UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE DE STRASBOURG

IGBMC – CNRS UMR 7104 – Inserm U 1258

THÈSE présentée par :

Sergio SARNATARO

soutenue le : **5 Juin 2020**

pour obtenir le grade de: **Docteur de l'université de Strasbourg**

Discipline/ Spécialité: Biologie Computationnelle

Cinétique de réactivation de la transcription et réorganisation de la structure de la chromatine après la mitose

Transcription reactivation kinetics and chromatin structure reorganization after mitosis

THÈSE dirigée par:

Dr. MOLINA Nacho — CR, IGBMC, Université de Strasbourg

RAPPORTEURS:

Dr. STADLER Michael — DR, Friedrich Miescher Institute, Basel (Switzerland)

Prof. ZAVOLAN Mihaela — DR, Biozentrum, Basel (Switzerland)

AUTRES MEMBRES DU JURY:

Dr. SEXTON Thomas — CR, IGBMC, Université de Strasbourg

"If you want to understand function, study structure"
F. Crick

Acknowledgements

I thank Nacho Molina for giving me the opportunity to have my PhD in his research group, and all the lab members for the useful discussions and the very positive work environment.

Contents

Thesis summary in English

All the cells in an organism, despite their wide distinct functions, have an identical genome, which contains the information to produce all the necessary proteins. During the development, the identity and the function of the cells are established through gene regulation, where specific genes are expressed in some cells and silenced in others. In addition, the environment of the cells can induce changes in gene expression, as response to stimuli and signals from external players. The regulation of the activity of a given gene depends on certain important regions of the genome: the promoter and the enhancers. The promoter is located in proximity of the transcription start site (TSS) of the gene and is typically few hundred base-pairs long. It contains binding sites for transcription factors (TFs) that recruit the RNA polymerase and the basal components of the transcriptional machinery. Enhancers are gene-distal regulatory regions, since they can be found up to some thousands of kilobases far from the gene, and play a crucial role in the regulation of gene expression and in the nuclear organization of the genome by promoting physical contacts between promoters and enhancers and the recruitment of the transcriptional machinery. How these important regions get in contact despite their great genomic distance, together with the entire three dimensional architecture of the DNA, have been deeply studied in the last decade. Some key concepts about these studies are reported hereafter.

If we stretched the entire human DNA, we would obtain a linear length of about 2 meters. All this DNA is contained into a cell nucleus with a diameter of about $5\mu m$. These values provide an idea of how complex is the architecture of the DNA inside the nucleus. Indeed, DNA is bound to proteins, called histones, which have the fundamental role of packaging the DNA in a more condensed manner. Together with the DNA, histones form a reinforced complex, the so-called chromatin.

Chromosomes are the highest level of chromatin organization. Different organisms show different number of chromosomes, and each chromosome can be present in a different number of copy (ploidy). The largest part of the eukaryotic organisms is diploid, i.e. they present two homologous copies per chromosome. It has been observed that each chromosome occupies a specific position in the nucleus, called chromosome territory (CT). Furthermore, chromatin can show two structural forms that are tightly linked to gene regulation: euchromatin, rich of genes, more dispersed and less compacted,

and heterochromatin, much more condensed and closer to the nuclear envelope. Remarkably, heterochromatin is usually less accessible to TFs and polymerase and consequently transcriptionally silent compared to euchromatin. Finally, two genomic regions that are at a great genomic distance can be physically close in the 3D space. This important feature of the three dimensional structure of chromatin plays a fundamental role in the interaction between enhancers and promoters and has been shown it has a direct consequence on gene regulation.

In the last decade, new experimental approaches based on high-throughput sequencing have been developed in order to further investigate genome-wide the topological properties of chromatin and its impact on the regulation of transcription. These approaches are based on the chromosome conformation capture (3C) technique, and enable the estimation of the frequency of the interaction between different loci across a cell population. Briefly, the four key steps of 3C-based techniques are as follows: firstly, crosslinking of cells is performed, and the segments of chromatin that are physically close in space are linked by covalent bonds. Then, there is a fragmentation process on the cross-linked chromatin, by using digestion enzymes. The generated fragments are then ligated and form hybrid DNA molecules. Finally, a purification allows to detect and analyze the pairwise interactions, that can be quantified. Hi-C experiments are one of the most used 3C-based techniques. The results of the Hi-C experiments are shown as matrices, where each entry Hij corresponds to a value that is proportional to the frequency of interaction between the locus i and the locus j. To visualize the experimental data, the matrices are presented as heatmaps, where the color represents the frequency of interaction. In the standard Hi-C experiments, the values Hij represent an average over a cell population, while recent work led to the single cell Hi-C maps (scHi-C), where the matrices refers to a single cell, unmasking a significant variability among the cells. By analyzing Hi-C data, some important chromatin features have been discovered, concerning the 3D organization of chromatin in the cell nucleus, such as chromatin loops and topologically associated domains (TADs). A chromatin loop emerges when two regions of the same chromosome, but at a great genomic distance, show a strong interaction. Looking at a heatmap, chromatin loops appear as isolated points with a strong interaction, and they play a crucial role in gene regulation, linking promoters and enhancers. Moreover, loops are conserved across different cell types and correlate with the presence of CTCF proteins and with the boundaries of another, important architectural feature, the topologically associated domains (TADs). Looking at a Hi-C map, TADs are visually recognizable as squares of enriched contacts along the diagonal, and they represent fundamental, isolated units along the genomic coordinates. Practically, they are characterized by the fact that the regions inside a domain show a strongly enriched interaction, while they interact much less with the regions outside.

Several studies proposed a scenario where the mentioned three dimensional features of chromatin in the cell nucleus and gene regulation are tightly associated and reciprocally involved, besides the contact between promoters and enhancers. However, the exact cause-consequence link is still unclear. This relationship is especially interesting in the context of the cell-cycle. During mitosis, in fact, a dramatic reorganization of the nucleus occurs: nuclear envelope is disassembled, chromatin is compacted and mitotic chromosomes are formed. As a consequence, long-range interactions and TADs are disrupted, most TFs and the basal transcriptional machinery are evicted from chro-

matin and transcription is globally downregulated. To ensure the cell well-functioning after division, transcription has to be re-initiated at the appropriate set of genes, once mitosis is completed. Interestingly, Hi-C experiments on synchronized cell populations have revealed the kinetic process by which different levels of chromatin structure are reformed after mitosis. Moreover, recently live-imaging and biochemistry experiments have shown that some TFs are capable of binding mitotic chromosomes. It is believed that this phenomenon, known as mitotic bookmarking, helps maintaining cell identity by propagating gene regulatory programs from mother to daughter cells. However, it is still unclear how bookmarking may influence the precise kinetics of chromatin structure reorganization and transcription re-activation after mitosis.

In this thesis, we tried to uncover the existing link between the three-dimensional organization of chromatin and the regulation of the transcriptional machinery, by combining computational analyses and mathematical modeling of data from high-throughput experiments, such as RNA-Seq and Hi-C. In particular, the aims of this research are to reveal the key regulators responsible for the reactivation of the transcription exiting the mitosis, and to infer the most important factors driving the structural reorganization of the chromatin through the cell-cycle. Thereafter, the main points and results of this study will be summarized.

Firstly, to study the hierarchical organization of the 3D structure of chromatin, we adopted a graph-based approach to detect communities in networks. To do so, we modeled chromatin as a network where nodes are chromatin regions and edges represent physical contacts between regions determined by Hi-C data. Then, to detect chromatin domains at different resolutions, we used the stability algorithm, which measures the regions-quality as community structures, leveraging on the dynamical evolution of a Markov process that takes place on the chromatin network across different time scales. The detected communities were then called diffusion associated domains (DADs). In this approach, the time represents a parameter which determines the resolution of the detected domains, identifying different hierarchical levels of the chromatin structure. This computational method was applied to published Hi-C data obtained at different stages of Drosophila Melanogaster (Dmel) embryogenesis. Although recent literature stated that the chromatin architecture of Dmel emerges with the onset of transcription activation in the zygote, while prior to zygotic genome activation the cell nucleus is mostly unstructured, by using our diffusion-like approach, we showed the presence of a "backbone" of the structure in the earliest developmental stages, with almost 68% of the three-dimensional architecture that is kept through the development. A similar analyis was performed on a population of synchronized HeLa cells by mitotic arrest, for which timing Hi-C datasets were obtained at different timepoints exiting mitosis, showing a DADs conservation level of almost 66% between mitotic chromatin and latest experimental timepoints.

These promising results suggested an early gene expression, and directed our research to investigate transcription reactivation during and exiting mitosis. To do so, we analyzed published data based on metabolic labeling of RNA (EU-RNA-Seq) of synchronized population of Human Hepatoma cells HUH7. In this study, the authors highlighted the presence of low levels of transcription during mitosis and the fact that housekeeping genes and not cell-specific genes are activated earlier during the mitotic exit. However, the study did not consider that mitotic-arrested cell populations pro-

gressively de-synchronize after the block release, and therefore the reported measurements are performed on mixture of cells at different internal cell-cycle times. We developed a mathematical model, assuming that after synchronization there is a stochastic lag time until cells can start again the cell-cycle progression, and that there is a certain average time t_m to complete the mitosis. We introduced the concept of *internal cell-cycle time*, defined as the effective cell-cycle time progression of every single cell, starting once the lag time is over. By using our mathematical model, we were able to deconvolve the expression of every gene given by the EU-RNA-Seq data to the internal cell-cycle time, in order to solve the uncertainty due to the progressive de-synchronization of the cell population. Moreover, using imaging data on the time evolution of the fraction of mitotic cells observed after synchronization treatment release, we could fit the mean and standard deviation of the distribution of lag times and the average time to complete the mitosis, that we estimated to be, respectively, 3.43, 0.74 and 67 minutes. That was crucial to individuate clusters of earlier activated genes, and to divide genes in groups based on their first activation peak of expression.

Another main goal of the research project was to find the key factors determining the expression kinetics. To do so, we developed a linear model assuming that the EU-RNA-Seq expression of genes at a given time point of the cell-cycle progression is a linear combination of the activities of all the TFs that can bind on their promoters. By knowing the deconvolved gene expression and integrating data on transcription factors motif affinities, we calculated the activity of every expressed TF. This has the major advantage of describing the problem of the reactivation with much less parameters, since we passed from the analysis of about 12000 genes to only about 330 TFs. Moreover, this analysis allowed to divide the TFs in groups according to the importance of their activity over the internal cell-cycle time, identifying key factors that are early active with respect to the others and that may play a crucial role in the transcription re-activation during and exiting mitosis. In addition, further analyses were performed to compare our results with the behavior of TFs that have been proposed as candidates for bookmarking activity. Interestingly, we did not see any correlation between known bookmarking factors as FOXA1 and the speed at which their target genes are reactivated. However, we identified around 60 TFs that are highly active during mitosis and represent new candidates of mitotic bookmarking factors. This surprising result leads to wonder what is the real purpose of the bookmarking, that apparently is not always directly involved in the expression of genes. Among the hypotheses, the role of the mitotic bookmarking could be structural, with some TFs binding the mitotic DNA to keep the chromatin open and accessible to other factors and to the RNA polymerase, in order to make possible the re-initiation of the transcription starting from the latest stages of the mitosis.

Lastly, to further test the mutual relationship between gene regulation and 3D structure of chromatin in the nucleus, we applied the linear model mentioned above to infer the transcription factors activity during formation and development of TAD boundaries. We used published time dependent Hi-C data on synchronized HeLa cells obtained during mitotic exit and early interphase. TADs detected 11 hours after the release of the mitotic arrest were considered as a reference, and a total of almost 4000 boundaries were found. The insulation score (IS) was used as a metric for the boundaries strength, and it was calculated at each earlier experimental timepoint in correspondence of the position of the reference boundaries. By knowing the IS and integrating data on transcription factors motif affinities on the boundaries, we inferred the activity

of almost 400 TFs. Thanks to this analysis we could identify new TFs that may play an important role in the formation of TADs after mitosis.

In conclusion, we developed a computational method that allows to detect communities in the chromatin with a network-based approach, by using an algorithm where the diffusion time plays a crucial role as a parameter determining the resolution of the detected chromatin domains, allowing to reveal different hierarchical levels of the three-dimensional chromatin structure. Furthermore, our studies provided a linear mathematical model to sort the TFs according to their importance on the reactivation of post-mitotic transcription, as well as find clusters of TFs that show similar activity dynamics over the cell-cycle. By applying our model to RNA-Seq data, we did not see any correlation between known bookmarking factors and the speed at which their target genes are reactivated. However, we identified around 60 TFs that are highly active during mitosis and represent new candidates of mitotic bookmarking factors.

Thesis summary in French

Toutes les cellules d'un organisme, en dépit de leurs fonctions très distinctes, ont un génome identique qui contient les informations nécessaires pour produire les protéines. Au cours du développement, l'identité et la fonction des cellules sont établies par la régulation des gènes, où des gènes spécifiques sont exprimés dans certaines cellules et inactivés dans d'autres. En outre, l'environnement des cellules peut induire des changements dans l'expression des gènes, en réponse à des stimuli et des signaux provenant de facteurs externes. La régulation de l'activité d'un gène donné dépend de certaines régions importantes du génome : le promoteur et les amplificateurs (enhancers). Le promoteur est situé à proximité du site d'initiation de la transcription (en anglais, transcription start site, TSS) du gène et est généralement long de quelques centaines de paires de bases. Il contient des sites de liaison pour les facteurs de transcription (FT) qui recrutent l'ARN polymérase et les composants basaux de la machinerie transcriptionnelle. Les amplificateurs sont des régions régulatrices distales du gène, puisqu'ils peuvent se trouver jusqu'à plusieurs milliers de kilobases du gène. Ils jouent un rôle crucial dans la régulation de l'expression des gènes et dans l'organisation nucléaire du génome en favorisant les contacts physiques entre les promoteurs et les amplificateurs, ainsi que le recrutement de la machinerie transcriptionnelle. La façon dont ces régions entrent en contact malgré les grandes distances génomiques qui les séparent, ainsi que toute l'architecture tridimensionnelle de l'ADN, ont été étudiées en profondeur au cours de la dernière décennie. Certains concepts clés de ces études sont présentés ci-après.

Si nous étirions l'ADN humain dans son intégralité, nous obtiendrions une longueur linéaire d'environ 2 mètres. Tout cet ADN est contenu dans un noyau de cellule d'un diamètre d'environ $5\mu m$. Ces valeurs donnent une idée de la complexité de l'architecture de l'ADN à l'intérieur du noyau. En effet, l'ADN est lié à des protéines, appelées histones, qui ont le rôle fondamental de compacter l'ADN de manière plus condensée. Avec l'ADN, les histones forment un complexe renforcé : la chromatine.

Les chromosomes sont le plus haut niveau d'organisation de la chromatine. Différents organismes présentent un nombre différent de chromosomes, et chaque chromosome peut être présent dans un nombre différent de copies (ploïdie). La plus grande partie des organismes eucaryotes est diploïde, c'est-à-dire qu'ils présentent deux copies homologues par chromosome. Il a été observé que chaque chromosome occupe une position spécifique dans le noyau, appelée territoire chromosomique (TC). De plus,

la chromatine peut présenter deux formes structurelles étroitement liées à la régulation des gènes: l'euchromatine, riche en gènes, plus dispersée et moins compactée, et l'hétérochromatine, beaucoup plus condensée et plus proche de l'enveloppe nucléaire. Il est remarquable que l'hétérochromatine soit généralement moins accessible aux FT et à la polymérase et, par conséquent, moins active sur le plan de la transcription que l'euchromatine. Enfin, deux régions génomiques qui sont à une grande distance génomique, peuvent être physiquement proches dans l'espace 3D. Cette caractéristique importante de la structure tridimensionnelle de la chromatine joue un rôle fondamental dans l'interaction entre les activateurs et les promoteurs et il a été démontré qu'elle a une conséquence directe sur la régulation des gènes.

Au cours de la dernière décennie, de nouvelles approches expérimentales basées sur le séquençage à haut débit (en anglais, next generation sequencing, NGS) ont été développées afin d'étudier plus en détail les propriétés topologiques de la chromatine à l'échelle du génome et son impact sur la régulation de la transcription. Ces approches sont basées sur la technique de capture de la conformation chromosomique (3C) et permettent d'estimer la fréquence de l'interaction entre différents loci dans une population cellulaire. En bref, les quatre étapes clés des techniques basées sur la 3C sont les suivantes: tout d'abord, la réticulation des cellules (crosslinking) est effectuée, et les segments de chromatine qui sont physiquement proches dans l'espace sont liés par des liaisons covalentes. Ensuite, il y a un processus de fragmentation sur la chromatine réticulée, en utilisant des enzymes de digestion. Les fragments générés sont ensuite ligaturés et forment des molécules d'ADN hybrides. Enfin, une purification permet de détecter et d'analyser les interactions par paires, qui peuvent être quantifiées.

Les expériences Hi-C sont l'une des techniques les plus utilisées basées sur le 3C. Les résultats des expériences Hi-C sont présentés sous forme de matrices, où chaque entrée Hij correspond à une valeur qui est proportionnelle à la fréquence d'interaction entre le locus i et le locus j. Pour visualiser les données expérimentales, les matrices sont présentées sous forme de cartes de chaleur (heatmaps), où la couleur représente la fréquence d'interaction. Dans les expériences Hi-C standard, les valeurs Hij représentent une moyenne effectuée sur une population de cellules, alors que des travaux récents ont permis l'élaboration de cartes Hi-C à cellule unique (en anglais, single-cell Hi-C, scHi-C), où les matrices permettent d'effectuer ces mesures sur une seule cellule. Ces dernières ont permis de mettre en évidence une variabilité significative entre les cellules. En analysant les données Hi-C, certaines caractéristiques importantes de la chromatine ont été découvertes, notamment sur son organisation tridimensionnelle qui présente des boucles (chromatin loops) et des domaines topologiquement associés (en anglais, Topologically Associated Domains, TAD). Une boucle de chromatine émerge lorsque deux régions du même chromosome, séparées par une grande distance génomique, présentent une forte interaction. Sur une carte de chaleur, les boucles de chromatine apparaissent comme des points isolés montrant une forte interaction. Elles jouent un rôle crucial dans la régulation des gènes en liant les promoteurs et les amplificateurs. De plus, les boucles sont conservées dans différents types de cellules. Ces sites présentent le plus souvent un enrichissement pour la protéine CTCF et se retrouvent près des limites des domaines topologiquement associés (TAD). En regardant une carte de chaleur Hi-C, les TAD sont visuellement reconnaissables comme des carrés de contacts enrichis le long de la diagonale. Ils représentent des unités fondamentales, isolées le long des coordonnées génomiques. Concrètement, ils se caractérisent par le fait que les régions à l'intérieur d'un domaine montrent une interaction fortement en-

richie, alors qu'elles interagissent beaucoup moins avec les régions à l'extérieur.

Outre les contacts entre les promoteurs et les activateurs, plusieurs études ont proposé un scénario dans lequel les caractéristiques tridimensionnelles de la chromatine dans le noyau cellulaire et la régulation des gènes sont étroitement associées et réciproquement impliquées. Toutefois, les relations de cause à effet ne sont pas encore très claires. Cette relation est particulièrement intéressante dans le contexte du cycle cellulaire. En effet, au cours de la mitose, une réorganisation radicale du noyau se produit : l'enveloppe nucléaire est désassemblée, la chromatine est compactée et des chromosomes mitotiques sont formés. En conséquence, les interactions à longue distance et les TAD sont perturbés, la plupart des FT et la machinerie de transcription basale sont expulsées de la chromatine et la transcription est globalement régulée à la baisse. Pour assurer le bon fonctionnement de la cellule après la division, la transcription doit être relancée au niveau de l'ensemble des gènes appropriés, une fois la mitose terminée. Il est intéressant de noter que les expériences de Hi-C sur des populations de cellules synchronisées ont révélé le processus cinétique par lequel différents niveaux de structure de la chromatine sont reformés après la mitose. De plus, des expériences récentes d'imagerie et de biochimie ont montré que certains FT sont capables de se lier aux chromosomes mitotiques. On pense que ce phénomène, connu sous le nom de mitotic bookmarking (mise en signet mitotique), aide à maintenir l'identité des cellules en propageant les programmes de régulation des gènes des cellules mères aux cellules filles. Cependant, on ne sait toujours pas comment le mitotic bookmarking peut influencer la cinétique précise de la réorganisation de la structure de la chromatine et de la réactivation de la transcription après la mitose.

Dans cette thèse, nous avons tenté de découvrir le lien existant entre l'organisation tridimensionnelle de la chromatine et la régulation de la machinerie transcriptionnelle, en combinant des analyses informatiques et la modélisation mathématique de données provenant d'expériences à haut débit telles que le RNA-Seq et le Hi-C. Ces recherches visent principalement à mettre en évidence les principaux régulateurs responsables de la réactivation de la transcription à la sortie de la mitose, et à identifier les facteurs les plus importants de la réorganisation structurelle de la chromatine dans le cycle cellulaire. Les principaux points et résultats de cette étude seront résumés ci-après.

Tout d'abord, afin d'étudier l'organisation hiérarchique de la structure 3D de la chromatine, nous avons adopté une approche basée sur les graphes pour détecter les régions les plus fortement connectées (appelées communautés) dans les réseaux. Pour ce faire, nous avons modélisé la chromatine comme un réseau où les nœuds sont des régions de chromatine et les liens représentent les contacts physiques entre les régions déterminées par les données Hi-C. Ensuite, pour détecter les domaines de chromatine à différentes résolutions, nous avons utilisé l'algorithme de stabilité (stability algorithm). Cet algorithme évalue chaque région pour déterminer son appartenance éventuelle aux communautés en s'appuyant sur l'évolution dynamique d'un processus de Markov qui se déroule sur le réseau de chromatine à différentes échelles de temps. Les communautés détectées sont appelées domaines associés à la diffusion (en anglais, diffusion associated domains, DAD). Dans cette approche, le temps représente un paramètre qui détermine la résolution des domaines détectés, en identifiant les différents niveaux hiérarchiques de la structure de la chromatine. Cette méthode de calcul a été appliquée sur des données publiées de Hi-C obtenues à différents stades de l'embryogenèse de la drosophile mélanogaster (Dmel). La littérature récente indique que l'architecture de la chromatine de Dmel émerge avec le début de l'activation de la transcription dans le

zygote, alors qu'avant l'activation du génome zygotique le noyau cellulaire est essentiellement non structuré. Cependant, en utilisant notre approche de diffusion, nous avons montré la présence d'un "squelette" de structure même à l'état mitotique, avec près de 68% de l'architecture tridimensionnelle conservée tout au long du développement.

Une analyse similaire a été effectuée sur une population de cellules HeLa synchronisées pendant la sortie mitotique, pour laquelle des ensembles de données Hi-C de synchronisation ont été obtenus à différents points temporels sortant de la mitose, montrant un niveau de conservation des DAD de près de 66 $\%$ entre la chromatine mitotique et les derniers points temporels expérimentaux.

Ces résultats prometteurs suggèrent une expression génétique précoce, et ont orienté nos recherches vers l'étude de la réactivation de la transcription pendant et après la mitose. Pour ce faire, nous avons analysé des données publiées, basées sur le marquage métabolique de l'ARN (EU-RNA-Seq) d'une population de cellules synchronisées d'hépatome humain HUH7. Dans cette étude, les auteurs ont souligné la présence de faibles niveaux de transcription pendant la mitose et le fait que les gènes domestiques (housekeeping genes) et non spécifiques aux cellules sont activés plus tôt pendant la sortie de la mitose. Cependant, l'étude n'a pas pris en compte le fait que les populations de cellules arrêtées en mitose se désynchronisent progressivement après la reprise du cycle, et les mesures rapportées sont donc effectuées sur un mélange de cellules à différents stades du cycle cellulaire interne. Nous avons développé un modèle mathématique, en supposant qu'après la synchronisation, il y a un temps de latence stochastique jusqu'à ce que les cellules puissent recommencer la progression du cycle cellulaire, et qu'il y a un certain temps moyen t_m pour terminer la mitose. Nous avons introduit le concept de "temps de cycle cellulaire interne", défini comme la progression effective du temps de cycle cellulaire de chaque cellule, commençant une fois que le temps de latence est terminé. En utilisant notre modèle mathématique, nous avons pu déconvoluer l'expression de chaque gène issue des données EU-RNA-Seq en fonction du temps du cycle cellulaire interne, afin de résoudre l'incertitude due à la désynchronisation progressive de la population cellulaire. De plus, en utilisant les données d'imagerie sur l'évolution temporelle de la fraction de cellules mitotiques observée après l'arrêt du traitement de synchronisation, nous avons pu ajuster la moyenne et l'écart-type de la distribution des temps de latence et du temps moyen pour achever la mitose, que nous avons estimé à respectivement 3.43, 0.74 et 67 minutes. Cette étape était cruciale pour identifier les groupes de gènes activés précocement ainsi que pour regrouper les gènes en fonction du moment où apparait leur premier pic d'activation. Un autre objectif majeur du projet de recherche était de trouver les facteurs clés déterminant la cinétique d'expression. Pour ce faire, nous avons développé un modèle linéaire dans lequel les données d'expression EU-RNA-Seq d'environ 12 000 gènes de la lignée cellulaire HUH7 à un point donné de la progression du cycle cellulaire résulte de la combinaison des activités des FT pour lesquelles un motif de liaison est connu et pourrait se lier aux promoteurs du gène. En connaissant l'expression des gènes calculée par rapport à la durée du cycle cellulaire interne et en intégrant les données sur les affinités des motifs de FT, nous avons déduit l'activité de chaque FT exprimé. Ceci a l'avantage majeur de décrire le problème de la réactivation avec beaucoup moins de paramètres, puisque nous sommes passés de l'analyse d'environ 12 000 gènes à environ 330 FT seulement. De plus, cette analyse a permis de diviser les FT en groupes selon l'importance de leur activité au cours du cycle cellulaire interne, en identifiant

les facteurs clés qui sont activés précocement par rapport aux autres et qui peuvent jouer un rôle crucial dans la réactivation de la transcription après la sortie de la mitose.

En outre, des analyses supplémentaires ont été effectuées pour comparer nos résultats avec le comportement des FT qui ont été proposés comme candidats à l'activité de bookmarking. En particulier, nous avons montré que les gènes associés au facteur de transcription FOXA1, observés au microscope dans les chromosomes mitotiques, atteignent leur pic d'expression plus tard par rapport à la moyenne de tous les gènes. Ce résultat surprenant amène à se demander quel est le but réel du bookmarking, qui apparemment n'est pas toujours directement impliqué dans l'expression des gènes. Parmi les hypothèses, le rôle du mitotic bookmarking pourrait être structurel, avec quelques FT liant l'ADN mitotique pour maintenir la chromatine ouverte et accessible à d'autres facteurs et à l'ARN polymérase, afin de rendre possible la reprise de la transcription à partir des dernières étapes de la mitose.

Enfin, pour tester davantage la relation mutuelle entre la régulation des gènes et la structure 3D de la chromatine dans le noyau, nous avons appliqué le modèle linéaire décrit ci-dessus pour déduire l'activité des facteurs de transcription pendant la formation et le développement des limites des TAD. Nous avons utilisé des données de synchronisation Hi-C publiées sur des cellules HeLa synchronisées pendant la sortie mitotique, afin de pouvoir accéder aux données concernant la sortie de l'arrêt de la prométaphase, jusqu'à 12 heures plus tard. Ces mesures ont été réalisées en 16 points au cours du temps. Les TAD à 11h sont considérés comme le point de référence, avec un total de près de 4000 limites identifiées. Le score d'isolement (en anglais, insulation score, IS) a été utilisé comme mesure de la robustesse de ces limites. Il a été calculé en fonction de la position des limites de référence pour chaque étape temporelle. En connaissant l'IS et en intégrant les données sur l'affinité des motifs présents sur les limites, nous avons déduit l'activité de près de 400 FT. Grâce à cette analyse, nous avons pu identifier de nouveaux FT qui pourraient jouer un rôle important dans la formation des TAD après la mitose. Cependant, notre modèle n'a pas détecté la présence de CTCF dont plusieurs études antérieures ont démontré la forte corrélation avec la formation des limites des TAD. Des recherches supplémentaires sont donc nécessaires pour valider la puissance de notre méthode et la liste des nouveaux TF identifiés.

En conclusion, nous avons développé une méthode de calcul qui permet de détecter des communautés dans la chromatine par une approche de réseau qui utilise un algorithme où le temps de diffusion joue un rôle crucial comme paramètre déterminant la résolution des domaines de chromatine détectés. Cela permet de révéler différents niveaux hiérarchiques de la structure tridimensionnelle de la chromatine. En outre, nos études ont fourni un modèle mathématique linéaire permettant de trier les FT selon leur importance pour la réactivation de la transcription, ainsi que de trouver des groupes de FT qui présentent une dynamique d'activité similaire au cours du cycle cellulaire. En appliquant notre modèle aux données RNA-Seq, nous n'avons vu aucune corrélation entre les facteurs de signet connus et la vitesse à laquelle leurs gènes cibles sont réactivés. Cependant, nous avons identifié environ 60 FT qui sont très actifs pendant la mitose et représentent de nouveaux candidats de mitotic bookmarking.

List of Figures

1

Introduction

Premise 1: as decided in the *Official Bulletin of the Ministry of Education, 2016, n. 35 (September 29th, 2016, cf 3.2.2 and 4.2.3)*, figures taken from other papers must have a maximum definition of 400 x 400 pixels and a maximum resolution of 72 DPI. In order to follow these rules, some picture from this chapter might be not of high quality. However, the reference of every picture is indicated in the caption, so then we refer to the original papers for better resolution figures.

Premise 2: according to the decision of the *Administrative Council of the University of Strasbourg (November 24, 2009)*, a part of the thesis must be written in French. Then, besides the summary, also captions of figures from this chapter are in French.

In the last decade, several studies proposed a scenario where the three-dimensional organization of chromatin in the cell nucleus and gene regulation are tightly associated and reciprocally involved. This relationship is especially interesting in the context of the cell-cycle, where a dramatic, structural reorganization of the nucleus occurs, affecting the functional activities of the cells. However, the exact cause-consequence nexus is still unclear.

In this thesis, we tried to uncover the existing link between the three-dimensional organization of chromatin and the regulation of the transcriptional machinery, by combining computational analyses and mathematical modeling of data from high-throughput experiments, such as RNA-Seq and Hi-C. In particular, the aims of this research are to reveal the key regulators responsible for the reactivation of the transcription exiting the mitosis, and to infer the most important factors driving the structural reorganization of the chromatin through the cell-cycle.

In this chapter, some concepts and recent experiments from literature will be presented, in order to put our study in the right context and provide a comprehensive biological view.

1.1 Gene regulation

Distinct types of cells in the same organism have the same genome, which contains complete information to transcribe any molecules of RNA and proteins. During the development, identity and function of the cells are established by regulation of the transcription, with specific genes that are expressed in some cells and silenced in others. In addition, the environment of the cells can induce changes in gene expression [1], as response to stimuli and signals from external players.

The regulation of the activity of a given gene depends on some important regions of the genome: the promoter and the enhancers. The promoter is located in proximity of the Transcription Start Site (TSS) of the gene and is few hundreds base-pairs long. It is a docking platform for Transcription Factors (TFs), RNA polymerase and the other components of the transcriptional machinery. The enhancers are gene-distal regulators, since they can be found up to some thousands of kilobases far from the gene, and play a crucial role in the regulation of the gene expression and in the nuclear organization, in the scenario of the physical contacts between promoters and enhancers [4].

There are different levels at which the control of the gene expression can occur [1]: *transcriptional control*, that is how many times the transcription of a specific gene occurs; *RNA processing control*, controlling the splicing; the *transport and localization of RNA*, regulating exportation of completed mRNA outside the nucleus and its localization in the cytosol; *translational control*, establishing mRNAs in cytoplasm that have to be translated; *control of the degradation of mRNA*; *protein activity control*, concerning protein activity after their production.

In principle, each of these steps can be regulated in order to establish which genes are finally expressed. However, for the majority of the genes, initiating the RNA transcription is the key step for controlling the whole process, being the only way to avoid the synthesis of unnecessary intermediate products.

The recent, fast development of new experimental technologies such as new generation sequencing (NGS) allowed to measure the quantity and sequences of RNA in a population of cells (RNA-Seq) or in a single cell (scRNA-Seq), at a given time. Putting together these techniques with time-course experiments represents a stimulating challenge for a quantitative analysis of the dynamics of gene regulation.

1.2 3D organisation of the chromatin in the nucleus

In the cell nucleus of eukaryotic organisms, DNA is bound to some proteins, called histones, which have the fundamental role of packaging the DNA in a more condensed manner. Together with the DNA, histones form a reinforced complex, called chromatin. Moreover, it has been shown that histone modifications play a crucial role in gene regulation [2].

However, the interplay between DNA and histones is not the only organisational level of the chromatin. For example, if we stretched the entire human DNA, we wuold obtain a linear lenght of about 2 meters, and these DNA is contained into a cell nucleus with a diameter of about $2\mu m$: these values provide an idea of how complex is the architecture of the chromatin inside the nucleus.

Chromosomes are the higest level of this architecture. Different organisms show different number of chromosomes, and each chromosome can be present in a different number of copy (ploidy). The largest part of the eukaryotic organisms are diploid, i.e.

they present two homologous copies of each chromosome. For instance, human cells have 23 different couples of chromosomes while mice have 20. Chromosomes occupy a specific position inside the nucleus, called *chromosome territories (CTs)* [3], as shown in fig. 1.1.

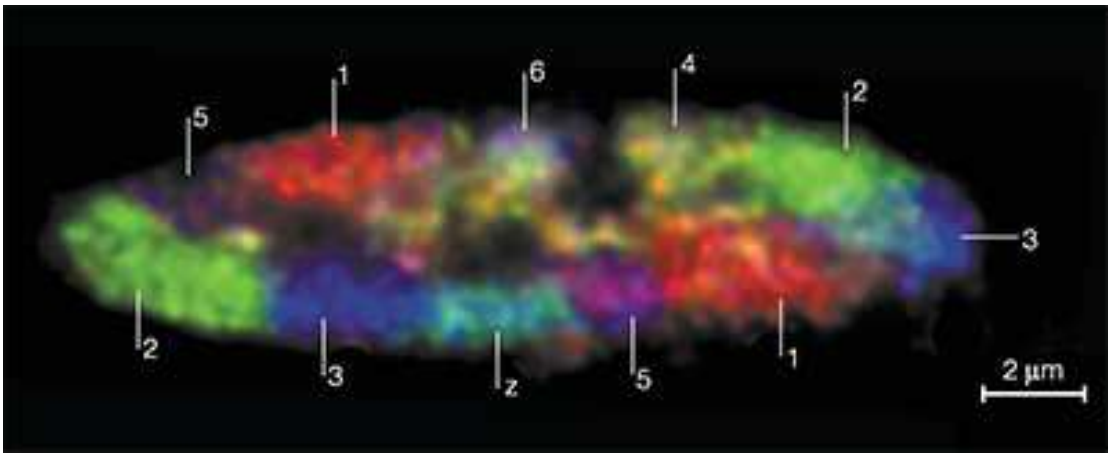


Figure 1.1. Territoires chromosomiques dans le noyau de fibroblastes de poulet, observés par microscopie par hybridation fluorescente in situ (FISH). Différentes couleurs indiquent différents chromosomes. Les chromosomes homologues sont situés dans différents territoires. Adaptée de [3]

Another organisational level is strictly linked to the gene regulation. In fact, chromatin can show two structural forms: *euchromatin*, rich of genes, more dispersed and less compacted, and *heterochromatin*, much more condensed and closer to the nuclear envelope. For its structural form, heterochromatin is usually less accessible to polymerase and consequently gene-poor, if compared to the euchromatin.

Recently, new high-resolution experimental techniques led to discover other organizational levels of the chromatin in the cell nucleus, as it will be explained in the next sections.

1.2.1 3C-based techniques

As discussed above, microscopy techniques using fluorescence (FISH) have been used to uncover the organisation of the cell nucleus, revealing the existence of chromosome territories. Despite these techniques have been a crucial step for the comprehension of the general architecture of the nucleus, some technical limitations made them inadequate for the whole understanding of the chromatin organization. For instance, the high specificity of the probe sequences and the low resolution do not allow to reveal the genome-wide pattern of the chromatin [5].

In the last few years, some new experimental approaches based on the high-throughput sequencing data have been developed, in order to uncover the genome-wide topological properties of the chromatin and the impact of such a structure on the functional mechanisms of the cells. These approaches leverage the *chromosome conformation capture* (3C) techniques, and enable the estimation of the frequency of the interaction between different loci across a cell population. More attention is given to the fact that two genomic regions that are at a great genomic distance could be phisically close in the 3D space. This is an important feature which play a fundamental

role in the interaction between enhancers and promoters, shedding further light on the gene regulation.

These techniques will be briefly presented hereafter. All the 3C based techniques present the same 4 key steps, as shown in fig. 1.2, panel a. Firstly, crosslinking of cells is performed, and segments of chromatin that are physically close in space are linked by covalent bonds. Then, there is a fragmentation process on the crosslinked chromatin, by using some digestion enzymes such as HindIII, DpnII or NcoI. The generated fragments are then ligated to form hybride DNA molecules. Finally, a purification allows to detect and analyse the pairwaise interactions, that can be quantified. This last step (detection and quantification) differentiates the distinct 3C based approaches [6]. Hybrid DNA

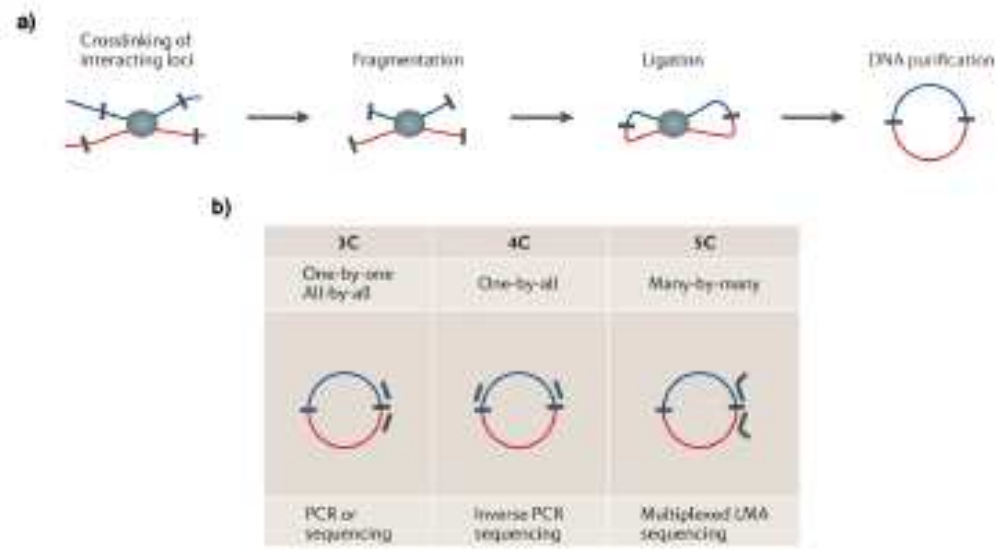


Figure 1.2. Présentation des techniques basées sur 3C. **a:** Étapes expérimentales des techniques basées sur 3C. **b:** Détection 3C, 4C et 5C du produit de la ligature. Adaptée de [6]

molecules produced by the ligation process are detected one by one by using PCR in classical 3C experiments, through the use of specific primers. 3C is mostly used to test candidate interacting loci, such as promoters and enhancers. 4C experiments are based on inverse PCR. Here, the interactions of a single locus are analysed genome-wide. 5C is instead the analysis of *many-vs-many* segments, analysing the long-range interactions of all the restriction fragments, up to some megabases of distance. A brief summary of these techniques is shown in fig. 1.2, panel b.

1.2.2 Hi-C experiments

In 2009, Lieberman-Aiden and others [10] proposed *Hi-C*, a new experimental 3C-based method to investigate the long range interactions of the chromatin in the cell nucleus. After the crosslinking of the cells by using formaldehyde, the digestion takes place leaving a biotin residue on the staggered ends. After the ligation, the DNA molecules are purified and sheared, and the biotin residues are pulled down. The reconstruction of the frequency of interactions genome-wide is allowed by the biotinylated junctions, which make recognizable the origin fragments, sequenced on the genome.

Resolution of Hi-C experiments is given by the depth of the sequencing: deeper the sequencing, shorter can be the windows of fixed length the genome is divided into. A schematic outlook of Hi-C method is shown in fig. 1.3.

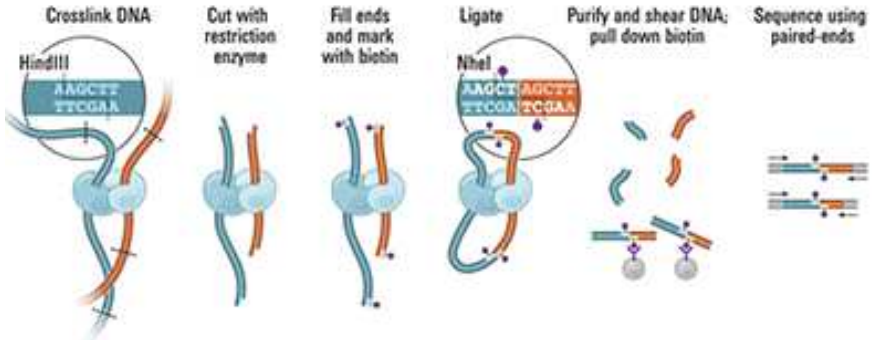


Figure 1.3. Étapes fondamentales des expériences Hi-C. De gauche à droite: réticulation; digestion par des enzymes de restriction; remplir les extrémités décalées et marquer avec de la biotine; ligature; purification; séquencage. Adaptée de [10]

Results of Hi-C experiments are shown as matrices **H**, where each entry H_{ij} corresponds to a value that is proportional to the frequency of interaction between the locus i and the locus j. To facilitate an immediate comprehension of the data, the matrices are presented as heatmaps, with different colours meaning different frequency of interaction (see fig. 1.4 for an example). In standard Hi-C experiments, the values H_{ij} reprensent an average of a cell population, while recent work led to the *single cell Hi-C* (scHi-C), where the matrices refer to single cells [7], unmasking a significant variability among different cells. Hi-C data are mostly used to analyse intra-chromosomal interactions (also called *in-cis* interactions), but some literature about inter-chromosomal interactions (*in-trans*) analysis can be found, such as [9].

1.2.3 Chromatin loops

By analysing Hi-C data, some important chromatin features have been discovered, concerning the 3D organisation of the chromatin in the cell nucleus. A *chromatin loop* emerges when two regions of the same chromosome, but at a great genomic distance,

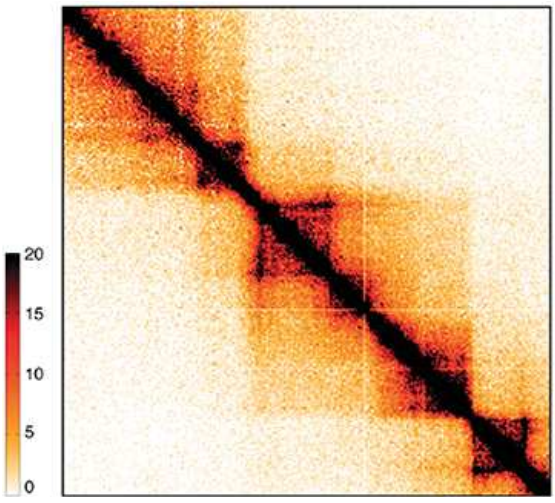


Figure 1.4. Matrice de contact Hi-C provenant de cellules souches embrioniques de souris (mESC), chromosome 10, 38-40,5 Mb, à 10 ko de résolutions. À gauche, une barre de couleur indique l'intensité des interactions. Adaptée de [15]

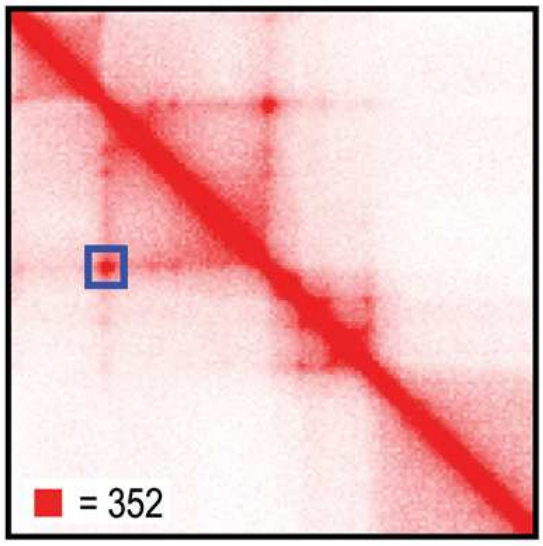


Figure 1.5. Exemple d'une boucle de chromatine sur une carte thermique Hi-C, mise en évidence par le carré bleu. Dans le coin inférieur gauche, la valeur maximale de la matrice est indiquée. Modifiée à partir de [8]

show a strong interaction. Looking at a heatmap, chromatin loops appear as isolated points with a strong interaction. An example is given in the fig. 1.5. As shown in [8], chromatin loops play a crucial role in the gene regulation, linking promoters and enhancers. Moreover, loops are conserved across different cell types and correlate with the presence of CTCF proteins and with the boundaries of the *Topologically Associated Domains*, that will be introduced in the section 1.2.5.

1.2.4 A/B compartments

By using *Principal Component Analysis* (PCA) on correlation matrices of Hi-C data, Lieberman-Aiden and others [10] showed that loci in the chromosomes can be divided into two categories, called *A* and *B*. Regions associated with the *A* and *B* compartments, having a size ranging from few Mb up to almost 10 Mb, are characterized by a strong self-interaction. In fact, contacts between loci belonging to same categories are highly enriched with respect to the loci belonging to different categories, giving to the Pearson correlation Hi-C matrix a charateristic *plaid pattern*. In addition, Lieberman-Aiden and other showed that *A* and *B* categories correlate with euchromatin and heterochromatin respectively (fig. 1.6).

1.2.5 Self-interacting domains

An other, smaller organizational level of the chromatin inside the nucleus is represented by the self-interacting domains, as shown by Dixon and others [11]. They are characterized by the fact that the regions inside a domain show a strongly enriched interaction, while they interact much less with the regions outside. The typical size of the domains showed in [11] is $0.5 - 1.0Mb$. Diverse nomenclatures and methods have been used to show the presence of self-interacting domains, such as *Topological Associated Domains*, TADs ([12]) or generically *contact domains* ([8]), the latter ones with a smaller average length of about $200kb$. Other algorithms that should be mentioned have been developed by Zhan and others ([13]) and by Fraser and others ([14]), where a hierarchical structure of domains-within-domains (metaTADs) have been revealed.

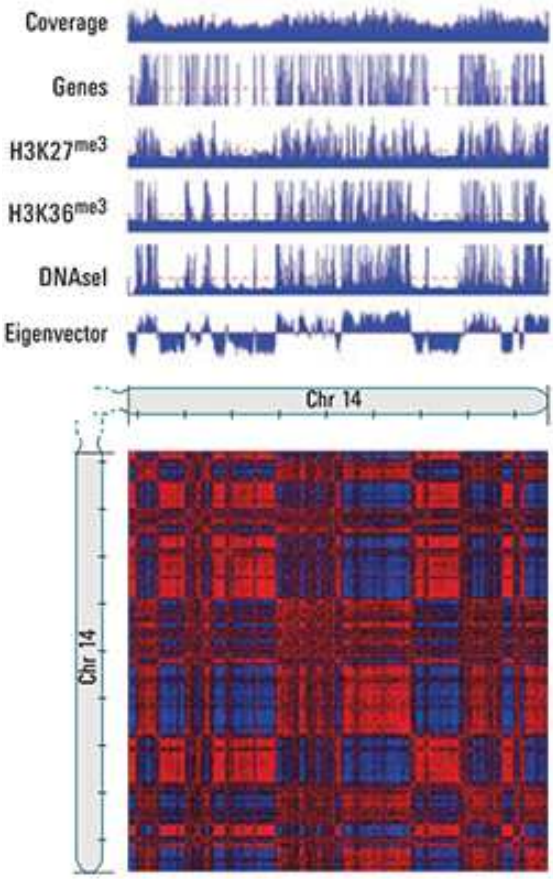


Figure 1.6. En bas, la corrélation de Pearson de la carte thermique Hi-C du chromosome 14 (lignée cellulaire humaine GM06990) à 100 kb de résolution. En haut, les caractéristiques de l'euchromatine en corrélation avec le vecteur propre (composant principal). Adaptée de [10]

Looking at a Hi-C heatmap, domains are visually recognizable as squares of enriched contacts along the diagonal (fig. 1.7), and they represent fundamental, isolated units along the genomic coordinate. From the epigenetic point of view, the domain boundaries correlate with the presence of CTCF, and loci inside the same domains present correlation in histone modifacations for eight different factors ([8]). Despite the relative abundance of algorithms and methods to identify the domains, the biological mechanisms that regulate their formation is still unclear, and investigating such mechanisms by crossing omics and structural data is an open and attractive field.

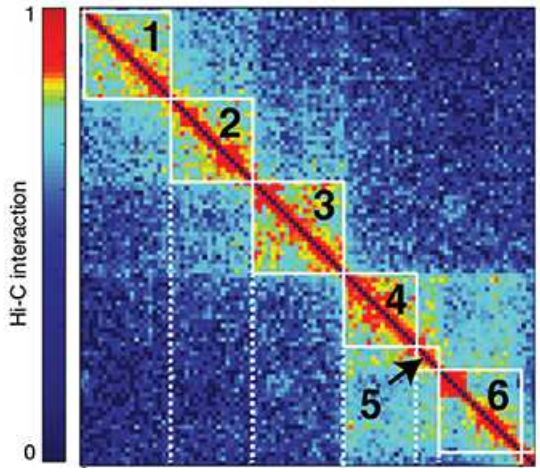


Figure 1.7. Cellules souches embrioniques de souris (mESC), carte thermique Hi-C du chromosome 2, 53-58 Mb. Les valeurs vont du bleu profond (faible interaction) au rouge fort (forte interaction). En utilisant l'algorithme *Index de directionnalité (DI)* [14], 6 domaines ont été trouvés, comme le montrent les carrés blancs le long de la diagonale. Adaptée de [14]

1.3 Gene regulation and 3D structure of the chromatin

Several studies proposed a scenario where the 3D structure of the chromatin in the cell nucleus and the regulation of the genes are hardly associated and reciprocally involved, as explained by Vermunt and others in the review [16]. At the higest level, as seen in 1.2.4, chromatin regions tend to be segregated into two distinct compartments, called *A* and *B*, correlating with euchromatin and heterochromatin chromatin, respectively. Moreover, we already mentioned in section 1.1 that 3D structure of the chromatin and in particular chromatin loops are fundamental to intermediate the contact between gene promoters and distal enhancers.

Besides, some literature ([17][18][19]) highlighted the impact of some structural changes in terms of deaseses, without any variation in the genome sequences. Also, it has been proved that human cancer can correlate with an important level of structural alterations [20].

Taken together, these results shed light on the importance of the 3D structure on the gene regulation and, eventually, on the functional defects that could affect the health of the cells. Understanding which chromosomal regions are in contact and the dynamics of the 3D structure during the cell development is crucial.

In the section 1.5 we will present some studies about the 3D structure of the cell nucleus through the cell-cycle.

1.4 Gene regulation and cell cycle

Cell division is a complex process which needs a perfect replication of the entire genome. Mature cells present a quite robust pattern of gene regulation, that is challenged by the division [24]. To maintain their own identity, cells have to reorganize their structure and their epigenomic features through the cell cycle. It has been shown ([25][26]) that defects and epigenomic instability during this process cause replication stress, which could lead to dangerous alterations in the chromatin trasmitted to the doughter cells. A pictorial representation of this scenario is proposed in fig 1.8. Hereafter in this section, we will briefly present the principal steps of the cell cycle, with particular attention to the mitosis and introducing some models and discoveries concerning the maintenance of the epigenomic pattern and cell identity during cell division and replication.

Figure 1.8. Une représentation picturale du cercle vicieux établissant quand il y a des défauts dans la transmission des informations épigénomiques de la mère à la cellule fille. Adaptée de [24]

1.4.1 Cell cycle

Every organism adopts the same strategy to make a new cell, that is duplicating an already existing cell. Regardless of the cell type, an ordered sequence of steps, together called *cell cycle*, characterizes cell reproduction.

Two main phases characterize the cell cycle: the *S phase*, with the *S* being for *synthesis*, where the DNA is duplicated; and the *M phase*, with the *M* being for *mitosis*, where the chromosomes are segregated and cell divides. *M phase* in turn can be separated in two subphases, i.e. nuclear division, also called *mitosis*, with the formation of two daughter nuclei containing the copied chromosomes, and the *cytokinesis*, where cell divides in two [1].

Besides S and M phases, some *gap phases* are required for some cell cycles. In particular, G_1 takes place between *M* and *S*, and G_2 between *S* and *M*.

Together, G_1, *S* and G_2 are called *interphase*. To provide an idea of the temporal duration, here are some numbers: in culture human cells, interphase takes about 23 hours out of a total of 24 hours, with mitosis lasting for 1 hour. Figure 1.9 provides a summary of the cell cycle. If there are no propitious condition, such as unfavorable

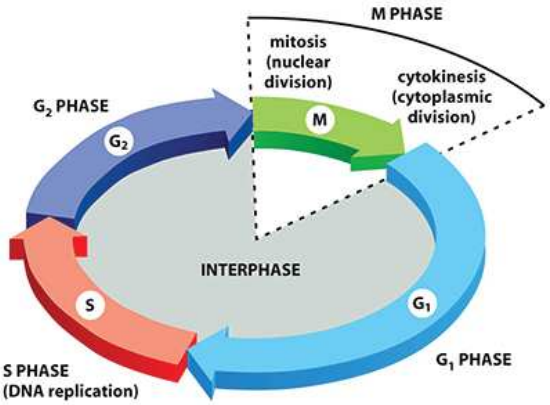


Figure 1.9. Les quatre phases principales du cycle cellulaire. Adaptée de [1]

extracellular environment, before G_1, an other gap phase (G_0) can occur. In principle, G_0 can last several hours or, in extreme cases, cells can remain stuck in G_0 until they die. But once cells reach a *restriction point*, at the end of G_1, they are committed to replicate the DNA [1].

In the next section, we will provide some little information more about the mitosis, that is a crucial step of the cell cycle, where cells undergo the most dramatic changes.

1.4.2 Mitosis

Five stages characterize the mitosis: *prophase, prometaphase, metaphase, anaphase* and *telophase*.

During *prophase*, the condensation of the copied chromosomes occurs, and pairs of rigid rod-shaped chromosomes, called *sister chromatids*, are formed.

In *prometaphase*, the nucleare membrane breaks down, and microtubules invade the nucleus, to form a bipolar array called *mitotic spindle*, to which chromosomes attach, starting active movements.

In *metaphase*, chromosomes allign at the equator of the mitotic spindle.

During *anaphase*, sister chromatids simultaneosly unattach to form the daughter chromosomes, and each of them is pulled toward a pole of the spindle.

Finally, in *telophase*, the two sets of chromosomes reach the poles of the spindle, and undergo a decondensation. A new nuclear membrane forms around each of the set. Cell is now ready for the cytokinesis, consisting in the division of the mother cell in two daughters.

From a regulatory point of view, the highly-condensed shape of the mitotic chromosomes seems to make the transcription of the genes very difficult [21]. The historical scenario presents the mitosis as a transcriptional silent stage, with the eviction of most of the transcription factors and the arrest of the activity of the RNA polymerases [29].

This leads to the question of how the cell identity is kept and how the transcription is reinitiated after the mitosis. Different theories have been developed to address the problem, such as a chromatin-properties approach (DNA methylation or histone modification, [29]).

In the next section, we will introduce the *mitotic bookmarking*, an hypotesis by which a subset of transcription factors binds the DNA during mitosis, giving rise to the reactivation of the transcription.

1.4.3 Mitotic Bookmarking

As explained in the previous section, mitosis corresponds to the most striking reorganisation of the cell nucleus. The nuclear envelope breaks down, increasing the diffusive volume of the transcription factors, then reducing the local concentration and making more difficult the specific binding between regulators and DNA. In addition, condensed and rod-shaped chromosomes are observed [23] (see fig. 1.10). Therefore, in principle, mitotic chromatin is not favorable for the standard gene regulation machinery. As a consequence, most of the transcription factors are evicted and the transcription is donwregulated [27][28]. However, it has been shown that some TFs are capable of binding mitotic chromosomes. It is believed that this phenomenon, known as *mitotic bookmarking* [23], helps conveying gene regulatory information from mother to daughter cells. To understand the mitotic bookmarking theory with respect the other possible approaches, we report the very explicative fig. 1.11.

Although several transcription factors that have been identified as potential candidate for the bookmarking activity, only few of them have been confirmed to be specifically bound to the DNA, thanks to experimental observations. For example, Kadauke and others [29] found that GATA1 shows mitotic binding on genes responsible for hematopoietic regulation; also, Caravaca and others [30], found that FOXA1 remains specifically bound to the mitotic chromosomes in hepatocytes. Other experimental observations concerning transcription factors being strong candidate to be bookmarking

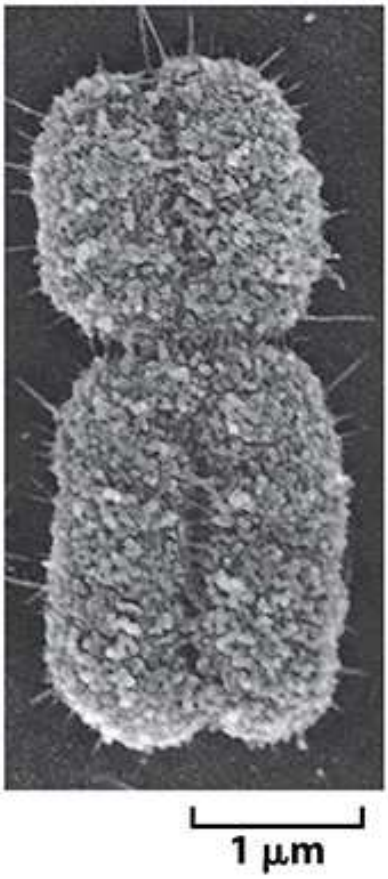


Figure 1.10. Chromosome mitotique de cellule humain. Figure obtenue par micrographie électronique à balayage. Adaptée de [1]

factors, are reported in the fig. 1.12.

1.4.4 Large scale studying on mitotic bookmarking

As described in the previous section, by using diverse experimental techniques such as ChIP-seq and fluorescence microscopy, some TFs have been found binding mitotic chromosomes. ChIP-Seq experiments allow to identify specific binding on the DNA, while fluorescence microscopy identifies the association between TFs and DNA with no respect to enrichments on specific target regions. However, it is believed that is the non-specific rather than specific DNA binding to drive the site specific search of the transcription factors [31][32], and some evidences hint that microscopy observation of TFs on the mitotic chromosomes are due to non-specific bindings [32].

On this basis, Raccaud and others [32] used live cell fluorescent microscopy to investigate the non-specific binding properties of almost 500 TFs in mouse embrionic stem cells (mESC). They measured the mitotic chromosome binding (MCB), by using Mitotic Bound Fraction (MBF) as a metric (see fig. 1.13), and proposed three main categories to rank the analyzed transcription factors, based on visual examination of the fuorescent signal of TFs with respect to the signal of cytoplasm: *depleted*, *intermediate* and *enriched*, with the TFs signal that is lower, equal and higher than the signal in the cytoplasm, respectively.

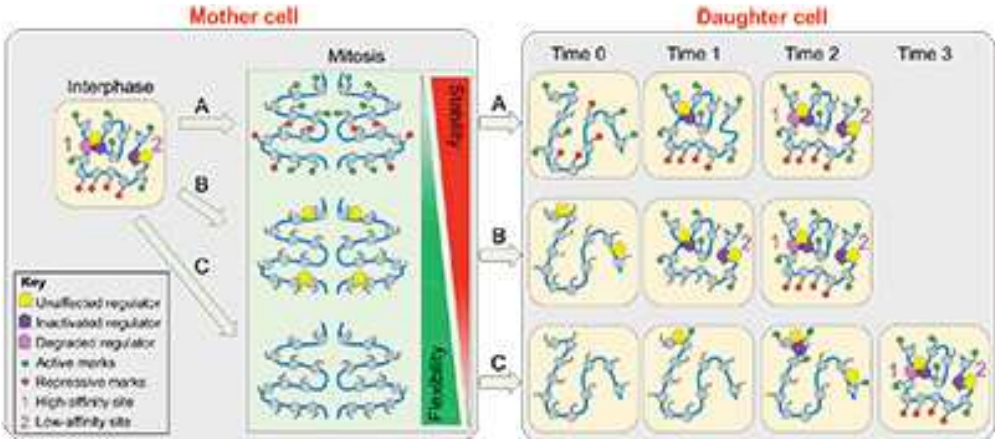


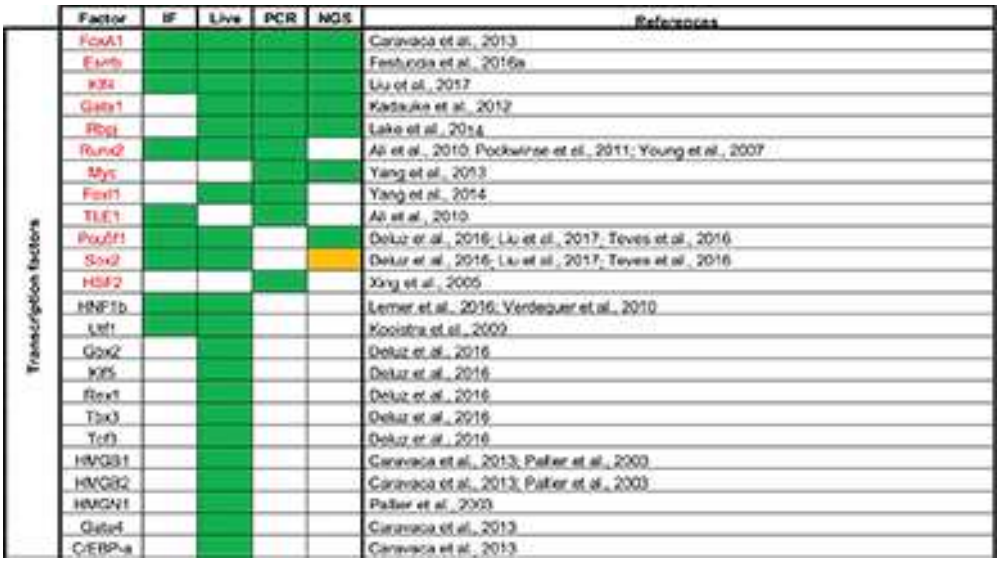
Figure 1.11. Sur cette figure, les cases jaunes indiquent les cellules en interphase, tandis que la case verte représente la cellule mitotique dans trois scénarios possibles, c'est-à-dire que toutes les marques de chromatine sont maintenues (A); seul un sous-ensemble de régulateurs, les *facteurs de bookmarking mitotiques*, sont maintenu (B); tous les régulateurs sont expulsés (C). A et B sont des scénarios plus stables que C, tandis que B et C sont plus flexibles que A. Cela signifie que le cas B, représentant le *bookmarking mitotique*, montre les avantages de A et C. Adaptée de [23]

1.4.5 Transcription waves during re-activation of the cell cycle

In 2017, also Palozola and others [28] tried to unmask the mechanisms to maintain the cell identity during mitosis, and to uncover the reactivation hierarchy, i.e. which genes are priorly reactivated with respect to the others, during the mitotic exit and the early interphase. To do so, they quantified pulse-labeled nascent transcripts (EU-RNA-Seq experiments) in HUH7 human hepatoma cells, previously synchronized by inducing mitotic arrest with nocodazole. They defined a transcription timing by collecting data during the arrest, the mitotic-exit and in asynchronous cells. A summary of experiment and timing is proposed in fig. 1.14.

In this study, the authors highlighted the presence of low levels of transcription during mitosis and the fact that housekeeping genes and not cell-specific genes are activated earlier during mitotic exit. In fact, first to be reactivated were genes categorized as lumen/envelope in gene ontology (GO) nomenclature. Just after mitosis, genes involved in basic cell structure and in the cell-growth are found, followed by adhesion genes and, lastly, genes responsible for cell-cycle and DNA-replication, consistent with the fact that cells are ready to enter in S phase.

However, the study did not consider that mitotic-arrested cell populations progressively de-synchronize after the block release and therefore the reported measurements are performed on mixture of cells at different internal cell-cycle times. This point was addressed in our research, and it will be discussed in details in chapter 3.



	Factor	IF	Live	PCR	NGS	References
Transcription factors	FoxA1					Caravaca et al., 2013
	Ets1					Festuccia et al., 2016a
	Klf4					Liu et al., 2017
	Gata1					Kadauke et al., 2012
	Rbpj					Lake et al., 2014
	Runx2					Ali et al., 2010; Pockwinse et al., 2011; Young et al., 2007
	Myc					Yang et al., 2013
	Foxl1					Yang et al., 2014
	TLE1					Ali et al., 2010
	Pou5f1					Deluz et al., 2016; Liu et al., 2017; Teves et al., 2016
	Sox2					Deluz et al., 2016; Liu et al., 2017; Teves et al., 2016
	HSF2					Xing et al., 2005
	HNF1b					Lerner et al., 2016; Verdeguer et al., 2010
	Lef1					Kioussis et al., 2002
	Gcx2					Deluz et al., 2016
	Klf5					Deluz et al., 2016
	Rex1					Deluz et al., 2016
	Tbx3					Deluz et al., 2016
	Tcf3					Deluz et al., 2016
	HMGB1					Caravaca et al., 2013; Pallier et al., 2003
	HMGB2					Caravaca et al., 2013; Pallier et al., 2003
	HMGN1					Pallier et al., 2003
	Gata4					Caravaca et al., 2013
	C/EBP-a					Caravaca et al., 2013

Figure 1.12. Une liste des TF qui ont été observés liés aux chromosomes mitotiques. En haut, la technique expérimentale utilisée pour la détection est rapportée. IF et Live se réfèrent respectivement à l'immunofluorescence microscopique et à l'imagerie microscopique en direct. De plus, des études d'immunoprécipitation ont été prises en compte, provenant d'analyses PCR et NGS (via ChIP-Seq). Les candidats les plus robustes sont écrits en rouge. Cellule jaune: résultat débattu. Modifiée à partir de [23]

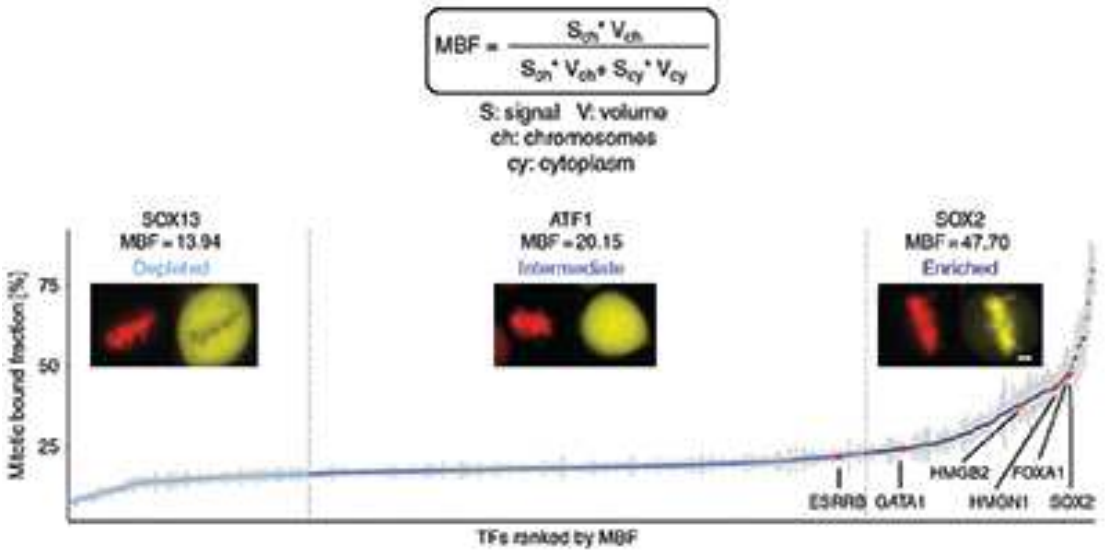


Figure 1.13. En haut, la définition mathématique de la fraction liée mitotique (MBF). En bas, le MBF pour 501 TF, regroupés en trois catégories: *depleted*, *intermediate* et *enriched*. Un exemple d'image de microscopie est montré pour chacune des catégories, le signal jaune représentant les TF et le signal rouge l'ADN. Le nom d'un intermédiaire et de certains TF enrichis a été signalé. Modifiée à partir de [32]

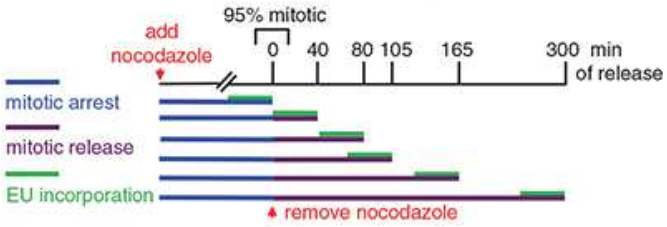


Figure 1.14. Schéma des expériences EU-RNA-Seq avec les points temporels analysés. Adaptée de [28]

1.5 3D structure and cell development

Here, we introduce two different studies proposing an analysis of chromatin structure by using Hi-C experimental approaches, in order to reveal the developmental stages where the formation of different organizational features occurs. Importantly, data provided by the studies presented hereafter, will be used in our analyses, in chapters 2 and 4.

In 2017, Hug and others [33] analyzed Hi-C data from early development of Drosophila melalogaster (Dmel) embryo.

The development of Dmel is characterized by a series of 13 nuclear cycles (nc), until the zygoyic embryo cellularizes and the zygotic genome activation (ZGA) takes place, in correspondence of nc14. Then, a new recruitment of RNA polymease II (RNA Pol II) on a large-scale occurs, with the activation of transcriptional activity. Analysis of Hi-C data from embryos at different developmental stages surrounding the zygotic genome activation, highlighted dramatic changes in the chromatin structure in correspondence of the ZGA: TADs establishment correlates with the activation of the gene expression, while before ZGA the genome does not present any relevant structure. Importantly, after a pharmacological inhibition of RNA Pol II, the establishment of TADs was not precluded, but only some contact properties of the domains and the co-localization of the boundaries with housekeeping gene-enriched regions were affected. This hints that the mechanism of TAD formation does not depend on transcription, but transcription is fundamental to maintain a proper organization of the chromatin.
The second study we mentioned above was published by Abramo and others in 2019 [34], where the authors used nocodazole induced mitotic arrest to synchronize HeLa cells in prometaphase. Then, the arrest was released and cells could start again progressing into the cell-cycle. To investigate the changes of the structural conformation of the chromatin in the cell nucleus between mitotic exit and G1 phase, fractions of the cells were collected at different time points between the arrest (time $t = 0$hours) and $t = 12$hours. Then, timing Hi-C analyses were performed. The authors showed that the telophase represents a critical transitional step between the mitosis and the interphase, from a conformational point of view. In fact, during mitosis, the chromosomes are folded into helically organized array of nested loops, that are mediated by condensin, a protein complex known to be fundamental for the segregation of the chromosomes during the mitosis [35]. However, these loops are lost by telophase, and a loops-free intermediate state is formed. By cytokinesis, TADs appear as well as new loops, mediated by the cohesin, a protein complex regulating the separation of sister chromatids during the division of the cell [36]. Boundaries of the compartments are established early too, while the long-range compartmentalization requires much more time, and continues for hours after cells entering in G1.

To sum up, there is a critical transition during telophase, where the mediation of the chromatin folding pass from the condensin to the cohesin.
Taken together, these results highlight once more the deep correlation between three-dimensional organization of the cell nucleus and the gene regulation, as discussed in the section 1.3.

1.6 Inferring the Gene Regulatory Networks

When a gene is expressed, its product can induce the activation of other genes with further expressions and products, with a chain of events which affect all the biological function of cell. The entire biological system is governed by regulators which interact with each other, such as DNA, RNA, TFs or other proteins. This interactions can be direct or indirect, i.e. by involving their expression product [37]. All these regulators and their reciprocal interactions form a complex network, called *gene regulatory network* (GRN, see fig. 1.15 for an example). The usage of experimental data on interactions between the regulators and the inference of their mutual relationship is fundamental to understand the regulatory mechanism and reconstruct and model the GRN [38]. A full comprehension of the GRNs is still very challenging, but the recent exper-

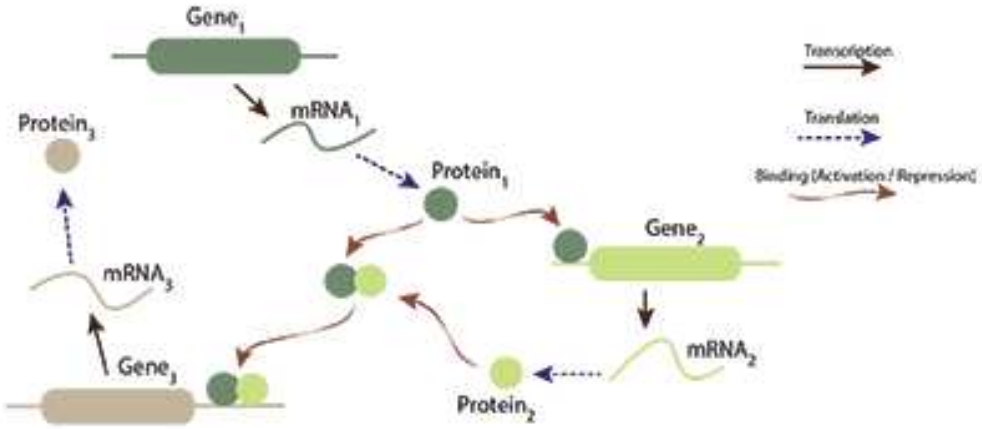


Figure 1.15. Un exemple illustré de réseau de régulation des gènes. De [39]

imental advance in the field of the high-throughput techniques increased the parallel understanding of the gene regulatory netwkork: ChiP-Seq, RNA-Seq, miRNA-Seq are nowdays used to uncover the reciprocal mechanisms governing the regulators and to investigate the dependencies between transcription factors and target genes [37]. Several computational methods have been developed in the last few years, using different mathematical and statistical approaches, such as information theory models, boolean networks, differential equations models, baysian networks and neural networks [40]. Each of them leverages on different assumptions and depicts distinct characteristic of the described gene regulatory network. What we will briefly described in the next section is a method which aims to identify key regulatory factors to infer the activity of the TFs.

1.6.1 Identify key regulatory factors to infer activity of TFs

In 2014, Balwierz and others [41] proposed ISMARA (*Integrated System for Motif Activity Response Analysis*), a method which allows to model the gene expression in terms of predictions of regulatory sites. ISMARA is a completely automatic computational tool, freely usable by uploading data on a website. The principle is that, given as an input genome-wide gene expression data in the form of RNA-Seq or microarray (miRNA) across different samples (e.g. different cell conditions), ISMARA identifies the key regulatory factors driving the expression. The method counts on two previously collected

datasets: an annotation of promoters in human and mouse, and a complete set of transcription factors binding sites (TFBSs), predicted for all the promoters by using a comparative genomic Bayesian approach.
ISMARA can be summarized by the following linear model:

$$E_{ps} = \sum_{m} N_{pm} \cdot A_{ms} + c_p + \tilde{c}_s \tag{1.6.1}$$

In the equation 1.6.1, E_{ps} are the data provided by the user, i.e. the signal associate to the promoter p in the sample s, in the form of RNA-Seq or microarray (miRNA). The matrix N_{pm} contains the information about the number of sites for the motif m in the promoter p. A_{ms} is the activity of the motif m at the sample (condition) s. Finally, c_p is the basal level for the promoter and $\tilde{c}_s$ is the basal level for the sample s.
To sum up, ISMARA is able to explain the signal E_{ps} by using the binding sites N_{pm} and the activities A_{ms}, that are inferred by the method and that represent the output of the tool. Moreover, the profiles of the activities are sorted according to their significance in the model, so that the most important motifs can be easily individuated.

A modified version of this linear model has been used in the context of our research, and will be presented in the chapter 3, with all the mathematical details.

2

A graph-based approach to detect domains in chromatin

As shown in the section 1.2.5, topologically associated domains (TADs) are one of the organizational levels of chromatin. Several algorithms have been developed to detect TADs, such as *Directionality Index, DI* [11] and *Arrowhead algorithm* [8]. However, a limitation of these approaches is that they detect domains at a certain length-scale, without revealing any higher or lower organizational level. Then, in the following years, other methods have been published to detect chromatin domains at different length-scale, revealing a hierarchical structure. In particular, in 2015, Fraser and colleagues [14] discovered the existence of *metaTADs*, a domains-within-domains structure covering all the possible genomic scales, from small sub-domains to the entire chromosome. Also, in 2017, Zhan and colleagues [13] proposed *CaTCH*, an algorithm identifying hierarchical trees of chromatin domains, starting from Hi-C data. However, the existence of preferential hierarchical levels and scales is still debated.
In this chapter, a computational method to dectect domains at all possible genomic scales will be presented. It leverages on Markov process and random walk theory, as well as on a network-based approach.

The algorithm, called *stability algorithm*, was originally developed to detect communities in graphs. Thus we modelled the chromatin as a network and applied the stability algorithm on two different published Hi-C datasets: in the first case, on Hi-C datasets obtained at different stages of Drosophila Melanogaster (Dmel) embryogenesis; in the second case, on Hi-C datasets obtained from time-dependent experiments on synchronized HeLa cells during mitotic exit, analyzing data from mitotic arrest up to 11 hours later, for a total of 16 timepoints.

By using the stability algorithm, we showed the presence of a *backbone* of the structure in Dmel even before the zygotic genome activation (ZGA), suggesting an early activation of transcription, although recent literature stated that chromatin architecture emerges with the onset of transcription activation in the zygote, while prior to ZGA the cell nucleus is mostly unstructured [33].

In addition, analysis on HeLa cells exiting mitosis revealed the presence of mitotic 3D structure that is conserved across the mitotic exit up to the latest available timepoints, i.e. several hours after the release of the mitotic block.

In the next section, we will briefly provide some fundamental concepts about Markov processes and random walks, that will allow a full comprehension of the principles behind the algorithm, described in the section 2.2. Then, our results will be presented in sections 2.3 and 2.4.

2.1 Markov processes and random walks

A Markov chain is a stochastic model that aims to describe a succession of events or states with a fundamental properties: the transition to the next state depends exclusively on the current state (*memoryless property* or *Markov property*). In other word, any prediction on the future event can be made knowing only the present state of the system. If the Markov chain takes place on a countable state space during time, it is called *Markov process*.
A *random walk* (RW) is a mathematical object describing the trajectory given by a sequence of random steps on a certain space. The simplest example of RW is the one taking place on the $\mathbb{Z}$ space of integers: at each step, only two events are possible, i.e. -1 and $+1$. When the mathematical space is a graph, the random walk is a special case of Markov chain [42]. This is the case which will be treated in the next section.

2.2 Community detection using the stability of a graph partition

Given a network characterized by a group of nodes linked by some connections (edges), a *community* is defined as a subgroup of nodes more densely connected if compared with nodes outside the community. Several computational and mathematical methods have been developed to detect communities in networks, i.e. to detect partitions. However, determining how much stable are the network partitions and measuring the quality of the found communities is still an open problem. In 2012, Delmotte and Schaub proposed an algorithm [43], freely available and written in Matlab. It is based on previous theoretical studies, such as [44] and [45], and aims to detect communities in networks by defining a measure for the *stability of a graph partition*, leveraging on the properties of a random walk. A brief recup of the mathematical concepts behind the algorithm will be presented hereafter, while for further details we refer to the cited literature.
If we have a graph G whose edges $A_{i,j}$ between each pair of nodes i and j are described by the matrix A, we can write the standard dynamics of a RW as follows:

$$P(t) = e^{-Mt} P_0 \tag{2.2.1}$$

where $P(t)$ is the probability vector of being in a state at time t, and M is the matrix A normalized on the total number of degrees, so that $M_{i,j} = \frac{A_{i,j}}{\sum_i A_{i,j}}$. It can be shown that such a dynamics converges to a stationary distribution $\pi = \frac{d^t}{2m}$, with $d = A\mathbf{1}$, i.e. the vector of the degrees, and $2m \equiv \mathbf{1}^T d$, i.e. the sum over all degrees.

If $N \times N$ is the size of the matrix A, we can introduce the presence of c communities by using the matrix H, with a size $N \times c$, such that $H_{i,j} \in \{0, 1\}$, i.e. 1 if the node i belongs to the cluster j, 0 otherwise.

Moreover, by introducing the term $\Pi = diag(\pi)$, we can define the *clustered autocovariance* of the process at a time t as follows:

$$R(t, H) = H^T [\Pi P(t) - \pi^T \pi] H \tag{2.2.2}$$

Here, $R(t, H)$ is a matrix with size $c \times c$, whose diagonal elements provides the propensity of a random walk of remaining in the starting community at time t. Then,

we can define the *stability of the partition* H as:

$$r(t, H) = \text{trace}R(t, H) \tag{2.2.3}$$

Thus, the optimization of 2.2.3 provides a list of optimal clusterings, with coarser communities as the time increases, i.e. with the time playing the role of resolution parameter. Furthermore, it can be mathematically demonstrated that 2.2.3 can be rewritten as the *modularity* of a network $G(t)$ that depends on time, where the modularity is a quality function which compares the density of edges inside a community to edges between communities, in order to find the best partition of a weighted graph [46]. This last property makes possibile the use of the *Louvain method* [47] for the modularity optimization, that is implemented in the stability algorithm.

We used this computational method to detect domains in the chromatin. To do so, we modelled chromatin as a network where nodes are chromatin regions and edges represent physical contacts between regions determined by Hi-C data. The detected communities were then called *diffusion associated domains (DADs)*. As shown, with this approach time represents a parameter which determines the resolution of detected domains, identifying different hierarchical levels of the chromatin structure.

2.3 Identifying DADs in Drosophila Melanogaster embryo development

We applied the stability algorithm systematically to *in situ* Hi-C data from [33], already presented in section 1.5 (public database *ArrayExpress: E-MTAB-4918*). We did that for developmental stages *nuclear cycle 12* (nc12), *nuclear cycle 13* (nc13), *nuclear cycle 14* (nc14), $3 - 4$ *hours post fertilization* (3-4hpf) and for *mitosis*, with a resolution of $10kb$. Fastq files were downloaded and and alligned by using *bowtie2* [49] on the reference genome *Dmel r0.07*. Files were then processed by using the pipeline *HiCUP 0.5.9* [48]. A genome digest file was produced by using *hicup digester* and the MboI digestion enzyme sequence. Following parameters were set in *HiCUP 0.5.9*: *Threads: 8*, *Quiet: 0*, *Keep: 0*, *Zip: 1*, *Longest: 800*, *Shortes: 150*. After processing, lowest 5-percentile of the Hi-C data were exluded to correct for the background noise, and Hi-C data were normalized by using ICE normalization [50], through the *iced* library in python [51].

We obtained the dynamics of the diffusion associated domains (DADs) for all chromosomes, by setting the following parameters in the Matlab code of the stability algorithm: *Linearised Stability*, *Normalised laplacian*, *Verbose mode: Yes*, *Number of Louvain iterations: 100*, *Precision used:* $1e - 09$. 54 time values from an exponential distribution were chosen, from $10^{-3.3}$ to 10^2, with an exponential step of 0.1. Regions without Hi-C signal were excluded from simulation.

As a result, given a Hi-C dataset, for each diffusion time (that we will call timestep from now on) we have a list of n indices to assign to the regions/nodes of the chromatin/network, indicating the community (DAD) they belong to. This allows to build a temporal dynamic of DADs formation, corresponding to a hierarchical structure at different size-scales, from the smallest one (at timestep $t = 0$ the number of regions corresponds to the number of DADs) to the biggest one (at the end, all the regions belong to the same community, taking the entire chromosome). Note that there is no constraint for contiguity of the DADs, so we can obtain DADs formed by non contiguous chromatin regions.

DADs were represented by using the Matlab library *imagesc*. As a comparison, Hi-C maps were generated by using the python function *heatmaps* from the library *seaborn*. Results for chromosomes 2, 3, 4 and X are shown, respectively, in the figures 2.1, 2.2, 2.3, 2.4. These results have to be commented in the light of what Hug and colleagues stated in their article [33], i.e. that the emergence of the three-dimensional structure of the chromatin occurs in correspondence with the onset of the zygotic genome activation (ZGA), around the nuclear cycle 14, while in the earlier stages the chromatin is mostly unstructured. Also, by a visual inspection of the normalized Hi-C maps at 10 kilobases of resolution, the authors of the cited study affirmed that there is a noticeable difference in the 3D structure of the chromatin up to nuclear cycle 13, if compared with the later developmental stages. In contrast, a visual inspection of detected DADs, seemed to reveal that an organized structure of the chromatin exists in the earlier stages of the embryogenesis. Besides, also mitotic chromosomes show well defined DADs if compared with the other analyzed stages, suggesting that a hierarchical organization of the chromatin makes an appearance during mitosis.

To quantify these observations and to compare the DADs-like structure of chromosomes at different stages, we used the Adjusted Rand Index (ARI), a version of the Rand Index [52] adjusted for a chance-correction. The Rand Index R masures the similarity between different clusterings or partitions, by assigning a value between 0 and 1.

Quantitatively: let us consider a set of n nodes $S = \{s_1, s_2, ...s_n\}$ and two partitions of S, namely $P_1 = \{X_1, X_2, ..., X_r\}$ and $P_2 = \{Y_1, Y_2, ..., Y_s\}$, containing, respectively, r and s subsets of S. The Rand Index R is then given by the following equation:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} \tag{2.3.1}$$

where a is the number of pairs of nodes in S that are in the same subset both in P_1 and in P_2, b the number of pairs of nodes in S that are in different subsets both in P_1 and in P_2, c is the number of pairs of nodes in S that are in the same subset in P_1 but in a different subset in P_2 and d is the number of pairs of nodes in S that are a different subset in P_1 but in the same subset in P_2. As mentioned above, the Adjusted Rand Index is adjusted for a chance-correction. We have then, in general:

$$ARI = \frac{R - \text{expected(R)}}{\max(\text{R}) - \text{expected(R)}} \tag{2.3.2}$$

In principle, different models can be used to calculate an expected R. Traditionally, the ARI is obtained with a permutation model for clusterings, i.e. the size and the number of the clusters of a given partition are fixed, and random clusterings are obtained by shuffling of the elements among the clusters. For more mathematical details, we refer to [53].

To compare DADs-based structure of different developmental stages at any diffusion time, we used the python function *adjusted_rand_score* from the library *sklearn.metrics.cluster*. We used the developmental stages *3-4hpf* as a reference, as it is the latest one among the analyzed datasets. The DADs-like structure of all the other stages was then compared with the *3-4hpf* structure, by calculating the ARI. The results are shown in the figure 2.5. In addition, an average over the simulation timesteps was calculated for every chromosome and over all chromosomes. The results are reported in figure 2.6. Even if the ARI score of mitotic stage was always the lowest one, it was still significant. In fact, we obtained a mitotic ARI of 0.68, 0.63, 0.64, 0.69, 0.68 and 0.34

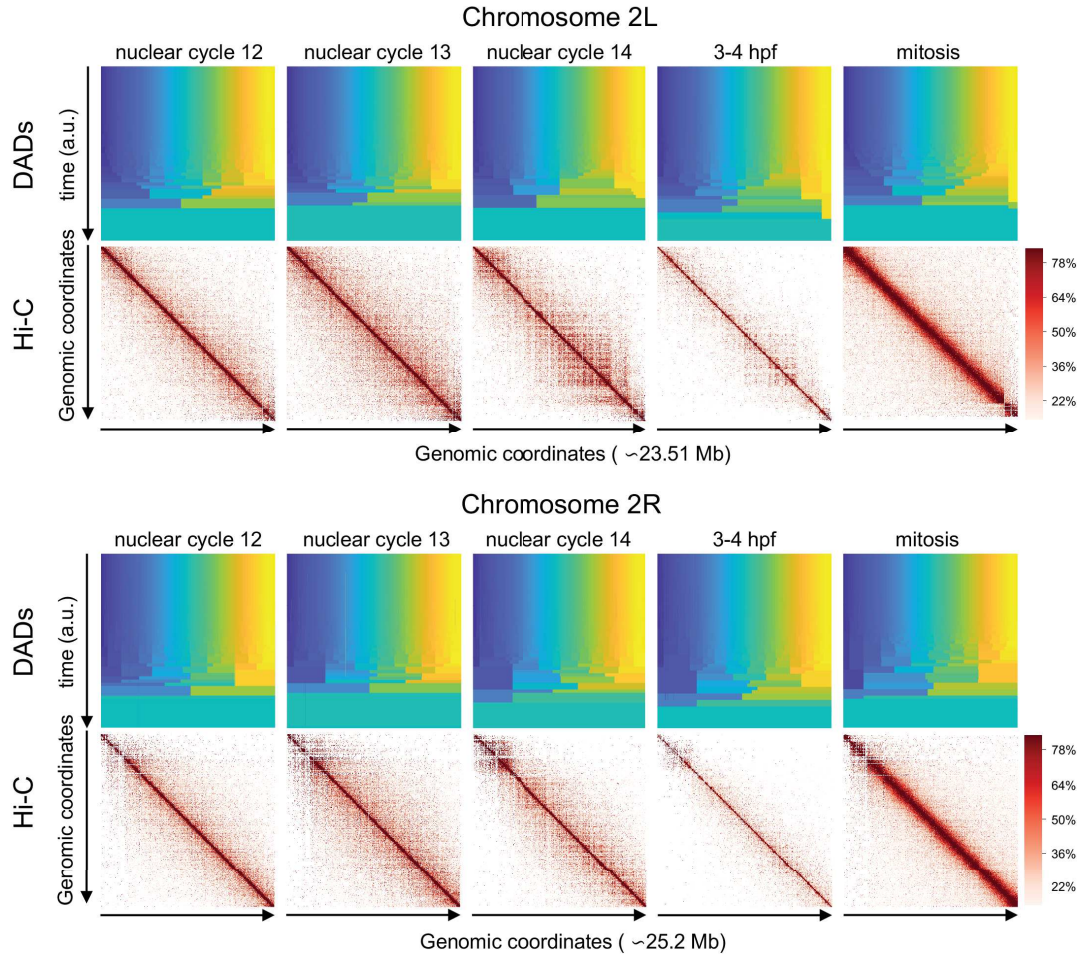


Figure 2.1. Diffusion associate domains (DADs) and Hi-C maps for Dmel Chromosome 2L (left arm, top) and 2R (right arm, bottom). DADs were represented assigning different colours to different communities detected over the genomic coordinates (horizontal axis), over the simulation timesteps progression of the algorithm (vertical axis). For the first timesteps, each region corresponds to a community. Then, as the timesteps increase, coarser communities are detected by the stability algorithm, until a saturation is reached, i.e. the entire network becomes one single community, represented by one single color. Hi-C heatmaps are reported as a reference, with the colorbar on the bottom right providing the percentage with respect to the maximum value of the Hi-C matrix.

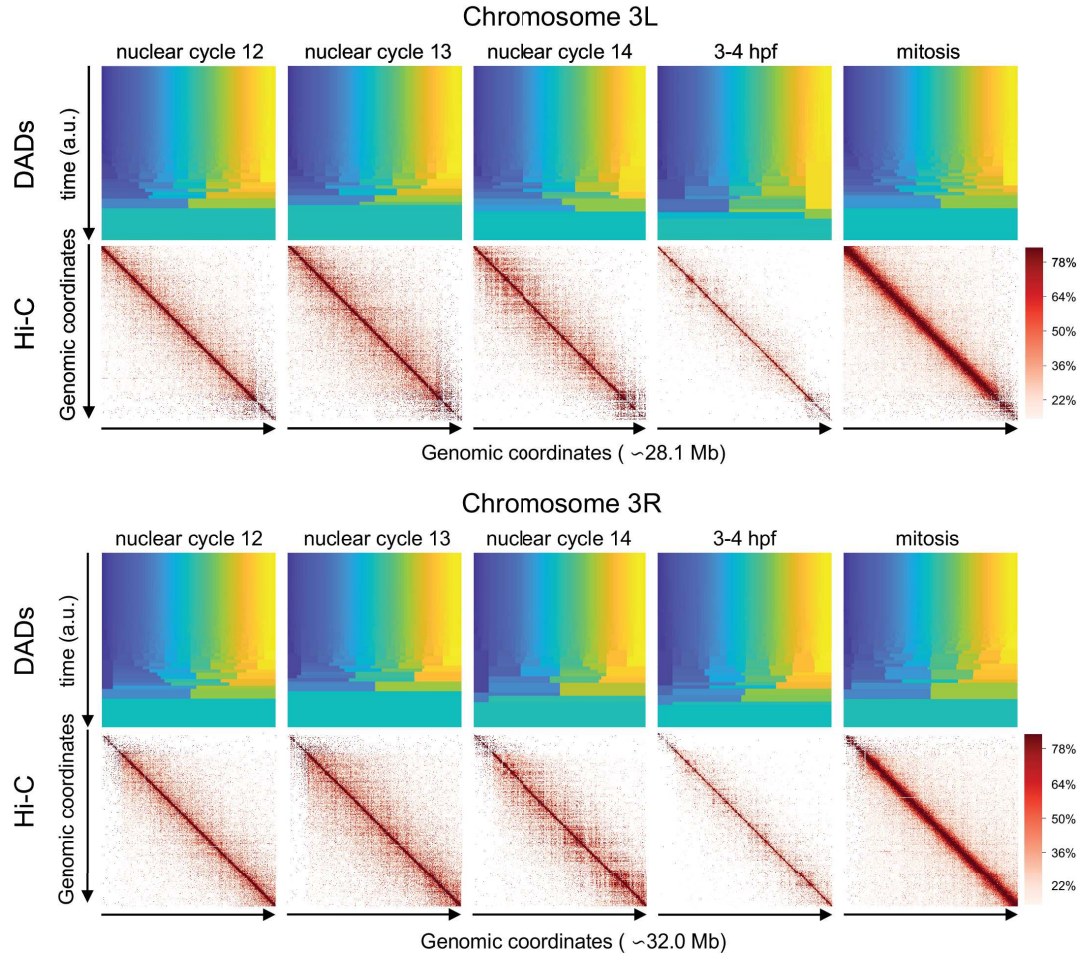


Figure 2.2. Diffusion Associate Domains (DADs) and Hi-C maps for Dmel Chromosome 3L (left arm, top) and 3R (right arm, bottom). See the description in figure 2.1

.

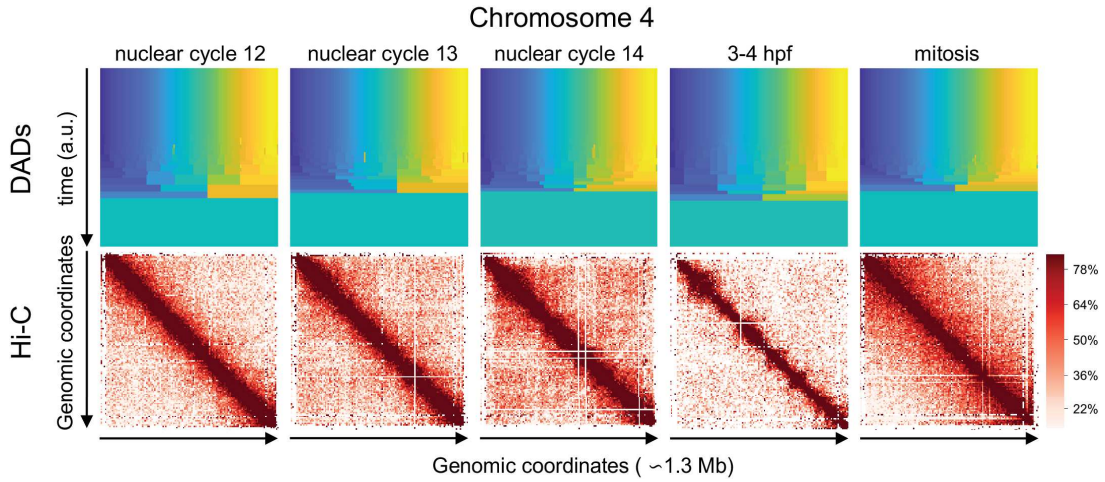


Figure 2.3. Diffusion Associate Domains (DADs) and Hi-C maps for Dmel Chromosome 4. See the description in figure 2.1

.

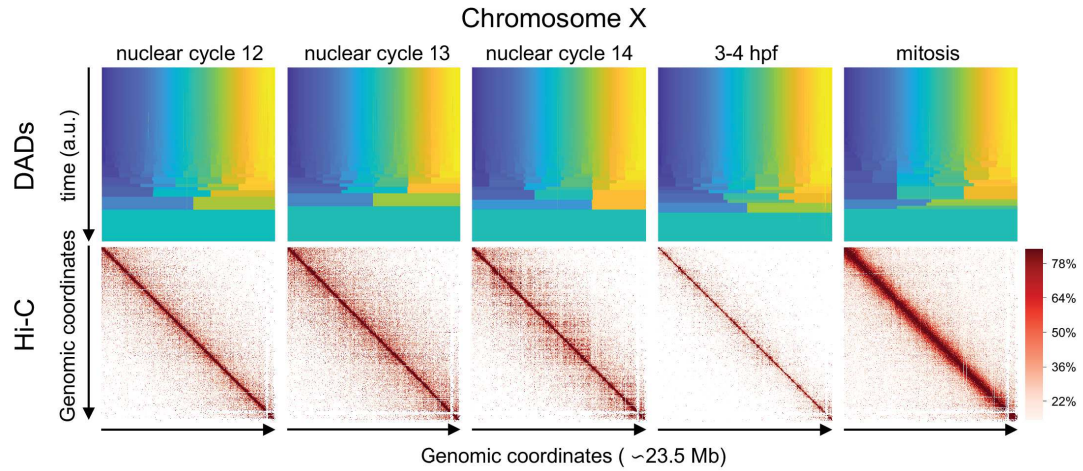


Figure 2.4. Diffusion Associate Domains (DADs) and Hi-C maps for Dmel Chromosome X. See the description in figure 2.1

for, respectively, chromosomes 2L, 2R, 3L, 3R, X and 4.
These results may sugges that Dmel chromatin maintains a certain memory of its organizational structure through the embryonic development, even during the earliest stages. In fact, taken DADs detected in cells at 3-4 hpf as a reference, similarities from 63% to 80% were obtained in comparison with DADs detected at earlier developmental stages. Lower results were obtained only for chromosome 4, with similarities going from 34% to 55%. Furthermore, 68% was obtained by averaging the mitotic similarity over all chromosomes, excluding the chromosome 4 (which instead shows a mitotic similarity score of 34%). These results highlight the presence of a *backbone* of the structure even in the mitotic state and, in the light of the strict relationship between the three-dimensional organization of the chromatin and the transcriptional activity, may suggest an early activated transcription.

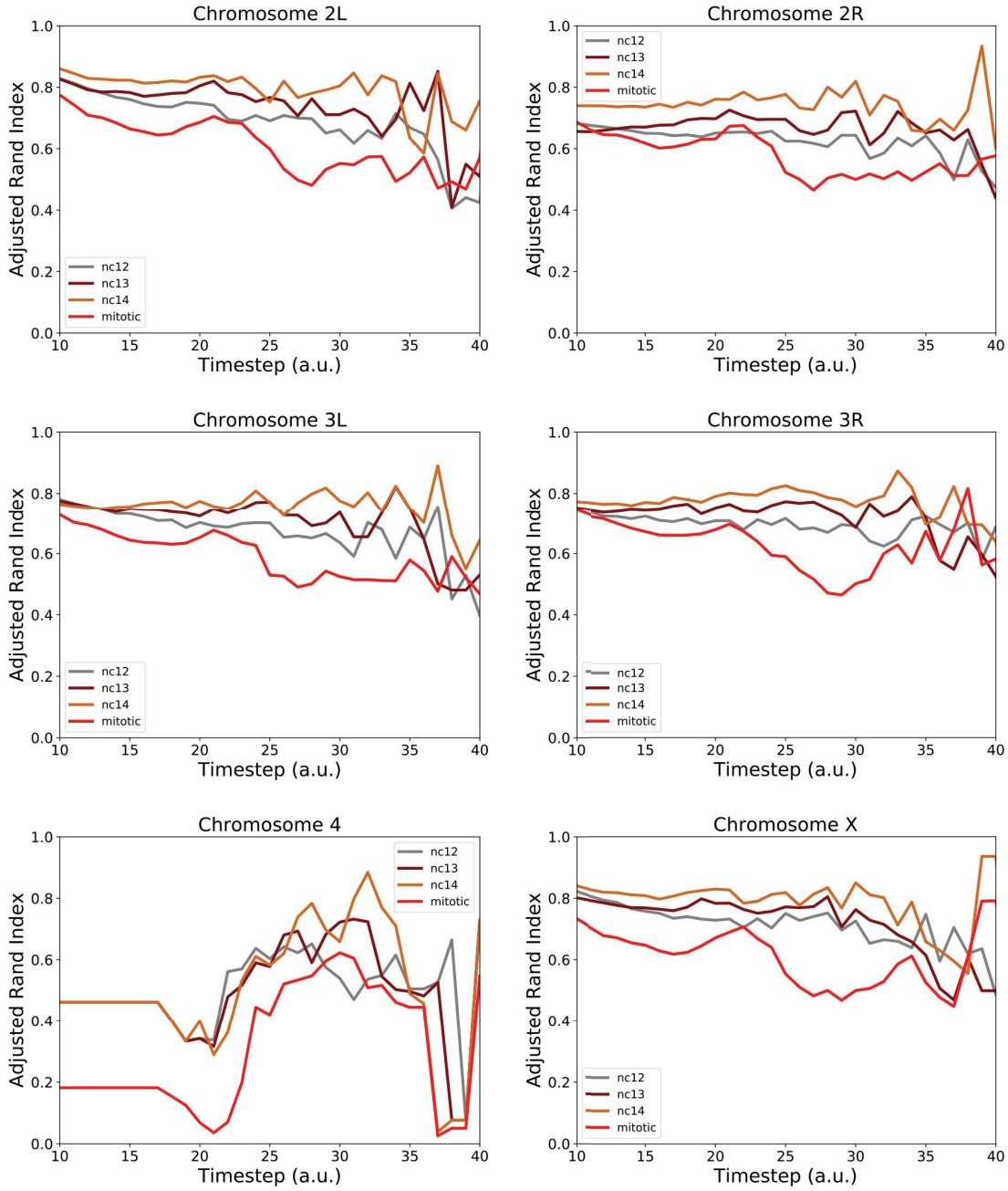


Figure 2.5. ARI scores for Dmel DADs-based structure detected at different developmental stages with respect to 3-4 hours post fertilization. The ARI score versus simulation timesteps of stability algorithm was calculated to compare the partitions generated by the DADs detected at 3-4 hours after the fertilization with all the earlier stages: nuclear cycle 12 (nc12), nuclear cycle 13 (nc13), nuclear cycle 14 (nc14) and mitosis.

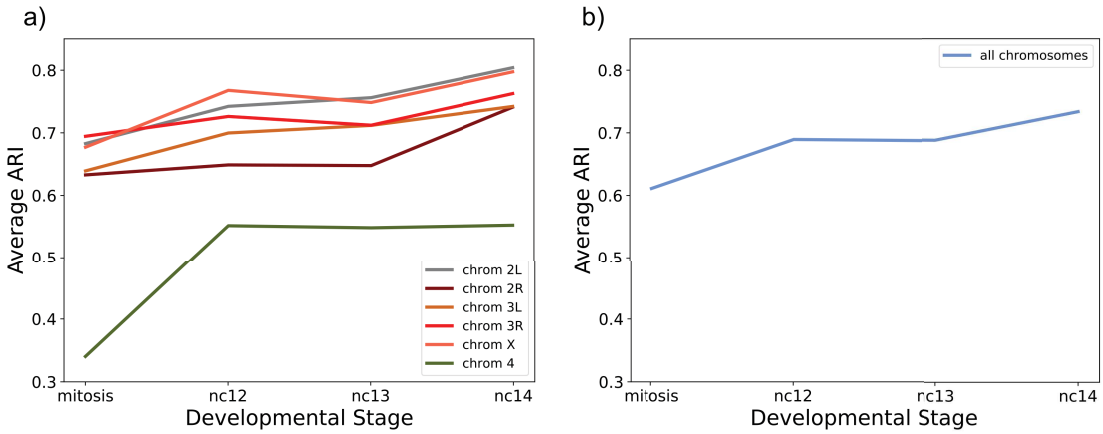


Figure 2.6. Average ARI score for Dmel DADs-based structure. **a)** The average ARI score per chromosome was calculated over the simulation timesteps of the algorithm. Each colour corresponds to the chromosome indicated by the legend. **b)** The ARI score over all chromosomes was calculated by averaging the curves shown in panel a.

2.4 Identifying DADs in HeLa cells during mitotic exit

In 2019, Abramo and others [34] published time-dependent Hi-C data on synchronized HeLa cells during mitotic exit, that we already presented in section 1.5 (GEO accession: GSM3909714). The authors of the study provided Hi-C datasets at the time of the release from mitotic arrest, up to some hours later, for a total of 16 timepoints. We downloaded Hi-C data in *cool* format and accessed them by using the *cooler* package of *python*. Lowest 5-percentile of data were excluded, in order to correct for the background noise. Hi-C data were then normalized by using ICE normalization [50], through the *iced* library in python [51].

The stability algorithm was applied after setting all the parameters as shown in the section 2.3. As well DADs and Hi-C heatmaps were represented by using the same tools and packages shown in section 2.3.

As an example, in figure 2.7 results are shown for the chromosome 10, for 15 analysed timepoints. In analogy with what has been shown in the previous section, we wanted to compare the structure of the earliest experimental timepoints exiting mitosis with the latest ones. To do that, we compared the DADs-like structure of chromosomes at different timepoints, by using the Adjusted Rand Index (ARI) seen in equation 2.3.2. We used the timepoint 11hours as a reference, and the DADs-like structure of all the previous timepoints was then compared with that reference structure. As an expamle, results for 4 chromosomes are reported in figure 2.8. In addition, an average over the simulation timesteps was calculated for every chromosome and over all chromosomes. The results are reported in figure 2.9. Finally, we wondered which length scales of the hierarchical organization were earlier reformed exiting mitosis. To do that, for every timestep T of the algorithm, we took the set s of DADs detected at T, and calculated the ARI over the experimental timepoints averaged over all chromosomes, normalizing all the values between 0 and 1. Results are shown in figure 2.10.

Then, experimental timepoints corresponding to first reaching of 0.5 (half of their height) for every timestep were collected. Results are shown in figure 2.11.

Noticeably, a local minimum is visible, correspondig to resolution timesteps 23,24 and 25. These timesteps correspond to an average length of $0.95Mb$, $1.30Mb$ and $1.75Mb$ respectively, suggesting that organizational structures at these length scales are the ones that reform faster with respect to the others.

In conclusion, by detecting sistematically DADs on time-dependent Hi-C datasets from HeLa cells, we showed that chromatin maintains a certain memory of its organizational structure during the mitotic exit, showing a high percentage of structure that is kept from mitosis to the latest experimental timepoints. Quantitatively, taken DADs detected at experimental timepoint 11hours as a reference, similarities from 40% to 85% were obtained in comparison with DADs detected at earlier timepoints. Furthermore, 66% was obtained by averaging the mitotic similarity over all chromosomes. Again, these results highlight the presence of a *backbone* of the structure even in mitosis and, in the light of the strict relationship between the three-dimensional organization of the chromatin and the transcriptional activity, may suggest the possibility that transcription occurs during mitosis and early interphase.

A further investigation on transcription reactivation during and exiting mitosis will be shown in the next chapter.

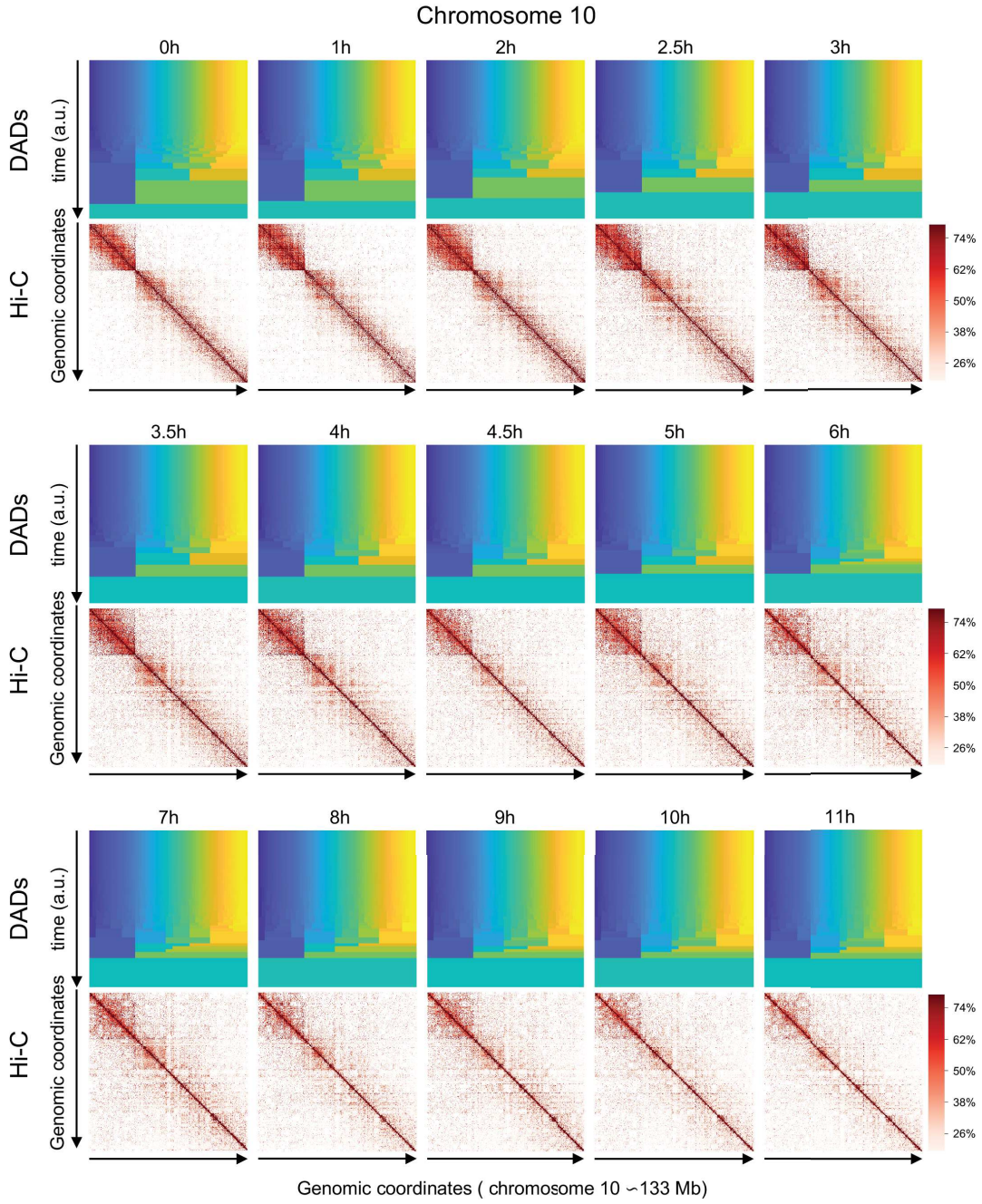


Figure 2.7. Diffusion associate domains (DADs) and Hi-C maps for HeLa Chromosome 10. DADs were represented assigning different colours to different communities detected over the genomic coordinates (horizontal axis), over the simulation timesteps progression of the algorithm (vertical axis). For the first timesteps, each region corresponds to a community. Then, as the timestep increase, coarser communities are detected by the stability algorithm, until a saturation is reached, i.e. the entire network becomes one single community, represented by one single color. Hi-C heatmaps are reported as a reference, with the colorbar on the bottom right providing the percentage with respect to the maximum value of the Hi-C matrix.

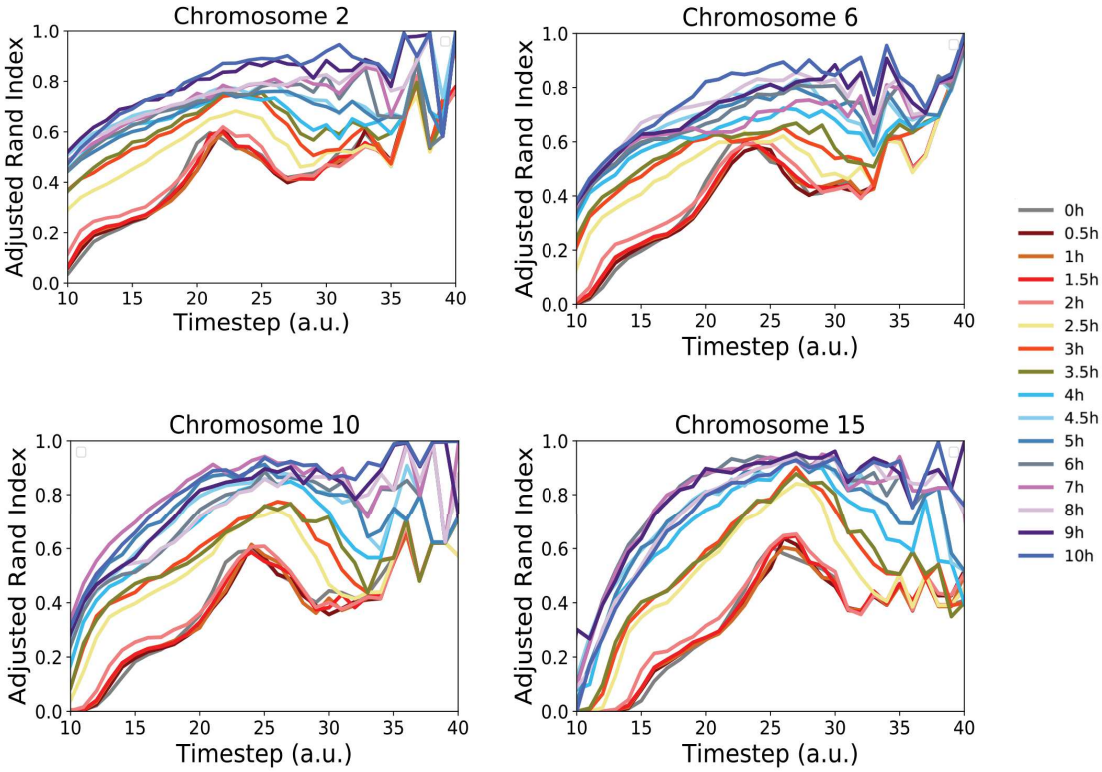


Figure 2.8. ARI scores for Hela DADs-based structure at different experimenta timepoints with respect to 11 hours after the release of synchronization. The ARI score versus simulation timesteps of stability algorithm was calculated to compare the partitions generated by DADs detected at 11 hours after the release of the synchronization with all the earlier timepoints, from 0 hours to 10 hours, as indicated by the legend on the right.

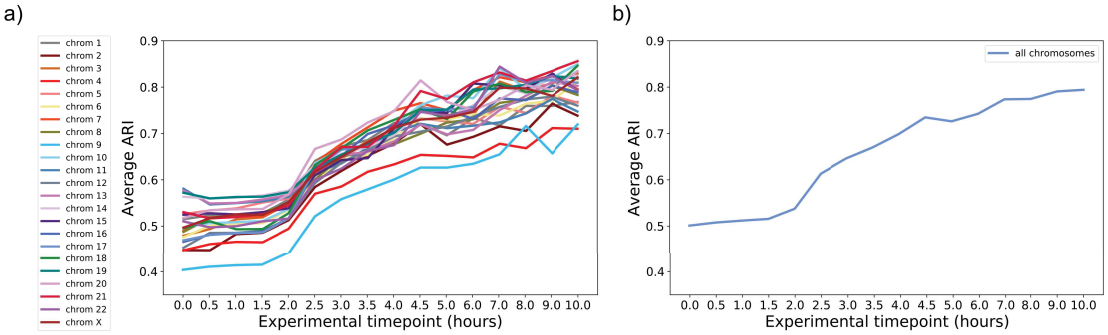


Figure 2.9. Average ARI score for HeLa DADs-based structure. **a)** The average ARI score per chromosome was calculated over the simulation timesteps of the algorithm. Each colour corresponds to the chromosome indicated by the legend on the left. **b)** The ARI score over all chromosomes was calculated by averaging the curves shown in panel a.

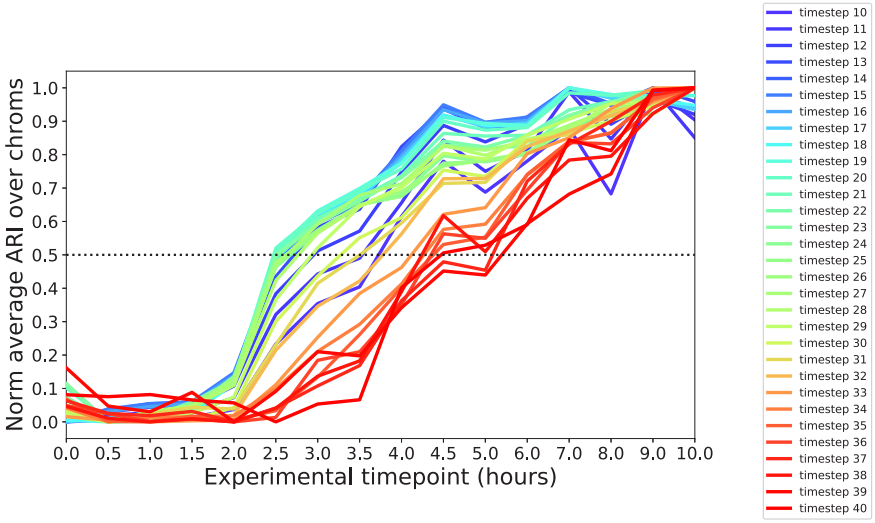


Figure 2.10. ARI in function of experimental timepoints for different timesteps of stability algorithm. ARI has been calculated over all experimental timepoints, and an average over all chromosomes has been considered here. ARI values have been normalized between 0 and 1. Each curve corresponds to the algorithm timestep indicated in the legend on the right. The horizontal, dotted line has corresponds to 0.5, in order to highlights which resolution reaches first the half of its height.

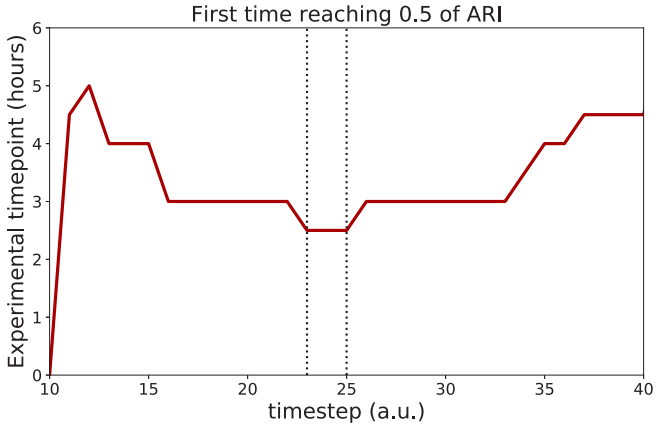


Figure 2.11. Differen velocity in structure reformation at different length scales. Experimental timepoints corresponding to first reaching of 0.5 (i.e. half of their height) were collected and plotted in function of the timestep of the algorithm. Dotted vertical lines highlight local minimum range, between timesteps 23 and 25.

3

Regulation of transcription reactivation kinetics exiting mitosis

Premise: this chapter has the form of a stand-alone study. In fact, it has been submitted as an independent article and is now under review. Only font and numerical continuity of references, equations and figures has been changed here with respect to the submitted work, in order to make everything consistent with the rest of the thesis.

Authors

Sergio Sarnataro, Andrea Riba, Nacho Molina

3.1 Abstract

Proliferating cells experience a global reduction of transcription during mitosis, yet their cell identity is maintained and regulatory information is propagated from mother to daughter cells. Mitotic bookmarking by transcription factors has been proposed as a potential mechanism to ensure the reactivation of transcription at the proper set of genes exiting mitosis. Recently, mitotic transcription and waves of transcription reactivation have been observed in synchronized populations of human hepatoma cells. However, the study did not consider that mitotic-arrested cell populations progressively desynchronize leading to measurements of gene expression on a mixture of cells at different internal cell-cycle times. Moreover, it is not well understood yet what is the precise role of mitotic bookmarking on mitotic transcription as well as on the transcription reactivation waves. Ultimately, the core gene regulatory network driving the precise transcription reactivation dynamics remains to be identified. To address these questions, we developed a mathematical model to correct for the progressive desynchronization of cells and estimate gene expression dynamics with respect to a cell-cycle pseudotime. Furthermore, we used a multiple linear regression model to infer transcription factor activity dynamics. Our analysis allows us to characterize waves of transcription factor activities exiting mitosis and identify a core gene regulatory network responsible of the transcription reactivation dynamics. Moreover, we identified more than 60 transcription factors that are highly active during mitosis and represent new candidates of mitotic bookmarking factors which could represent relevant therapeutic targets to control cell proliferation.

3.2 Introduction

Proliferating cells show a global downregulation of transcription during mitosis. This results from the combination of three main processes: 1) nuclear envelope breakdown leading to an increase of the volume that transcription factors (TFs) and the RNA polymerases II (RNAPII) can explore and therefore a decrease of their local concentration around gene promoters; 2) major reorganization of chromatin architecture characterized by chromosome condensation, repositioning of nucleosomes in some regulatory regions, loss of long-range interaction between enhancers and promoters and disassembling of topological associated domains (TADs); and, 3) TF-DNA binding inactivation through postranscriptionally regulated phosphorylation. As a consequence, most TFs and the RNAPII are evicted from mitotic chromosomes and RNA synthesis is drastically reduced [60].

In spite of this global decrease of gene expression during mitosis, proliferating cells are able to maintain their cell identity and propagate regulatory transcriptional programs from mother to daughter cells [56]. Mitotic bookmarking has been proposed as a potential mechanism that could be involved in the transmission of regulatory information during the cell-cycle [23]. Indeed, a significant fraction of TFs are able to remain bound to chromatin during mitosis [32]. These mitotic-bound factors (MFs) show faster interactions with mitotic chromatin than in interphase as reduced residence times have been reported. It is believed that non-specific chromatin or protein-protein interactions between MFs and chromosome coating proteins can explain this fast observed dynamics [32, 31]. However, it has been shown for a handful of MFs, known as bookmarking factors (BFs) [30, 54, 55, 64], their ability to interact specifically with at least a fraction of their interphase target sites during mitosis, indicating that chromosomes are not as compacted as previously thought [55]. In fact, chromatin accessibility and nucleosomes landscape during mitosis remain unchanged on bookmarked regions bound by known BFs [57, 58]. This ability of BFs to maintain chromatin structure locally could promote a quick transcription reactivation exiting mitosis.

Transcription dynamics during mitosis and early G1 phase has recently been measured by metabolic labeling of RNA (EU-RNA-Seq) in synchronized population of Human Hepatoma cells HUH7[28]. Remarkably, this study showed a low but detectable transcription activity during mitosis in up to 8000 genes. Furthermore, transcription reactivation occurred in intense waves exiting mitosis and early G1 phase. However, the study did not take into account that mitotic-arrested cell populations progressively desynchronized once the block was released. As a consequence, RNA measurements are performed on mixture of cells at different internal cell-cycle times. Moreover, it is not understood yet what is the precise role of mitotic bookmarking on mitotic transcription and the transcription reactivation waves. Ultimately, the core gene regulatory network driving the precise transcription reactivation dynamics remains to be identified.

In this paper we developed mathematical models and computational methods to address these open questions. First, in order to correct for the progressive desynchronization of cell populations we assumed that there is a stochastic lag time until a cell can restart the cell-cycle progression again. We characterized the distribution of lag times by analyzing how the observed fraction of mitotic cells evolves over time after the mitotic block is released. This allows us to deconvolve the EU-RNA-Seq data and produce gene expression profiles with respect to a cell-cycle pseudotime and classify the different waves of transcription reactivation in relationship with the cell-cycle progression

instead of the experimental time. Moreover, we identified the key TFs determining the transcription reactivation dynamics. To do that, we developed an ISMARA-like model [41] assuming that the expression of genes at a given time point of the cell-cycle progression is a linear combination of the activities of all the TFs that can bind on their promoters. By knowing the deconvolved gene expression and integrating data on transcription factors motif affinities, we calculated the activity of every expressed TF and its role in transcription reactivation exiting mitosis. Indeed, this analysis allows us to divide TFs in groups according to their peak of activity with respect to the cell-cycle pseudotime and identify a core regulatory network of TFs responsible of the observed transcription waves. Interestingly, we do not see a strong correlation between known BFs as FOXA1 and the speed at which their target genes are reactivated. However, we identified around 60 TFs that are highly active during mitosis and represent new candidates of mitotic bookmarking factors.

3.3 Results

3.3.1 Deconvolution of gene expression data from desynchronized cell populations

In 2017, Palozola et al. published a study based on metabolic labeling of RNA (EU-RNA-Seq) of prometaphase synchronized population of Human Hepatoma cells (HUH7) by arresting cell-cycle progression [28] with nocodazole. EU-RNA-Seq experiments were performed to measure newly synthesized transcripts at 0 minutes, 40 minutes, 80 minutes, 105 minutes, 165 minutes and 300 minutes after mitotic block release as well as for an asyncronous cell population. In this study, the authors highlighted the presence of low levels of transcription during mitosis and the fact that housekeeping genes and not cell-specific genes are activated earlier during the mitotic exit. We reanalyzed the EU-RNA-Seq datasets and characterized the expression dynamics at the gene level. By performing k-means clustering on the gene expression profiles, we identified 5 different clusters, presenting diverse transcription reactivation dynamics over the experimental time. Similarly as the authors reported one of these group showed a peak in the expression at 40 minutes, while others showed a later transcription reactivation (see Fig. 3.1, panel a).

Notably, the study did not consider that mitotic-arrested cell populations progressively desynchronize after washing out nocodazole and therefore the reported measurements are performed on mixture of cells at different internal cell-cycle times. In addition, at every experimental time point there is contamination from cells that escape mitotic block. We developed a mathematical model to correct for the desynchronization and the contamination of non-synchronized cells. To do so, we assumed that after mitotic block release there is a stochastic lag time until cells can start again the cell-cycle progression that is log-normal distributed with a certain mean μ and standard deviation σ. We introduced the concept of *internal cell-cycle pseudotime* τ, defined as the effective cell-cycle time progression of a cell, starting once the lag time is over. We then assumed that there is an average time τ_{mit} that cells need to complete mitosis. Finally, we fitted the parameters of the model τ_{mit}, μ and σ using data from cell imaging reporting how the fraction of observed mitotic cells evolves over time after the mitotic block is released [28]. This led to an estimated median lag time of 30 minutes and an average time to complete mitosis of 67 minutes (see Fig. 3.1, panel b and c and Methods

for a detailed mathematical derivation).

By applying our model, we deconvolved the time-dependent EU-RNA-Seq data and mapped them onto the internal cell-cycle pseudotime τ. As a result we obtained gene expression dynamics with respect to the cell-cycle progression, allowing us to highlight the transition between mitosis and early G1 phase. Again, we identified 5 different clusters of genes showing distinct transcription reactivation dynamics over the cell-cycle pseudotime τ. Strikingly, around 2000 genes showed an expression wave very early during mitosis, presumably around metaphase, while a large fraction of genes reach their reactivation peak just before exiting mitosis, during telophase or during the transition to early G1 phase, as shown in Fig. 3.1, panel d. In summary, our analysis allows us to correct for desynchronization of cell populations and study gene expression dynamics with respect to the cell-cycle pseudotime highlighting the waves of transcription in relationship with the transition between mitosis and interphase.

3.3.2 Transcription factor activity dynamics during mitosis and early G1 phase

The transcription waves identified in the previous section are driven by regulatory transcriptional programs mainly activated by transcription factors (TFs). To understand which ones among all TFs are in fact the principal drivers of the transcription reactivation dynamics, we developed an ISMARA-like approach [41]. Thus, we assumed that the normalized log-transformed expression $e_{g\tau}$ of a gene g at cell-cycle pseudotime τ can be obtained as a linear combination of the cell-cycle dependent activities $A_{f\tau}$ of all TFs f that can potentially regulate the gene. The model can be summarized by the following equation:

$$e_{g\tau} = \sum_{f} N_{gf} A_{f\tau} \tag{3.3.1}$$

where the values N_{gf} represent the entries of a matrix $\mathbf{N}$ containing the number of binding sites for the TF f associated with promoter of the gene g, taking into account the affinity between the motif of f and the sequence of the gene promoter [59]. From the analysis, we excluded TFs associated to unexpressed genes. Furthermore, to avoid overfitting we introduced a regularization term that enforces smooth TF activities over the cell-cycle time and we calibrated using a cross-validation approach. For further mathematical details we refer to Methods. Our analysis allows us to infer the activity of 332 TFs. This can be understood as a dimensionality reduction approach as we describe the problem of transcription reactivation with much fewer parameters, since we pass from the analysis of thousands of genes to only hundreds of TFs (see Fig. 3.2, panel a). To analyze the activity dynamics we divide them in 3 clusters, according to their profile over τ (see Methods). We showed that almost 19% of TFs present positive activity during mitosis, with a peak in the first minutes, and then progressive decrease of activity. Conversely, 36% of TFs present a negative activity during mitosis, and then a high activity in early G1. Lastly, the remaining 45% of TFs show a moderate amplitude in their dynamics suggesting that they play a minor role on transcription reactivation dynamics (the results are shown in Fig. 3.2, panel b). Among TFs that are active during mitosis, we obtain known bookmarking factors as C/EBP, HSF1, TBP, GATA1 and ESRRβ [29, 28, 61, 62, 23] reassuring that our approach is able to identify relevant TFs. Indeed, activities of TFs that are annotated to the Gene Ontology category cell-cycle show an intense dynamics during mitosis and early G1 phase (see supp. Fig. 3.S4). In-

terestingly, by sorting TFs according to when their highest peak of activity occurs, we observed waves of activity suggesting an intrinsic TF hierarchy with respect to their role on the temporal reactivation of transcription after mitosis (see Fig. 3.2 panel b).

Determining the molecular mechanisms underlying the TF activity dynamics that we inferred goes beyond the scope of this study. However, we can have a first clue by analyzing the correlation between the TF activity and the expression of the corresponding TF gene. Indeed, a strong correlation indicates that changes in transcription may be responsible for changes in activity. On the contrary, low correlations may suggest that postranscriptional regulation is required to explain the TF activity dynamics. In Fig. 3.2 panel c, we show activities of mitotic- and early G1-active TFs with a high amplitude dynamics together with their expression profiles. Interestingly, TBP, TAF1 and FOSL2 show a high positive correlation indicating that their activities may be regulated at the transcriptional level. On the other hand, SOX13 and HNF1A show a clear delay between expression and activity which could reflect the delay on the accumulation of active protein due to mRNA and protein half-lives or postranscriptional regulation. Strikingly, POU5F1 shows a strong negative correlation which suggests that may act mainly as a repressor. In summary, our analysis not only allows us to identify the activity dynamics of key TFs involved in transcription reactivation, but also provides preliminary hints on the molecular mechanisms that may be involved in such dynamics.

3.3.3 Bookmarking and transcription reactivation kinetic

Next, we investigated the role of mitotic bookmarking in the transcription reactivation dynamics. To do that, we analyzed the expression of genes associated to FOXA1, a liver-specific factor and one of the first identified bookmarking factors. We used mitotic ChIP-Seq data from a study of Caravaca et al. [30]. We selected the genes associated to FOXA1 ChIP-Seq peaks (see Methods) and we calculated the average expression of these genes and compared it with the overall average gene expression. Surprisingly, genes associated to FOXA1 reach their activation peak later than the overall peak of gene expression that occurs during the transition between mitosis and early G1 phase (see Fig. 3.3, panel a). Then, we compared the activity of FOXA1 with the average activity of all TFs, revealing a negative peak during mitosis (see Fig. 3.3, panel b), in accordance with the results shown in Fig. 3.3, panel a. These results suggest that FOXA1, despite its presence on mitotic chromosomes through specific and non-specific interactions, is not sufficient to promote quick transcription reactivation. However it may play a structural function by keeping the chromatin open to promote binding of other TFs.

To scale up this analysis we took advantage of a recent large scale study by Raccaud et al. in 2019 [32]. The authors were able to systematically measure the mitotic chromosome binding of 501 TFs in mouse fibroblast cells by live-imaging cell lines carrying exogenous florescence constructs. The mitotic bound fraction (MBF) was defined as the fraction of fluorescence signal located on mitotic chromosomes over the total cell signal. According to this score the TFs were divided in three categories (enriched, intermediate and depleted) indicating their capacity to bind mitotic chromosomes and their potential to be bookmarking factors. We then assumed that human TFs in HUH7 cells behave similar as their mouse paralogs and assigned the corresponding MBF score. We hypothesized that genes regulated by TFs with high MBF should be ready to be reactivated earlier. To test this, we calculated a MBF weighted average score (MWAS) for each promoter as the average MBF of all the TFs that regulate a given gene promoter weighted by the number of their binding sites. Then, we divided genes in high and

low MWAS and we calculated the average expression of these two groups (see Methods for further information). Genes associated to high MWAS, i.e genes that tend to be regulated by TFs with high MBF, did not show a faster reactivation dynamics but a significant larger expression during early G1 phase (see Fig. 3.3, panel c and d). Consistently, we showed no significant difference between the MBF distribution of TFs with high activity during mitosis or during early G1 (supplementary Fig. 3.S5). These results indicate that there is an absence of correlation between TF mitotic binding and TF mitotic activity and quick transcription reactivation of their target genes. We cannot rule out that the absence of correlation could be due to the fact that MBF scores were measured with an artificial system in a different cell line of a different organism. However, the absence of correlation suggests that previously reported mitotic bound factors, as in the case of FOXA1, may have a structural function by keeping chromatin open during mitosis and the speed of transcription reactivation may be then regulated by other determinants.

Next, we studied whether promoter architecture could be one of the determinants of early transcription reactivation. Surprisingly, just the total number of binding sites within the gene promoter is a strong feature to predict early or late reactivation. Indeed, average expression of genes with large number of binding sites shows a quick transcription reactivation during mitosis, in contrast to a reactivation during early G1 of genes with small number of binding sites (see Fig. 3.3, panel e and f). Two non-exclusive mechanisms could explain why strong promoters reactivate earlier: first, gene promoters with more TF binding sites may be easier to kept accessible during mitosis as more TFs could compete against nucleosomes leading to nucleosome free regions. Second, large number of binding sites may facilitates that TFs find the promoters and thus increase the chance to recruit the transcriptional machinery. Finally, as expected, genes that have a large number of binding sites for TFs with a high inferred activity during mitosis (high mitotic activity weighted average score, MAWAS), showed a high mitotic transcription and a quick transcription reactivation (see Fig. 3.3, panel g and h). Therefore, we believe that our method allows to identify new bookmarking factors that should not only bind mitotic chromosomes but be able to bind specific DNA binding sites during mitosis. In addition, we predict that promoters with large number of binding sites for these TFs should show a higher degree of chromatin accessibility during mitosis.

3.3.4 Identification of the Core Regulatory Network responsible for the transcription reactivation after mitotic exit

Next, we wanted to identify the TFs, among the 332 for which we are able to infer activities, that have a major role on the reactivation of transcription exiting mitosis. Namely, the key TFs that if perturbed may affect more significantly the measured gene expression patterns. To do so, we calculated the fraction of explained variance as a measure of the performance of our model to fit the data. Then, we defined a TF importance score as the reduction on fraction of explained variance when the TFs is removed as an explanatory variable from the multiple linear regression model (see Methods for further details). Furthermore, in Fig. 3.4 we show a Core Regulatory Network (CRN) where the nodes are formed by the 5% top most important TFs and the links represent potential regulatory interactions between the selected TFs according to the presence of TF binding sites in their promoters. Interestingly, the CRN shows a large number of regulatory links (149 connections) while random networks with the same number of TFs

produce a smaller number of connections (36 on average). Thus, the high interconnectiviy of our CRN suggests that the identified TFs may be related functionally. Indeed, some of them have been reported to be involved in cell-cycle or cell proliferation and growth as ATF1, FOS, CEBPZ, SP3 and KLF4 [63]. Moreover, the CRN structure shows multiple feedback loops rather than a hierarchical network as one could expect taking into account the observed sequential waves of transcription reactivation. This type of structure has the potential to show cycling dynamics which may be important not only for the reactivation after mitosis but for the regulation of transcription across the whole cell-cycle. In conclusion, we predict that these TFs could represent relevant therapeutic targets to control cell proliferation.

3.3.5 Genes within in the same TAD show a higher correlation on the reactivation kinetics

The connection between transcription and the 3D structure of chromatin is currently a very active field of research. During mitosis, topologically associated domains (TADs) are disrupted and rebuilt at different dynamics during the transition between mitosis and G1 phase [65]. However, the causal connection between transcription reactivation and chromatin structure reformation exiting mitosis is not known yet. As a first attempt to investigate this relationship, we analyzed the correlation between transcription reactivation profiles of genes belonging to the same TAD. To do that, we took two TAD lists identified in human IMR90 and ES cells, from Hi-C experiments [11]. Although these are different cells lines than the HUH7, it has been shown that TADs are highly conserved between different cell-types and even different organisms [11]. We obtained a total of 2290 and 3061 TADs respectively. For each TAD in both lists, we identified the expressed genes that are located within its limits, finding respectively 10225 and 9849 genes. Pearson correlation coefficients were calculated between all pairs of genes within the same TAD. As a control, expressed genes were randomly located into TADs respecting the total number of genes in each TAD. Distribution of correlation coefficients as well as distribution of random coefficients are shown in Fig. 3.5. Interestingly, genes belonging to the same TAD show a higher correlation than random pairs of genes, indicating that transcription reactivation dynamics are similar between genes that are located near in space within the same local self-interacting chromatin 3D structure. Finally, we hypothesized that TADs containing genes characterized by a quick transcription reactivation should show a faster reformation exiting mitosis. Further experimental work would be required to validate our hypothesis.

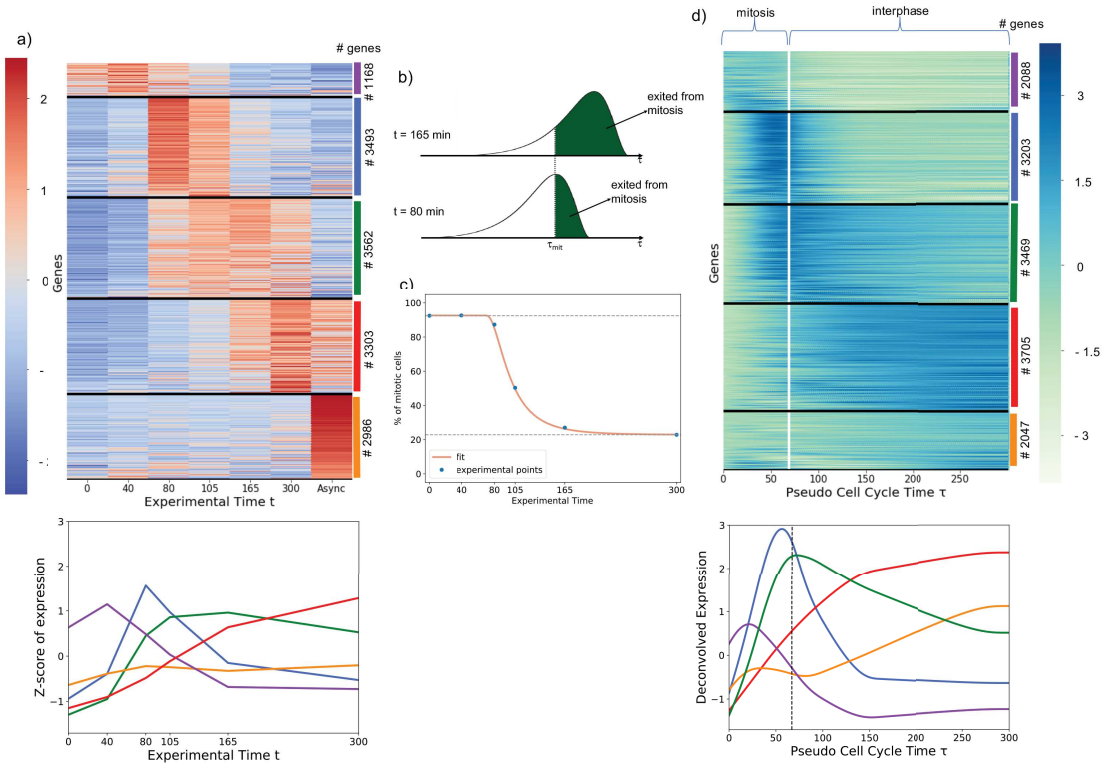


Figure 3.1. Deconvolution of gene expression data of synchronized cell population leads to dynamic expression profile respect to cell-cycle average profile. **a**: Genes can be divided in groups according to their dynamic expression profiles over time. Each row corresponds to a gene. Black horizontal lines divide the different clusters of genes. The color scale represents the level of the z-score of expression of each gene, as shown by the colorbar on the left. On the right, the number of genes for each cluster is indicated as well as the color corresponding to the cluster expression average shown on the bottom panel. **b**: We assume that cells have to wait a stochastic, log-normally distributed lag time to start again the cell-cycle progression after the release of the chemical cell-cycle arrest by nocodazole. Here, a pictorial representation of lag time distribution, where the green part of area represents the fraction of cells that already exited mitosis at two different experimental time points. The dashed line indicates time τ_{mit} that cells need to complete mitosis. **c**: Blue dots: quantification of cells showing condensed (mitotic) and decondensed (non-mitotic) chromatin after synchronization release (data from [28]); Orange line: model fitting used to infer τ_{mit} and the parameters of the log-normal distribution, σ and μ. **d**: After the deconvolution, genes were divided in groups according the their dynamic expression profile over the internal pseudo cell cycle time τ. The vertical white line represents τ_{mit} and separates ideally mitosis from interphase in early G1 phase. The color scale represents the level of the z-score of expression of each gene, as shown by the colorbar on the right. On the right, the number of genes for every cluster is indicated as well as the color corresponding to the cluster expression average shown on the bottom to one of the curves on the bottom.

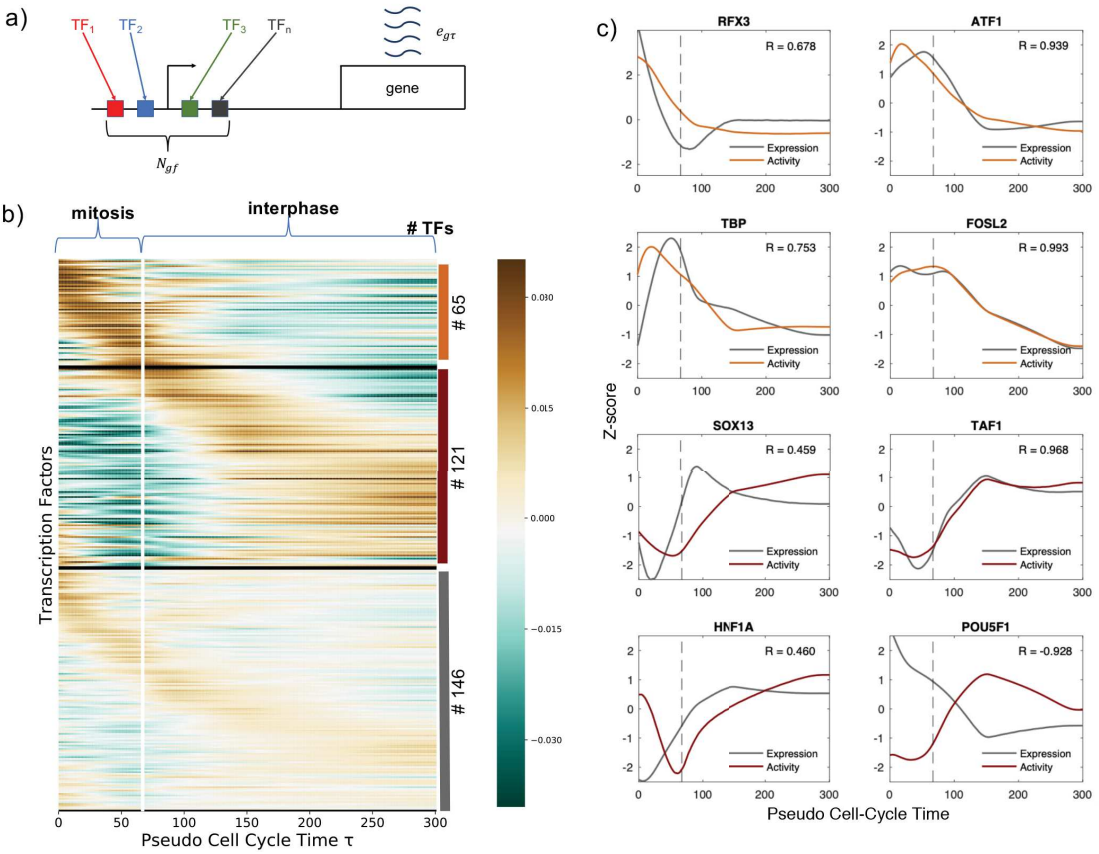


Figure 3.2. Transcription factor activity dynamics during mitosis and early G1 phase. **a**: Schematic representation of the model: the expression $e_{g\tau}$ of the gene g at cell-cycle pseudotime τ is a linear combination of the activities of different transcription factors f binding the promoter of g. N_{gf} represents the entries of a matrix **N** containing the number of sites for TF f associated with promoter of the gene g, taking into account the affinity between the motif of f and the sequence of the promoter. **b**: TFs can be divided in groups according to their activity dynamics over the cell-cycle pseudotime τ. The vertical white line represents τ_{mit}, and indicates the transition between mitosis and interphase. On the right, the number of TFs belonging to each cluster is indicated. **c**: Activities of mitotic-active (orange curves) and early-G1-active (red curves) TFs that show a high amplitude dynamics. Grey lines show the gene expression dynamics of the corresponding TF genes. Pearson correlation coefficients between the TF activites and expressions are shown in each panel. Dashed lines represent τ_{mit}.

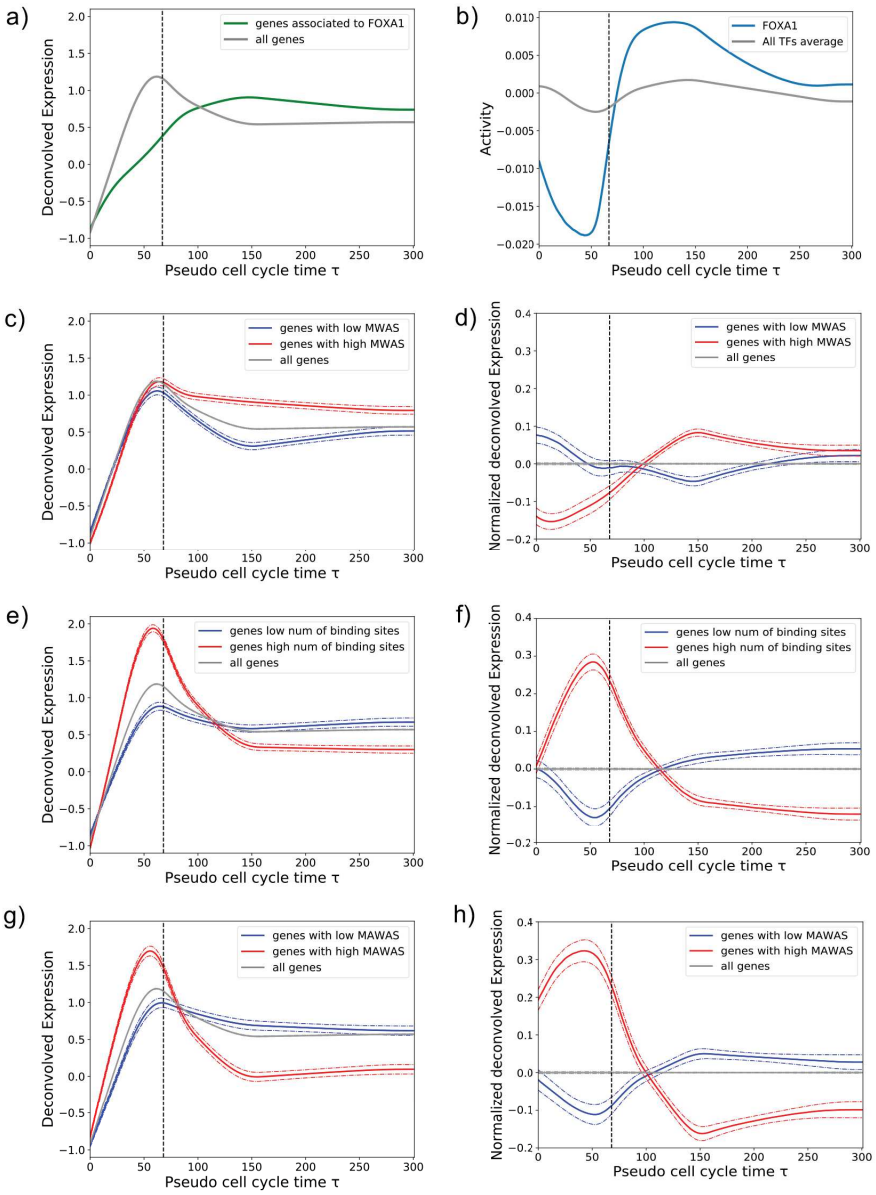


Figure 3.3. Bookmarking and transcription reactivation dynamics. **a**: The average expression of all genes (grey line) was compared with the average expression of FOXA1 target genes during mitosis (green line). **b**: The activity of FOXA1 (blue line) in comparison with the average activity of all TFs (grey line). **c**: Gene expression pattern as a function of the promoter MWAS (MBF weighted average score). The average expression of all genes (grey line) was compared to the average expression of genes whose promoters tend to be regulated by TFs with high MBF (red line) and low MBF (blue line) (see Methods). **d**: The same as in panel c, but gene expression patterns have been processed as done when applied the linear model to infer the activity, as described in Methods. **e**: Gene expression pattern as a function of the total number of TF binding sites in the promoter. The average expression of genes with large (red) and low (blue) number of promoter binding sites are compared to the overall averege expression (grey line). **f**: The same as in panel e, but gene expression patterns have been processed as done when applied the linear model to infer the activity, as described in Methods. **g**: Gene expression pattern as a function of the promoter MAWAS (mitotic activity weighted average score, see Methods). The average expression of genes with high (red) and low (blue) MAWAS are compared to the overall averege expression (grey line). **h**: The same as in panel g, but gene expression patterns have been processed as done when applied the linear model to infer the activity, as described in Methods. Dashed vertical lines in all panels indicate τ_{mit}.

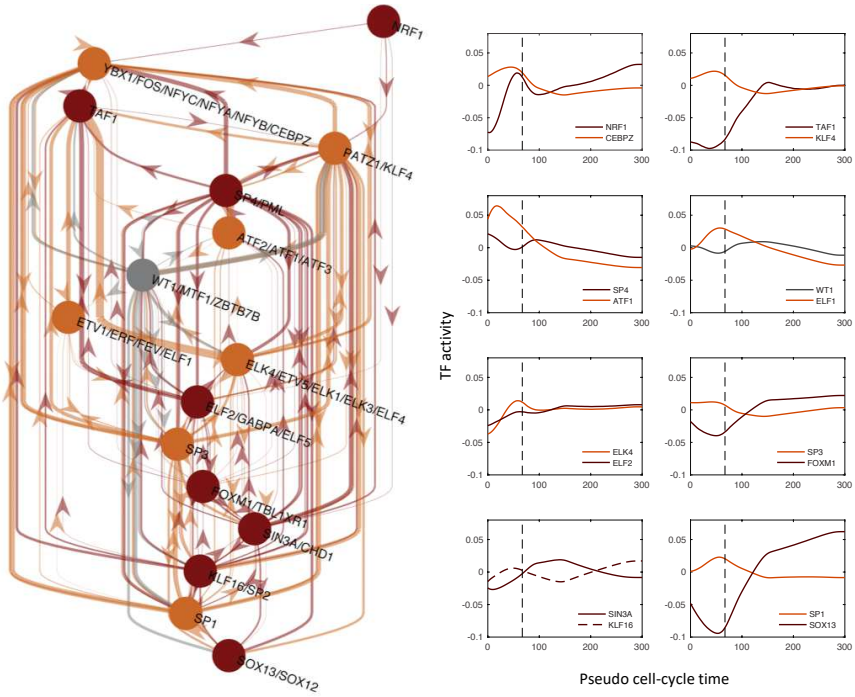


Figure 3.4. Identification of the Core Regulatory Network responsible for the transcription reactivation after mitotic exit. Inferred core regulatory network (CRN) by selecting the top 5% of the TFs according to their importance in explaining the gene expression patterns and their inferred time-dependent activities. Colours indicate the cluster to which TFs belong, in accordance with the in Fig. 3.2, panel b, vertical coloured bars on the right. Note that the majority of TF motifs are associated uniquely to single TF, while some other motifs are shared by more than one TF, precluding the inference of single TF activities of motifs which can be potentially bound by more than one TFs.

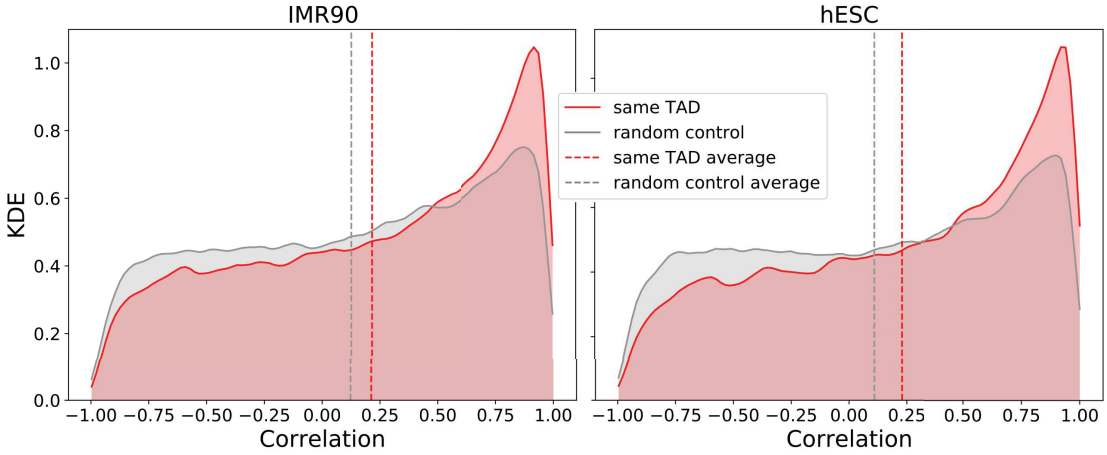


Figure 3.5. Genes belonging to the same TADs show an higher correlation in expression. Correlations between gene expressions for genes belonging to the same TAD were calculated, for two different TADs sets (IMR90, hESC cells). In red, the KDE of the distribution of the correlation is plotted, in comparison with a random model (grey plot). Dashed vertical lines represent the average values of the corresponding distributions.

3.4 Conclusion

Cell identity maintenance in proliferating cells is a biological process that has crucial implications in developmental biology, regenerating medicine and cancer. Nevertheless, the precise molecular mechanisms responsible for the transmission of regulatory information from mother to daughter cells are not fully understood yet. Mitotic bookmarking by TFs through specific DNA binding on mitotic chromosomes has been proposed as a mechanism to reinforce cell identity maintenance during cell division. In this paper we studied the regulation of transcription reactivation exiting mitosis and the connection with mitotic bookmarking.

First, we reanalyzed time-dependent EU-RNA-Seq data on synchronized cell populations by a mitotic arrest, to correct for the progressive desynchronization of cells after block release. This allowed us to estimate gene expression profiles with respect to a cell cycle pseudotime with an explicitly defined transition between mitosis and early G1 phase. Remarkably, we identified a set of genes that show a very early wave of transcription reactivation during mitosis. However, the majority of genes showed a peak of transcription at telophase or during the transition between mitosis and G1.

Next, we estimated TF activity dynamics of 332 expressed TFs by fitting a multiple linear model to the deconvolved gene expression profiles. We observed time-dependent waves of TF activities suggesting an intrinsic TF hierarchy with respect to their role on transcription reactivation after mitosis. In addition, we investigated whether TFs previously reported to bind mitotic chromosomes were responsible for a faster reactivation dynamics. Surprisingly, we did not find a strong correlation between genes regulated by mitotic bound TFs and the speed of reactivation. However, our approach allowed us to identify around 60 TFs that are highly active during mitosis and represent new candidates of mitotic bookmarking factors. Therefore, we predict that the interactions of these factors with their specific target sites during mitosis are the molecular mechanisms responsible for mitotic transcription and transcription reactivation. Moreover, we hypothesize that these specific interactions may also play an important role maintaining chromatin accessibility on mitotic chromosomes. Further experimental work would be needed to validate our hypothesis and predictions.

Moreover, we reconstructed a core regulatory network underlying the dynamics of transcription reactivation exiting mitosis, by selecting the key TFs that showed the highest explanatory power in our multiple linear regression model. Then, we propose a list of candidates to be the crucial players in the process of reactivating the gene expression in the first stages of the interphase, ensuring the cell identity. We predict that these TFs could represent relevant therapeutic targets to control cell proliferation. Further experiments are required to validate our predictions and prove the active role of TFs on chromatin accessibility and 3D structure.

3.5 Methods

3.5.1 Fitting of model parameters for deconvolution of gene expression data

To estimate gene expression dynamics with respect to an internal cell-cycle pseudotime, we assumed that after the release of the synchronization there is a stochastic lag time until cells can start again the cell-cycle progression. According to our model, the cell-cycle progression of a cell is represented by the internal cell-cycle pseudotime $\tau = t - \eta$, where t is the experimental time and η is the stochastic lag time that the cell had to

wait until the cell-cycle progression was restarted again. We further assumed that the lag time η is log-normally distributed with a certain mean μ and standard deviation σ. Then, the probability of finding a cell in the population with an internal cell-cycle pseudotime τ at a given experimental time t can be written as:

$$P(\tau|t)d\tau = \frac{1}{\sqrt{2\pi\sigma^2}(\tau - t)} e^{-\frac{(\log(t-\tau)-\mu)^2}{2\sigma^2}} d\tau \tag{3.5.1}$$

Assuming that cells require an average time τ_{mit} to complete mitosis we can calculate the fraction of cells waiting for mitosis to be finished as $q(t) = \int_0^{\tau_{\text{mit}}} P(\tau|t)d\tau$ and solving the integral we obtain:

$$q(t) = \begin{cases} \frac{1}{2} - \frac{1}{2}\text{erf}(\frac{\log(t-\tau_{\text{mit}})-\mu}{\sqrt{2}\sigma}) & t > \tau \\ 1 & t \leq \tau \end{cases} \tag{3.5.2}$$

Then, we fit the parameters of the stochastic model (τ_{mit}, μ and σ) by using data from [28] on the time evolution of the number of mitotic cells observed after synchronization treatment release. To do so, we define the likelihood of the data based on the assumption that the cell counts follow a binomial distribution with probability $q(t)$. Thus,

$$\mathcal{L} = \prod_i q(t_i)^{n_i^{\text{mit}}} (1 - q(t_i))^{n_i^{\text{tot}} - n_i^{\text{mit}}} \tag{3.5.3}$$

where n_i^{tot} and n_i^{mit} are, respectively, the total number of cells and the number of cells in mitosis counted at experimental time $t_0 = 0$ minutes, $t_1 = 40$ minutes, $t_2 = 80$ minutes, $t_3 = 105$ minutes, $t_4 = 165$ minutes and $t_5 = 300$ minutes. Then, the log likelihood can be written as:

$$\log\mathcal{L} = \sum_i n_i^{\text{mit}} \log\left(q(t_i)\right) + \sum_i (n_i^{\text{tot}} - n_i^{\text{mit}}) \log\left(1 - q(t_i)\right) \tag{3.5.4}$$

By performing an optimization of Eq. 3.5.4 (python *scipy optimize* package, *nelder-mead* algorithm), we can infer the parameters τ_{mit}, μ and σ, obtaining, respectively, 67min, 3.43 and 0.74. Once the parameters have been inferred, the probability $P(\tau|t)$ is fully determined and, therefore, we can recover the gene expression with respect to the internal cell-cycle pseudotime τ using the following convolution equation:

$$r_g(t) = \int_0^t E_g(\tau) P(\tau|t) d\tau \tag{3.5.5}$$

where $r_g(t)$ represents the expression of the gene g at experimental time t (given by the EU-RNA-Seq data), $E_g(\tau)$ is the expression of the same gene g at the cell-cycle pseudotime τ. This equation basically reflects that the gene expression measured at a certain experimental time is the population average over the expressions of cells at different cell-cycle times. In case of perfect synchronization over time, the probability $P(\tau|t)$ would become a Dirac delta function and the gene expression in both times would be the same. Furthermore, we took into account the fact that the samples were contaminated by a fraction $\pi_M = 0.23$ of cells that never exited mitosis and a fraction $\pi_I = 0.075$ of cells that did not response to the mitotic block and stayed in interphase (see Fig. 3.1, panel c). It means that only a fraction $\pi_C = 1 - \pi_M - \pi_I$ starts again the cell cycle progression within the duration of the experiment. Then, this can be summarized by

describing the measured gene expression as a mixture of the three cell populations as follows:

$$r_g(t) = \pi_C \int_0^t E_g(\tau) P(\tau|t) d\tau + \pi_M E_g(0) + \pi_I E_g^I \tag{3.5.6}$$

$$r_g^a = f_{\text{mit}} \int_0^{\tau_{\text{mit}}} E_g(\tau) d\tau / \tau_{\text{mit}} + f_I E_g^I \tag{3.5.7}$$

where, E_g^I is the average expression during interphase and an extra equation is included to relate the gene expression r_g^a measured on an asynchronous cell population as a weighted average of the gene expression during mitosis and during interphase where the weights reflect the fraction of the cell-cycle duration T_C that cells expend on average in each phase, i.e $f_{\text{mit}} = \tau_{\text{mit}}/T_C$ and $f_I = 1 - f_{\text{mit}}$.

Then, to perform the deconvolution we discretized the cell-cycle pseudotime into small intervals ($\delta\tau = 1$ min) and expressed the Eq. 3.5.6 and 3.5.7 into matricial from: $\mathbf{r}_g = M\mathbf{E}_g$, where the expression vectors are defined as $\mathbf{r}_g = (r_g(0), r_g(t_0), r_g(t_1), \ldots, r_g^a)$ and $\mathbf{E}_g = (E_g(0), E_g(\delta\tau), E_g(2\delta\tau), \ldots, E_g^I)$ and the matrix M is the sum of three components: $M = M_C + M_M + M_I$ that account for the three distinct cell populations. First, the cell-cycle matrix M_C models the desynchronization of the cells that re-enter the cell cycle, as in Eq. 3.5.5, and can be written as:

$$M_C = \begin{pmatrix} & & & 0 \\ & \pi_C P & & 0 \\ & & & \cdots \\ & & & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \tag{3.5.8}$$

where P is the discrete version of Eq. 3.5.2. Second, the mitotic matrix M_M adds to the model the contribution of the cells that are still in mitosis by mapping them into $\tau = 0$. Its explicit form is:

$$M_M = \pi_M \begin{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} & \begin{matrix} 0 \\ 0 \\ \cdots \\ 0 \end{matrix} \\ \begin{matrix} 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \end{matrix} & 0 \end{pmatrix} \tag{3.5.9}$$

And third, the interphase matrix M_I exploits the asynchronous dataset to infer the average expression levels during interphase. The matrix takes the following form:

$$M_I = \begin{pmatrix} \begin{pmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} & \begin{matrix} \pi_I \\ \pi_I \\ \cdots \\ \pi_I \end{matrix} \\ \begin{matrix} \frac{1}{T_C} & \frac{1}{T_C} & \cdots & \cdots & \cdots & \cdots & \cdots \end{matrix} & f_I \end{pmatrix} \tag{3.5.10}$$

where the cell cycle duration T_C is set to $24h$ [1].

Hence, the deconvolution problem can be understood as a multiple linear regression and, therefore, we can infer the gene expression in the space of the cell-cycle pseudotime by optimizing the following quadratic loss function:

$$L_{SM} = \sum_{g} \mid \mathbf{r}_g - M\mathbf{E}_g \mid^2 + \lambda \sum_{gi} \mid E_{g,\tau+1} - E_{g,\tau} \mid^2 \tag{3.5.11}$$

where we added a smooth Ridge regularization term to be able to solve the overrepresented linear model and avoid overfitting. Then, the solution is $\mathbf{E}^*_g = (Q + \lambda I)^{-1} M^T \mathbf{r}_g$, where $Q = M^T M$ and I is the regularization matrix.

Finally, To choose the parameter λ, we calculated the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) scores in function of λ, as follows [66, 67]:

$$\text{AIC} = N_E \cdot \chi^2 + 2N_G \cdot D \tag{3.5.12}$$

$$\text{BIC} = N_E \cdot \chi^2 + 2N_G \cdot D \cdot \log(N_E) \tag{3.5.13}$$

where N_E is the number of experimental time points, N_G the total number of genes, $\chi^2 = \sum_g \mid \mathbf{r}_g - M\mathbf{E}^*_g \mid^2$ is the minimum error and D is the degree of freedoms that, for a multiple linear regression model with smooth Ridge regularitation, can be calculated as $D = \text{Tr}(M((Q + \lambda I)^{-1} M^T)$. The BIC score tends to introduce a stronger penalty producing a solution more robust against overfitting, therefore we chose $\lambda = 0.79$ that minimizes the BIC score (see supp. Fig.3.S1).

3.5.2 Visualization of the gene expression through heatmaps

To represent the gene expression as shown in Fig. 3.1 panel a, processed EU-RNA-Seq data at the transcript level from [28] were used. Transcript FPKMs from the same gene were then grouped to obtain gene level EU-RNA-Seq data, and all the genes with a low expression on the asynchronous sample (< 36 FPKM) were excluded (see supp. Fig. 3.S1, panel a). Then, a z-score was calculated, correcting each FPKM value by subtracting the mean μ_g and dividing by the standard deviation σ_g, both μ_g and σ_g calculated over the corresponding gene. Genes were divided into 5 clusters (the optimum number to obtain significantly different profiles) according to their z-score over time, by using the *KMeans* tool from *sklearn* python library. The heatmap was represented by using *seaborn* python library, ordering the genes of each cluster according to their norm with respect to the corresponding cluster average expression.

3.5.3 Inference of transcription factor activities

We developed an ISMARA-like model [41] where the expression of a given gene with respect to the cell-cycle pseudotime can be obtained as a linear combination of time-dependent activities of all TFs that can potentially bind its promoter. First, a as proposed in [41], we preprocessed our data as follows: to revert the z-score transformation performed above we multiplied the gene expression values $E_{g\tau}$ by the standard deviation σ_g and added the average μ_g. Second, in order to calculate the log2 expression for all genes, we add a pseudo-count to the corrected $E_{g\tau}$ values, i.e. for every given time τ, we ranked all the values higher than zero and we calculated the 5th percentile pc_τ. We then added pc_τ to the corresponding $E_{g\tau}$. After that, we calculated $\hat{e}_{g\tau}$, i.e. normalized values of the gene expression at a cell-cycle pseudotime τ, as follows:

$$\hat{e}_{g\tau} = \log_2 \left[10^6 \cdot \frac{E_{g\tau}}{\sum_{g'} E_{g'\tau}} \right] \tag{3.5.14}$$

We further normalized the expression of genes across pseudotime and genes resulting in $e_{g\tau} = \hat{e}_{g\tau} - \langle \hat{e}_g \rangle - \langle \hat{e}_\tau \rangle + \langle\langle \hat{e} \rangle\rangle$. Finally, we write the linear model as:

$$e_{g\tau} = \sum_f N_{gf} A_{f\tau} \tag{3.5.15}$$

where the value N_{gf} represents the number of binding sites for the TF f on the gene promoter g, taking into account the affinity between the motif of f and the sequence of the promoter, and the unknown parameter $A_{f\tau}$ is the activity of the TF f at a given cell-cycle pseudotime τ. The binding site matrix is further normalized to ensure $\sum_g N_{gf} = 0$. Note that the TF activities are then zero mean variables.

Then we used least square fitting to obtain the TF activites. To avoid overfitting we included a Ridge regularization penalty. To estimate the weight of the regularization we calculated the Mean Square Error (MSE) for a training and a test datasets and we performed a 80-20 cross-validation. A regularization factor $\lambda = 443$ was chosen, corresponding to the minimum of the MSE of the test dataset (see supp. Fig. 3.S2). In addition, we calculated the explained variance (EV) of the model, $EV = \frac{(e_{g\tau} - e^{th}_{g\tau})^2}{(e_{g\tau} - \mu_g)^2}$, where $e^{th}_{g\tau}$ is the theoretical expression of the gene g at internal cell cycle time τ, i.e. calculated using the inferred activity $A_{f\tau}$ and the matrix N_{gf}, and μ_g is the mean among all the values $e_{g\tau}$. We obtained a regularization factor $\lambda = 443$, corresponding to the maximum of the EV of the test dataset (supp. Fig. 3.S2), in accordance with the minimum obtained for the MSE.

3.5.4 Visualization of the TFs activities through heatmaps

To represent the TFs activities as shown in Fig.3.2, the TFs were divided into 3 clusters according to their activity dynamics over τ. First, we calculated the standard deviation over time of TF activities as $\sigma_f = \sqrt{\sum_\tau A^2_{f\tau}}$ and classify a TF as high amplitude dynamic if $\sigma_f > 0.07$. Second, we sorted TFs according to when their maximum activity peak occurred and defined a TF as mitotic active if the peak appeared before τ_{mit}. Therefore, TFs were classified as either mitotic active, early-G1 active or non-dynamic. The heatmap in Fig.3.2 was represented by using *seaborn* python library, ordering the TFs in each cluster by the the first reached maximum over τ.

To represent the TFs activity as shown in supp. Fig. 3.S3, we inferred the mean activity $\bar{A}_f$ of each factor f, as described in [41], and then we added $\bar{A}_f$ to the corresponding $A_{f\tau}$ values. Then, they were divided in 3 groups by performing a k-means, through the *KMeans* tool from *sklearn* python library.

To represent the TFs activity as shown in supp. Fig. 3.S4 only TFs corresponding to genes belonging to the Gene Ontology (GO) category (Cell Cycle - GO:0007049) were considered.

3.5.5 Core Regulatory Network

To build the core regulatory network (CRN) we selected the TFs that showed a high degree of explanatory power to reproduce the gene expression dynamics. To do that, we assigned a score for each TF based on its contribution to the explained variance by calculating a *reduced* explained variance $EV_f = \frac{(e_{g\tau f} - e'^{th}_{g\tau f})^2}{(e_{g\tau f} - \mu_g)^2}$, i.e. the EV as shown in the section 3.5.3 but removing from the model the corresponding TF f. Then, we defined

the importance score of a TF by the ratio $\frac{EV_f}{EV}$ and we ranked all TFs according to this score taking into account that the smaller is the ratio the higher is the impact of the TF on the explanatory power of the model. The figure was then generated by using digraph library in Matlab, by selecting relevant TFs and corresponding genes in the **N** matrix.

3.5.6 Genes associated to TFs

To establish which genes are associated to FOXA1, as shown in Fig. 3.3, panel a, we took into account the mitotic ChIP-Seq peaks of FOXA1 from [30]. Then, for every peak, we selected the nearest expressed gene, using as references the corresponding TSS and the average point of the selected peak. So, we obtained a list of expressed genes that we defined as genes bound by FOXA1 during mitosis.

To establish which genes tend to be regulated by TFs with high or low MBF [32], as shown in Fig. 3.3, panels c and d, we calculated what we called *MBF weighted average scores* (MWAS) as follows: for each gene g, we took the number of binding sites N_{gf} corresponding to the factors f for which we know the MBF score. Each of these values was multiplied for the corresponding MBF, and then they were summed all together. Finally, this sum was divided by the total number of binding sites, i.e. the sum $\sum_f N_{gf}$. This score is what we called MWAS. Then we ranked the genes according to the MWAS, and we removed the ones with MWAS $= 0$. The 10% of the genes with the highest MWAS were then considered associated with enriched TFs ("enriched genes"), while the 10% of the genes with the lowest MWAS were considered associated with the depleted TFs ("depleted genes").

To obtain genes enriched or depleted in binding sites, we calculated for each gene promoter the total number of binding sites, i.e. $\sum_f N_{gf}$ and then the 10% of the genes with largest number of binding sites and the 10% of the genes with smallest number were considered to calculate average expression profiles as shown in Fig. 3.3 panel e and f.

To obtain genes with large number of binding sites for TFs with a high inferred activity during mitosis, as as shown in Fig.3.3, panels g and h, we calculated what we called *(high mitotic activity weighted average score* (MAWAS), i.e. we adopted the same procedure used to calculate the MWAS, but taking into account average mitotic activity instead of MBF score for each TF.

3.6 Supplementary Figures

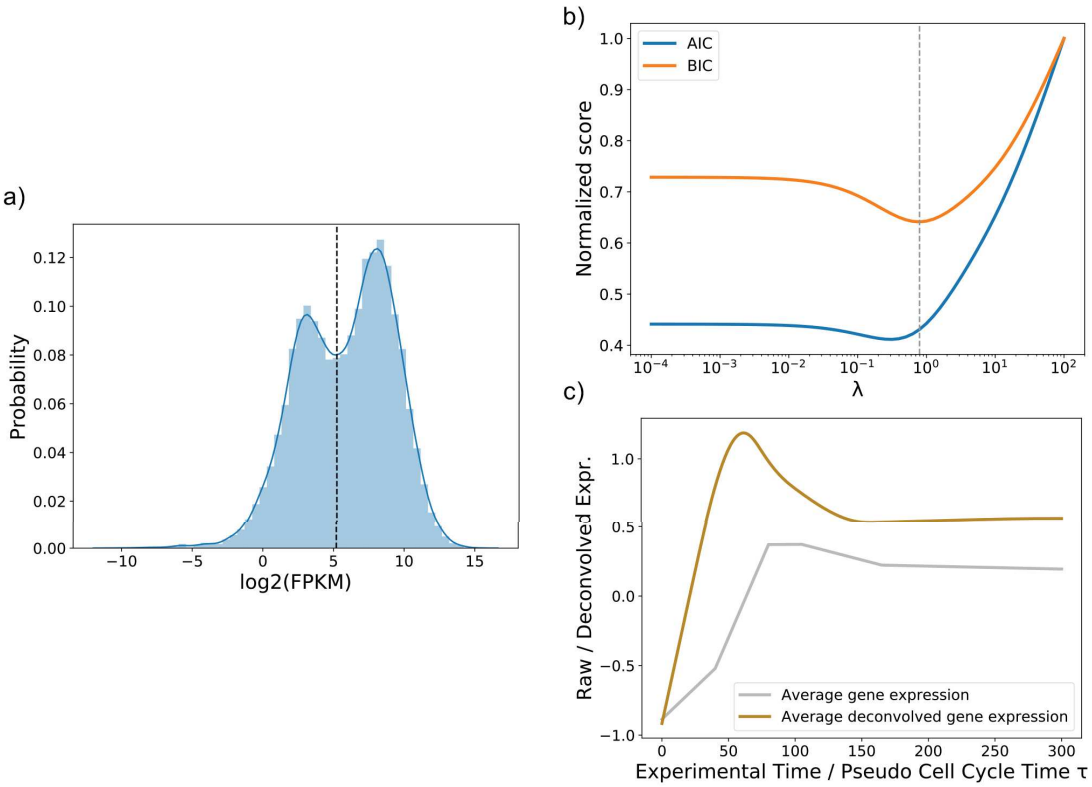


Figure 3.S1. Data processing for deconvolution. **a**: Log2 histogram of FPKM reads at gene level for asynchronous data. The dashed vertical line represents the threshold we considered to process our data: genes with asynchronous $FPKM < 36.76$ were excluded. On the vertical axis, kernel density estimation is indicated. **b**: AIC and BIC scores were calculated in order to establish the best λ parameter for the regularization of the deconvolution process (see Methods). Both AIC and BIC showed a minimum, and we choose $\lambda = 0.79$, corresponding to the BIC minimum (dashed vertical line). **c**: Average gene expression of convolved (grey line) and deconvolved (yellowish line) data were represented on the same plot.

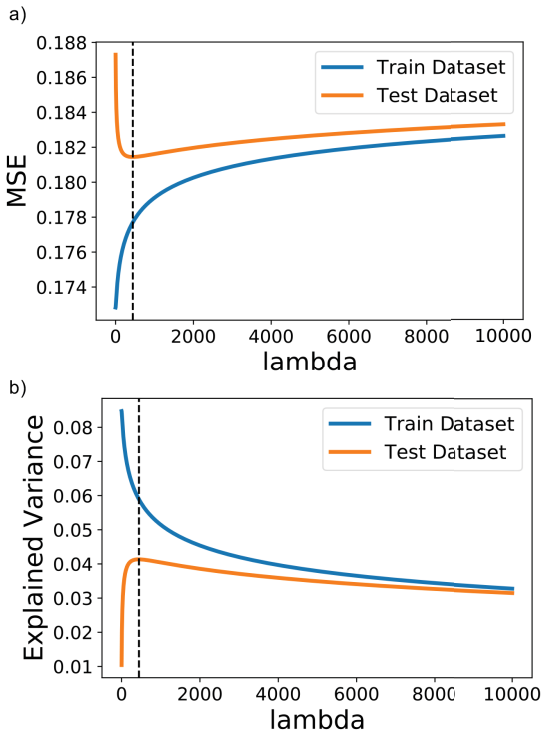


Figure 3.S2. Cross validation of the linear model. **a**: A cross-validation 80/20 was performed to find the best λ regularization parameter for inferring the TFs activity (see section 3.5.3). A value $\lambda = 443$ was chosen (dashed vertical line), corresponding to the minimum of the Mean Squared Error (MSE) of the test dataset (see Methods). **b**: The same analysis shown in the panel a was performed by using Explained Variance (EV) instead of MSE. The dashed vertical line corresponds to the maximum $\lambda = 443$, in accordance with the minimum MSE.

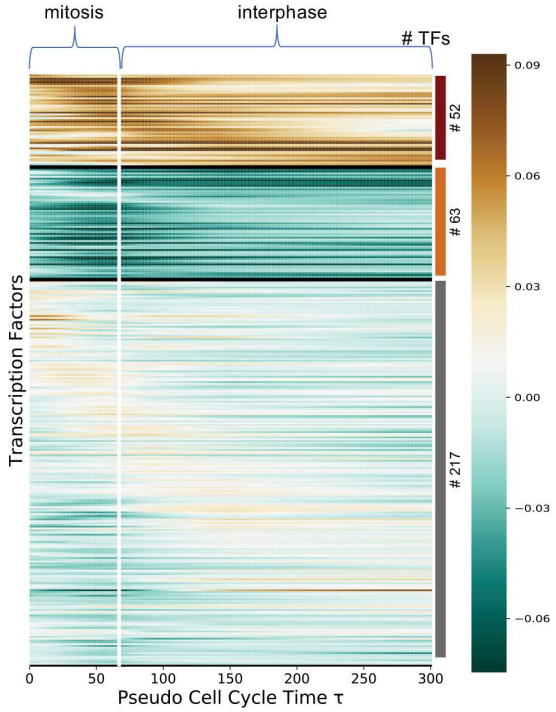


Figure 3.S3. Transcription factors dynamics taking into account their average activity. Here, the mean activity $\bar{A}_f$ of each factor f was added to $A_{f\tau}$, and then TFs were clustered according to the profile of their activity over the internal pseudo cell cycle time τ. The vertical white line represents τ_{mit}, and separates ideally the mitosis from the interphase. On the right, the number of TFs for every cluster is indicated.

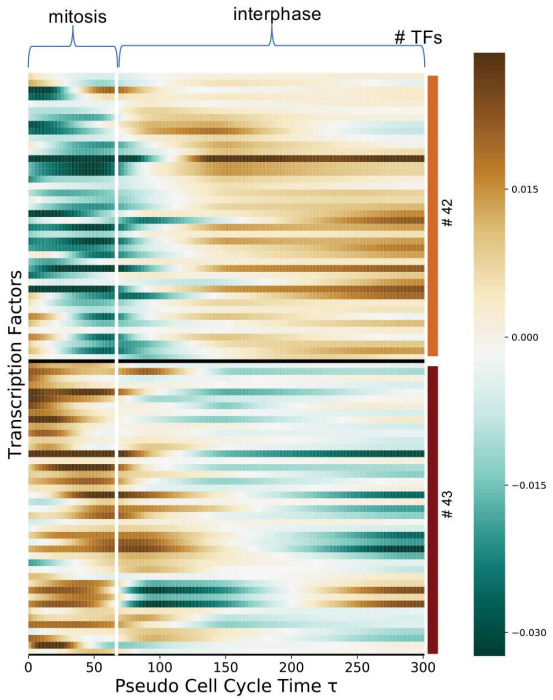


Figure 3.S4. Transcription factors dynamics taking into account only cell-cycle GO category. Here, only TFs associated to genes belonging to the Gene Ontology (GO:0007049) category have been shown and clustered. In this case, only 2 main groups of TFs have been individuated, and both of them show a significant activity change over τ. The vertical white line represents τ_{mit}, and separates ideally the mitosis from the interphase. On the right, the number of TFs for every cluster is indicated.

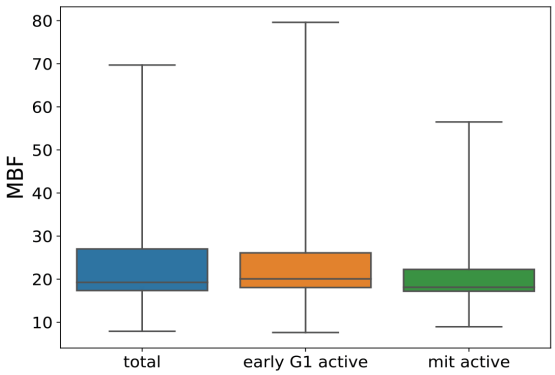


Figure 3.S5. Average MBF for mitotic and early G1 active transcription factors. Boxplots showing the average MBF for TFs with higher activity during mitotis (green box) and during early G1 (orange box) respectively, in comparison with the average of all TFs (blue box).

4

A preliminary analysis on 3D structure reformation after mitosis

The connection between transcription and the 3D structure of chromatin is currently a very active field of research. During mitosis, topologically associated domains (TADs) are disrupted and rebuilt at different dynamics during the transition between mitosis and G1 phase [65]. However, the causal connection between transcription reactivation and chromatin structure reformation exiting mitosis is not known yet.

Here, we applied the linear ISMARA-like model described in chapter 3 to infer the activity of transcription factors during the formation and the development of the boundaries of TADs. As we will show in the next sections, thanks to this preliminary analysis we could identify new TFs that may play an important role in the formation of TADs after mitosis, introducing new possible regulators in the scenario of the 3D structure organization of the chromatin.

4.1 Transcription factors dynamics in boundaries reformation during and exiting mitosis

The insulation score (IS) of a genomic region is defined as the average interaction occurring in a certain vicinity of the region. Minima of IS reflect high insulation, and it's the most common way to classify TAD boundaries [68]. Using the IS, equation 3.3.1 can be modified to describe the process driving the boundaries reformation exiting mitosis. To do so, we assume that IS_{bt}, the insulation score of the region b at time t, is a linear combination of the activity A_{ft} of different TFs f at time t. Then we have:

$$IS_{bt} = \sum_{f} K_{bf} A_{ft} \tag{4.1.1}$$

where the values K_{bf} represent the entries of a matrix **K** containing the number of sites for TFs f on the regions b, taking into account the affinity between the motif of f and the sequence of the regions.

4.2 Results

The equation 4.1.1 has been sistematically applied to the dataset published in [34] and already presented and used in chapter 2, section 2.4, as described thereafter.

We downloaded Hi-C data in *cool* format, and we processed them by using *HiCExplorer* tools [69]. *hicConvertFormat* was used to convert *cool* files in *h5* files at $32kb$ of resolution. ICE normalization was performed by using *hicCorrectMatrix*. We decided to use TADs at 11 hours after the release of the mitotic block as a reference: each of the corresponding 4321 boundaries represents then a region b of length $32kb$ in the equation 4.1.1. TADs were called by using *hicFindTADs*, by using the following parameters: minDepth=96000, maxDepth=192000, step=32000, thresholdComparison=0.05, delta=0.01, correctForMultipleTesting=fdr. *hicFindTADs* was also used to calculate the insulation score (IS) in correspondence of the 4321 reference regions for datasets at previous experimental timepoints (0, 0.5, 1, 2, 2.5, 3, 3.5, 4.0, 4, 4.5, 5, 6, 7, 8, 9 and 10 hours after the release of the mitotic block). To choose a reliable set of parameters, we tested different combinations and compared, by a visual inspection, the TAD boundaries obtained in correspondence of a sample region of asynchronous cells ($36.5 - 70$Mb of chromosome 14) , with ChIP-Seq peaks for CTCF, that is indicated in literature as correlating with TADs boundaries [70, 71]. To plot boundaries in comparison with ChIP-Seq peaks, we used *hicPlotTADs*. ChIP-Seq data for CTCF for HeLa cells in *wig* format were downloaded from GEO accession: GSM3619487, and then converted in *bigwig* format by using the tool *wigToBigWig* [72]. Lengths of each human chromosome, needed to run *wigToBigWig*, were downloaded from *UCSC* database [73].

The application of 4.1.1 also requests the knowledge of the values K_{bf}, that have been obtained by using data from *Homer* database [74]. Here, possible motif positions genome-wide on human genome are shown for a total of 425 TFs, by taking into account the affinity between the motif and the sequences of the genome. Then, for each TF, we obtained a list of possible hits on the human genome, and each of these hits is characterized by a *zero-or-one occurrences per sequence* (zoops) type score [75]. Given a TF f, we selected only the hits with a score above the average μ_f, calculated over all the scores corresponding to f. Then, for every region b, we counted how many hits of f fall inside b: this count indicates the value K_{bf}. We found an average of 3106 hits per region, with a standard deviation of 214. By doing that and by knowing the insulation scores IS_{bt}, we could infer the activities A_{ft}, by using the same data normalization, mathematical and computational procedures seen in chapter 3, section 3.3.2. We performed a 80-20 cross-validation, and a regularization factor $\lambda = 1623$ was chosen.

We divided the TFs in 3 clusters, according to their profile over experimental timepoints. We showed that almost 28% of TFs present a positive activity during the first 2 hours after the block release, and then an almost constant negative value. Conversely, 13% of TFs present a negative acrivity during the first 2 hours post-mitotic arrest, and then an almost constant positive value. Lastly, remaining 59% of TFs show as well a significant activity, but that cannot be classified in the previous groups. The results are shown in Fig. 4.1, panel a.

Then, to find the key regulators candidate to be responsible for boundaries reformation exiting mitosis, we calculated the total explained variance ($EV \simeq 45\%$) and the reduced EV, as seen in section 3.3.4, and we selected the 5% most important TFs, that are shown in Fig. 4.1, panel b.

Importantly, some of them have been reported to be involved in cell differentiation or cell proliferation and growth, such as HOXA9, STAT3 and IL21, while mutation or deficiencies of some others (GLIS3, SOX9, MYF5) are known to may cause diseases [63], confirming the link between changes in topological domains and disorders, already mentioned in section 1.3. In addition, we also found ESRRβ, already seen in chapter 3

to be a bookmarking factor.

Furthermore, our model detected DR1: the protein encoded by the corresponding gene has been reported as a repressor which binds the TBP promoter, establishing a mechanisms of altered DNA conformation that affects the rate of RNA Pol II transcription [63].

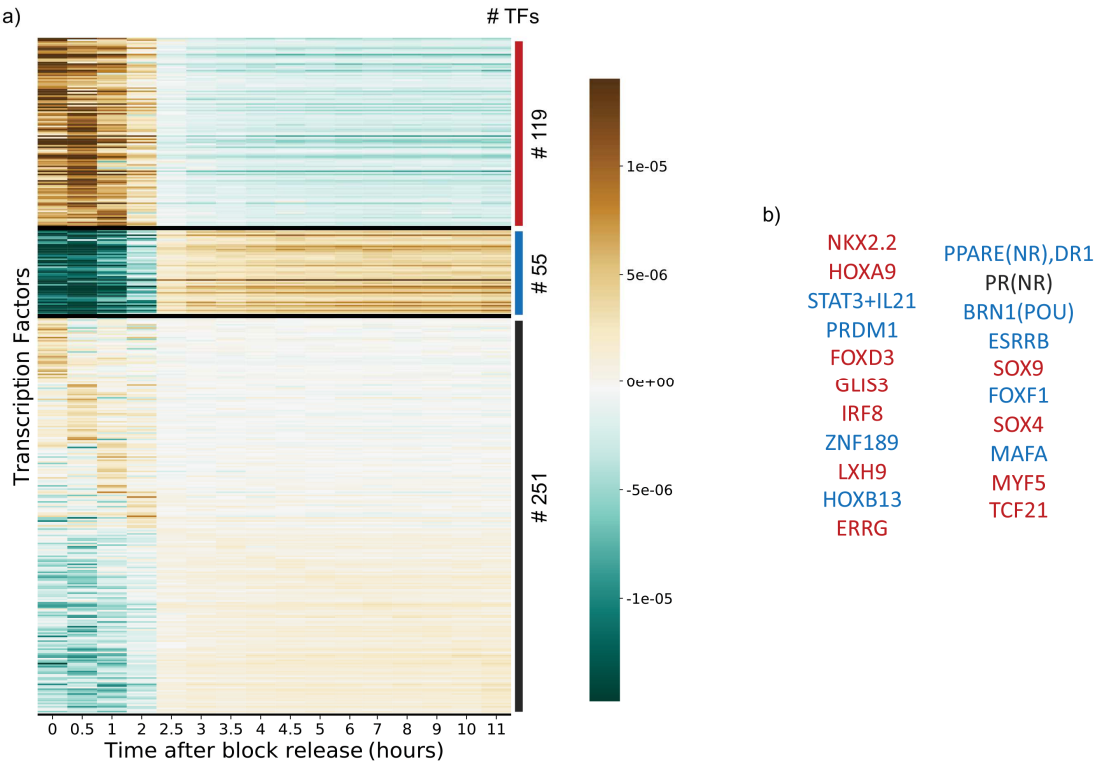


Figure 4.1. Transcription factors dynamics for boundaries reformation. **a**: TFs can be divided in clusters according to the profile of their activity over time after the release of the mitotic arrest. Here, each row corresponds to a TF, and the color scale represents the level of activity, as shown by the colorbar on the right, where also the number of TFs of each cluster is reported. Finally, an identification colour is assigned to each cluster (vertical bars on the right). **b**: Names of most important 5% of TFs have been reported, according to their reduced explained variance. The colour of each name corresponds to the identification colour used in panel a.

4.3 Conclusions

In conclusion, we aimed to infer regulators responsible for reformation of topological associated domains boundaries after mitosis in HeLa cells. To do so, we applied a linear model combining motifs of 425 human transcription factors and Hi-C datasets obtained at different timepoints during mitosis and early G1. This allowed us to derive a set of regulators candidate to drive the boundaries reformation post-mitosis, establishing a certain level of insulation between domains, and then ensuring one of the three-dimensional features of the chromatin.

5

Discussion and future perspectives

In this thesis, we tried to uncover the existing link between the three-dimensional organization of chromatin and the regulation of the transcriptional machinery, by combining computational analyses and mathematical modeling of data from high-throughput experiments, such as RNA-Seq and Hi-C.

We applied a diffusion-based method to detect diffusive associated domains (DADs) on Hi-C datasets obtained at different stages of Drosophila Melanogaster (Dmel) embryogenesis. Although mitosis is a phase for which chromatin has been mostly described as unstructured and transcriptionally silent in literature, our computational analyses showed important evidences suggesting the presence of a "backbone" of structural features in mitotic chromosomes. In fact, almost 68% of DADs are conserved from mitosis to latest embryo developmental stages.

A similar analyis was perfomed on Hi-C datasets at different timepoints exiting mitosis obtained from populations of synchronized HeLa cells, where we showed a conservation level of DADs of almost 66% between chromatin in mitotic state and 11 hours after the release of the mitotic block.

These results highlight the existence of a non random mitotic structure, that is kept until latest stages of cell development. Importantly, in the light of the strict relationship between the three-dimensional organization of the chromatin and the transcriptional activity, this may suggest the occurrence of gene expression during mitosis and early interphase. This important point has been further investigated and addressed in our studies.

To do that, we reanalyzed time-dependent EU-RNA-Seq data on synchronized cell populations by a mitotic arrest, to correct for the progressive desynchronization of cells after block release. This allowed us to estimate gene expression profiles with respect to a cell-cycle pseudotime with an explicitly defined transition between mitosis and early G1 phase. Remarkably, we identified a set of genes that show a very early wave of transcription reactivation during mitosis. However, the majority of genes showed a peak of transcription at telophase or during the transition between mitosis and G1.

Next, we estimated TF activity dynamics of 332 expressed TFs by fitting a multiple linear model to the deconvolved gene expression profiles. We observed time-dependent waves of TF activities suggesting an intrinsic TF hierarchy with respect to their role on transcription reactivation after mitosis. In addition, we investigated whether TFs previously reported to bind mitotic chromosomes were responsible for a faster reactivation

kinetics. Surprisingly, we did not find a strong correlation between genes regulated by mitotic bound TFs and the speed of reactivation. However, our approach allowed us to identify around 60 TFs that are highly active during mitosis and represent new candidates of mitotic bookmarking factors. Therefore, we predict that the interactions of these factors with their specific target sites during mitosis are the molecular mechanisms responsible for mitotic transcription.

Moreover, we reconstructed a core regulatory network underlying the dynamics of transcription reactivation exiting mitosis, by selecting the key TFs that showed the highest explanatory power in our multiple linear regression model. Then, we propose a list of candidates to be the crucial players in the process of reactivating the gene expression in the first stages of the interphase, ensuring the cell identity. We predict that these TFs could represent relevant therapeutic targets to control cell proliferation.

In addition, our work aimed to infer regulators responsible for reformation of topological associated domains boundaries after mitosis in the synchronized population of HeLa cells that we mentioned in the first part and that we used to detect DADs. To do so, we developed a linear model combining motifs of 425 human transcription factors and the Hi-C datasets obtained at different timepoints during mitosis and early G1. This preliminary analysis allowed us to derive a set of regulators candidate to drive the boundaries reformation post-mitosis, establishing a certain level of insulation between domains, and then ensuring one of the three-dimensional features of the chromatin.

However, we are currently working to improve our method, and one of the goals is to integrate ATAC-Seq data. In fact, one of the limitations we encountered is that several TFs can potentially bind thousands of binding sites (averagely 3000) for each of the boundaries we analyzed, that have a length of $32kb$. This could affect the results, taking into account a huge amount of potential bindings regardless the actual accessibility of the binding sites. By using ATAC-Seq data we could exclude from the analysis the motifs compatible with non accessible sequences, improving the accuracy of the analysis.

Also, imaging experiments by using fluorescent microscopy would be helpful to assess whether the new factors that we proposed for being mitotic active can effectively bind the mitotic chromatin. As well, Chip-Seq analyses could be performed to check if these TFs specifically bind chromatin during mitosis.

Certainly, one of the limitations of our study was that we tried to investigate the correlation between transcription reactivation kynetics and boundaries reformation exiting mitosis by using datasets from different cell lines. Ideally, time-dependent EU-RNA-Seq experiments on HeLa cells or, alternatively, time-dependent Hi-C experiments on HUH7 cells should be performed to further explore the connection between gene expression and topological domains establishment. For example, analyzing both the dynamics of genes and boundaries they belong to, we could detect temporal shifts, which would shed light on the cause-consequence nexus, that is still unknown. However, as a preliminary test, this analysis will be performed in the next future by using the available data on HUH7 and HeLa cells, taking into account the high conservation level of boundaries across different cell lines.

Bibliography

[1] Alberts, Bruce, et al. "Molecular Biology of the Cell 4th edn (New York: Garland Science)." Ann Bot 91 (2002): 401.

[2] Bannister, Andrew J., and Tony Kouzarides. "Regulation of chromatin by histone modifications." Cell research 21.3 (2011): 381.

[3] Cremer, Thomas, and Christoph Cremer. "Chromosome territories, nuclear architecture and gene regulation in mammalian cells." Nature reviews genetics 2.4 (2001): 292.

[4] Bulger, Michael, and Mark Groudine. "Functional and mechanistic diversity of distal transcription enhancers." Cell 144.3 (2011): 327-339

[5] Han, Jinlei, Zhiliang Zhang, and Kai Wang. "3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering." Molecular Cytogenetics 11.1 (2018): 21.

[6] Dekker, Job, Marc A. Marti-Renom, and Leonid A. Mirny. "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data." Nature Reviews Genetics 14.6 (2013): 390.

[7] Nagano, Takashi, et al. "Single-cell Hi-C reveals cell-to-cell variability in chromosome structure." Nature 502.7469 (2013): 59

[8] Rao, Suhas SP, et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." Cell 159.7 (2014): 1665-1680.

[9] Sarnataro, Sergio, et al. "Structure of the human chromosome interaction network." PloS one 12.11 (2017): e0188201.

[10] Lieberman-Aiden, Erez, et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." science 326.5950 (2009): 289-293.

[11] Dixon, Jesse R., et al. "Topological domains in mammalian genomes identified by analysis of chromatin interactions." Nature 485.7398 (2012): 376.

[12] Nora, Elphège P., et al. "Spatial partitioning of the regulatory landscape of the X-inactivation centre." Nature 485.7398 (2012): 381.

[13] Zhan, Yinxiu, et al. "Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes." Genome research 27.3 (2017): 479-490.

[14] Fraser, James, et al. "Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation." Molecular systems biology 11.12 (2015).

[15] Redolfi, Josef, et al. "DamC reveals principles of chromatin folding in vivo without crosslinking and ligation." Nature structural and molecular biology (2019): 1

[16] Vermunt, Marit W., Di Zhang, and Gerd A. Blobel. "The interdependence of gene-regulatory elements and the 3D genome." Journal of Cell Biology 218.1 (2018): 12-26.

[17] Franke, Martin, et al. "Formation of new chromatin domains determines pathogenicity of genomic duplications." Nature 538.7624 (2016): 265.

[18] Lupiáñez, Darío G., et al. "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions." Cell 161.5 (2015): 1012-1025.

[19] Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. "Structural variation in the 3D genome." Nature Reviews Genetics 19.7 (2018): 453.

[20] Yang, Lixing, et al. "Diverse mechanisms of somatic structural variations in human cancer genomes." Cell 153.4 (2013): 919-929.

[21] Timmers, HT Marc, and C. Peter Verrijzer. "Mitotic Chromosomes: Not So Silent After All." Developmental cell 43.2 (2017): 119-121.

[22] Kadauke, Stephan, and Gerd A. Blobel. "Mitotic bookmarking by transcription factors." Epigenetics & chromatin 6.1 (2013): 6.

[23] Festuccia, Nicola, et al. "Mitotic bookmarking in development and stem cells." Development 144.20 (2017): 3633-3645.

[24] Alabert, Constance, and Anja Groth. "Chromatin replication and epigenome maintenance." Nature reviews Molecular cell biology 13.3 (2012): 153.

[25] Halazonetis, Thanos D., Vassilis G. Gorgoulis, and Jiri Bartek. "An oncogene-induced DNA damage model for cancer development." science 319.5868 (2008): 1352-1355.

[26] De, Subhajyoti, and Franziska Michor. "DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes." Nature biotechnology 29.12 (2011): 1103.

[27] Ma, Yiqin, Kiriaki Kanakousaki, and Laura Buttitta. "How the cell cycle impacts chromatin architecture and influences cell fate." Frontiers in genetics 6 (2015): 19.

[28] Palozola, Katherine C., et al. "Mitotic transcription and waves of gene reactivation during mitotic exit." Science 358.6359 (2017): 119-122.

[29] Kadauke, Stephan, et al. "Tissue-specific mitotic bookmarking by hematopoietic transcription factor GATA1." Cell 150.4 (2012): 725-737.

[30] Caravaca, Juan Manuel, et al. "Bookmarking by specific and nonspecific binding of FoxA1 pioneer factor to mitotic chromosomes." Genes & development 27.3 (2013): 251-260.

[31] Hettich, Johannes, and J. Christof M. Gebhardt. "Transcription factor target site search and gene regulation in a background of unspecific binding sites." Journal of theoretical biology 454 (2018): 91-101.

[32] Raccaud, Mahe, et al. "Mitotic chromosome binding predicts transcription factor properties in interphase." Nature communications 10.1 (2019): 487.

[33] Hug, Clemens B., et al. "Chromatin architecture emerges during zygotic genome activation independent of transcription." Cell 169.2 (2017): 216-228.

[34] Abramo, Kristin, et al. "A chromosome folding intermediate at the condensin-to-cohesin transition during telophase." Nature cell biology 21.11 (2019): 1393-1402.

[35] Hirano, Tatsuya. "Condensin-based chromosome organization from bacteria to vertebrates." Cell 164.5 (2016): 847-857.

[36] Wood, Andrew J., Aaron F. Severson, and Barbara J. Meyer. "Condensin and cohesin complexity: the expanding repertoire of functions." Nature Reviews Genetics 11.6 (2010): 391-404.

[37] Liu, Enze, Lang Li, and Lijun Cheng. "Gene Regulatory Network Review." (2019): 155-164.

[38] Ristevski, Blagoj. "A survey of models for inference of gene regulatory networks." Nonlinear Anal Model Control 18.4 (2013): 444-465.

[39] Maksimov, Nikolai. "Algorithms for gene regulatory networks reconstruction." (2015).

[40] Delgado, Fernando M., and Francisco Gómez-Vela. "Computational methods for Gene Regulatory Networks reconstruction and analysis: A review." Artificial intelligence in medicine 95 (2019): 133-145.

[41] Balwierz, Piotr J., et al. "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs." Genome research 24.5 (2014): 869-884.

[42] Aldous, David, and James Fill. "Reversible Markov chains and random walks on graphs." (1995).

[43] https://github.com/michaelschaub/PartitionStability

[44] Delvenne, J-C., Sophia N. Yaliraki, and Mauricio Barahona. "Stability of graph communities across time scales." Proceedings of the national academy of sciences 107.29 (2010): 12755-12760.

[45] Delvenne, Jean-Charles, et al. "The stability of a graph partition: A dynamics-based framework for community detection." Dynamics On and Of Complex Networks, Volume 2. Birkhäuser, New York, NY, 2013. 221-242.

[46] Newman, Mark EJ. "Modularity and community structure in networks." Proceedings of the national academy of sciences 103.23 (2006): 8577-8582.

[47] Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.

[48] Wingett, S., et al. "HiCUP: pipeline for mapping and processing Hi-C data. F1000Res 4: 1310." (2015)

[49] Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." Nature methods 9.4 (2012): 357.

[50] Imakaev, Maxim, et al. "Iterative correction of Hi-C data reveals hallmarks of chromosome organization." Nature methods 9.10 (2012): 999.

[51] Servant, Nicolas, et al. "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing." Genome biology 16.1 (2015): 259.

[52] Rand, William M. "Objective criteria for the evaluation of clustering methods." Journal of the American Statistical association 66.336 (1971): 846-850.

[53] Hubert, Lawrence, and Phipps Arabie. "Comparing partitions." Journal of classification 2.1 (1985): 193-218.

[54] Deluz, Cédric, et al. "A role for mitotic bookmarking of SOX2 in pluripotency and differentiation." Genes & development 30.22 (2016): 2538-2550.

[55] Festuccia, Nicola, et al. "Transcription factor activity and nucleosome organization in mitosis." Genome research 29.2 (2019): 250-260.

[56] Luo, Huaibing, et al. "Cell identity bookmarking through heterogeneous chromatin landscape maintenance during the cell cycle." Human molecular genetics 26.21 (2017): 4231-4243.

[57] Javasky, Elisheva, et al. "Study of mitotic chromatin supports a model of bookmarking by histone modifications and reveals nucleosome deposition patterns." Genome research 28.10 (2018): 1455-1466.

[58] Ginno, Paul Adrian, et al. "Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape." Nature communications 9.1 (2018): 1-12.

[59] Arnold, Phil, et al. "MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences." Bioinformatics 28.4 (2012): 487-494.

[60] Palozola, Katherine C., Jonathan Lerner, and Kenneth S. Zaret. "A changing paradigm of transcriptional memory propagation through mitosis." Nature Reviews Molecular Cell Biology 20.1 (2019): 55-64.

[61] Lee, Yoon-Jin, et al. "HSF1 as a mitotic regulator: phosphorylation of HSF1 by Plk1 is essential for mitotic progression." Cancer research 68.18 (2008): 7550-7560.

[62] Teves, Sheila S., et al. "A stable mode of bookmarking by TBP recruits RNA polymerase II to mitotic chromosomes." Elife 7 (2018): e35621.

[63] Stelzer, Gil, et al. "The GeneCards suite: from gene data mining to disease genome sequence analyses." Current protocols in bioinformatics 54.1 (2016): 1-30.

[64] Festuccia, Nicola, et al. "Mitotic binding of Esrrb marks key regulatory regions of the pluripotency network." Nature cell biology 18.11 (2016): 1139-1148.

[65] Naumova, Natalia, et al. "Organization of the mitotic chromosome." Science 342.6161 (2013): 948-953.

[66] Akaike, Hirotugu. "A new look at the statistical model identification." IEEE transactions on automatic control 19.6 (1974): 716-723.

[67] Schwarz, Gideon. "Estimating the dimension of a model." The annals of statistics 6.2 (1978): 461-464.

[68] Crane, Emily, et al. "Condensin-driven remodelling of X chromosome topology during dosage compensation." Nature 523.7559 (2015): 240-244.

[69] Ramírez, Fidel, et al. "High-resolution TADs reveal DNA sequences underlying genome organization in flies." Nature communications 9.1 (2018): 1-15.

[70] Pombo, Ana, and Niall Dillon. "Three-dimensional genome architecture: players and mechanisms." Nature reviews Molecular cell biology 16.4 (2015): 245-257.

[71] Yu, Miao, and Bing Ren. "The three-dimensional organization of mammalian genomes." Annual review of cell and developmental biology 33 (2017): 265-289.

[72] Kent, W. James, et al. "BigWig and BigBed: enabling browsing of large distributed datasets." Bioinformatics 26.17 (2010): 2204-2207.

[73] https://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes

[74] Heinz, Sven, et al. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." Molecular cell 38.4 (2010): 576-589.

[75] Bailey, Timothy L., and Charles Elkan. "The value of prior knowledge in discovering motifs with MEME." Ismb. Vol. 3. 1995.

Université de Strasbourg

Sergio SARNATARO

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

Cinétique de réactivation de la transcription et réorganisation de la structure de la chromatine après la mitose

Résumé en français:
Dans cette thèse, nous avons tenté de découvrir le lien existant entre l'organisation tridimensionnelle de la chromatine et la régulation de la machinerie transcriptionnelle, en combinant des analyses informatiques et la modélisation mathématique de données provenant d'expériences à haut débit telles que le RNA-Seq et le Hi-C. Ces recherches visent principalement à mettre en évidence les principaux régulateurs responsables de la réactivation de la transcription à la sortie de la mitose, et à identifier les facteurs les plus importants de la réorganisation structurelle de la chromatine dans le cycle cellulaire.
Mots clés: Organisation tridimensionnelle de la chromatine, régulation de la machinerie transcriptionnelle, Hi-C, RNA-Seq, mitose.

Resume in English:
In this thesis, we tried to uncover the existing link between the three-dimensional organization of chromatin and the regulation of the transcriptional machinery, by combining computational analyses and mathematical modeling of data from high-throughput experiments, such as RNA-Seq and Hi-C. In particular, the aims of this research are to reveal the key regulators responsible for the reactivation of the transcription exiting the mitosis, and to infer the most important factors driving the structural reorganization of the chromatin through the cell-cycle.
Key words: Three-dimensional organization of chromatin, regulation of the transcriptional machinery, RNA-Seq, Hi-C, mitosis.