved in a transient de-differentiation from pre-B cells into an intermediate state, and re-differentiation into macrophages. **b**: Among the set of stably expressed and differentially marked genes contributing to PC2 (“DM only” genes in Supplementary Figure 6a), number of genes with variable chromatin profiles for increasing numbers of histone marks. For instance, 79 genes present changes in three histone modifications along transdifferentiation. **c**: Example of a stably expressed gene (*TALDO1*) contributing to PC2 in the PCA in Figure 1c, and showing significant changes in some chromatin profiles along transdifferentiation. Expression and chromatin tracks from one biological replicate are displayed, as well as normalized line plots averaging the signal from the two replicates. Profiles of HMM states are shown at the bottom.

Supplementary Table 1

histone marks	silent		expressed	
	unmarked	marked	unmarked	marked
H3K4me1	1,165 (75.1%)	387 (24.9%)	52 (0.5%)	10,644 (99.5%)
H3K4me2	1,225 (78.9%)	327 (21.1%)	45 (0.4%)	10,651 (99.6%)
H3K9ac	1,500 (96.6%)	52 (3.4%)	130 (1.2%)	10,566 (98.8%)
H3K27ac	1,478 (95.2%)	74 (4.8%)	159 (1.5%)	10,537 (98.5%)
H3K4me3	1,488 (95.9%)	64 (4.1%)	183 (1.7%)	10,513 (98.3%)
H3K36me3	1,444 (93.0%)	108 (7.0%)	354 (3.3%)	10,342 (96.7%)
H4K20me1	1,348 (86.9%)	204 (13.1%)	1,851 (17.3%)	8,845 (82.7%)
H3K9me3	990 (63.8%)	562 (36.2%)	7,201 (67.3%)	3,495 (32.7%)
H3K27me3	1,371 (88.3%)	181 (11.7%)	9,421 (88.1%)	1,275 (11.9%)

Supplementary Table 1: Numbers of unmarked and marked genes within the sets of 1,552 silent and 10,696 expressed genes — See also Figure 1 and Supplementary Figure 1. For a given histone mark, unmarked genes have no peaks called at any time-point in the region of interest, while marked genes have peaks called in the region of interest in at least one time-point (see Methods).

Supplementary Table 2: GO terms significantly enriched among genes contributing to Principal Components 1 and 2 — See also Figure 1 and Supplementary Figure 1. The list of terms refers to Biological Processes.

Supplementary Table 3

histone marks	unmarked	marked		differentially marked		
		stably	differentially	positively c.	uncorrelated	negatively c.
H3K4me1	41 (0.5%)	4,591 (57.2%)	3,398 (42.3%)	1,491 (43.9%)	1,457 (42.9%)	450 (13.2%)
H3K4me2	32 (0.4%)	5,074 (63.2%)	2,924 (36.4%)	1,248 (42.7%)	1,316 (45.0%)	360 (12.3%)
H3K9ac	107 (1.3%)	2,996 (37.3%)	4,927 (61.4%)	3,239 (65.8%)	1,548 (31.4%)	140 (2.8%)
H3K27ac	135 (1.7%)	2,835 (35.3%)	5,060 (63.0%)	3,065 (60.6%)	1,761 (34.8%)	234 (4.6%)
H3K4me3	150 (1.9%)	4,310 (53.7%)	3,570 (44.4%)	2,048 (57.4%)	1,402 (39.3%)	120 (3.3%)
H3K36me3	283 (3.5%)	4,801 (59.8%)	2,946 (36.7%)	2,472 (83.9%)	459 (15.6%)	15 (0.5%)
H4K20me1	1,403 (17.5%)	2,484 (30.9%)	4,143 (51.6%)	2,782 (67.2%)	1,256 (30.3%)	105 (2.5%)
H3K9me3	5,363 (66.8%)	1,698 (21.1%)	969 (12.1%)	253 (26.1%)	615 (63.5%)	101 (10.4%)
H3K27me3	6,988 (87.0%)	362 (4.5%)	680 (8.5%)	40 (5.9%)	347 (51.0%)	293 (43.1%)

Supplementary Table 3: Decision-tree labelling of differentially expressed genes — See also Figures 3-4, Supplementary Figure 3. Left side: numbers of unmarked and marked genes within the set of DE genes. Marked genes are further separated into genes that are either stably or differentially marked (i.e. have stable or variable chromatin profiles during transdifferentiation). The percentages refer to the total number of DE genes (n = 8,030). Right side: within the set of differentially marked genes, we distinguish between genes that are positively correlated, uncorrelated or negatively correlated with gene expression over time (see Methods). The percentages in this case are computed with respect to the number of differentially marked genes found for each histone modification.

Supplementary Table 4

histone mark	silent			stably expressed		
	unmarked	marked		unmarked	marked	
		stably	differentially		stably	differentially
H3K4me1	1,165 (75.1%)	114 (7.3%)	273 (17.6%)	11 (0.4%)	1,711 (64.2%)	944 (35.4%)
H3K4me2	1,225 (78.9%)	91 (5.9%)	236 (15.2%)	13 (0.5%)	1,904 (71.4%)	749 (28.1%)
H3K9ac	1,500 (96.6%)	15 (1%)	37 (2.4%)	23 (0.9%)	1,338 (50.2%)	1,305 (48.9%)
H3K27ac	1,478 (95.2%)	28 (1.8%)	46 (3%)	24 (0.9%)	1,197 (44.9%)	1,445 (54.2%)
H3K4me3	1,488 (95.9%)	30 (1.9%)	34 (2.2%)	33 (1.2%)	1,741 (65.3%)	892 (33.5%)
H3K36me3	1,444 (93%)	78 (5%)	30 (1.9%)	71 (2.7%)	2,204 (82.7%)	391 (14.7%)
H4K20me1	1,348 (86.9%)	88 (5.7%)	116 (7.5%)	448 (16.8%)	1,221 (45.8%)	997 (37.4%)
H3K9me3	990 (63.8%)	445 (28.7%)	117 (7.5%)	1,838 (68.9%)	558 (20.9%)	270 (10.1%)
H3K27me3	1,371 (88.3%)	138 (8.9%)	43 (2.8%)	2,433 (91.3%)	125 (4.7%)	108 (4.1%)

Supplementary Table 4: Absent, stable and differential chromatin marking over time among silent and stably expressed genes — See also Figure 3, Supplementary Figure 3. Numbers of unmarked and marked genes within the sets of 1,552 silent and 2,666 stably expressed genes. Marked genes are further separated into genes that are either stably or differentially marked.

Supplementary Table 5

Experiment ID	Accession file ID	Replicate	Biosample term name
ENCSR000CON	ENCFF369ZNM	1	A549
ENCSR000CON	ENCFF627QMV	2	A549
ENCSR000CTV	ENCFF485EUP	1	B cell
ENCSR000CTV	ENCFF231GYC	2	B cell
ENCSR000CUC	ENCFF299BIL	1	CD14-positive monocyte
ENCSR000CUC	ENCFF397DFK	2	CD14-positive monocyte
ENCSR000AED	ENCFF902UYP	1	GM12878
ENCSR000AED	ENCFF550OHK	2	GM12878
ENCSR000CPE	ENCFF004HYK	1	HepG2
ENCSR000CPE	ENCFF401KRE	2	HepG2
ENCSR000CPH	ENCFF172GIN	1	K562
ENCSR000CPH	ENCFF768TKT	2	K562
ENCSR000CPT	ENCFF009GDJ	1	MCF-7
ENCSR000CPT	ENCFF885LEQ	2	MCF-7

Supplementary Table 5: ENCODE PolyA+ RNA-seq experiments in seven cell types — See also Supplementary Figure 4. The ENCODE accession numbers allow to uniquely identify the experiment and gene expression quantification file (tsv) on the ENCODE portal (https://www.encodeproject.org/).

Supplementary Table 6

Histone mark	Experiment ID	Accession file ID	Biosample term name
H3K27ac	ENCSR000AUI	ENCFF268BMM	A549
H3K27me3	ENCSR000AUK	ENCFF368SNX	A549
H3K4me1	ENCSR000AUM	ENCFF761SFV	A549
H3K4me2	ENCSR000AVI	ENCFF260MGY	A549
H3K4me3	ENCSR000DPD	ENCFF820IQP	A549
H3K9ac	ENCSR000ASV	ENCFF649ABE	A549
H3K9me3	ENCSR000AUN	ENCFF900ULD	A549
H4K20me1	ENCSR000AUO	ENCFF505MWT	A549
H3K27ac	ENCSR000AUP	ENCFF041HKG	B cell
H3K27me3	ENCSR162DGX	ENCFF428KOX	B cell
H3K36me3	ENCSR424XBP	ENCFF649XGE	B cell
H3K4me1	ENCSR290YLQ	ENCFF778RHF	B cell
H3K4me2	ENCSR000AUY	ENCFF615MAT	B cell
H3K4me3	ENCSR878JSF	ENCFF225QYU	B cell
H3K9ac	ENCSR799SLA	ENCFF890NWX	B cell
H3K9me3	ENCSR005WWZ	ENCFF281WGS	B cell
H4K20me1	ENCSR000AVJ	ENCFF856EUL	B cell
H3K27ac	ENCSR000ASJ	ENCFF239LOH	CD14-positive monocyte
H3K27me3	ENCSR000ASK	ENCFF930KLN	CD14-positive monocyte
H3K36me3	ENCSR000ASL	ENCFF108MXF	CD14-positive monocyte
H3K4me1	ENCSR000ASM	ENCFF673ZGJ	CD14-positive monocyte
H3K4me3	ENCSR000ASN	ENCFF691MBD	CD14-positive monocyte
H3K9ac	ENCSR000ATF	ENCFF994MCP	CD14-positive monocyte
H3K9me3	ENCSR000ASP	ENCFF236ADT	CD14-positive monocyte
H4K20me1	ENCSR000ASQ	ENCFF887JRI	CD14-positive monocyte
H3K27ac	ENCSR000AKC	ENCFF690GQK	GM12878
H3K27me3	ENCSR000DRX	ENCFF103NGB	GM12878
H3K36me3	ENCSR000DRW	ENCFF144MAY	GM12878
H3K4me1	ENCSR000AKF	ENCFF378FBA	GM12878
H3K4me2	ENCSR000AKG	ENCFF514YHH	GM12878
H3K4me3	ENCSR057BWO	ENCFF296PTF	GM12878

H3K9ac	ENCSR000AKH	ENCFF637GBK	GM12878
H4K20me1	ENCSR000AKI	ENCFF154MVT	GM12878
H3K27me3	ENCSR000DUE	ENCFF034QJR	HepG2
H3K36me3	ENCSR000DUD	ENCFF370NTL	HepG2
H3K4me1	ENCSR000APV	ENCFF095ZHO	HepG2
H3K4me2	ENCSR000AMC	ENCFF948LWD	HepG2
H3K4me3	ENCSR575RRX	ENCFF229PGV	HepG2
H3K9ac	ENCSR000AMD	ENCFF129YID	HepG2
H3K9me3	ENCSR000ATD	ENCFF997LPG	HepG2
H4K20me1	ENCSR000AMQ	ENCFF031AYD	HepG2
H3K27me3	ENCSR000EWB	ENCFF233ODK	K562
H3K36me3	ENCSR000DWB	ENCFF514DBT	K562
H3K4me1	ENCSR000EWC	ENCFF359WWB	K562
H3K4me2	ENCSR000AKT	ENCFF168NKC	K562
H3K4me3	ENCSR668LDD	ENCFF465RJJ	K562
H3K9ac	ENCSR000EVZ	ENCFF257END	K562
H3K9me3	ENCSR000APE	ENCFF361WTS	K562
H3K27ac	ENCSR000EWR	ENCFF040ZCD	MCF-7
H3K27me3	ENCSR000EWP	ENCFF825FPO	MCF-7
H3K4me1	ENCSR493NBY	ENCFF158SKW	MCF-7
H3K4me2	ENCSR875KOJ	ENCFF651IUJ	MCF-7
H3K4me3	ENCSR000DWJ	ENCFF530SPD	MCF-7
H3K9ac	ENCSR056UBA	ENCFF636NEF	MCF-7
H3K9me3	ENCSR000EWQ	ENCFF348ISZ	MCF-7
H4K20me1	ENCSR639RHG	ENCFF052ILJ	MCF-7

Supplementary Table 6: ENCODE histone ChIP-seq experiments in seven cell types — See also Supplementary Figure 4. The ENCODE accession numbers allow to uniquely identify the experiment and peak call file (bigBed) on the ENCODE portal (https://www.encodeproject.org/).

Supplementary Table 7: Catalog of 12,248 protein-coding genes analyzed in this study. For each gene we provide the level of expression at 0 hours p.i. (average of the normalized levels from the two biological replicates), the type of expression profile (silent / stably expressed / bending / down-regulated / peaking / up-regulated) and the chromatin marking status (unmarked / stably marked / differentially marked) with respect to each of the nine histone marks. In the case of DE genes, we further specify the type of relationship with gene expression over time (positively correlated / uncorrelated / negatively correlated), as well as the corresponding chromatin cluster (1: stable / uncorrelated marking; 2: positively correlated marking; 3: absence of marking).

CHAPTER 2

The genomic location of regulatory elements plays a role in tissue-specific gene expression

The level and spatio-temporal pattern of expression of a gene are determined by a combination of regulatory elements that dictate its transcriptional activation. Key elements known to regulate gene expression are located within introns of their target genes. Nevertheless, it is unclear whether this is a sporadic feature or a pattern of biological relevance. By leveraging the ENCODE registry of candidate *cis*-regulatory elements (cCREs), we have identified sets of common and tissue-specific distal cCREs characterized by epigenetic signatures of enhancer activity. We uncover that distal regulatory activity shared among tissues is more frequently located in intergenic regions. In contrast, tissue-specific elements accumulate in introns. Up to approximately 20% of these intronic tissue-specific elements carry eQTLs detected in the same tissue, and their target genes perform tissue-specific functions, especially in brain and muscle. Remarkably, the target gene is only in roughly 50% of the cases the same hosting the intronic enhancer, an observation that disentangles the presence of intronic REs from the regulation of the host gene. Sequence elements, such as transcription factor binding sites, do not seem to play a role in the intronic preference of tissue-specific enhancers. Although at a lower rate than brain and muscle adult tissues, enhancers active in differentiated tissues of the embryo are also more frequently located in introns, compared to elements specific to embryonic stem cells or shared among developmental samples.

Borsari B.*, Villegas-Mirón P.*, Laayouni H., Segarra-Casas A., Bertranpetit J., Guigó R. and Acosta S. (2020). Intronic enhancers regulate the expression of genes involved in tissue-specific functions and homeostasis.

Submitted. Available on *bioRxiv*: `https://doi.org/10.1101/2020.08.21.260836`

Intronic enhancers regulate the expression of genes involved in tissue-specific functions and homeostasis

Beatrice Borsari*[1], Pablo Villegas-Mirón*[2], Hafid Laayouni[2], Alba Segarra-Casas[2], Jaume Bertranpetit[2], Roderic Guigó[1,3] and Sandra Acosta†[2]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Catalonia, Spain

[2]Institut de Biologia Evolutiva (UPF-CSIC), Universitat Pompeu Fabra, Dr. Aiguader 88, Barcelona 08003, Catalonia, Spain

[3]Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, Barcelona 08003, Catalonia, Spain

Running title:

Intronic enhancers lead tissue-specific regulation.

Keywords:

enhancers, introns, gene regulation, tissue function, tissue patterning

*Beatrice Borsari and Pablo Villegas-Mirón contributed equally to this work.

†Correspondence should be addressed to E-mail: sandra.acosta@upf.edu (Sandra Acosta)

Abstract

Tissue function and homeostasis reflect the gene expression signature by which the combination of ubiquitous and tissue-specific genes contribute to the tissue maintenance and stimuli-responsive function. Enhancers are central to control this tissue-specific gene expression pattern. Here, we explore the correlation between the genomic location of enhancers and their role in tissue-specific gene expression. We found that enhancers showing tissue-specific activity are highly enriched in intronic regions and regulate the expression of genes involved in tissue-specific functions, while housekeeping genes are more often controlled by intergenic enhancers. Notably, an intergenic-to-intronic active enhancers continuum is observed in the transition from developmental to adult stages: the most differentiated tissues present higher rates of intronic enhancers, while the lowest rates are observed in embryonic stem cells. Altogether, our results suggest that the genomic location of active enhancers is key for the tissue-specific control of gene expression.

Introduction

Multiple layers of molecular and cellular events tightly control the level, time and spatial distribution of expression of a particular gene. This wide range of mechanisms, known as gene regulation, defines tissue-specific gene expression signatures (Melé et al., 2015), which account for all the processes controlling the tissue function and maintenance, namely tissue homeostasis. Both the level and spatio-temporal pattern of expression of a gene are determined by a combination of regulatory elements (REs) controlling its transcriptional activation. Most genes contributing to tissue-specific expression signatures are actively transcribed in more than one tissue, but at different levels and with distinct patterns of expression in time and space, suggesting that the regulation of these genes is different across tissues. Nevertheless, approximately 10-20% of all genes are ubiquitously expressed (*housekeeping genes*), and they are involved in basic cell maintenance functions (Pervouchine et al., 2015; Zabidi et al., 2015; Eisenberg and Levanon, 2013).

cis-REs (CREs) are distributed across the whole genome, and their chromatin status correlates with the transcriptional control they exert over their target genes (Chen et al., 2019; Hawkins et al., 2010; Choukrallah et al., 2015). The activation of CREs depends on several epigenetic features, including combinations of different transcription factors' binding sites, and it is positively correlated with the H3K27ac histone modification signal (Heinz et al., 2015; Heintzman et al., 2007). Epigenetic features in specific tissues may change throughout the life-span of individuals. During development, embryos undergo dramatic morphological and functional changes. These changes shape cell fate and identity as a result of tightly regulated transcriptional programs, which in turn are intimately associated with CREs' activity and chromatin dynamics (Shlyueva et al., 2014; Bonev et al., 2017; Rand and Cedar, 2003; Gilbert et al., 2003).

Notably, key CREs known to regulate gene expression have been reported to locate in introns of their target genes (Ott et al., 2009; Kawase et al., 2011). However, it is unknown whether this is either a sporadic feature associated with certain types of genes – for instance long genes, such as HBB (β-globin) (Gillies et al., 1983) or CFTR (Ott et al., 2009) –, a common regulatory mechanism to most genes (Khandekar et al., 2007; Levine, 2010), or a pattern of biological significance. To delve into this question, we analyzed the genomic location of CREs across a panel of 87 adult and embryonic human cell types available from the Encyclopedia of DNA Elements (ENCODE) Project (Abascal et al., 2020). We found that highly shared CREs are mostly intergenic, while tissue-specific CREs tend to accumulate in introns. The prevalence of intronic CREs correlates with the level of specialization of the tissues, with the more differentiated ones presenting enrichment of intronic CREs. Moreover, intronic CREs target genes involved in tissue-specific functions and homeostasis, suggesting their implication in the functional specificity of tissues.

Results

Enhancer-like regulatory elements define tissue-specific signatures

We leveraged the cell type-agnostic registry of candidate *cis*-Regulatory Elements (cCREs) generated for the human genome (hg19) by the ENCODE Project. We focused on the set of 991,173 cCREs classified

as Enhancer-Like Signatures (ELSs), defined as DNAse hypersensitive sites supported by the H3K27ac epigenetic signal, and assessed their presence-absence patterns across 60 adult cell type-specific catalogues (Table S1; see Methods). We first explored the data with multidimensional scaling (MDS), which uncovered tissue-specific presence-absence patterns (Fig. S1A). Indeed, the separation of samples driven by ELSs' activity was comparable to the one obtained from the analysis of Genotype-Tissue Expression (GTEx) data (Melé et al., 2015), with blood and brain as the most diverging samples. This suggests a correlation between gene regulation mechanisms orchestrated by ELSs and tissue-specific gene expression patterns, which has been previously described (Pennacchio et al., 2007; Ernst et al., 2011).

Interestingly, we observed that the proportion of active ELSs located in intergenic regions was positively correlated with the number of samples in which ELSs were active (Spearman's ρ = 0.55; *p* value = 6.2e-06; Fig. 1A), suggesting a functional role for the genomic location of ELSs. Thus, to untangle the relationship between genomic location and cell-type specificity of ELSs, we selected a subset of 25 samples that clustered into 5 main groups – iPSCs, fibro/myoblasts, muscle, blood and brain samples (Fig. 1B-C; Table S1, *Samples' Cluster*) – according to their MDS proximity and consistently with their tissue of origin and function. This curated subset of samples allowed us to study enhancer activity in a tissue-specific manner, and compare it with regulatory mechanisms shared among tissues. Tissues represented by only one sample were not included in the subsequent analysis. Indeed, the fact that the *ad hoc* tissues' functional clustering is supported by tissue-specific enhancer signatures suggests a direct link between ELSs' activity and the regulation of tissue-specific functions. We defined *tissue-active* ELSs as those active in $\geq$ 80% of the samples within a given cluster (Table S2, **Tissue-active ELSs**; see Methods). As expected, in some cases we observed shared regulatory activity between tissues, in other words a fraction of ELSs active in a given cluster were also active in samples belonging to other clusters. For instance, approximately 1,700 blood-active ELSs were also active in all the seven brain samples (Fig. S1B). Because of this overlap, we defined sets of *tissue-specific* ELSs (Table S2, see Methods) as those active in $\geq$ 80% of the samples within the tissue cluster and in at most one sample outside the cluster. Due to their small size, for iPSCs and fibro/myoblasts clusters we considered as tissue-specific those ELSs active exclusively within their clusters (see Methods). The overlap of tissue-specific ELSs with samples from other clusters is depicted in Fig. 1D. The majority of brain- and blood-specific ELSs were active only within their tissue cluster (71.9% and 62.3%, respectively), while a considerable fraction (52.0%) of muscle-specific ELSs was shared with one sample from other clusters, mostly with fibro/myoblast samples (33.1%). This is consistent with the samples' MDS proximity observed in Fig. 1B, suggesting a functional relevance of the genes regulated by shared ELSs. In addition, we identified a set of 208 ELSs active in all the 25 samples (Table S2, **Common ELSs**).

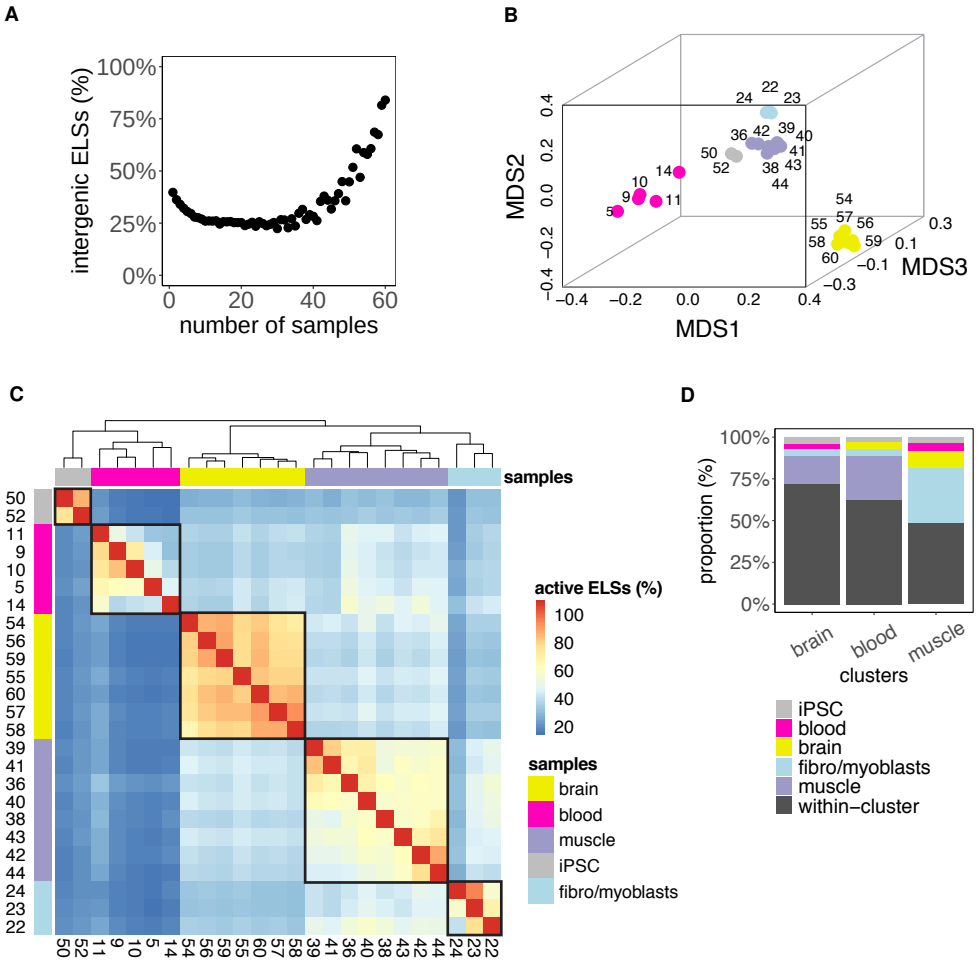


Figure 1. **A**: Highly-shared ELSs are more frequently located in intergenic regions. The scatter plot represents the proportion of intergenic ELSs active in increasing numbers of human adult samples (Spearman's ρ = 0.55; p value = 6.2e-06). **B**: MDS distribution of human adult samples defined by ELSs' activity. Analogous representation to Figure S1A for the subset of 25 selected adult human samples. **C**: Samples' clustering defined by ELSs' presence-absence patterns (clustering method: *complete*; clustering distance: *euclidean*). The heatmap represents the percentage of ELSs active in row *i* that are also active in column *j*. For this analysis we considered 268,214 of the 991,173 ELSs that were active in at least 2 of the 25 selected human adult samples. The correspondence between samples and numbers is reported in Table S1. **D**: Tissue-specific ELSs. The barplot represents the type of samples found within sets of brain-, blood- and muscle-specific ELSs. As described in Methods (section *Tissue-active, tissue-specific and common ELSs*), most of tissue-specific ELSs are only active in the samples of the corresponding cluster ("within-cluster", *black*), but a few of them may be active in at most one outer sample (i.e. a sample that does not belong to the tissue cluster, *coloured*). IPSCs- and fibro/myoblasts-specific ELSs are not represented, since we did not allow outer samples given their small cluster sizes (2 and 3, respectively; see Methods).

The genomic location of regulatory elements correlates with their tissue-homeostatic functions

We next explored the genomic location of the sets of common and tissue-specific ELSs. While common ELSs were preferentially located in intergenic regions (63.4%, Fig. 2A), the majority of muscle- and brain-specific ELSs fell inside introns (71.6% and 74.0%, respectively; Fig. 2A). These significant differences in genomic distribution between tissue-specific and common regulatory elements (Table S3) are consistent with our initial observation of a high sharing rate of intergenic ELSs across samples (Fig. 1A). In contrast, the iPSCs, fibro/myoblasts and blood clusters – which comprise undifferentiated, non-specialized or more heterogeneous cell types, respectively – showed a more even distribution of tissue-specific ELSs between intergenic and intronic regions (Fig. 2A). Overall, we observed a scarcity of exonic ELSs (Fig. 2A, Table S4).

Genes harboring tissue-specific ELSs may present distinctive features, including differences in intron length and density. To rule out any bias in our analyses, we compared these features between genes hosting common and tissue-specific ELSs. While the number of introns per hosting gene was comparable across groups (Kruskal-Wallis p value test = 0.98; Fig. S2A), we reported significant differences in the median intron length per gene (Kruskal-Wallis p value test $<$ 2.2e-16; Fig. S2A). Moreover, we observed significant differences in the intronic ELSs' density (Kruskal-Wallis p value test $<$ 2.2e-16), with higher values for brain and muscle, suggesting that the enrichment of tissue-specific ELSs in intronic regions is not biased by the intron length (Fig. S2A).

We subsequently explored whether the genes harboring tissue-specific intronic ELSs perform functions associated with tissue homeostasis maintenance and response to stimuli. We performed a Gene Ontology (GO) enrichment analysis on the genes containing tissue-specific intronic ELSs. Indeed, the enrichment of terms associated with tissue-specific cellular components is consistent with the ELSs' identity (Table S5). For instance, genes hosting brain-specific ELSs perform functions associated with synapses and axons, while in the case of muscle and blood we found significant terms related to sarcolemma, Z-disc and contractile fibers, and immunological synapses and cell membranes, respectively. Conversely, genes harboring common ELSs reported terms related to non-specific cell membrane composition (Table S5). Although this suggests an implication of intronic ELSs in tissue-specific functions, likely through tissue-specific gene regulation mechanisms, there is no proven association of intronic ELSs being direct regulators of their host genes.

To address this issue, we integrated our ELS analysis with the catalogue of expression Quantitative Trait Loci (eQTLs) provided by the Genotype-Tissue Expression (GTEx) Project (Aguet et al., 2017). Among the 35,275 common and tissue-specific ELSs, 5,941 overlap with a significantly associated eQTL-eGene pair, hereafter referred to as eQTL-ELSs. The proportion of eQTL-ELSs was similar among groups, with the exception of iPSCs, which are not represented in the GTEx sampling collection (Fig. S2B). This allowed us to leverage the eQTL-ELSs pairs to explore the biological function of the genomic distribution of ELSs, focusing on eQTLs regulating gene expression in the four GTEx categories matching our samples' clusters (fibroblasts, blood, muscle and brain subregions; see Methods). In line with the above-mentioned results, highly specialized tissues such as brain and muscle showed the highest proportion of intronic *vs* intergenic

ELSs hosting eQTLs detected in the corresponding tissue: brain (2,105 (78%) *vs* 593 (22%)), muscle (349 (76%) *vs* 108 (24%)), fibro/myoblasts (289 (63%) *vs* 163 (37%)), blood (138 (59%) *vs* 94 (41%)) (Figure 2B). Conversely, common eQTL-ELSs were more frequently located in intergenic elements (5 (25%) *vs* 15 (75%)) (data not shown). Overall, these results indicate a potential functional role of the genomic distribution of ELSs in the regulation of tissue-specific gene expression. Still, although there is a clear trend of eQTL-ELSs' specificity per tissue, many of these eQTLs are not exclusive to a single tissue. For this reason, we validated our observations with a GO enrichment analysis on the sets of genes associated with intronic and intergenic eQTL-ELSs. GO analysis on muscle- and brain-specific eQTL-ELSs showed a clear prevalence of tissue-specific homeostatic functions for those genes targeted by intronic eQTL-ELSs (for instance, muscle: carbohydrate and amino acid metabolism; brain: cell projection and organization) (Table S6). On the contrary, in the case of blood we found significantly enriched GO terms only for genes targeted by intergenic eQTL-ELSs (Table S6). This might be due to the fact that blood comprises different cell types and can be considered a more heterogeneous tissue. Overall, these results suggest that intronic eQTL-ELSs are involved in the regulation of genes controlling tissue-specific functions and tissue homeostasis.

Next, we wanted to understand the relationship between the intronic ELSs and their harboring genes. Of note, the proportion of intronic eQTL-ELSs targeting their host genes was comparable among groups of samples, but always below 54.3% (Fig. 2C). Most interestingly, eQTL-ELSs regulating the expression of the host gene are associated with tissue-specific functions, with genes involved in axonal components for the brain (e.g. NRCAM), actin cytoskeleton for fibroblasts (e.g. FMN1) or contractility-related terms for muscle (e.g. SYNM). However, those targeting the expression of non-hosting genes are involved in homeostatic functions not directly associated with the tissue function. For instance, the brain presents significant terms related to the splicing proteins (e.g. SF3A1, SF3B1), a widely extended process in the brain and responsible of the fine tuning of several brain functions (Vuong et al., 2016) (Fig. 2D). Overall, this suggests that other mechanistic strategies may account for the intronic preference of regulatory elements in highly specialized tissues.

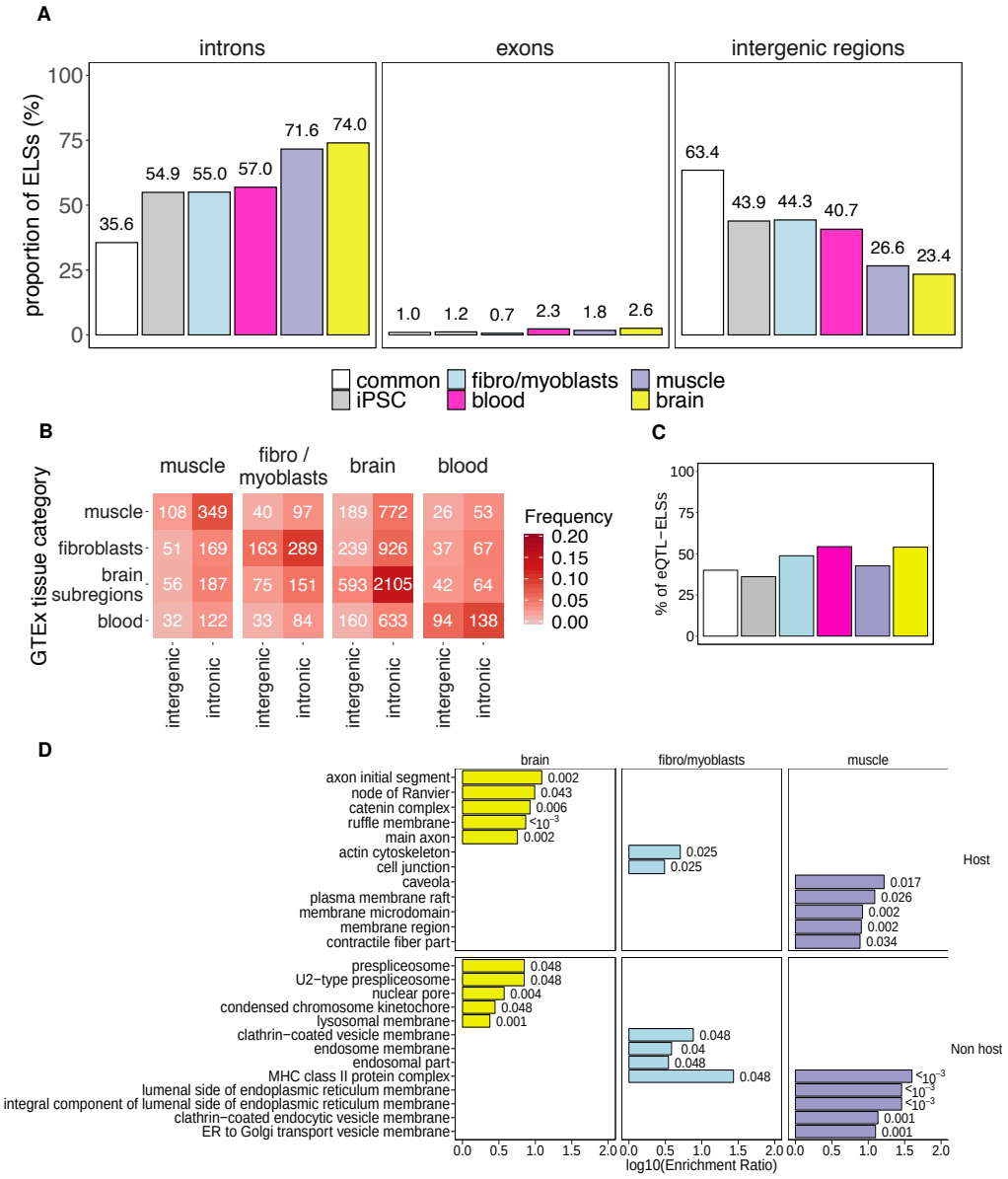


Figure 2. **A**: Proportions of common and tissue-specific ELSs identified in the 25 selected human adult samples that overlap intronic, exonic and intergenic regions. **B**: Number of intergenic and intronic muscle-, fibro/myobblasts-, brain- and blood-specific ELSs harboring eQTLs detected in Muscle, Fibroblasts, Brain subregions and Blood GTEx samples. Coloured cells represent the proportion of eQTL-ELSs over the total amount of tissue-specific ELSs within each group. **C**: Proportions of common and tissue-specific eQTL-ELSs targeting their host genes. These proportions were computed over the total amount of intronic eQTL-ELSs within each group. **D**: Top five enriched GO terms associated with the hosting and non-hosting eQTL-ELSs regulated genes. *P* value (FDR corrected) is reported for each enriched term.

The enrichment of transcription factor binding sites in tissue-specific ELSs is independent of their genomic location

The activation of ELSs is a dynamic process depending, mainly, on its accessible chromatin to be bound by transcription factors (TFs). Thus, tissue-specific gene expression programs may be controlled by the underlying signature of TFs-ELSs pairing (Schmitt et al., 2016). We next wondered whether the specific distribution of ELSs, i.e. intronic *vs* intergenic, was associated with a different transcription factor binding site (TFBS) signature that could account for their tissue specificity. For this purpose we explored, using the software HOMER (Heinz et al., 2010), TFBSs differences between intronic and intergenic ELSs that were either common or specific to a given tissue. We observed a high sharing rate of TFBSs between intronic and intergenic ELSs, suggesting that there is a strong prevalence of certain transcriptional programs in each tissue independently of the genomic location of ELSs. Notably, there are no enriched TFBSs in common ELSs, either intronic or intergenic (Fig. 3A and Table S7). Amongst the TFBSs enriched in the tissue-specific intronic and intergenic ELSs, there are some that are well known to control tissue-specific homeostatic events, such as FLI1 and RUNX in blood controlling adult endothelial hemogenesis (Lis et al., 2017), and POU6F1 (Brn5), SOX4 and SOX8 in brain controlling the adult neural plasticity (McClard et al., 2018). POU5F1 (Oct4) is required for iPSCs reprogramming, and MEIS1 in muscle is key for cardiomyogenesis (Dupays et al., 2015). Although a great number of the TFs identified in our analysis are known for shaping the functions of certain tissues, the vast majority of these TFs are ubiquitously or widely expressed in several tissues (Fig. 3B), suggesting that the tissue-specificity of gene regulation does not arise from the transcription factor’s potential to bind an ELS, but most likely from the genomic localization of the ELSs.

The genomic location of developmental ELSs is not associated with tissue specificity

Tissue-specific homeostatic features vary dramatically among different adult tissues. For instance, blood comprises a number of cell types characterized by heterogeneous functions and high turnover. On the other hand, muscles are formed by fewer cell types, mainly dedicated to the same function and with limited cell division capacity. The maintenance of tissue homeostasis is ensured by quiescent adult stem cells with features similar to their developmental native lineage (Rué and Martinez Arias, 2015; Biteau et al., 2011). During development, tissues mature to fully reach their functional capacity in adulthood. Still, whether the regulatory features of a given tissue are reminiscent of their developmental lineage remains largely unknown. For this reason, we assessed the activity of the 991,173 cell type-agnostic ELSs across 27 embryonic samples (Table S8). The correlation between the percentage of intergenic ELSs and the number of samples in which ELSs are active was lower compared to adult samples (Spearman’s ρ = 0.38; p value = 0.054; Fig. 4A). MDS analysis highlighted three main groups of embryonic samples: stem cells (ESC), neural progenitors, and a heterogeneous group of more differentiated cell types (Fig. 4B; Table S8, Samples’ Group). The three groups of samples were associated with 3,112, 784 and 1,166 specific ELSs, respectively (Table S9). Although the majority of these ELSs were active only within the corresponding cluster, we reported that 26.2% of the

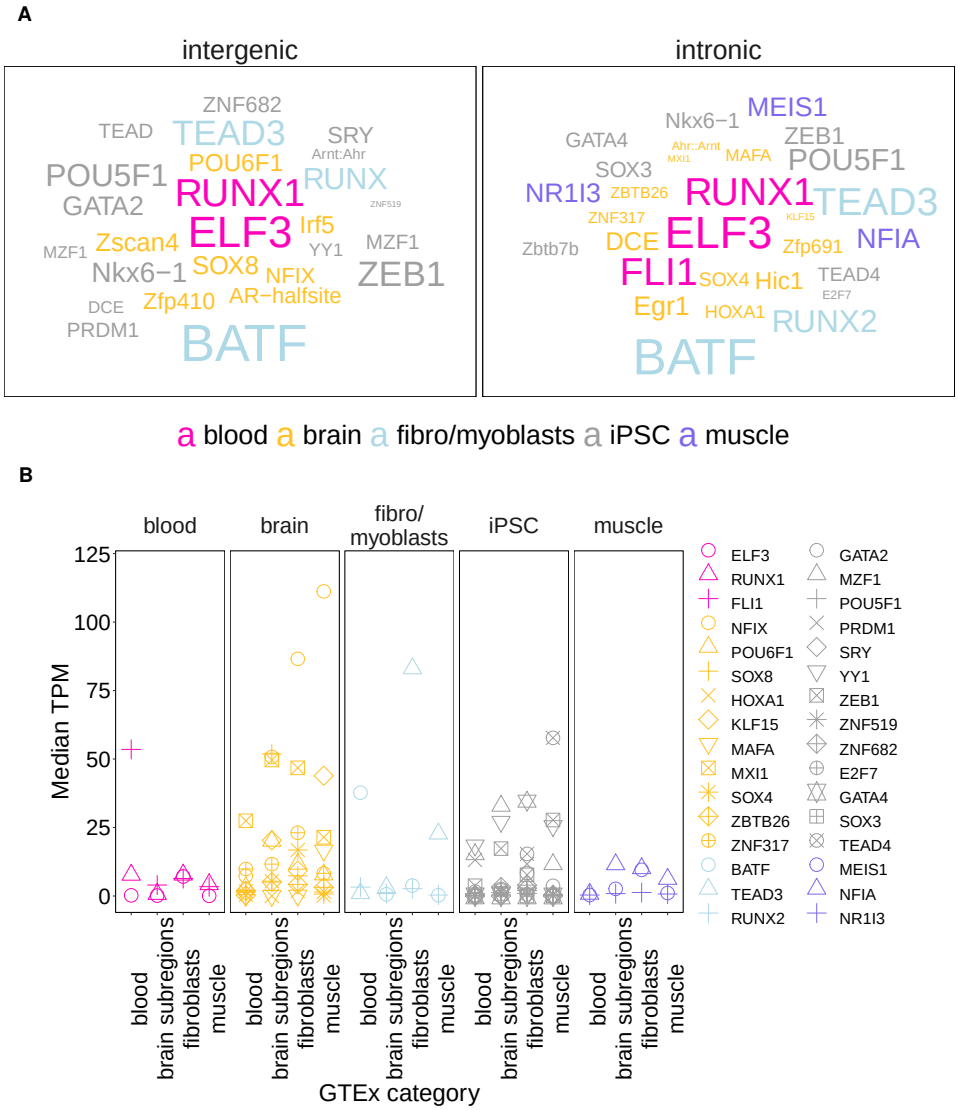


Figure 3. **A**: Word cloud reporting the TFBSs significantly enriched in intronic and intergenic tissue-specific ELSs. No significant TFBSs were found in common ELSs. The size of the word represents the significance of TFBSs enrichment. **B**: Median expression, in the four matching GTEx tissues categories, of the TFs associated with significantly enriched TFBSs in each cluster.

neural progenitors-specific ELSs were also active in one ESC sample (Fig. S3A). On the contrary, we identified only 94 ELSs common to all embryonic samples (Table S9). The proportion of specific intronic ELSs was higher for neural progenitors and differentiated tissues (60.3% and 60.6%, respectively; Fig. 4C) compared to ESC-specific (50.9%) and common (38.3%) ELSs, but lower with respect to clusters of adult muscle and brain samples (71.6% and 74.0%, respectively, Fig. 2A). As in the case of adult samples, we observed a scarcity of exonic ELSs (Fig. 4C, Table S11), while we could not find significant associations between the frequency of group-specific intronic ELSs and features of intron length and density (Figs. S3B). On the other hand, the density of ELSs per introns (Fig. S3B) was similar to the one observed in adult samples (Fig. S2A).

When studying the genes harboring developmental group-specific intronic ELSs, we observed that they are enriched in functions consistent with the corresponding adult tissue (Table S12). For instance, the ones hosting neural progenitors-specific ELSs are enriched in neural development-related terms, such as axonogenesis and dendritic spine organization. Notably, genes harboring developmental common ELSs are enriched in protein complexes like nBAF and SWI/SNF, known developmental chromatin remodelers (Alver et al., 2017).

Lastly, in an attempt to define the amount of regulatory activity shared by embryonic and adult samples as an indicator of the reminiscent embryonic function in adult tissue homeostasis, we computed, for specific and common embryonic ELSs, the number of adult tissues in which they were found active. As expected, whereas ELSs specific to stem cells and neural progenitors were active in a limited set of adult samples, embryonic differentiated tissues reported a higher degree of shared regulatory activity with adult cell types. Moreover, ELSs active in all embryonic samples (common) were also active in the majority of adult samples (Fig. 4D). Overall, these results show that the genomic location of ELSs is dynamic throughout development, and shifts towards intronic localization during tissue maturation.

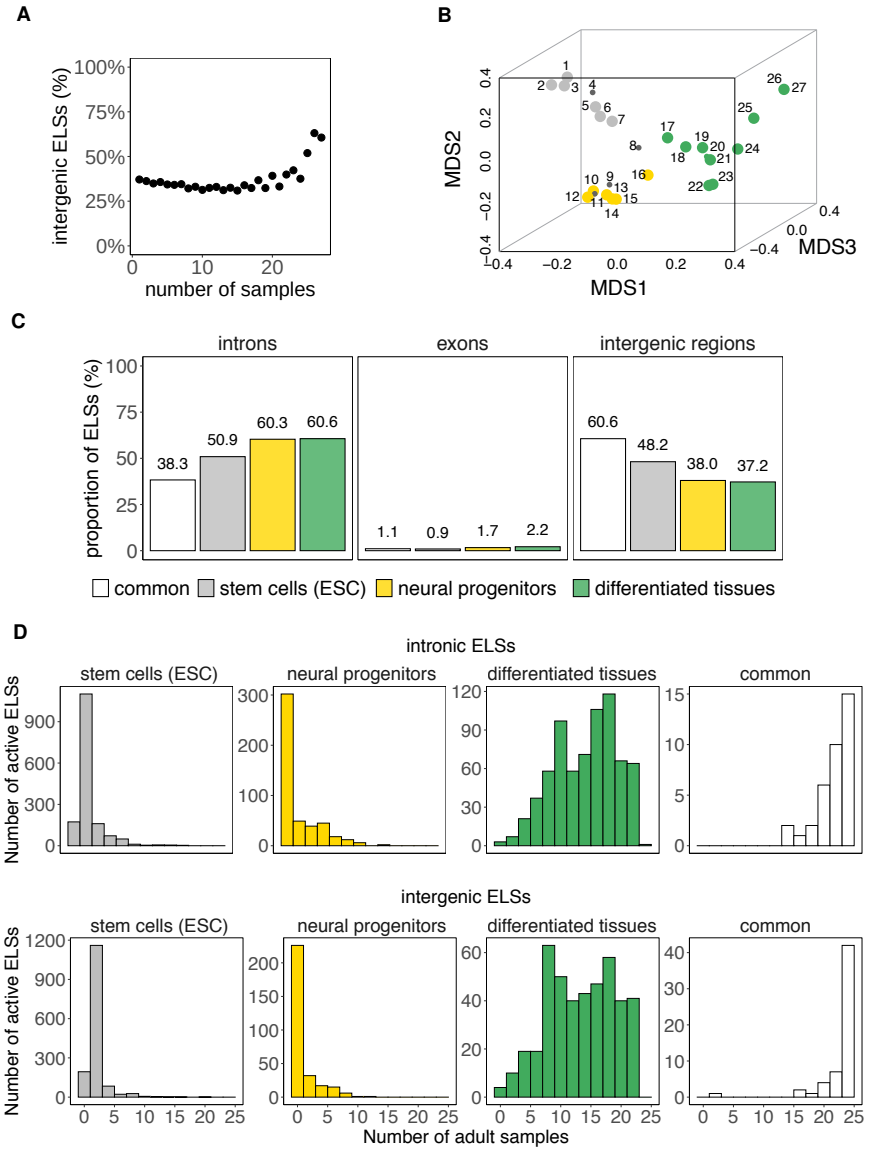


Figure 4. **A**: Scatter plot representing the percentage of intergenic ELSs active in increasing numbers of human embryonic samples (Spearman's ρ = 0.38; p value = 0.054). The degree of correlation between ELSs' sharing and the percentage of intergenic ELSs is lower compared to the one observed for adult samples. **B**: MDS representation of the dissimilarities between the 27 human embryonic samples according to the pattern of activity of ELS-cREs (analogous to Fig. 1B). The correspondence between samples and numbers is reported in Table S8. The MDS highlights 3 main groups of embryonic samples. **C**: Proportions of common and group-specific ELSs identified in embryonic samples that overlap intronic, exonic and intergenic regions. **D**: Rate of sharing of intronic (upper panel) and intergenic (lower panel) ELSs between embryonic and adult samples. The histogram represents the number of selected adult samples (n = 25) in which embryonic ELSs are active.

Discussion

In this study, we show the central role of intronic Enhancer-Like Signatures (ELSs) in the control of tissue-specific expression signatures. Tissue-specific homeostasis is a dynamic process encompassing the coordinated expression in time and space of a wealth of genes, mainly controlled by active ELSs. The ENCODE project reports that about half of these ELSs are intergenic, and 38% are intronic (ENCODE SCREEN Portal: `https://screen-v10.wenglab.org/`, section “About”). The enrichment in intronic ELSs in the most specialized tissues observed in our study, independently of the sequence – in terms of transcription factor binding sites – suggests an important role of the genomic location of ELSs. Since Heitz described in 1928 (Heitz, 1928) euchromatin as chromosomal regions enriched in genes, and heterochromatin as inactive or passive chromatin regions, this dual definition has been shaped throughout the years but it still remains vastly correct (De Laat and Duboule, 2013; DeMare et al., 2013; Ernst and Kellis, 2010). Intergenic regions are often regulatorily silenced, and this happens more frequently in adult than embryonic tissues (Heinz et al., 2015). A similar correlation is observed in our data, since embryonic ELSs are not as frequently found in intronic elements as in adults, suggesting that the maturation and tissue commitment correlates with the ELS distribution across the whole genome. One could hypothesize that the enriched presence of intronic ELSs in specialized tissues is advantageous for the control of the gene expression signature of a particular tissue, for instance granting ELSs accessibility in open DNA regions (genes) and avoiding leaky activity of ELSs. Introns have been long observed as gene expression regulators throughout different mechanisms (Rose, 2019; Chorev and Carmel, 2012; Shaul, 2017). Introns regulatory potential has been longly associated with the regulation of the host gene’s expression in several different ways, often related to alternative splicing, intron retention (Jacob and Smith, 2017), non-sense mediated decay (Lewis et al., 2003), and even with the control of transcription initiation via recruitment of RNA Polymerase II (Bieberstein et al., 2012). However, here we found that about half of the eQTL-ELSs located in introns do not regulate the expression of the host gene. This is important regulatory information since it disentangles the presence of intronic ELSs from the regulation of the host gene, opening new opportunities to identify the regulatory mechanisms controlling tissue-specific gene expression. Overall, our results suggest that the genomic distribution of tissue-specific active ELSs is not stochastic and mainly overlaps with intronic elements. The opposite happens to active ELSs common to all tissues. These results suggest that intronic enhancers play a role in the regulation of gene expression in a tissue-specific manner.

Methods

The ENCODE registry of candidate *cis*-Regulatory Elements

The cell type-agnostic registry of human candidate *cis*-Regulatory Elements (cCREs) available from the ENCODE portal corresponds to a subset of 1,310,152 representative DNase hypersensitivity sites (rDHSs) in the human genome with epigenetic activity further supported by histone modification (H3K4me3 and H3K27ac) or CTCF-binding data (`https://screen-v10.wenglab.org/`; section "About"). It comprises 991,173 Enhancer-Like Signatures (ELS), 254,880 Promoter-Like Signatures (PLS), and 64,099 CTCF-only Signatures. In addition, cell type-specific catalogues are provided for those cell types with available DNase and ChIP-seq ENCODE data.

Selection of cCREs with enhancer-like signature (ELS) across human samples

We downloaded the set of 1,310,152 cell type-agnostic cCREs for human assembly 19 (hg19) from the ENCODE SCREEN webpage (`https://screen-v10.wenglab.org/`; file ID: ENCFF788SJC). From the ENCODE portal (`www.encodeproject.org/matrix/?type=Annotation&encyclopedia_version=ENCODE+v4&annotation_type=candidate+Cis-Regulatory+Elements&assembly=hg19`), we retrieved cell type-specific registries of cCREs for 60 adult and 27 embryonic human samples with available DNase data and ChIP-seq H3K4me3 and H3K27ac data. The ENCODE File Identifiers for the adult and embryonic datasets are reported in Table S1 and S8, respectively. We focused on the 991,173 cell type-agnostic cCREs with ELS activity, and generated a binary table in which we assessed, for a given cCRE, the presence/absence of ELS activity annotation (column 9 = "255, 205, 0") in each of the 60 adult and 27 embryonic samples. A binary distance matrix between all pairs of adult samples was used to perform multidimensional scaling (MDS) in three dimensions. This resulted in the selection of 25 adult samples. The same procedure was applied, independently, to the embryonic samples. In this case, IMR-90, mesendoderm, mesodermal cell, endodermal cell and ectodermal cell samples were not included in subsequent analyses.

Intersection of ELSs with genes, introns, exons and intergenic regions

Genes, exons and introns' coordinates were obtained from GENCODE v19 annotation (`https://www.gencodegenes.org/human/release_19.html`). The overlap between ELSs and genes, exons and introns was computed using BEDTools intersectBed v2.27.1 (Quinlan and Hall, 2010). The proportions of ELSs overlapping intronic segments (Figs. 2A, 4C) also include a limited set of ELSs overlapping both intronic and exonic regions (common adult ELSs: 2.4%; iPSCs-specific ELSs: 3.1%; fibro/myoblasts-specific ELSs: 4.5%; blood-specific ELSs: 5.6%; muscle-specific ELSs: 4.4%; brain-specific ELSs: 7.4%; common embryonic ELSs: 7.4%; differentiated tissues-specific ELSs: 5.1%; neural progenitors-specific ELSs: 5.0%; ESC-specific ELSs: 3.2%). On the other hand, we defined as exonic ELSs those intersecting exclusively exonic regions (Figs. 2A, 4C). The overlap of ELSs with intergenic regions was obtained by intersecting the former with the genes' coordinates using the BEDTools intersectBed option -v.

Tissue-active, tissue-specific and common ELSs

Tissue-active ELSs are ELSs active (see Methods section *Selection of cCREs with enhancer-like signature (ELS) across human samples*) in $\geq$ 80% of the samples within a given group of samples (blood = 4/5; muscle = 6/8; brain = 6/7; stem cells = 5/6; neural progenitors = 5/6; differentiated tissues = 8/10). Because of the small sample size, we required iPSCs- and fibro/myoblasts-ELSs to be active in 100% of the samples (2/2; 3/3). Tissue-specific ELSs are tissue-active ELSs that are active in 0 (iPSCs, fibro/myoblasts) or at most 1 (all other groups) outer samples (i.e. samples outside the considered group). Common adult and embryonic ELSs are ELSs active in 100% of the samples (25/25 and 22/22, respectively). To rule out indirect effects of ELS activity related to promoter regions, we discarded common and tissue-specific ELSs overlapping any annotated Transcription Start Site (TSS, ± 2Kb) in GENCODE v19.

Assessing enhancer regulatory activity

ELSs were annotated by using the GTEx v7 (Aguet et al., 2017)) significant variant-gene pairs from 46 different tissues (number of samples with genotype $\geq$ 70). Only single-tissue eQTL-eGene associations with a qval $\leq$ 0.05 were used. Similar GTEx tissues were grouped in unique categories in order to consider the most complete catalogue of eQTL-eGene pairs per group of samples. These categories were named as follows: fibroblasts (Skin Not Sun Exposed Suprapubic, Cells Transformed Fibroblasts), blood (Whole Blood, Spleen), muscle (Skeletal Muscle), brain subregions (all brain subregions, Pituitary Gland, Nerve Tibial), cardiovascular (Heart Atrial Appendage, Heart Left Ventricle, Artery Aorta, Artery Coronary, Artery Tibial), digestive (Liver, Pancreas, Small Intestine Terminal Ileum, Stomach, Colon Sigmoid, Colon Transverse, Esophagus Gastroesophageal Junction, Esophagus Mucosa, Esophagus Muscularis, Adipose Subcutaneous, Adipose Visceral Omentum), gland (Adrenal Gland, Thyroid, Minor Salivary Gland), breast (Breast Mammary Tissue), lung (Lung), sexual (Ovary, Prostate, Testis, Uterus, Vagina). Bedtools (Quinlan and Hall, 2010) was used to intersect the tissue-specific ELSs' coordinates with the *cis*-eQTLs' positions in the considered genomic locations (intronic and intergenic). We kept all eQTL-eGene pairs that were found significantly associated with the matching eQTL-ELS's tissue category (brain, blood, muscle and fibro/myoblasts). In the case of iPSCs-specific and common ELSs, we considered those eQTL-eGene pairs that were significantly reported in all the tissues. The resulting intersected ELSs were considered as being responsible for the regulation of the associated eGene. The functional enrichment of the ELSs' target genes was performed by the online utility WebGestalt (Liao et al., 2019).

cis-Regulatory Elements and Transcription Factor Binding Sites

Transcription factor binding sites (TFBSs) were predicted by using the motif discovery software HOMER (Heinz et al., 2010) This program performs a differential motif discovery by taking two sets of genomic regions (findMotifGenome.pl script) and identifying the motifs that are enriched in one set of sequences relative to a background list of regions. We analysed the tissue-specific ELSs' binding motifs by considering the ELS regions from all the other tissues as background. We searched for 6-mer and 7-mer length motifs as a way

to focus on enriched core motif sequences and avoid redundancy from longer motifs with similar functions. A hypergeometric test and FDR correction were applied for the motif enrichment. Only significantly enriched motifs were considered in the subsequent analysis. The word size in Figure 3A is proportional to the significance of the enrichment, it is calculated as the difference of sequence frequencies where the TFBS is found in the target and background lists of regions. The functionality of the predicted TFBSs was assessed by analysing the tissue-specific expression of the transcription factors that bind to them. GTEx expression data (v7) was analysed for those transcription factors whose TFBSs were reported as significant by HOMER in all tissues and genomic locations.

Data access

All ENCODE data used in this study is publicly available on the ENCODE portal (`www.encodeproject.org/`). GTEx gene expression and eQTL data is available on the GTEx portal (`www.gtexportal.org`).

Acknowledgments

B.B. is supported by the fellowship 2017FI_B00722 from the Secretaría d’Universitats i Recerca del Departament d’Empresa i Coneixement (Generalitat de Catalunya) and the European Social Fund (ESF). P.V-M. is supported by an FPI PhD fellowship (FPI-BES-2016-077706) part of the ”Unidad de Excelencia María de Maeztu” funded by the MINECO (ref: MDM-2014-0370). S.A. is supported by a fellowship from the Secretaría d’Universitats i Recerca del Departament d’Empresa i Coneixement (Generalitat de Catalunya) (BP-2017-00176). We thank the ENCODE and GTEx Consortia for data production. We thank Diego Garrido-Martín (R. Guigó Lab) for valuable statistical advice.

Author Contributions: S.A., B.B, and P.V-M. designed the study, analyzed the data and wrote the manuscript with feedback from all the authors. H.L. and A.S-C. analyzed the data.

Competing interest statement

The authors declare no competing interests.

References

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., Akiyama, J. A., Jammal, O. A., Amrhein, H., Anderson, S. M., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710.

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y. S., Parsana, P., Segrè, A. V., et al. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

Alver, B. H., Kim, K. H., Lu, P., Wang, X., Manchester, H. E., Wang, W., Haswell, J. R., Park, P. J., and Roberts, C. W. (2017). The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers. *Nature Communications*, 8.

Bieberstein, N. I., Oesterreich, F. C., Straube, K., and Neugebauer, K. M. (2012). First exon length controls active chromatin signatures and transcription. *Cell Reports*, 2(1):62–68.

Biteau, B., Hochmuth, C. E., and Jasper, H. (2011). Maintaining tissue homeostasis: Dynamic control of somatic stem cell activity. *Cell Stem Cell*, 9(5):402–411.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J. P., Tanay, A., et al. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, 171(3):557–572.

Chen, C., Yu, W., Tober, J., Blobel, G. A., Speck, N. A., and Correspondence, K. T. (2019). Spatial Genome Re-organization between Fetal and Adult Hematopoietic Stem Cells. *Cell Reports*, 29(12):4200–4211.

Chorev, M. and Carmel, L. (2012). The function of introns. *Frontiers in Genetics*, 3.

Choukrallah, M. A., Song, S., Rolink, A. G., Burger, L., and Matthias, P. (2015). Enhancer repertoires are reshaped independently of early priming and heterochromatin dynamics during B cell differentiation. *Nature Communications*, 6.

De Laat, W. and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, 502(7472):499–506.

DeMare, L. E., Leng, J., Cotney, J., Reilly, S. K., Yin, J., Sarro, R., and Noonan, J. P. (2013). The genomic landscape of cohesin-Associated chromatin interactions. *Genome Research*, 23(8):1224–1234.

Dupays, L., Shang, C., Wilson, R., Kotecha, S., Wood, S., Towers, N., and Mohun, T. (2015). Sequential Binding of MEIS1 and NKX2-5 on the Popdc2 Gene: A Mechanism for Spatiotemporal Regulation of Enhancers during Cardiogenesis. *Cell Reports*, 13(1):183–195.

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49.

Gilbert, N., Boyle, S., Sutherland, H., Heras, J. d. L., Allan, J., Jenuwein, T., and Bickmore, W. A. (2003). Formation of facultative heterochromatin in the absence of HP1. *The EMBO Journal*, 22(20):5540–5550.

Gillies, S. D., Morrison, S. L., Oi, V. T., and Tonegawa, S. (1983). A Tissue-specific Transcription Enhancer Element Is Located in the Major Intron of a Rearranged Immunoglobulin Heavy Chain Gene. *Cell*, 33(3):717–728.

Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E., Kuan, S., Luu, Y., Klugman, S., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell*, 6(5):479–491.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–89.

Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, 16(3):144–154.

Heitz, E. (1928). Das Heterochromatin der Moose. *Jahrbücher für wissenschaftliche Botanik*, 69.

Jacob, A. G. and Smith, C. W. (2017). Intron retention as a component of regulated gene expression programs. *Human Genetics*, 136(9):1043–1057.

Kawase, S., Imai, T., Miyauchi-Hara, C., Yaguchi, K., Nishimoto, Y., Fukami, S. I., Matsuzaki, Y., Miyawaki, A., Itohara, S., and Okano, H. (2011). Identification of a novel intronic enhancer responsible for the transcriptional regulation of musashi1 in neural stem/progenitor cells. *Molecular Brain*, 4(1).

Khandekar, M., Brandt, W., Zhou, Y., Dagenais, S., Glover, T. W., Suzuki, N., Shimizu, R., Yamamoto, M., Lim, K. C., and Engel, J. D. (2007). A Gata2 intronic enhancer confers its pan-endothelia-specific regulation. *Development*, 134(9):1703–1712.

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Current Biology*, 20(17):754–763.

Lewis, B. P., Green, R. E., and Brenner, S. E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 100(1):189–192.

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, 47(W1):W199–W205.

Lis, R., Karrasch, C. C., Poulos, M. G., Kunar, B., Redmond, D., Duran, J. G., Badwe, C. R., Schachterle, W., Ginsberg, M., Xiang, J., et al. (2017). Conversion of adult endothelium to immunocompetent haematopoietic stem cells. *Nature*, 545(7655):439–445.

McClard, C. K., Kochukov, M. Y., Herman, I., Liu, Z., Eblimit, A., Moayedi, Y., Ortiz-Guzman, J., Colchado, D., Pekarek, B., Panneerselvam, S., et al. (2018). POU6f1 mediates neuropeptide-dependent plasticity in the adult brain. *Journal of Neuroscience*, 38(6):1443–1461.

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., et al. (2015). The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665.

Ott, C. J., Blackledge, N. P., Kerschner, J. L., Leir, S. H., Crawford, G. E., Cotton, C. U., and Harris, A. (2009). Intronic enhancers coordinate epithelial-specific looping of the active CFTR locus. *Proceedings of the National Academy of Sciences of the United States of America*, 106(47):19934–19939.

Pennacchio, L. A., Loots, G. G., Nobrega, M. A., and Ovcharenko, I. (2007). Predicting tissue-specific enhancers in the human genome. *Genome Research*, 17(2):201–211.

Pervouchine, D. D., Djebali, S., Breschi, A., Davis, C. A., Barja, P. P., Dobin, A., Tanzer, A., Lagarde, J., Zaleski, C., See, L. H., et al. (2015). Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nature Communications*, 6.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rand, E. and Cedar, H. (2003). Regulation of imprinting: A multi-tiered process. *Journal of Cellular Biochemistry*, 88(2):400–407.

Rose, A. B. (2019). Introns as gene regulators: A brick on the accelerator. *Frontiers in Genetics*, 9.

Rué, P. and Martinez Arias, A. (2015). Cell dynamics and gene expression control in tissue homeostasis and development. *Molecular Systems Biology*, 11(2):792.

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., et al. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports*, 17(8):2042–2059.

Shaul, O. (2017). How introns enhance gene expression. *International Journal of Biochemistry and Cell Biology*, 91:145–155.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286.

Vuong, C. K., Black, D. L., and Zheng, S. (2016). The neurogenetics of alternative splicing. *Nature Reviews Neuroscience*, 17(5):265–281.

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518(7540):556–559.

Supplementary Figures and Tables

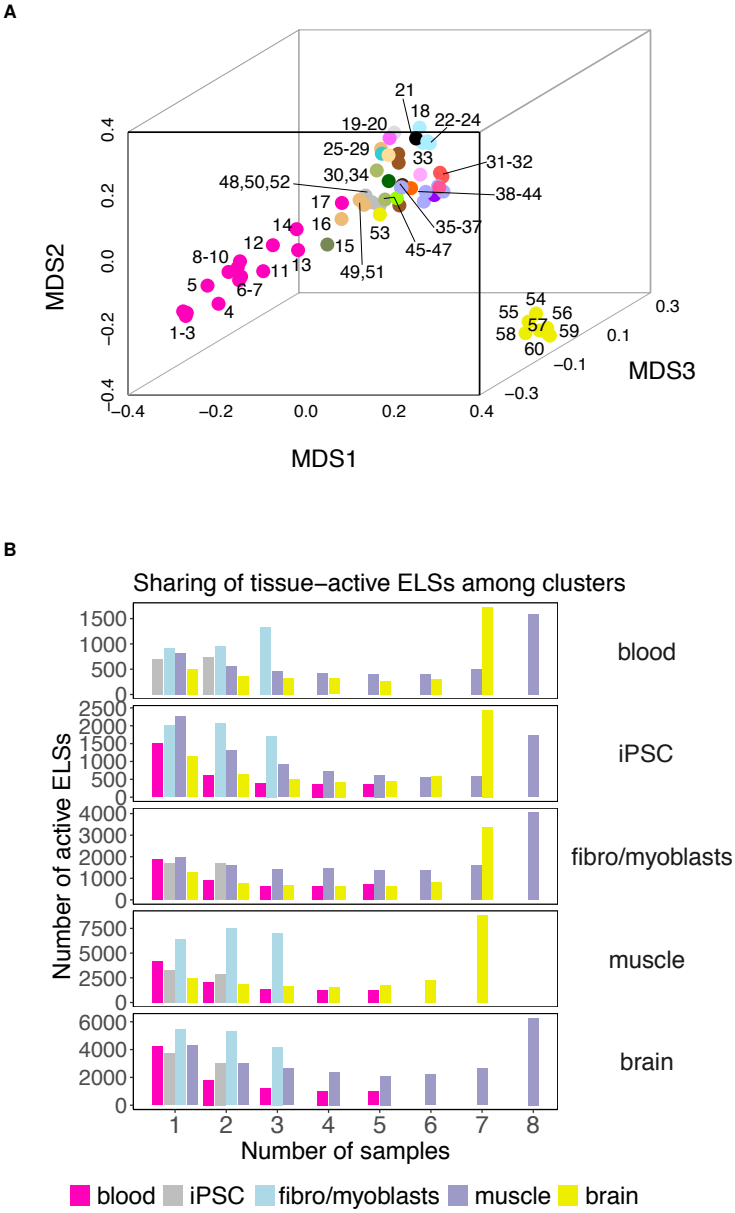


Fig. S1. **A**: Multidimensional scaling (MDS) representation of the dissimilarities between the 60 human adult samples based on the pattern of activity of ELS-cCREs. The binary distance between a given pair of samples was computed considering presence/absence vectors of the 991,173 ELS cCREs. The correspondence between samples and numbers is reported in Table S1. **B**: Tissue-active ELSs can be shared among tissues from other clusters. The histogram represents, within the different sets of tissue-active ELSs per cluster(panels 1-5), the number of ELSs also active in samples belonging to other clusters, indicating that only a proportion of ELSs are tissue-specific.

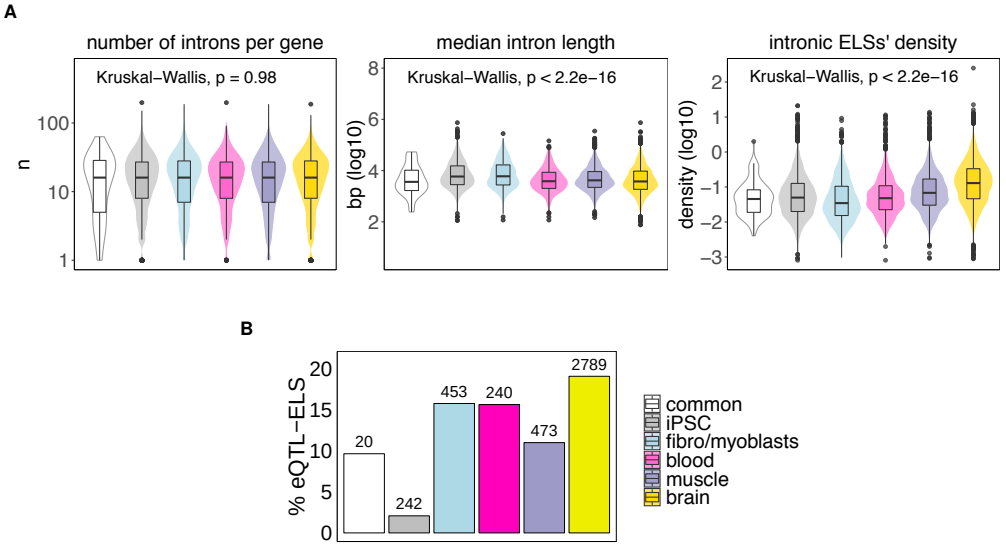


Fig. S2. **A**: Features of genes hosting either common or tissue-specific intronic ELSs identified in adult samples: (1) number of introns per hosting gene (2) median intron length per hosting gene (3) intronic ELSs' density, defined as the ratio between the number of ELSs intersecting a given intron and the size of the intron (Kb). **B**: Proportion of eQTL-ELSs with respect to the total amount of ELSs in each group.

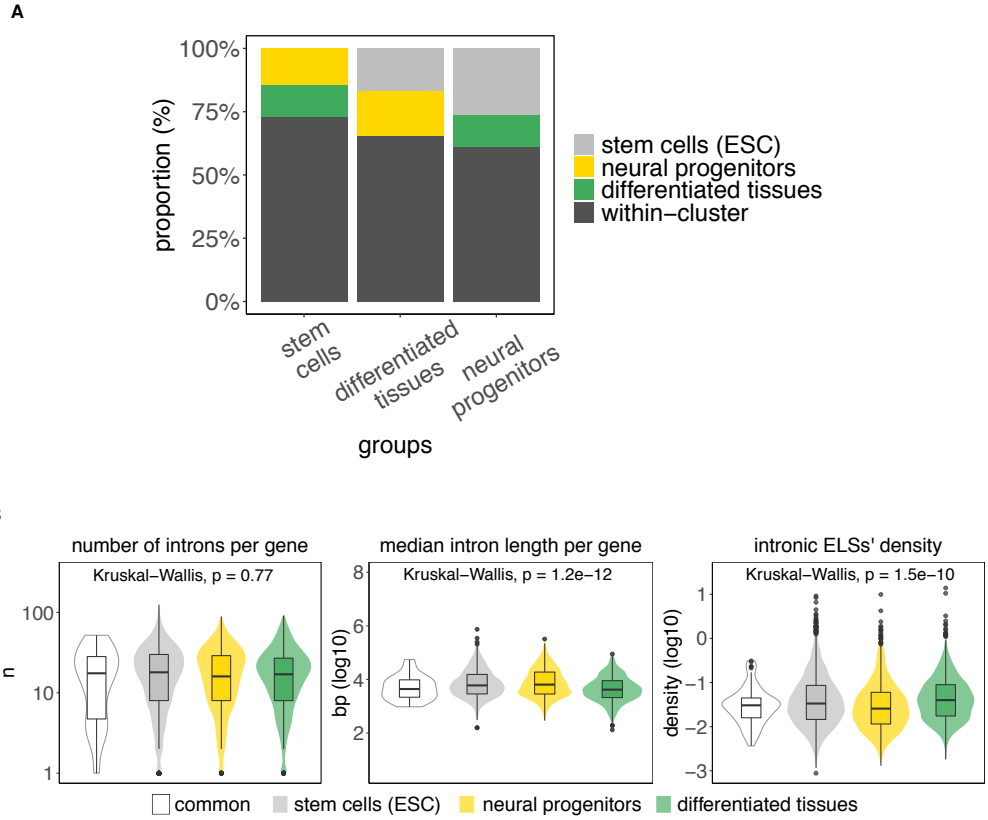


Fig. S3. **A**: Group-specific ELSs in embryonic samples (analogous to Figure 1D). The barplot represents the type of outer samples observed within sets of stem cells-, differentiated tissues- and neural progenitors-specific ELSs. **B**: Features of genes hosting either common or specific intronic ELSs identified in embryonic samples (analogous to Figure S2A): (1) number of introns per hosting gene (2) median intron length per hosting gene (3) density of ELSs per intron, defined as the ratio between the number of ELSs intersecting a given intron and the size of the intron (Kb).

	Biosample Term Name	Biosample Type	Samples' Cluster	ENCODE File ID
1	DOHH2	cell line	-	ENCFF980MYQ
2	SU-DHL-6	cell line	-	ENCFF772BBT
3	OCI-LY7	cell line	-	ENCFF250ICR
4	OCI-LY1	cell line	-	ENCFF625TTV
5	B-cell	primary cell	blood	ENCFF379TAE
6	Karpas-422	cell line	-	ENCFF106UHI
7	OCI-LY3	cell line	-	ENCFF281DXP
8	GM12878	cell line	-	ENCFF895ZSL
9	T-cell	primary cell	blood	ENCFF098NHL
10	natural killer cell	primary cell	blood	ENCFF529UWB
11	peripheral blood mononuclear cell	primary cell	blood	ENCFF509DPX
12	MM.1S	cell line	-	ENCFF971FPO
13	Loucy	cell line	-	ENCFF565KAA
14	CD14-positive monocyte	primary cell	blood	ENCFF967MJU
15	spleen	tissue	-	ENCFF821ESA
16	mucosa of rectum	tissue	-	ENCFF759YFL
17	K562	cell line	-	ENCFF095YVI
18	GM23248	cell line	-	ENCFF298RII
19	PC-3	cell line	-	ENCFF523DDS
20	HeLa-S3	cell line	-	ENCFF477PXA
21	ACC112	cell line	-	ENCFF618KHO
22	fibroblast of lung	primary cell	fibro/myoblasts	ENCFF495RTY
23	skeletal muscle myoblast	primary cell	fibro/myoblasts	ENCFF037UZZ
24	myotube	in vitro differentiated cells	fibro/myoblasts	ENCFF120MMC
25	HCT116	cell line	-	ENCFF664XAH
26	MCF-7	cell line	-	ENCFF446ZBB
27	stomach	tissue	-	ENCFF992HIZ
28	pancreas	tissue	-	ENCFF681HOL
29	body of pancreas	tissue	-	ENCFF768JUC
30	right lobe of liver	tissue	-	ENCFF476MEG
31	aorta	tissue	-	ENCFF178GDW
32	thoracic aorta	tissue	-	ENCFF257XAQ
33	ovary	tissue	-	ENCFF586NXH
34	thyroid gland	tissue	-	ENCFF296SZK
35	esophagus	tissue	-	ENCFF442HYL
36	muscle layer of duodenum	tissue	muscle	ENCFF862BGI
37	gastrocnemius medialis	tissue	-	ENCFF322RAX
38	subcutaneous abdominal adipose tissue	tissue	muscle	ENCFF725QLM
39	rectal smooth muscle tissue	tissue	muscle	ENCFF093MDL
40	vagina	tissue	muscle	ENCFF904XYE
41	stomach smooth muscle	tissue	muscle	ENCFF726JTT
42	skeletal muscle tissue	tissue	muscle	ENCFF311MNY
43	right cardiac atrium	tissue	muscle	ENCFF278RUJ
44	gastrocnemius medialis	tissue	muscle	ENCFF863OGG
45	endocrine pancreas	tissue	-	ENCFF055CJM
46	lung	tissue	-	ENCFF598QTT
47	liver	tissue	-	ENCFF645PQQ
48	GM23338	cell line	-	ENCFF651YUN
49	mucosa of rectum	tissue	-	ENCFF403IPC
50	iPS-18a	cell line	iPSC	ENCFF920QRH
51	colonic mucosa	tissue	-	ENCFF867TJN
52	iPS-20b	cell line	iPSC	ENCFF231KWX
53	bipolar neuron	in vitro differentiated cells	-	ENCFF045GKW
54	middle frontal area 46	tissue	brain	ENCFF070EXF
55	caudate nucleus	tissue	brain	ENCFF508GKP
56	angular gyrus	tissue	brain	ENCFF942KAC
57	layer of hippocampus	tissue	brain	ENCFF159NZA
58	substantia nigra	tissue	brain	ENCFF233VRB
59	temporal lobe	tissue	brain	ENCFF810IQU
60	cingulate gyrus	tissue	brain	ENCFF494WCN

Table S1. ENCODE catalogues of cell type-specific candidate cis-Regulatory Elements (cCREs) for 60 human adult samples. The accession number (ENCODE File ID) allows to uniquely identify the catalogue on the ENCODE portal (`https://www.encodeproject.org/`). The color palette was inspired by the Genotype Tissue Expression (GTEx) Project.

Samples	Tissue-active ELSs	Tissue-specific ELSs
blood	6,589	1,539
iPSC	23,743	11,666
fibro/myoblasts	18,695	2,882
muscle	29,013	4,313
brain	40,221	14,667

Samples	Common ELSs
all	208

Table S2. [upper panel] Number of ELSs active in each of the 5 clusters of 25 selected human adult samples. Tissue-active ELSs are those active in 100% (iPSC, fibro/myoblasts) or $\geq$ 80% (all other clusters) of the samples within a cluster. Tissue-specific ELSs are tissue-active ELSs that are active in 0 (iPSC, fibro/myoblasts) or at most 1 (all other clusters) outer sample (i.e. a sample that does not belong to the considered cluster). [lower panel] Number of ELSs active in 100% of the 25 selected human adult samples (common ELSs).

Genomic location	Tissue cluster	FDR	Odds ratio	Confidence interval
	iPSC	5.7e-08	0.45	0.34 - 0.61
	fibro/myoblasts	1.2e-07	0.45	0.33 - 0.61
intronic	blood	2.0e-08	0.42	0.3 - 0.57
	muscle	5.1e-25	0.22	0.16 - 0.3
	brain	9.0e-30	0.19	0.14 - 0.26
	iPSC	7.3e-01	0.79	0.31 - 1.67
	fibro/myoblasts	3.5e-01	0.64	0.25 - 1.38
exonic	blood	2.0e-02	0.40	0.16 - 0.88
	muscle	1.5e-01	0.53	0.21 - 1.14
	brain	8.2e-04	0.31	0.12 - 0.66
	iPSC	5.7e-08	2.22	1.66 - 2.99
	fibro/myoblasts	1.6e-07	2.18	1.62 - 2.96
intergenic	blood	2.0e-09	2.52	1.85 - 3.46
	muscle	1.9e-26	4.79	3.55 - 6.49
	brain	6.5e-33	5.69	4.24 - 7.66

Table S3. For each cluster of samples we assessed, with Fisher's exact test, significant differences in the proportions of common *vs* tissue-specific ELSs that overlap intronic, exonic and intergenic regions. *P* value (FDR-corrected), odds ratio and confidence interval are reported for each test.

Group	Genes ∩ ELSs			
	Introns	Exons	Both	Total
blood	548 (82.90%)	30 (4.54%)	83 (12.56%)	661
iPSC	2,008 (83.98%)	59 (2.47%)	324 (13.55%)	2,391
fibro/myoblasts	912 (86.36%)	19 (1.80%)	125 (11.84%)	1,056
muscle	977 (83.08%)	36 (3.06%)	163 (13.86%)	1,176
brain	1,647 (64.34%)	153 (5.98%)	760 (29.69%)	2,560
common	57 (90.48%)	2 (3.17%)	4 (6.35%)	63

Table S4. Number of genes whose introns and / or exons intersect tissue-specific and common ELSs identified in adult samples.

Group	GO term	Description
Brain	BP: 0007266	Rho protein signal transduction
	BP: 0007417	Central nervous system development
	BP: 0007411	Axon guidance
	BP: 0099111	Microtubule-based transport
	BP: 0007265	Ras protein signal transduction
	CC: 0044309	Neuron spine
	CC: 0044295	Axonal growth cone
	CC: 0099091	Postsynaptic specialization, intracellular component
	CC: 0099055	Integral component of postsynaptic membrane
	CC: 0099240	Intrinsic component of synaptic membrane
Blood	BP: 0046777	Protein autophosphorylation
	BP: 0002521	Leukocyte differentiation
	BP: 0051338	Regulation of transferase activity
	BP: 0043370	Regulation of CD4-positive, alpha-beta T cell differentiation
	BP: 0050867	Positive regulation of cell activation
	CC: 0001772	Immunological synapse
	CC: 0009898	Cytoplasmic side of plasma membrane
	CC: 0098562	Cytoplasmic side of membrane
	CC: 0048471	Perinuclear region of cytoplasm
	CC: 0030667	Secretory granule membrane
Muscle	BP: 0007266	Rho protein signal transduction
	BP: 0007417	Central nervous system development
	BP: 0051216	Cartilage development
	BP: 0048705	Skeletal system morphogenesis
	BP: 0007265	Ras protein signal transduction
	CC: 0042383	Sarcolemma
	CC: 0098589	Membrane region
	CC: 0030018	Z disc
	CC: 0043292	Contractile fiber
	CC: 0097517	Contractile actin filament bundle
Common	BP: 1990776	Response to angiotensin
	BP: 1901699	Cellular response to nitrogen compound
	CC: 0045121	Membrane raft
	CC: 0098857	Membrane microdomain
	CC: 0098589	Membrane region
	CC: 0005899	Insulin receptor complex

Table S5. Significantly enriched GO terms associated with genes hosting intronic ELSs identified in adult samples. Only the top five enriched terms are shown for each analysis. (BP: Biological Process; CC: Cellular Component).

Group	Genomic location	GO term	Description
Brain	Intergenic	CC:0015630	microtubule cytoskeleton
		CC:0043005	neuron projection
	Intronic	BP:0030031	cell projection assembly
		BP:0000226	microtubule cytoskeleton organization
		BP:0030030	cell projection organization
		CC:0071547	piP-body
		CC:0032420	stereocilium
		CC:0016235	aggresome
		MF:0098632	cell-cell adhesion mediator activity
		MF:0015631	tubulin binding
		MF:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
Muscle	Intergenic	BP:0006067	ethanol metabolic process
		CC:0042611	MHC protein complex
		CC:0030990	intraciliary transport particle
		CC:0010008	endosome membrane
	Intronic	BP:0009069	serine family amino acid metabolic process
		BP:1901137	carbohydrate derivative biosynthetic process
		BP:1901135	carbohydrate derivative metabolic process
		CC:0071556	integral component of lumenal side of endoplasmic reticulum membrane
		CC:0032154	cleavage furrow
		CC:0030665	clathrin-coated vesicle membrane
		MF:0032190	acrosin binding
		MF:0032395	MHC class II receptor activity
		MF:0016755	transferase activity, transferring amino-acyl groups
Blood	Intergenic	BP:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II
		BP:0060333	interferon-gamma-mediated signaling pathway
		BP:0002757	immune response-activating signal transduction
		CC:0042613	MHC class II protein complex
		CC:0071556	integral component of lumenal side of endoplasmic reticulum membrane
		CC:0032588	trans-Golgi network membrane
		MF:0032395	MHC class II receptor activity
	Intronic	-	-
iPSC	Intergenic	-	-
	Intronic	-	-
Fibro/myoblasts	Intergenic	-	-
	Intronic	CC:0015629	actin cytoskeleton
Common	Intergenic	BP:0010760	negative regulation of macrophage chemotaxis
	Intronic	CC:0009986	cell surface

Table S6. Significantly enriched GO terms associated with the intergenic and intronic eQTL-ELSs’ target genes. Only the three top enriched terms are shown for each analysis (BP: Biological Process; CC: Cellular Component; MF: Molecular Function).

Group	Genomic location	Transcription factors
Brain	Intronic	SOX4, Hic1, Zfp691, MAFA, ZBTB26, DCE, ZNF317, MXI1,Egr1, HOXA1, Ahr::Arnt, KLF15
	Intergenic	SOX8, Zscan4, NFIX, Zfp410, Irf5, AR-halfsite, POU6F1
Blood	Intronic	ELF3, RUNX1, FLI1
	Intergenic	ELF3, RUNX1
Muscle	Intronic	NFIA, MEIS1, NR1I3
	Intergenic	-
Fibro-myoblasts	Intronic	BATF, TEAD3, RUNX2
	Intergenic	BATF, TEAD3, RUNX
iPSC	Intronic	POU5F1, ZEB1, TEAD4, SOX3, Nkx6-1, Zbtb7b, GATA4, E2F7
	Intergenic	POU5F1, ZEB1, Nkx6-1, GATA2, MZF1, YY1, SRY, ZNF519, ZNF682, TEAD, DCE, MZF1, PRDM1, Arnt:Ahr
Common	Intronic	-
	Intergenic	-

Table S7. Transcription factors corresponding to the significantly enriched transcription factor binding sites (TFBSs) reported by HOMER in each group of ELSs and genomic location.

	Biosample Term Name	Biosample Type	Samples' Group	ENCODE File ID
1	HUES6	cell line	stem cells (ESC)	ENCFF205SDB
2	HUES64	cell line	stem cells (ESC)	ENCFF180QLH
3	HUES48	cell line	stem cells (ESC)	ENCFF086FKD
4	mesendoderm	in vitro differentiated cells	-	ENCFF620BVM
5	H9	cell line	stem cells (ESC)	ENCFF021HBJ
6	H9	cell line	stem cells (ESC)	ENCFF505OUS
7	H1	cell line	stem cells (ESC)	ENCFF051OUV
8	mesodermal cell	in vitro differentiated cells	-	ENCFF250CGY
9	endodermal cell	in vitro differentiated cells	-	ENCFF138DOQ
10	neuroepithelial stem cell	in vitro differentiated cells	neural progenitors	ENCFF138OGZ
11	ectodermal cell	in vitro differentiated cells	-	ENCFF332EYK
12	radial glial cell	in vitro differentiated cells	neural progenitors	ENCFF593TNG
13	neural progenitor cell	in vitro differentiated cells	neural progenitors	ENCFF112ZGF
14	mid-neurogenesis radial glial cells	in vitro differentiated cells	neural progenitors	ENCFF376XBS
15	neural stem progenitor cell	in vitro differentiated cells	neural progenitors	ENCFF455CQW
16	neural cell	in vitro differentiated cells	neural progenitors	ENCFF477EUQ
17	smooth muscle cell	in vitro differentiated cells	differentiated tissues	ENCFF281QON
18	thymus	tissue	differentiated tissues	ENCFF059PHA
19	adrenal gland	tissue	differentiated tissues	ENCFF840ANN
20	IMR-90	cell line	-	ENCFF469PXS
21	fibroblast of lung	primary cell	differentiated tissues	ENCFF292NZP
22	muscle of trunk	tissue	differentiated tissues	ENCFF800YES
23	muscle of leg	tissue	differentiated tissues	ENCFF941JIE
24	stomach	tissue	differentiated tissues	ENCFF198WHL
25	hepatocyte	in vitro differentiated cells	differentiated tissues	ENCFF093BQM
26	large intestine	tissue	differentiated tissues	ENCFF903RGX
27	small intestine	tissue	differentiated tissues	ENCFF543DVJ

Table S8. ENCODE catalogues of cell type-specific candidate cis-Regulatory Elements (cCREs) for 27 human embryonic samples. The accession number (ENCODE File ID) allows to uniquely identify the catalogue on the ENCODE portal (`https://www.encodeproject.org/`).

Samples	Group-active ELSs	Group-specific ELSs
stem cells (ESC)	7,561	3,112
neural progenitors	4,332	784
differentiated tissues	4,046	1,166

Samples	Common ELSs
all	94

Table S9. [upper panel] Number of ELSs active in each of the 3 groups of 22 selected human embryonic samples. Group-active ELSs are those active in $\geq$ 80% of the samples within a group. Group-specific ELSs are group-active ELSs that are active in at most 1 outer sample (i.e. a sample that does not belong to the considered group). [lower panel] Number of ELSs active in 100% of the 22 selected human embryonic samples (common ELSs).

Genomic location	Samples' Group	FDR	Odds ratio	Confidence interval
intronic	stem cells (ESC)	3.1e-02	0.60	0.38-0.93
	neural progenitors	1.3e-04	0.41	0.26-0.65
	differentiated tissues	9.9e-05	0.40	0.25-0.63
exonic	stem cells (ESC)	7.6e-01	1.14	0.03-7.05
	neural progenitors	1.0e+00	0.64	0.01-4.34
	differentiated tissues	8.1e-01	0.49	0.01-3.07
intergenic	stem cells (ESC)	3.1e-02	1.66	1.07-2.6
	neural progenitors	9.9e-05	2.51	1.59-4.01
	differentiated tissues	9.9e-05	2.60	1.66-4.11

Table S10. For each group of samples we assessed, with Fisher's exact test, significant differences in the proportions of common *vs* group-specific ELSs that overlap intronic, exonic and intergenic regions. *P* value (FDR-corrected), odds ratio and confidence interval are reported for each test.

Group	Genes ∩ ELSs			
	Introns	Exons	Both	Total
stem cells (ESC)	907 (89.27%)	21 (2.07%)	88 (8.66%)	1016
neural progenitors	359 (87.56%)	13 (3.17%)	38 (9.27%)	410
differentiated tissues	492 (86.16%)	24 (4.2%)	55 (9.63%)	571
common	33 (82.5%)	1 (2.5%)	6 (15%)	40

Table S11. Number of genes whose introns and / or exons intersect group-specific and common ELSs identified in embryonic samples.

Group	GO term	Description
Neural progenitors	BP: 0060291	Long-term synaptic potentiation
	BP: 0050770	Regulation of axonogenesis
	BP: 0097061	Dendritic spine organization
	CC: 0008328	Ionotropic glutamate receptor complex
	CC: 0098878	Neurotransmitter receptor complex
	CC: 0014069	Postsynaptic density
	MF: 0004970	Ionotropic glutamate receptor activity
	MF: 0005089	Rho guanyl-nucleotide exchange factor activity
	MF: 0008013	Beta-catenin binding
Differentiated tissues	BP: 1900020	Positive regulation of protein kinase C activity
	BP: 1900040	Regulation of interleukin-2 secretion
	BP: 0060766	Negative regulation of androgen receptor signaling pathway
	CC: 0098651	Basement membrane collagen trimer
	CC: 0098644	Complex of collagen trimmers
	CC: 0005583	Fibrillar collagen trimer
	MF: 0044548	S100 protein binding
	MF: 0035252	UDP-xylosyltransferase activity
	MF: 0030020	Extracellular matrix structural constituent conferring tensile strength
Stem cells (ESC)	BP: 0042908	Xenobiotic transport
	BP: 0045986	Negative regulation of smooth muscle contraction
	BP: 0098698	Postsynaptic specialization assembly
	CC: 0099092	Postsynaptic density, intracellular component
	CC: 0031304	Intrinsic component of mitochondrial inner membrane
	CC: 0008328	Ionotropic glutamate receptor complex
	MF: 0008146	Sulfotransferase activity
	MF: 0005547	Phosphatidylinositol-3,4,5-triphosphate binding
	MF: 0070300	Phosphatidic acid binding
Common	CC: 0071565	nBAF complex
	CC: 0016514	SWI/SNF complex
	CC: 0070603	NI/SNF superfamily-type complex

Table S12. Significantly enriched GO terms associated with the genes harboring intronic ELSs identified in embryonic samples. Only the top three enriched terms are shown in each analysis (BP: Biological Process; CC: Cellular Component; MF: Molecular Function).

CHAPTER 3

When to cut? Analyzing the timing of splicing

The availability of different types of NGS datasets, including eCLIP, ChIP-seq and fractional RNA-seq, has recently prompted research towards understanding the role of chromatin and RNA binding proteins (RBPs) in RNA processing events. Splicing of primary transcripts occurs mostly in the nucleus prior to export to the cytosol. In virtue of this, we have implemented a methodology that, based on the proportion of RNA-seq reads mapped to a pair of splicing junctions in the nuclear and cytosolic compartments, classifies the corresponding intron as co-transcriptionally or post-transcriptionally spliced. We have applied this method to a panel of 13 human cell lines for which fractional RNA-seq is available from the ENCODE portal. The fraction of introns undergoing post-transcriptional splicing dramatically varies across cellular conditions. We observe, for a subset of introns, that the classification in either of the two groups is shared across the majority of cell lines. Besides, co-transcriptionally spliced introns are more abundant, as previously reported. Nevertheless, a considerable fraction of introns, especially within protein-coding genes, switch from co-transcriptional to post-transcriptional splicing among cell types. We have integrated these results with the analysis of ENCODE eCLIP and ChIP-seq datasets available for a number of RBPs, transcription factors (TFs), chromatin modifiers and histone modifications. We observe a preferential binding of components of the spliceosome machinery to post-transcriptionally spliced introns, consistent with the delayed processing of these introns. We are currently developing machine learning classifiers to predict co- *vs* post-transcriptional splicing based on binding patterns of RBPs and epigenetic features.

Borsari B., Peña Castillo L. and Guigó R. Variation and constraint in the timing of splicing.

In preparation.

Variation and constraint in the timing of splicing

Beatrice Borsari[1], Lourdes Peña Castillo[2] and Roderic Guigó[1,3]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
[2]Departments of Computer Science and Biology, Memorial University of Newfoundland, St. John’s, Canada
[3]Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain

Introduction

The large availability of different types of Next Generation Sequencing datasets, including eCLIP and fractional RNA-seq, as well as a plethora of ATAC-seq data, and ChIP-seq data for histone modifications and DNA-binding factors, has prompted research towards understanding the role of chromatin and RNA-binding proteins in RNA processing events. A relationship between nucleosome positioning and exon recognition was described more than a decade ago, with higher nucleosome occupancy rates in exons compared to introns reported across multiple species (Andersson et al., 2009, Schwartz et al., 2009, Tilgner et al., 2009, Wilhelm et al., 2011). This exonic nucleosome preference does not seem to be transcription-dependent, since it has also been observed in transcriptionally silent genes. Instead, H3K36me3 marking at exons has been shown to correlate with exon expression and inclusion (Kolasinska-Zwierz et al., 2009, Spies et al., 2009).

On the other hand, the availability of fractional RNA-seq experiments allows to track the degree of splicing completion transcriptome-wide, and to compare it between the nuclear and cytosolic compartments. Based on metrics such as *cosi* and *cosit* (Pervouchine et al., 2013, Tilgner et al., 2012), it is possible to compute the degree of splicing completeness for a given exon or intron, respectively, therefore to assess whether it has been spliced co-transcriptionally (i.e. it displays comparable and high degrees of splicing completion in both compartments) or post-transcriptionally (it shows, instead, a higher degree of splicing completion in the cytosol compared to the nucleus). It has been reported that splicing in the human genome is predominantly co-transcriptional, although inefficient for lncRNAs, which sometimes appear unspliced also in the cytosolic RNA-seq fraction (Tilgner et al., 2012). A 5’-to-3’ trend has also been observed, with downstream splicing events in the transcript more prone to occur post-transcriptionally. Furthermore, post-transcriptionally spliced exons have been described as enriched in canonically active chromatin marks, nucleosome occupancy and Pol II occupancy, compared to exons that are rapidly excised from the transcript (Tilgner et al., 2012).

Here, we analyze differences in splicing completion between nucleus and cytosol across a panel of 13 human cell lines, and explore patterns of chromatin and RNA binding features associated with intronic segments that are retained for a longer time within the transcript.

Results

With the aim of investigating the timing of splicing across different cellular conditions, we obtained fractional polyA+ RNA-seq profiles generated by the ENCODE project (Davis et al., 2018, Dunham et al., 2012) in 13 human cell lines (Djebali et al., 2012). Because splicing of primary transcripts mostly occurs in the nucleus prior to export to the cytosol (Buckley et al., 2014, Naftelberg et al., 2015), we took advantage of this spatial constraint as a proxy to study differences in the splicing time of introns.

For each cell line, we quantified splice junctions with the Integrative Pipeline for Splicing Analyses (IPSA, `https://github.com/guigolab/ipsa-nf`), independently for nuclear and cytosolic compartments. Introns with splice junctions lowly represented in both nuclear and cytosolic RNA-seq fractions were considered to be rapidly excised from the transcript (co-transcriptionally spliced, cs; Supplementary Figures 1a-b). On the contrary, introns displaying higher enrichment of junction reads in the nucleus compared to the cytosol were considered as retained for a longer time in the transcript (post-transcriptionally spliced, ps). Given the accumulating evidence for widespread intron retention events (Braunschweig et al., 2014, Pandya-Jones et al., 2013), we also considered the case of introns comparably retained within both nuclear and cytosolic transcripts (unprocessed introns) (Supplementary Figures 1a-b).

To rule out confounding effects due to alternative splicing events, we focused on a subset of 73,064 introns constitutively annotated in Gencode version 24 (Supplementary Figure 1c). Thus, for every cell line, we classified the expressed introns in this subset as co-/post-transcriptionally spliced or unprocessed, based on the proportion of reads supporting splicing incompletion and completion in the two cellular fractions (Supplementary Figures 1b, d, e). Introns that, counterintuitively, displayed a higher rate of splicing completion in the nucleus compared to the cytosol were considered artifacts (Supplementary Figures 1a, b, d, e; see Methods).

In our analyses, cs introns are more abundant than ps introns (Figures 1a-b, Supplementary Figure 1f), consistent with previous observations (Tilgner et al., 2012), while the fraction of unprocessed or artifactually spliced intronic sequences is overall negligible. Our classification is supported by the previously described intron-centric measure of splicing completion *cosit* (Supplementary Figure 1f). Remarkably, we observed subsets of introns constitutively cs (n = 16,798; 23%; Figures 1a, 1c) or ps (n = 12,751; 17%; Figures 1a, 1d) across all cell lines, while approximately 15% of the analyzed introns switch from co-transcriptional to post-transcriptional splicing, especially in endothelial cells and MCF-7, GM12878, K562 and IMR-90 cell lines (Figures 1a-b). Altogether, this suggests that, for a subset of introns, their splicing time is tightly regulated across different cell types. Constitutive ps introns are shorter, belong to shorter transcripts which present fewer exons and are more expressed than those carrying constitutively cs introns (Figure 1e). Moreover, ps introns are characterized by a higher GC content and present weaker acceptor (but not donor) sites (Supplementary Figure 1g). Genes hosting constitutive ps introns are enriched in functions associated with RNA processing and ncRNA metabolism (Figure 1f, right side), in line with mechanisms of splicing autoregulation previously described (Pervouchine et al., 2019). In contrast, genes carrying cs introns tend to be involved in cell division and DNA-based processes (Figure 1f, left side).

We next investigated whether cs and ps introns present specific patterns of RNA- and DNA-binding factors. To this end, we took advantage of the ENCODE collection of eCLIP experiments (Van Nostrand et al., 2020), which have assayed a number of RNA-binding proteins in K562 and HepG2 cell lines. We specifically focused on a subset of 73 RBPs profiled in both K562 and HepG2, and compared their enrichment between the sets of cs and ps introns identified in each cell line (Figure 2a). We observed differential binding of most RBPs between cs and ps introns in both cell lines, with higher frequency of peaks in the set of ps introns. The proportion of cs and ps introns bound by a given RBP is indeed consistent between K562 and HepG2. Nonetheless, only for a subset of RBPs – specifically KHSRP, RBM22, SF3B4, AQR, EFTUD2, PRPF8, which are well known members of the spliceosome machinery (source: UniProt, `https://www.uniprot.org/`) – this proportion is above 10% in both cell lines. We next assessed whether the enrichment of eCLIP peaks in ps compared to cs introns is statistically significant, and to which extent it is also shared by unprocessed introns. Thus, for each of the 73 RBPs, we pairwise compared the proportions of introns with eCLIP peaks between cs, ps, and unprocessed introns (two-sided Fisher's exact test, p-value < 0.01, odds-ratio < 0.56 or odds-ratio > 1.8). We identified 11 RBPs consistently more bound to ps and unprocessed introns than cs introns (Supplementary Figure 2). Of note, these RBPs are involved in RNA processing functions not related to splicing, such as RNA capping, cleavage and polyadenylation, but also RNA transport and methylation, and regulation of RNA translation (source: UniProt). Therefore, the binding of these factors to intronic sequences can be considered a general feature of introns retained for a longer time in the transcript, independently of their splicing outcome. On the other hand, four (RBM22, SF3B4, AQR, PRPF8) of the six aforementioned RBPs appear to selectively bind ps introns. SF3B4 is involved in the formation of splicing Complex A, while RBM22, AQR and PRPF8 – which have been reported to interact – contribute to the formation of splicing Complexes B and C (source: UniProt). This is consistent with these RBPs localizing in ps but not in unprocessed introns. On the other hand, exons flanking ps and cs introns show enrichment in a distinct set of RBPs, including FXR2, ZNF800, GRWD1, RPS3, UCHL5 and BUD13, which besides appear more evenly distributed between ps and cs exons (Figure 2b). Overall, this suggests that binding patterns of distinct RBPs characterize introns and exons, and that the enrichment of specific splicing factors correlates with the time required to excise introns.

We performed similar analyses with the ENCODE collection of ChIP-seq datasets for transcription factors (TFs) and chromatin modifiers, available for five of the considered cell lines. Nonetheless, we observed a scarcity of binding events in both cs and ps introns (Supplementary Figures 3a-b), which in all cases did not involve more than 10% of the total introns, pointing to a comparatively minor role of DNA-binding factors in discriminating the differential timing of splicing events. On the other hand, when focusing on histone modifications, we found that H3K36me3 marks a large fraction of the analyzed introns – often more than 30% (Figure 3a) –, and especially in cell lines characterized by higher rates of post-transcriptional splicing (such as GM12878, endothelial cells and K562), H3K36me3 is significantly more enriched in ps than cs introns (Figure 3b). Moreover, this mark appears to be enriched in spliced rather than unprocessed introns consistently across all cell lines (Figure 3b). In contrast, marking by H3K79me2, another modifi-

cation that broadly covers actively transcribed gene bodies, is less abundant at introns than H3K36me3, and more often enriched in ps/unprocessed than cs introns. Instead, H3K4me1 marks unspliced introns, being comparably depleted from cs and ps introns. In a few cell lines we also reported H4K20me1 signal enriched in ps compared to cs introns (Figures 3a-b). Exons flanking cs and ps introns are also marked by H3K36me3 at similar or even higher rates than introns, but show lower frequency of H3K79me2, H3K4me1 and H4K20me1 peaks (Supplementary Figure 3c). Overall, this suggests that marking of introns by distinct histone modifications can be associated with differences in splicing time (co-/post-transcriptionally: H3K36me3) and efficiency (retained/excised: H3K4me1, H3K36me3).

Besides patterns of DNA- and RNA-binding events, we also investigated whether other features are associated with the marked differences in co- *vs* post-transcriptional splicing observed across distinct cell types. To do so, we computed Spearman's correlation coefficients between the rate of co-transcriptional splicing (Figure 1b) and the nuclear or cytosolic expression levels of genes across the 13 cell lines. We identified three genes (*CTNS*, *SLC6A15* and *PSAT1*; Figure 4a) whose expression across cell lines is significantly and positively correlated (FDR < 0.1, Spearman's $\rho > 0.7$) with the frequency of co-transcriptional splicing. Among others, these genes are associated with functions related to amino-acid transport, which prompted us to investigate whether the timing of splicing of protein-coding RNAs is more tightly regulated in different conditions, compared to non-coding RNAs. We thus focused on those introns expressed in at least 9 out of the 13 (excluding artifactually spliced introns). Among them, we identified 927 introns belonging to 238 uniquely non-coding genes. We thus selected an equal number of uniquely protein-coding genes (*n* introns = 1,079) displaying similar distributions of expression levels (Supplementary Figure 4a). In order to analyze switches from co-transcriptional to post-transcriptional splicing or intron retention, we considered subsets of introns classified as either cs or ps/unprocessed in an equivalent number of cell lines (from 3 to 6, i.e. introns classified as cs in $\geq$3 cell lines and as ps/unprocessed in $\geq$3 other cell lines), and ensured that the overall rate of post-transcriptional splicing/intron retention across cell lines (Supplementary Figure 4b) was consistent with the one initially observed (Figure 1b). When considering these subsets, we found that protein-coding genes more frequently carry introns with differential timing of splicing across cell lines, compared to non-coding genes (Figure 4b). On the other hand, when considering introns constitutively classified as cs (Figure 4c) or ps (Figure 4d) (*n* cell lines $\geq$ 11, 12 or 13), we found the opposite trend, with introns of non-coding genes less prone to changes in splicing time. Although these results do not directly explain the differences in the timing of splicing observed among cell lines, they open the possibility that this process may be more tightly regulated in the case of protein-coding RNAs, perhaps as a function of the amount of mature transcripts available for translation.

Ongoing work

We are currently working on the implementation of different types of machine learning algorithms to classify cs *vs* ps introns, taking into account patterns of epigenetic features and RBPs that characterize both the

introns and their flanking exons. Within this framework, we are additionally integrating splicing QTLs, which we have reported in a previous work to be preferentially located in ps introns (Garrido-Martín et al., 2020, in press), as well as sequence-derived features (e.g. trinucleotide frequency).

Acknowledgments

We thank Dmitri Pervouchine and Reza Sodaei for insightful suggestions. We thank the ENCODE Consortium, in particular Gingeras (fractional RNA-seq), Graveley (eCLIP), Bernstein, Farnham, Fu, Keles, Myers, Stamatoyannopoulos, Snyder, Weissman and White (ChIP-seq) laboratories for data production.

References

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Research*, 19(10):1732–1741.

Braunschweig, U., Barbosa-Morais, N. L., Pan, Q., Nachman, E. N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., and Blencowe, B. J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, 24(11):1774–1786.

Buckley, P. T., Khaladkar, M., Kim, J., and Eberwine, J. (2014). Cytoplasmic intron retention, function, splicing, and the sentinel RNA hypothesis. *Wiley Interdisciplinary Reviews: RNA*, 5(2):223–230.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, 46(D1):D794–D801.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–108.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genetics*, 41(3):376–381.

Naftelberg, S., Schor, I. E., Ast, G., and Kornblihtt, A. R. (2015). Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annual Review of Biochemistry*, 84:165–198.

Pandya-Jones, A., Bhatt, D. M., Lin, C. H., Tong, A. J., Smale, S. T., and Black, D. L. (2013). Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression. *RNA*, 19(6):811–827.

Pervouchine, D., Popov, Y., Berry, A., Borsari, B., Frankish, A., and Guigó, R. (2019). Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. *Nucleic Acids Research*, 47(10):5293–5306.

Pervouchine, D. D., Knowles, D. G., and Guigó, R. (2013). Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics*, 29(2):273–274.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nature Structural and Molecular Biology*, 16(9):990–995.

Spies, N., Nielsen, C. B., Padgett, R. A., and Burge, C. B. (2009). Biased Chromatin Signatures around Polyadenylation Sites and Exons. *Molecular Cell*, 36(2):245–254.

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nature Structural & Molecular Biology*, 16(9):996–1001.

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J. Y., Cody, N. A., Dominguez, D., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 583(7818):711–719.

Wilhelm, B. T., Marguerat, S., Aligianni, S., Codlin, S., Watt, S., and Bähler, J. (2011). Differential patterns of intronic and exonic DNA regions with respect to RNA polymerase II occupancy, nucleosome density and H3K36me3 marking in fission yeast. *Genome Biology*, 12(8).

Figure 1

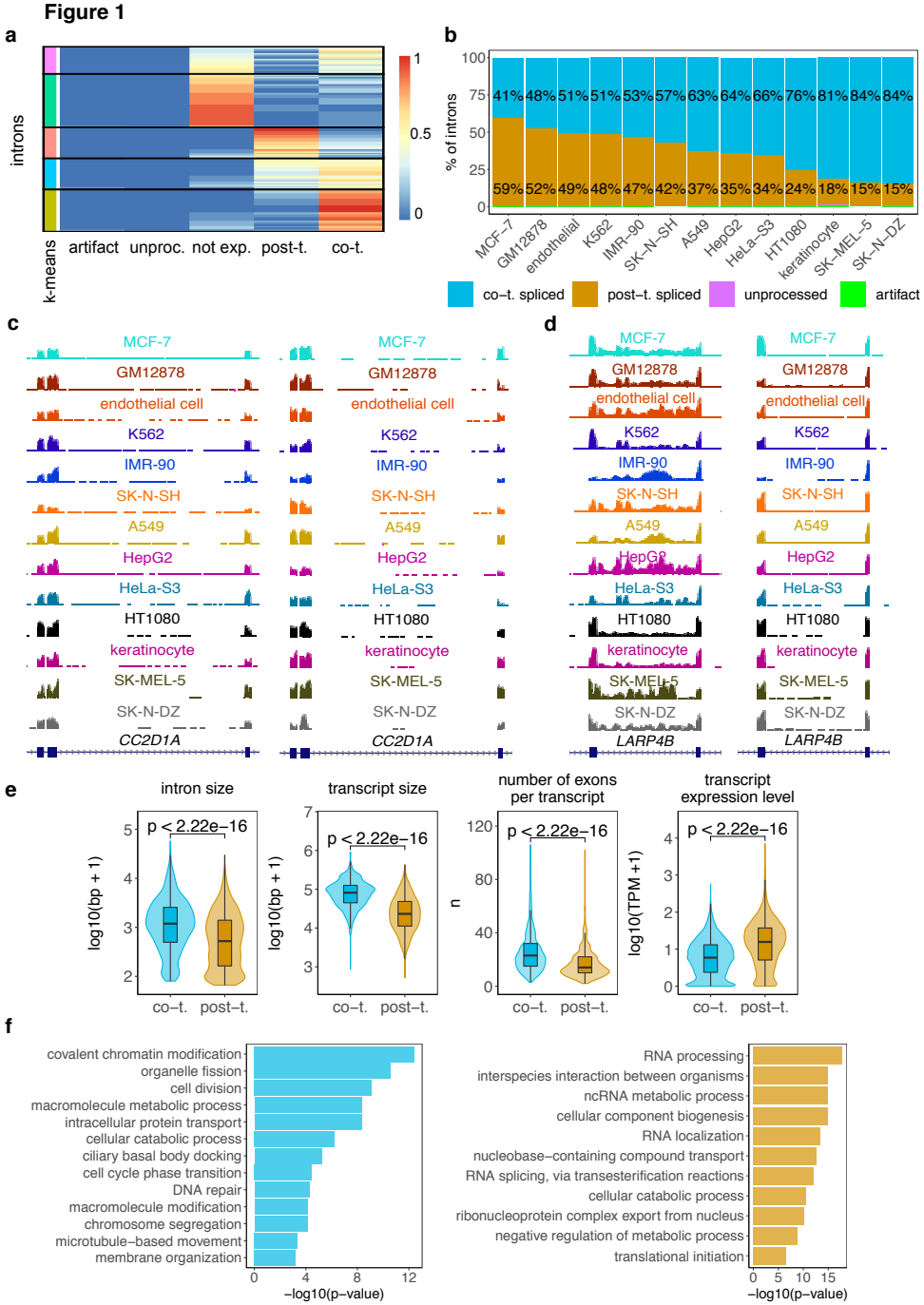

a
introns
k-means
artifact
unproc.
not exp.
post-t.
co-t.
1
0.5
0
b
% of introns
100
75
50
25
0
41%
48%
51%
51%
53%
57%
63%
64%
66%
76%
81%
84%
84%
59%
52%
49%
48%
47%
42%
37%
35%
34%
24%
18%
15%
15%
MCF-7
GM12878
endothelial
K562
IMR-90
SK-N-SH
A549
HepG2
HeLa-S3
HT1080
keratinocyte
SK-MEL-5
SK-N-DZ
co–t. spliced
post–t. spliced
unprocessed
artifact
c
MCF-7
GM12878
endothelial cell
K562
IMR-90
SK-N-SH
A549
HepG2
HeLa-S3
HT1080
keratinocyte
SK-MEL-5
SK-N-DZ
CC2D1A
d
LARP4B
e
intron size
transcript size
number of exons per transcript
transcript expression level
log10(bp + 1)
n
log10(TPM +1)
p < 2.22e−16
co−t.
post−t.
f
covalent chromatin modification
organelle fission
cell division
macromolecule metabolic process
intracellular protein transport
cellular catabolic process
ciliary basal body docking
cell cycle phase transition
DNA repair
macromolecule modification
chromosome segregation
microtubule−based movement
membrane organization
−log10(p−value)
RNA processing
interspecies interaction between organisms
ncRNA metabolic process
cellular component biogenesis
RNA localization
nucleobase−containing compound transport
RNA splicing, via transesterification reactions
cellular catabolic process
ribonucleoprotein complex export from nucleus
negative regulation of metabolic process
translational initiation

Figure 1. **a**: Heatmap representing the frequency at which the selected 73,064 non-redundant introns (rows) are either not expressed, or classified in one of the four groups (artifact / unprocessed / co-t. / post-t. spliced) across the 13 cell lines. K-means clustering of introns based on these frequencies highlights five main clusters (from top to bottom): 1) introns that are either not expressed or cs (n = 10,925); 2) introns that are not expressed in the majority of cell lines (n = 21,027); 3) introns that are constitutively ps (n = 12,751); 4) introns that are either cs or ps in roughly the same number of cell lines (n = 11,563); 5) introns that are constitutively cs (n = 16,798). **b**: Stacked barplot depicting the proportion of cs, ps, artifactually spliced and unprocessed introns in the 13 cell lines. **c**: Example of an intron within gene *CC2D1A* that is constitutively cs (RNA-seq uniquely mapped reads in nucleus and cytosol displayed on the left and on the right, respectively; chr19:13,900,786-13,930,701). **d**: Example of an intron within gene *LARP4B* that is constitutively ps (chr10:832,426-855,618). **e**: Distributions of intron size, transcript size, number of exons per transcript and transcript expression levels. **f**: Gene Ontology enriched terms for genes hosting constitutive cs and ps introns.

Figure 2

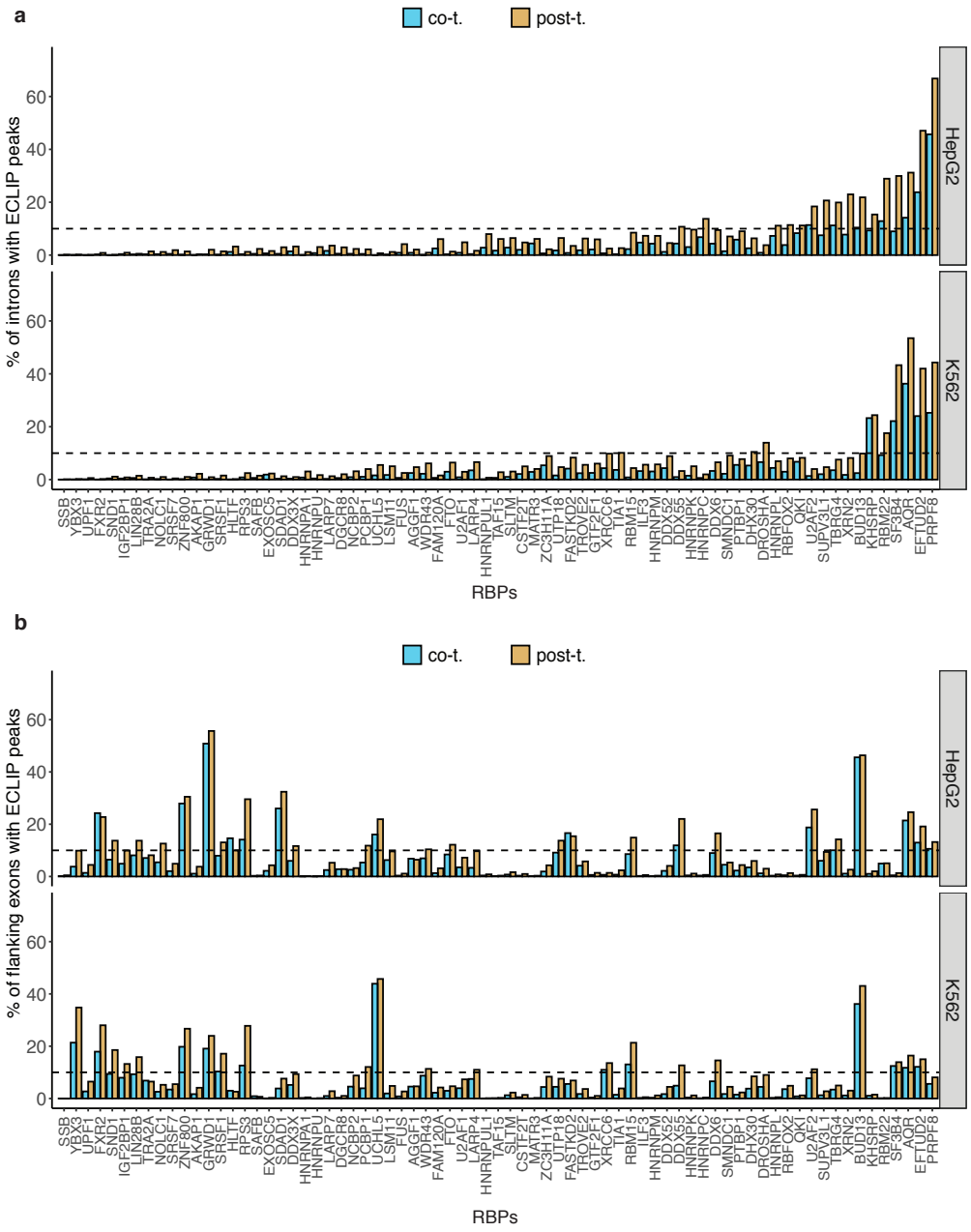

a
co-t.
post-t.
% of introns with ECLIP peaks
HepG2
K562
RBPs
b
co-t.
post-t.
% of flanking exons with ECLIP peaks
HepG2
K562
RBPs

Figure 2. **a**: Barplot depicting the proportion of cs and ps introns in HepG2 and K562 cell lines that display eCLIP peaks. Only the 73 RBPs profiled with eCLIP in both cell lines are represented. **b**: Analogous representation for the exons flanking cs and ps introns.

Figure 3

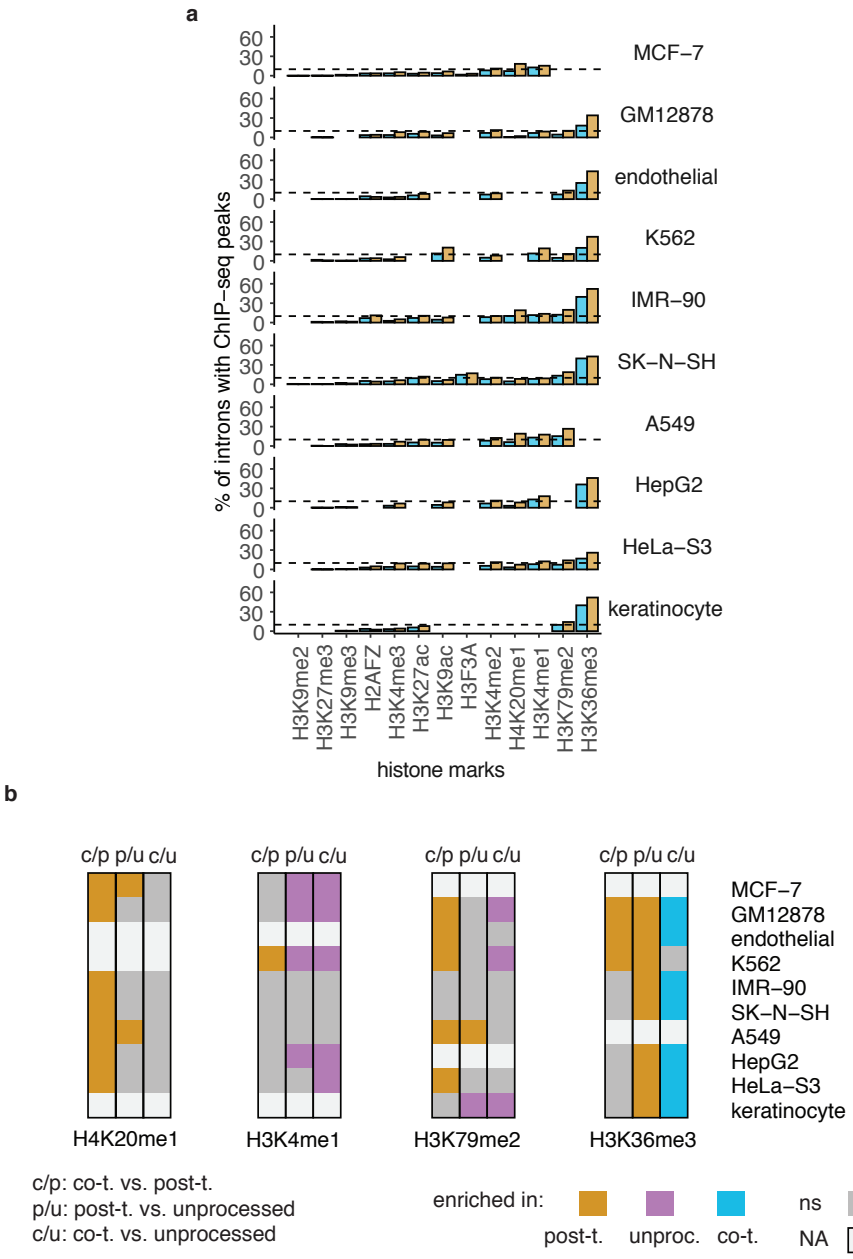

a
% of introns with ChIP−seq peaks
60
30
0
MCF−7
GM12878
endothelial
K562
IMR−90
SK−N−SH
A549
HepG2
HeLa−S3
keratinocyte
H3K9me2
H3K27me3
H3K9me3
H2AFZ
H3K4me3
H3K27ac
H3K9ac
H3F3A
H3K4me2
H4K20me1
H3K4me1
H3K79me2
H3K36me3
histone marks
b
c/p p/u c/u
H4K20me1
H3K4me1
H3K79me2
H3K36me3
MCF−7
GM12878
endothelial
K562
IMR−90
SK−N−SH
A549
HepG2
HeLa−S3
keratinocyte
c/p: co-t. vs. post-t.
p/u: post-t. vs. unprocessed
c/u: co-t. vs. unprocessed
enriched in:
post-t.
unproc.
co-t.
ns
NA

Figure 3. **a**: Barplot depicting the proportion of cs and ps introns that display peaks of histone modifications. Three cell lines were excluded from the analysis because of the absence of ChIP-seq experiments. **b**: Heatmaps representing histone modifications significantly enriched in groups of ps and unprocessed introns. We pairwise compared the frequency of a given histone mark among groups of cs vs ps (c/p), cs vs unprocessed (c/u), or ps vs unprocessed (p/u) introns (two-sided Fisher's exact test, p-value < 0.01, odds-ratio < 0.56 or odds-ratio > 1.8). Significant enrichment in a group of introns is color-coded. Absence of significant enrichment is shown in gray. Unavailable (NA) ChIP-seq experiments are shown in white.

Figure 4

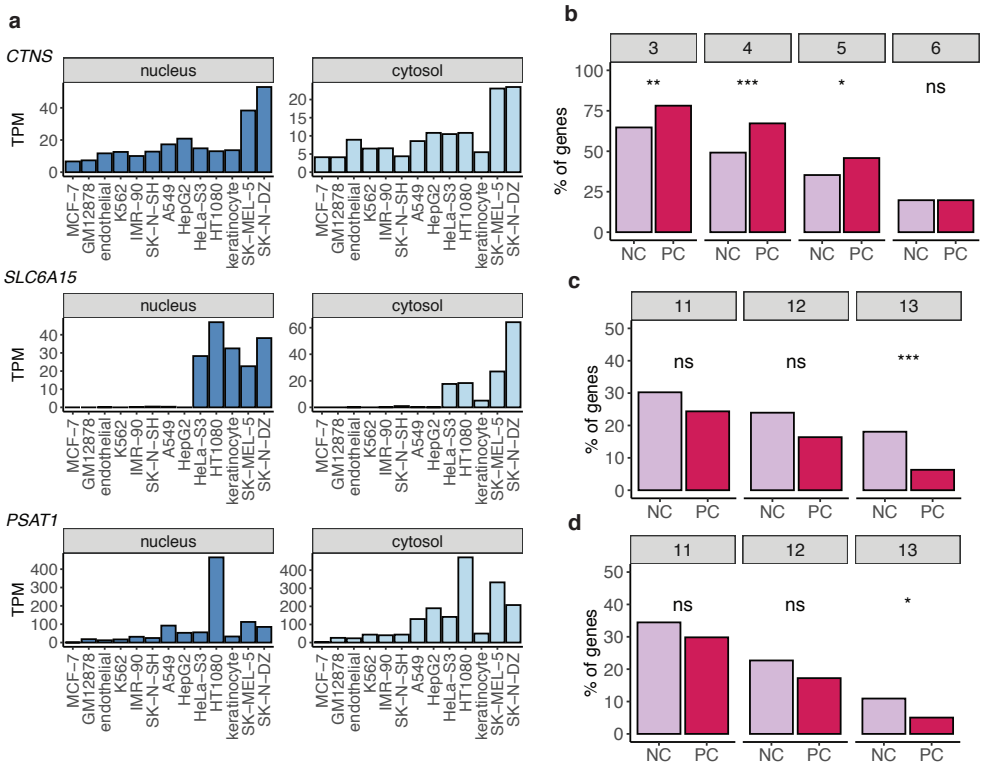
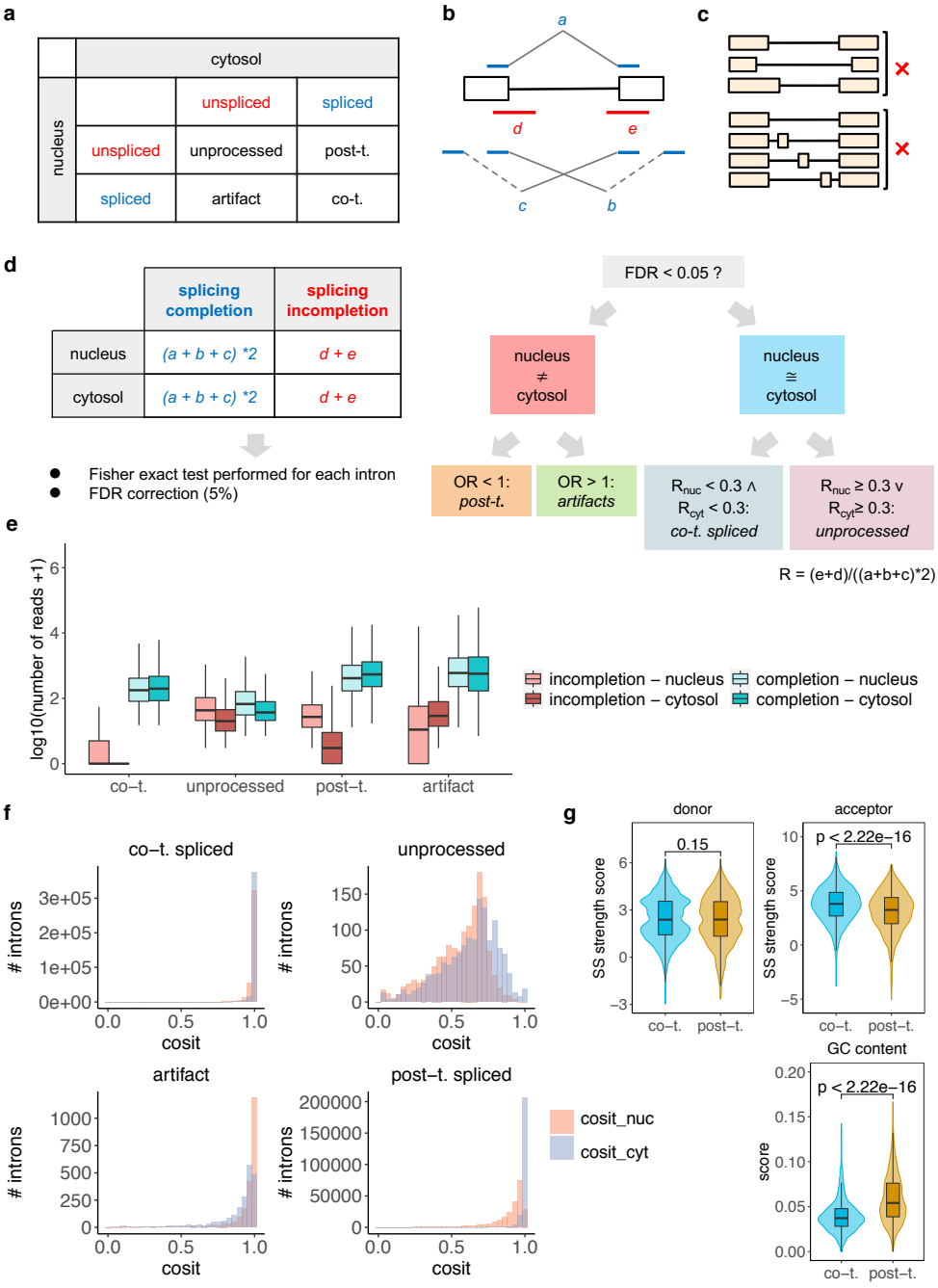

a
CTNS
nucleus
cytosol
TPM
SLC6A15
PSAT1
MCF−7
GM12878
endothelial
K562
IMR−90
SK−N−SH
A549
HepG2
HeLa−S3
HT1080
keratinocyte
SK−MEL−5
SK−N−DZ
b
3
4
5
6
% of genes
**
***
*
ns
NC
PC
c
11
12
13
d

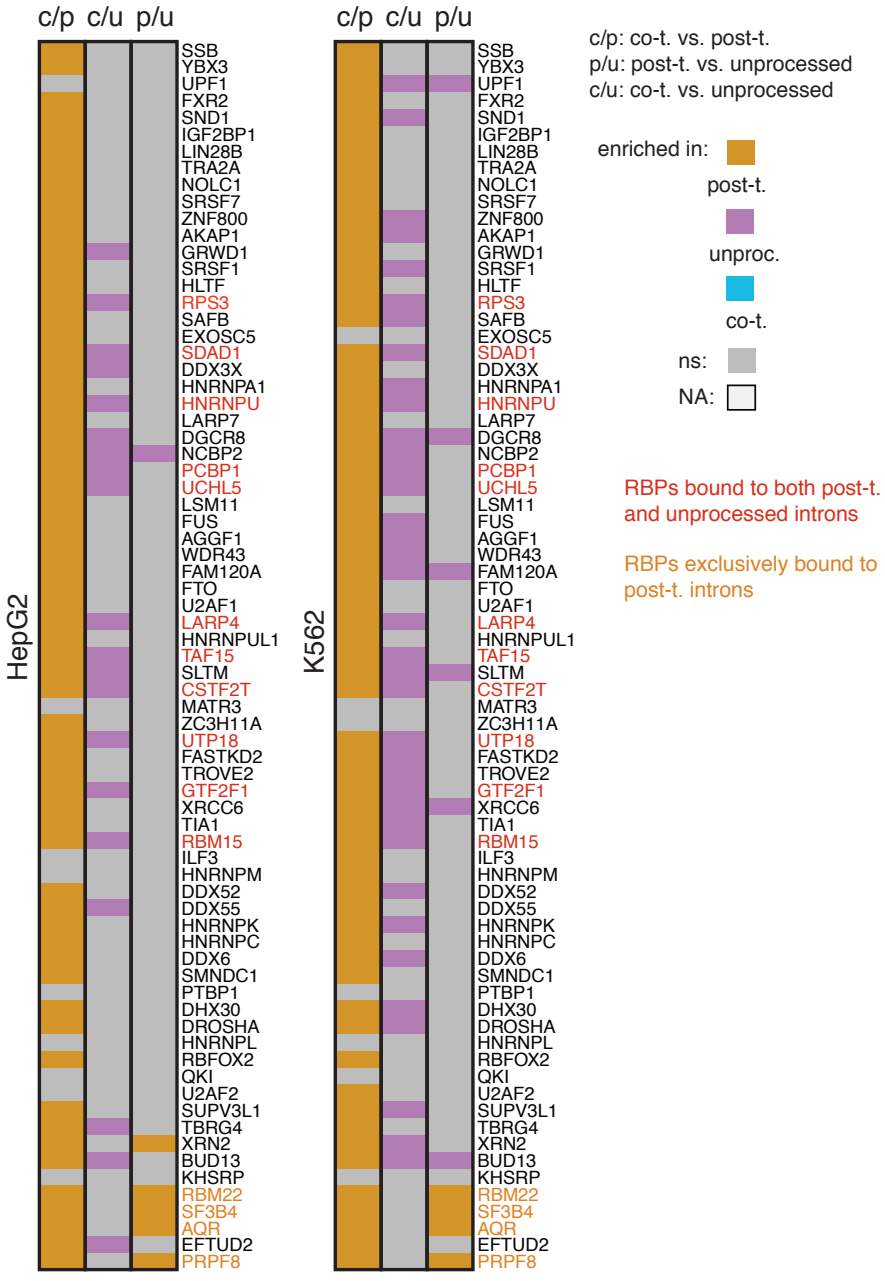
Figure 4. **a**: Genes whose expression profile across the 13 cell lines (either in the nuclear or cytosolic fractions) is significantly positively correlated (FDR < 0.1, Spearman's $\rho > 0.7$) with the rate of co-transcriptional splicing. **b**: Proportion of non-coding (NC) and protein-coding (PC) genes containing at least one intron switching from co-transcriptional to post-transcriptional splicing or intron retention in increasing numbers (3-6) of cell lines. **c**: Proportion of non-coding and protein-coding genes containing at least one intron constitutively ps or unprocessed. **d**: Analogous representation for genes containing introns constitutively cs.

Supplementary Figure 1



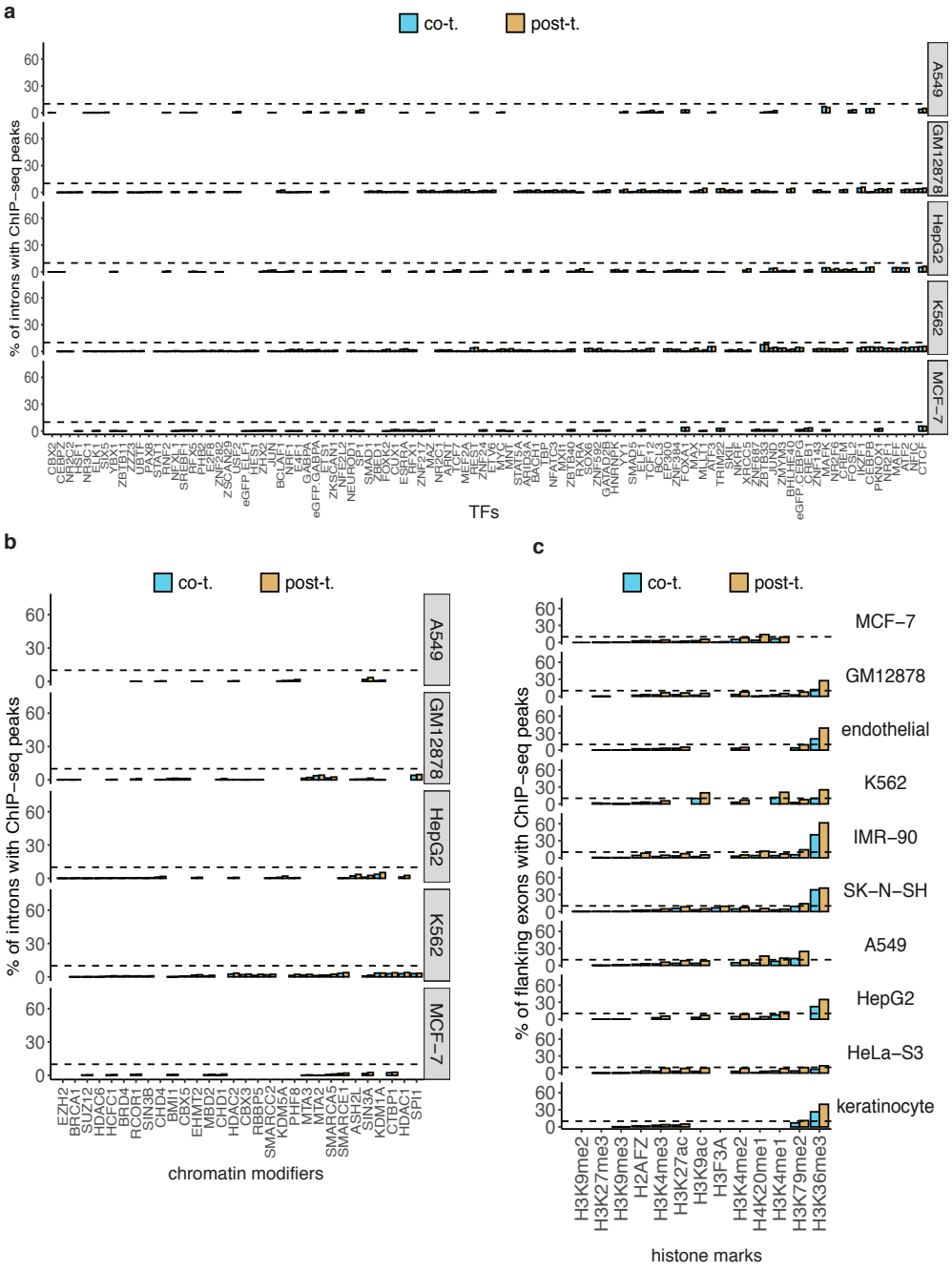
Supplementary Figure 1. **a**: Summary of the definition of co-transcriptionally spliced, post-transcriptionally spliced, artifactually spliced and unprocessed introns according to their splicing status in the nucleus and cytosol. **b**: Splice junction RNA-seq reads detected by IPSA that support either splicing incompletion (*d* and *e*; red) or completion (*a*, *b*, *c*; blue). **c**: Schematic representation of the types of introns excluded from the analysis to avoid confounding effects due to alternative splicing events. **d**: [right side] For each intron we used the depicted contingency table to perform Fisher's exact test, in order to assess the differential proportions of reads supporting splicing completion between the nuclear and cytoplasmic compartments. FDR correction was applied exclusively on the selected set of 73,064 non-redundant introns (i.e. same chromosome, start, end, strand). [left side] Among the introns that display significantly different proportions of splicing completion reads between the two compartments (FDR $<$ 0.05), we distinguished between those that are post-transcriptionally (odds-ratio $<$ 1) and artifactually (odds-ratio $>$ 1) spliced. Those introns that are not differentially spliced between the nuclear and cytosolic compartments (FDR $\geq$ 0.05) were classified as either co-transcriptionally spliced ($R_{nuc} < 0.3$ AND $R_{cyt} < 0.3$) or unprocessed ($R_{nuc} \geq 0.3$ OR $R_{cyt} \geq 0.3$), with R = (e+d)/((a+b+c)*2). **e**: Distribution of number of reads supporting splicing incompletion and completion in the two cellular fractions for the four groups of introns. **f**: Distributions of *cosit* values in nucleus and cytosol for the four groups of introns (*cosit* = 0: absence of splicing; *cosit* = 1; complete splicing). **g**: Distributions of splice donor and acceptor sites strength, as well as GC content, for constitutive cs and ps introns.

Supplementary Figure 2



Supplementary Figure 2. **a**: Heatmaps representing RBPs significantly enriched in groups of ps and unprocessed introns in HepG2 and K562 cell lines. We pairwise compared the binding frequency of a given RBP between groups of cs *vs* ps (c/p), cs *vs* unprocessed (c/u), or ps *vs* unprocessed (p/u) introns (two-sided Fisher's exact test, p-value < 0.01, odds-ratio < 0.56 or odds-ratio > 1.8). Significant enrichment in a given group of introns is color-coded. We highlighted in red the RBPs that, consistently in the two cell lines, are significantly more bound to both ps and unprocessed introns, and in orange those RBPs enriched only in ps introns. Lack of significantly different binding is shown in gray.

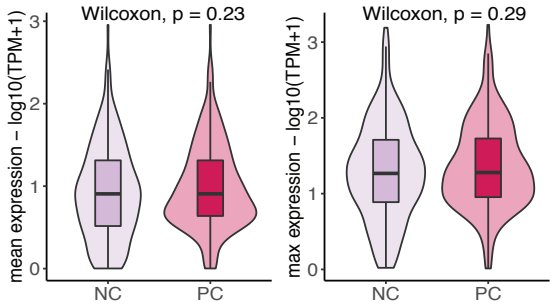
Supplementary Figure 3



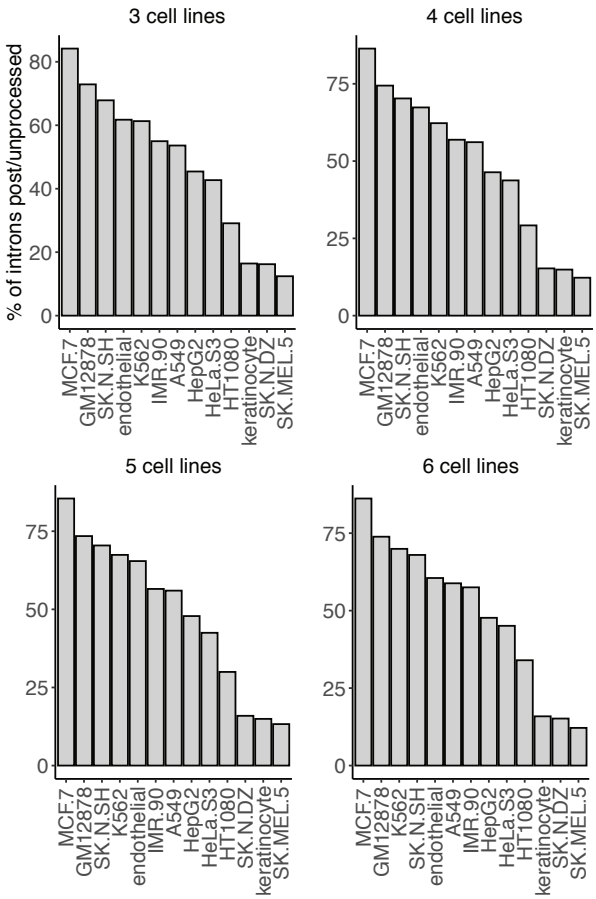
Supplementary Figure 3. **a**: Barplot depicting the proportion of cs and ps introns that display TFs' peaks. We focused on the five cell lines more extensively characterized by ENCODE ChIP-seq experiments, and reported TFs profiled in at least two cell lines. **b**: Analogous representation for chromatin modifiers.

Supplementary Figure 4

a



b



Supplementary Figure 4. **a**: Distributions of mean and maximum expression levels for the sets of uniquely non-coding (NC) and uniquely protein-coding (PC) genes. **b**: Barplot reporting, for subsets of introns switching classification in increasing numbers of cell lines, the proportion of ps/unprocessed introns found within each cell line. For instance, in the first barplot (“3 cell lines”) we are considering introns that are classified as cs in at least 3 cell lines, and as ps/unprocess in at least 3 other cell lines. We observe that the amount of delayed/absent splicing (ps/unprocessed) across cell lines is consistent with the general trend initially observed with the entire set of introns.

DISCUSSION

This thesis work initially focused on understanding the interplay between gene expression and histone marks over time, by analyzing a rich set of transcriptomic and chromatin maps generated during a controlled transdifferentiation process, as depicted in Chapter 1. In addition to the extensive data resource, we believe that our findings uncover some general principles that govern the relationship between histone marks and gene expression, integrating the diverse and apparently contradictory observations in the field. Although they lack a causal role on gene expression, histone marks correlate with certain biochemical activities, and their genomic enrichment can help identify loci with potential regulatory functions. With this perspective, in Chapter 2 we moved from a time-oriented to a space-oriented analysis of the epigenome, and investigated the genomic location of distal regulatory elements in developmental and adult tissues. Furthermore, the observation that histone marks are deposited after *de novo* gene activation motivated us to assess their contribution to the timing of RNA processing events, extending the analysis of the transcriptome not only to different cell types but also to distinct subcellular compartments (Chapter 3). Altogether, this work represents a valuable contribution to understanding how chromatin features relate to the transcriptome that we measure in a population of cells.

Time-resolved studies are key to understand the interplay between gene expression and chromatin

As described in the introductory section, there is evidence of multiple and sometimes contradictory ways in which expression and chromatin relate to each other. This is likely due to the limited number of available time-resolved studies that allow to assess how transcription and epigenetic features change in response to stimuli or developmental cues. Because of this, in the first chapter of the thesis we take advantage of a rich data resource generated in our lab during the transdifferentiation of pre-B cells into macrophages. This trans-

differentiation model is characterized by widespread transcriptomic changes (Rapino et al., 2013), which make it a suitable system to investigate how closely gene expression is coupled with epigenetic features.

Before commenting upon some of the findings of this work which, for space constraints, were not deeply addressed in the discussion of the manuscript, I would like to spend a few words on the methodology that we have applied to generate matrices of histone modifications' signals. This has been a challenging part of the study, mainly because of the lack of a well established protocol to assign signals of histone modifications to a gene, as it instead exists for gene expression. What hinders the task is the fact that the distribution of the signal around the promoter or over the gene body varies among histone marks and sometimes also across genes, making it difficult to analyze maps of several histone modifications in a systematic and homogeneous way across multiple conditions. We hope that the methodology that we employ – which browses through all annotated TSSs of a gene, combines discrete regions (peaks) with continuous pile-up signals, and applies a joint normalization on time-points and replicates – may foster the debate about the topic and help the epigenomics field establish a consensus pipeline.

Our analyses first attempted to solve the gap between steady-state and time-series studies. In fact, at individual time-points we observed strong, genome-wide correlated patterns between expression and canonical active histone marks, while at the level of individual genes the degree of correlation between expression and a given mark is substantially lower. Because time-course correlations cannot be computed for silent genes, we reasoned that these and the genes stably expressed during transdifferentiation, which tend to be constitutively void of histone modifications and marked, respectively (Pervouchine et al., 2015), may account for this discrepancy, as well as for the high steady-state correlations obtained upon randomization of ChIP-seq data. Indeed, after removing these two sets of genes we found that the steady-state and time-course correlation measurements give more consistent results, suggesting that correlation analyses can lead to very different results depending on the subsets of genes considered (i.e. silent, stably expressed or differentially expressed – DE). This first set of analyses also uncovered two unexpected findings. First, while weakly negative in steady-state conditions (which include silent genes), the degree of correlation be-

tween expression and H3K9me3 becomes positive when measured over time and taking into account expressed genes. Until very recently, there was only one evidence (in cancer cells) of deposition of H3K9me3 at promoter regions associated with active gene expression (Wiencke et al., 2008). This year, one paper reported that this modification at promoters is actually compatible with gene expression in early mouse development, while associated with repressed expression at later stages of development (Burton et al., 2020). This indeed supports our results of a dual behavior of H3K9me3. Second, of all canonically active histone marks, H4K20me1 is perhaps the least profiled one in epigenomic studies, and it is not included in the set of histone modifications recommended by the International Human Epigenome Consortium (IHEC). In our analyses, a lower proportion of expressed genes are actually marked by H4K20me1 (83% *vs* 99% averaged over the other six active modifications), but we indeed observed that H4K20me1 is the mark characterized by the highest correlation with gene expression over time, which suggests that it may be a candidate epigenetic feature to monitor in other experiments.

A novelty of this study is that, for the first time to our knowledge, the epigenomic data have been segmented longitudinally over the time of the process, and not along the genome sequence as previously done (Ernst and Kellis, 2012, Hoffman et al., 2012, Song and Chen, 2015, Zhang et al., 2016, Zhang and Hardison, 2017). Indeed, the sequence of time-points represents a natural way to accommodate the transitions modeled by HMMs, and at the same time allows to summarize the combinations of histone marks found both around the promoter and over the body of the gene. The chromatin states that we have identified represent the limited number of histone modifications that characterize human genes during transdifferentiation. An interesting observation is that the most dramatic changes in expression are observed for genes transitioning from low to active marking states, while genes displaying incremental deposition of histone marks (i.e. switch from active to strong marking state) are more numerous but show comparatively lower fold-changes. This provided a first hint that the mechanisms behind histone marks’ deposition may vary depending on the level of activation of the gene. Another less disclosed observation here is that a number of stably expressed and silent genes actually change chromatin state during the process, a result that was further supported by the analysis of individual marks. The magnitude of these changes is, in some cases, comparable to,

or even larger than the one observed for DE genes. While in the case of stably expressed genes we observed mostly changes in lysine acetylation, silent genes mainly show variations in H3K4me1 and H3K4me2. GO term analysis (data not shown for space constraints) revealed that these genes are enriched in functions related to development and differentiation, and further analysis is required to rule out that their promoters are involved in distal enhancer activity.

HMMs are suitable to summarize the epigenetic status of genes, and in our case suggested a rather coordinated mechanism of chromatin marking that limits all possible combinations of histone marks ($2^9 = 512$) to a few ones. Nevertheless, these combinations and the signal of the corresponding marks are an averaged measure across all genes. Therefore, they do not inform as to the contributions of each histone modification to the state transitions observed in individual genes, which are characterized by distinct levels and moments of activation, and perform different functions. For this reason, and motivated by the need to relate changes in chromatin to changes in expression, we clustered the genes based on their chromatin marking status and its relationship with gene expression, taking into account the behaviors of individual marks. This strategy highlighted three main modes of association of chromatin marks and gene expression, which actually resume the different and sometimes contradictory observations reported in the literature: besides the expected positive correlation, we report changes in gene expression either uncoupled from or in the absence of marking. One of the key points of this study is that for the first time we are able to interpret these distinct modes of regulation in a general framework that relates the expression life and function of a gene to its chromatin status. In this sense, housekeeping and cell type- (in our case macrophage-) specific genes, characterized by uncoupled, and absent or correlated marking, respectively, represent the two end points of this timeline. At the initial stage of activation, genes are characterized by a strong association between expression and histone marks, which is lost as we monitor genes in more advanced activation stages. These latter genes still undergo dramatic changes in expression but keep a rather stable epigenetic status, that we speculatively attribute to a saturation of modified lysine residues. Within this general framework we also assess the order in which histone marks are deposited with respect to expression. In doing so, we demonstrate that the strong correlation at the initial stage of activation is not due to histone modifications being instructive for transcription. In fact, with the exception

of H3K4me1 and H3K4me2, all other marks increase concomitantly or more frequently after gene expression. Thus, we basically recapitulate in the same cellular model interspersed findings about the lack of causality of some histone marks (Dorighi et al., 2017, Douillet et al., 2020, Rickels et al., 2017, Zhang et al., 2020), while extending these findings also to four additional modifications. This order of events in chromatin marking also sheds some light on the lately deposited gene body marks, in particular H4K20me1. Furthermore, we observe that at later stages of the up-regulation process, increases in expression and histone marks are more frequently concomitant. While it is tempting to think of a progressive synchronization of these two processes over time, we also acknowledge that the last points monitored during transdifferentiation are more distant, and therefore we cannot rule out a lack of time resolution.

Tissue-specific regulatory activity accumulates in introns

In the second chapter, we leveraged the ENCODE registry of cCREs (Abascal et al., 2020) to investigate the genomic location of distal regulatory elements in adult and developmental human samples. We initially observed that highly shared cCREs that display enhancer-like signatures (ELSs) are more frequently located in intergenic regions, compared to ELSs active in a smaller number of samples. We thus set out to identify groups of tissue-specific and common ELSs, by focusing on a curated subset of samples (iPSCs, fibro/myoblasts, blood, muscle and brain cell types) that form discrete multidimensional and hierarchical clusters. We validated our initial observation, by showing that common and tissue-specific ELSs are more frequently found in intergenic and intronic sequences, respectively. Remarkably, we reported that the proportion of intronic ELSs specific to a given cluster of samples correlates with the level of specialization of the corresponding tissue. This suggests that less differentiated or specialized cell types show a more balanced distribution of regulatory elements between intergenic and intronic regions.

By integrating these analyses with the GTEx (Aguet et al., 2017) catalog of expression QTLs (eQTLs), we made two key findings. First, up to roughly 20% of tissue-specific intronic enhancers contain eQTLs detected in the same tissue (eQTL-ELSs). Second, espe-

cially in brain and muscle, these eQTLs are associated with genes that perform tissue-specific functions, which is not the case for the genes targeted by intergenic eQTL-ELSs. Nevertheless, only in approximately 50% of the cases the target genes are the same that host the tissue-specific eQTL-ELSs. This piece of information brings in the hypothesis that the tissue-specificity of intronic regulatory activity is not only required to target the expression of the host gene. Moreover, while host genes perform tissue-specific functions, non-host genes are instead associated with mechanisms implicated in the homeostasis of the tissue, such as alternative splicing in neural cell types (Vuong et al., 2016). One property that characterizes ESCs is indeed their more accessible epigenetic landscape, which is progressively lost during differentiation. In this perspective, and because no differences in TF binding sites are observed between tissue-specific intergenic and intronic enhancers, we can speculate that intronic regulatory activity is an advantageous mechanism developed by specialized tissues. In other words, in differentiated tissues the regulatory activity may concentrate in a limited number of chromatin hubs, which can overlap with introns of key tissue-specific genes and orchestrate the expression of regulators of tissue homeostasis via long-range interactions. This strategy may discourage, on one side, investing an unnecessary amount of energy in ubiquitous chromatin remodeling, and on the other side avoid leaky expression in tissues in which a core set of tissue-specific genes is not expressed. In line with these findings, we observe a similar trend of intronic ELSs specific to more differentiated embryonic tissues, compared to enhancers shared among embryonic samples or specific to ESCs.

The timing of splicing is tightly regulated across cell states

In Chapter 3, we analyzed ENCODE (Davis et al., 2018, Dunham et al., 2012) fractional RNA-seq, eCLIP and ChIP-seq datasets to explore the regulation of splicing time across a panel of 13 cell lines. More precisely, we took advantage of the spatial constraint between nucleus and cytosol as a proxy to study differences in splicing completion for more than 70,000 introns annotated in the human genome. Thus, we classified introns expressed in each cell line as co- or post-transcriptionally spliced (cs and ps, respectively), unprocessed (i.e.

retained) or artifactually spliced (i.e. spliced at higher rates in the nucleus than in the cytosol).

Although widespread co-transcriptional splicing has been previously reported, one of our findings is that there is strong variation in the timing of splicing among cell lines, which range from predominant (MCF-7: 59%) to limited (SK-MEL-5, SK-N-DZ: 15%) post-transcriptional splicing. One could argue that the multiple testing correction burden may affect the detection of introns differentially spliced between the two compartments. In other words, cell lines characterized by a higher amount of expressed introns would show higher frequency of co-transcriptional splicing, and the other way around. Nevertheless, we observed negative correlation between the proportion of co-transcriptionally spliced introns and the number of expressed introns (data not shown), which rules out technical artifacts. Instead, these differences in splicing time seem to moderately reflect the germ layer of origin of cells. For instance, mesoderm-derived cell lines – such as GM12878, K562 and endothelial cells – show a prevalence of post-transcriptional splicing, whereas ectoderm-derived cell lines – like keratinocytes, SK-MEL-5 and SK-N-DZ – are characterized by the highest rates of co-transcriptional splicing. IMR-90, A549 and HepG2, which are of endodermal origin, display intermediate profiles. Nevertheless, exceptions to this pattern are represented by MCF-7 and SK-N-SH (ectodermal) and HT1080 (mesodermal).

We reported a number of features that distinguish ps introns from cs introns, including smaller intron and transcript size, higher transcript expression and GC content, and weaker acceptor splice sites. We also found that genes containing ps introns are often involved in RNA processing, in line with previously reported mechanisms of splicing autoregulation (Pervouchine et al., 2019). Of note, the expression of genes associated with functions related to amino-acid transport positively correlates with the rate of co-transcriptional splicing across cell lines. In addition, introns of protein-coding genes switch from co-transcriptional to post-transcriptional splicing/intron retention more frequently than those belonging to non-coding genes. Overall, we may speculate that some feedback mechanisms related to translation could regulate the timing of splicing events.

A growing body of evidence points to a role of RNA-binding proteins in splicing and other RNA processing events (Van Nostrand et al., 2020). Indeed, we found that RBPs’ binding patterns markedly differ

between introns and their flanking exons, and that a number of RBPs involved in RNA processing functions other than splicing bind to both ps and unprocessed introns. On the other hand, four members of the spliceosome machinery (RBM22, SF3B4, AQR and PRPF8) were found enriched exclusively in ps introns. This is somehow expected, given that ps introns are retained longer within the primary transcript (Garrido-Martín et al., 2020, in press). Thus, one could conclude that the lack of RBPs peaks on cs introns is mostly due to their quick excision from the transcript. Nonetheless, we have observed enriched binding of these factors to ps introns even when selecting subsets of ps and cs introns with comparable RNA-seq coverage in the nuclear fraction (data not shown). This suggests that the increased binding frequency of these four factors – which are involved in the formation of subsequent splicing complexes – may not merely be the consequence of a protracted presence of ps introns within the transcript. Instead, their longer residence time on ps introns could be interpreted as the cause of a delay in splicing completion.

Our sets of ps introns are, by construction, retained for a longer time within the transcript, most probably after the polyadenylation step, given that we are using polyA+ RNA-seq experiments. This is actually in agreement with studies of splicing and 3’ end cleavage and polyadenylation kinetics, which highlight the presence of a population of not-fully spliced but already 3’ end mature transcripts in the fraction of RNA attached to chromatin (Bhatt et al., 2012, Pandya-Jones et al., 2013). Thus, our set of ps introns belong to transcripts, many of which are likely to be still attached to chromatin. This motivated us to investigate whether specific epigenetic features, such as TFs, chromatin modifiers and histone marks, are associated with differential timing of splicing. We found that a few histone modifications are more abundant within introns than exons (H3K79me2, H3K4me1, H4K20me1), and in some cell lines are associated with differences in splicing time (H3K36me3 enriched in ps introns) and efficiency (H3K4me1 and H3K36me3 enriched in unprocessed and spliced introns, respectively). We are currently validating the importance of these features and of the above-mentioned RBPs with machine learning classifiers trained to discriminate between co-transcriptional and post-transcriptional splicing.

CONCLUSIONS

The work presented in this thesis investigates the role of epigenetic marks in transcriptional regulation, by analyzing the interplay between gene expression and histone modifications over time and across different cellular conditions.

Here is a summary of the main contributions of this thesis:

- We have monitored the transcriptome by RNA-seq, and the epigenome by ChIP-seq of nine histone modifications, at twelve time points during the induced transdifferentiation of human pre-B cells into macrophages.
- Analysis of these data reveals that:
  - A limited number of combinations of histone modifications define the major chromatin states marking human genes.
  - Genes tend to remain in the same chromatin state throughout transdifferentiation, irrespective of changes in gene expression.
  - A substantial amount of chromatin changes are not accompanied by changes in gene expression, and therefore the contribution of epigenetic modifications to cell state transition cannot be fully recapitulated by transcriptomic profiles.
  - There is a positive association between gene expression and histone marks only at the time of initial gene activation.
  - At this time, gene activation is preceded by deposition of H3K4me1 and H3K4me2, and followed by other canonically active histone modifications.
  - Subsequent changes in gene expression, comparable or even stronger than those at initial gene activation, are uncoupled from chromatin changes.
- We have analyzed the location of distal regulatory elements in the human genome, and its relationship with widespread and tissue-specific gene expression patterns.

- Specifically, we report that:
  - Highly shared regulatory elements are mostly intergenic, while those specific to a given tissue tend to accumulate in introns.
  - Intronic regulatory elements target genes involved in tissue-specific functions and homeostasis.
  - The prevalence of intronic regulatory elements correlates with the degree of specialization of the tissue.
- We have analyzed the timing of splicing across a panel of human cell lines, and investigated patterns of RNA-binding proteins and epigenetic features related to this phenomenon.
- Specifically, we report that:
  - The proportion of introns undergoing post-transcriptional splicing dramatically varies across cellular conditions.
  - There are sets of introns characterized by constrained timing of splicing (either co- or post-transcriptional) across multiple conditions.
  - There is a subset of introns that switch from co-transcriptional to post-transcriptional splicing, and more often belong to protein-coding genes.
  - Components of the spliceosome machinery selectively bind to post-transcriptionally spliced introns.

BIBLIOGRAPHY

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., Akiyama, J. A., Jammal, O. A., Amrhein, H., Anderson, S. M., Andrews, G. R., Antoshechkin, I., Ardlie, K. G., Armstrong, J., Astley, M., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710.

Adelman, K. and Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature reviews. Genetics*, 13(10):720–731.

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y. S., Parsana, P., Segrè, A. V., Strober, B. J., Zappala, Z., Cummings, B. B., Gelfand, E. T., Hadley, K., et al. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

Alabert, C., Barth, T. K., Reverón-Gómez, N., Sidoli, S., Schmidt, A., Jensen, O. N., Imhof, A., and Groth, A. (2015). Two distinct modes for propagation of histone PTMs across the cell cycle. *Genes and Development*, 29(6):585–590.

Alabert, C. and Groth, A. (2012). Chromatin replication and epigenome maintenance. *Nature Reviews Molecular Cell Biology*, 13(3):153–167.

Almouzni, G. and Cedar, H. (2016). Maintenance of epigenetic information. *Cold Spring Harbor Perspectives in Biology*, 8(5).

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., and Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural and Molecular Biology*, 18(12):1435–1440.

Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C., and Komorowski, J. (2009). Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Research*, 19(10):1732–1741.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., et al. (2014a). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Andersson, R., Refsing Andersen, P., Valen, E., Core, L. J., Bornholdt, J., Boyd, M., Heick Jensen, T., and Sandelin, A. (2014b). Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nature Communications*, 5.

Andersson, R. and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, 21(2):71–87.

Andersson, R., Sandelin, A., and Danko, C. G. (2015). A unified architecture of transcriptional regulatory elements. *Trends in Genetics*, 31(8):426–433.

Annunziato, A. T. (2015). The fork in the road: Histone partitioning during DNA replication. *Genes*, 6(2):353–371.

Avery, O. T., Macleod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of Experimental Medicine*, 79(2):137–158.

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H. F., John, R. M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., and Fisher, A. G. (2006). Chromatin signatures of pluripotent cell lines. *Nature Cell Biology*, 8(5):532–538.

Bannister, A. J., Schneider, R., Myers, F. A., Thorne, A. W., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *Journal of Biological Chemistry*, 280(18):17732–17736.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837.

Barth, T. K. and Imhof, A. (2010). Fast signals and slow marks: the dynamics of histone modifications. *Trends in Biochemical Sciences*, 35(11):618–626.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S. C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell*, 143(3):470–484.

Batut, P., Dobin, A., Plessy, C., Carninci, P., and Gingeras, T. R. (2013). High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Research*, 23(1):169–180.

Beck, D. B., Oda, H., Shen, S. S., and Reinberg, D. (2012). PR-set7 and H4K20me1: At the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. *Genes and Development*, 26(4):325–337.

Beltran, S., Angulo, M., Pignatelli, M., Serras, F., and Corominas, M. (2007). Functional dissection of the ash2 and ash1 transcriptomes provides insights into the transcriptional basis of wing phenotypes and reveals conserved protein interactions. *Genome Biology*, 8(4).

Beltran, S., Blanco, E., Serras, F., Pérez-Villamil, B., Guigó, R., Artavanis-Tsakonas, S., and Corominas, M. (2003). Transcriptional network controlled by the trithorax-group gene ash2 in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6):3293–3298.

Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P., Liu, J. S., Kouzarides, T., and Schreiber, S. L. (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13):8695–8700.

Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., Gingeras, T. R., Schreiber, S. L., and Lander, E. S. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–181.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., and Lander, E. S. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2):315–326.

Bhatt, D. M., Pandya-Jones, A., Tong, A. J., Barozzi, I., Lissner, M. M., Natoli, G., Black, D. L., and Smale, S. T. (2012). Tran-

script dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*, 150(2):279–290.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Blackledge, N. P., Farcas, A. M., Kondo, T., King, H. W., McGouran, J. F., Hanssen, L. L., Ito, S., Cooper, S., Kondo, K., Koseki, Y., Ishikura, T., Long, H. K., Sheahan, T. W., Brockdorff, N., Kessler, B. M., et al. (2014). Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. *Cell*, 157(6):1445–1459.

Blackledge, N. P., Fursova, N. A., Kelley, J. R., Huseyin, M. K., Feldmann, A., and Klose, R. J. (2020). PRC1 Catalytic Activity Is Central to Polycomb System Function. *Molecular Cell*, 77(4):857–874.

Blahnik, K. R., Dou, L., Echipare, L., Iyengar, S., O'Geen, H., Sanchez, E., Zhao, Y., Marra, M. A., Hirst, M., Costello, J. F., Korf, I., and Farnham, P. J. (2011). Characterization of the Contradictory Chromatin Signatures at the 3' exons of Zinc finger genes. *PLoS ONE*, 6(2).

Blanco, E., González-Ramírez, M., Alcaine-Colet, A., Aranda, S., and Di Croce, L. (2020). The Bivalent Genome: Characterization, Structure, and Regulation. *Trends in Genetics*, 36(2):118–131.

Bonev, B. and Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, 17(11):661–678.

Borde, V., Robine, N., Lin, W., Bonfils, S., Géli, V., and Nicolas, A. (2009). Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO Journal*, 28(2):99–111.

Boutz, P. L., Bhutkar, A., and Sharp, P. A. (2015). Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes and Development*, 29(1):63–80.

Boveri, T. (1904). *Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns.* G. Fischer, Jena.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Names, A., Temper, V., Razin, A., and Cedar, H. (1994). Spl elements protect a CpG island from de novo methylation. *Nature*, 371(6496):435–438.

Braunschweig, U., Barbosa-Morais, N. L., Pan, Q., Nachman, E. N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., and Blencowe, B. J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*, 24(11):1774–1786.

Braunstein, M., Rose, A. B., Holmes, S. G., David Allis, C., and Broach, J. R. (1993). Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. *Genes and Development*, 7(4):592–604.

Briggs, R. and King, T. J. (1952). Transplantation of living nuclei from blastula cells into enucleated frogs’ eggs. *Proceedings of the National Academy of Sciences of the United States of America*, 38(5):455–463.

Brinkman, A. B., Gu, H., Bartels, S. J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., and Stunnenberg, H. G. (2012). Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Research*, 22(6):1128–1138.

Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S. L., Van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, 150(6):1209–1222.

Bungard, D., Fuerth, B. J., Zeng, P. Y., Faubert, B., Maas, N. L., Viollet, B., Carling, D., Thompson, C. B., Jones, R. G., and Berger, S. L. (2010). Signaling kinase AMPK activates stress-promoted transcription via histone H2B phosphorylation. *Science*, 329(5996):1201–1205.

Burton, A., Brochard, V., Galan, C., Ruiz-Morales, E. R., Rovira, Q., Rodriguez-Terrones, D., Kruse, K., Le Gras, S., Udayakumar, V. S., Chin, H. G., Eid, A., Liu, X., Wang, C., Gao, S., Pradhan, S., et al. (2020). Heterochromatin establishment during early mammalian

development is regulated by pericentromeric RNA and characterized by non-repressive H3K9me3. *Nature Cell Biology*, 22(7):767–778.

Cantone, I. and Fisher, A. G. (2013). Epigenetic programming and reprogramming during development. *Nature Structural and Molecular Biology*, 20(3):282–289.

Cao, K., Collings, C. K., Morgan, M. A., Marshall, S. A., Rendleman, E. J., Ozark, P. A., Smith, E. R., and Shilatifard, A. (2018). An Mll4/COMPASS-Lsd1 epigenetic axis governs enhancer function and pluripotency transition in embryonic stem cells. *Science Advances*, 4(1):eaap8747.

Cao, R., Tsukada, Y. I., and Zhang, Y. (2005). Role of Bmi-1 and Ring1A in H2A ubiquitylation and hox gene silencing. *Molecular Cell*, 20(6):845–854.

Cao, R. and Zhang, Y. (2004). The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Current Opinion in Genetics and Development*, 14(2):155–164.

Carlone, D. L., Lee, J.-H., Young, S. R. L., Dobrota, E., Butler, J. S., Ruiz, J., and Skalnik, D. G. (2005). Reduced Genomic Cytosine Methylation and Defective Cellular Differentiation in Embryonic Stem Cells Lacking CpG Binding Protein. *Molecular and Cellular Biology*, 25(12):4881–4891.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6):626–635.

Carrillo Oesterreich, F., Preibisch, S., and Neugebauer, K. M. (2010). Global analysis of nascent rna reveals transcriptional pausing in terminal exons. *Molecular Cell*, 40(4):571–581.

Carrozza, M. J., Li, B., Florens, L., Suganuma, T., Swanson, S. K., Lee, K. K., Shia, W. J., Anderson, S., Yates, J., Washburn, M. P., and Workman, J. L. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, 123(4):581–592.

Cenik, B. K. and Shilatifard, A. (2020). COMPASS and SWI/SNF complexes in development and disease. *Nature Reviews Genetics*, 22(1):38–58.

Chen, K., Johnston, J., Shao, W., Meier, S., Staber, C., and Zeitlinger, J. (2013). A global change in RNA polymerase II pausing during the Drosophila midblastula transition. *eLife*, 2.

Chen, T. and Dent, S. Y. (2014). Chromatin modifiers and remodellers: Regulators of cellular differentiation. *Nature Reviews Genetics*, 15(2):93–106.

Cheng, B., Li, T., Rahl, P. B., Adamson, T. E., Loudas, N. B., Guo, J., Varzavand, K., Cooper, J. J., Hu, X., Gnatt, A., Young, R. A., and Price, D. H. (2012). Functional association of gdown1 with RNA polymerase II poised on human genes. *Molecular Cell*, 45(1):38–50.

Cheng, J., Blum, R., Bowman, C., Hu, D., Shilatifard, A., Shen, S., and Dynlacht, B. D. (2014). A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. *Molecular Cell*, 53(6):979–992.

Chih, L. L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S. L., Friedman, N., and Rando, O. J. (2005). Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biology*, 3(10).

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J. H., Skalnik, D., and Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes and Development*, 26(15):1714–1728.

Conerly, M. L., Teves, S. S., Diolaiti, D., Ulrich, M., Eisenman, R. N., and Henikoff, S. (2010). Changes in H2A.Z occupancy and DNA methylation during B-cell lymphomagenesis. *Genome Research*, 20(10):1383–1390.

Cooper, S., Dienstbier, M., Hassan, R., Schermelleh, L., Sharif, J., Blackledge, N. P., DeMarco, V., Elderkin, S., Koseki, H., Klose, R., Heger, A., and Brockdorff, N. (2014). Targeting Polycomb to Pericentric Heterochromatin in Embryonic Stem Cells Reveals a Role for H2AK119u1 in PRC2 Recruitment. *Cell Reports*, 7(5):1456–1470.

Core, L. and Adelman, K. (2019). Promoter-proximal pausing of RNA polymerase II: A nexus of gene regulation. *Genes and Development*, 33(15-16):960–982.

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., and Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, 46(12):1311–1320.

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21931–21936.

Cui, K., Zang, C., Roh, T. Y., Schones, D. E., Childs, R. W., Peng, W., and Zhao, K. (2009). Chromatin Signatures in Multipotent Human Hematopoietic Stem Cells Indicate the Fate of Bivalent Genes during Differentiation. *Cell Stem Cell*, 4(1):80–93.

Curado, J., Iannone, C., Tilgner, H., Valcárcel, J., and Guigó, R. (2015). Promoter-like epigenetic signatures in exons displaying cell type-specific splicing. *Genome Biology*, 16(1).

Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R. M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H. A., Christiansen, L., Qiu, X., Steemers, F. J., Trapnell, C., Shendure, J., and Furlong, E. E. (2018). The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, 555(7697):538–542.

Dao, L. T., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., Alomairi, J., Martin, D., Torres, M., Fernandez, N., Soler, E., et al. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics*, 49(7):1073–1081.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan,

A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., et al. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, 46(D1):D794–D801.

De Almeida, S. F., Grosso, A. R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., Andrau, J. C., Ferrier, P., and Carmo-Fonseca, M. (2011). Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. *Nature Structural and Molecular Biology*, 18(9):977–983.

De La Mata, M., Alonso, C. R., Kadener, S., Fededa, J. P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., and Kornblihtt, A. R. (2003). A slow RNA polymerase II affects alternative splicing in vivo. *Molecular Cell*, 12(2):525–532.

de la Serna, I. L., Ohkawa, Y., Berkes, C. A., Bergstrom, D. A., Dacwag, C. S., Tapscott, S. J., and Imbalzano, A. N. (2005). MyoD Targets Chromatin Remodeling Complexes to the Myogenin Locus Prior to Forming a Stable DNA-Bound Complex. *Molecular and Cellular Biology*, 25(10):3997–4009.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biology*, 8(5):e1000384.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., Dean, A., and Blobel, G. A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–1244.

Deng, W., Rupon, J. W., Krivega, I., Breda, L., Motta, I., Jahn, K. S., Reik, A., Gregory, P. D., Rivella, S., Dean, A., and Blobel, G. A. (2014). Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell*, 158(4):849–860.

Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K. L., and Ren, B. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods*, 14(6):629–635.

Dorighi, K. M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B. S., Nady, N., Still, C. D., Garcia, B. A., Adelman, K., and Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Molecular Cell*, 66(4):568–576.

Douillet, D., Sze, C. C., Ryan, C., Piunti, A., Shah, A. P., Ugarenko, M., Marshall, S. A., Rendleman, E. J., Zha, D., Helmin, K. A., Zhao, Z., Cao, K., Morgan, M. A., Singer, B. D., Bartom, E. T., et al. (2020). Uncoupling histone H3K4 trimethylation from developmental gene expression via an equilibrium of COMPASS, Polycomb and DNA methylation. *Nature Genetics*, 52(6):615–625.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Elgin, S. C. and Reuter, G. (2013). Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. *Cold Spring Harbor Perspectives in Biology*, 5(8):a017780.

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., and Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, 539(7629):452–455.

Erhardt, S., Su, I. H., Schneider, R., Barton, S., Bannister, A. J., Perez-Burgos, L., Jenuwein, T., Kouzarides, T., Tarakhovsky, A., and Azim Surani, M. (2003). Consequences of the depletion of zygotic and embryonic enhancer of zeste 2 during preimplantation mouse development. *Development*, 130(18):4235–4248.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825.

Ernst, J. and Kellis, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping

and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49.

Evans, K. J., Huang, N., Stempor, P., Chesney, M. A., Down, T. A., and Ahringer, J. (2016). Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. *Proceedings of the National Academy of Sciences of the United States of America*, 113(45):E7020–E7029.

Felsenfeld, G. (2014). A brief history of epigenetics. *Cold Spring Harbor Perspectives in Biology*, 6(1):a018200.

Ferrai, C., Torlai Triglia, E., Risner-Janiczek, J. R., Rito, T., Rackham, O. J., Santiago, I., Kukalev, A., Nicodemi, M., Akalin, A., Li, M., Ungless, M. A., and Pombo, A. (2017). RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation. *Molecular Systems Biology*, 13(10):946.

Fisher, C. L. and Fisher, A. G. (2011). Chromatin states in pluripotent, differentiated, and reprogrammed cells. *Current Opinion in Genetics and Development*, 21(2):140–146.

Flemming Walther (1882). *Zellsubstanz, Kern und Zelltheilung*. F.C.W. Vogel, Leipzig.

Foulds, C. E., Feng, Q., Ding, C., Bailey, S., Hunsaker, T. L., Malovannaya, A., Hamilton, R. A., Gates, L. A., Zhang, Z., Li, C., Chan, D., Bajaj, A., Callaway, C. G., Edwards, D. P., Lonard, D. M., et al. (2013). Proteomic analysis of coregulators bound to ERα on DNA and Nucleosomes reveals coregulator dynamics. *Molecular Cell*, 51(2):185–199.

Fraser, N. W., Sehgal, P. B., and Darnell, J. E. (1978). DRB-induced premature termination of late adenovirus transcription. *Nature*, 272(5654):590–593.

Fueyo, R., Iacobucci, S., Pappa, S., Estarás, C., Lois, S., Vicioso-Mantis, M., Navarro, C., Cruz-Molina, S., Reyes, J. C., Rada-Iglesias, Ã., de la Cruz, X., and Martínez-Balbás, M. A. (2018). Lineage specific transcription factors and epigenetic regulators mediate TGFβ-dependent enhancer activation. *Nucleic acids research*, 46(7):3351–3365.

Gariglio, P., Bellard, M., and Chambon, P. (1981). Clustering of RNA polymerase B molecules in the 5' moiety of the adult β-globin gene of hen erythrocytes. *Nucleic Acids Research*, 9(11):2589–2598.

Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., and Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1-2):377–390.

Gates, L. A., Shi, J., Rohira, A. D., Feng, Q., Zhu, B., Bedford, M. T., Sagum, C. A., Jung, S. Y., Qin, J., Tsai, M. J., Tsai, S. Y., Li, W., Foulds, C. E., and O'Malley, B. W. (2017). Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *Journal of Biological Chemistry*, 292(35):14456–14472.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., et al. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, 330(6012):1775–1787.

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E. E. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 512(1):96–100.

Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., Gregory, L., Lonie, L., Chew, A., Wei, C. L., Ragoussis, J., and Natoli, G. (2010). Identification and Characterization of Enhancers Controlling the Inflammatory Gene Expression Program in Macrophages. *Immunity*, 32(3):317–328.

Gilchrist, D. A., Dos Santos, G., Fargo, D. C., Xie, B., Gao, Y., Li, L., and Adelman, K. (2010). Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, 143(4):540–551.

Gilchrist, D. A., Fromm, G., dos Santos, G., Pham, L. N., Mcdaniel, I. E., Burkholder, A., Fargo, D. C., and Adelman, K. (2012). Regulating the regulators: The pervasive effects of Pol II pausing on stimulus-responsive gene networks. *Genes and Development*, 26(9):933–944.

Gilchrist, D. A., Nechaev, S., Lee, C., Ghosh, S. K. B., Collins, J. B., Li, L., Gilmour, D. S., and Adelman, K. (2008). NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes and Development*, 22(14):1921–1933.

Girard, C., Will, C. L., Peng, J., Makarov, E. M., Kastner, B., Lemm, I., Urlaub, H., Hartmuth, K., and Luhrmann, R. (2012). Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nature Communications*, 3.

Gonzalez-Sandoval, A., Towbin, B. D., Kalck, V., Cabianca, D. S., Gaidatzis, D., Hauer, M. H., Geng, L., Wang, L., Yang, T., Wang, X., Zhao, K., and Gasser, S. M. (2015). Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in C. elegans Embryos. *Cell*, 163(6):1333–1347.

Greer, C. B., Tanaka, Y., Kim, Y. J., Xie, P., Zhang, M. Q., Park, I. H., and Kim, T. H. (2015). Histone Deacetylases Positively Regulate Transcription through the Elongation Machinery. *Cell Reports*, 13(7):1444–1455.

Gromak, N., West, S., and Proudfoot, N. J. (2006). Pause Sites Promote Transcriptional Termination of Mammalian RNA Polymerase II. *Molecular and Cellular Biology*, 26(10):3986–3996.

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., Greenside, P., Srivas, R., Phanstiel, D. H., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J. K., Bustamante, C. D., et al. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–1065.

Gruss1, C., Wu, J., Koller, T., and Sogo2, J. M. (1993). Disruption of the nucleosomes at the replication fork. *The EMBO Journal*, 12(12):4533–4545.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell*, 130(1):77–88.

Hajkova, P., Ancelin, K., Waldmann, T., Lacoste, N., Lange, U. C., Cesari, F., Lee, C., Almouzni, G., Schneider, R., and Surani, M. A. (2008). Chromatin dynamics during epigenetic reprogramming in the mouse germ line. *Nature*, 452(7189):877–881.

Halfon, M. S., Carmena, A., Gisselbrecht, S., Sackerson, C. M., Jimé Nez, F., Baylies, M. K., and Michelson, A. M. (2000). Ras Pathway Specificity Is Determined by the Integration of Multiple Signal-Activated and Tissue-Restricted Transcription Factors. *Cell*, 103(1):63–74.

Hanna, J., Saha, K., Pando, B., Van Zon, J., Lengner, C. J., Creyghton, M. P., Van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature*, 462(7273):595–601.

Hannah, A. (1951). Localization and Function of Heterochromatin in Drosophila Melanogaster. *Advances in Genetics*, 4(C):87–125.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318.

Henikoff, S. and Shilatifard, A. (2011). Histone modification: Cause or cog? *Trends in Genetics*, 27(10):389–396.

Hezroni, H., Tzchori, I., Davidi, A., Mattout, A., Biran, A., Nissim-Rafinia, M., Westphal, H., and Meshorer, E. (2011). H3K9 histone acetylation predicts pluripotency and reprogramming capacity of ES cells. *Nucleus*, 2(4):300–309.

Hödl, M. and Basler, K. (2012). Transcription in the absence of histone H3.2 and H3K4 methylation. *Current Biology*, 22(23):2253–2257.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476.

Hon, G., Wang, W., and Ren, B. (2009). Discovery and Annotation of Functional Chromatin Signatures in the Human Genome. *PLoS Computational Biology*, 5(11):e1000566.

Hoogenkamp, M., Lichtinger, M., Krysinska, H., Lancrin, C., Clarke, D., Williamson, A., Mazzarella, L., Ingram, R., Jorgensen, H., Fisher, A., Tenen, D. G., Kouskoff, V., Lacaud, G., and Bonifer, C. (2009). Early chromatin unfolding by RUNX1: a molecular explanation for differential requirements during specification versus maintenance of the hematopoietic gene expression program. *Blood*, 114(2):299–309.

Houston, S. I., McManus, K. J., Adams, M. M., Sims, J. K., Carpenter, P. B., Hendzel, M. J., and Rice, J. C. (2008). Catalytic function of the PR-Set7 histone H4 lysine 20 monomethyltransferase is essential for mitotic entry and genomic stability. *Journal of Biological Chemistry*, 283(28):19478–19488.

Howe, F. S., Fischl, H., Murray, S. C., and Mellor, J. (2017). Is H3K4me3 instructive for transcription activation? *BioEssays*, 39(1):1–12.

Hu, D., Gao, X., Cao, K., Morgan, M. A., Mas, G., Smith, E. R., Volk, A. G., Bartom, E. T., Crispino, J. D., Di Croce, L., and Shilatifard, A. (2017). Not All H3K4 Methylations Are Created Equal: Mll2/COMPASS Dependency in Primordial Germ Cell Specification. *Molecular Cell*, 65(3):460–475.

Inoue, A. and Zhang, Y. (2011). Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science*, 334(6053):194–194.

Jacob, E., Hod-Dvorai, R., Schif-Zuck, S., and Avni, O. (2008). Unconventional association of the polycomb group proteins with cytokine genes in differentiated T helper cells. *Journal of Biological Chemistry*, 283(19):13471–13481.

Jakobsen, J. S., Braun, M., Astorga, J., Gustafson, E. H., Sandmann, T., Karzynski, M., Carlsson, P., and Furlong, E. E. (2007). Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. *Genes and Development*, 21(19):2448–2460.

Javierre, B. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., Freire-Pritchett, P., Spivakov, M., Fraser, P., Burren,

O. S., Cutler, A. J., Todd, J. A., Wallace, C., Wilder, S. P., Kreuzhuber, R., et al. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5):1369–1384.

Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492.

Jørgensen, S., Schotta, G., and Sørensen, C. S. (2013). Histone H4 Lysine 20 methylation: Key player in epigenetic regulation of genomic integrity. *Nucleic Acids Research*, 41(5):2797–2806.

Kaikkonen, M. U., Spann, N. J., Heinz, S., Romanoski, C. E., Allison, K. A., Stender, J. D., Chun, H. B., Tough, D. F., Prinjha, R. K., Benner, C., and Glass, C. K. (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular Cell*, 51(3):310–325.

Kang, H., Shokhirev, M. N., Xu, Z., Chandran, S., Dixon, J. R., and Hetzer, M. W. (2020). Dynamic regulation of histone modifications and long-range chromosomal interactions during postmitotic transcriptional reactivation. *Genes and Development*, 34(13-14):1–18.

Karmodiya, K., Krebs, A. R., Oulad-Abdelghani, M., Kimura, H., and Tora, L. (2012). H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics*, 13(1).

Keogh, M. C., Kurdistani, S. K., Morris, S. A., Ahn, S. H., Podolny, V., Collins, S. R., Schuldiner, M., Chin, K., Punna, T., Thompson, N. J., Boone, C., Emili, A., Weissman, J. S., Hughes, T. R., Strahl, B. D., et al. (2005). Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell*, 123(4):593–605.

Kephart, D. D., Marshall, N. F., and Price, D. H. (1992). Stability of Drosophila RNA polymerase II elongation complexes in vitro. *Molecular and Cellular Biology*, 12(5):2067–2077.

Khodor, Y. L., Menet, J. S., Tolan, M., and Rosbash, M. (2012). Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. *RNA*, 18(12):2174–2186.

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187.

Kim, Y. J., Greer, C. B., Cecchini, K. R., Harris, L. N., Tuck, D. P., and Kim, T. H. (2013). HDAC inhibitors induce transcriptional repression of high copy number genes in breast cancer through elongation blockade. *Oncogene*, 32(23):2828–2835.

Kirmizis, A., Santos-Rosa, H., Penkett, C. J., Singer, M. A., Vermeulen, M., Mann, M., Bähler, J., Green, R. D., and Kouzarides, T. (2007). Arginine methylation at histone H3R2 controls deposition of H3K4 trimethylation. *Nature*, 449(7164):928–932.

Klemm, S. L., Shipony, Z., and Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220.

Koche, R. P., Smith, Z. D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B. E., and Meissner, A. (2011). Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell Stem Cell*, 8(1):96–105.

Kohlmaier, A., Savarese, F., Lachner, M., Martens, J., Jenuwein, T., and Wutz, A. (2004). A chromosomal memory triggered by Xist regulates histone methylation in X inactivation. *PLoS Biology*, 2(7).

Koike, N., Yoo, S. H., Huang, H. C., Kumar, V., Lee, C., Kim, T. K., and Takahashi, J. S. (2012). Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science*, 338(6105):349–354.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genetics*, 41(3):376–381.

Krejčí, J., Uhlířová, R., Galiová, G., Kozubek, S., Šmigová, J., and Bártová, E. (2009). Genome-wide reduction in H3K9 acetylation during human embryonic stem cell differentiation. *Journal of Cellular Physiology*, 219(3):677–687.

Krumm, A., Meulia, T., Brunvand, M., and Groudine, M. (1992). The block to transcriptional elongation within the human c-myc gene is

determined in the promoter-proximal region. *Genes and Development*, 6(11):2201–2213.

Krysinska, H., Hoogenkamp, M., Ingram, R., Wilson, N., Tagoh, H., Laslo, P., Singh, H., and Bonifer, C. (2007). A Two-Step, PU.1-Dependent Mechanism for Developmentally Regulated Chromatin Remodeling and Transcription of the c-fms Gene. *Molecular and Cellular Biology*, 27(3):878–887.

Kuang, Z., Cai, L., Zhang, X., Ji, H., Tu, B. P., and Boeke, J. D. (2014). High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast. *Nature Structural and Molecular Biology*, 21(10):854–863.

Landry, J. R., Mager, D. L., and Wilhelm, B. T. (2003). Complex controls: The role of alternative promoters in mammalian genomes. *Trends in Genetics*, 19(11):640–648.

Lane, N., Dean, W., Erhardt, S., Hajkova, P., Surani, A., Walter, J., and Reik, W. (2003). Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis*, 35(2):88–93.

Larschan, E., Alekseyenko, A. A., Gortchakov, A. A., Peng, S., Li, B., Yang, P., Workman, J. L., Park, P. J., and Kuroda, M. I. (2007). MSL Complex Is Attracted to Genes Marked by H3K36 Trimethylation Using a Sequence-Independent Mechanism. *Molecular Cell*, 28(1):121–133.

Larschan, E., Bishop, E. P., Kharchenko, P. V., Core, L. J., Lis, J. T., Park, P. J., and Kuroda, M. I. (2011). X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature*, 471(7336):115–118.

Laskey, R. A. and Gurdon, J. B. (1970). Genetic content of adult somatic cells tested by nuclear transplantation from Cultured Cells. *Nature*, 228(5278):1332–1334.

Lauberth, S. M., Nakayama, T., Wu, X., Ferris, A. L., Tang, Z., Hughes, S. H., and Roeder, R. G. (2013). H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell*, 152(5):1021–1036.

Lavarone, E., Barbieri, C. M., and Pasini, D. (2019). Dissecting the role of H3K27 acetylation and methylation in PRC2 mediated control of cellular identity. *Nature Communications*, 10.

Le Martelot, G., Canella, D., Symul, L., Migliavacca, E., Gilardi, F., Liechti, R., Martin, O., Harshman, K., Delorenzi, M., Desvergne, B., Herr, W., Deplancke, B., Schibler, U., Rougemont, J., Guex, N., et al. (2012). Genome-Wide RNA Polymerase II Profiles and RNA Accumulation Reveal Kinetics of Transcription and Associated Epigenetic Changes During Diurnal Cycles. *PLoS Biology*, 10(11).

Leatham-Jensen, M., Uyehara, C. M., Strahl, B. D., Matera, A. G., Duronio, R. J., and McKay, D. J. (2019). Lysine 27 of replication-independent histone H3.3 is required for Polycomb target gene silencing but not for gene activation. *PLoS Genetics*, 15(1).

Lenstra, T. L., Benschop, J. J., Kim, T. S., Schulze, J. M., Brabers, N. A., Margaritis, T., van de Pasch, L. A., van Heesch, S. A., Brok, M. O., Groot Koerkamp, M. J., Ko, C. W., van Leenen, D., Sameith, K., van Hooff, S. R., Lijnzaad, P., et al. (2011). The Specificity and Topology of Chromatin Interaction Pathways in Yeast. *Molecular Cell*, 42(4):536–549.

Lettice, L. A., Williamson, I., Wiltshire, J. H., Peluso, S., Devenney, P. S., Hill, A. E., Essafi, A., Hagman, J., Mort, R., Grimes, G., DeAngelis, C. L., and Hill, R. E. (2012). Opposing Functions of the ETS Factor Family Define Shh Spatial Expression in Limb Buds and Underlie Polydactyly. *Developmental Cell*, 22(2):459–467.

Levine, M. (2011). Paused RNA polymerase II as a developmental checkpoint. *Cell*, 145(4):502–511.

Li, X., Wang, X., He, K., Ma, Y., Su, N., He, H., Stolc, V., Tongprasit, W., Jin, W., Jiang, J., Terzaghi, W., Li, S., and Xing, W. D. (2008). High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell*, 20(2):259–276.

Liber, D., Domaschenz, R., Holmqvist, P. H., Mazzarella, L., Georgiou, A., Leleu, M., Fisher, A. G., Labosky, P. A., and Dillon, N. (2010). Epigenetic priming of a Pre-B Cell-Specific enhancer through binding of Sox2 and Foxd3 at the ESC stage. *Cell Stem Cell*, 7(1):114–126.

Local, A., Huang, H., Albuquerque, C. P., Singh, N., Lee, A. Y., Wang, W., Wang, C., Hsia, J. E., Shiau, A. K., Ge, K., Corbett, K. D., Wang, D., Zhou, H., and Ren, B. (2018). Identification of H3K4me1-associated proteins at mammalian enhancers. *Nature Genetics*, 50(1):73–82.

Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996–1000.

Lupien, M., Eeckhoute, J., Meyer, C. A., Wang, Q., Zhang, Y., Li, W., Carroll, J. S., Liu, X. S., and Brown, M. (2008). FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell*, 132(6):958–970.

Lv, X., Han, Z., Chen, H., Yang, B., Yang, X., Xia, Y., Pan, C., Fu, L., Zhang, S., Han, H., Wu, M., Zhou, Z., Zhang, L., Li, L., Wei, G., et al. (2016). A positive role for polycomb in transcriptional regulation via H4K20me1. *Cell Research*, 26(5):529–542.

Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., Law, T., Lareau, C., Hsu, Y.-C., Regev, A., and Buenrostro, J. D. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, 183(4):1103–1116.

Macleod, D., Charlton, J., Mullins, J., and Bird, A. P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes and Development*, 8(19):2282–2292.

Madamba, E. V., Berthet, E. B., and Francis, N. J. (2017). Inheritance of Histones H3 and H4 during DNA Replication In Vitro. *Cell Reports*, 21(5):1361–1374.

Mali, P., Chou, B. K., Yen, J., Ye, Z., Zou, J., Dowey, S., Brodsky, R. A., Ohm, J. E., Yu, W., Baylin, S. B., Yusa, K., Bradley, A., Meyers, D. J., Mukherjee, C., Cole, P. A., et al. (2010). Butyrate greatly enhances derivation of human induced pluripotent stem cells by promoting epigenetic remodeling and the expression of pluripotency-associated genes. *Stem Cells*, 28(4):713–720.

Margaritis, T., Oreal, V., Brabers, N., Maestroni, L., Vitaliano-Prunier, A., Benschop, J. J., van Hooff, S., van Leenen, D., Dargemont, C., Géli, V., and Holstege, F. C. (2012). Two Distinct Repressive Mechanisms for Histone 3 Lysine 4 Methylation through Promoting 3'-End Antisense Transcription. *PLoS Genetics*, 8(9).

Margueron, R., Justin, N., Ohno, K., Sharpe, M. L., Son, J., Drury, W. J., Voigt, P., Martin, S. R., Taylor, W. R., De Marco, V., Pirrotta, V., Reinberg, D., and Gamblin, S. J. (2009). Role of the polycomb

protein EED in the propagation of repressive histone marks. *Nature*, 461(7265):762–767.

Marshall, N. F. and Price, D. H. (1992). Control of formation of two distinct classes of RNA polymerase II elongation complexes. *Molecular and Cellular Biology*, 12(5):2078–2090.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. *Nature*, 403(6769):501–502.

McKay, D. J., Klusza, S., Penke, T. J., Meers, M. P., Curry, K. P., McDaniel, S. L., Malek, P. Y., Cooper, S. W., Tatomer, D. C., Lieb, J. D., Strahl, B. D., Duronio, R. J., and Matera, A. G. (2015). Interrogating the function of metazoan histones using engineered gene clusters. *Developmental Cell*, 32(3):373–386.

McManus, S., Ebert, A., Salvagiotto, G., Medvedovic, J., Sun, Q., Tamir, I., Jaritz, M., Tagoh, H., and Busslinger, M. (2011). The transcription factor Pax5 regulates its target genes by recruiting chromatin-modifying proteins in committed B cells. *EMBO Journal*, 30(12):2388–2404.

Meers, M. P., Henriques, T., Lavender, C. A., McKay, D. J., Strahl, B. D., Duronio, R. J., Adelman, K., and Matera, A. G. (2017). Histone gene replacement reveals a posttranscriptional role for H3K36 in maintaining metazoan transcriptome fidelity. *eLife*, 6.

Mercer, E. M., Lin, Y. C., Benner, C., Jhunjhunwala, S., Dutkowski, J., Flores, M., Sigvardsson, M., Ideker, T., Glass, C. K., and Murre, C. (2011). Multilineage Priming of Enhancer Repertoires Precedes Commitment to the B and Myeloid Cell Lineages in Hematopoietic Progenitors. *Immunity*, 35(3):413–425.

Mi, W., Guan, H., Lyu, J., Zhao, D., Xi, Y., Jiang, S., Andrews, F. H., Wang, X., Gagea, M., Wen, H., Tora, L., Dent, S. Y., Kutateladze, T. G., Li, W., Li, H., et al. (2017). YEATS2 links histone acetylation to tumorigenesis of non-small cell lung cancer. *Nature Communications*, 8(1).

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs, A., Leproust, E., Follows, G. A., et al. (2015). Mapping long-range promoter contacts

in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47(6):598–606.

Mikkelsen, T. S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B. E., Jaenisch, R., Lander, E. S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. *Nature*, 454(7200):49–55.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560.

Min, I. M., Waterfall, J. J., Core, L. J., Munroe, R. J., Schimenti, J., and Lis, J. T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes and Development*, 25(7):742–754.

Morgan, T. H. (1911). An attempt to analyze the constitution of the chromosomes on the basis of sex-limited inheritance in Drosophila. *Journal of Experimental Zoology*, 11(4):365–413.

Morillon, A., Karabetsou, N., Nair, A., and Mellor, J. (2005). Dynamic lysine methylation on histone H3 defines the regulatory phase of gene transcription. *Molecular Cell*, 18(6):723–734.

Mullen, A. C., Orlando, D. A., Newman, J. J., Lovén, J., Kumar, R. M., Bilodeau, S., Reddy, J., Guenther, M. G., Dekoter, R. P., and Young, R. A. (2011). Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell*, 147(3):565–576.

Muller, H. J. (1930). Types of visible variations induced by X-rays in Drosophila. *Journal of Genetics*, 22(3):299–334.

Nechaev, S. and Adelman, K. (2008). Promoter-proximal Pol II: When stalling speeds things up. *Cell Cycle*, 7(11):1539–1544.

Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., Kheradpour, P., Eaton, M. L., Loriaux, P., Sealfon, R., Li, Z., Ishii, H., Spokony, R. F., Chen, J., Hwang, L., et al. (2011). A cis-regulatory map of the Drosophila genome. *Nature*, 471(7339):527–531.

Neugebauer, K. M. (2019). Nascent RNA and the coordination of splicing with transcription. *Cold Spring Harbor Perspectives in Biology*, 11(8).

Ng, H. H., Robert, F., Young, R. A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Molecular Cell*, 11(3):709–719.

Ninova, M., Godneeva, B., Chen, Y. C. A., Luo, Y., Prakash, S. J., Jankovics, F., Erdélyi, M., Aravin, A. A., and Fejes Tóth, K. (2020). The SUMO Ligase Su(var)2-10 Controls Hetero- and Euchromatic Gene Expression via Establishing H3K9 Trimethylation and Negative Feedback Regulation. *Molecular Cell*, 77(3):571–585.

Ninova, M., Tóth, K. F., and Aravin, A. A. (2019). The control of gene expression and cell identity by H3K9 trimethylation. *Development (Cambridge)*, 146(19).

Nishioka, K., Rice, J. C., Sarma, K., Erdjument-Bromage, H., Werner, J., Wang, Y., Chuikov, S., Valenzuela, P., Tempst, P., Steward, R., Lis, J. T., David Allis, C., and Reinberg, D. (2002). PR-Set7 Is a Nucleosome-Specific Methyltransferase that Modifies Lysine 20 of Histone H4 and Is Associated with Silent Chromatin The N-terminal tails of the core histone proteins protrude. *Molecular Cell*, 9(6):1201–1213.

Oda, H., Okamoto, I., Murphy, N., Chu, J., Price, S. M., Shen, M. M., Torres-Padilla, M. E., Heard, E., and Reinberg, D. (2009). Monomethylation of Histone H4-Lysine 20 Is Involved in Chromosome Structure and Stability and Is Essential for Mouse Development. *Molecular and Cellular Biology*, 29(8):2278–2295.

Olek, A. and Walter, J. (1997). The pre-implantation ontogeny of the H19 methylation imprint. *Nature*, 17(3):275–276.

Osterwalder, M., Barozzi, I., Tissiéres, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Akiyama, J. A., et al. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, 554(7691):239–243.

Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S., and Natoli, G.

(2013). Latent enhancers activated by stimulation in differentiated cells. *Cell*, 152(1-2):157–171.

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. *Current Biology*, 10(8):475–478.

Paige, S. L., Thomas, S., Stoick-Cooper, C. L., Wang, H., Maves, L., Sandstrom, R., Pabon, L., Reinecke, H., Pratt, G., Keller, G., Moon, R. T., Stamatoyannopoulos, J., and Murry, C. E. (2012). A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell*, 151(1):221–232.

Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G. A., Stewart, R., and Thomson, J. A. (2007). Whole-Genome Analysis of Histone H3 Lysine 4 and Lysine 27 Methylation in Human Embryonic Stem Cells. *Cell Stem Cell*, 1(3):299–312.

Pandya-Jones, A., Bhatt, D. M., Lin, C. H., Tong, A. J., Smale, S. T., and Black, D. L. (2013). Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression. *RNA*, 19(6):811–827.

Papp, B. and Plath, K. (2013). Epigenetics of reprogramming to induced pluripotency. *Cell*, 152(6):1324–1343.

Pappa, S., Padilla, N., Iacobucci, S., Vicioso, M., De La Campa, E. Ã., Navarro, C., Marcos, E., De La Cruz, X., and Martínez-Balbás, M. A. (2019). PHF2 histone demethylase prevents DNA damage and genome instability by controlling cell cycle progression of neural progenitors. *Proceedings of the National Academy of Sciences of the United States of America*, 116(39):19464–19473.

Park, S. H., Park, S. H., Kook, M. C., Kim, E. Y., Park, S., and Lim, J. H. (2004). Ultrastructure of human embryonic stem cells and spontaneous and retinoic acid-induced differentiating cells. *Ultrastructural Pathology*, 28(4):229–238.

Pasini, D., Bracken, A. P., Hansen, J. B., Capillo, M., and Helin, K. (2007). The Polycomb Group Protein Suz12 Is Required for Embryonic Stem Cell Differentiation. *Molecular and Cellular Biology*, 27(10):3769–3779.

Pengelly, A. R., Copur, Ã., Jäckle, H., Herzig, A., and Müller, J. (2013). A histone mutant reproduces the phenotype caused by loss of histone-modifying factor polycomb. *Science*, 339(6120):698–699.

Pérez-Lluch, S., Blanco, E., Carbonell, A., Raha, D., Snyder, M., Serras, F., and Corominas, M. (2011). Genome-wide chromatin occupancy analysis reveals a role for ASH2 in transcriptional pausing. *Nucleic Acids Research*, 39(11):4628–4639.

Pérez-Lluch, S., Blanco, E., Tilgner, H., Curado, J., Ruiz-Romero, M., Corominas, M., and Guigó, R. (2015). Absence of canonical marks of active chromatin in developmentally regulated genes. *Nature Genetics*, 47(10):1158–1167.

Pervouchine, D., Popov, Y., Berry, A., Borsari, B., Frankish, A., and Guigó, R. (2019). Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. *Nucleic Acids Research*, 47(10):5293–5306.

Pervouchine, D. D., Djebali, S., Breschi, A., Davis, C. A., Barja, P. P., Dobin, A., Tanzer, A., Lagarde, J., Zaleski, C., See, L. H., Fastuca, M., Drenkow, J., Wang, H., Bussotti, G., Pei, B., et al. (2015). Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. *Nature Communications*, 6.

Pimentel, H., Parra, M., Gee, S. L., Mohandas, N., Pachter, L., and Conboy, J. G. (2016). A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Research*, 44(2):838–851.

Piunti, A. and Shilatifard, A. (2016). Epigenetic balance of gene expression by polycomb and compass families. *Science*, 352(6290).

Plath, K., Fang, J., Mlynarczyk-Evans, S. K., Cao, R., Worringer, K. A., Wang, H., De la Cruz, C. C., Otte, A. P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. *Science*, 300(5616):131–135.

Plet, A., Eick, D., and Blanchard, J. (1995). Elongation and premature termination of transcripts initiated from c-fos and c-myc promoters show dissimilar patterns. *Oncogene*, 10(2):319–328.

Podlaha, O., De, S., Gonen, M., and Michor, F. (2014). Histone Modifications Are Associated with Transcript Isoform Diversity in Normal and Cancer Cells. *PLoS Computational Biology*, 10(6).

Poepsel, S., Kasinath, V., and Nogales, E. (2018). Cryo-EM structures of PRC2 simultaneously engaged with two functionally distinct nucleosomes. *Nature Structural and Molecular Biology*, 25(2):154–162.

Pradeepa, M. M., Grimes, G. R., Kumar, Y., Olley, G., Taylor, G. C., Schneider, R., and Bickmore, W. A. (2016). Histone H3 globular domain acetylation identifies a new class of enhancers. *Nature Genetics*, 48(6):681–686.

Pradeepa, M. M., Sutherland, H. G., Ule, J., Grimes, G. R., and Bickmore, W. A. (2012). Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genetics*, 8(5).

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., Schierup, M. H., and Jensen, T. H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, 322(5909):1851–1854.

Proudfoot, N. J. (2011). Ending the message: Poly(A) signals then and now. *Genes and Development*, 25(17):1770–1782.

Rada-Iglesias, A. (2018). Is H3K4me1 at enhancers correlative or causative? *Nature Genetics*, 50(1):4–5.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–285.

Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., Sharp, P. A., and Young, R. A. (2010). C-Myc regulates transcriptional pause release. *Cell*, 141(3):432–445.

Raisner, R., Kharbanda, S., Jin, L., Jeng, E., Chan, E., Merchant, M., Haverty, P. M., Bainer, R., Cheung, T., Arnott, D., Flynn, E. M., Romero, F. A., Magnuson, S., and Gascoigne, K. E. (2018). Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. *Cell Reports*, 24(7):1722–1729.

Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M., Durham, T., Zhang, X., Donaghey, J., Epstein, C. B., Regev, A., et al. (2011). Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell*, 147(7):1628–1639.

Rapino, F., Robles, E. F., Richter-Larrea, J. A., Kallin, E. M., Martinez-Climent, J. A., and Graf, T. (2013). C/EBPα Induces Highly Efficient Macrophage Transdifferentiation of B Lymphoma and Leukemia Cell Lines and Impairs Their Tumorigenicity. *Cell Reports*, 3(4):1153–1163.

Rasmussen, E. B. and Lis, J. T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proceedings of the National Academy of Sciences of the United States of America*, 90(17):7923–7927.

Reverón-Gómez, N., González-Aguilera, C., Stewart-Morgan, K. R., Petryk, N., Flury, V., Graziano, S., Johansen, J. V., Jakobsen, J. S., Alabert, C., and Groth, A. (2018). Accurate Recycling of Parental Histones Reproduces the Histone Modification Landscape during DNA Replication. *Molecular Cell*, 72(2):239–249.

Rickels, R., Herz, H. M., Sze, C. C., Cao, K., Morgan, M. A., Collings, C. K., Gause, M., Takahashi, Y. H., Wang, L., Rendleman, E. J., Marshall, S. A., Krueger, A., Bartom, E. T., Piunti, A., Smith, E. R., et al. (2017). Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nature Genetics*, 49(11):1647–1653.

Riddle, N. C., Minoda, A., Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Tolstorukov, M. Y., Gorchakov, A. A., Jaffe, J. D., Kennedy, C., Linder-Basso, D., Peach, S. E., Shanower, G., Zheng, H., Kuroda, M. I., Pirrotta, V., et al. (2011). Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome Research*, 21(2):147–163.

Robertson, A. G., Bilenky, M., Tam, A., Zhao, Y., Zeng, T., Thiessen, N., Cezard, T., Fejes, A. P., Wederell, E. D., Cullum, R., Euskirchen, G., Krzywinski, M., Birol, I., Snyder, M., Hoodless, P. A., et al. (2008). Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Research*, 18(12):1906–1917.

Rougvie, A. E. and Lis, J. T. (1988). The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell*, 54(6):795–804.

Ruthenburg, A. J., Allis, C. D., and Wysocka, J. (2007). Methylation of Lysine 4 on Histone H3: Intricacy of Writing and Reading a Single Epigenetic Mark. *Molecular Cell*, 25(1):15–30.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E. E. (2007). A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes and Development*, 21(4):436–449.

Sandmann, T., Jensen, L. J., Jakobsen, J. S., Karzynski, M. M., Eichenlaub, M. P., Bork, P., and Furlong, E. E. (2006). A Temporal Map of Transcription Factor Activity: Mef2 Directly Regulates Target Genes at All Stages of Muscle Development. *Developmental Cell*, 10(6):797–807.

Santos, F., Hendrich, B., Reik, W., and Dean, W. (2002). Dynamic reprogramming of DNA methylation in the early mouse embryo. *Developmental Biology*, 241(1):172–182.

Santos-Rosa, H., Schneider, R., Bannister, A., Sherriff, J., Bernstein, B., Emre, N., Schreiber, S., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. *Nature*, 419(6905):407–411.

Sardina, J. L., Collombet, S., Tian, T. V., Gómez, A., Di Stefano, B., Berenguer, C., Brumbaugh, J., Stadhouders, R., Segura-Morales, C., Gut, M., Gut, I. G., Heath, S., Aranda, S., Di Croce, L., Hochedlinger, K., et al. (2018). Transcription Factors Drive Tet2-Mediated Enhancer Demethylation to Reprogram Cell Fate. *Cell Stem Cell*, 23(5):727–741.

Schaaf, C. A., Misulovin, Z., Gause, M., Koenig, A., Gohara, D. W., Watson, A., and Dorsett, D. (2013). Cohesin and Polycomb Proteins Functionally Interact to Control Transcription at Silenced and Active Genes. *PLoS Genetics*, 9(6).

Schmidt, U., Basyuk, E., Robert, M. C., Yoshida, M., Villemin, J. P., Auboeuf, D., Aitken, S., and Bertrand, E. (2011). Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: Implications for alternative splicing regulation. *Journal of Cell Biology*, 193(5):819–829.

Schneider, J., Wood, A., Lee, J. S., Schuster, R., Dueker, J., Maguire, C., Swanson, S. K., Florens, L., Washburn, M. P., and Shilatifard, A. (2005). Molecular regulation of histone H3 trimethylation by COMPASS and the regulation of gene expression. *Molecular Cell*, 19(6):849–856.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B. M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W., Dimitrova, E., Dimond, A., Edelman, L. B., Elderkin, S., Tabbada, K., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, 25(4):582–597.

Schotta, G., Sengupta, R., Kubicek, S., Malin, S., Kauer, M., Callén, E., Celeste, A., Pagani, M., Opravil, S., De La Rosa-Velazquez, I. A., Espejo, A., Bedford, M. T., Nussenzweig, A., Busslinger, M., and Jenuwein, T. (2008). A chromatin-wide transition to H4K20 monomethylation impairs genome integrity and programmed DNA rearrangements in the mouse. *Genes and Development*, 22(15):2048–2061.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nature Structural and Molecular Biology*, 16(9):990–995.

Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., Young, R. A., and Sharp, P. A. (2008). Divergent transcription from active promoters. *Science*, 322(5909):1849–1851.

Seki, Y., Hayashi, K., Itoh, K., Mizugaki, M., Saitou, M., and Matsui, Y. (2005). Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. *Developmental Biology*, 278(2):440–458.

Shachar, S., Voss, T. C., Pegoraro, G., Sciascia, N., and Misteli, T. (2015). Identification of Gene Positioning Factors Using High-Throughput Imaging Mapping. *Cell*, 162(4):911–923.

Shao, G.-B., Ding, H.-M., Gong, A.-H., and Xiao, D.-S. (2008). Inheritance of Histone H3 Methylation in Reprogramming of Somatic Nuclei Following Nuclear Transfer. *Journal of Reproduction and Development*, 54(3):233–238.

Shi, X., Kachirskaia, I., Walter, K. L., Kuo, J. H. A., Lake, A., Davrazou, F., Chan, S. M., Martin, D. G., Fingerman, I. M., Briggs, S. D., Howe, L. A., Utz, P. J., Kutateladze, T. G., Lugovskoy, A. A., Bedford, M. T., et al. (2007). Proteome-wide analysis in Saccharomyces cerevisiae identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36. *Journal of Biological Chemistry*, 282(4):2450–2455.

Shikata, D., Yamamoto, T., Honda, S., Ikeda, S., and Minami, N. (2020). H4K20 monomethylation inhibition causes loss of genomic integrity in mouse preimplantation embryos. *Journal of Reproduction and Development*, 66(5):411–419.

Siersbæk, R., Nielsen, R., John, S., Sung, M. H., Baek, S., Loft, A., Hager, G. L., and Mandrup, S. (2011). Extensive chromatin remodelling and establishment of transcription factor hotspots during early adipogenesis. *EMBO Journal*, 30(8):1459–1472.

Silva, J., Mak, W., Zvetkova, I., Appanah, R., Nesterova, T. B., Webster, Z., Peters, A. H., Jenuwein, T., Otte, A. P., and Brockdorff, N. (2003). Establishment of histone H3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Developmental Cell*, 4(4):481–495.

Simpson, R. T. (1978). Structure of Chromatin Containing Extensively Acetylated H3 and H4. *Cell*, 13(4):691–699.

Small, S., Blair, A., and Levine, M. (1992). Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO Journal*, 11(11):4047–4057.

Smith, E., Lin, C., and Shilatifard, A. (2011). The super elongation complex (SEC) and MLL in development and disease. *Genes and Development*, 25(7):661–672.

Smith, H. F., Roberts, M. A., Nguyen, H. Q., Peterson, M., Hartl, T. A., Wang, X. J., Klebba, J. E., Rogers, G. C., and Bosco, G. (2013). Maintenance of interphase chromosome compaction and homolog pairing in Drosophila is regulated by the condensin Cap-H2 and its partner Mrg15. *Genetics*, 195(1):127–146.

Smith, Z. D., Chan, M. M., Mikkelsen, T. S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase

of DNA methylation in the early mammalian embryo. *Nature*, 484(7394):339–344.

Smith, Z. D. and Meissner, A. (2013). DNA methylation: Roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220.

Soares, L. M., He, P. C., Chun, Y., Suh, H., Kim, T. S., and Buratowski, S. (2017). Determinants of Histone H3K4 Methylation Patterns. *Molecular Cell*, 68(4):773–785.

Song, J. and Chen, K. C. (2015). Spectacle: Fast chromatin state annotation using spectral learning. *Genome Biology*, 16(1):33.

Soufi, A., Donahue, G., and Zaret, K. S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell*, 151(5):994–1004.

Spies, N., Nielsen, C. B., Padgett, R. A., and Burge, C. B. (2009). Biased Chromatin Signatures around Polyadenylation Sites and Exons. *Molecular Cell*, 36(2):245–254.

Spitz, F. and Furlong, E. E. (2012). Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626.

Sridharan, R., Tchieu, J., Mason, M. J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., and Plath, K. (2009). Role of the Murine Reprogramming Factors in the Induction of Pluripotency. *Cell*, 136(2):364–377.

Steward, M. M., Lee, J. S., O'Donovan, A., Wyatt, M., Bernstein, B. E., and Shilatifard, A. (2006). Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nature Structural and Molecular Biology*, 13(9):852–854.

Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41–45.

Strobl, L. J. and Eick, D. (1992). Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. *EMBO Journal*, 11(9):3307–3314.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14(1):43–59.

Sun, Z., Zhang, Y., Jia, J., Fang, Y., Tang, Y., Wu, H., and Fang, D. (2020). H3K36me3, message from chromatin to DNA damage repair. *Cell and Bioscience*, 10(1).

Sutton, W. S. (1902). On the morphology of the chromosome group in brachystola magna. *The Biological Bulletin*, 4(1):24–39.

Sutton, W. S. (1903). The chromosomes in heredity. *The Biological Bulletin*, 4(5):231–250.

Tagoh, H., Schebesta, A., Lefevre, P., Wilson, N., Hume, D., Busslinger, M., and Bonifer, C. (2004). Epigenetic silencing of the c-fms locus during B-lymphopoiesis occurs in discrete steps and is reversible. *EMBO Journal*, 23(21):4275–4285.

Talasz, H., Lindner, H. H., Sarg, B., and Helliger, W. (2005). Histone H4-lysine 20 monomethylation is increased in promoter and coding regions of active genes and correlates with hyperacetylation. *Journal of Biological Chemistry*, 280(46):38814–38822.

Tamburri, S., Lavarone, E., Fernández-Pérez, D., Conway, E., Zanotti, M., Manganaro, D., and Pasini, D. (2020). Histone H2AK119 Mono-Ubiquitination Is Essential for Polycomb-Mediated Transcriptional Repression. *Molecular Cell*, 77(4):840–856.

Taylor, G. C., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M. M., and Bickmore, W. A. (2013). H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Research*, 23(12):2053–2065.

Therizols, P., Illingworth, R. S., Courilleau, C., Boyle, S., Wood, A. J., and Bickmore, W. A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science*, 346(6214):1238–1242.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., et al. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Tie, F., Banerjee, R., Stratton, C. A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M. O., Scacheri, P. C., and Harte, P. J. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development*, 136(18):3131–3141.

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nature Structural & Molecular Biology*, 16(9):996–1001.

Torres-Padilla, M. E., Parfitt, D. E., Kouzarides, T., and Zernicka-Goetz, M. (2007). Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature*, 445(7124):214–218.

Tremblay, K. D., Duran, K. L., and Bartolomei, M. S. (1997). A 5 2-Kilobase-Pair Region of the Imprinted Mouse H19 Gene Exhibits Exclusive Paternal Methylation throughout Development. *Molecular and Cellular Biology*, 17(8):4322–4329.

Trojer, P. and Reinberg, D. (2007). Facultative Heterochromatin: Is There a Distinctive Molecular Signature? *Molecular Cell*, 28(1):1–13.

Trompouki, E., Bowman, T. V., Lawton, L. N., Fan, Z. P., Wu, D. C., Dibiase, A., Martin, C. S., Cech, J. N., Sessa, A. K., Leblanc, J. L., Li, P., Durand, E. M., Mosimann, C., Heffner, G. C., Daley, G. Q., et al. (2011). Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell*, 147(3):577–589.

Ulianov, S. V., Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., Penin, A. A., Logacheva, M. D., Imakaev, M. V., Chertovich, A., Gelfand, M. S., Shevelyov, Y. Y., and Razin, S. V. (2016). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Research*, 26(1):70–84.

Ullrich, S. and Guigó, R. (2020). Dynamic changes in intron retention are tightly associated with regulation of splicing factors and proliferative activity during B-cell development. *Nucleic acids research*, 48(3):1327–1340.

Vakoc, C. R., Sachdeva, M. M., Wang, H., and Blobel, G. A. (2006). Profile of Histone Lysine Methylation across Transcribed Mammalian Chromatin. *Molecular and Cellular Biology*, 26(24):9185–9195.

Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., Murata, M., Nishiyori, H., Lazarevic, D., Motti, D., Marstrand, T. T., Tang, M. H. E., Zhao, X., Krogh, A., Winther, O., et al. (2009). Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Research*, 19(2):255–265.

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J. Y., Cody, N. A., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, 583(7818):711–719.

Vermeulen, M., Mulder, K. W., Denissov, S., Pijnappel, W. W., van Schaik, F. M., Varier, R. A., Baltissen, M. P., Stunnenberg, H. G., Mann, M., and Timmers, H. T. M. (2007). Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell*, 131(1):58–69.

Villar, D., Flicek, P., and Odom, D. T. (2014). Evolution of transcription factor binding in metazoans-mechanisms and functional implications. *Nature Reviews Genetics*, 15(4):221–233.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858.

Voigt, P., LeRoy, G., Drury, W. J., Zee, B. M., Son, J., Beck, D. B., Young, N. L., Garcia, B. A., and Reinberg, D. (2012). Asymmetrically modified nucleosomes. *Cell*, 151(1):181–193.

Voigt, P., Tee, W. W., and Reinberg, D. (2013). A double take on bivalent promoters. *Genes and Development*, 27(12):1318–1338.

Vuong, C. K., Black, D. L., and Zheng, S. (2016). The neurogenetics of alternative splicing. *Nature Reviews Neuroscience*, 17(5):265–281.

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. A., Winston, F., Buratowski, S., and Handa, H. (1998). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes and Development*, 12(3):343–356.

Waddington, C. H. (1953). Epigenetics and evolution. *Symposia of the Society of Experimental Biology*, 7:186–199.

Wakabayashi, K.-i., Okamura, M., Tsutsumi, S., Nishikawa, N. S., Tanaka, T., Sakakibara, I., Kitakami, J.-i., Ihara, S., Hashimoto, Y., Hamakubo, T., Kodama, T., Aburatani, H., and Sakai, J. (2009). The Peroxisome Proliferator-Activated Receptor γ/Retinoid X Receptor α Heterodimer Targets the Histone Modification Enzyme PR-Set7/Setd8 Gene and Regulates Adipogenesis through a Positive Feedback Loop. *Molecular and Cellular Biology*, 29(13):3544–3555.

Wamstad, J. A., Alexander, J. M., Truty, R. M., Shrikumar, A., Li, F., Eilertson, K. E., Ding, H., Wylie, J. N., Pico, A. R., Capra, J. A., Erwin, G., Kattman, S. J., Keller, G. M., Srivastava, D., Levine, S. S., et al. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*, 151(1):206–220.

Wang, A., Kurdistani, S. K., and Grunstein, M. (2002). Requirement of Hos2 histone deacetylase for gene activity in yeast. *Science*, 298(5597):1412–1414.

Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R. S., and Zhang, Y. (2004). Role of histone H2A ubiquitination in Polycomb silencing. *Nature*, 431(7010):873–878.

Wang, Z., Zang, C., Cui, K., Schones, D. E., Barski, A., Peng, W., and Zhao, K. (2009). Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes. *Cell*, 138(5):1019–1031.

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Peng, W., Zhang, M. Q., and Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903.

Waszak, S. M., Delaneau, O., Gschwind, A. R., Kilpinen, H., Raghav, S. K., Witwicki, R. M., Orioli, A., Wiederkehr, M., Panousis, N. I., Yurovsky, A., Romano-Palumbo, L., Planchon, A., Bielser, D., Padioleau, I., Udin, G., et al. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell*, 162(5):1039–1050.

Weinert, B. T., Narita, T., Satpathy, S., Srinivasan, B., Hansen, B. K., Schölz, C., Hamilton, W. B., Zucconi, B. E., Wang, W. W., Liu, W. R., Brickman, J. M., Kesicki, E. A., Lai, A., Bromberg, K. D., Cole, P. A., et al. (2018). Time-Resolved Analysis Reveals Rapid Dynamics and Broad Scope of the CBP/p300 Acetylome. *Cell*, 174(1):231–244.

Wiencke, J. K., Zheng, S., Morrison, Z., and Yeh, R. F. (2008). Differentially expressed genes are marked by histone 3 lysine 9 trimethylation in human cancer cells. *Oncogene*, 27(17):2412–2421.

Wilhelm, B. T., Marguerat, S., Aligianni, S., Codlin, S., Watt, S., and Bähler, J. (2011). Differential patterns of intronic and exonic DNA regions with respect to RNA polymerase II occupancy, nucleosome density and H3K36me3 marking in fission yeast. *Genome Biology*, 12(8).

Wilson, E. (1925). *The cell in development and heredity.* Macmillan, New York.

Wong, J. J., Ritchie, W., Ebner, O. A., Selbach, M., Wong, J. W., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., Thoeng, A., Khoo, T. L., Bailey, C. G., Holst, J., and Rasko, J. E. (2013). Orchestrated intron retention regulates normal granulocyte differentiation. *Cell*, 154(3):583–595.

Wossidlo, M., Arand, J., Sebastiano, V., Lepikhov, K., Boiani, M., Reinhardt, R., Schöler, H., and Walter, J. (2010). Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. *The EMBO Journal*, 29(11):1877–1888.

Wysocka, J., Swigut, T., Xiao, H., Milne, T. A., Kwon, S. Y., Landry, J., Kauer, M., Tackett, A. J., Chait, B. T., Badenhorst, P., Wu, C., and Allis, C. D. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442(7098):86–90.

Xhemalce B, Dawson MA, and Bannister AJ (2011). Histone modifications. In Meyers R, editor, *Encyclopedia of Molecular Cell Biology and Molecular Medicine*, volume 1. John Wiley and Sons.

Xi, H., Shulha, H. P., Lin, J. M., Vales, T. R., Fu, Y., Bodine, D. M., McKay, R. D., Chenoweth, J. G., Tesar, P. J., Furey, T. S., Ren, B., Weng, Z., and Crawford, G. E. (2007). Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genetics*, 3(8):1377–1388.

Xu, J., Pope, S. D., Jazirehi, A. R., Attema, J. L., Papathanasiou, P., Watts, J. A., Zaret, K. S., Weissman, I. L., and Smale, S. T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):12377–12382.

Xu, J., Watts, J. A., Pope, S. D., Gadue, P., Kamps, M., Plath, K., Zaret, K. S., and Smale, S. T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. *Genes and Development*, 23(24):2824–2838.

Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., and Handa, H. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, 97(1):41–51.

Yang, X., Noushmehr, H., Han, H., Andreu-Vieyra, C., Liang, G., and Jones, P. A. (2012). Gene Reactivation by 5-Aza-2'-Deoxycytidine-Induced Demethylation Requires SRCAP-Mediated H2A.Z Insertion to Establish Nucleosome Depleted Regions. *PLoS Genetics*, 8(3):e1002604.

Yuh, C.-H., Ransick, A., Martinez, P., Britten, R. J., and Davidson, E. H. (1994). Complexity and organization of DNA-protein interactions in the 5'-regulatory region of an endoderm-specific marker gene in the sea urchin embryo. *Mechanisms of Development*, 47:165–186.

Zaret, K. S. and Carroll, J. S. (2011). Pioneer transcription factors: Establishing competence for gene expression. *Genes and Development*, 25(21):2227–2241.

Zavolan, M., Van Nimwegen, E., and Gaasterland, T. (2002). Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Research*, 12(9):1377–1385.

Zeitlinger, J., Simon, I., Harbison, C. T., Hannett, N. M., Volkert, T. L., Fink, G. R., and Young, R. A. (2003). Program-Specific Distribution of a Transcription Factor Dependent on Partner Transcription Factor and MAPK Signaling. *Cell*, 113(3):395–404.

Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Research*, 21(8):1273–1283.

Zhang, J. A., Mortazavi, A., Williams, B. A., Wold, B. J., and Rothenberg, E. V. (2012). Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell*, 149(2):467–482.

Zhang, K., Lin, W., Latham, J. A., Riefler, G. M., Schumacher, J. M., Chan, C., Tatchell, K., Hawke, D. H., Kobayashi, R., and Dent, S. Y. (2005). The Set1 methyltransferase opposes Ipl1 Aurora kinase functions in chromosome segregation. *Cell*, 122(5):723–734.

Zhang, T., Zhang, Z., Dong, Q., Xiong, J., and Zhu, B. (2020). Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology*, 21(1).

Zhang, X., Bernatavichute, Y. V., Cokus, S., Pellegrini, M., and Jacobsen, S. E. (2009). Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome Biology*, 10(6).

Zhang, Y., An, L., Yue, F., and Hardison, R. C. (2016). Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Research*, 44(14):6721–6731.

Zhang, Y. and Hardison, R. C. (2017). Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Research*, 45(17):9823–9836.

Zhao, X. D., Han, X., Chew, J. L., Liu, J., Chiu, K. P., Choo, A., Orlov, Y. L., Sung, W. K., Shahab, A., Kuznetsov, V. A., Bourque, G., Oh, S., Ruan, Y., Ng, H. H., and Wei, C. L. (2007). Whole-Genome Mapping of Histone H3 Lys4 and 27 Trimethylations Reveals Distinct Genomic Compartments in Human Embryonic Stem Cells. *Cell Stem Cell*, 1(3):286–298.

Zheng, Y., Thomas, P. M., and Kelleher, N. L. (2013). Measurement of acetylation turnover at distinct lysines in human histones identifies long-lived acetylation sites. *Nature Communications*, 4.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J. Q., and Tian, D. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*, 10.