Analysis of chromosome conformation data and application to cancer

Nicolas Servant

PhD Thesis

Analysis of chromosome conformation data and application to cancer

Nicolas Servant

Institut Curie, INSERM U900, Mines ParisTech, PSL University

Ecole Doctorale "Complexité du vivant" - UPMC

Supervisors: Dr. Emmanuel Barillot[1,2,3,6], Pr. Edith Heard[1,4,5,6]

[1]Institut Curie [2]INSERM, U900 [3]Mines ParisTech [4]CNRS UMR3215 [5]INSERM U934 [6]Paris Sciences Lettre Research University, Paris, France

2017 Novembre 22th

Membres du jury de Thèse

1. Président : Pr. Stéphane Le Crom
2. Rapporteur : Dr. Carl Herrmann
3. Rapporteur : Dr. Jean-Charles Cadoret
4. Examinateur : Dr. Morgane Thomas-Chollier
5. Examinateur : Dr. Daan Noordermeer
6. Examinateur : Pr. Daniel Gautheret
7. Directeur : Dr. Emmanuel Barillot
8. Directeur : Pr. Edith Heard

Remerciements

Nombreuses sont les personnes qui me demandent les raisons pour lesquelles je me suis lancé dans ce projet, si tard, et sans que cela ne soit requis dans ma position actuelle. Après 13 années passées dans la Recherche, je pense simplement en avoir eu besoin. Envie de tenter l'aventure, de relever un défi, ou tout simplement d'aller au bout des choses. Une chose est certaine, je suis heureux d'avoir pris cette décision, et d'avoir mené ce projet à son terme. Ces années m'ont permis d'aborder un sujet passionnant et tout ce travail n'aurait pas été possible sans de nombreuses personnes que je tiens à remercier ici.

Les permiers sont évidemment les membres de ma Troupe ; Laurence, Quentin et Elsa. A Laurence, merci d'être à mes cotès depuis toutes ces années, de me supporter et de m'encourager, parfois contre vents et marées. A Quentin et Elsa, merci pour vos sourires et vos mots doux. J'espère que quelque part, ces quatres années et ces longs week-ends passés à me voir mener ce projet vous aurons appris quelque chose et transmis certaines valeures.

Je souhaite ensuite remercier les membres de mon jury de thèse, qui ont accepté de me conseiller, d'évaluer ce travail et d'être présent à son dénouement.

Je tiens également sincérement à remercier Edith, et Emmanuel. Depuis tout ce temps que je travaille à leur coté, ce sont eux qui m'ont permis de m'accomplir professionnellement. Emmanuel, merci de m'avoir toujours tiré vers le haut depuis ce jour de 2004 ou je suis arrivé en stage, et jusqu'à aujourd'hui. Enfin Edith, je pense que personne ne m'a autant appris sur le plan scientifique et humain que toi. Tu représentes tout ce que la Science devrait être ; excellence, mais aussi simplicité et modestie. Travailler à vos cotés est un honneur et un plaisir, et j'éspère que cela continuera le plus longtemps possible.

Ce travail n'aurait jamais été possible sans de nombreux collaborateurs que je

ne pourrai pas tous citer. Il y a néanmoins quelques personnes sur lesquelles je veux m'attarder. Il y a tout d'abord Nelle, que je tiens particulièrement à remercier, et qui a réellement été au centre de ce travail. Merci à toi d'avoir été mon binôme durant ces quatres années. Il est claire que je n'aurai pas été capable d'accomplir ce travail sans notre collaboration et nos longs échanges. Nos carrières n'en sont qu'à leur début, et j'espère que nous aurons encore de nombreuses occasions de travailler ensemble. Je tiens également à remercier Jean-Philippe pour son implication dans tous ces projets. Ton aide, tes explications et ta motivation ont été, comme toujours, très précieuses. Enfin, j'aimerai remercier les gens qui m'ont permis d'apprendre et de progresser, avec qui j'ai travaillé sur ces projets, toujours avec le sourire et dans la bonne humeur ; Felix, Rafael, Joke, Mayra, Eric, Aurélie(s) et bien d'autres.

Sans tous les citer, je souhaiterai également remercier les membres de mon équipe. Faire cette thèse, en parallèle de ma position à l'Institut, n'a pas toujours été simple. Je tiens à les remercier pour leur compréhension et pour leur aide. J'espère avoir désormais plus de temps à leur consacrer, et je suis certain que nous avons encore de belles choses a réaliser ensemble dans les années a venir.

Je remercie également les personnes qui ont pris le temps de relire ce manuscrit et de me faire leur retour.

Enfin, j'aimerai finir par remercier tous les autres, amis, parents ou collègues, qui de prêt ou de loin, ont été à mes cotés au cours de ces années de thèse ! Merci à vous !

Résumé détaillé

L'organisation nucléaire de la chromatine n'est pas aléatoire. Sa structure est parfaitement contrôlée, suivant un modèle hiérarchique avec différents niveaux d'organisation et de compaction. A large échelle, chaque chromosome occupe son propre espace au sein du noyau. A plus fine résolution, un chromosome est subdivisé en compartiments actifs ou répressifs, caractérisés par un état de la chromatine plus ou moins compact. A l'échelle du méga-base, cette organisation hiérarchique peut encore être divisée en domaines topologiques (ou TADs), jusqu'à la caractérisation de boucle d'ADN facilitant les interactions entre promoteurs et régions régulatrices. Très brièvement, et bien que les méchanismes exactes restent à déterminer, il a récemment été démontré que l'organisation spatiale de la chromatine dans une cellule normale joue un rôle primordial dans la régulation et l'expression des gènes. L'organisation en domaines topologiques implique la présence de complexes protéiques insulateurs tel que CTCF/cohésine. Ces facteurs jouent un rôle de barrière en restreignant et favorisant les interactions entre éléments régulateurs et gènes à l'intérieur d'un domaine, tout en limitant les interactions entre domaines. De cette façon, deux régions appartenant au même domaine topologique pourront fréquemment interagir, alors que deux régions appartenant à des domaines distincts auront une très faible probabilité d'interaction.

Dans la cellule cancéreuse, l'implication de l'épigénome et de l'organisation spatiale de la chromatine dans la progression tumorale reste à ce jour largement inexplorée. Certaines études récentes ont toutefois démontré qu'une altération de la conformation de l'ADN pouvait être associée à l'activation de certains oncogènes. Même si les mécanismes exacts ne sont pas encore connus, cela démontre que l'organisation de la chromatine est un facteur important de la tumorigenèse, permettant, dans certains cas, d'expliquer les méchanismes moléculaires à l'origine de la dérégulation de certains gènes. Parmi les cas rapportés, une alération des régions insulatrices (ou frontières) entre domaines topologiques permettrait à des régions normalement éloignées spatialement de se retrouver en contact, favorisant ainsi l'activation de certains gènes. Une

caractérisation systématique de la conformation spatiale des génomes cancéreux pourrait donc permettre d'améliorer nos connaissances de la biologie des cancers.
Les techniques haut-débit d'analyse de la conformation de la chromatine sont actuellement largement utilisées pour caractériser les interactions physiques entre régions du génome. Brièvement, ces techniques consistent à fixer, digérer, puis liguer ensemble deux régions du génome spatialement proches. Les fragments d'ADN chimériques ainsi générés peuvent alors être séquencés par leurs extrémités, afin de quantifier le nombre de fois où ces régions ont été trouvées en contact. Parmi les différentes variantes de ces techniques, le Hi-C associé à un séquençage profond permet l'exploration systématique de ces interactions à l'échelle du génome, offrant ainsi une vue détaillée de l'organisation tri-dimensionnelle de la chromatine d'une population cellulaire.
Comme la majorité des applications de séquençage, un certain nombre de biais techniques comme le taux de GC, la taille des fragments de restriction ou la complexité des régions génomiques, ont été identifiés. Afin de corriger ces biais, plusieurs méthodes ont été proposées. A ce jour, les méthodes itératives dites de 'matrix balancing' sont très largement utilisées, essentiellement en raison de leur facilité de mise en oeuvre, et du fait qu'elles ne formulent aucune hypothèse a priori sur les biais à corriger.
Le travail présenté ici a pour but de développer des méthodes et outils bioinformatiques permettant l'analyse et l'exploitation des données Hi-C, et en particulier des données Hi-C issues de cellules cancéreuses. Ce travail s'articule autour de deux axes ; i) le traitement bioinformatique des données Hi-C, ii) et la normalisation de données Hi-C cancer. Enfin un dernier chapitre présente nos premiers résultats concernant l'étude de la sur-expression de l'oncogène *c-myc* et son impact sur le statut du chromosome X inactif dans les tumeurs du foie murine.

La mise en place de nouvelles applications biologiques, comme l'analyse de la conformation de la chromatine, doit s'accompagner de développements bioinformatiques permettant l'exploitation de ces données, mais également la reproductibilité des analyses d'une étude à l'autre. Parmi les applications de séquençage actuellement disponibles, le Hi-C permet de générer une quantité très importante de données pouvant atteindre plusieurs milliards de lectures de sequençage par échantillon. Le traitement de ces données nécessite donc des outils bioinformatiques dédiés, capable d'extraire de l'information en un temps raisonnable, et sur des environnements informatiques de

capacité variable. Afin de répondre à ce besoin, nous avons développé HiC-Pro, un nouvel outil bioinformatique d'analyse de données Hi-C, à la fois optimisé et flexible. HiC-Pro prend en charge les données Hi-C en sortie du séquenceur, jusqu'à l'obtention de cartes d'interaction inter et intra-chromosomales normalisées. Brièvement, HiC-Pro permet ; i) d'aligner les données sur un génome de référence, ii) de détecter les produits d'interaction valides, iii) de générer les cartes d'interactions, et iv) de normaliser ces données. Toutes ces étapes s'accompagnent de contrôles qualité dédiés permettant d'identifier de potentiels problèmes expérimentaux pendant la préparation des échantillons. HiC-Pro propose de corriger les données des biais expérimentaux en appliquant une méthode de normalisation itérative (ICE) récemment publiée, dont l'implémentation a été largement revue et optimisée.

HiC-Pro est simple d'utilisation, et permet l'analyse simultanée de plusieurs échantillons en une simple ligne de commande. Il est optimisé offrant si besoin, la possibilité d'analyser les données de façon parallélisée sur un environnement de calcul haute performance, tout en se basant sur un format de données efficace, réduisant au maximum l'espace de stockage dédié aux données processées. Enfin, HiC-pro est compatible avec tous les protocoles basés sur la technique Hi-C actuellement disponibles, incluant les protocoles sans enzyme de restriction (DNAse Hi-C, Micro-C) ainsi que les données de conformation ciblée (HiChIP, capture-C, capture Hi-C).

HiC-Pro est également le seul pipeline d'analyse Hi-C intégrant un mode allèle-spécifique. Si les génotypes parentaux sont connus et spécifiés, HiC-pro utilise une stratégie dédiée pour l'alignement des lectures de séquençage consistant à masquer les positions relatives aux SNPs hétérozygotes. Sur la base des ces informations, il est ensuite possible d'assigner les interactions allèles-spécifiques à l'un ou l'autre des génotypes parentaux. Ce mode d'analyse aboutit à la génération de cartes d'interaction propres à chacun des génotypes parentaux.

Finalement, HiC-Pro est aujourd'hui largement utilisé par la communauté scientifique, offrant un outil fiable et simple d'utilisation pour l'analyse de données Hi-C. Il est documenté, et s'appuie sur un groupe de discussion au sein duquel les utilisateurs peuvent poser leurs questions ou proposer de nouvelles fonctionnalités. HiC-pro est un outil dédié à l'analyse primaire de données Hi-C, et n'a donc pas vocation à proposer des méthodes d'analyse secondaires, telle que la détection de domaines topologiques ou la comparaison de cartes d'interaction. Toutefois, plusieurs efforts ont été réalisés

pour assurer sa compatibilité avec d'autres logiciels, notamment pour la visualisation de données Hi-C.

Dans un second temps, nous avons commencé à explorer les données de conformation de la chromatine issue de lignées tumorales. Les données cancer nécessitent en général des méthodes d'analyse dediées, permettant notamment de prendre en comptes les altérations du nombre de copies d'ADN fréquemment observées dans ces échantillons. Pour ce faire, nous avons tout d'abord mis en place un modèle de simulation permettant d'estimer les effets du nombre de copie sur les cartes d'interaction inter et intra-chromosomales. Ce modèle prend en entrée un jeu de donnés réel et des altérations génomiques à simuler, et retourne les données Hi-C attendues en présence de ces altérations génomiques. Sur la base de ces simulations, nous avons rapidement mis en évidence que les méthodes iteratives de normalisation des données Hi-C généralement utilisées, ne permettent pas de corriger efficacement les données en présence d'altération du nombre de copies d'ADN. Plus surprenant, l'application de ces méthodes peut aboutir à une surcorrection des données aboutissant à une inversion du signal entre régions amplifiées et perdues.

Les méthodes actuelles ne permettant pas de normaliser efficacement ces données, nous avons entrepris de développer une nouvelle méthode de normalisation. Toutefois, la façon d'appréhender ce type de biais dans les données Hi-C peut être sujet à discussion. En effet, la majorité des méthodes développées pour d'autres applications de séquençage de lignées tumorales ont pour but de retirer l'effet lié au nombre de copie d'ADN, considérant ce signal comme du bruit non informatif. Si cette approche fait sens pour certaines questions, il peut également être souhaitable de corriger les données Hi-C des biais expérimentaux, tout en conservant l'effet du nombre de copie. Cela pourrait permettre par exemple d'étudier plus en détails les effets du nombre de copies sur les profils d'interactions, ou de reconstruire la structure tri-dimensionnelle des génomes cancéreux. Nous avons alors développé une nouvelle approche basée sur une extension des algorithmes de correction itératifs préalablement proposés. Cette méthode repose tout d'abord sur l'identification des points de cassure délimitant les segments d'ADN et leur nombre de copies. Ces derniers peuvent soit être détectés via une source externe de données (puces à ADN, séquençage de génome complet) soit être inférés directement par une procédure de segmentation des données Hi-C que nous avons mis en place. Notre méthode permet ensuite de retirer les biais expérimentaux en proposant au choix, de

préserver ou de corriger le signal du nombre de copie. Conserver l'information du nombre de copie revient à appliquer une méthode de correction itérative localement, par segment d'ADN. Alors que la correction du nombre de copie revient à ajuster les fréquences d'interactions en fonction de la distance génomique pour chaque paire de segments d'ADN en fonction de leur nombre de copie. Nous avons finalement démontré l'intérêt de ces nouvelles approches pour les études de conformation de la chromatine en les appliquant sur plusieurs jeux de données Hi-C cancer.

Dans une troisième partie, nous nous sommes intéressés aux liens entre conformation du génome et expression dans le contexte de cancer du foie chez la souris. Pour ce faire, nous avons utilisé un modèle murin Tet-Myc induisant une sur-expression de l'oncogène *c-myc*, aboutissant à l'apparition rapide de tumeurs du foie chez la souris. Les souris Tet-Myc de fond génétique FVB ont été croisées avec une autre souche murine (Castaneous) engendrant des souris hybrides Tet-Myc FVB/Cast. Ce modèle offre donc une opportunité unique, non seulement d'étudier l'impact de l'expression de *c-myc* sur l'organisation de la chromatine des tumeurs du foie, mais également de pouvoir le faire en distinguant les génotypes parentaux et donc les deux formes (active et inactive) du chromosome X chez la femelle. En effet, le rôle du chromosome X dans la tumorigenèse a été démontré au travers de nombreuses études qui ont abouti à l'identification de plusieurs génes d'intérêt sur le chromosome X actif. Toutefois, l'implication du chromosome X inactif reste peu connu. Quelques travaux récents ont néanmoins décrit un changement d'état du chromosome X inactif dans les cancers du sein, aboutissant à l'expression anormale de certains gènes.

A partir du modèle hybride Tet-Myc, nous avons pu étudier l'expression et la structure du chromosome X de quatre échantillons tumoraux. Ces échantillons ont été comparés à des données de cellules hépatiques normales de souris pour lesquelles le chromosome inactif correspond toujours à l'allèle Castaneous, l'autre allèle (Bl6) étant muté et ne pouvant pas s'inactiver. Dans un premier temps, nous avons confirmé le mode d'action de *c-myc* dans ces échantillons. En effet, lorsque *c-myc* est induit et sur-exprimé, sa protéine est retrouvée sur une très large majorité des séquences promotrices des gènes exprimés, laissant suggérer que *c-myc* agit comme un amplificateur de l'expression globale des gènes.

Nous nous sommes ensuite focalisés sur le statut du chromosome X inactif. A partir des données d'expression de chaque allèle du chromosome X et de leur enrichissment

en marque d'histone H3K27me3, nous avons pu conclure que, dans les tumeurs du foie induites par la sur-expression de *c-myc*, le chromosome X inactif n'était pas réactivé et conservait donc son état globalement inactif. Toutefois, nous avons tout de même observé un nombre plus élevé de gènes capables d'échapper à l'inactivation du chromosome X sur certaines tumeurs, validant ainsi les observations faites sur le cancer du sein.

Nous avons alors étudié l'organisation tri-dimensionnelle du chromosome X à partir de données Hi-C. De façon surprenante, nous avons observé que dans les cellules du foie (normales et tumorales), le chromosome X inactif présente plus de structure qu'attendu, alors que le chromosome X actif possède moins de TADs que dans d'autres lignées récemment étudiées. Dans les tumeurs, les différences d'organisation entre X inactif et actif semblent encore moins importantes. De plus, il a récemment été décrit que le chromosome X inactif est organisé en méga-domaines divisant sa structure en deux larges blocs dont la frontière se situe à proximité du macro-satellite DXZ4. Dans les tumeurs du foie induites par la sur-expression de *c-myc*, nous avons observé une diminution du signal au locus DXZ4 laissant supposer une perte de méga-domaines dans une sous-population cellulaire. Il est donc envisageable que le nombre anormal de gènes capables d'échapper à l'inactivation du chromosome X dans ces tumeurs soit lié aux différences de structure du chromosome X inactif observées.

A partir du modèle murin Tet-Myc hybrid, nous avons ainsi pu conduire une analyse poussée, allèle spécifique, de l'expression gènique, et de l'état de la chromatine du chromosome X inactif dans les tumeurs du foie induites par la sur-expression de *c-myc*. Nos premiers résultats confirment que le X inactif à une structure altérée dans ces tumeurs, associée à un état transcriptionel plus labile. Le projet est néanmoins toujours en cours. Des analyses complémentaires sont nécessaires pour valider ces observations.

Pour conclure, ces travaux ont permis de mettre en évidence l'importance d'utiliser des méthodes efficaces et dédiées à l'analyse de données Hi-C en général, et plus particulièrement aux données Hi-C cancer. Nous avons développé de nouveaux outils et approches bioinformatiques permettant d'adresser en partie les enjeux de ces analyses. Toutes les méthodes et outils développés sont accessibles à la communauté scientifique. L'étude de la conformation spatiale de la chromatine des cancers est une approche prometteuse qui devrait nous permettre dans les années à venir de mieux appréhender

les mécanismes moléculaires à l'origine de la dérégulation des cellules cancéreuses. De nombreux défis restent à être relevés dans l'espoir que ces études aboutissent un jour à une meilleure prise en charge des patients.

Abstract

The chromatin is not randomly arranged into the nucleus. Instead, the nuclear organization is tightly controlled following different organization levels. Recent studies have explored how the genome is organized to ensure proper gene regulation within a constrained nuclear space. However, the impact of the epigenome, and in particular the three-dimensional topology of chromatin and its implication in cancer progression remain largely unexplored. As an example, recent studies have started to demonstrate that defects in the folding of the genome can be associated with oncogenes activation. Although the exact mechanisms are not yet fully understood, it demonstrates that the chromatin organization is an important factor of tumorigenesis, and that a systematic exploration of the three-dimensional cancer genomes could improve our knowledge of cancer biology in a near future. High-throughput chromosome conformation capture methods are now widely used to map chromatin interaction within regions of interest or across the genome. The Hi-C technique empowered by next generation sequencing was designed to explore intra and inter-chromosomal contacts at the whole genome scale and therefore offers detailed insights into the spatial arrangement of complete genomes.

The aim of this project was to develop computational methods and tools, that can extract relevant information from Hi-C data, and in particular, in a cancer specific context. The presented work is divided in three parts. First, as many sequencing applications, the Hi-C technique generates a huge amount of data. Managing these data requires optimized bioinformatics workflows able to process them in reasonable time and space. To answer this need, we developped HiC-Pro, an optimized and flexible pipeline to process Hi-C data from raw sequencing reads to normalized contact maps. HiC-Pro maps reads, detects valid ligation products, generates and normalizes intra- and inter-chromosomal contact maps. In addition, HiC-Pro is compatible with all current Hi-C-based protocols. HiC-Pro is now widely used by the community, providing an easy and efficient way of processing raw Hi-C data.

Then, we investigated the ability of current normalization methods to correct cancer Hi-C data from systematic biases. We therefore proposed a simulation model, to estimate the effect of copy number variants on Hi-C contact maps. Using these simulated data, we demonstrated that the current approaches, while very effective on fully diploid genome, fail to correct for unwanted effects in the presence of copy number variations. We therefore proposed a simple extension of matrix balancing methods that properly models the copy-number variation effect. Our approach can either retain the copy-number variation effect or remove it. We showed that this leads to better downstream analysis of the three-dimensional structure of rearranged genome.

Finally, we started to investigate the effect of the *c-myc* oncogene over-expression on the epigenetic status of the inactive X chromosome in mouse liver tumors. Using an hybrid Tet-Myc mouse model that over-express the *c-myc* oncogene and a normal Xist +/- hybrid mouse model, we explored the chromatin conformation, the gene expression and the histone modifications of *myc*-induced liver tumors and normal liver cells using allele-specific analysis. Our first results confirm that the inactive X chromosome is transcriptionally more labile, with more expression from the inactive X in some tumor samples. Surprisingly, the Hi-C data of both normal and tumor liver samples reveal more TAD-like structures on the inactive X chromosome, while the active X chromosome has less TADs than previously observed on other cell types. In addition, the mega-domains boundary of the inactive X chromosomes described in proliferative cells is not clearly observed in liver tumor cells. It is therefore tempting to associate the changes observed on the inactive X chromosome organization with the changes observed at the expression level in the liver tumor induced by *c-myc* over-expression.

Taken together, our results highlight the importance of using efficient methods for Hi-C data analysis in general and for cancer Hi-C data in particular. We have developped new bioinformatics approaches and tools that address part of these challenges. All methods and tools are available and can be used by the community. We also started to analyse conformation data from *c-myc* inducible tumor model in order to better characterize the link between topology and gene regulation in cancer. We hope that this work therefore paves the way for further exploration of the three-dimensional architecture of cancer genomes.

Contents

List of Figures

LIST OF FIGURES

1

Introduction

1.1 Epigenetic and diseases

The term *'epigenetics'* was introduced by Conrad Waddington in the early 1940s to refer to the molecular mechanisms regulating the expression of a genotype into a specific phenotype. According to Waddington, *'Epigenetics is a landscape in which a cell can go down different pathways and have a different fate according to the interactions between genes and their environment'*. In the 1970s-1980s, epigenetics included the notion of heritability and was defined as *'the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence'* (Ct and Morris (2001)). After decades of debate and research, epigenetics is today commonly used to describe chromatin-based modifications that regulate information from the DNA template.

1.1.1 Epigenetics and regulation

1.1.1.1 The chromatin fibre

The genome of eukaryotes is composed of several chromosomes, each containing a linear molecule of DNA. Although the number and size of chromosomes can vary between species, their structure remains the same in all eukaryotes. Each human cell contains approximatively 2 meters of DNA that are compacted in a 10 microns nucleus. Understanding how the DNA molecule can be fitted in such a small space without getting

tangled up is a fascinating question that have occupied researchers for many years. The packaging of DNA involves specialized proteins that bind and fold the DNA molecule, providing different levels of organization. But amazingly, although the DNA is extremely folded, it is also compacted in such a way that many proteins can access it to promote cellular functions such as replication, DNA repair or transcription.

The compaction of eukaryote genomes starts when 147 base pairs (bp) of DNA are wrapped around an histone octamer to form a nucleosome. The nucleosomes are then packaged, with linker DNA around 20-50bp in length, into a thread known as chromatin. The nucleosomes are formed by the core histone proteins H2A, H2B, H3 and H4. Note that some histone variants can replace the core histones in some cases (see Kornberg and Lorch (1999) for a review). The histone H1 is not part of the nucleosome structure, but instead binds the linker DNA region, helping stabilize the chromatin fibre. This level of nucleosome compaction is sometimes described as a 10 nm 'beads on a string' chromatin fibre (Oudet et al. (1975)). This first packaging shortens the DNA length approximately six fold, but the chromatin is still too long to fit into the nucleus. The chromatin fibre is therefore coiled again into a 30 nm fibre structure called a 'solenoid', containing six nucleosomes per turn. This fibre is finally looped and coiled again, leading to the familiar shape of chromosomes (Figure 1.1).

In 1928, Emil Heitz first described that active euchromatin and inactive heterochromatin occupy distinct nuclear environment (Heitz (1928)). The euchromatin characterized regions where the DNA is accessible, usually presented as open chromatin, associated with active transcriptional states. Whereas the heterochromatin is compacted and therefore difficult to access. The heterochromatin regions are usually associated with repeated elements, repressed genes and can be further divided into constitutive or facultative heterochromatin. Constitutive heterochromatin is mainly composed of telomeres and centromeres and refers to regions which are always condensed. Facultative heterochromatin can switch between different states in a development-specific manner, losing its compacted structure and becoming transcriptionally active. A typical example is the inactivation of the X chromosome in mammalian female cells (Wutz (2011)).

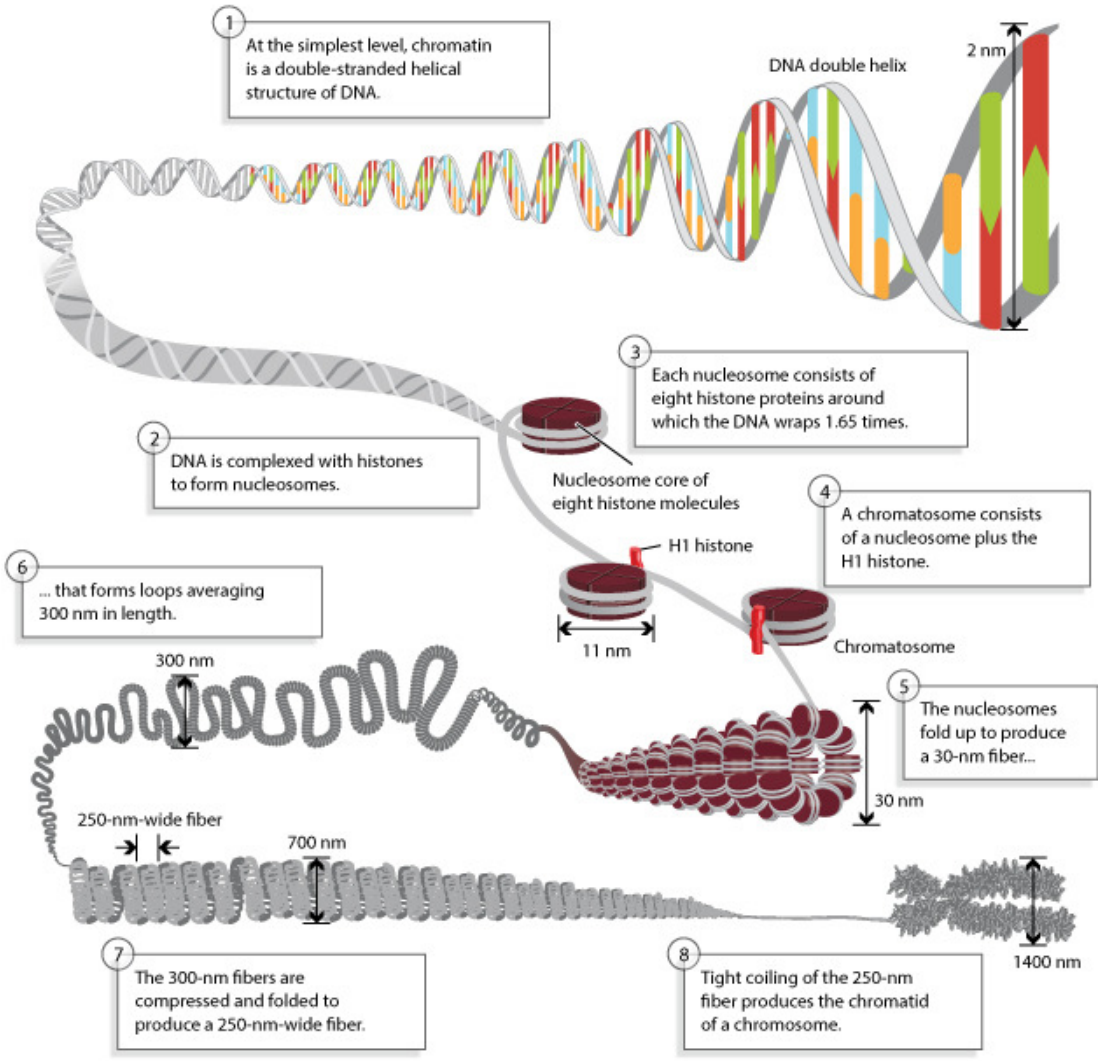


Figure 1.1: Folding the genome - The DNA molecule is coiled into a chromatin fibre with different level of compaction. First, the DNA is wrapped around histones proteins to form the nucleosomes. The chromatin fibre, composed of histones and DNA, is then coiled into a 30 nm structure, which is looped and coiled again to produce a 250 nm-wide fibre that will finally be condensed to form the chromosome. From Annunziato (2008).

The chromatin folding varies during the life cycle of the cell. In non-dividing cells (interphase), the euchromatin is decondensed and mainly appears as 30 nm chromatin fibres. This open structure allows genes to be transcribed, and DNA to be replicated in preparation for cell division. Then, when cells start to divide, their chromosomes become highly condensed, and the transcription is globally silenced.

Finally, the nucleosome positioning along with histone variants defines the primary structure of chromatin. However, the chromatin structure itself can vary from a genomic position to another, or between different cell types. These specific chromatin configurations are complex and maintained through epigenetic controls involving the chromatin machinery. Both the DNA and the histones can be subjected to a diversity of chemical modifications, therefore enabling a first level of epigenetic diversity (see Zhou et al. (2011) for a review).

1.1.1.2 DNA methylation

The first epigenetic mechanism and the most widely studied, is the DNA methylation. The DNA methylation is a chemical process that adds a methyl group to the 5th carbon of cytosine (5mC). This process occurs almost exclusively in the context of CpG sites, which are DNA regions characterized by a cytosine nucleotide located next to a guanine and linked by a phosphate. The CpG sites can also clustered in CpG island which are largely associated with gene promoters. The methylation pattern is established during embryonic development by the DNA methyltransferases enzymes (*Dnmt3A*, *Dnmt3B*, *Dnmt3C*, and *Dnmt3L*), and is then transmitted through cell division by the maintenance methyltransferase *Dnmt1* (Figure 1.2). Thus, the methylation has long been seen as a stable form of epigenetic cellular memory. However, reduced levels of methylation were also observed during development in mammals. This loss of methylation can either occur passively or actively (see Chen and Riggs (2011) for a review). While passive DNA methylation is usually established through DNA replication in absence of maintenance methylation pathways, active demethylation is associated with the action of ten-eleven translocation (*TET*) proteins (Tahiliani et al. (2009)). *TET* proteins convert the 5mC methyl group into 5-hydroxymethylcytosine (5hmC) which is the first oxidative product generated during active demethylation (see Schbeler (2015) for a review, Figure 1.2a).

In normal cells, DNA methylation occurs predominantly in repetitive regions, including short and long interspersed transposable elements (SINEs, LINEs). This DNA methylation pattern ensures that transposons remain in a silenced state. At the CpG-island level, the methylation is in general associated with stable gene silencing such as genomic imprinting. Genomic imprinting is characterized by the methylation of one of

the two parental alleles, resulting in a monoallelic expression of either the paternal and maternal copy of the gene.

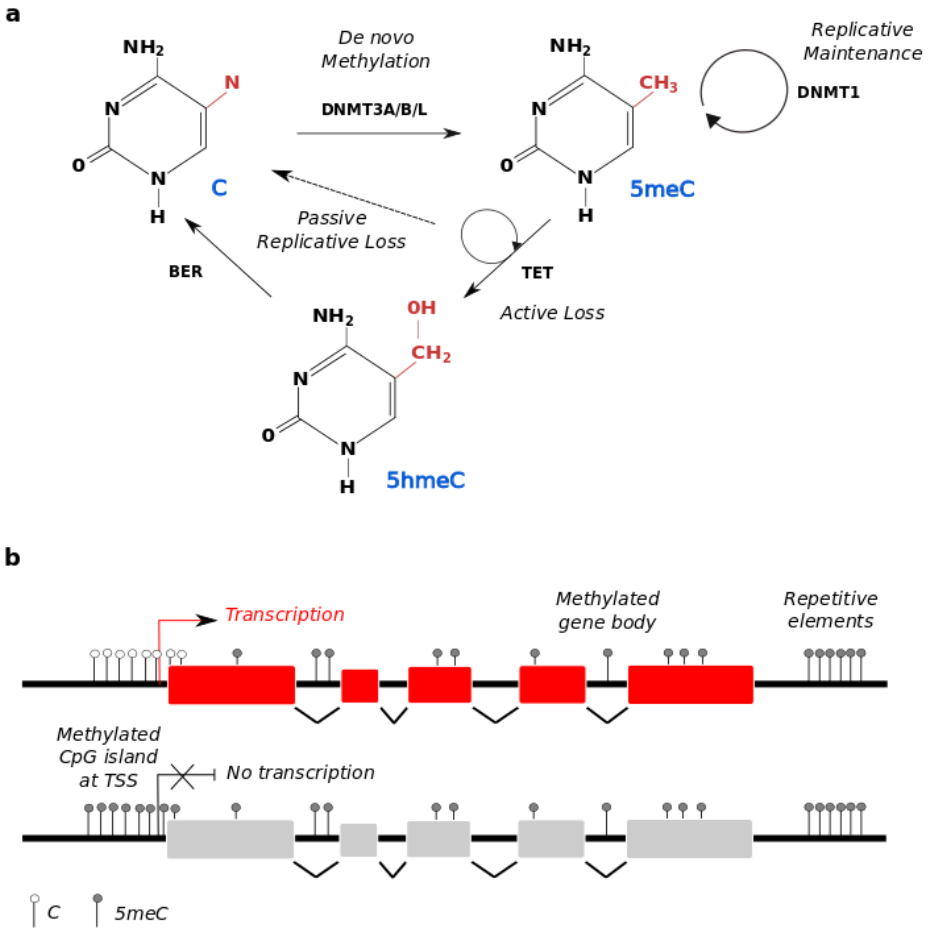


Figure 1.2: DNA methylation - a. The DNA methylation is established by DNA methyltransferases enzymes (*Dnmt3A,B,C,L*) that add a methyl group on the cytosine (5mc) or that maintain the methylated status (*DNMT1*). In addition, the methylation can be either passively or actively lost. Passive loss of methylation is associated to replication, whereas active loss requires the action of base-excision repair (BER) pathway or ten-eleven translocation (TET) proteins. TETs convert the 5mc methyl group into 5-hydroxymethylcytosine (5hmc) which is the first step of active demethylation. **b.** When it occurs at the gene promoter, the methylation is associated to gene silencing. However, the methylation of the gene body has been described to be associated with RNA splicing and elongation.

However, when the DNA methylation occurs at the gene body, it has been proposed

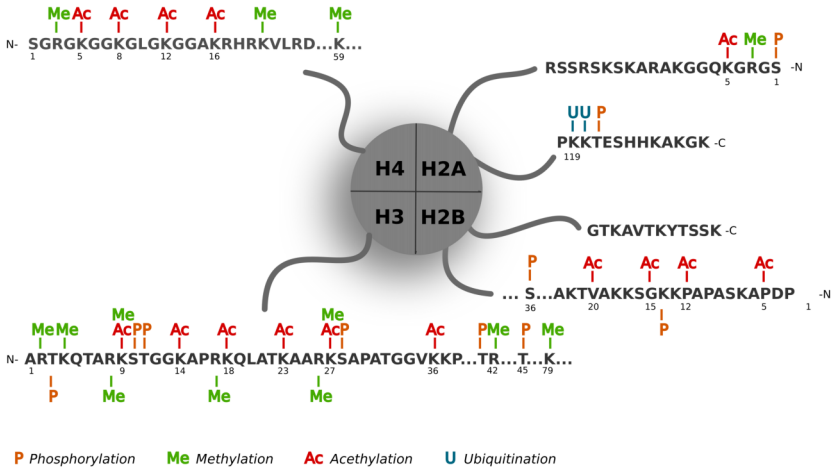
to facilitate genes transcription. This leads to an interesting paradox in which methylation at the promoter level is correlated with expression silencing, whereas methylation in the gene body is positively correlated with expression (Figure 1.2b). The function of gene body methylation is still unclear. It was first thought that this methylation was a mechanism to silence repetitive DNA elements which can be located within the gene body. Alternative roles such as elongation efficiency or RNA splicing were more recently proposed (see Jones (2012) for a review).

Finally, the DNA methylation can modulate the gene expression using different mechanisms. One obvious way is the presence of the methyl group that interferes with the binding of transcription factors involved in transcription activation. In addition, the methylation can also promote the recruitment of methyl-CpG-binding domain (MBD) proteins, which are themselves involved in the recruitment of histone-modifiers and chromatin-remodelling complexes that are able to alter the chromatin structure and regulate the transcriptional activity (see Li and Zhang (2014) for a review).

1.1.1.3 Histone modifications

Histones are key players of epigenetic, and are all subject to post-transcriptional modifications. The histone tails may undergo several types of modification, that play an important role in chromatin compaction or transcription control. These modifications are involved in the regulation of many different cellular processes including gene expression, DNA repair, replication or chromatin compaction (see Kouzarides (2007) for a review). Histone modifications involve different residues such as lysine, arginine, and serine, as well as different chemical changes such as methylation (at multiple degrees), acetylation, phosphorylation, or ubiquitination (Figure 1.3). These modifications involve the action of chromatin modifiers, able to add (writers), interpret (readers) or remove (erasers) histone modifications. Interestingly, writers are often able to bind pre-existing histone modifications, therefore leading to a positive or negative feedback control of their activity (Zhang et al. (2015)). The chemical conversions require multiple histone-modifiers enzymes (mainly methyltransferases or kinases), which are usually specific to one residue (see Biswas and Rao (2017) for a review). Readers are able to specifically recognize the histone marks leading to the recruitment of various complexes involved in gene transcription, replication or chromatin remodelling (see Musselman

et al. (2012) for a review). Most of these modifications are dynamic, and erasers, that can remove these changes, have also been identified.



Modification	Histone	Residus	Function
Acethylation	*H2A*	*K5*	*transcriptional activation*
	H2B	*K5, K12, K15, K20*	*transcriptional activation*
	H3	*K9,K14,K23*	*histone deposition*
		K9,K14,K18,K23,K27,K36	*transcriptional activation*
		K14,K18,K23	*DNA repair*
		K18	*DNA replication*
	H4	*K5,K12*	*histone deposition*
		K5,K8,K12,K16	*transcriptional activation*
		K5,K8,K12,K16	*DNA repair*
Methylation	*H1*	*K26*	*transcriptional silencing*
	H2A	*R3*	*transcriptional activation/repression*
	H3	*R17,R26,R42,K4,K36,K79*	*transcriptional activation*
		R8	*transcriptional activation/repression*
		R2,K9,K27	*transcriptional repression*
	H4	*R3*	*transcriptional activation/repression*
		K59	*transcriptional repression*
		K20	*transcriptional repression, DNA replication*
Phosphorylation	*H1*	*S27*	*transcriptional activation*
	H2A	*S1,T120*	*mitosis*
		S1,T120	*transcriptional repression*
	H2AX	*S139,T142*	*DNA repair*
		T142	*apoptosis*
	H2B	*S14*	*DNA repair*
		S14	*apoptosis*
		S36	*transcriptional activation*
	H3	*T3,S10,T11,S28*	*mitosis*
		S10,T41	*transcriptional activation*
		T45	*apoptosis*
Ubiquitination	*H2A*	*K119*	*spermatogenesis*
	H2B	*K120*	*meiosis*

Figure 1.3: Known histone modifications in mammals - Histone modifications involve different residues, histones, and chemical changes. These modifications result in changes in the accessibility of the chromatin and are associated with different regulatory functions. Adapted from https://www.cellsignal.com

The mechanisms behind histone modifications are not yet fully characterized. How-

ever, a general idea is that the histone modifications are able to affect the contact between histones of adjacent nucleosomes by changing their charge, and are therefore able to adapt the chromatin structure. Among the different histone tail modifications, acethylation has the potential to open the chromatin, and is usually associated with euchromatin, promoter accessibility and active transcription (Grunstein (1997)). By contrast, lysine mono, di and tri methylation can be linked to either active or inactive chromatin. In general, methylation of histone H3 lysine 4 (H3K4), H3K36, and H3K79 is associated with an active state, whereas H3K9, H3K27 and H4K20 modifications correlate with repression and heterochromatin (Barski et al. (2007); Kouzarides (2007)). All these histones modifications involve complex mechanisms. As a result, any aberration in these chromatin modifiers can have a strong effect on normal cell regulation and have been reported as altered in many diseases, including cancer.

1.1.2 Epigenetic alterations in common diseases and cancer

While many diseases have been shown to have a genetic component, our appreciation of the epigenetic complexity and plasticity is more recent and has significantly increased over the last few years, paving the way to potential new epigenetics therapies. It is now established that the initiation and progression of many diseases, including cancer, heart disease, metabolic or neurological disorders, are controlled by both genetic and epigenetic events (see Heerboth et al. (2014) for a review). An important concept is that epigenetic modifications are, by definition, reversible, and can modulate the cell activity by switching a gene on or off. This idea represents an incredible hope for treatments, where it may be possible to use this switch to reverse a disease phenotype.
Over the past years, several consortium have emerged in order to characterize the epigenetic landscape of normal and disease samples. Among these initiatives, the International Human Epigenome Consortium (IHEC) and the BLUEPRINT consortium provide an access to high-resolution epigenetics data for several normal and disease human cell types. To date, more than 8500 datasets related to gene expression, methylation, or histone variants have been generated. In the same way, The Cancer Genome Atlas (TCGA) is dedicated to cancer and also provides genetic and epigenetic profiles for thousand of samples. The goal of the TCGA consortium was first to characterize a large number of tumor types at the genetic level (DNA rearrangement, Single Nu-

cleotide Variations (SNV)), but rapidly, expression and methylation profiles were also generated for more than 10000 samples.

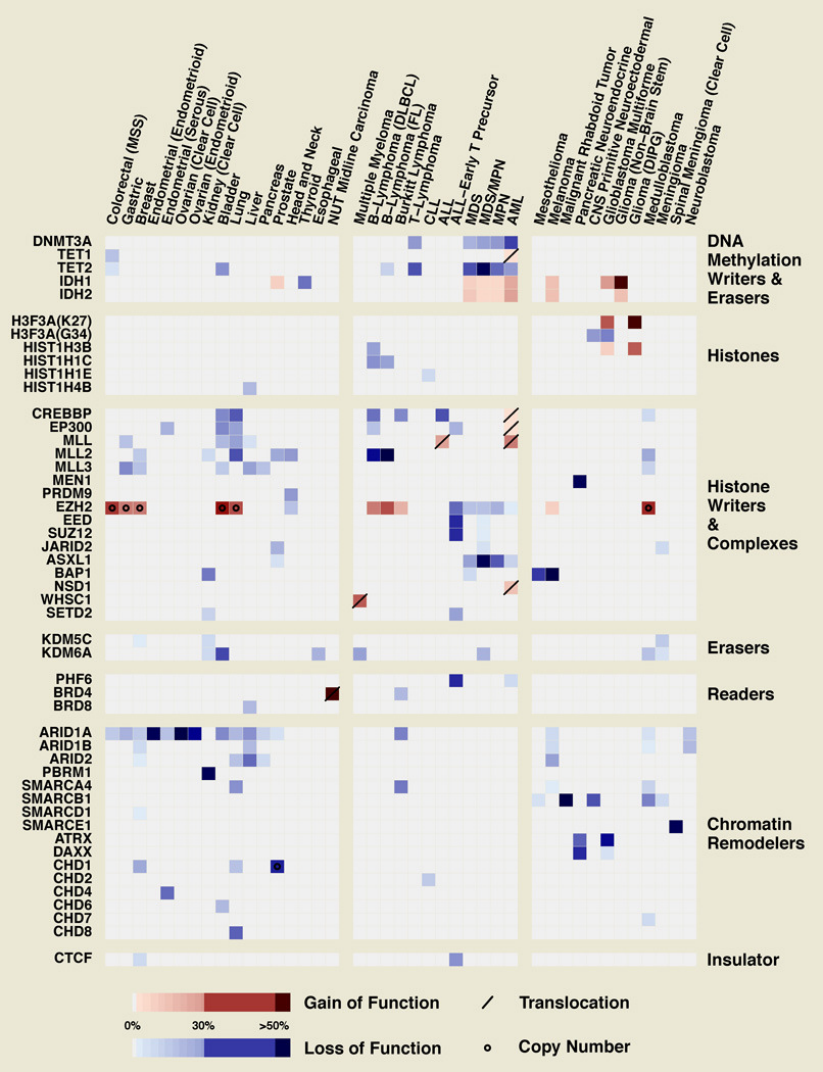


Figure 1.4: Association between mutations in driver genes and epigenetic alterations in cancer. - The mutations involving a gain of function are presented in red, whereas the ones leading to a loss of function are in blue. The darkness represents the frequency of the mutations. From Shen and Laird (2013).

Epigenetic regulations are involved in many normal cellular processes, with a strong impact on gene expression. As previously presented, the regulation of the chromatin state involves different mechanisms which can be broadly classified into methylation,

histone modifications and chromatin structure (Portela and Esteller (2010)). Alterations in these mechanisms are associated with most aspects of diseases, from their initiation to their progression (Figure 1.5).

1.1.2.1 Alteration of methylation status

Disruption in methylation is associated with many types of diseases including cancer. The loss of imprinted expression through gain or loss of DNA methylation has been described in several syndromes (see Robertson (2005); Zoghbi and Beaudet (2016) for a review). In cancer, cells are usually characterized by a global loss of DNA methylation. The most common consequence of this hypomethylation is the over-expression of oncogenes and growth factors which are involved in various hallmarks of cancer. This overall hypomethylation of tumor cells is associated with a loss of methylation from repetitive regions of the genome. Transposable elements, as LINE1 elements, are then reactivated, transcribed and can therefore integrate others sites of the genome, leading to mutations and chromosome instability. In addition, chromosomal rearrangements such as translocation or deletion were shown to be associated with hypomethylation (Esteller (2008), Figure 1.5).

On the other hand, hypermethylation of normally unmethylated CpG islands of gene promoters can also occur. The DNA hypermethylation can affect genes involved in multiple mechanisms, such as tumor suppressor genes, transcription factors, DNA repair or miRNAs pathways, and are usually specific to a cancer type (see Baylin and Jones (2011); Esteller (2008); Kulis and Esteller (2010) for a review). A typical example is the hypermethylation of the promoter of the retinoblastoma tumor-suppressor gene (RB1) involved in retinoblastoma cancers. In the same context, the promoter hypermethylation of the gene CDKN2A leads to uncontrolled cell cycle progression which is commonly observed in many different tumors (Herman et al. (1995)).

Many abnormal epigenetic events can be related to genetic alterations. This is for instance the case when tumor-specific mutations occur in genes involved in methylation pathways (Figure 1.5a). As an example, recurrent mutations in the *Dnmt3A* gene have been detected in 22% of patients with acute myeloid leukaemia (AML). The majority of mutations appears at the R882 locus, and are heterozygous (Ley et al. (2010)). Still in AML, chromosomal translocations involving the *Mll* and *Tet1* genes were also reported in a subset of patients. Mutations in isocitrate dehydrogenase 1 (*Idh1*) and *Idh2* genes

were also reported resulting in an hypermethylation phenotype (Figueroa et al. (2010)).

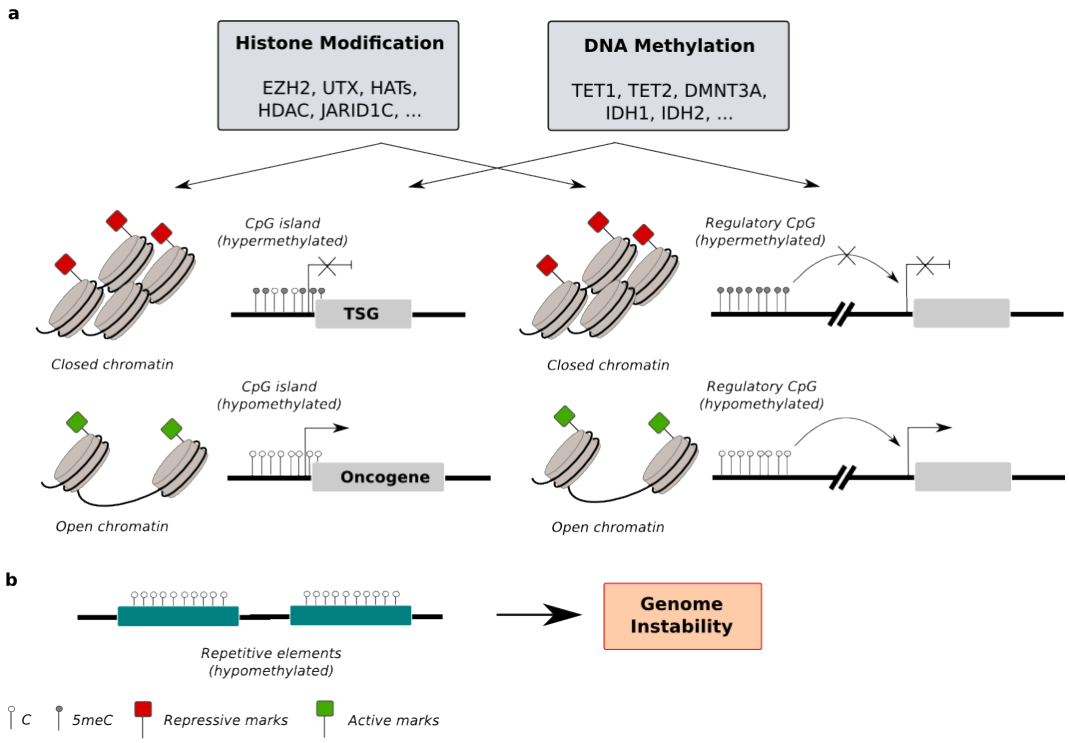


Figure 1.5: Methylation and Histone modifications in cancer - a. Alterations in genes involved in histone modification or DNA methylation have been described in various types of cancer. Oncogenes can be over-expressed when the chromatin is open, and their promoter is unmethylated. On the contrary, tumor suppressor genes (TSG) are repressed by hypermethylation of their promoter and compaction of the chromatin, leading to their transcriptional inactivation. **b.** In addition, the genome of cancer cell frequently undergoes global hypomethylation at repetitive sequences. This loss of methylation contributes to reactivation of transposable element, which might contribute to genomic instability that characterizes the tumor cells.

The *Tet2* gene is usually described as a tumor-suppressor gene (TSG). Mutations of *Tet2* were reported in AML and other myeloid neoplasms. However, the mechanism and the extent to which *Tet2* mutations affect DNA methylation remain an open question. While an increase in 5-mC is expected in the context of *Tet2* inactivation, a couple of studies have reported a decrease of the methylation (Roy et al. (2014)). In short, it seems that the *Tet2* inactivation has complex and multiple consequences on the

cell regulation. As an example, recent studies have proposed that *Tet2* could regulate gene transcription not only through DNA demethylation, but also through chromatin modification (see Roy et al. (2014) for a review).
Whatever its functions, aberrant DNA methylation profiles have been observed in many syndromes and cancer types, at various stages of disease progression. It therefore represents a promising tool for diagnosis and prognosis of patients, as well as for the development of new therapeutic targets.

1.1.2.2 Disruption of histone variants

Changes in histone modifications pattern have been reported in numerous diseases-associated studies. For instance, hypermethylation of the CpG islands in the promoter regions of tumor-suppressor genes in cancer cells is commonly associated with modifications of histone tails such as deacetylation of histones H3 and H4, loss of H3K4 trimethylation, and gain of H3K9 methylation and H3K27 trimethylation (Fraga et al. (2005)). These modifications in histone variants are usually associated with misregulation of the histone methyltransferases, demethylases, acetyltransferases (HATs) or deacetylases (HDACs) (Figure 1.5a).
One of the most prominent alteration of histone modification in cancer cells is a global loss of histone acetylation. This change can be explained by a decrease of HATs activity or by an increase of the HDACs activity. By removal of acetyl groups from histones, HDACs promotes chromatin compaction and therefore prevent the transcription of genes that encode proteins involved in tumorigenesis. Conversely, mutations or chromosomal rearrangements of HATs proteins have been reported in colon, uterus, lung and leukaemia cancer (Portela and Esteller (2010); Ropero and Esteller (2007)).
Misregulations of methyltransferases and demethylases have also been reported in various cancers. As an example, mutations in *Setd2* methyltransferase or *Utx*, *Jarid1C* histone demethylases have been reported in renal carcinomas (Staller (2010)). Mutations in the histone methyltransferase *Ezh2* were also reported in several cancer types. *Ezh2* is a subunit of the PRC2 complex, that methylates lysine 27 of histone H3 to promote transcriptional silencing. *Ezh2* was found to be over-expressed in breast cancer, bladder cancer, endometrial cancer, and melanoma (see Chase and Cross (2011) for a review). Its gain-of-function results in an increase of H3K27 trimethylation, leading to a repressed expression of Polycomb targets. In addition, *Ezh2* has been found to be

correlated with proliferation markers. However, the link between *Ezh2* and proliferation is still under investigation. Recently, Wassef et al. proposed that the high level of *Ezh2* in tumors is a consequence, and not a cause, of proliferation. In addition, low *Ezh2* expression relative to proliferation is linked to poor prognosis in breast cancer, and is associated with metastasis.

1.1.2.3 Methylation and chromatin blocks in cancer

In addition to local chromatin changes, several studies reported that the normal and cancer genomes can be partitioned into large domains harbouring similar epigenetic changes. These large genomic domains have been described as large organized chromatin lysine modifications (LOCKS), lamina-associated domains (LADs) or long-range epigenetic activation or silencing of regions (LREA/LRES) (Bert et al. (2013); Timp and Feinberg (2013)).

LOCKs and LADs refer to heterochromatin, silenced regions and largely overlap with each other. LOCKs have been described as regions enriched for heterochromatin modifications, such as histone H3 lysine 9 dimethylation, whereas LADs were described as regions associated with proteins in the nuclear lamina. Both types of regions are associated with a high level of DNA methylation in normal cells. Genomic blocks associated with other histone marks have also been reported, such as H3K9me3 and H3K27me3 in human fibroblast compared to human embryonic stem cells (Hawkins et al. (2010)). In cancer cells, the size of LOCKs is reduced, and is associated with a general disorganization of the nuclear membrane. Large regions of hypomethylation have also been identified, and globally correspond the LOCKs and LADs (Timp and Feinberg (2013)). In the same way, LRES have been described in various cancer types including colorectal, bladder, non-small cell lung cancer, breast, and prostate (Taberlay et al. (2016)). These regions can span several mega-bases and are characterized by a gain of repressive histone marks, a loss of active histone marks and an hypermethylation of CpG islands within the region (Clark (2007)). Conversely, LREA are active regions of several mega-bases, characterized by gain of active chromatin (H3K9ac) marks and loss of repressive marks (H3K27me3) in cancer. Both types of region include cancer-associated genes and have therefore been proposed as a mechanism to explain gene misregulation in cancer.

1.2 The genome organization, a new key player of epigenetic regulation ?

The way by which eukaryote genomes are compacted into their nuclei and how it impacts the cell function remains a fundamental mystery of cell biology. Eukaryote organisms are composed of billion of cells, all with nearly identical genome but distinct phenotypes and functions. It is now widely accepted that epigenetic mechanisms are responsible for these differences both at the genomic and transcriptomic level.
During the last decade, many progress have been made to understand the functional implications of DNA methylation, histone modifications, and chromatin remodelling events as well as their impact on gene regulation (see section 1.1.1.1). In the same time, it also becomes obvious that these epigenomic modifications does not suffice to fully understand all the ways in which the genome can be related to such a high variety of epigenomes and cell types. It is now clear that the different epigenomes also depend on differences in chromatin organization. Understanding how the DNA molecule is compacted in three-dimension paves the way to many fundamental questions as the role of this organization on the molecular and phenotypic portrait of the cells.
The genome structure and the chromosomal organization in the nucleus have been first studied in early nineties with the emergence of microscopy techniques. A century after, the improvement of microscopy techniques and the emergence of high-throughput genomics allow to deeply explore the genome organization and its impact on the cell function. However, many aspects are not yet fully understood. Exploring the genome organization therefore remains an active and fascinating field of research.

1.2.1 Principle of genome architecture

1.2.1.1 Basic principles of genome organization uncovered by microscopy

The chromosomes and the nuclear structure have been commonly explored through DNA imaging technologies, based on electron or light microscopy. Microscopy techniques have revealed that the three-dimensional location of chromatin is not random. The history starts at the late 19th century (see Cremer and Cremer (2006b) and Cremer and Cremer (2006a) for reviews). In 1882, Wather Flemming introduced the terms 'chromatin' and 'mitosis'. For the first time, he described in detail the nuclear division, and the fact that these chromatin coils are able to segregate into the newly forming

daughter nuclei. In 1888, these chromatin threads were named 'chromosome' by Wilhelm Waldeyer. In 1885, Carl Rabl suggested a territorial organization of interphase chromosomes. Twenty years later, in 1909, Theodor Boveri suggested the notion of 'chromosome territory' arguing that each chromosome occupies a distinct space in the nucleus during interphase. During the 1950s to 1980s, different models of chromosomes organization were proposed and debated (see Cremer and Cremer (2006a) for a review). Finally, the final proof of the chromosome territory concept was proposed by Stephen Stack in 1978.

Nowadays, one of the most popular experimental approach to investigate chromosomal structure and positioning in eukaryotes cell is the DNA fluorescent in situ hybridization (FISH, Volpi and Bridger (2008)). FISH, combined with fluorescence microscopy, largely contributes to enrich our knowledge in three-dimensional genome organization. FISH can directly measure the physical distances between two DNA loci and visualize their position within the nucleus. FISH uses fluorophore-labelled DNA (or RNA probes for RNA-FISH), which are designed to target a specific genomic loci (or transcript). Depending on the genomic loci to look at, both maternal and paternal loci can be visualized. While FISH techniques continue to be improved in parallel to the emergence of new microscopy techniques, they remain limited to the study of a small number of genetic loci in parallel and do not allow a comprehensive analysis of nuclear architecture of the complete genome. However, future developments in this field, such as high-throughput imaging approach (HIPMap, Shachar et al. (2015)) and super-resolution microscopy will probably allow to overcome these limitations (Wang et al. (2016)).

1.2.2 Different level of genomic organization

The genome architecture is organized as a hierarchical model. At larger scale, chromosomes are organized within the nucleus, occupying their own space. At finer scales, the chromosome itself is organized into compartments of active/close chromatin, which can further be divided into topological domains, up to a single DNA loop resolution (see Bonev and Cavalli (2016); Gibcus and Dekker (2013); Ramani et al. (2016) for reviews, Figure 1.6).

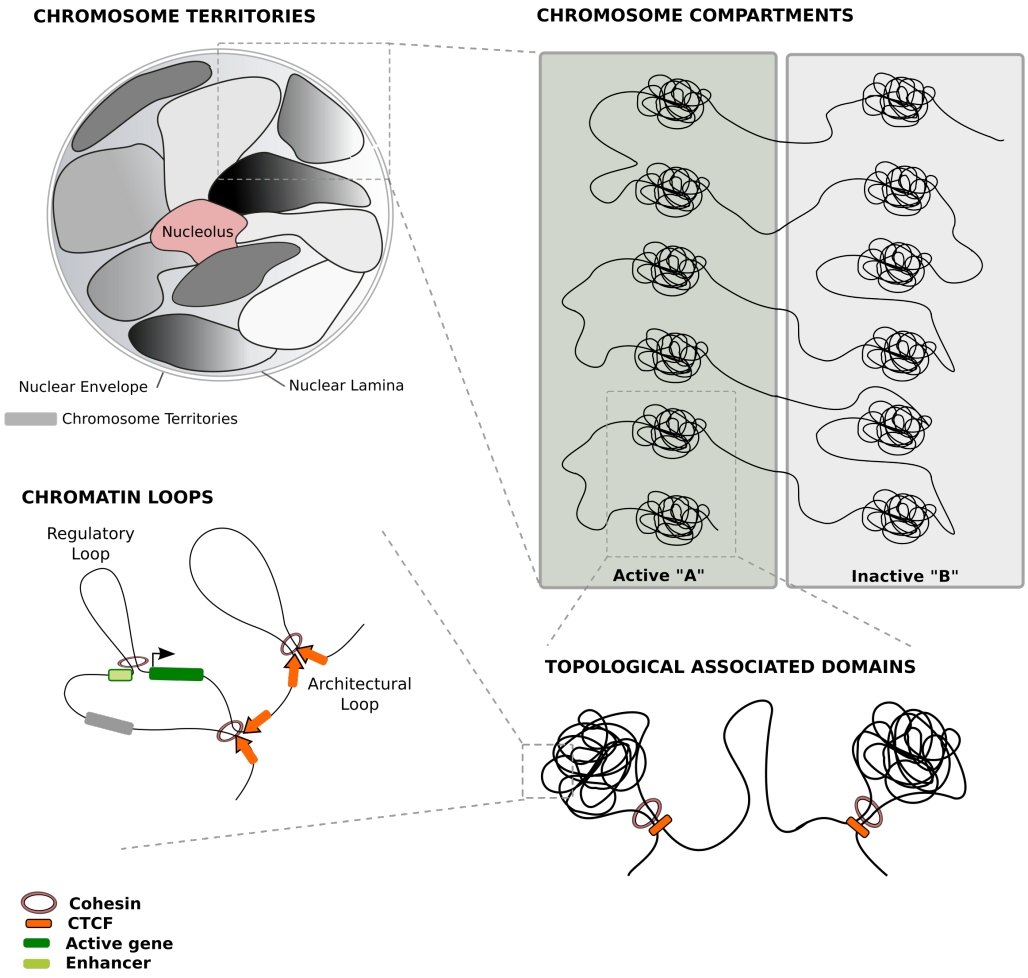


Figure 1.6: Genome organization - The genome is organized as a hierarchical model. Homologous chromosomes contain their own territory in the nucleus where small, gene-rich chromosomes tend to co-localize. Then, the chromosome itself can be divided in compartments of active (A) and inactive (B) chromatin. Regions with similar epigenetic signatures are characterized by stronger inter-domain interactions. The chromatin is further organized in topological associated domains (TADs). CTCF proteins are shown as orange rectangles and loop-extrusion complexes such as cohesin are depicted as red circles. Finally, at higher resolution, different types of chromatin loops can potentially occur within a TAD. Architectural loops are formed between a forward and reverse CTCF sites. Regulatory loops bring closer the genes and their regulatory elements such as enhancer in order to promote the transcription. Adapted from Bonev and Cavalli (2016).

1.2.2.1 Chromosome territories

In mammalian genomes, each chromosome occupies its own nuclear space, called chromosome territories (CTs, Cremer and Cremer (2010)). The fact that each chromosome invades a distinct nuclear space raises the question of the stochasticity of this organization. Are CTs randomly organized, or is there any pattern in this organization, associated by common molecular functions or genomic features ?

Recent advances in the field clearly demonstrate that the chromosomes are arranged in a non-random way. One assumption is that the arrangement of chromosomes in the nucleus may facilitate cellular functions to occur in an efficient manner. This is for example, the case of the inactive X chromosome that is able to directly interact with the nuclear lamina (Chen et al. (2016)).

Recent developments in molecular biology techniques demonstrated that intra-chromosomal contact frequencies are much more frequent than inter-chromosomal contacts (Lieberman-Aiden et al. (2009)), therefore demonstrating that CTs are discrete structures within the nucleus (Figure 1.6). However, and even if this is much less frequent, it is also clear that neighbouring chromosomes can overlap with each other and that chromatin loops from one territory can invade the body of the neighbouring territory (Branco and Pombo (2006)). One typical example of contacts occurring between CTs is the interaction patterns of centromeres and telomeres observed in different organisms such as Yeast (Varoquaux et al. (2015)) or Drosophila (Hou et al. (2012)).

The position of chromosomes within the nucleus seems to follow a subtle length to gene-density ratio. Small and gene-rich chromosomes tend to cluster together nearby the center of the nucleus, whereas large and gene-poor chromosomes are usually located at the nuclear periphery (Boyle et al. (2001)). Interestingly, the chromosome 18, which is small but gene-poor does not interact frequently with the other small chromosomes (Croft et al. (1999); Lieberman-Aiden et al. (2009)). This supports the idea that chromosomes organization is somehow influenced by various nuclear processes such as transcription or DNA replication/repair. As an example, the nucleoli were described as a specialized sub-nuclear structure, formed from several chromosomes, and involved in expression by RNA polymerase I or III (Nmeth et al. (2010)).

Interestingly, the spatial configuration of CTs relative to each other has been described to be tissue-specific (Parada et al. (2004)). Relocating specific chromosomes to nu-

clear periphery can have critical consequences on gene expression, therefore suggesting a possible link between CTs and diseases (Finlan et al. (2008)). In the same way, a number of studies have investigated the proximity of chromosomes and their implication in common translocations in cancers (Branco and Pombo (2006); Meaburn et al. (2007)). These results support the hypothesis that translocations occur in interphase nuclei between chromosomes that occupy close nuclear space.

1.2.2.2 Chromosome compartments

In the nucleus and within CTs, euchromatin tends to be spatially separated from heterochromatin. Gene-rich and open chromatin regions are separated from gene-poor and closed regions, therefore suggesting a functional compartmentalization of chromosomes (Fraser and Bickmore (2007); Naumova and Dekker (2010)).

In order to study the genome architecture in a more systematic way and at high scale, high-throughput 3C-based methods were proposed (see section 1.3). In 2009, Lieberman-Aiden et al. published the first Hi-C experiment, allowing to explore genome-wide contact frequencies. Based on this first analysis, two types of chromosome compartments, called A and B, were identified at the mega-base scale on the basis of preferential interaction with each other (Figure 1.6). Regions within each compartment mostly interact with regions from the same compartment type. Further analysis of these two classes of compartments demonstrated that A compartments are enriched in active and open chromatin, whereas B compartments are compacted and repressed (Kalhor et al. (2011); Lieberman-Aiden et al. (2009)).

Interestingly, the three-dimensional genome organization appears to be closely associated with the one-dimensional chromatin states (Mourad and Cuvier (2015)). Recently, Fortin and Hansen proposed to use genome-wide methylation, DNA hypersensibility sequencing, and assay for transposase-accessible chromatin (ATAC) sequencing data as a predictor of A/B compartments. DNAse, ATAC sequencing as well as methylation signal are highly correlated with open chromatin and A compartments regions. In addition, Hi-C data at very high resolution suggested that these two major compartments can be further divided into at least five different sub-compartments with specific patterns of histone modifications (Rao et al. (2014)). At a resolution of 25kb, compartments A can be partitioned into 2 sub-compartments (A1, A2), which can be differentiated by replication timing, GC content, gene length or H3K9me3 enrichment.

In the same way, compartment B can be further partitioned into 3 sub-compartments (B1, B2, B3) with different replication timing, H3K27me3, and nucleolus-associated domains (NADs) enrichments.
Finally, chromosome compartments were described as dynamic structures. Switches between compartments A and B have been reported during cell differentiation, therefore demonstrating a high degree of plasticity in the establishment of the A/B compartments across cell types (Dixon et al. (2015)). These changes in structure are usually associated with modifications in gene expression, and have also been reported in cancer cell lines (Barutcu et al. (2015)).

1.2.2.3 Topological associated domains

The characterization of topological associated domains (TADs) is one of the most exiting discovery of these last years. TADs are usually described as a functional unit of chromosome organization (Dixon et al. (2012); Nora et al. (2012)). Within a TAD, genomic loci are able to interact with each other, whereas contacts with adjacent domains are much less frequent.
TADs have been identified in several species including mammals, drosophila, *C. elegans*, *S. pombe* or *C. crescentus* (see Dekker and Heard (2015) for a review). Looking at a Hi-C contact map, TADs are represented by square domains along the diagonal of the count matrix (Figure 1.7). These squares represent the topological domains where genes and regulatory elements can interact together, and are separated by boundaries with specific insulating properties, that structurally distinguish two adjacent TADs. In mammals, TADs were first described with an average size around the mega-base scale (Dixon et al. (2012)). However, their exact size is not clearly defined. Depending on the data resolution, different studies reported TADs with a median size between 900kb to 185kb. It should be noted that TADs are usually called from high-throughput 3C based experiments, which provide a view of chromosomal architecture averaged among a cell population (Figure 1.7). Data resolution, as well as computational algorithms can explain a part of variability in their definition (Filippova et al. (2014)). Nevertheless, recent studies suggested that TADs can be defined as hierarchical structures, which can be further partitioned into smaller sub-mega-base domains, often called sub-TADs (Fraser et al. (2015); Phillips-Cremins et al. (2013)).

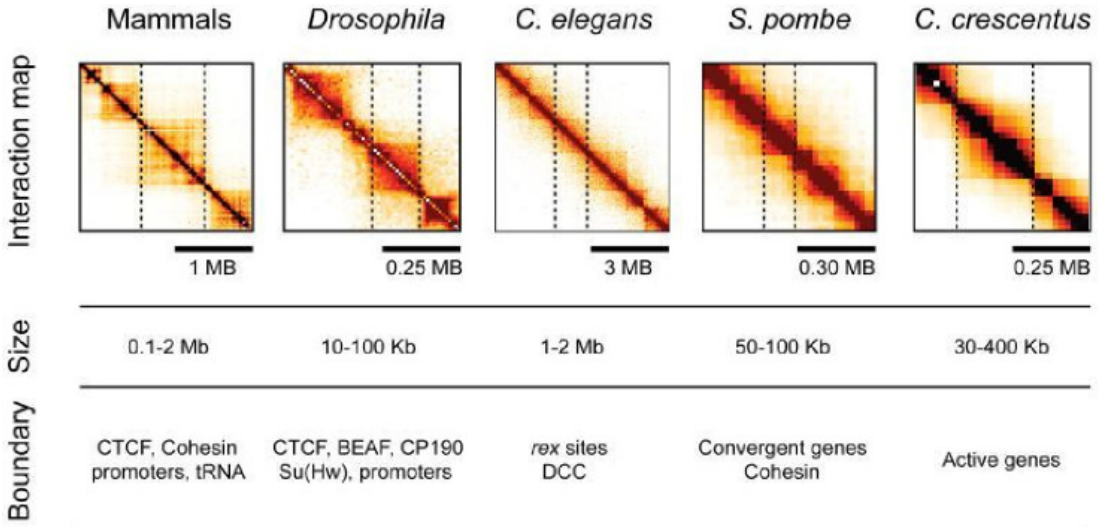


Figure 1.7: Topological associated domains (TADs) in different organisms - Contact maps represented TADs in multiple organisms. Adapted from Dekker and Heard (2015).

Although the exact mechanisms at the origin of TADs and sub-TADs structures are not fully understood, it appears that TADs boundaries are often enriched for insulator proteins such as CCCTC-binding factor (CTCF) and cohesin (see section 1.2.3). In addition, the TADs boundaries correlate with many genomic features such as histone modifications, gene expression or replication timing suggesting a link between the establishment of TADs and the regulation of neighbouring elements (Dixon et al. (2012); Nora et al. (2012); Pope et al. (2014)). As for chromosome compartments, models have been proposed to predict TADs boundaries using histone marks, chromosome accessibility or CTCF binding sites data (Huang et al. (2015); Zhu et al. (2016)). Interestingly, deletion of genes involved in histone modifications, does not impact the TADs structure (Nora et al. (2012)), therefore meaning that TADs are not the results of histone modifications, but that, on the contrary, histone modifications occurs on pre-existing TADs.

TADs are usually described as functional units of genome organization. Interactions between regulatory elements such as enhancers and promoters, mainly occur within the same TAD. Enhancers represent regulatory DNA elements of a few hundred base pairs, that positively control the transcriptional activity of genes located a few kilobases away. Enhancers are able to attract specific transcription factors and to contact

upstream or downstream promoter region of target genes. This mechanism creates a loop in chromatin structure that bring enhancers and its target gene in close spatial proximity. While transcription doesn't seem to be strictly required for TAD formation, genes located within the same TAD tend to have coordinated expression during cell differentiation (Nora et al. (2012)).

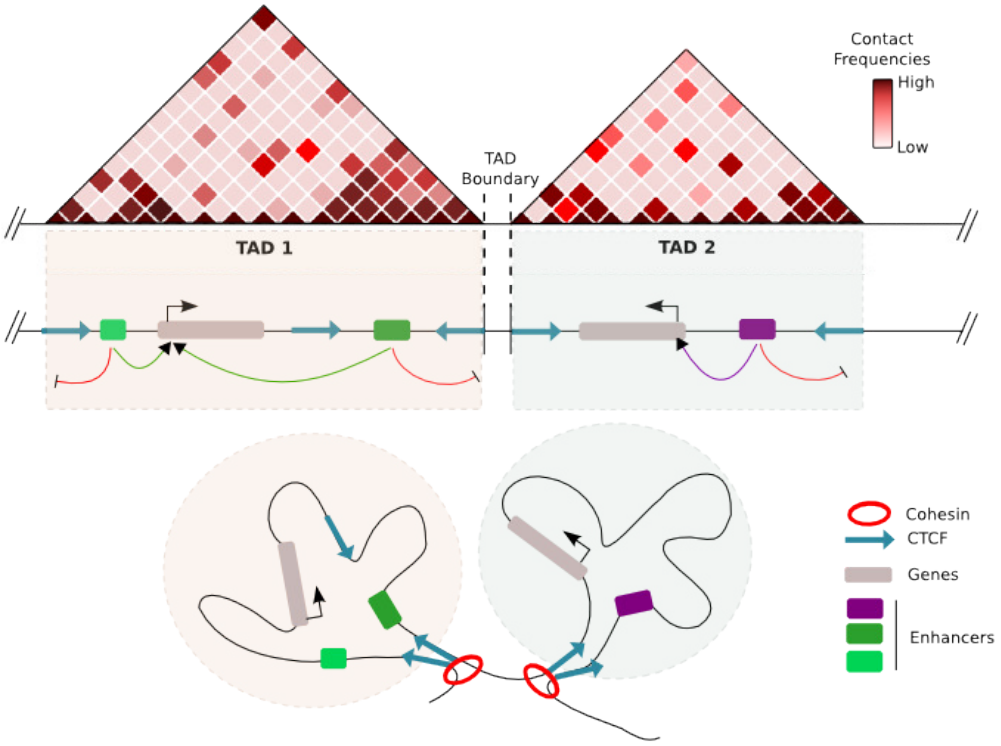


Figure 1.8: Topological associated domains (TADs) as functional units - TADs are frequently described as functional unit of organization. Using 5C or Hi-C data, TADs can be seen as triangular blocks along the diagonal, which restrict the influence of regulatory elements to genes within the same TAD. The TAD boundaries have been observed to be associated with convergent CTCF and cohesin binding sites.

In the same way, reporter genes integrated hundreds of kilo-bases apart but in the same TAD, may show the same expression patterns, whereas this correlation rapidly decreases when they are located beyond TAD boundaries (Symmons et al. (2014)). Thus, TADs are thought to play a role in promoting interactions in three-dimensional space between regulatory elements belonging to the same domain, therefore decreasing

the probability that non-specific or damaging interactions occur.

Moreover, recent studies have demonstrated a strong correlation between replication domains and chromatin structure during interphase. Early and late replication domains have been first described to be associated with open/close chromatin compartments (Dixon et al. (2012); Ryba et al. (2010)). Finer compartments classification was proposed to subdivide active (A) compartments into A1 compartments, mostly associated with early constitutive replication domains, and A2 compartments characterized with early cell type-specific replication domains. In the same way, close compartments (B) can be sub-divided into B1 compartments, associated with early/late cell type-specific replication domains, whereas B2 and B3 sub-classes are mainly enriched in late constitutive and cell type-specific replication domains (Dileep et al. (2015); Rao et al. (2014)). In addition, almost all TADs boundaries can be associated with replication domains boundaries demonstrating a role in TADs as units of replication timing (Pope et al. (2014). Inter-TAD interactions are also more frequent between regions of similar replication timing (Yaffe et al. (2010)). Interestingly, not all replication domains boundaries match a TADs boundary, thus raising the question of the definition and resolution of TADs calling. Finally, the current model is that TADs could act as a functional replication unit. When a TAD is replicated early, its boundaries will be associated with timing transition regions and the TAD will switch to an active compartment state, increasing its contact frequency with other early-replicating TADs. The replication is then passively extended to adjacent later-replicating TADs (Pope et al. (2014)). Interestingly, Dileep et al. recently explored the temporal establishment of TADs with respect to replication timing. Their results suggest that TADs are established during early G1 phase when the replication-timing program is re-established, consistently with previous reports on the chromatin organization of mitotic chromosomes (Naumova et al. (2013)). Interestingly, the TADs structure is conserved during G2 phase while determinants of replication timing are absent. Altogether, these results highlight the role of chromatin organization and TADs in particular, into the establishment of the replication-timing program.

Surprisingly, the TADs structure appears to be highly conserved among different cell types and organisms, with around 50 to 70% of common boundaries (Dixon et al. (2012); Rao et al. (2014)). This observation is surprising given the previous evidence of cell specific chromosome compartments organization. It is therefore tempting to

speculate that the mechanisms leading to TADs and chromatin compartments organization are not the same. One model would be that chromatin compartments are highly dependant on chromatin modifications, whereas TADs and their boundaries rather rely on insulators and architectural proteins such as CTCF and cohesin. However, at higher resolution, sub-TADs seem to be much more dynamic. Recently, Dixon et al. described changes at the sub-TADs level during cell differentiation and observed that intra-TAD contacts increased when a compartment switches from an inactive (B) to an active (A) state, and decrease when a compartment is repressed. This observation means that although the TADs boundaries are largely conserved, cell type specific loops between promoters and regulatory elements can occur within invariant TADs.

Although significant advances have been recently made in our understanding of TADs structure, their functions are not yet fully explored. As previously discussed, our current definition of TADs mainly depends on data resolution and computational algorithms used to detect them (see section 1.4.4.3 for details). In addition, most of the recent studies were done on averaged cell population, therefore ignoring the cell-to-cell variability of genome architecture. Although single cell approaches have been proposed, they currently do not have the resolution to robustly measure chromatin folding at the sub-megabase resolution (Nagano et al. (2013)).

Polymer physics and computational models were therefore proposed to help in validating existing assumptions of chromosomal organization (see Imakaev et al. (2015) for a review). Recently, Fudenberg et al. proposed a loop extrusion model as a mechanism for TAD formation. This model implies loop-extruding factors such as cohesin to form larger loops, stalled by boundary elements such as CTCF insulators.

1.2.2.4 Chromatin loops

Chromatin loops represent the finer scale of genome organization. Recent high resolution Hi-C data demonstrated that architectural loops are formed between TAD boundaries. Although the exact mechanisms underlying the looping of TAD boundaries are largely unknown, several studies have reported the role of CTCF and cohesin complexes at the sites that anchor these loops (Dixon et al. (2012); Rao et al. (2014)). These results suggest that around 38% of TADs are enclosed by a chromatin loop at their boundaries. In addition, TADs without loops are usually flanked by loop domains

and may therefore be a consequence of constrained architecture of neighbouring domains (Rao et al. (2014)).

In contrast to these architectural loops, regulatory loops can be formed between enhancers and promoters within a single TAD (Figure 1.6). So far, two types of enhancer-promoter loops have been described ; the pre-existing loops and the loops formed de novo (see Bouwman and de Laat (2015) for a review). Pre-existing loops allows proximity in three-dimensional spaces of genes and their regulatory elements independently from their expression status. This mechanism is believed to facilitate the cell response to developmental stimuli (Jin et al. (2013)). By contrast, de novo loops were observed at cell type specific enhancers, and are highly variable, suggesting the existence of specific chromatin interaction structures between cell types (Jin et al. (2013); Rao et al. (2014)). In practice, it seems that pre-existing loops cover interactions which are located up to 1 Mb away from each other, whereas de novo loops mainly cover short genomic distances. Finally, tissue-specific enhancer-promoter loops have been described and depend on the association of specific transcription factors with transcriptional co-activator and insulator proteins (Kagey et al. (2010)).

1.2.3 The role of CTCF/cohesin complex

In the past decade, many studies reported the crucial role of CTCF in chromatin architecture. CTCF is a conserved zinc finger nucleic acid binding protein initially characterized as a transcription factor able to modulate gene expression. Based on recent advance on genome organization research, CTCF was later characterized as the main insulator protein in vertebrates which has the ability to block enhancer-promoter interactions (see Ong and Corces (2014) for a review). The TADs boundaries and anchors of chromatin loops are frequently associated with the binding of CTCF. This observation highlights the fundamental role of CTCF in chromatin architecture, although many CTCF sites do not seem to be related to looping or chromatin folding.

One of the main CTCF interactor is the cohesin complex. Cohesin is a protein complex composed of four core subunits ; SMC1A (structural maintenance of chromosomes protein 1A), SMC3 (structural maintenance of chromosomes protein 3), RAD21 (double-strand-break repair protein rad21 homologue), and either STAG1 or STAG2 (cohesin subunit SA1/2). Cohesin plays important roles in establishing and regulating chromatin organization and usually overlap with CTCF binding sites (Wendt et al. (2008)),

demonstrating a functional link between them. Cohesin has been shown to be required to stabilize most CTCF-mediated chromosomal contacts. Indeed, depletion of cohesin abolished loops between CTCF sites, and decreases intra-TADs contacts (see Merkenschlager and Nora (2016) for a review).

The CTCF/cohesin complex is a key player of the genome architecture. Cohesin depletion abrogates looping between CTCF sites and reduces interactions within TADs, but does not impact chromosome compartments (Seitan et al. (2013)). The CTCF protein is able to bind the DNA and to recognize a non-palindromic motif typically written as 5'-CCACNAGGTGGCAG-3'. Interestingly, CTCF sites involved in chromatin looping are almost always in a convergent linear orientation (Rao et al. (2014)). Depletion or inversion of CTCF site at the TADs boundaries can lead to a disruption of local compartmentalization and chromatin loops (see section 1.2.5).
In the context of X chromosome inactivation, the deletion of a TAD boundary in the X inactivation center led to a fusion of adjacent TADs, impacting the expression of neighboring genes (Nora et al. (2012)). In the same way, a deletion of CTCF sites within the Hox clusters results in the spreading of active chromatin marks into a usually repressed domain (Narendra et al. (2015)). Recently, Nora et al. proposed an auxin inducible degron system allowing an acute and reversible depletion of endogenous CTCF in mouse ES cells. Depletion of CTCF triggers a dramatic loss of TAD insulation and formation of chromatin loops, therefore demonstrating that CTCF is absolutely essential for TADs organization. As already mentioned, the CTCF deletion does not affect higher-order genomic compartmentalization, suggesting independent mechanisms between chromosome compartments and TADs insulation (Nora et al. (2016)).

1.2.4 Chromosomal organization is dynamic

The chromosome organization is a dynamic process, which fluctuates between cells as well as in time and space. Single-cell Hi-C provided a first glimpse of cell-to-cell variability in chromosome organization, showing that the chromatin domains at the megabase-scale seem to be conserved between cells, whereas variable cell-to-cell structures are observed at larger scales (Nagano et al. (2013)). Although the current techniques suffer from a lack of resolution to explore the chromatin organization at the

sub-mega-base scale, several studies reported that cells seem to maintain their organization in domains at the mega-base scale, but show variable cell-to-cell organization at larger scales (Goetze et al. (2007); Nagano et al. (2013)). Therefore, the position of TADs boundaries appears to be consistent between cells. These observations have been further confirmed in a recent study reporting an integrative analysis of chromatin contact maps in 21 primary human tissues and cell types (Schmitt et al. (2016)). In addition, high-resolution imaging recently indicated that at higher resolution, loop domains appear to be variable from one cell to another (Giorgetti et al. (2014)). This is in agreement with FISH experiments showing that genes adopt different locations in different cells, and that the distances between pairs of loci might also change between cells. This cell-to-cell variability has also been explored using polymer physics and modelling, indicating that chromatin loops might not be stable structures but more probabilistic events (Fudenberg and Mirny (2012); Giorgetti et al. (2014)). In this case, the TADs structure that appears using Hi-C data could emerge from an ensemble of conformation that comes from many possible structures across the cell population. Interestingly, this variability in chromatin structure could therefore contribute to cell-to-cell transcriptional variability, indicating that contacts between enhancer and promoter would also arise in a probabilistic way (Krijger and de Laat (2013)).

1.2.4.1 Example of the X chromosome inactivation

One particular example of changes in chromatin organization occurs during the X chromosome inactivation (XCI) in mammals. In 1961, Mary Lyon first made the hypothesis that one X chromosome is inactivated in each female cell (LYON (1961)). More than 50 years later, the precise mechanism leading to XCI remains elusive and many related questions are still being explored.

X inactivation occurs during the development of early embryo and ensures dosage compensation between female (XX) and male (XY). Two different forms of XCI have been described ; imprinted and random XCI. In mouse, imprinted inactivation of the paternal X occurs in early female embryo just before implantation (Mak et al. (2004)). This paternal X silencing is maintained in extra-embryonic tissues (placenta) but is reverted in the inner cell mass at the origin of the embryo. Then, at the blastocyst stage, random XCI occurs leading to the inactivation of either the paternal or the maternal X chromosome. The inactive chromosome is then stably maintained and transmitted

through cell divisions to future generation of cells.

Many molecular actors have been described as key players of XCI over recent years (see Augui et al. (2011) for a review). Among them, the master regulation locus, also known as X inactivation center (Xic), is composed of the *Xist* non-coding RNA and its regulators. The *Xist* gene is expressed from the future inactive X and is essential for both imprinted and random X-inactivation. Its long non-coding RNA spreads along the inactive X chromosome (Xi) leading to an almost complete transcriptional silencing associated with chromatin modifications and spatial reorganization of the chromosome (Figure 1.9a). How *Xist* RNA is able to spread along the Xi is still not fully understood. One hypothesis would be that LINE-1, which are over-represented on the X chromosome may serve as 'anchors' elements, thus allowing the RNA to coat the chromosome (Lyon (2003)). Recently, another model was proposed where *Xist* first targets gene-rich regions before spreading to gene-poor inter-regions (Simon et al. (2013)). In addition, *Xist* was reported to be associated with distal regions that are closed to its locus in three-dimensional space (Engreitz et al. (2013)).

Among the different elements that defined the XIC, The *Xist* anti-sense gene (*Tsix*) is now known to be involved in *Xist* regulation (see Maclary et al. (2013) for a review). *Tsix* is first expressed in females from both X chromosomes prior to X inactivation. Then, during X inactivation, *Tsix* is down regulated from the Xi and is expressed from the active X chromosome (Xa). *Xist* and *Tsix* therefore show mutually exclusive expression from the Xi and Xa chromosomes respectively. In addition, *Tsix* seems to play a role during the random XCI by controlling the choice of which X-chromosome will be inactivated (Lee and Lu (1999)).

Xist is thus able to physically coat the Xi chromosome and to modify its chromatin structure. *Xist* is thought to function by recruiting chromatin modifiers that will help in establishing the heterochromatic state of the Xi chromosome. Among its different partners, *Xist* is known to recruit the Polycomp group proteins.

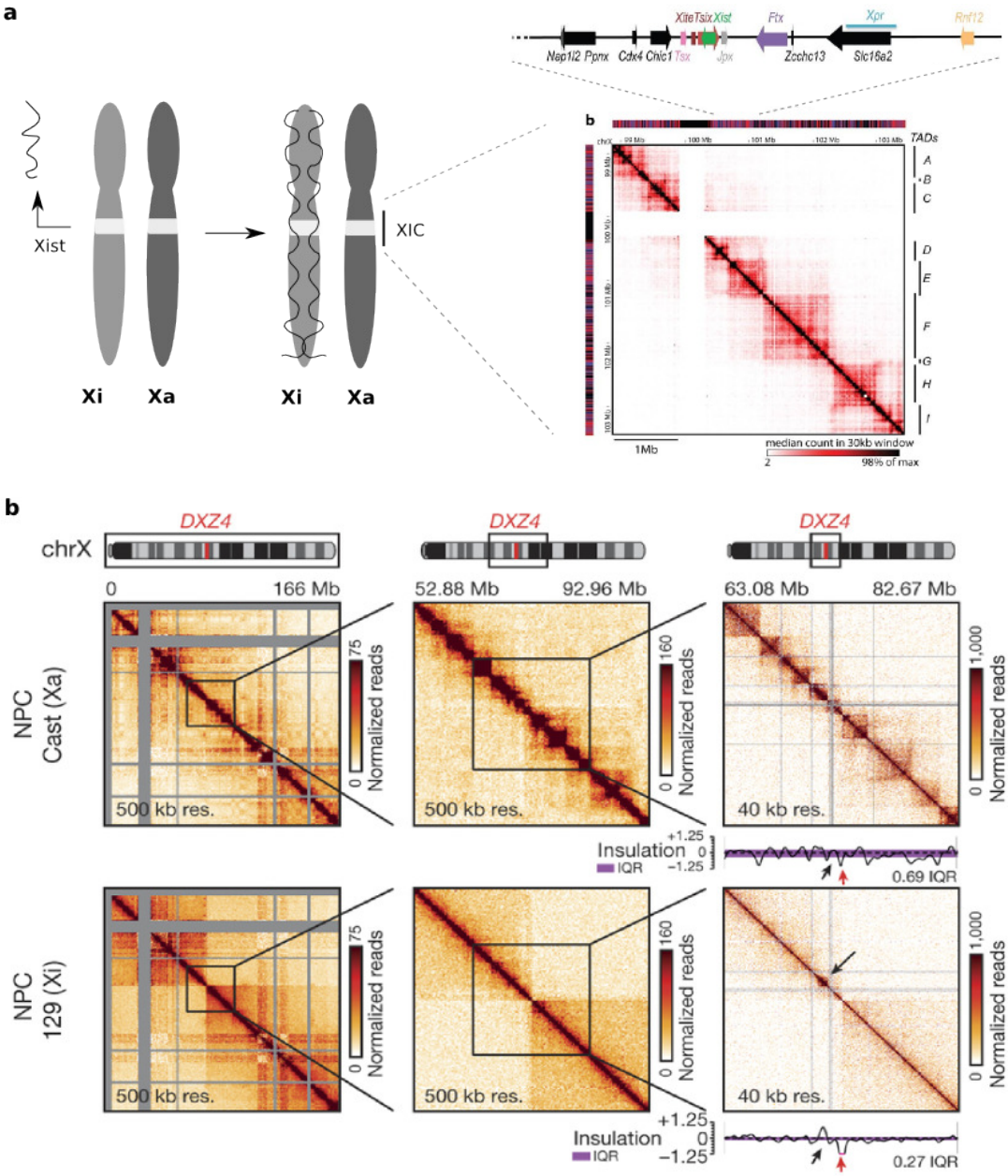


Figure 1.9: Genome organization of X chromosome in mouse - a. In female, one X chromosome is inactivated (Xi) by the coating of the *Xist* RNA, expressed from the X inactivation center (XIC). In 2012, Nora et al. described the three-dimensional structure of the active X chromosome (Xa) in mouse ES cell using 5C experiments. The XIC covers a region of a few mega-bases, with distinct TADs structures. **b.** At the chromosome level, the use of Hi-C technique allows to characterize the structure of Xi that is organized in two mega-domains, with almost no TAD structures, and divided at the *DXZ4* locus. Unlike the Xi, the Xa is characterized by a chromatin structure with chromosome compartments and TADs all along the chromosome. Adapted From Giorgetti et al. (2016); Nora et al. (2012)

The Polycomp group proteins are composed of two complexes, respectively called polycomb repressive complex 1 (PRC1) and polycomb repressive complex 2 (PRC2) (see Margueron and Reinberg (2011); Schwartz and Pirrotta (2013) for reviews). The PRC1 complex is involved in chromatin compaction through mono-ubiquitylation of histone H2A, such as ubiquitination of lysine 119 in histone H2A (H2AK119ub). The PRC2 complex also contributes to chromatin accessibility, through methylation of histone H3 at lysine 27 (H3K27me3). Both histone modifications are known to be enriched on the Xi. Conversely, active histone marks such as acetylation of histone H3 and H4 are depleted from the inactivated chromosome.

Interestingly, all genes are not affected by X inactivation. Some of them, called escapees, are able to maintain a bi-allelic expression from the Xa and Xi chromosomes (Berletch et al. (2010)). About 3% and 15% of genes are able to escape X inactivation in mouse and human respectively. The identity and the number of escapees vary from a species to another. In addition, some, but not all, escaping genes are known to have a copy on the Y chromosome, as for example *Kdm5C* or *Kdm5d*. These genes therefore have two expressed alleles in both male and female cells.

In addition, while most genes have a stable inactivation pattern, a subset of genes, called facultative escapees, escape XCI in a tissue or context specific manner (see Berletch et al. (2011) for a review). In contrast with the general compacted chromatin state observed on Xi chromosome, regions that escape XCI are euchromatic and usually characterized by an active state and H3/H4 acetylation. Histone mark associated with gene transcription such as H3K4me3 are also usually enriched at escapees locus (see Heard and Disteche (2006) for a review). In practice, the chromatin structure may have an important role in excluding the escaping regions out of the compacted inactivated X structure. This also suggests a potential role for insulator proteins, such as CTCF, that could block the spreading of heterochromatin or prevent the repressive marks from being added in escaping regions (Filippova et al. (2005)).

Recently, significant progress has been made in understanding the chromatin structure of X chromosome. In 2012, the presence of TADs, defined as mega-base scale functional domains, has been first reported on the Xic locus (Nora et al. (2012), Figure 1.9a). Using Carbon copy 3C (5C, see section 1.3) experiments over a 4.5 Mb region centred on the Xic, Nora et al. demonstrated that the promoters of *Xist* and *Tsix* fall into two distinct TADs. This organization promotes the interactions of *Xist* with its positive

regulators in one TAD, while segregating it from the *Tsix* promoter and its regulators. Within these two TADs, known interactions have been confirmed and new long-range interactions were identified, such as interaction between *Tsix* and the long non-coding RNA *Linx*. In addition, genes within the same TAD show coordinated expression, with an opposite regulation of *Xist* and *Tsix* TADs during early differentiation in mouse (Nora et al. (2012)).

Improvement in resolution of high-throughput chromosome conformation capture (Hi-C, see section 1.3) recently allows to explore the chromatin structure of the entire Xa and Xi chromosomes in mouse and human (Deng et al. (2015); Giorgetti et al. (2016); Rao et al. (2014)). The Xa chromosome is characterized by a structure similar to autosomes chromatin organization, with prominent open and closed (A/B) chromosome compartments and TAD organization at the mega-base scale. On the contrary, the Xi chromosome is condensed and is usually located within the nucleus near the nuclear membrane or the nucleolus (Zhang et al. (2007)). TADs are globally absent from the Xi chromosome. However, genes that escape from X inactivation conserve an open chromatin and active transcription states, and therefore harbour a TAD-like structure (Giorgetti et al. (2016), Figure 1.9b).

The Xi chromosome presents a global loss of local structure and is divided in two large mega-domains where loci in the same mega-domain tend to co-localize. These mega-domains are separated by a 200kb region including the *DXZ4* macrosetellite in both mouse and human. Although the genomic context is not the same between species, this region is partially conserved between mouse and human (Deng et al. (2015)), and more globally between eutherian mammals (Darrow et al. (2016)). The role of *DXZ4* in the segregation of these two mega-domains is not clearly understood. So far, it has been shown that this region is able to bind CTCF and is associated with the nucleolus exclusively on the Xi homologue (Deng et al. (2015)). In addition, the *DXZ4* locus seems to be involved in the establishment of super loops spanning dozens of mega-bases of the genome, establishing contacts between *DXZ4* and several loci including the *Xist* and *Firre* loci (Darrow et al. (2016)). Depletion of this region leads to the fusion of the two mega-domains but does not affect XCI establishment (Giorgetti et al. (2016)).

To conclude, the X chromosomes present a specific chromatin organization, that changes during cell differentiation. Before the establishment of XCI, both X chromosomes are characterized by open/close chromosome compartments and TADs organization. After

XCI, the Xi harbours a specific condensed structure characterized by the presence of two mega-domains and a global loss of local structure.

1.2.4.2 Changes during cell differentiation

Understanding the impact of chromatin organization during cell differentiation is now up for debate. How dynamic changes occur during cell differentiation or cell development ? What is the impact of the genome structure on the cell identity ? How changes in chromatin organization correlates with changes in gene expression observed during differentiation ? In order to answer these questions, recent studies investigated the genetic and epigenetic mechanisms that are responsible for the control of cell differentiation. In this context, embryonic stem (ES) cells are usually a good system model. ES cells are pluripotent and can therefore differentiate into many other cell types. ES cells are characterized by a global accessible chromatin landscape, pointing to a model in which the open chromatin structure maximizes the cell plasticity, and is therefore essential in establishing pluripotency (Meshorer et al. (2006)).

Recently, Dixon et al. explored changes in chromatin structure between human ES cells and its derived lineages; mesendoderm, mesenchymal stem cells, neural progenitor cells and trophoblast-like cells. Their study reports a significant plasticity in the establishment of A/B compartments. Therefore, around 36% of chromatin compartments were able to switch between active and inactive states (or vice versa) in at least one of the lineages analysed. Although these switches in chromosome compartments seem correlated with gene expression, most genes stay unaffected and only a subset of genes modulates their expression in accordance with compartment switching. This might be explained by the fact that we are looking at population-based experiments. It is therefore tempting to speculate that the chromosome comportments pattern which is observed reflects a stochastic tendency of compartments to interact, rather than a set of deterministic contacts between specific loci.

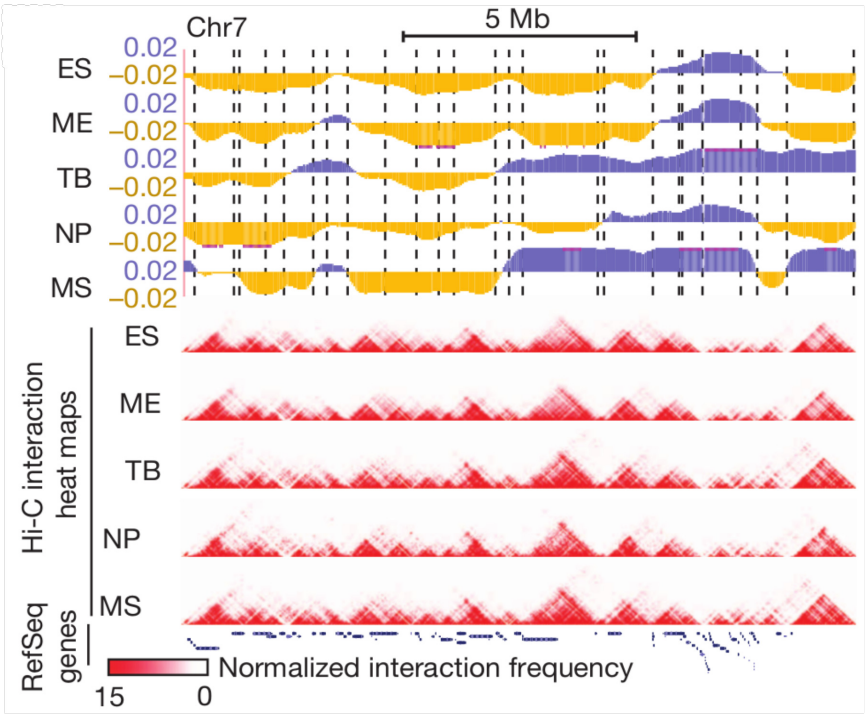


Figure 1.10: Dynamic of chromatin structure during differentiation of human ES cells (From Dixon et al. (2015)) - Chromosome compartments are dynamic during cell differentiation, switching from active (A, blue) compartments, to inactive (B, yellow) compartments. In addition, although changes within TADs have been described during differentiation, their positioning remains stable between cell types.

In agreement with previous studies, TADs boundaries are highly conserved between cell types and do not change during cell differentiation. However, a subset of TADs undergo within-domain changes in interaction frequency during differentiation. These changes correlated positively with active chromatin marks (H3K27ac), the density of enhancer elements (H3K4me1) and gene expression. In the same way, they correlate negatively with repressive chromatin marks (H3K27me3, H3K9me3).

At the level of autosomal chromosomes, both copies of homologous chromosomes have highly comparble A/B compartment patterns. As expected, regions involved in imprinted expression present a significant increase in the variability of A/B compartments. As previously discussed, the main changes between homologous chromosomes during ES cells differentiation occur at the X chromosome level in females (see section 1.2.4.1). On differentiated cells, one the two X chromosomes is inactivated. The chromosome

structure of the Xi is therefore completely disrupted and is characterized by a loss of local structures and the presence of mega-domains splitting the chromosome in two highly condensed blocks (Giorgetti et al. (2016)).

1.2.4.3 Changes during the mammalian cell cycle

Recent advances in high-throughput and 'C'-based technologies have allowed to rapidly accumulate new insights about chromosome organization in different cell types and contexts. Among the different unanswered questions, understanding what are the mechanisms that determine the dynamic of chromosome compartments or TADs remains challenging. In the same context, and until recently, the chromatin structure and the organization of mitotic chromosomes have remained largely elusive. In order to answer this question, one recent study deeply explored the dynamic of chromosome architecture during the cell cycle (Naumova et al. (2013)). We reviewed and summarized this work in Giorgetti, Servant, Heard (2013) (see Annexe 1.6.1).

Chromatin organization is generally studied in non-synchronous cells, of which interphase cells represent the biggest proportion. In this stage, chromosomes are organized in A/B compartments and TADs, as described above. Naumova et al. used 5C and Hi-C technologies to explore the chromosome conformation during early G1-, mid-G1-, S- and M-phase cells. Their results suggest that mitotic chromatin organization differs dramatically from all other cell cycle stages. In mitotic cells, previously described organization levels disappear. In metaphase, chromosome compartments as well as TADs structures appear to be absent throughout the entire genome. After cell division, the chromosomes decondense and reposition themselves in the nucleus. Chromatin structures then re-appear in early G1.

This capacity of erasing the chromatin structure and re-establishing it upon exit from mitosis raises the question of the mechanisms involved in such plasticity. The concept of 'mitotic bookmarking' has therefore been proposed (see Kadauke and Blobel (2013) for a review). The main idea is that some sequence-specific DNA binding factors (such as CTCF/cohesin), histone marks or DNA methylation remain to the mitotic chromatin, to allow a rapid emergence of the structure in early G1 phase. It is therefore tempting to speculate that these bookmarking factors are also able to propagate the structural information to daughter cells.

Interestingly, the probability of observing a contact between two loci in mitotic chromosome suddenly falls to zero at a distance of approximately 10 Mb, therefore suggesting that chromatin is organized differently above and below 10Mb. This observation suggests two levels of organization of mitotic chromosomes. Using polymer physics and modelling, the mitotic chromosome organization was finally described as a combination of linear ordering that separate loci by more than 10 Mb, together with consecutive 80-120kb loops (Naumova et al. (2013)).

1.2.5 The 3D genome as a driver of disease-associated gene expression

Our understanding of chromatin organization has exploded in the last decade, allowing to better characterize the different mechanisms involved in epigenetic modifications. Given the important advances that these studies have provided into our comprehension of the epigenetic regulation of normal cells, their application to a disease context offers the possibility to explore the impact that perturbations in 3D genomic organization might have on cell regulation and disease (see Krijger and de Laat (2016) for a review). At the genetic level, common diseases or cancer are frequently associated with the acquirement of variants, located within genes, regulatory regions or non coding regions. In the context of cancer, different alterations can characterized a tumor and are usually caused by a few functional events (driver events), which occur among many non-functional events (passenger events), mainly located in the non-coding part of the genome (Vogelstein et al. (2013)).

So far, little attention has been paid to non-coding variants, as their interpretation, their link to the disease, and their functional impact remain difficult to assess. Recently, genetic and epigenetic alterations in the non-coding part of the genome, including distal regulatory elements such as enhancers or insulators, have been reported and found to impact gene expression in cancer (Taberlay et al. (2014)).

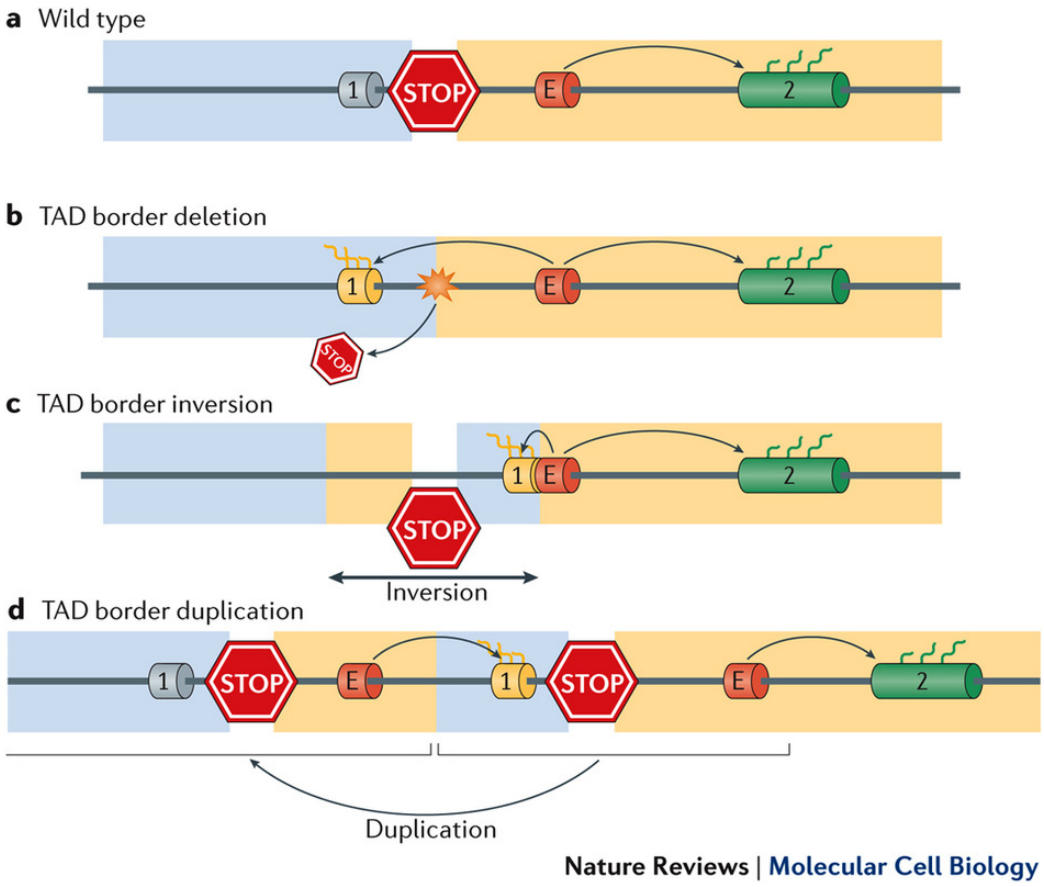


Figure 1.11: Mechanisms of TADs disruption in disease (From Krijger and de Laat (2016)) - **a.**Example of neighbouring TADs (blue and yellow). The gene 2 is activated by its enhancer (E), whereas the gene 1 is not expressed. **b.** The disruption of the TADs boundary results in the merging of the two TADs. Gene 1 can therefore be activated by the enhancer. **c,d** Examples of chromosomal rearrangements (inversion, duplication) leading to enhancer hijacking and gene misregulation.

This new field of investigation has led to intense interest in the spatial proximity and 3D organization of cancer genomes. In this context, enhancers can be inactivated by somatic mutations, or completely deleted, leading to a transcriptional silencing of its target genes. Or conversely, complex structural rearrangements can also relocate an enhancer to a new genomic environment, leading to new interactions and to activation of neighbouring genes like proto-oncogenes. This mechanism is usually called 'enhancer hijacking', and can involve different types of structural rearrangements, such as translo-

cation bringing enhancers from one chromosome in proximity of target genes in another chromosome, deletion of DNA region between active enhancers and proto-oncogenes or inversions of DNA regions switching enhancer in proximity to proto-oncogenes (see Beroukhim et al. (2016) for a review). A typical example has been described in medulloblastoma where the *Gfi1* gene family was juxtaposed to active enhancer elements therefore initiating oncogenic activity (Northcott et al. (2014)).

Interactions between enhancers and target gene promoters occur within the limit of the TADs they belong to. A disruption of a TAD boundary can therefore have an impact on the local chromatin organization, allowing new interactions, and possibly new gene regulation. Recently, Franke et al. investigated to effect of DNA duplication on TADs organization and gene expression. Using mouse model, they demonstrated that the duplication of an inter-TAD region, could lead to the formation of new TAD, therefore affecting the original chromatin organization. Interestingly, when this duplication contains the *Kcnj2* gene and the *Sox9* regulatory region, the formation of the new TAD will results in the activation of *Kcnj2* by the *Sox9* regulatory elements and is associated with a limb malformation phenotype.

In cancer, alterations of CTCF insulators sites and TAD boundaries have also been reported with dramatic effects on the genes regulation. Mutations of the CTCF/cohesin complexes have been reported as a driver events involved in numerous type of cancer (Corces and Corces (2016), Figure 1.12). In lymphoblastic leukaemia, recurrent deletions of CTCF sites have been identified, leading to inappropriate enhancer-promoter interactions (Hnisz et al. (2016)). Among them, deletion of CTCF sites around the proto-oncogenes *Tal1* or *Lmo2* induced their over-expression. In the same way, a recurrent deletion of a genomic region spanning the *TP53* gene has been described in prostate cancer (Taberlay et al. (2016)). This deletion results in a division of a single TAD into two smaller TADs. On the contrary, larger chromosomal rearrangements can also merge the TADs that carry the chromosomal breakpoints, leading to new interactions (Krijger and de Laat (2016)).

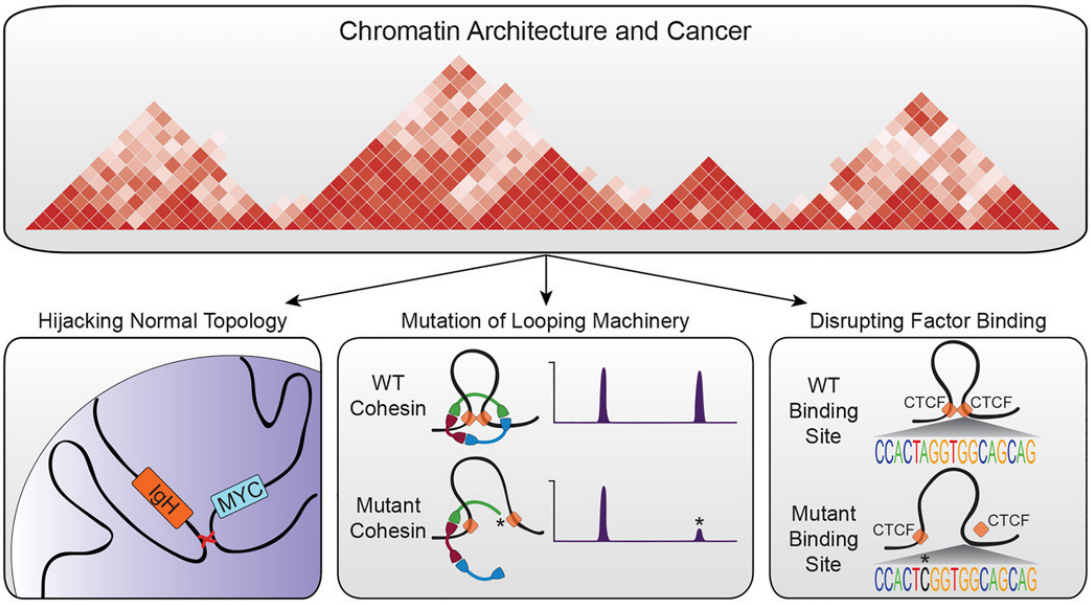


Figure 1.12: Impact of CTCF/cohesin mutations on chromatin architecture - The chromosome architecture can be associated to cancer in different ways. On the left, the MYC/IGH loci have been described in close proximity in some normal cell types, therefore, encouraging genomic translocation between the two loci reported in lymphoma. Mutation in cohesin complexes or CTCF binding sites, can also have dramatic effect on the binding of transcription factors (illustrated by ChIP-seq peaks), or the formation of chromatin loops. From Corces and Corces (2016).

In addition to functional changes due to structural rearrangements, other mechanisms involving insulators binding have been explored in cancer. One of the most obvious is the acquisition of somatic mutations at CTCF/cohesin-binding sites. These mutations are frequently observed in various cancer types, and affect the chromatin structure by inhibiting the binding of the CTCF protein (Katainen et al. (2015), Poulos et al. (2016)). In the same way, hypermethylation of CTCF sites have also been reported in IDH mutant gliomas leading to a loss of insulation between TADs and aberrant gene expression. Among the altered genes, the disruption of boundaries lead to an activation of the *Pdgfra* glioma oncogene (Flavahan et al. (2016)).
In addition, somatic mutations in cohesin complex proteins, that play critical role in TADs organization and chromosome looping, have been described in various types of cancer (see Hill et al. (2016) for a review). Mutations in genes encoding components of the cohesin complex cause developmental disorders and cancer in humans. Muta-

tions in *Stag2* protein have been reported in leukaemia, bladder cancer, glioblastoma, Ewing sarcoma, melanoma, cervical carcinoma, and haematological cancers (Hill et al. (2016)). As for CTCF, mutations in cohesin have an obvious effect on gene expression changes of crucial oncogenes or tumor suppressors. In addition to its role in genome organization and insulation, cohesin also plays a role in homologous recombination and DNA repair. Mutations in cohesin complex therefore increase the genome instability due to deficient DNA replication and/or repair.

Thus, changes in chromatin organization at different scales are now considered as key players in cancer, as well as important potential biomarkers. This suggests that disruption of chromatin architecture could be at the origin of tumors formation. Many recent studies started to investigate the link between local structure and gene regulation in cancer. These works rise new exciting questions, paving the way to new prognostic and predictive factors in rare diseases and cancer.

1.2.6 Changes in chromatin structure between normal and tumor cells

Most of the mechanisms related to chromatin organization and cancer presented above relies on local disruption of specific interactions between target genes and their regulatory elements. However, at the genomic scale, little is known about potential changes in chromatin organization that occur during tumorigenesis. Barutcu et al. recently investigated differences in chromatin organization between the MCF-10A mammary epithelial and the MCF-7 breast cancer cell lines. Their study reported changes in the physical proximity of gene-rich, small chromosomes which tend to be spatially closer in the MCF10a cells compared to normal MCF7 cells. Interestingly, the decrease in interaction frequency between small chromosomes in MCF7 cells is associated with a higher fraction of open compartments in the same chromosomes, wich are themselves associated to higher genes expression, including genes involved in known oncogenic pathways.

Interestingly, while around 12% of compartments switches occur between MCF10a and MCF7 cells, their TADs boundaries appear to be strongly conserved (85%). Contrastingly, a higher number of smaller TADs have been reported in prostate cancer cells (LNCaP) compared to normal prostate cells (PrEC), leading to cancer-specific TADs boundaries occuring at regions harbouring copy number changes (Taberlay et al. (2016)). In prostate cancer cells, changes at the TADs level are frequently associated

with small deletions of a few megabases (up to 10 kb) located at the new TADs boundaries. As an example, Taberlay et al. reported the formation of new TADs at the *TP53* locus, which is commonly deleted in these tumors. The formation of these new cancer-specific TADs is associated with abnormal regulation of neighboring genes.

These studies therefore demonstrated the interest of integrating information about the chromatin structure into cancer genomic projects. Being able to explore changes at the chromosomal compartments and TADs levels is a powerful approach to better understand the molecular mechanisms involved in tumorigenesis and the link between genome alteration, epigenetic marks and gene expression.

1.3 Chromosome conformation capture techniques

Two different types of experimental approaches are mainly used to explore the genome organization in three-dimensional space. Imaging techniques such as fluorescence in situ hybridization (DNA FISH, Volpi and Bridger (2008)) were extensively used in the past and are still considered as the state-of-the-art techniques to investigate the genome structure. In 2002, the development of the Chromosome Conformation Capture (3C) technique by J. Dekker has revolutionized our ability to explore the three-dimensional architecture of genome (Dekker et al. (2002)). Since then, the 3C technique and its derived approaches, have provided important insights into our comprehension of chromatin architecture.

The differences between the two types of methods mainly rely on the throughput and the resolution of the analysis. In comparison with microscopy methods, 3C-based techniques allow a systematic measure of DNA structure between multiple genomic loci, at high resolution and in a cell population. Broadly speaking, many studies have reported concordant results between 3C-based techniques and 3D DNA FISH, regardless the investigated spatial range or technical approach used. In practice, the two approaches are commonly used to validate each other. However, discrepancies between the two types of methods have also been found (Williamson et al. (2014)). Despite various technical issues and potential biases that can explain some of the observed differences, it is important to notice that both techniques do not measure exactly the same signal (see Dekker (2016); Giorgetti and Heard (2016) for reviews). 3C-based methods only detect events where the two loci are close in three-dimensional space (i.e.. closer than a certain distance). This signal is represented by an observed contact probability between two loci, but can, in theory, be detected in only a small fraction of cells. DNA FISH, on the other hand, allows a quantitative evaluation of the observed distances between two loci inside single cells. It therefore allows to estimate a distribution of distances between pairs of loci within a cell population (Giorgetti and Heard (2016)). This distribution of observed distances is characterized by a mean and a variance. The mean thus reflects the distance observed in the majority of cells, and the variance, the cell-to-cell variability of these distances. This fundamental difference in the two types of methods can lead to different interpretation. Nevertheless, with appropriate cautions and experimental design, the two methods can also be seen as complementary

approaches that can together bring comprehensive insights into the organization of the genome.

Since the development of the 3C-based techniques, many variants have been proposed (Table 1.1). The following section aims at presenting these 3C-based techniques and their specificities.

1.3.1 3C-based techniques

The 3C technique introduced by J. Dekker (Dekker et al. (2002)) relies on the idea that digestion and re-ligation of fixed chromatin allow to estimate the probability of contact between two genomic loci. The first experimental steps are common to many 3C derived methods. All these methods have been described in details in several excellent reviews (Barutcu et al. (2016); de Wit and de Laat (2012); Denker and de Laat (2016); Ramani et al. (2016)).

First, the chromatin is fixed, most often using formaldehyde. Next, the fixed chromatin is digested. Although different variants of digestion protocol were proposed (see section 1.3.2), this step usually involved restriction enzymes that cut the DNA more or less frequently, following a specific DNA motif. After digestion, the cross-linked fragments are ligated together in diluted condition that favour ligation of the juxtaposed DNA fragments. Therefore, loci that are close in space, should be ligated to each other to generate linear or circular chimeric DNA fragments composed of the two interacting loci. This final template is the input of all 3C-based methods (Figure 1.13).

The initial 3C experiment is usually described as 'one-to-one approach'. In that sense, the goal is to quantify the contact probability between two targeted genomic loci. Ligation junctions are amplified by Polymerase Chain Reaction (PCR) using primers that specifically hybridize the end of the two targeted fragments. Thus, the 3C experiment is driven by an a-priori knowledge about the genomic regions that should interact. In addition, the interpretation of the results is critical. Appropriate controls such as data normalization, background of random collisions and control of PCR efficiency are mandatory (Dekker (2006)). When properly controlled, the 3C experiment remains a powerful method to quantify the contact frequency among two genomic loci.

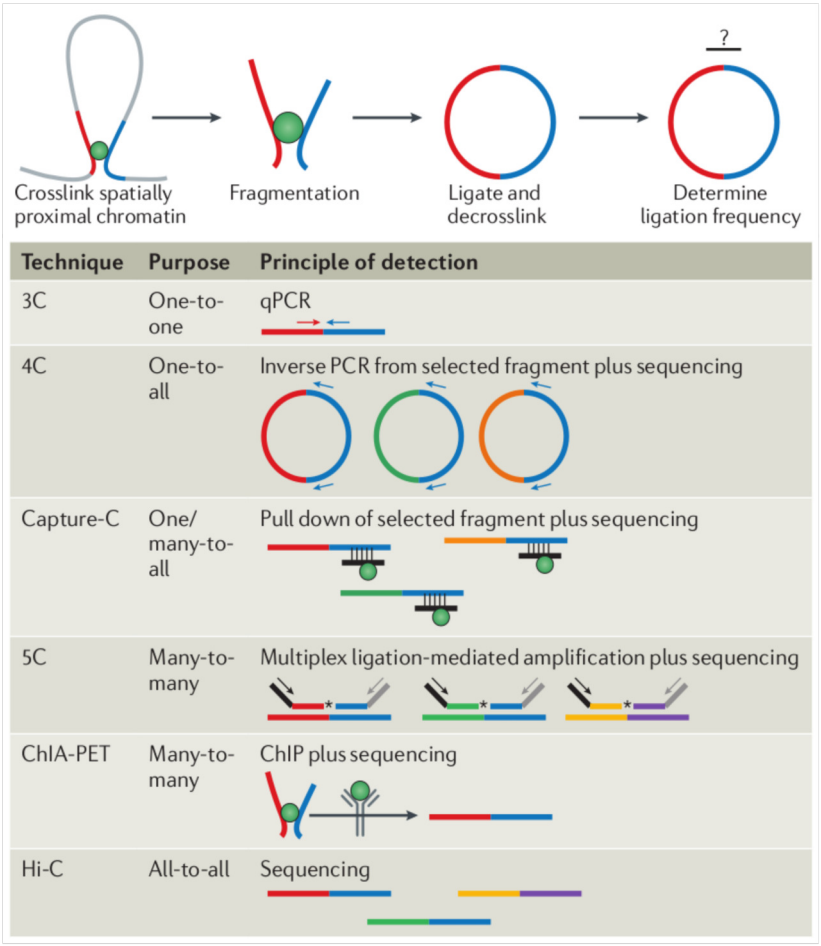


Figure 1.13: Main steps of Chromosome Conformation techniques (From Krijger and de Laat (2016)) - Chromosome conformation techniques are all based on cross-linking of DNA fragments that are closed in space. After fragmentation, these interacting fragments are ligated together and de-crosslinked before quantification. The principle of detection then varies from a technique to another.

A couple of years after the introduction of 3C experiment, the Circular Chromosome Conformation Capture (4C) was proposed as a 'one-versus-all' measure of contact frequency. While the original 3C experiment is usually limited to short range interaction (up to a distance of $\sim$ 1 Mb), the 4C experiment allows to detect chromatin interaction between a single restriction fragment with the rest of the genome, and therefore regulatory elements distant from several mega-bases away. Different protocols and adaptation of the 4C have been proposed (Sexton et al. (2012); Simonis et al. (2006);

van de Werken et al. (2012); Wrtele and Chartrand (2006); Zhao et al. (2006)). The 4C experiment is based on the same principles than the 3C experiment. One fundamental difference is the generation of circular DNA during ligation. One must nevertheless note that the 4C protocols are often based on a second round of digestion followed by another ligation step resulting in smaller circularized molecules. These molecules are then used as a template for reverse PCR using primers specific to the viewpoint of interest, that will amplified the interactor sequence ligated with the targeted locus. The amplified regions can then be identified using high-throughput technologies such as micro-array or sequencing. As results, the 4C signal is expected to decrease with genomic distance from the viewpoint to finally reaches a certain level of noise for larger distances. Dedicated computational methods were then developed in order to detect loci that significantly interact with the viewpoint, taking into account the linear distance between them.

Although recent 4C protocols can now be multiplexed to study multiple viewpoints in a single experiment, the method is not designed to explore the landscape of all possible interactions within a given genomic region. To answer this need, a 'many-versus-many' version of the 3C, named Carbone Copy Chromosome Conformation Capture (5C) was proposed (Dostie et al. (2006)). Briefly, the 5C is based on the same first steps than the 3C experiment. However, it allows to simultaneously interrogate several hundred of loci, spanning a genomic region of several mega-bases. To this end, 5C primers that are complementary to the restriction fragments of interest are designed. Both forward and reverse primers are used. The mixture of 5C primers is then hybridized to the 3C library. Only pairs of forward and reverse primers anneal next to each other in a head-to-head manner, can be ligated. The 5C library can be then amplified and analysed using high-throughput technology such as micro-array or sequencing. The output is a contact map of many-versus-many interaction frequencies. The contact maps are characterized by a strong diagonal representing higher contact frequencies between loci that are linearly close in the genome.

In practice, a 5C experiment is limited to a genomic region of a few mega-bases. In addition, the design of hundred of primers remains an important constraint of the method. To overcome these limitations and to be able to capture a view of all genome-wide interactions, Lieberman-Aiden et al. developed the Hi-C technique. The Hi-C is therefore an 'all-versus-all' chromosome conformation technique. Again, the protocol remains

very similar to the 3C method. In Hi-C, the 5' overhang left after digestion is filled with biotin-labelled nucleotides. The library of ligated fragments is then shared and the biotin fragments are pull-down, ensuring an enrichment of ligated fragments. These fragments are sequenced using a paired-end strategy, where the sequencing paired reads are respectively associated to the two interacting loci. After processing, the results is a genome-wide interaction matrix which reports the number of time two fragments interact together. The resolution achieved with the Hi-C technique mainly depends on the sequencing depth. In the original Hi-C study, Lieberman-Aiden et al. generated around 9 millions of paired-end (PE) reads to extract 1Mb resolution Hi-C contact maps. More recently, Rao et al. (2014) went up to 5 billion of PE reads per sample to reach the 1kb resolution. These differences are motivated by the biological question and the level of chromatin organization to explore. But it also has an obvious consequence on the cost of the experiment.

In parallel to 3C-based techniques, another type of method was developed to be more cost-effective and to enrich for DNA associated with specific DNA-binding proteins. The ChIA-PET technique is a combination of 3C and chromatin immunoprecipation sequencing (ChIP-seq). The main idea is to use an antibody to pull down ligation junctions bound by a protein of interest (Fullwood et al. (2009); Li et al. (2014)). Thus, in theory, the ChIA-PET can be used to detect the protein factor binding sites as well as the long-range contact associated with the protein of interest. In comparison to Hi-C, which gives a three-dimensional view of genome organization, the ChIA-PET gives information about the potential role of a protein in structuring the genome organization. The same idea was recently improved and a new technique named HiChiP was proposed (Mumbach et al. (2016)). The main differences between ChIA-PET and HiChiP is that the DNA ligation is established in situ, within the nucleus, then followed by chromatin immunoprecipitation on the 3C library. Comparing both approaches shows that HiChIP is much more efficient in capturing informative contacts (Mumbach et al. (2016)).

1.3.2 Hi-C-based techniques and variants

Since its development, the Hi-C technique has been widely used to characterize chromatin structure in many different contexts. And in the same time, new developments have been made leading to so-called Hi-C variants (Table 1.1).

One of the first improvement in the Hi-C protocol was the choice of the restriction enzyme, which has a direct impact on the resolution and the size of the digested fragments. The first Hi-C experiment was based on HindIII, a restriction enzyme recognizing 6 base pairs (bp), and able to cut the human genome, in average, every 3.7kb. Recent studies now tend to improve the data resolution by using a 4-cutter enzymes such as MboI or DpnII, able to cut the genome every 435 bp in average. Other methods have then been proposed for chromatin fragmentation. DNase I was first used instead of restriction enzymes for the fragmentation of cross-linked DNA (Ma et al. (2015)). This method offers the main advantage to be independent of a restriction enzyme. The small size and random distribution of DNA fragments within a DNase Hi-C library provide a better genome coverage than an restriction enzyme-based Hi-C library. In addition, PCR duplicates can be more easily distinguished from independent ligation events, therefore leading to a better quantification of observed contacts (Ma et al. (2015)). However, the downstream analysis is also more approximative as non-informative contacts are also more complicated to filter out (see section 1.4). Similarly, micrococcal nuclease (MNase) has also been used to achieve a nucleosome-level resolution of chromatin organization in Yeast (Hsieh et al. (2015)).

Although the Hi-C technique has been extensively used these last years, the cost of the method to reach a sufficient resolution remains a limitation. For instance, in order to characterized TADs in mammals, a resolution of 20/40kb is required, representing around 400 million of PE reads. The sequencing depth should again be increased to characterize single chromatin loops, up to several billions of PE reads per sample. Based on the first Hi-C protocol, also called dilution Hi-C (Lieberman-Aiden et al. (2009)), around half of the data are expected to be used after processing to build the contact maps. One way of reducing the cost is therefore to improve the efficiency of the protocol and to generate more valid interaction products at fixed sequencing depth. In that sense, the in situ Hi-C protocol proposed by Rao et al., was a real advance in the field. The main difference with the dilution Hi-C protocol is that the ligation is performed inside the nuclei. This modification reduces the frequency of spurious contacts due to random ligation in dilute solution, and therefore improves the efficiency of the protocol (Nagano et al. (2015)).

Another way to reduce the cost of the Hi-C experiment and to increase the resolution, is to focus on a subset of chromatin contacts such as a targeted genomic region, a locus

of interest, or even specific categories of sequences such as promoter regions. Although the 4C or 5C methods could answer part of this need and are less expensive, they also suffer from low throughput and complex primers design. To overcome these limitations, targeted methods have been proposed by applying capture technologies to 3C or Hi-C libraries. The capture-C experiment was developed to interrogate interactions at hundreds of loci, at high-resolution, in a single experiment (Hughes et al. (2014)). The capture-C protocol is based on standard 3C libraries which are then hybridized with biotinylated oligos specific to a set of viewpoints of interest. The results is a 4C-like profile of interactions for each viewpoint. Recently, the technique was improved (Next Generation (NG) capture-C, Davies et al. (2016)) to include a second round of hybridization to the baits, in order to reduce the fraction of off-target reads. However, the capture-C still suffers from a couple of biases mainly due to capture efficiency or over-representation of ligations between independently captured sites.

Following the same idea, capture protocols have also been developed directly from Hi-C libraries. One of the main advantage in this case, is that the protocol can, in theory, be enriched in valid ligation products through biotin pull-down as in Hi-C. As an example, the DNase Hi-C protocol was combined with a DNA capture technology to specifically enriched the ligation products into contacts associated with genomic loci of interest (Ma et al. (2015)). This type of capture protocol can be extended to thousand of sequences as promoter regions. Promoter capture Hi-C was developed to identify distal sequences that significantly interact with gene promoters (Mifsud et al. (2015)). The technique involves thousands of biotinylated RNA oligomers specific to annotated promoters, used to enriched the Hi-C library with ligation fragments involved in promoter contacts. Thus, the method allows to investigate regulatory enhancer-promoter contacts at the genome-scale.

Method	Main features	References
3C	The founding method of the 3C family of techniques; for detecting chromatin interactions between a pair of genomic loci	Dekker et al. (2002)
4C	For detecting chromatin interactions between one locus and the rest of the genome	Simonis et al. (2006); Zhao et al. (2006)
5C	For detecting chromatin interactions between multiple selected loci	Dostie et al. (2006)
ChIA-PET	For detecting genome-wide chromatin interactions mediated by a particular protein	Fullwood et al. (2009); Tang et al. (2015)
Hi-C	For mapping whole-genome chromatin interactions in a cell population; proximity ligation is carried out in a large volume	Lieberman-Aiden et al. (2009)
Capture-C	Combines 3C with a DNA capture technology; equivalent to a high-throughput version of 4C	Hughes et al. (2014)
Capture-Hi-C	Combines Hi-C with a DNA capture technology; equivalent to a high-throughput version of 5C	Mifsud et al. (2015)
Targeted DNase Hi-C	Combines DNase or in situ DNase Hi-C with a DNA capture technology	Ma et al. (2015)
TCC	Similar to Hi-C, except that proximity ligation is carried out on a solid phase-immobilized proteins	Kalhor et al. (2011)
Single-cell Hi-C	For mapping chromatin interactions at the single-cell level	Nagano et al. (2013)
DNase Hi-C	Chromatin is fragmented with DNase I; proximity ligation is carried out in solid gel	Ma et al. (2015)
In situ Hi-C	Proximity ligation is carried out in the intact nucleus	Rao et al. (2014)
Micro-C	Chromatin is fragmented with micrococcal nuclease	Hsieh et al. (2015)
In situ DNase Hi-C	Chromatin is fragmented with DNase I; proximity ligation is carried out in the intact nucleus	Deng et al. (2015)

Table 1.1: Overview of 3C-based techniques (adapted from Ramani et al. (2016))

Finally, capture-Hi-C protocols can also be used to focus on a genomic region of interest at very high resolution (Franke et al. (2016)). In this context, capture-Hi-C can be seen as an improvement of the 5C technique, with the throughput and efficiency of modern Hi-C techniques. As previously, the main limitation of this method remains the cost and the quality of the capture technology.

All the 3C-based methods presented above were extensively used to assess functional implications of genome structure. Therefore, the choice of the method mainly depends on the question to address. The different aspects to consider are basically the number of loci of interest, the desired resolution, the ease of the protocol and the cost. Among the different protocols, in situ methods are more efficient, and thus highly recommended. Then, Hi-C remains the method of choice to have a comprehensive overview of the genome organization. However, and even if the sequencing cost is still decreasing, raising a very high resolution in Hi-C data is currently expensive. Capture performed on Hi-C libraries therefore represents an interesting option, especially to extract a detailed contact map of a locus of interest. Thus, Capture-Hi-C on a genomic region of a few mega-bases might be a good alternative to standard 5C protocol, although the cost of the capture design is not cheap. In the context of single viewpoints, capture-C technique may be preferred to 4C as the quantification of ligation looks more accurate (Davies et al. (2016)). On the other hand, the 4C sequencing is now based on a robust protocol, which might be cheaper than capture-C and leads to a higher fraction of usable reads.

1.4 How bioinformatics can help in understanding the genome organization ?

The emergence of sequencing technologies enabled us to make a leap forward in life sciences. As a reminder, the sequencing of the first human genome costed around 3 billion dollars and took more than 10 years to be achieved in 2003. Currently, the sequencing of human genome is accessible to any lab for a few thousand dollars. However, the users of these technologies usually underestimate the efforts which are then required once the data are available. How to manage these data ? How to extract relevant information from them ? Or to efficiently process them ? Finally, what is the real cost of such analysis ? Here are the reasons why computer sciences, such as bioinformatics, are now essential to provide a support to biology.

Among the different 3C-based technologies, the Hi-C is by far the one that generates the most data. Nevertheless, it is still far from being able to interrogate all interacting loci genome-wide. As an example, using a 6bp cutting restriction enzyme on human genome, there are almost 840 000 restriction fragments, leading to a potential interaction space of 7e11 pairwise interactions. This number reaches the 5e13 possible pairwise interactions with a 4bp cutting enzyme. Thus, being able to achieve a sufficient resolution is challenging, but also depends on the biological question. If the goal is to measure large scale structures of human genome organization, such as chromosome compartments, then a resolution of 1 Mb is usually sufficient, requiring a few millions read pairs. If the question requires to look into topological domains, a minimum resolution of 40kb is advised, requiring a few hundreds million read pairs. Finally, in order to study higher level of chromatin structure such as chromatin loops, several billions of sequencing pairs are required to cover the genome. The sequencing data generated will therefore have a size in the order of hundreds of gigabytes (Gb) to terabytes (Tb). Thus, high performance computing and big data management are needed to process these data. Developing dedicated computational solutions is challenging but necessary to extract relevant information and explore the genome structure (Shavit et al. (2016)).

The sequencing depth is therefore one of the most critical point to consider during high-throughput C experiments, but it is not the only one. The protocol itself is an important factor. For instance, recent advances in Hi-C protocols such as the in-situ

Hi-C protocol (see section 1.3.2), have significantly increased the proportion of valid 3C products used to build the contact maps, therefore allowing to reduce the sequencing depth, and therefore the cost of the experiment. On the contrary, current capture-C protocols suffer from the absence of biotin pull-down during the library preparation. As a consequence, a large fraction of sequencing reads are unligated, non-informative fragments. Another important factor to highlight is the library complexity which is defined by the number of unique 3C products that exist in a Hi-C library. The library complexity therefore mainly depends on the number of cells, the quantity of starting material and the quality of the library. A library with low complexity will quickly saturate with increasing sequencing depth, leading to an incomplete view of the interaction landscape.

1.4.1 Typical workflow for Hi-C data processing

A guide of Hi-C data analysis has been recently proposed by the Dekker's lab (Lajoie et al. (2015)). A typical analysis workflow for Hi-C data can be divided into four main steps; reads mapping, selection of valid 3C products, generation of contact matrices and data normalization.

1.4.1.1 Alignment on a reference genome

The analysis of any 'C' technology associated with high-throughput sequencing starts with raw sequencing files. Usually, the sequencing of the 3C ligation fragments is performed using paired-end sequencing (PE), reading 50 to 100 bp from each 5' end of the molecule. But, in practice the ligation products can also be sequenced using single-end reads, assuming that the read is long enough to cover both interacting sites. This strategy is still commonly used in 5C experiments. In addition, the current sequencing technologies now allow to generate longer reads, up to 250 bp from both 5' end. Using such read length can improve the mapping of repeated regions which are usually difficult to explore with short reads, and can ease the detection of allele-specific interactions by increasing the probability of covering phased single-nucleotide polymorphisms (SNPs) (Li et al. (2017)).

In theory, the sequencing reads can be aligned on a reference genome with any alignment software. However, Hi-C reads have two specificities that have to be taken into account. First, the two paired reads have to be aligned independently on the genome.

Indeed, all recent alignment software propose a PE mode, allowing to align both ends of the fragment in a single process. However, this type of alignment relies on an expected fragment size distribution, and therefore on an expected distance between the two reads of the same fragment. In 3C-based experiments, the distance between the two reads of a ligation product can vary from a few bases to several mega-bases, or can even imply different chromosomes in the case of inter-chromosomal contacts. For this reason, it is therefore not advised to use the alignment software in PE mode, but rather to align independently the two reads on the reference genome.

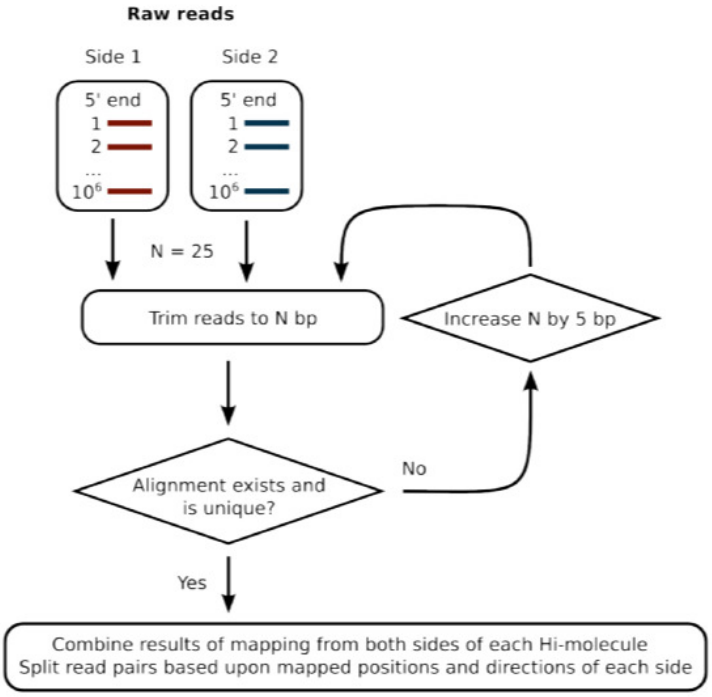


Figure 1.14: Iterative mapping procedure (From Imakaev et al. (2012)) - This mapping strategy is still commonly used to align Hi-C reads on a reference genome. Reads are first truncated to 25 bp, aligned to the genome and extended as much as possible through an iterative procedure. This strategy allows to align chimeric reads spanning the ligation site.

Then, depending on the position of the ligation site, the size of ligation fragments after shearing and the length of sequencing reads, it is also possible that one of two sequencing read spans the ligation site. In this case, the two interacting loci will be part

of the same chimeric read, and will therefore require a dedicated mapping strategy. In order to solve this problem, an iterative mapping approach has been proposed (Imakaev et al. (2012)). The idea is to start by aligning the first 25 bp of each sequencing read. Reads that do not aligned at a unique position on the genome are then extended by 5 more nucleotides and re-aligned. The same process is repeated iteratively until all reads were uniquely aligned or until the reads cannot be further extended. This strategy therefore allows alignment the 5' of chimeric read, until the extension process reaches the ligation site. However, this procedure is time and memory consuming as several iterations are usually required, especially for longer reads. In addition, the extension of 5bp is arbitrary, and by definition, the method does not allow to consider reads coming from repeated regions of the genome.

1.4.1.2 Detection of valid 3C products

Once aligned on a reference genome, each sequencing read can be assigned to a restriction fragment. The restriction fragments are defined by in silico digestion of the reference genome by the restriction enzyme(s) used during the library preparation. Both reads are expected to map near a restriction site, and with a distance within the range of molecule size distribution after shearing. Fragments with a size outside the expected range can be discarded but are usually the result of random breaks or star activity of the enzyme, and can therefore be included in downstream analysis. Low mapping quality reads, and singletons can be discarded. Although looking at interactions from repeated elements could be of interest, reads aligned at multiple loci on the genome are usually also discarded from the analysis.

After assigning each of the PE reads to restriction fragments, a selection of valid 3C products is performed. The valid 3C products correspond to sequencing read pairs involving two different restriction fragments, for which we can in silico reconstruct the ligation event. Read pairs mapped to the same restriction fragment can either be classify as unligated fragments (i.e. 'dangling end' fragments) or circularized fragments (i.e. 'self-circle' fragments). The proportion of these non-informative fragments are very important to assess the overall quality of the experiment. For instance, the fraction of dangling-end pairs is a good indicator of the efficiency of the ligation step during the library preparation. Finally, duplicated read pairs aligned at exactly the same genomic coordinates are usually associated to PCR artefact, and filtered out.

This step of the analysis therefore allows to generate a list of ligation products or contacts, and how many times these contacts occur in our experiment. Except for single-cell Hi-C, all other protocols are based on cells population. This point is critical for data interpretation. For instance, if all cells of a population had a distinct genome organization, the observed interaction frequencies would be an averaged, smoothed, interaction maps, with little structure. This doesn't mean that the genome of each cell is not compacted and structured, but only that if such organization exists, it is not consistent between cells.

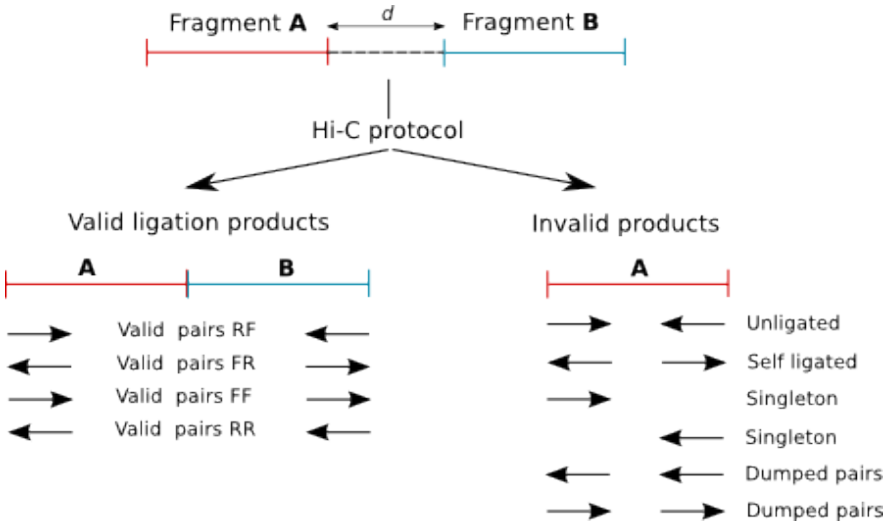


Figure 1.15: Selection of valid 3C products - A valid ligation product necessarily involves two distinct restriction fragments (A and B). Then, according to the ligation event, 4 different sequencing products are expected (Forward (F) Reverse (R), RF, RR, FF). In theory, as the ligation is a random process, a similar fraction of each type of product is expected. Invalid pairs can also be classified in different types ; unligated, self-circle or singleton. These products can be distinguished using the sequencing reads orientation. Dumped pairs represent here the pairs that cannot be explained and are likely to be mapping artefacts. Note that additional filter can be applied, for instance based on the minimum distance d between the sequencing reads.

In theory, the first steps of the analysis presented above can be applied to any Hi-C-based experiments, such as dilution Hi-C, in-situ Hi-C, capture-C or capture-Hi-C. In the context of Hi-C protocols without restriction enzyme digestion (DNase Hi-C, Ma et al. (2015)), invalid pairs can be filtered out using the distance between the two sequencing reads.

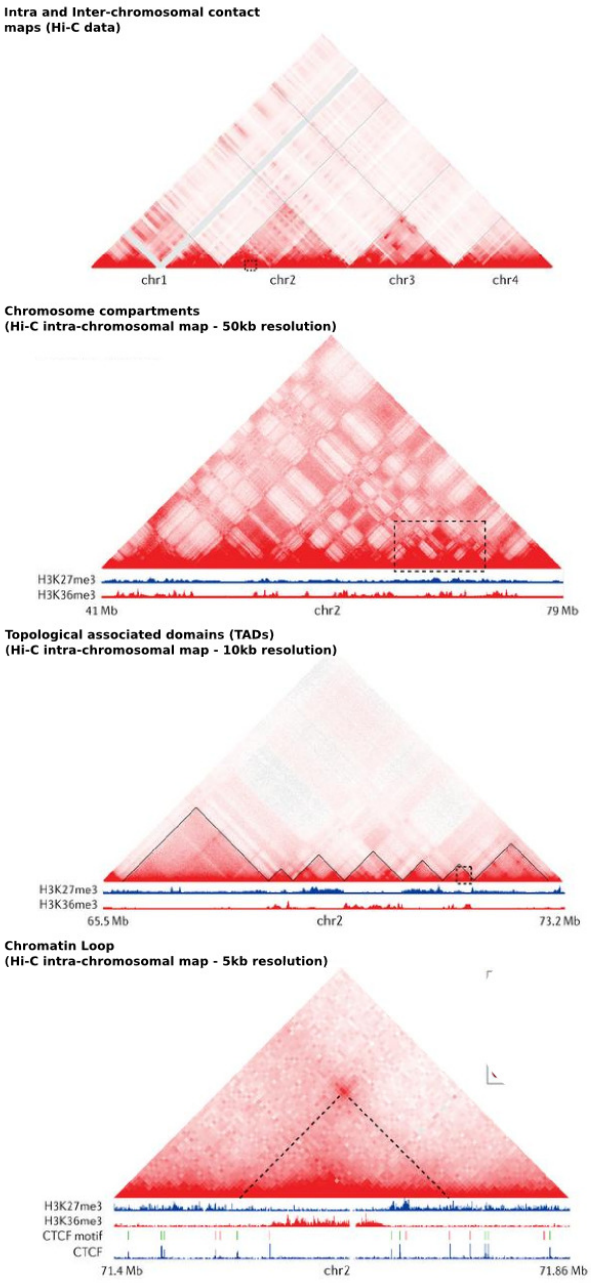


Figure 1.16: Hierarchical organization of the genome - Hi-C contact maps from GM12878 cells (Rao et al. (2014)) and chromatin immunoprecipitation sequencing (ChIP-seq) tracks for H3K36me3 (red) and H3K27me3 (blue) histone marks at different resolutions. From top to bottom ; i) intra and inter-chromosomal maps of chromosome 1 to 4 ii) Intra-chromosomal map of chromosome 2 and visualization of chromosome compartments at a 50kb resolution iii) Visualization of TADs at 10kb resolution iv) chromatin loop domain. Adapted from Bonev and Cavalli (2016).

In practice, discarding all interactions occurring in a range of 1kb allows to filter most of the invalid products.

1.4.1.3 Contacts maps

In order to generate the contact maps, the genome is then divided into bins of the desired size, and the number of contacts observed between each pair of bins is summarized by the sum of interactions involving the overlapping restriction fragments.
Binning the data reduces the resolution and the complexity of observed interactions, but conversely, increases the signal to noise ratio. Data are usually binned at multiple resolutions, from 1Mb, up to 10 or 5kb according to the sequencing depth. Both intra and inter-chromosomal maps are generated. A typical intra-chromosomal maps is characterized by a strong diagonal and contact frequencies that decrease with the linear distance. In addition, a higher contact frequency is expected from loci which are on the same chromosome (i.e. intra-chromosomal, cis contacts), compared to contact between distinct chromosomes (i.e. inter-chromosomal, trans contacts). These observations are in adequacy with the organization of chromosomes in territories where each chromosome occupies a distinct space in the nucleus, therefore promoting intra-chromosomal contacts. Thus, the simple ratio of the number of intra versus inter-chromosomal contacts is also commonly used as a quality control of Hi-C experiment. A noisier Hi-C experiment will usually result in a lower cis/trans ratio of contacts. However, it is also interesting to note that this ratio can change according to the organism or the biological context. This is for instance the case in the context of centromere co-localization (Varoquaux et al. (2015)) or cancer genomes (Barutcu et al. (2015)).

1.4.1.4 Hi-C data normalization

As any other high-throughput data, Hi-C data contain different biases mainly due to GC content, mappability and effective fragment length (Yaffe and Tanay (2011), Figure 1.17). An appropriate normalization method is therefore mandatory to correct these biases.

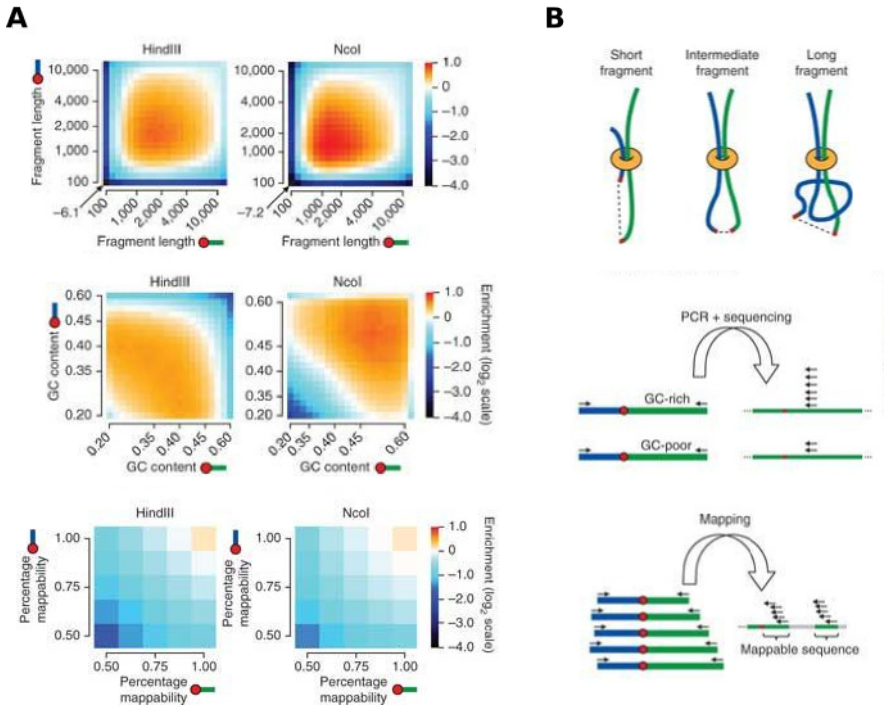


Figure 1.17: Sources of bias on Hi-C data - A. Systematic biases such as GC content, fragment lengths and mappability have been identified in Hi-C experiments. **B.** Effect of these biases on ligation efficiency, PCR and reads mapping. Adapted from Yaffe and Tanay (2011)

In the past few years, several methods and packages have been developed to normalize Hi-C data (see Ay and Noble (2015) for a review). These methods can be divided in two main classes, respectively based on explicit-factor correction or on matrix balancing algorithms. Explicit-factor normalization methods require a priori knowledge of the biases in Hi-C data. Yaffe and Tanay first proposed a non-parametric model to estimate the probability of observing a contact between two loci given their GC content, mappability and fragment length. The main limitation of this method is its computational cost, which make it very difficult to use in practice. Then, Hu et al. (2012) proposes a much faster explicit correction method, based on poisson or negative binomial regression model which can be applied at the bin resolution, and give similar performance compared to the Yaffe and Tanay method. In brief, the method estimates an average GC content, mappability and effective fragment length, for each genomic bin, and then used a generalized liner model to correct these biases in Hi-C contact

map. In term of computational time, the methods is more than 1000 times faster compared to the original Yaffe and Tanay method (Hu et al. (2012)).

Method	Model Assumption	References
Yaffe & Tanay	explicit-factor correction	Yaffe and Tanay (2011)
HiCNorm	explicit-factor correction	Hu et al. (2012)
ICE	equal visibility	Imakaev et al. (2012)
SCN	equal visibility	Cournac et al. (2012)
Knight and Ruiz	equal visibility	Rao et al. (2014)

Table 1.2: Methods to correct Hi-C experiments from systematic biases. Two main types of method have been proposed to correct Hi-C data from systematic biases, either based on explicit-factor correction or matrix balancing algorithms. The first one corrects for a-priori known biases, such as GC content, mappability and fragment length, whereas the second is able to correct for any bias under the assumption of equal visibility.

In addition to these methods, other strategies based on matrix balancing algorithm have been proposed (Cournac et al. (2012); Imakaev et al. (2012); Rao et al. (2014)). Contrary to the explicit factor correction methods, one interest of the matrix balancing approaches is that it does not assume any specific source of bias and should therefore be able to correct for all factors affecting contact frequencies. However, these methods are based on two strong assumptions. First, it assumes that the bias observed between two interacting loci can be simplified as the product of the two locus-specific biases. Then, it assumes that all fragments or bins should have the same number of contacts genome-wide. The latter is usually refered as an assumption of 'equal visibility'. In practice, it means that the total sum of genome-wide contact for each locus is expected to be the same after data normalization.

Among these methods, the iterative correction (ICE, Imakaev et al. (2012)) is one of the most popular method for Hi-C data normalization. In short, the method iterates to reach a uniform genome-wide coverage along the genome. The matrix of contact probabilities, M, for all given pairs of regions (i,j) is normalized such as $\sum_{i,i\neq j,j\pm 1} M_{ij} = 1$ for each region j. Although its assumption may require further exploration and can be discussed in some cases, the ICE normalization has been widely used by recent studies due to its conceptual simplicity, parameter-free algorithm and ability to correct for

unknown biases. In addition, the method can be applied to protocols that are not based on restriction enzyme digestion, such as DNase Hi-C.

1.4.2 Computational considerations

As already discussed, the computational resources required to process Hi-C data mainly depend on the sequencing depth and the resolution of Hi-C contact maps. Using the pipeline proposed by Imakaev et al. (2012), the processing of 400 Millions Hi-C sequencing pairs requires around 30 hours. However, a couple of computational considerations can also help in improving the efficiency of Hi-C data analysis. As an example, the first steps of Hi-C analysis workflow, from the raw sequencing files to the list of valid interaction products can, in theory, be parallelized in order to achieve a significant speed up in the Hi-C processing.
Once processed, a contact map is defined as a matrix of counts associated with a description of the genomic bins. Although the size of the Hi-C processed data is usually much smaller than the original data, manipulating such big matrices can become a real challenge. For instance, a human 20kb genome-wide map is represented by a square matrix with 150 000 rows and columns which can be difficult to manage for further analysis. Normalization algorithms typically require a significant amount of memory to load and process contact matrices, and dividing the algorithm into independent tasks using multiple cores is not always straightforward. In order to reduce the data storage, several pipelines and consortium like ENCODE proposed to use the hdf5 format. This format is indeed much more efficient that the text format to store the data, but its use implies computational constraints such as dedicated libraries and conversion to simpler format to explore or analyse the data. Other solutions can be proposed to reduce the disk space of Hi-C data and to speed up their processing (see our work Chapter 2).

1.4.3 Available solutions for Hi-C data processing

From 2012 to 2016, we observed a boom in the development of computational pipelines or methods for Hi-C data analysis. Among these tools, many are dedicated to a single task, such as normalization, or statistical confidence of contacts. A few are able to process Hi-C data from raw reads to ready-to-use contact frequencies. The *hiclib* package (Imakaev et al. (2012)) is still a reference in the field as it was the first available pipeline to process Hi-C data from raw sequencing reads to normalized contact maps. Its main

disadvantage is that *hiclib* is a python library which therefore requires programming skills. In addition, the *hiclib* library has some limitation to analyse and normalize very high resolution data. We, and others, therefore proposed to develop new bioinformatics pipelines to process these data (see Chapter 2 and Table 1.3). The tools presented in Table 1.3 are restricted to pipeline for Hi-C data processing. A more complete list of tools for Hi-C analysis and visualization is available on the 4D nucleome portal.

Tool	Mapping	Mapping Type	Read-pair Filtering	Normalization	Year	Maintained
hiclib	√	Iterative	√	Balancing	2012	√
HiCUP	√	Pre-Truncation	√	-	2012	√
HOMER	-	-	√	distance	2012	-
HiC-inspector	√	-	√	-	2015	-
HIPPIE	√	-	√	-	2015	-
HiC-Box	√	Iterative	√	Balancing	2015	√
HiCdat	√	-	√	Balancing, HiCNorm	2015	√
HiFive	-	-	√	Balancing	2015	√
HiC-Pro	√	Trimming	√	Balancing	2015	√
TADbit	√	Iterative	√	Balancing	2016	√
Juicer	-	Post-process	√	Balancing	2016	√

Table 1.3: Tools for Hi-C data processing. We restrict this list of tools to pipeline starting from raw sequences or aligned data, and generating ready-to-use contact information. We note as 'maintained' the tools with at least one update in 2016. Adapted from Ay and Noble (2015)

All these pipelines start from raw sequencing reads or aligned data. The mapping strategy can vary from a tool to another. Several pipelines use the iterative mapping procedure proposed by Imakaev et al.. Others, like the *HICUP* pipeline, propose to cut the 3' of reads after the ligation site before mapping (Wingett et al. (2015)). This method therefore avoids multiple mapping steps. In the same way the *Juicer* pipeline (Durand et al. (2016b)) proposes to run a local mapping approach, and then to post-

process the chimeric reads. Then, all methods propose to filter the read pairs, in order to select the valid 3C products. The classes of filtered read pairs can vary from a tool to another with more or less details about the fraction of reads in each class. Finally, all presented tools propose to remove systematic biases from the contact matrices, with the exception of *HICUP* which just returns the list of valid interactions. Normalization methods can be different from a tool to another, but most of them are based on matrix balancing algorithm or explicit factor correction like *HiCNorm* (Hu et al. (2012)).
These tools are not all maintained. And some are not designed to support very high resolution data. In this context, the Juicer pipeline was built on high-performance computing technologies. It is therefore able to process in parallel billions of reads, but required a dedicated cluster architecture.
Some tools, such as Juicer or TADbit also propose additional downstream analysis such as the detection of chromosome compartments and TADs.

1.4.4 Downstream analysis and interpretation of Hi-C data

Downstream analysis of Hi-C data mainly consists in extracting interaction patterns from the contact maps. In the beginning of Hi-C experiments, many different methods were proposed to achieve the same goal. This is partially explained by the fact that there was no explicit definition of the patterns we were looking for. Over the years, some methods were more frequently applied on different datasets and contexts, therefore validating their interest. Apart from the detection of significant contacts, the methods presented thereafter were implemented in our HiTC R package (Servant et al. (2012), see Annexe 1.6.2)). The HiTC R package is frequently updated to provide user-friendly implementations of current methods for 5C and Hi-C data analysis.

1.4.4.1 Distance-dependent contact frequency

The first pattern which can be extracted form the intra-chromosomal contact matrices is a distance-dependent decay of contact frequency. This pattern can be easily identify on the contact maps, appearing as a decrease of the contact frequency when one moves away from the diagonal (Figure 1.18A). It follows the intuition that regions which are linearly closed on the genome are more likely to interact, and thus, that the contact frequencies decrease as the genomic distances increase. The distance-dependent decay

of contact frequency is largely used in polymer physics to model the organization of the chromatin fibre into the cell nucleus (Fudenberg and Mirny (2012), Figure 1.18B). In addition, several downstream analysis methods first propose to remove this distance-dependent effect in order to highlight significant long-range contacts. Several methods can be used to estimate the distance-dependent effect. The simplest is to use the mean of observed contacts per genomic distance. The contact maps are then normalized using the ratio of observed versus expected counts (O/E). Other methods usually consist in fitting a function to the counts-distance relationship and to use this function to represent it. Several fitting functions, such as the Loess regression, have been successfully applied (Ay et al. (2014); Nora et al. (2012); Varoquaux et al. (2014)).

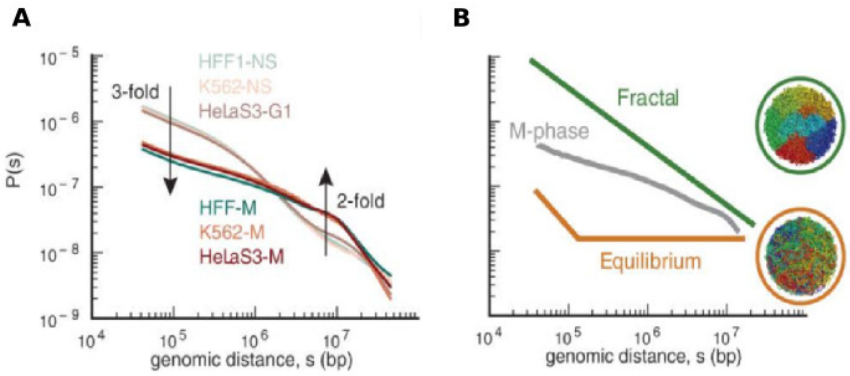


Figure 1.18: Contact probability as a function of genomic distance - A. Distance-interaction frequency plots for interphase and mitotic cells. In metaphase, the slope of the curve changes at 10 Mb, demonstrating a decrease of contact frequencies when loci are separated by more than 10Mb.**B.** This information can be then used in polymer physics in order to model the genome organization in 3D. Adapted from Naumova et al. (2013).

1.4.4.2 Detection of chromosomal compartments

One of the first downstream analysis proposed on large-scale contact maps (1Mb resolution), is the chromosome compartment calling. For each intra-chromosomal contact map, the chromosome compartment profile appears as a checker-board-like interaction pattern, shifting from blocks with either high and low interaction frequency (Figure 1.16). This pattern can be seen as genomic regions that alternate along the chromosome, and where the contact frequencies are higher between regions of the same type

compared to regions of different types. These genomic regions are usually presented as A/B chromosome compartments (see section 1.2.2.2). One way to identify these compartments is to apply a Principal Component Analysis (PCA) on the correlation matrix of the distance-corrected intra-chromosomal contact maps (Figure 1.19).

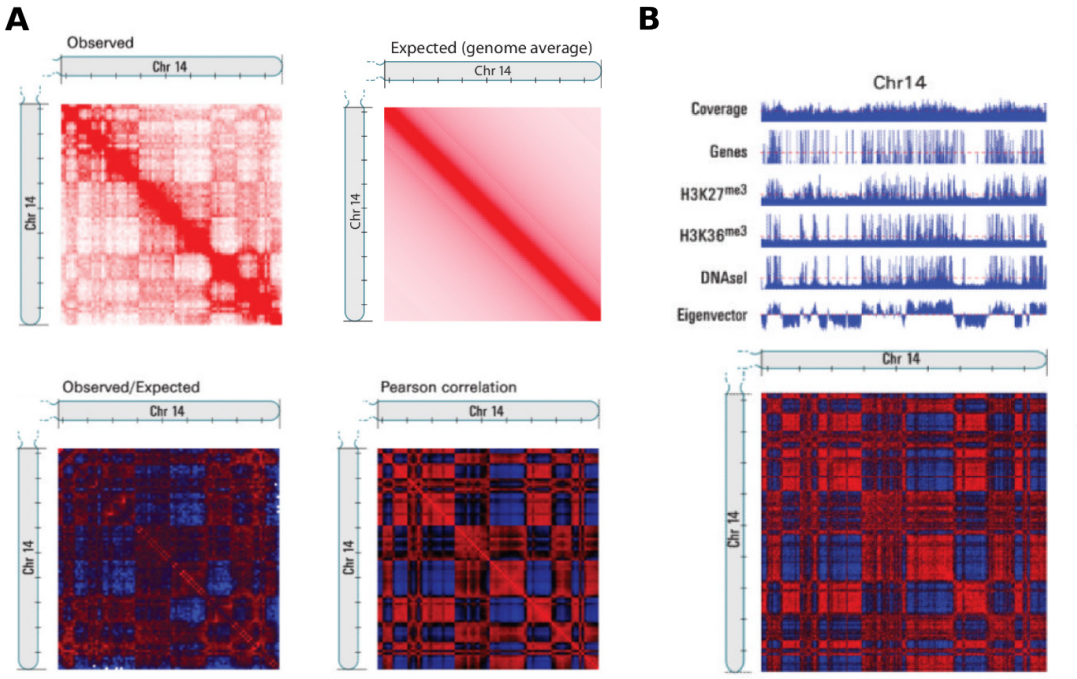


Figure 1.19: Detection of chromosomal compartments using Principle Component Analysis - A. Intra-chromosomal contact maps are first normalized to remove the estimated distance-dependent effect. The Pearson correlation matrix of normalized contact frequency then shows that genomic regions of the same type are highly correlated together. **B.** The PCA is then applied on the correlation matrix. The first principal component (eigenvector) is represented together with histone marks and gene expression. Adapted from Lieberman-Aiden et al. (2009).

The PCA analysis defines the first component as the one which maximize the variance between the genomic bins and therefore which better separate the A/B compartments. The variable loadings of the PCA can be seen as a score with positive or negative values representing the two types of chromosome compartments. However, this score does not define which type of chromosome compartments is active (A) and which one is inactive (B). To do so, it is common to use the gene enrichment under the assumption that active chromatin is usually associated with gene-rich regions and

inactive chromatin with gene-poor regions. In addition, these compartments have been found to be correlated with chromatin state, such as histone marks, replication timing or DNA accessibility (Lieberman-Aiden et al. (2009)).

1.4.4.3 TADs calling

The TADs calling remains a challenging part of the analysis. As the definition of what is exactly a TAD is not yet clear, different methods will give different results (Figure 1.20A). It is therefore very difficult to evaluate and compare these methods, even if there were developed to detect the same type of pattern (Forcato et al. (2017)). Among the available methods, the directionality index (DI, Dixon et al. (2012)), and the insulation score (IS, Crane et al. (2015)) are intuitive methods that were directly inferred from the current TADs definition. Both methods are based on the calculation of contact differences between upstream and downstream regions (Figure 1.20).
For any bin i, A_i is the number of contacts involving i and the upstream region of size L, and B_i is the number of contacts involving i and the downstream region of size L, the DI score is calculated as :

$$DI_i = \frac{B_i - A_i}{|B_i - A_i|}\left(\frac{(A_i - E_i)^2}{E_i} + \frac{(B_i - E_i)^2}{E_i}\right)$$

where E_i is the mean contact frequency between A_i and B_i such as $E_i = \frac{(A_i+B_i)}{2}$. Based on this calculation, the DI is expected to be minimal at the end of TADs, and to be maximal at the beginning of TADs (Figure 1.20C,D). Boundaries are then detected and associated with each other to form a TAD using an Hidden Markov Model (Dixon et al. (2012)).
Following the same idea, the IS is calculated as the average number of contacts that occur across this bin within a defined genomic range. Keeping the notations previously defined, A and B are respectively the upstream/downstream regions of locus i, C is the contact matrix and k is the number of bins in a genomic range of size L, the IS is calculated as :

$$IS_i = \frac{\sum_{a\in A, b\in B} C(a,b)}{k^2}$$

The IS is therefore expected to be minimal at TADs boundaries (Figure 1.20C,E). Local minima are then detected and use as boundaries in order to define TADs regions between consecutive boundaries. IS is usually used as a continuous signal along the

chromosome to visualize the strength of TADs boundaries. Difference in IS can also be used to detect changes in TADs boundaries between conditions (Crane et al. (2015), Figure 1.20E).

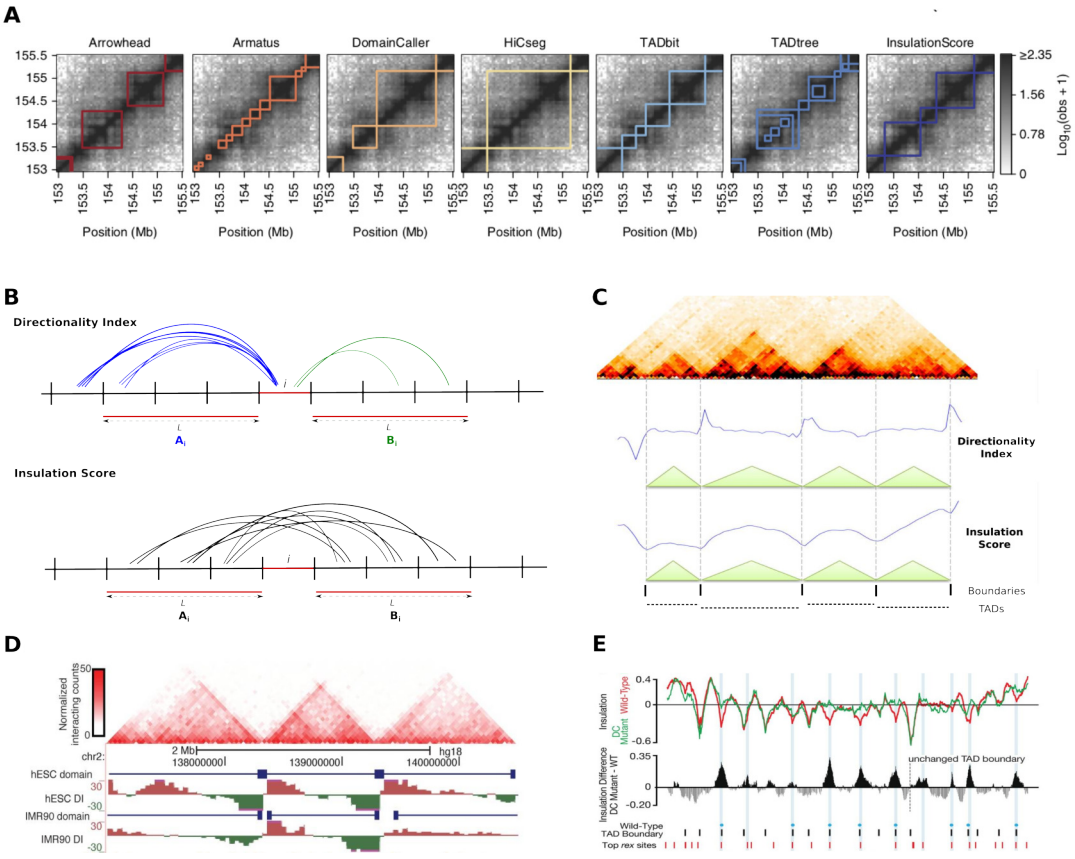


Figure 1.20: TADs calling based on Directionality Index (DI) and Insulation Score (IS). - **A.** Comparison of TADs callers. TADs detected by different software on the GM12878 Hi-C sample (Rao et al. (2014)) (chr1:153,000,000155,500,000) at 40-kb resolution (from Forcato et al. (2017)) **B.** Rational of DI and SI scores calculation for locus i. A and B are upstream/downstream regions of size L. The DI score is based on contacts between i and neighbouring regions whereas the IS is based on the average contacts over the bin i. **C.** The DI score is expected to be maximal (resp. minimal) at the beginning (resp. end) of TADs. The IS score is always minimal at TADs boundaries **D.** Example of TADs calling using DI index applied on human ES and IMR90 cells. **E.** Example of TADs calling using the IS applied on wild type and mutant DC. Adapted from Crane et al. (2015); Dixon et al. (2012); Lajoie et al. (2015)

Although these two methods are intuitive and usually give good results, their parameters have to be manually tuned and are based on prior knowledge and expectation. Therefore, others more sophisticated computational algorithms have been proposed to detect TADs (see Dali and Blanchette (2017) for a review). As already discussed, all these methods give very different results, but the lack of accepted biological definition of TADs, combined with the absence of experimental data that could be used as gold standard, make the comparison difficult. Of note, a few methods are now able to detect hierarchical TADs structure (i.e.. TADs, sub-TADs), consistent at different resolutions. These methods are of growing interest to infer the hierarchical chromatin structure of domains at different scales (Zhan et al. (2017)).

1.4.4.4 Detection of significant interactions

Finally, the last type of pattern which can be extracted from intra-chromosomal Hi-C contact maps is a list of significant mid/long-range interactions. This type of point interactions (or peaks) are likely to represent fine-scale contacts between regulatory elements. Being able to detect such events requires high resolution Hi-C data at nearly kilo-base resolution. All the methods proposed to detect such peaks rely on an efficient modelling of the background signal (see Schmitt et al. (2016) for a review). The background is usually modelled based on the observed distance-dependant effects and systematic bias factors. Some methods such as FitHiC, define a background model in which the expected contact frequency is estimated by fitting the contact frequencies at a given linear genomic distance (Ay and Noble (2015)). Other recent methods, such as HICCUPS use a local background model (Rao et al. (2014)). Then, given the background model, the methods test the significance of individual pairwise contacts. The results is a set of loci that interact more frequently than expected by the background model. All these methods finally propose to correct for multiple testing issues using standard approaches.

However, from a biological point of view, careful evaluation is always advised for interpreting the statistical confidence of such approaches and to distinguish real point interactions from noise and false positives. As an example, these methods may miss some significant contacts occurring between loci which are closed on the linear genome. In this case, although the interaction can be biologically of interest, the contact frequency may not be much higher than expected by the distance, and might therefore

not be statistically significant. On the other hand, a very small increase in contact frequency of loci which are very far away from each other can be statistically significant, but usually occurs in only a small fraction of cells. Additional controls such as biological replicates can help in reducing the number of false interactions detected by these methods ((Lajoie et al., 2015)).

1.5 Thesis project

The use of Hi-C based techniques have exploded these recent years. And so far, these methods have mainly been applied to explore the genome organization of normal cells. My thesis project was conducted with two main objectives : i) developing computational methods and resources, that can be used by the community to easily extract relevant information from Hi-C data, and ii) applying this technique to cancer samples in order to better characterize the relationship between genetic alterations in cancer and the genome organization.

My project was therefore divided in three main parts. First, when I started in 2013, the iterative mapping pipeline (Imakaev et al. (2012)) was the only publicly available tool to process Hi-C data from raw sequencing reads to normalized contact maps. We rapidly identified the limitations of this tool and therefore proposed to develop a new solution, named HiC-Pro, which is presented in Chapter 2. HiC-pro was published in December 2015 in *Genome Biology*, and is currently one of the most popular pipeline for Hi-C data processing, therefore demonstrating the need for such computational solution in this field.

Then, we decided to start exploring Hi-C data from cancer samples. We first proposed a simulation model to estimate the effect of copy number on Hi-C contact maps. Using this model, we identified the challenges related to cancer Hi-C analysis and demonstrated that the current methods usually applied for Hi-C data normalization were not designed to study cancer genome. These observations were validated on public cancer Hi-C dataset from Breast cancer cell line MCF7 and T47D. We therefore proposed new strategies, and applied them to several datasets including our own data on uveal melanoma. This work is presented in Chapter 3 and is publicly available on *BioRXiv*.

Finally, in the Chapter 4, we started to explore the link between chromosome structure and gene expression in tumors. We used an hybrid inducible mouse model that overexpresses the *c-myc* oncogene, leading to the development of liver tumors. As part of a collaborative project, we started to explore the impact of the c-myc expression on the genome organization and the gene expression of the inactive X chromosome using a set of heterogeneous data including expression (RNA-seq), histone modifications (ChIP-seq) and chromosome conformation (Hi-C).

1.6 Appendices

1.6.1 Changes in the organization of the genome during the mammalian cell cycle (Giorgetti et al., Genome Biology, 2013)

Giorgetti et al. Genome Biology 2013, 14:142
http://genomebiology.com/2013/14/12/142

Genome Biology

RESEARCH HIGHLIGHT

Changes in the organization of the genome during the mammalian cell cycle

Luca Giorgetti[1*], Nicolas Servant[2] and Edith Heard[1]

Abstract

By using chromosome conformation capture technology, a recent study has revealed two alternative three-dimensional folding states of the human genome during the cell cycle.

The packaging conundrum at ever-higher levels of scrutiny

Understanding the spatial organization of chromosomes represents a major quest, not only for our comprehension of how a length of 2 m of DNA can be packaged into a nucleus of just a few microns, but also because the physical interactions occurring within and between chromosomes are thought to play an important role in gene regulation, DNA replication and genome stability. Over recent years, the development of chromosome conformation capture (3C)-based techniques, together with the emergence of next-generation sequencing, have changed our view of nuclear organization. High-throughput conformation capture techniques have been developed to study the physical interactions of the chromatin fiber with or within genomic regions spanning from a few megabases (5C technique) to a whole-genome scale (4C and Hi-C techniques). In 2009, Eric Lander, Job Dekker and colleagues [1] described the first human chromosomal architecture at a resolution of 1 Mb. Then, four years later, owing to the rapid evolution of sequencing capabilities, Bing Ren and colleagues [2] published a high-resolution map of the 'chromatin interactome' in human fibroblasts at a resolution of 5 to 10 kb.

The picture that emerges from this blossoming area of research is that metazoan chromosomes are characterized by a nested hierarchy of structural layers (Figure 1, left panel). Within each chromosome territory, compartments originally termed 'A' and 'B', each of several megabases, tend to associate within each single chromosome, reflecting the preferential colocalization of active, gene-rich regions and their segregation from gene-poor, inactive regions [1]. More recently, a further sub-megabase scale of genome partitioning into topologically associating domains (TADs) has been reported [3-5]. Sequences within TADs tend to interact more frequently than with any other surrounding region of the genome. These domains, spanning a few hundred kilobases, are stable across different cell types, suggesting that they are an inherent property of mammalian genomes. Furthermore, within TADs, a network of cell-type-specific physical interactions between potential regulatory sequences has been shown to take place [6]. Consistent with this, a recent study using high-resolution Hi-C showed that enhancer-promoter contacts almost always occur within TADs [2].

Despite the pace at which new details of mammalian chromosome organization are accumulating, many fundamental questions remain unanswered. For example, what is the cell-to-cell variability in the structures that give rise to compartments, TADs and promoter-enhancer interactions? How important is each of those structural layers for the regulation of transcription? What are the molecular mechanisms that determine the appearance of TADs and compartments? When are these layers established during development? Are these different levels of chromosome organization present throughout the cell cycle, or are they established at a particular phase? Finally, what is the structure of mitotic chromosomes? Are the hierarchical layers of folding maintained during mitosis, or do chromosomes acquire a different organization, as suggested by light and electron microscopy? A recent exciting study has shed light on these last questions [7].

Compartments and TADs are erased during mitosis

Although DNA fluorescence *in situ* hybridization (FISH) coupled to super-resolution microscopy has previously

* Correspondence: luca.giorgetti@curie.fr
[1]Institut Curie, CNRS UMR3215, INSERM U934 26 rue d'Ulm, Paris F-75248, France
Full list of author information is available at the end of the article

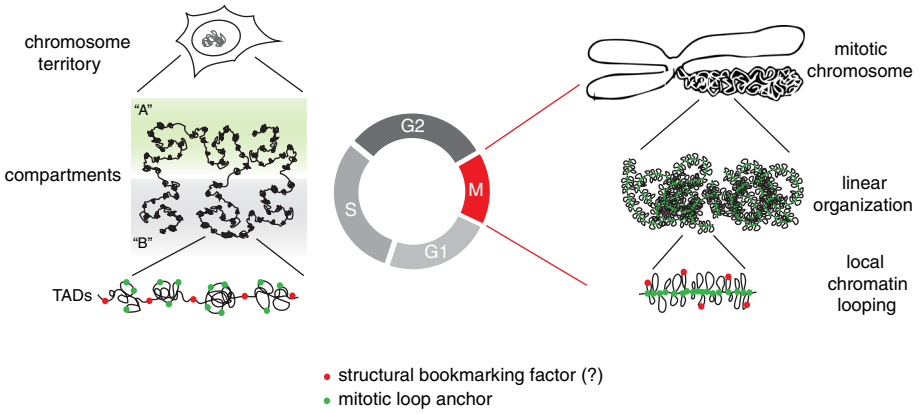


Figure 1 Cell-cycle dependence of chromosome structure. Left: high-resolution chromosome-conformation capture technologies have revealed that each mammalian chromosome territory can be decomposed into multi-megabase compartments where preferential interactions occur between active gene-rich and inactive gene-poor regions of the chromosome. Compartments can be further partitioned at the submegabase scale into topological associating domains (TADs) spanning a few hundred kilobases each. Dekker and colleagues [7] show that compartments and TADs are stable across the cell cycle from G1 to late S phase. Right: in metaphase, the compartment and TAD structures disappear and are replaced by a cylindrically condensed structure of linearly arranged short-range loops [7], which occur on a shorter genomic scale (80 kb on average) than TADs. It is hypothesized that the persistence of bookmarking factors on the mitotic chromosome might facilitate the rapid re-emergence of compartments and TADs upon exit of the cells from mitosis.

been used in murine cells to investigate the presence of TADs within the X-inactivation center during the cell cycle from G1/S phase to mitotic prophase [4], a detailed molecular analysis of chromatin interactions throughout the cell cycle had not, until recently, been reported. Dekker and colleagues [7] now report the results of performing carbon-copy chromosome conformation capture (5C) and Hi-C experiments in synchronized human HeLa cells at different phases of the cell cycle. Using 5C technology spanning the whole of chromosome 21 at a resolution of 250 kb, the authors first studied the long-range chromatin interactions of early G1-, mid-G1-, S- and M-phase cells. Interaction maps of these different stages showed that the mitotic interaction pattern differs dramatically from that at all other stages, thus revealing two distinct chromosome folding states during the cell cycle. Using Hi-C (at a resolution of 40 kb) on mitotic and mid-G1-stage cells [7], the investigators further found that both compartments [1] and TADs [3,4] appear to be absent throughout the entire genome in metaphase, becoming detectable only in early G1 phase and remaining unchanged throughout interphase. These observations on mitotic chromosomes were repeated and validated in another cell line and in primary human fibroblasts. Importantly, the authors showed that, apart from metaphase, compartments and TADs are not restricted to any specific cell cycle phase and they are easily detected in early G1, thus ruling out the possibility that the patterns observed in 5C or Hi-C could be due to a superposition of alternative cell-cycle phase-specific conformations.

The finding that TADs and compartments apparently disappear during metaphase, and reappear in early G1, raises the questions of why they are lost and also what mechanism ensures their prompt re-establishment upon exit from mitosis. It is known that certain transcription factors and histone modifications or associated factors can act as 'bookmarking' factors on mitotic chromosomes [8] by remaining bound to the condensed chromatin polymer in order to ensure the propagation of transcriptional states to daughter cells. It is therefore tempting to speculate that structural bookmarking factors that propagate organizational information to daughter cells might similarly exist. Although the molecular bases of compartments and TADs remain elusive, accumulating evidence suggests that long-range interactions between genomic sites bound by the proteins CTCF and cohesin contribute to the organization of chromatin architecture at the TAD scale, either by associating with TAD boundaries [3] or by supervising a network of long-range interactions inside single TADs that might stabilize the structure within TADs [6]. Interestingly, both CTCF and cohesin have been shown to associate with mitotic chromosomes [8], suggesting that mutual interactions between CTCF and/or cohesin binding sites might readily occur upon exit from mitosis, thus enabling TADs to be established immediately at the beginning of G1 phase. Other factors that might remain bound during mitosis to the positions of TAD or

compartment boundaries include Polycomb group proteins, which have been shown to significantly overlap with TAD boundaries [5] on mitotic chromosomes in *Drosophila* [9]

A polymer view of a mitotic chromosome

The apparent loss of partitioning into TADs and compartments observed in mitotic chromosomes by Dekker and colleagues [7] suggests that they contain unusually few structural details compared with those of their interphase counterparts. However, the authors cleverly exploited the quantitative information present in the Hi-C mitotic chromosome data to build a structural model. The foundation of their approach is that Hi-C or 5C interaction maps represent the frequency at which every pair of genomic loci along the chromatin polymer encounter each other, averaged over millions of cells. It is therefore possible to use models from polymer physics to simulate virtual Hi-C (or 5C) data and compare them with data from experiments. Using such an approach, the thermodynamic ensemble of configurations of a model chromosome can be generated by computer simulations, and a contact map is retrieved by averaging the contacts of each pair of loci over all simulated polymer conformations, allowing a straightforward comparison with experimental data.

In the case of mitotic chromosome data, the Hi-C or 5C contact maps are homogeneous, with no sign of the regular patterning due to TADs and compartments that characterize the interphase contact maps [7]. Dekker and colleagues also noted that the contact probability between two loci gradually decreases with increasing genomic distance (albeit significantly more slowly than on interphase chromosomes), and then suddenly falls off to zero at approximately 10 Mb. This unusual behavior cannot be explained by any simple polymer models. Building on the well-established experimental evidence describing the quantitative properties of mitotic chromosomes (such as cylindrical symmetry, linear organization of chromatids and chromatin packing density), the authors thus set out to build various alternative models for the chromatin organization within mitotic chromosomes, simulated all of them and tested each model's predictions. In order to reproduce the unexpected fall-off in contact probabilities for loci that are separated by >10 Mb, they had to impose that the chromatin fiber is linearly organized, so that genomic loci belonging to distal parts of the chromosome (that is, separated by >10 Mb) cannot be brought into close spatial proximity. Furthermore, they had to impose that the chromatin fiber is arranged in an array of consecutive loops, each spanning approximately 80 kb, in order for their simulations to account correctly for the gentle decrease in contact frequencies in the genomic length range between 0 and 10 Mb. Thus, amongst all possible models tested, it was this combination of linear organization and consecutive looping that best accounted for the behavior of the contact probability of mitotic chromosomes over all genomic length scales. Interestingly, the authors further discovered that they could simultaneously reproduce the scaling of contact probabilities and the homogeneity of the mitotic contact maps only when they allowed the size of the loops to be different (and stochastic) in each of the polymer configurations.

Future perspectives

The study by Dekker and colleagues represents a powerful combination of biochemical investigation and polymer modeling [7]. The prowess of such an approach is that it can provide insights into structure that would not have been obtained without modeling. The picture that emerges, thanks to their study, is that of a mitotic chromosome composed of consecutive loops of chromatin, the size of which is on average 80 kb, although this varies from cell to cell, which are arranged in a linear fashion eventually resulting in an effective cylindrical chromosomal volume (Figure 1, right panel). This picture is both qualitatively and quantitatively in agreement with previously proposed 'loops-on-a-scaffold' models of mitotic chromosomes based on microscopy and biochemical assays.

Although the molecular mechanisms that could mediate this pervasive local looping remain elusive, the model predicts that cell-to-cell variability of looping events has a sizeable effect in organizing the structure of mitotic chromatin. Indeed, the potential importance of fluctuations in interphase chromosome structure was also recently revealed by an adaptation of the Hi-C technique to single cells [10]. In the future, this type of single-cell approach, together with physical modeling and quantitative analysis, could allow the exploration of chromosome structure in situations where cell-to-cell variability might be crucial, such as during early development.

Abbreviations

3C: Chromosome conformation capture; 4C: Circularized chromosome conformation capture; 5C: Carbon-copy chromosome conformation capture; Hi-C: High-throughput conformation capture; TAD: Topologically associating domain.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

LG was supported by an EMBO Fellowship (ALTF 1559–2011); work in the lab of EH is supported by the "Ligue Nationale contre le cancer", the EpiGeneSys FP7 257082 Network of Excellence, ERC Advanced Investigator award 250367 and EU FP7 MODHEP EU grant no. 259743.

Author details
[1]Institut Curie, CNRS UMR3215, INSERM U934 26 rue d'Ulm, Paris F-75248, France. [2]Institut Curie, INSERM U900, Bioinformatics and Computational Systems Biology of Cancer, 26 rue d'Ulm, Paris F-75248, France.

Published: 24 December 2013

References

1. Lieberman-Aiden E, Berkum NL V, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326:**289–293.
2. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B: **A high-resolution map of the three-dimensional chromatin interactome in human cells.** *Nature* 2013, **503:**290–294.
3. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485:**376–380.
4. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen N, Dekker J, Heard E: **Spatial partitioning of the regulatory landscape of the X-inactivation centre.** *Nature* 2012, **485:**381–385.
5. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the *Drosophila* genome.** *Cell* 2012, **148:**458–472.
6. Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, Ong C-T, Hookway TA, Guo C, Sun Y, Bland MJ, Wagstaff W, Dalton S, McDevitt TC, Sen R, Dekker J, Taylor J, Corces VG: **Architectural protein subclasses shape 3D organization of genomes during lineage commitment.** *Cell* 2013, **153:**1281–1295.
7. Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J: **Organization of the mitotic chromosome.** *Science* 2013, **342:**948–953.
8. Kadauke S, Blobel GA: **Mitotic bookmarking by transcription factors.** *Epigenetics Chromatin* 2013, **6:**6.
9. Follmer NE, Wani AH, Francis NJ: **A Polycomb group protein is retained at specific sites on chromatin in mitosis.** *PLOS Genet* 2012, **8:**e1003135.
10. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P: **Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.** *Nature* 2013, **502:**59–64.

doi:10.1186/gb4147
Cite this article as: Giorgetti *et al.*: Changes in the organization of the genome during the mammalian cell cycle. *Genome Biology* 2013 **14**:142.

1.6.2 HiTC: exploration of high-throughput C experiments (Servant et al. Bioinformatics, 2012)

BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 21 2012, pages 2843–2844
doi:10.1093/bioinformatics/bts521

Sequence analysis

Advance Access publication August 24, 2012

HiTC: exploration of high-throughput 'C' experiments

Nicolas Servant[1,2,3,*], Bryan R. Lajoie[4], Elphège P. Nora[1,5,6], Luca Giorgetti[1,5,6], Chong-Jian Chen[1,2,3,5,6], Edith Heard[1,5,6], Job Dekker[4] and Emmanuel Barillot[1,2,3]

[1]Institut Curie, F-75248 Paris, France, [2]INSERM, U900, F-75248 Paris, France, [3]Ecole des Mines ParisTech, F-77300 Fontainebleau, France, [4]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA, [5]CNRS UMR3215, F-75248 Paris, France and [6]INSERM U934, F-75248 Paris, France

Associate Editor: Alex Bateman

ABSTRACT

Summary: The R/Bioconductor package *HiTC* facilitates the exploration of high-throughput 3C-based data. It allows users to import and export 'C' data, to transform, normalize, annotate and visualize interaction maps. The package operates within the Bioconductor framework and thus offers new opportunities for future development in this field.

Availability and implementation: The R package *HiTC* is available from the Bioconductor website. A detailed vignette provides additional documentation and help for using the package.

Contact: nicolas.servant@curie.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 5, 2012; revised on August 10, 2012; accepted on August 14, 2012

1 INTRODUCTION

The three-dimensional organization of chromosomes and the physical interactions occurring along and between them play an important role in the regulation of gene activity. Over the past 10 years, the development of Chromosome Conformation Capture (3C)-based techniques has changed our view of nuclear organization (de Wit and de Laat, 2012). With the emergence of next-generation sequencing, high-throughput conformation capture techniques, such as Circular 3C (4C; Simonis *et al.*, 2006), 3C Carbon-Copy (5C; Dostie *et al.*, 2006) or more recently Hi-C (Lieberman-Aiden *et al.*, 2009), have been developed to study the physical interactions between many loci in parallel.

While the use of high-throughput 'C' techniques is expected to increase in coming years (Dixon *et al.*, 2012; Nora *et al.*, 2012), bioinformatic methods and software to analyze such data are still lacking. Here, we present the R/Bioconductor package *HiTC* that enables users to visualize and explore high-throughput 'C' data. One advantage of the *HiTC* package is that it operates within the open source Bioconductor framework (Gentleman *et al.*, 2004) and thus offers new opportunities for future developments in this field. The *HiTC* package is aimed at biologists interested in investigating their data and at biostatisticians involved in the development of new statistical methods which can be applied to C data.

*To whom correspondence should be addressed.

2 AVAILABLE FUNCTIONALITIES

The *HiTC* package provides a variety of functionalities to handle high-throughput C data and is especially suited for visualization and basic transformations of 5C and Hi-C data (Supplementary Fig. S1). Here, we present some of the main functionalities of the package.

2.1 Importing C data

Two distinct datasets are included in the package. The first one is a 5C dataset (Nora *et al.*, 2012), corresponding to the X inactivation center obtained in Mouse ES cells (GSE35721) and the second is the Hi-C data from chromosome 14 (GSE18199) published by Lieberman-Aiden *et al.* (2009). Both Hi-C and 5C datasets can be imported using a defined csv format. In addition, the *HiTC* package is fully compatible with data from the my5C web tool (Lajoie *et al.*, 2009).

2.2 Quality control

Because of the polymer nature of chromatin, a Hi-C or 5C experiment is expected to be dominated by signal from neighboring restriction fragments in *cis*. Quality control provides simple descriptive statistics and graphical outputs to check the prevalence of *cis*- and *trans*-chromosomal interactions, and to assess whether the expected higher frequency between sites located near each other in the linear genome is verified.

2.3 Visualization

An interaction map is a two-dimensional heatmap representation of the matrix of 5C or Hi-C counts, whose entries correspond to the number of times two restriction fragments in a given genomic region have been ligated in 3C and sequenced as a pair. The *HiTC* package proposes a list of options to define the appropriate data visualization, such as contrast, color or counts trimming. Two different views are provided: a square heatmap view and a triangular view (Fig. 1). The latter is particularly useful for aligning interaction maps and genomic and epigenomic features.

2.4 Interaction map transformation

Depending on the experimental resolution and/or the desired genomic scale to be visualized, each pixel of an interaction map can correspond to a single restriction fragment, several restriction fragments or genomic intervals of any given size (and

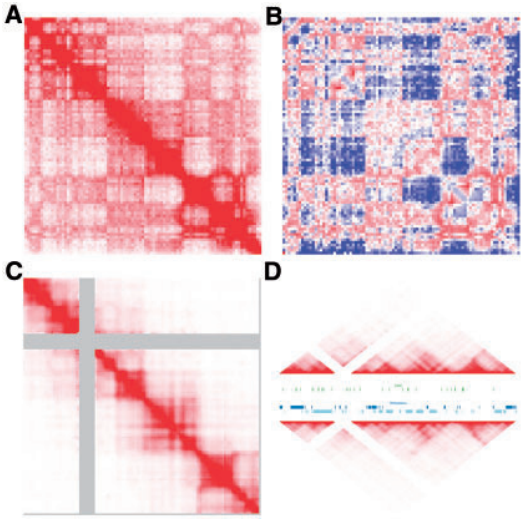


Fig. 1. Visualization of interaction maps. (**A**) Binned interaction map (1 Mb) of Hi-C data (chr14, GSE18199). (**B**) Binned and normalized by expected counts interaction map (1 Mb) of the Hi-C data. (**C**) Heatmap view of the ESC E14 5C interaction map (GSE35721). (**D**) Comparison of the ESC E14 and PGK 5C interaction maps. The genes and CTCF regions from both strands are displayed in blue and green, respectively

therefore various numbers of restriction fragments). For example, 5C allows interaction frequencies to be assessed for each pair of restriction fragments present in the pool of 5C oligonucleotides. The Hi-C protocol, on the contrary, does not necessarily yield counts for every single pair of restriction fragments, especially when analyzing large genomes. Hi-C results are thus typically displayed for genomic bins of an arbitrary size. The *HiTC* package provides a binning function to address the interaction map transformation. For instance, *HiTC* enables the same 5C dataset to be displayed either at the restriction-fragment resolution or after binning in 100 Kb or 1 Mb bins, and these bins can be chosen to partially overlap or not.

2.5 Interaction map normalization

As mentioned earlier, at small genomic distances, pairs of restriction fragments that are close to each other in the linear genome will give higher signal than fragments that are further apart. This leads to most counts mapping to the heatmap diagonal. When considering any given pair of restriction fragments, it can therefore be informative to assess whether the observed counts are above what would be expected given their genomic distance. The *HiTC* package includes a basic normalization function that estimates the interaction counts one would expect if the signal was only dependent on the genomic distance between the interacting loci (Fig. 1B). This calculation is based on Lowess averaging of the observed interaction counts as applied by Bau *et al.* (2011).

3 CONCLUSION

Although we are still far from understanding the exact relationship between chromosome conformation and gene or genome regulations, breakthrough technologies are now available for the systematic and detailed analysis of nuclear organization. The analysis of chromosome conformation capture datasets is quite complex and requires the development of computational tools, including dedicated statistical methods and visualization software, such as the one we propose here. We wish to emphasize that appropriate interpretation of high-throughput C data can require pre-processing of the data, in order to eliminate systematic biases that can be introduced by the experimental protocol or that can arise from the intrinsic properties of the genome, such as a non-homogenous distribution of restriction sites (Yaffe and Tanay, 2011; Zhang and McCord, 2012). The R/BioConductor package *HiTC* proposes a powerful and extensible framework for visualizing and exploring high-throughput C data. It is able to handle both 5C and Hi-C data and offers new functionalities such as standard import, data transformation and integrative visualization methods. While pre-processing and visualization tools started to emerge, other methods such as bias correction, samples comparison or data integration can be further investigated. In this way, the HiTC package provides a flexible basis for further developments by the community.

ACKNOWLEDGEMENTS

The authors thank Joern Toedling, Pierre Gestraud and Marc Carlson for useful discussions and for their help in creating the R package.

Funding: The French program 'investissement d'avenir' action bioinformatique (ABS4NGS project), the Ministère de la Recherche et de l'Enseignement Supérieur and the ARC (to E.P.N.), EU EpiGeneSys FP7 Network of Excellence no. 257082, the Fondation pour la Recherche Medicale, ANR, ERC Advanced Investigator award no. 250367 and EU FP7 SYBOSS (242129) and MODHEP (259743) (to E.H.). NIH (R01 HG003143) and W.M. Keck Foundation Distinguished Young Scholar Award (J.D. and B.R.L.).

Conflict of Interest: none declared.

REFERENCES

Bau,D. *et al.* (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, **18**, 107–114.

de Wit,E. and de Laat,W. (2012) A decade of 3c technologies: insights into nuclear organization. *Genes Dev.*, **26**, 11–24.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Dostie,J. *et al.* (2006) Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Lajoie,B.R. *et al.* (2009) My5c: web tools for chromosome conformation capture studies. *Nat. Methods*, **6**, 690–691.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Nora,E.P. *et al.* (2012) Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, **485**, 381–385.

Simonis,M. *et al.* (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat. Genet.*, **38**, 1348–1354.

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.

Zhang,Y. and McCord,R.P. (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, **148**, 908–921.

2

New strategy for Hi-C data processing

2.1 Rational of HiC-Pro development

Since 2009, and the publication of the first Hi-C paper (Lieberman-Aiden et al. (2009)), the use of the Hi-C technique has exploded, changing our vision of the genome organization. More importantly, the recent discoveries arising from its use have convinced that the genome architecture is a key factor of the cell regulation. Having a view of the genome organization is now essential to complete the epigenetic portrait of a cell population. And as the RNA-seq or ChIP-seq are now standard methods to explore genes expression or histone modifications, the Hi-C tends to become the method of reference to explore how the genome is organized. Therefore, it is essential to develop accurate computational methods, that can be applied to any Hi-C dataset, to process the raw data and easily extract the relevant biological information.

In 2012, Imakaev et al. proposed a first python library, named 'hiclib', able to process Hi-C data following the steps previously presented (see section 1.4.1). The hiclib library is organized in functions that can be used to align the reads, to filter out non informative read pairs, and to generate raw and normalized contact maps. Although hiclib provides good results, the first version was difficult to install and to use. And by definition, a python library requires computational skills, and did not directly provide a ready-to-use pipeline.

We therefore developed a new pipeline for Hi-C data processing, named HiC-Pro, based on the following guidelines :

- Simplicity. Being able to process a dataset in a simple command line.
- Modularity. Being able to run only part of the analysis.
- Flexibility. Having a simple pipeline architecture, facilitating future updates and improvements.
- Scalability and efficiency. Being able to process Hi-C at the terabyte scale, in a reasonable computational time.

HiC-Pro proposes a new mapping strategy which is much faster than the iterative mapping approach. The most time-consuming steps of the processing were optimized using C++ and python programming language. We also introduced a new text format for Hi-C contact maps, allowing to dramatically reduce the size of output files for high resolution Hi-C data. HiC-Pro was designed in collaboration with biologists in order to provide useful quality controls that can help to validate the different steps of the protocol. Finally, HiC-Pro is the only pipeline able to use the genotype information of a sample to directly build allele-specific contact maps.

For maintenance and sharing with the community, HiC-Pro is available on github. The software and all parameters are documented. A forum is also available where people can ask questions, and report needs or bugs. Since its publication, HiC-Pro has been used or cited in more than 40 papers (source from google scholar). It is also regularly cited among the most commonly used tool for Hi-C data processing (Davies et al. (2017)).

2.2 HiC-Pro: an optimized and flexible pipeline for Hi-C data processing

Servant et al. Genome Biology (2015) 16:259
DOI 10.1186/s13059-015-0831-x

Genome Biology

SOFTWARE Open Access

HiC-Pro: an optimized and flexible pipeline for Hi-C data processing

Nicolas Servant[1,2,3*], Nelle Varoquaux[1,2,3], Bryan R. Lajoie[4], Eric Viara[5], Chong-Jian Chen[1,2,3,6,7,8], Jean-Philippe Vert[1,2,3], Edith Heard[1,6,7], Job Dekker[9] and Emmanuel Barillot[1,2,3]

Abstract

HiC-Pro is an optimized and flexible pipeline for processing Hi-C data from raw reads to normalized contact maps. HiC-Pro maps reads, detects valid ligation products, performs quality controls and generates intra- and inter-chromosomal contact maps. It includes a fast implementation of the iterative correction method and is based on a memory-efficient data format for Hi-C contact maps. In addition, HiC-Pro can use phased genotype data to build allele-specific contact maps. We applied HiC-Pro to different Hi-C datasets, demonstrating its ability to easily process large data in a reasonable time. Source code and documentation are available at http://github.com/nservant/HiC-Pro.

Keywords: Chromosome conformation, Hi-C, Bioinformatics pipeline, Normalization

Introduction

High-throughput chromosome conformation capture methods are now widely used to map chromatin interactions within regions of interest and across the genome. The use of Hi-C has notably changed our vision of genome organization and its impact on chromatin and gene regulation [1, 2]. The Hi-C technique involves sequencing pairs of interacting DNA fragments, where each mate is associated with one interacting locus. Briefly, cells are crossed-linked, DNA is fragmented using a restriction enzyme [3] or a nuclease [4], and interacting fragments are ligated together. After paired-end sequencing, each pair of reads can be associated to one DNA interaction.

In recent years, the Hi-C technique has demonstrated that the genome is partitioned into domains of different scale and compaction level. The first Hi-C application has described that the genome is partitioned into distinct compartments of open and closed chromatin [3]. Higher throughput and resolution have then suggested the presence of megabase-long and evolutionarily conserved smaller domains. These topologically associating domains are characterized by a high frequency of intra-domain chromatin interactions but infrequent inter-domain chromatin interactions [5, 6]. More recently, very large data sets with deeper sequencing have been used to increase the Hi-C resolution in order to detect loops across the entire genome [7, 8].

As with any genome-wide sequencing data, Hi-C usually requires several millions to billions of paired-end sequencing reads, depending on genome size and on the desired resolution. Managing these data thus requires optimized bioinformatics workflows able to extract the contact frequencies in reasonable computational time and with reasonable resource and storage requirements. The overall strategy to process Hi-C data is converging among recent studies [9], but there remains a lack of stable, flexible and efficient bioinformatics workflows to process such data. Solutions such as the HOMER [10], HICUP [11], HiC-inspector [12], HiCdat [13] and HiC-box [14] pipelines are already available for Hi-C data processing. HOMER offers several functions to analyze Hi-C data but does not perform the mapping of reads nor the correction of systematic biases. HiCdat, HiC-inspector and HiCbox do not allow chimeric reads to be rescued during the mapping of reads. HICUP provides a complete pipeline until the detection of valid interaction products. Using HICUP together with the SNPsplit program [15] allows the extraction of allele-specific interaction products whereas all other solutions do not allow allele-specific analysis. The HiCdat and HiCbox packages offer a means of correcting contact maps for systematic

* Correspondence: nicolas.servant@curie.fr
[1]Institut Curie, Paris, France
[2]INSERM, U900, Paris, France
Full list of author information is available at the end of the article

biases. Finally, none of these software were designed to process very large amounts of data in a parallel mode. The hiclib package is currently the most commonly used solution for Hi-C data processing. However, hiclib is a Python library that requires programming skills, such as knowledge of Python and advanced Linux command line, and cannot be used in a single command-line manner. In addition, parallelization is not straightforward and it has limitations with regard to the analysis and normalization of very high-resolution data (Table 1).

Here, we present HiC-Pro, an easy-to-use and complete pipeline to process Hi-C data from raw sequencing reads to normalized contact maps. HiC-Pro allows the processing of data from Hi-C protocols based on restriction enzyme or nuclease digestion such as DNase Hi-C [4] or Micro-C [16]. When phased genotypes are available, HiC-Pro is able to distinguish allele-specific interactions and to build both maternal and paternal contact maps. It is optimized and offers a parallel mode for very high-resolution data as well as a fast implementation of the iterative correction method [17].

Results

HiC-Pro results and performance

We processed Hi-C data from two public datasets: IMR90 human cell lines from Dixon et al. [6] (IMR90) and from Rao et al. [7] (IMR90_CCL186). The latter is currently one of the biggest datasets available, used to generate up to 5-kb contact maps. For each dataset, we ran HiC-Pro and generated normalized contact maps at 20 kb, 40 kb, 150 kb, 500 kb and 1 Mb resolution. Normalized contact maps at 5 kb were only generated for the IMR90_CCL186 dataset. The datasets were either used in their original form or split into chunks containing 10 or 20 million read pairs.

Using HiC-Pro, the processing of the Dixon's dataset (397.2 million read pairs split into 84 read chunks) was completed in 2 hours using 168 CPUs (Table 2). Each chunk was mapped on the human genome using four CPUs (two for each mate) and 7 GB of RAM Processing the 84 chunks in parallel allows extraction of the list of valid interactions in less than 30 minutes. All chunks were then merged to generate and normalize the genome-wide contact map.

In order to compare our results with the hiclib library, we ran HiC-Pro on the same dataset, and without initial read splitting, using eight CPUs. HiC-Pro performed the complete analysis in less than 15 hours compared with 28 hours for the hiclib pipeline. The main difference in speed is explained by our two-step mapping strategy compared with the iterative mapping strategy of hiclib, which aligned the 35 base pair (bp) reads in four steps. Optimization of the binning process and implementation of the normalization algorithm led to a three-fold decrease in time to generate and normalize the genome-wide contact map.

The IMR90 sample from the Rao dataset (1.5 billion read pairs split into 160 read chunks) was processed in parallel using 320 CPUs to generate up to 5-kb contact maps in 12 hours, demonstrating the ability of HiC-Pro to analyze very large amounts of data in a reasonable time. At a 5-kb resolution, we observe the presence of chromatin loops as described by Rao et al. [7] (Figure S1 in Additional file 1). The merged list of valid interactions was generated in less than 7.5 hours. Normalization of the genome-wide contact map at 1 Mb, 500 kb, 150 kb, 40 kb, 20 kb and 5 kb was performed in less than 4 hours. Details about the results and the implementation of the different solutions are available in Additional file 1.

Finally, we compared the Hi-C processing results of hiclib and HiC-Pro on the IMR90 dataset. Although the processing and filtering steps of the two pipelines are not exactly the same, we observed a good concordance in the results (Fig. 1). Using default parameters, HiC-Pro is less stringent than hiclib and used more valid interactions to

Table 1 Comparing solutions for Hi-C data processing

	Mapping	Detection of valid interactions	Binning	Correction of systematic noise	Parallel implementation	Allele-specific analysis
HOMER		x	x			
HICUP	x	x				x
HiC-inspector	x[a]	x	x			
HiC-Box	x[a]	x	x	x		
HiCdat	x[a]	x	x	x		
Hiclib	x	x	x	x		
HiC-Pro	x	x	x	x	x	x

HOMER [10] offers several programs to analysis Hi-C data from aligned reads. [a]HiC-inpector [12], HiCdat [13] and HiC-Box [14] do not allow chimeric reads to be rescued during the mapping. HICUP [11] provides a complete pipeline until the detection of valid interaction products. It can be used together with the SNPsplit software [15] to extract allele-specific mapped reads. The hiclib Python library [17] can be applied for all analysis steps but requires good programming skills and cannot be used in a single command-line manner. None of these software enable very large amounts of data to be processed easily in a parallel mode. Note that HOMER, hiclib and HiCdat also offer additional functions for downstream analysis. In the case of HiC-Pro, the downstream analysis is supported by the HiTC BioConductor package [28]

Table 2 HiC-Pro performance and comparison with hiclib

Dataset	IMR90	IMR90	IMR90	IMR90_CCL186
Number of reads	397,200,000	397,200,000	397,200,000	1,535,222,082
Pipeline	hiclib	HiC-Pro	HiC-Pro parallel	HiC-Pro parallel
Number of input files	10	10	84	160
Number of jobs	1	1	42	80
Number of CPUs per job	8	8	4	4
Maximum memory	10	7	7	24
Wall time	28:24	14:32	02:15	11:49
Mapping	22:03	10:31	00:21	05:56
Filtering	00:30	03:10	00:05	00:36
Merge		00:20	00:18	00:50
Contacts maps	01:45	00:15	00:15	00:42
Normalization	04:06	01:16	01:16	03:49

HiC-Pro was run on the IMR90 Hi-C dataset from Dixon et al. and Rao et al. in order to generate contact maps at resolutions of 20 kb, 40 kb, 150 kb, 500 kb and 1 Mb. Contact maps at 5 kb were also generated for the IMR90_CCL186 dataset. The CPU time for each step of the pipeline is reported and compared with the hiclib Python library. The reported results include time of writing contact maps in text format. Times are minutes:seconds

build the contact maps. The two sets of normalized contact maps generated at different resolutions are highly similar (Fig. 1c). We further explored the similarity between the maps generated by the two pipelines by computing the Spearman correlation of the normalized intra-chromosomal maps. The average correlation coefficient across all chromosomes at different resolutions was 0.83 (0.65–0.95). Finally, since the inter-chromosomal data are usually very sparse, we summarized the inter-chromosomal signal using two one-dimensional coverage vectors of rows and columns [18, 19]. The average Spearman correlation coefficient of all coverage vectors between hiclib and HiC-Pro inter-chromosomal contact maps was 0.75 (0.46–0.98).

Implementation of the iterative correction algorithm

We provide an implementation of the iterative correction procedure which emphasizes ease of use, performance, memory-efficiency and maintainability. We obtain higher or similar performance on a single core compared with the original ICE implementation from the hiclib library (Table 2) and from the HiCorrector package [20] (Table 3).

The HiCorrector package provides a parallel version of the iterative correction for dense matrices. We therefore compared the performance of HiCorrector with the HiC-Pro normalization at different Hi-C resolutions (Table 3). All algorithms were terminated after 20 iterations for the purpose of performance comparison, as each iteration requires nearly the same running time. Choosing dense or sparse matrix-based implementation is dependent on the Hi-C data resolution and on the depth of coverage. Although our implementation can be run in either sparse or dense mode, the available data published at resolutions of 5–40 kb are currently characterized by a high degree of sparsity. At each level of Hi-C contact map resolution, we compared our dense or sparse implementation with the parallel and/or sequential version of HiCorrector. Our results demonstrate that using a compressed sparse row matrix structure is more efficient on high resolution contact maps (<40 kb) than using parallel computing on dense matrices. As expected for low resolution contact maps (1 Mb, 500 kb), using a dense matrix implementation is more efficient in time, although the gain, in practice, remains negligible.

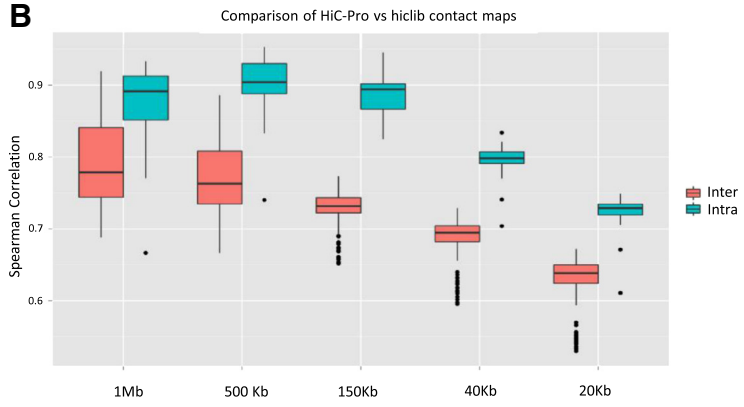
The code for the normalization is available as a standalone package (https://github.com/hiclib/iced) as well as being included in HiC-Pro. Our implementation based on sparse row matrices is able to normalize a 20-kb human genome map in less than 30 minutes with 5 GB of RAM (Table 3). Genome-wide normalization at 5 kb can be achieved in less than 2.5 hours with 24 GB of RAM. Thus, compared to existing solutions, our implementation substantially speeds up and facilitates the normalization of Hi-C data prior to downstream analysis.

Allele-specific contact maps

We used HiC-Pro to generate allele-specific contact maps for the human GM12878 cell line. Differences in paternal and maternal X chromosome organization were recently described, with the presence of mega-domains on the inactive X chromosome, which are not seen in the active X chromosome [7, 21, 22]. We used HiC-Pro to generate the maternal and paternal chromosome X contact maps of the GM12878 cell line using the Hi-C dataset published by Selvaraj et al. [23]. Phasing data were gathered from the Illumina Platinum Genomes Project [24]. Only good quality heterozygous phased

A

	hiclib	HiC-Pro
Total read pairs	397 194 480	397 194 480
Uniquely aligned read pairs	231 047 307 (58.17%)	257 502 619 (64.83%)
Self-Circle	1 569 902 (0.68%)	1 793 553 (0.69%)
Dangling-end	79 701 493 (34.49%)	94 024 488 (36.51%)
Valid interactions	141 686 863 (61.32%)	159 737 835 (62.03%)
Filtered valid interactions	107 977 460 (46.73%)	133 761 282 (51.9%)
Intra-chromosomal contacts	66 619 145 (61.69%)	85 694 952 (64.06%)
Inter-chromosomal contacts	41 358 315 (38.30%)	48 066 330 (35.93%)



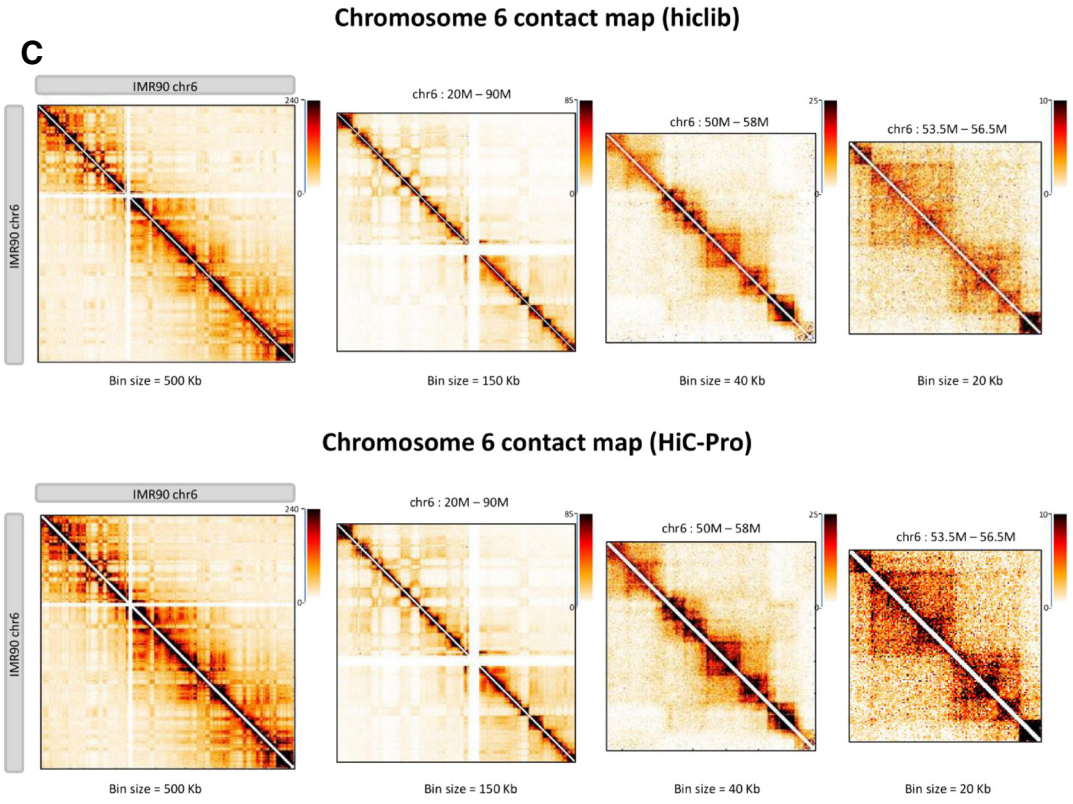


Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Comparison of HiC-Pro and hiclib processing. **a** Both pipelines generate concordant results across processing steps. The fraction of uniquely aligned read pairs is calculated on the total number of initial reads. Self-circle and dangling-end fractions are calculated on the total number of aligned read pairs. Intra- and inter-chromosomal contacts are calculated as a fraction of filtered valid interactions. **b** Boxplots of the Spearman correlation coefficients of intra- and inter-chromosomal maps generated at different resolutions by both pipelines. **c** Chromosome 6 contact maps generated by hiclib (*top*) and HiC-Pro (*bottom*) at different resolutions. The chromatin interaction data generated by the two pipelines are highly similar

single-nucleotide polymorphisms (SNPs) were selected. The final list contained 2,239,492 SNPs. We then masked the human genome hg19 by replacing the SNP position by an 'N' using the BEDTools utilities [25] and generated the new bowtie2 indexes. In practice, the allele-specific analysis can be easily performed by simply specifying to HiC-Pro the list of SNPs and the N-masked indexes for read alignment through the configuration file.

Among the initial 826 million read pairs, 61 % were classified as valid interactions by HiC-Pro. Around 6 % of valid interactions were then assigned to either the paternal or maternal genome and used to construct the haploid maps. As expected, the inactive X chromosome map is partitioned into two mega-domains (Fig. 2). The boundary between the two mega-domains lies near the DXZ4 micro-satellite.

Materials and methods

HiC-Pro workflow

HiC-Pro is organized into four distinct modules following the main steps of Hi-C data analysis: (i) read alignment, (ii) detection and filtering of valid interaction products, (iii) binning and (iv) contact map normalization (Fig. 3).

Mapping

Read pairs are first independently aligned on the reference genome to avoid any constraint on the proximity between the two reads. Most read pairs are expected to be uniquely aligned on the reference genome. A few percent, however, are likely to be chimeric reads, meaning that at least one read spans the ligation junction and therefore both interacting loci. As an alternative to the iterative mapping strategy proposed by Imakaev et al. [17], we propose a two-step approach to rescue and align those reads (Fig. 4a). Reads are first aligned on the reference genome using the bowtie2 end-to-end algorithm [26]. At this point, unmapped reads are mainly composed of chimeric fragments spanning the ligation junction. According to the Hi-C protocol and the fill-in strategy, HiC-Pro is then able to detect the ligation site using an exact matching procedure and to align back on the genome the 5′ fraction of the read. Both mapping steps are then merged in a single alignment file. Low mapping quality reads, multiple hits and singletons can be discarded.

Detection of valid interactions

Each aligned read can be assigned to one restriction fragment according to the reference genome and the selected restriction enzyme. Both reads are expected to map near a restriction site, and with a distance within the range of molecule size distribution after shearing. Fragments with a size outside the expected range can be discarded if specified but are usually the result of random breaks or star activity of the enzyme, and can therefore be included in downstream analysis [17]. Read pairs from invalid ligation products, such as dangling end and self-circle ligation, are discarded (Fig. 4b). Only valid pairs involving two different restriction fragments are used to build the contact maps. Duplicated valid pairs due to PCR artifacts can also be filtered out. Each read is finally tagged in a BAM file according to its mapping and fragment properties (Figure S2 in Additional file 1). In the context of Hi-C methods which are not based on restriction enzyme digestion, no filtering of restriction fragments is applied. The uniquely mapped read pairs are directly used to build the contact maps. However, one way to filter out artifacts such as self-ligation is to discard intra-chromosomal pairs below a given distance threshold [4]. HiC-Pro therefore allows these short range contacts to be filtered out.

Table 3 Performance of iterative correction on IMR90 data

	HiC-Pro – Iced (dense – 1 CPU)	HiC-Pro – Iced (sparse – 1 CPU)	HiCorrector – MES (dense – 1 CPU)	HiCorrector – MEP (dense – 8 CPUs)
IMR90 1Mbp	00:00:12	00:00:25	00:00:25	00:00:06
IMR90 500 kbp	00:00:40	00:01:30	00:02:15	00:00:22
IMR90 150 kbp	-	00:04:28	00:13:21	00:03:10
IMR90 40 kbp	-	00:07:19	02:35:34	00:35:43
IMR90 2 0kbp	-	00:08:36	12:57:17	02:34:05

HiC-Pro is based on a fast implementation of the iterative correction algorithm. We therefore compare our method with the MES (Memory-Efficient Sequential) and MEP (Memory-Efficient Parallel) algorithms of the HiCorrector software [20] for Hi-C data normalization (hours:minutes:seconds). All algorithms were terminated after 20 iterations (see Additional file 1 for details)

A

Total number of read pairs	826 414 879
Total number of valid pairs	503 587 609 (100%)
Number of pairs assigned to Paternal genome	28 472 217 (5.65%)
Number of pairs assigned to Maternal genome	28 432 091 (5.64%)
Number of trans Maternal/Paternal pairs	608 642 (0.12%)
Number of unassigned reads	446 007 709 (88.57%)
Number of conflicting reads	68 950 (0.01%)

B

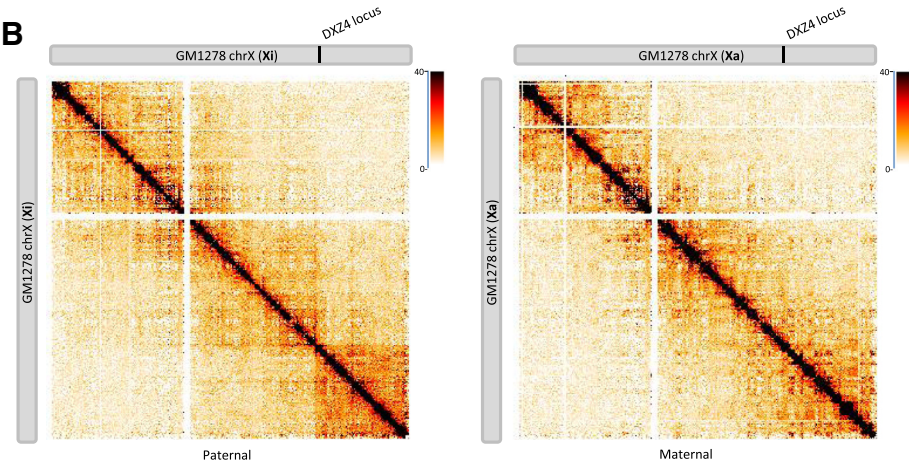


Fig. 2 Allele-specific analysis. **a** Allele-specific analysis of the GM12878 cell line. Phasing data were gathered from the Illumina Platinum Genomes Project. In total, 2,239,492 high quality SNPs from GM12878 data were used to distinguish both alleles. Around 6 % of the read pairs were assigned to each parental allele and used to build the allele-specific contact maps. **b** Intra-chromosomal contact maps of inactive and active X chromosome of the GM12878 cell line at 500-kb resolution. The inactive copy of chromosome X is partitioned into two mega-domains which are not seen in the active X chromosome. The boundary between the two mega-domains lies near the DXZ4 micro-satellite

Binning

In order to generate the contact maps, the genome is divided into bins of equal size, and the number of contacts observed between each pair of bins is reported. A single genome-wide interaction map containing both raw intra- and inter-chromosomal maps is generated for a set of resolutions defined by the user in the configuration file.

Normalization

In theory, the raw contact counts are expected to be proportional to the true contact frequency between two loci. As for any sequencing experiment, however, it is known that Hi-C data contain different biases mainly due to GC content, mappability and effective fragment length [18, 19]. An appropriate normalization method is therefore mandatory to correct for these biases. Over the last few years, several methods have been proposed using either an explicit-factor model for bias correction [19] or implicit matrix balancing algorithm [17, 27]. Among the matrix balancing algorithm, the iterative correction of biases based on the Sinkhorn-Knopp algorithm has been widely used by recent studies due to its conceptual simplicity, parameter-free nature and ability to correct for unknown biases, although its assumption of equal visibility across all loci may require further exploration. In theory, a genome-wide interaction matrix is of size $O(N^2)$, where N is the number of genomic bins. Therefore, applying a balancing algorithm on such a matrix can be difficult in practice, as it requires a significant amount of memory and computational time. The degree of sparsity of the Hi-C data is dependent on the bin size and on the sequencing depth of coverage. Even for extremely large sequencing coverage, the interaction frequency between intra-chromosomal loci is expected to decrease as the genomic distance between them increases. High-resolution data are therefore usually associated with a high level of sparsity. Exploiting matrix sparsity in the implementation can improve the performance of the balancing algorithm for high-resolution data. HiC-Pro proposes a fast sparse-based implementation of the iterative correction method [17], allowing normalization of genome-wide high-resolution contact matrices in a short time and with reasonable memory requirements.

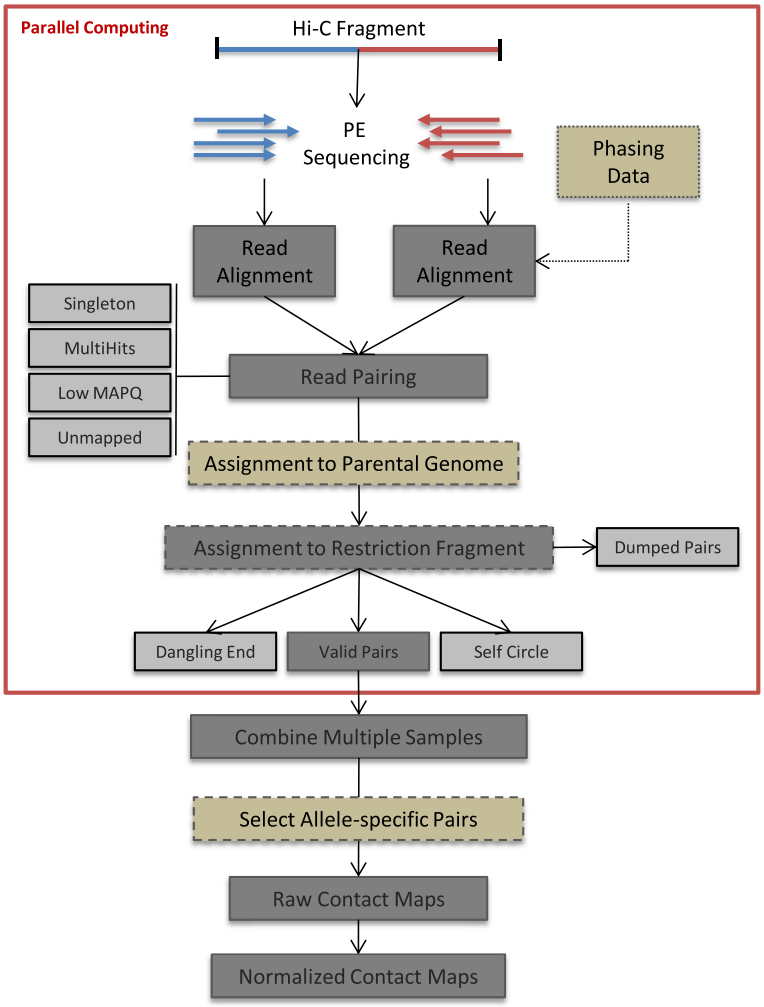


Fig. 3 HiC-Pro workflow. Reads are first aligned on the reference genome. Only uniquely aligned reads are kept and assigned to a restriction fragment. Interactions are then classified and invalid pairs are discarded. If phased genotyping data and N-masked genome are provided, HiC-Pro will align the reads and assign them to a parental genome. For the Hi-C protocol based on restriction enzyme digestion, the read pairs will then be assigned to a restriction fragment and invalid ligation products will be filtered out. These first steps can be performed in parallel for each read chunk. Data from multiple chunks are then merged and binned to generate a single genome-wide interaction map. For allele-specific analysis, only pairs with at least one allele-specific read are used to build the contact maps. The normalization is finally applied to remove Hi-C systematic bias on the genome-wide contact map. *MAPQ* Mapping Quality , *PE* paired end

Quality controls

To assess the quality of a Hi-C experiment, HiC-Pro performs a variety of quality controls at different steps of the pipeline (Fig. 5). The alignment statistics are the first available quality metric. According to the reference genome, a high-quality Hi-C experiment is usually associated with a high mapping rate. The number of reads aligned in the second mapping step is also an interesting control as it reflects the proportion of reads spanning the ligation junction. An abnormal level of chimeric reads can reflect a ligation issue during library preparation. Once the reads are aligned on the genome, the fraction of singleton or multiple hits is usually expected to be low. The ligation efficiency can also be assessed using the filtering of valid and invalid pairs. As ligation is a random process, it is expected that 25 % of each valid ligation class will be defined by distinct read pair orientation. In the same way, a high level of dangling-end or self-circle read pairs is associated with a bad quality experiment, and reveals a problem during the digestion, fill-in or ligation steps.

Additional quality controls, such as fragment size distribution, can be extracted from the list of valid interaction products (Figure S3 in Additional file 1). A high level of duplication indicates poor molecular complexity and a potential PCR bias. Finally, an important metric is the fraction of intra- and inter-chromosomal interactions, as

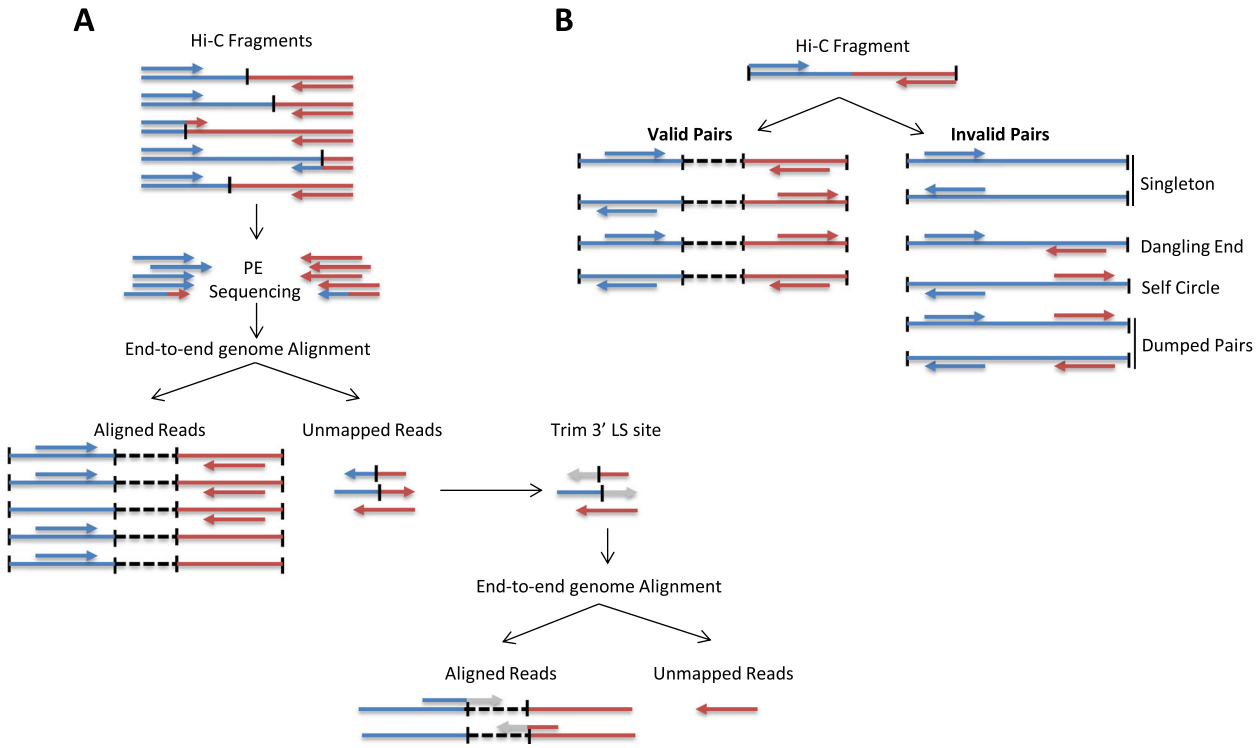


Fig. 4 Read pair alignment and filtering. **a** Read pairs are first independently aligned to the reference genome using an end-to-end algorithm. Then, reads spanning the ligation junction which were not aligned in the first step are trimmed at the ligation site and their 5′ extremity is realigned on the genome. All aligned reads after these two steps are used for further analysis. **b** According to the Hi-C protocol, digested fragments are ligated together to generate Hi-C products. A valid Hi-C product is expected to involve two different restriction fragments. Read pairs aligned on the same restriction fragment are classified as dangling end or self-circle products, and are not used to generate the contact maps. *PE* paired end, *LS Ligation Site*

well as long-range versus short-range intra-chromosomal interactions. As two genomic loci close on the linear genome are more likely to randomly interact, a strong diagonal is expected on the raw contact maps. A low quality experiment will result in a low fraction of intra-chromosomal interactions depending on the organism and the biological context. A high quality Hi-C experiment on the human genome is typically characterized by at least 40 % of intra-chromosomal interactions [9]. In the same way, a high quality experiment is usually characterized by a significant fraction (>40 %) of long-range intra-chromosomal valid pairs [7].

Speed and scalability

Generating genome-wide contact maps at 40 to 1 kb resolution requires a sequencing depth of hundreds of millions to multi-billions of paired-end reads depending on the organism [7, 8]. However, the main processing steps from read mapping to fragment reconstruction can be optimized using parallel computation of read chunks, significantly reducing the time taken by the Hi-C data processing. Next, all valid interactions are merged to remove the duplicates and to generate the final contact maps.

The user can easily run the complete analysis workflow with a single command line either on a single laptop or on a computer cluster. Analysis parameters are all defined in a single configuration file. In addition, HiC-Pro is modular and sequential, allowing the user to focus on a sub-part of the processing without running the complete workflow. In this way, HiC-Pro can also be used to complement other methods, for instance, by running the workflow from already aligned files, or by simply normalizing published raw contact maps.

The main steps of the pipeline are implemented in Python and C++ programming languages and are based on efficient data structures, such as compressed sparse row matrices for contact count data. Using an adequate data structure allows the data processing to be sped up as well circumvents memory limitations. In this way, HiC-Pro allows a genome-wide iterative correction to be run at very high resolution and in a short time. Our normalization implementation exploits *numpy*'s dense array format and fast operations, *scipy*'s sparse matrices representation and *Cython* to combine C and Python to reach the performance of C executables with the ease of use and maintainability of the Python language.

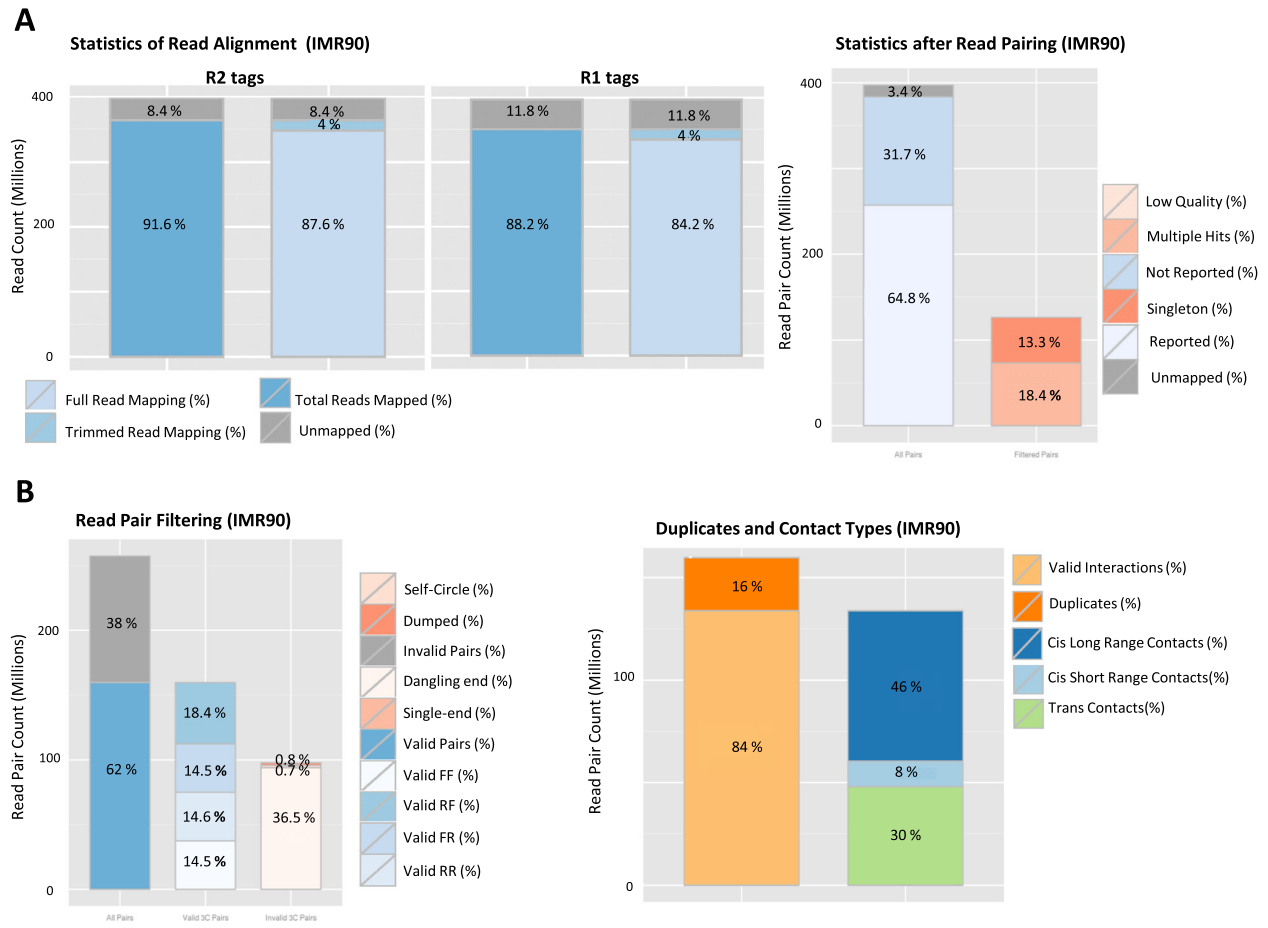


Fig. 5 HiC-Pro quality controls. Quality controls reported by HiC-Pro (IMR90, Dixon et al. data). **a** Quality control on read alignment and pairing. Low quality alignment, singleton and multiple hits are usually removed at this step. **b** Read pair filtering. Read pairs are assigned to a restriction fragment. Invalid pairs, such as dangling-end and self-circle, are good indicators of the library quality and are tracked but discarded for subsequent further analysis. The fractions of duplicated reads, as well as short range versus long range interactions, are also reported

Contact map storage

Genome-wide contact maps are generated for resolutions defined by the user. A contact map is defined as a matrix of contact counts and a description of the associated genomic bins and is usually stored as a matrix, divided into bins of equal size. The bin size represents the resolution at which the data will be analyzed. For instance, a human 20 kb genome-wide map is represented by a square matrix of 150,000 rows and columns, which can be difficult to manage in practice. To address this issue, we propose a standard contact map format based on two main observations. Contact maps at high resolution are (i) usually sparse and (ii) expected to be symmetric. Storing the non-null contacts from half of the matrix is therefore enough to summarize all the contact frequencies. Using this format leads to a 10–150-fold reduction in disk space use compared with the dense format (Table 4).

Allele-specific analysis

HiC-Pro is able to incorporate phased haplotype information in the Hi-C data processing in order to generate allele-specific contact maps (Fig. 2). In this context, the sequencing reads are first aligned on a reference genome for which all polymorphic sites were first N-masked. This masking strategy avoids systematic bias toward the reference allele, compared with the standard procedure where reads are mapped on an unmasked genome. Once aligned, HiC-Pro browses all reads spanning a polymorphic site, locates the nucleotide at the appropriate position, and assigns the read to either the maternal or paternal allele. Reads without SNP information as well as reads with conflicting allele assignment or unexpected alleles at polymorphic sites are flagged as unassigned. A BAM file with an allele-specific tag for each read is generated and can be used for further analysis. Then, we classify as allele-specific all pairs for which both reads

Table 4 Comparison of contact map formats

	Dense format (MB)	Sparse symmetric format (MB)
IMR90_CCL186 1 Mbp	27	49
IMR90_CL186 500 kbp	82	181
IMR90_CCL186 150 kbp	822	911
IMR90_CCL186 40 kbp	12,000	1900
IMR90_CL186 20 kbp	45,000	2600
IMR90_CL186 5 kbp	720,000	4200

Disk space for IMR90_CCL186 genome-wide contact maps generated using either the classical dense format or the sparse symmetric format at different resolutions

are assigned to the same parental allele or for which one read is assigned to one parental allele and the other is unassigned. These allele-specific read pairs are then used to generate a genome-wide contact map for each parental genome. Finally, the two allele-specific genome-wide contact maps are independently normalized using the iterative correction algorithm.

Software requirements

The following additional software and libraries are required: the bowtie2 mapper [26], R and the BioConductor packages *RColorBrewer*, *ggplot2*, *grid*, Samtools (>0.1.19), Python (>2.7) with the *pysam*, *bx.python*, *numpy* and *scipy* libraries, and the g++ compiler. Note that a bowtie2 version > 2.2.2 is strongly recommended for allele-specific analysis, because, since this version, read alignment on an N-masked genome has been highly improved. Most of the installation steps are fully automatic using a simple command line. The bowtie2 and Samtools software are automatically downloaded and installed if not detected on the system. The HiC-Pro pipeline can be installed on a Linux/UNIX-like operating system.

Conclusions

As the Hi-C technique is maturing, it is now important to develop bioinformatics solutions which can be shared and used for any project. HiC-Pro is a flexible and efficient pipeline for Hi-C data processing. It is freely available under the BSD licence as a collaborative project at https://github.com/nservant/HiC-Pro. It is optimized to address the challenge of processing high-resolution data and provides an efficient format for contact map sharing. In addition, for ease of use, HiC-Pro performs quality controls and can process Hi-C data from the raw sequencing reads to the normalized and ready-to-use genome-wide contact maps. HiC-Pro can process data generated from protocols based on restriction enzyme or nuclease digestion. The intra- and inter-chromosomal contact maps generated by HiC-Pro are highly similar to the ones generated by the hiclib package. In addition, when phased genotyping data are available, HiC-Pro allows the easy generation of allele-specific maps for homologous chromosomes. Finally, HiC-Pro includes an optimized version of the iterative correction algorithm, which substantially speeds up and facilitates the normalization of Hi-C data. The code is also available as a standalone package (https://github.com/hiclib/iced).

A complete online manual is available at http://nservant.github.io/HiC-Pro. The raw and normalized contact maps are compatible with the HiTC Bioconductor package [28], and can therefore be loaded in the R environment for visualization and further analysis.

Additional file

Additional file 1: The supplementary data file contains a description of the dataset used for this study as well as details about how the HiC-Pro, hiclib and HiCorrector software were used in practice. It also includes supplementary figures about the HiC-Pro results and output. (DOCX 1452 kb)

Abbreviations

PCR: polymerase chain reaction; SNP: single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

N.S., B.R., C.C., J.D., E.H. and E.B. designed the main steps of the Hi-C analysis workflow. N.V. and J.V. developed the new iterative correction implementation. N.S., E.V., B.R. and N.V. developed, optimized and tested the HiC-Pro pipeline. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Felix Krueger for useful discussion about allele-specific analysis, and Jesse Dixon and Neva Cherniavsky for their advice in defining the best GM12878 phasing data. This work was supported by the France Genomique National infrastructure (ANR-10-INBS-09), the Labex Deep, the European Research Coucil (SMAC-ERC-280032), the ERC Advanced Investigator award (ERC-250367), the European Commission (HEALTH-F5-2012-305626), the ABS4NGS project (ANR-11-BINF-0001), the National Human Genome Research Institute (R01 HG003143), and the Paris Alliance of Cancer Research Institutes (PACRI-ANR). J.D. is an Investigator of the Howard Hughes Medical Institute.

Author details

[1]Institut Curie, Paris, France. [2]INSERM, U900, Paris, France. [3]Mines ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, Fontainebleau, France. [4]Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA, USA. [5]Sysra, Yerres, France. [6]CNRS UMR3215, Paris, France. [7]INSERM U934, Paris, France. [8]Annoroad Gene Technology Co., Ltd, Beijing, China. [9]Howard Hughes Medical Institute, Program in Systems

Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA, USA.

Received: 10 August 2015 Accepted: 11 November 2015
Published online: 01 December 2015

References

1. de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. Genes Dev. 2012;26:11–24.
2. Barutcu AR, Fritz AJ, Sayyed KZ, van Wijnen AJ, Lian JB, Stein JL, et al. C-ing the genome: A compendium of chromosome conformation capture methods to study higher-order chromatin organization. J Cell Physiol. 2015; 1097–4652. doi:10.1002/jcp.25062.
3. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93. doi:10.1038/ng.947.
4. Ma W, Ay F, Lee C, Gulsoy G, Deng X, Cook S, et al. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of lincRNA genes in human cells. Nat Methods. 2015;12:71–8. doi:10.1038/nmeth.3205.
5. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. Nature. 2012;485(7398):381–5. doi:10.1038/nature11049.
6. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80. doi:10.1038/nature11082.
7. Rao SSP, Huntley MH, Durand NC, Bochkov SID, Robinson JT, Sanborn AL, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80. doi:10.1016/j.cell.2014.11.021.
8. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee Y, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013; 503(7475):290–4. doi:10.1038/nature12644.
9. Lajoie BR, Dekker J, Kaplan N. The hitchhiker's guide to hi-c analysis: Practical guidelines. Methods. 2015;72:65–75. doi:10.1016/j.ymeth.2014.10.031.
10. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Lasio P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576–89. doi:10.1016/j.molcel.2010.05.004.
11. HiCUP. http://www.bioinformatics.babraham.ac.uk/projects/hicup/.
12. Castellano G, Le Dily F, Hermoso Pulido A, Beato M, Roma G. Hi-Cpipe: a pipeline for high-throughput chromosome capture. bioRxiv 2015. Cold Spring Harbor Labs Journals. doi:10.1101/020636.
13. Schmid Marc W, Schmid MW, Grob S, Grossniklaus U. HiCdat: a fast and easy-to-use Hi-C data analysis tool. BMC Bioinformatics. 2015;16:277. doi:10.1186/s12859-015-0678-x.
14. HiCbox. https://github.com/koszullab/HiC-Box.
15. SNPsplit. http://www.bioinformatics.babraham.ac.uk/projects/SNPsplit/.
16. Hsieh TS, Weiner A, Lajoie B, Dekker J, Friedman N, Rando O. Mapping nucleosome resolution chromosome folding in yeast by micro-C. Cell. 2015;162(1):108–19. doi:10.1016/j.cell.2015.05.048.
17. Imakaev M, Fudenberg G, Patton McCord R, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of hi-c data reveals hallmarks of chromosome organization. Nat Methods. 2012;9(10):999–1003. doi:10.1038/nmeth.2148.
18. Yaffe E, Tanay A. Probabilistic modelling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43(11):1059–65. doi:10.1038/ng.947.
19. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu J. HiCNorm: removing biases in Hi-C data via poisson regression. Bioinformatics. 2012;28(23):3131–3. doi:10.1093/bioinformatics/bts570.
20. Li W, Gong K, Li Q, Alber Fand Zhou XJ. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale hi-c data. Bioinformatics. 2015;31(6):960–2. doi:10.1093/bioinformatics/btu747.
21. Minajigi A, Froberg JE, Wei C, Sunwoo H, Kesner B, Colognori D, et al. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. Science. 2015;349(6245):aab2276. doi:10.1126/science.aab2276.
22. Deng X, Ma W, Ramani V, Hill A, Yang F, Ay F, et al. Bipartite structure of the inactive mouse X chromosome. Genome Biol. 2015;16:152. doi:10.1186/s13059-015-0728-8.
23. Selvaraj S, Dixon JR, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation ans shotgun sequencing. Nat Biotechnol. 2013;31(12):1111–8. doi:10.1038/nbt.2728.
24. Platinum Illumina Genomes Project. http://www.illumina.com/platinumgenomes/.
25. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
26. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie2. Nat Methods. 2012;9:357–9. doi:10.1038/nmeth.1923.
27. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. BMC Genomics. 2012;13:436. doi:10.1186/1471-2164-13-436.
28. Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen CJ, Heard E, et al. HiTC: Exploration of high-throughput 'C' experiments. Bioinformatics. 2012;28(21):2843–4.

HiC-Pro: An optimized and flexible pipeline for Hi-C data processing.

N. Servant, N. Varoquaux, B. R. Lajoie, E. Viara, CJ. Chen, JP. Vert, E. Heard, J. Dekker, E. Barillot

SUPPLEMENTARY MATERIAL

I. Public dataset used.

We applied the HiC-Pro pipeline on three public dataset available on GEO.
The IMR90 Hi-C contact maps were first published by Dixon et al. at a resolution of 20Kb and 40Kb. The five run of IMR90 replicate 1 (GSM862724) were used and merged, for a total number of 397.2 million read pairs. We refer to this sample in the manuscript as IMR90.
More recently, Rao et al. generate genome-wide contact maps at a resolution of 1-5kb (GSE63525) for nine different cell lines. For the purpose of this paper, we applied HiC-Pro on the IMR90 cell line (GSM1551599, GSM1551600, GSM1551601, GSM1551602, GSM1551603, GSM1551604, GSM1551605). The combined samples represent a sequencing depth of 1.5 billion reads. We refer to this sample in the manuscript as IMR90_CCL186.
The allele specific analysis was performed using the human GM12878 Hi-C data published by Selveraj et al. (GSE48592). Phasing data were gathered from the Illumina Platinum Project v8.0.1 (http://www.illumina.com/platinumgenomes/).

II. Results and implementation

All pipelines and software were run on the high-performance computing resource of the Institut Curie. Each node has a total of 32 or 48 processors (Intel Xeon 2.2 GHz) and 128 GB memory.
The HiC-Pro version 2.6.0 was used and the *hiclib* library was downloaded from http://mirnylab.bitbucket.org/hiclib/. In order to compare the performance between both solutions, we run the pipeline described in the *hiclib*'s repository (hiclib/examples/pipeline2014/), on a single node with 8 CPUs. Following the *hiclib*'s help pages, the *binnedData* and *highResBinnedData* classes were respectively used for low (>100kb) and high resolution data (<100kb) as illustrated in the *testHighResHiC.py* script.
The HiC-Pro pipeline was run either in normal or parallel mode. HiC-Pro and *hiclib* were compared until the generation of genome-wide normalized contact maps at a resolution of 1Mb, 500Kb, 150Kb, 40Kb and 20Kb. Both pipelines were run with default parameters. The running time includes the export of contact maps in text format.

In order to compare the results generated by both pipelines, we calculated the Spearman correlation coefficient between HiC-Pro and hiclib intra and inter-chromosomal maps at different resolutions. By default hiclib is removing the matrix diagonal before normalizing the data. We therefore apply the same filter on the HiC-Pro contact maps to compare the results. The Spearman correlation coefficients were calculated between all intra-chromosomal maps. Since the inter-chromosomal contact maps are sparse, instead of measuring the correlation directly between the two maps, we computed the Spearman correlation of the one-dimensional coverage vectors of inter-chromosomal maps as proposed by Yaffe and Tanay (2011), and Hu et al. (2012). The results are available in Figure S4.

The HiCorrector package (version 1.1) was downloaded and compiled using openmpi-1.4.5. We compared the performance of the iterative correction algorithm included in HiC-Pro with HiCorrector on the Dixon et al. IMR90 dataset. We first split the dense matrix files using the *split_data_parallel* tool and the following command line; "*mpirun -np 8 split_data_paralllel DENSE_MATRIX_FILE NB_ROWS ./ 8 1024 job_id*" where *DENSE_MATRIX_FILE* is the path to the dense matrix and *NB_ROWS* the number of matrix rows. The genome wide contact maps were therefore split into 7 sub-matrices for 1M, 500Kb, 150Kb resolutions, 28 sub-matrices for the 40Kb resolution and 91 for the 20 Kb resolution.

The iterative correction was then applied using the ICE-MES and ICE-MEP methods on the genome-wide contact map. All algorithms were terminated after 20 iterations.

We ran the ICE_MEP method using the following parameters; "*mpirun -np 8 ic_mep –useSplitInputFiles –numRows=NB_RAWS –maxIteration=20 –numTask=8 –memSizePerTask=1024 –jobID=job_id*". The ICE_MES method was run using the following parameters; ""*ic_mes DENSE_MATRIX_FILE 5000 3115 20 0 0*".

The HiC-Pro normalization (1 CPU) was run using the ice script and the following parameters; "*--max_iter 20 –eps 1e-15 –filtering_perc 0*". The "*--dense*" option was added for the dense matrices. All input and output files were stored in the local *scratch* folder to limit the I/O time due to NFS system.

SUPPLEMENTARY FIGURES.

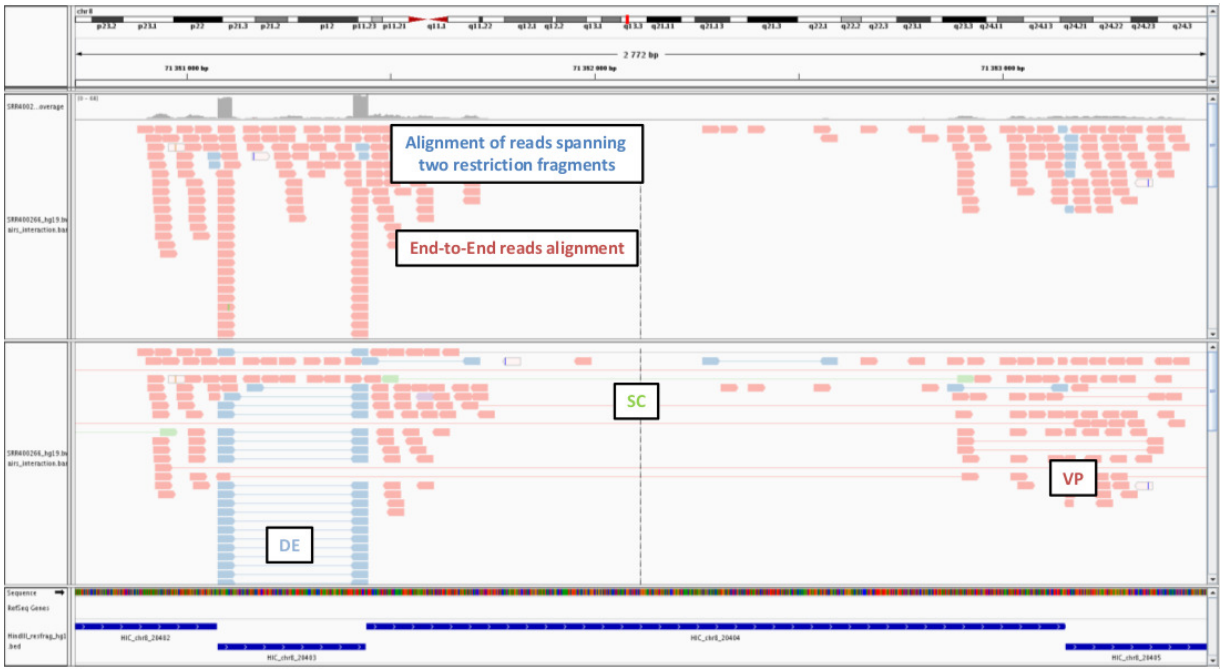


Figure S1: IGV screenshot of BAM file after mapping and fragment reconstruction. Top panel. The reads are colored according to the alignment procedure. Blue reads were trimmed before mapping, and flanked the restriction fragment borders. **Bottom panel.** Read pairs are colored according to their classification. Valid interactions are in red, dangling end in blue and self-circle ligation in green.

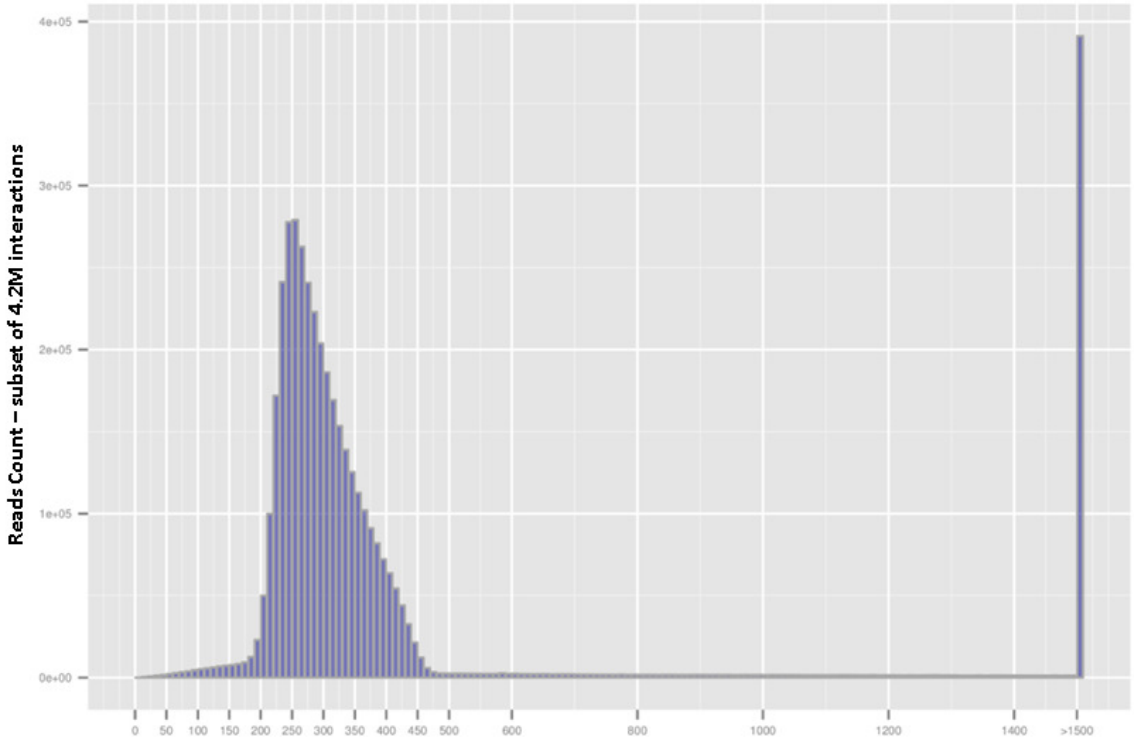


Figure S2: Size distribution of Hi-C ligation products generated by HiC-Pro (IMR90). Both reads are expected to map near a restriction site, and with a distance within the range of molecule size distribution after shearing. Fragments with a size outside the expected range can be discarded if specified in the HiC-Pro configuration file.

2.3 HiC-Pro roadmap and future developments

Since its first release, HiC-Pro has been regularly updated, providing new functionalities described below.

2.3.1 Support of all Hi-C-based protocols

HiC-Pro was first designed to process data based on dilution or in situ Hi-C protocols (see section 1.3 for details). These two protocols are based on restriction enzymes that digest the chromatin before ligation. During the analysis, the restriction fragments are used to distinguish valid ligation products which involved two distinct fragments, from non-informative read pairs aligned on the same restriction fragment.

In parallel to these protocols, other methods have then been proposed for chromatin fragmentation. For instance, DNase (Ma et al. (2015)) or MNase (Hsieh et al. (2015)) have also been used, providing a random distribution of digested fragments and a better resolution than 6bp restriction enzymes. HiC-Pro proposes a dedicated mode to process Hi-C data which are not based on restriction enzyme digestion. In this case, the mapping is performed in a single step, as the ligation motif is unknown. Then, non-informative read pairs are filtered based on the distance between the two reads. Reads separated from a distance smaller than 1kb are likely to be dangling-end or self-ligation fragments, and are discarded from the analysis. Applying the same filter on standard Hi-C data allows to remove 90% of non-informative pairs.

More recently, HiC-Pro was also updated to process capture-C and capture-Hi-C data. In both cases, the principle of these protocols is to enriched the 3C library in a set of regions of interest. Therefore, the first steps of the analysis are usually the same than for standard Hi-C protocols. In capture-C experiment, the interest is limited to one or several viewpoints. Thus, there is no interest in building genome-wide interaction maps. Using the stepwise mode of HiC-Pro, the data processing can be run until the detection of valid interaction products. Then, given a list of viewpoints, we designed a new HiC-Pro utility to extract all interactions involving the viewpoints of interest, and build a genome track that can be visualized as a 4C-like profile. We further compared the results of HiC-Pro with the CCanalyserV2 tool, whih is a pipeline dedicated to capture-C analysis (Davies et al. (2016)). Both tools give very concordant results,

therefore validating the ability of HiC-Pro to efficiently process capture-C data (Figure 2.1).

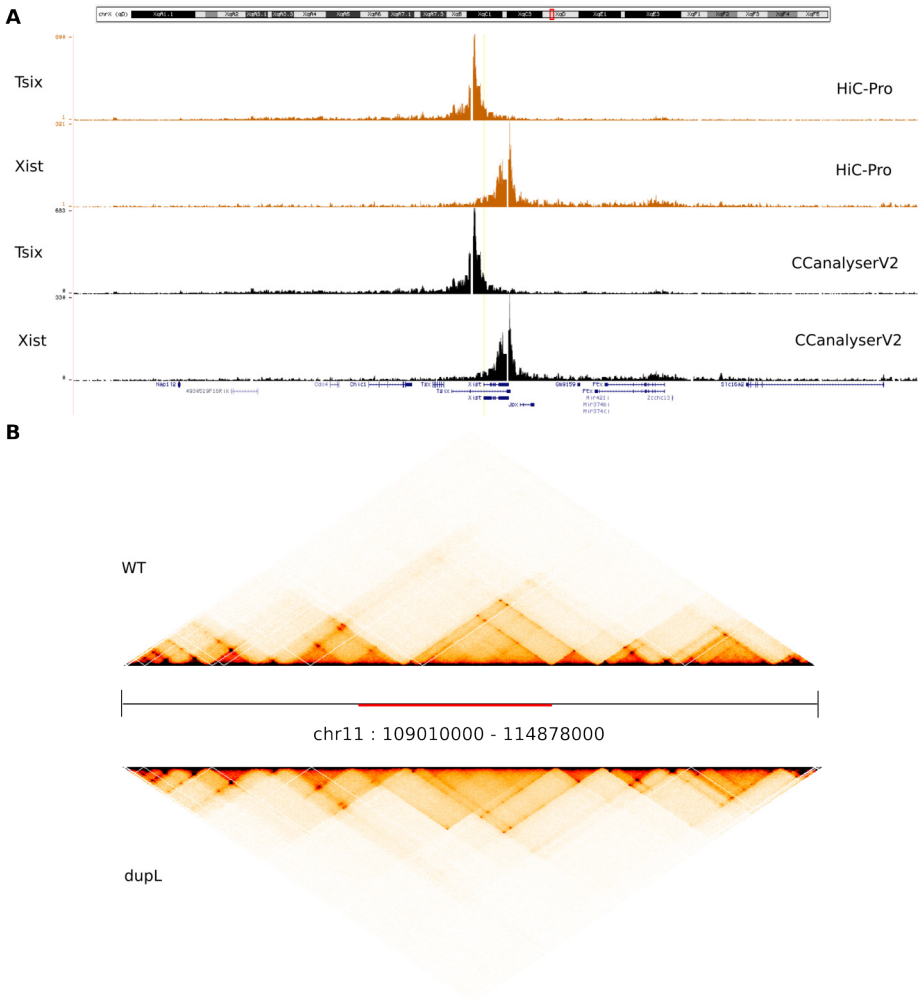


Figure 2.1: Analysing of capture-C and capture-Hi-C with HiC-Pro - A. Capture-C of Xist and Tsix genes on E14 Mouse cell line. Results of HiC-Pro (orange) and CCanalyserV2 tool (black, Davies et al. (2016)). **B.** Contact maps (10kb resolution) from capture-Hi-C data from Franke et al. and analyzed with HiC-Pro. The dupL sample is characterized by a genomic duplication (in red) which is not present in the normal (WT) sample.

Regarding the capture-Hi-C analysis, HiC-Po is able to use the coordinates of the targeted regions to filter out all off-target interactions. The output is a contact map of

intra-chromosomal interactions within the targeted region (Figure 2.1b).

In addition to Hi-C protocols, a recent improvement of the ChIA-PET, called HiChIP, was recently proposed (Mumbach et al. (2016)). The HiChIP is based on the same principle than the in situ Hi-C protocol. But once the contacts are established in the nucleus, an immunoprecipitation is performed on the library to directly capture the interactions associated with a protein of interest. In their paper, Mumbach et al. proposed to use HiC-Pro to process HiChIP data, as the goal of the processing is to filter out non-informative ligation products as in Hi-C data analysis. Note that in the context of HiChIP, the assumption that all fragments have, in average, the same contact frequency genome-wide is not valid and that it is therefore not adviced to use the ICE method to normalize the data.

2.3.2 Birth of a collaborative project

Developing open-source software is now current in research. However, encouraging people to directly contribute to the project is much more difficult. To do so, HiC-Pro is accompanied with a detailed manual. Each script is documented following the good practices in software development. In addition, HiC-Pro is organized in independant modules. Therefore, modifying one module does not require to understand all the other ones. Finally, HiC-Pro is available through an efficient version control system (Git), that manages and stores revisions of the project. The GitHub project is itself associated to a forum where people can ask questions, and share their experience and feedbacks on the software. One advantage of using such system is that people can easily clone the software sources, add new functionalities and proposed add-ons or corrections. All changes are carefully checked for compatibility by the system before merging.
In practice, several functions have been proposed by the community to improve HiC-Pro. This is for instance the case of the computational cluster support. In the first version, HiC-Pro was packaged to support the PBS/Torque cluster management system. Rapidly, other users proposed updates to support SLURM and SGE cluster management system. These updates are now fully integrated into the main branch.

2.3.3 Compatibility with others Hi-C tools

HiC-Pro is also compatible with several downstream analysis tools, such as HiCPlotter (Akdemir and Chin (2015)), Juicebox (Durand et al. (2016a)), or FitHiC (Ay et al.

(2014)) (Figure 2.2). To do so, we developed several utilities able to convert the HiC-Pro output format to other input formats. Of note, most of these developments have been made in collaboration with the developers of the other tools.

HiCPlotter is a recently published tool to visualize Hi-C data and to integrate different genomic tracks with the contact maps. In collaboration with the HiCPlotter developers, we proposed to update the HiCPlotter input file format to support the HiC-Pro contact maps. Users can now apply HiC-Pro to process their raw data, and visualize the contact maps using HiCPlotter (Figure 2.3A).

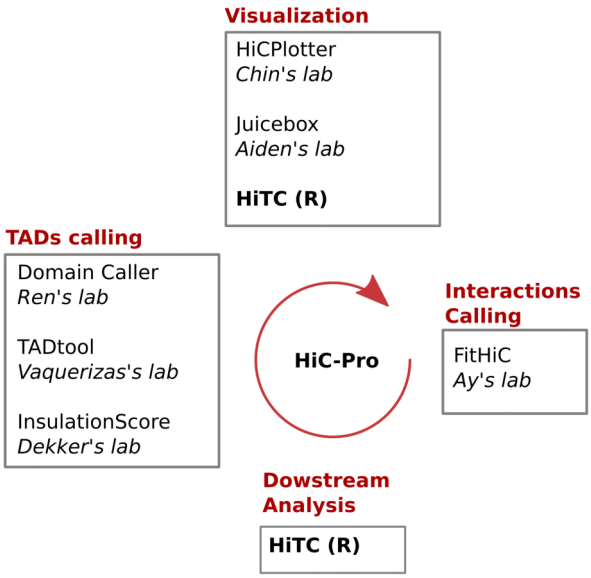


Figure 2.2: Compatibility of HiC-Pro with other tools - HiC-Pro (v.2.9.0) compatibility with other tools for visualization, TADs calling, interaction calling or other downstream analysis. Softwares that we developed are in bold.

Another powerful tool, Juicebox, has been recently proposed by Durand et al.. Juicebox allows to load contact maps in a dynamic way, to change the resolution of the data on-the-fly, up to the restriction fragments resolution, and to zoom in/out on the contact maps.

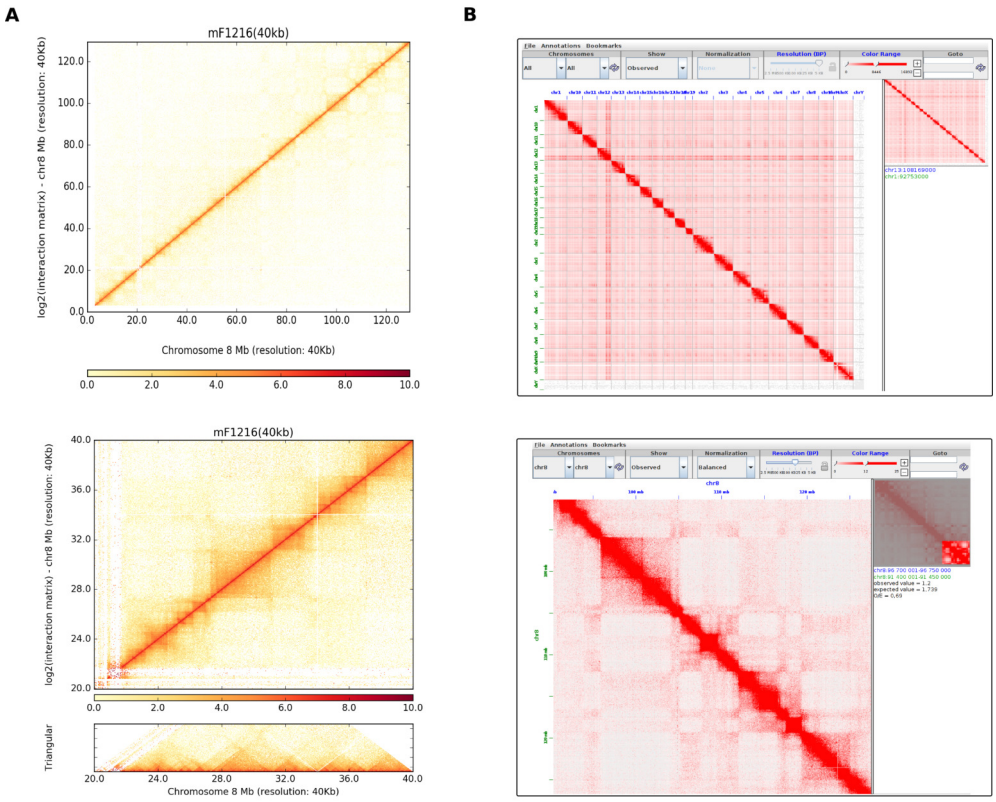


Figure 2.3: Visualization of HiC-Pro results with HiCPlotter and Juicebox - **A.** HiCPlotter visualization of intra-chromosomal contact maps (chr8) of Mouse ESC, generated with HiC-Pro. **B.** Dynamic visualization of contact maps generated by HiC-Pro with the Juicebox software (Mouse ESC cells).

Additional downstream analysis such as compartment calling are also available. Juicebox offers an access to pre-loaded datasets from many other studies, including data from a wild variety of organisms. Additional tracks from the ENCODE project are also available, and users can also load their own genomic tracks if needed. Juicebox is currently one of the most powerful tool for Hi-C data visualization. Its interface is easy to use, and the availability of external sources of data, makes it very attractive for biologists. Juicebox was first designed to visualize data from the Juicer processing pipeline, developed by the same team (Durand et al. (2016b)). But rapidly, the authors proposed a toolbox to generate Juicebox input file from a list of 3C interaction products. We therefore updated the HiC-Pro output to be direcly compatible with the Juicebox format. In addition, we provided a simple converter, starting from HiC-Pro

output files, and able to generate Juicebox input files. Users can now process their data with HiC-Pro and visualize their results using the Juicebox tool (Figure 2.3B).
In addition to visualization, HiC-Pro is also compatible with downstream analysis tools (Figure 2.2). For instance, we provided an utility to convert the HiC-Pro sparse triplet format into dense format for TADs calling as required by the Directionality Index method (Dixon et al. (2012), section 1.4.4.3). Other methods such as FitHiC (Ay and Noble (2015)), are also compatible with HiC-Pro. In collaboration with the FitHiC developers, we designed a converter to load HiC-Pro output into the FitHiC R package, and therefore to use it to call significant contacts in the data.
Finally, HiC-Pro is fully compatible with our HiTC R package (Servant et al. (2012)). The users can therefore load their interaction matrices into the R environment. We implemented into the HiTC R package a suite of functions for downstream analysis, including quality controls, statistics, visualization, and compartments calling (see section 1.6.2). In addition to these functions, being able to load the interaction matrices in R paves the way to new statistical developments in this field.

2.4 Discussion

While the 'C' techniques and the exploration of chromatin conformation mature, it is important to develop appropriate and robust method for data analysis and interpretation. It is now time to converge on the analysis protocol, rather than reinventing new analysis methods in each published paper. This is necessary to ensure reproducibility between studies and to compare and interpret results from different laboratories.
This is in this context that we developed HiC-Pro. HiC-Pro performs quality controls, reads alignments, detection of valid interaction products and generate raw and normalized ready-to-use contact maps. Each of these steps was optimized using appropriate data structure, programming langage and algorithms. In addition, HiC-Pro is able to run allele-specific Hi-C analysis, and to build allelic contact maps for the two parental alleles. To our knownledge, this is the only tool able to process Hi-C data from all protocols, regardless the fragmentation strategy. HiC-Pro can be used on a simple laptop or on a computational cluster to process very high-throughtput datasets. As illustrated above, we have made significant efforts to produce an efficient software, and to facilitate its use by the community. We have attached the greatest importance

to make it easy to use, documentated and compatible with other software. All these aspects certainly explain the current popularity of HiC-Pro, and take a step forward in the reproducibility and the comparison of data from different studies.

In the coming years, protocols and experiments to explore the genome organization will continue to be developed. Recent project like the 4D Nucleome Project will address these challenges. In praticular, new experimental standards and comparison of protocols will be performed. Standard on data analysis methods should also be addressed, and innovative algorithms for downstream analysis and 3D modelling will be developed. In the same way, it would be useful to run comparative studies on the Hi-C analysis methods, including data processing but also statistical analysis, and methods to infer the 3D structure of genomes. Even if the lack of validation data makes such comparison difficult, these results can help in consolidating some approaches by highligthing differences between them.

As protocols will continue to evolve, we will continue to update and develop HiC-Pro following the needs of the community. One point which is important is that HiC-Pro was built to efficiently process Hi-C data, and its role is limited to this task. We made the choice not to include any downstream analysis method into HiC-Pro, but rather to facilitate the link between HiC-Pro and other software. Among the future improvements of the pipeline, some users already proposed additional quality controls that we can add. We also plan to add additional options for capture-Hi-C processing, and to report more descriptive statistics such as the fraction of off-target interactions. Regarding the optimization, we identified a couple of points that can be further improved to make HiC-Pro faster. Among them, the last part of the analysis from the duplicates removal to the normalization is not yet parallelized. A simple optimization would be to run simultaneously several samples or resolution of Hi-C maps at the same time.

All these improvements will be available in the future versions of HiC-Pro.

3

Normalization of cancer Hi-C data

3.1 Challenges in Hi-C data normalization

Since the beginning of next-generation sequencing, many studies have reported the importance of data normalization in downstream analysis. Two main types of normalization have been proposed, either to correct experimental biases within a sample, or to adjust samples for technical differences such as sequencing depth before comparison. Among the most studied biases, the GC content and the mappability are common to all sequencing applications based on the Illumina system. Indeed, it has been reported that extreme base compositions, such as GC-poor or GC-rich sequences, lead to an uneven reads coverage across the genome. As an example, the reads coverage of a diploid whole genome sequencing is usually not uniform, but varies along the genome according to its base composition. In addition to GC content, the mappability introduces another bias in all applications that requires the alignment of sequencing reads on a reference genome. The mappability refers to the uniqueness of a sequence along the genome. For instance, repeated regions are usually much more difficult to study, mainly because the alignment tools are not able to reliably assign the reads to repeated genomic loci. Those regions are usually poorly covered or excluded from the analysis.
As expected, these two types of bias were also observed in Hi-C data (Yaffe and Tanay (2011)). In addition, another bias due to restriction fragment lengths was also introduced. Indeed, it has been shown that the length of DNA fragments after digestion is

highly correlated with the ligation efficiency. Fragments of different sizes are usually more difficult to ligate than fragments of the same size (see Figure 1.17, Yaffe and Tanay (2011)).

Correcting the data from these biases is an important challenge of the analysis, especially since it can have strong impacts on the results and on the biological conclusions or interpretations of the data. Regarding the Hi-C data, two main types of methods have been proposed (see section 1.4.1.4). The methods based on explicit-factor correction rely on a-priori known biases, that are determined empirically from the data and then corrected using appropriate statistical models. In addition, other methods based on matrix balancing algorithms have also been proposed. The main interest of the matrix balancing is that it does not require any a-priori knowledge about the biases to correct, but instead, assumes that the effect of all biases is captured in the sequencing coverage of each genome locus. Thus, the matrix balancing methods are based on the strong assumption that each locus should interact the same number of time, genome-wide, after biases are removed. Among the available matrix balancing algorithms, the ICE normalization (Imakaev et al. (2012)) is widely used, mainly because of its simplicity and parameter-free algorithm.

Finally, the aforementioned biases are true regardless the type of biological sample. However, in the context of cancer, an additional bias due to genome rearrangements and copy number has been reported in several sequencing applications such as whole genome or ChIP sequencing (Ashoor et al. (2013), Boeva et al. (2012)). So far, this bias has been poorly explored in Hi-C data, and it is tempting to speculate that the copy number also affects the contact frequency of rearranged loci.

We therefore proposed to explore the effect of copy number changes on Hi-C data. This step is crucial for proper interpretation of cancer Hi-C data. To do so, we first designed a simulation model that starts from real diploid contact maps, and simulates the effect of the copy number on the data. We demonstrated that our model gives concordant results with real cancer Hi-C data. We then explored the ability of current methods to normalize cancer Hi-C data. As the ICE normalization method is expected to correct the data from any source of bias, we applied this method on our simulated and on real cancer datasets. We demonstrated that the ICE algorithm is not suitable for cancer Hi-C data, because its assumption of equal visibility is no longer valid in this case.

We then proposed two new methods that take into account the copy number in the

normalization model. The first method, named LOIC (Local Iterative Correction), was designed to remove systematic biases while keeping the copy number information. The idea here is that the copy number can be seen as a biological information that can be used to understand the impact of local rearrangements on the chromatin structure (Harewood et al. (2017)). Nevertheless, we showed that in some cases, the copy number can affect the downstream analysis, and that being able to correct the contact maps is therefore of interest. We thus proposed a second method, CAIC (CNV Adjusted Iterative Correction), to correct for systematic biases including the copy number. We applied both methods to several datasets, including our own Hi-C data on uveal melanoma cell lines, to demonstrate their interest.

3.2 Effective normalization for copy number variation in Hi-C data

Effective normalization for copy number variation in Hi-C data

N. Servant [1,2,3,*†], N. Varoquaux [4,5,†], E. Heard [1,6,7], JP. Vert [1,2,3,†], E. Barillot [1,2,3,†]

October 25, 2017

[1]Institut Curie, Paris, France, [2]INSERM, U900, Paris, France, [3]Mines ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, Fontainebleau, France, [4]Department of Statistics, University of California, Berkeley, USA, [5]Berkeley Institute for Data Science, Berkeley, USA, [6]CNRS UMR3215, Paris, France, [7]INSERM U934, Paris, France, [8]Ecole Normale Superieure, Department of Mathematics and Applications, Paris, France.

Abstract

Normalization is essential to ensure accurate analysis and proper interpretation of sequencing data. Chromosome conformation data, such as Hi-C, is not different. The most widely used type of normalization of Hi-C data casts estimations of unwanted effects as a matrix balancing problem, relying on the assumption that all genomic regions interact as much as any other. Here, we show that these approaches, while very effective on fully haploid or diploid genome, fail to correct for unwanted effects in the presence of copy number variations. We propose a simple extension to matrix balancing methods that properly models the copy-number variation effects. Our approach can either retain the copy-number variation effects or remove it. We show that this leads to better downstream analysis of the three-dimensional organization of rearranged genome.

Background

The spatial organization of the genome and the physical interactions occurring within and between chromosomes are known to play an important role in gene regulation and in genome function in general. The organization and the folding of mammalian chromosomes within the nucleus involve multiple hierarchical chromatin structures (see Bonev and Cavalli [2016] for a review). At the megabase-scale, chromosomes are divided to genomic compartments of active and inactive chromatin, respectively associated with gene-rich, actively transcribed regions and gene-poor, silent regions [Lieberman-Aiden et al., 2009, Rao et al., 2014]. The arrangement of these compartments varies across physiological conditions and cell differentiation [Dixon et al., 2015, Barutcu et al., 2015]. At the sub-megabase scale, chromosomes are partitioned into

*To whom correspondence should be addressed
†Equally Contributed

topological associated domains (TADs). These functional units of regulation are well conserved both across cell types and between mammals [Nora et al., 2012, Dixon et al., 2012, Rao et al., 2014, Dixon et al., 2015]. TAD boundaries are frequently associated with the presence of the CTCF binding factor, itself also involved in the establishment of chromatin loops between convergent target sites [Rao et al., 2014]. These chromatin loops are also commonly associated with promoter-enhancer contacts and therefore with gene activation (see Bouwman and de Laat [2015] for a review).

Given the important recent insights that chromosome conformation techniques have provided into 3D genome organization in a normal context, the application of such approaches to a disease context offers great promises to explore the effect of perturbations in 3D genomic organization on cell regulation (see Krijger and de Laat [2016] for a review). At a high enough resolution, such techniques can be used to characterize links between disease-associated sequence variants and the gene regulatory landscape. For example, structural variants can disrupt the boundaries between TADs, and consequently act as driver events in the mis-regulation of associated gene expression [Franke et al., 2016, Lupiáñez et al., 2016].

Over the past decade, both major advances in high-throughput sequencing techniques and the availability of data from large patient cohorts across multiple cancer types have enabled a comprehensive and systematic exploration of genomic and epigenomic landscapes of a wide variety of cancers. While cancer has been shown to have a genetic component, our appreciation of the inherent epigenetic complexity is more recent and has dramatically increased over the last few years. At the genetic level, cancer is frequently associated with the sequential acquisition of somatic variants, both at single nucleotide and at the copy number levels [Ciriello et al., 2013]. The different alterations that characterize tumors are usually caused by a few functional driver events, which occur among many non-functional passenger events, mainly located in the non-coding part of the genome [Vogelstein et al., 2013]. One of the exciting discoveries that has emerged from systematic sequencing of cancer genomes was the high frequency of mutations in genes known to regulate epigenetic processes such as chromatin associated proteins, DNA methylation, or histone variants and modifications [Plass et al., 2013]. The contribution of altered epigenomes in the process of tumorigenesis is thus at last being unraveled thanks to the combination of genomic and epigenomic interrogation. More recently, genetic and epigenetic alterations in the non-coding part of the genome, including distal regulatory elements such as enhancers or insulators, have been reported and found to impact gene expression in cancer [Taberlay et al., 2014]. This has led to intense interest in the spatial proximity and 3D organization of cancer genomes. Losada [2014] reviews the effect of somatic mutations in cohesin complex proteins (which play a citical role in TADs organization and chromosome looping) in various types of cancer. Gröschel et al. [2014] and Taberlay et al. [2016] describe how disruptions in genome organization (respectively in leukemia and prostate cancer) lead to major epigenetic and transcriptional changes. Lastly, Hnisz et al. [2016], Weischenfeldt et al. [2017], Beroukhim et al. [2016] show how disruptions in long range DNA looping and genome rearrangements lead to enhancer hijacking and Flavahan et al. [2016] link insulator dysfunctions to oncogene activation in cancer. Thus, changes in chromosome conformation at different scales are now considered as key players in cancer, as well as important potential biomarkers.

Developing accurate and quantitative methods to analyze the chromatin conformation derived from disease-associated cells/tissues is therefore of increasing interest to a wide community of researchers and pathologists. In addition to standard microscopy approaches, several 3C-based methods have been proposed: these rely on digestion

and religation of fixed chromatin to estimate the probability of contact between two genomic loci (see Ramani et al. [2016] for a review). In Hi-C experiments, the contact frequencies between two genomic loci are roughly proportional to the reads counts observed between two regions after sequencing [Lieberman-Aiden et al., 2009]. However, as is the case for many high-throughput technologies, the raw contact frequencies are affected by technical biases such as GC content, mappability, or restriction fragment size [Yaffe and Tanay, 2011]. Estimating and correcting these biases is therefore an important step in ensuring accurate downstream analysis. In the past few years, several methods and packages have been developed to normalize Hi-C data (see Ay and Noble [2015] for a review). These methods fall into two main categories: explicit factor correction methods or matrix balancing algorithms (such as the iterative correction (ICE) method [Imakaev et al., 2012]). In the context of cancer Hi-C data, an additional perturbation related to chromosomal rearrangements must be considered. Amplified genomic regions have a greater chance of being pull-down during the library preparation, while genomic regions with lower copy numbers are more difficult to detect. To date, such copy number variants (CNVs) are usually ignored in cancer Hi-C data normalization, although they raise interesting and important questions both at the biological and methodological levels. The real impact of CNVs on contact frequencies remains difficult to assess. For instance, a tandem amplification does have a very different impact on local chromatin organization compared to the gain of a complete chromosome. Similarly, a genomic duplication could lead to different changes in contact frequencies depending on whether the event occurs within a TADs or across/at its boundary [Franke et al., 2016]. Addressing the question of CNVs during normalization is therefore an important challenge in the analysis of Hi-C data and its interpretation in the context of genetic and epigenetic misregulation in disease.

The question of how copy number signal should be treated mainly depends on the related biological questions. The first strategy is to consider the copy number effect as an unwanted effect, and to remove it during the normalization step [Wu and Michor, 2016]. This strategy indeed makes sense for the detection of a genome-wide list of significant contacts, or for the direct comparison of samples with different chromosomal rearrangement profiles. On the other hand, signal from copy number alterations can also be considered as an important biological information, that can be of interest for 3D modeling, genome reconstruction of cancer cells, or to simply further characterize the genomic landscape of a tumor [Harewood et al., 2017].

Here, we propose to further explore the impact of CNVs on Hi-C data and provide tools that deal with its effect on data normalization. First, we develop a model simulating large copy number rearrangements on a diploid Hi-C contact map. Using such simulated data, we demonstrate that the naive matrix balancing algorithm which is commonly used to normalize Hi-C data, cannot be applied to cancer Hi-C data. We then propose two methods that extend the ICE algorithm and correct the data from systematic biases, either considering the CNVs as a bias to remove or as an interesting signal to conserve in the data structure. Finally, we apply these methods to several disease associated Hi-C data sets, demonstrating their relevance.

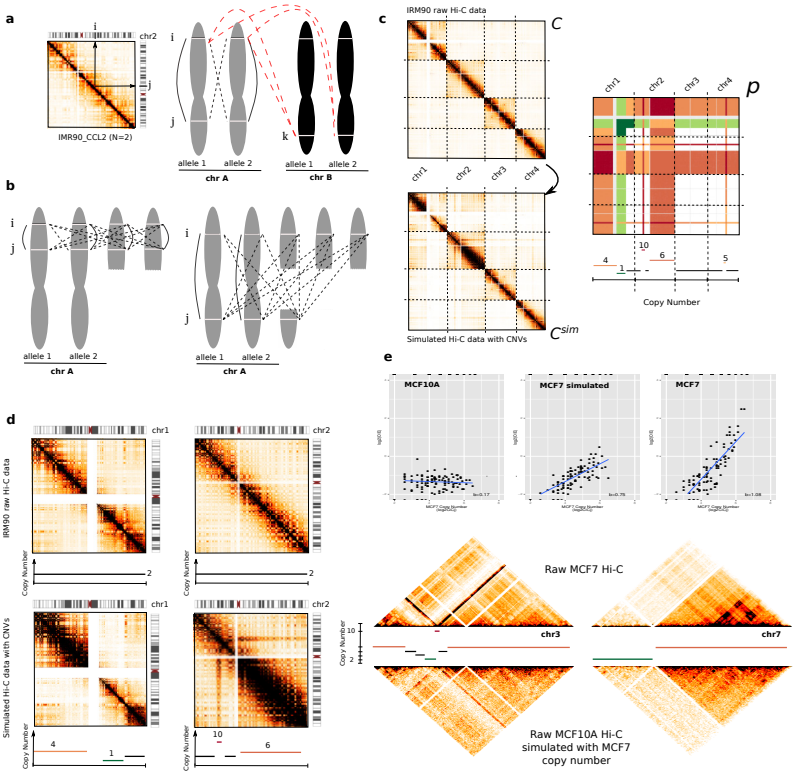


Figure 1: **Simulation of cancer Hi-C data.**
a. In diploid Hi-C data, the contact frequency measured between two loci i and j is equal to the sum of 2 *cis* interactions (black solid lines) occurring within an individual allele and of 2 *trans* interactions between homologous chromosomes (*transH*, black dashed lines). In addition, the contact frequency observed in *trans* between loci i and k is the sum of 4 interactions between non homologous chromosomes (red dashed lines). **b.** In the context of segmental rearrangement, these properties can be extended and generalized if loci i and j belong to the same DNA segment, or to different segments (see Methods and Figure S1). **c.** Simulation of cancer Hi-C data from normal diploid (C) data by calculating the scaling factor matrix (p). Colors in scaling factor matrix represents the level of gains (red) and loss (green) to simulate. For each interaction C_{ij}^{sim}, the simulated count is finally estimated using a binomial downsampling method of probability (see Methods). **d.** Intra-chromosomal maps of chromosome 1 and 2 before (top) and after (bottom) simulation of copy number changes. Copy number effects are characterized by blocks of high/lower signal. Overall, the simulation conserves the structure and the counts/distance properties of the Hi-C maps. **e.** Validation of the simulation model using Hi-C data from MCF10A cell line from which we simulated the expected copy number of MCF7 cancer cell line. We observed a positive correlation between the raw log2 O/E (Observed/Expected) ratios and log2 multiplicative copy number in 1 Mb resolution Hi-C maps, on both simulated and real MCF7 Hi-C data. Looking at the intra-chromosomal maps of chromosomes 3 and 8 demonstrates that our model efficiently simulates large copy number events.

Results

Simulating the effect of copy number variations on Hi-C data

Due to the large number of genomic and epigenomic factors possibly involved, predicting the true effect of copy-number variations on the 3D organization of the genome is challenging. We propose a simple mathematical model to simulate the effect of abnormal karyotypes on a diploid Hi-C data set by estimating the enrichment in interaction due to copy number variation (see Methods and Figure 1).

In order to validate our simulation model, we leverage available Hi-C data from two epithelial cell lines: the MCF7 breast cancer cell line and the MCF10A nearly diploid, non-tumorigenic cell line [Barutcu et al., 2015]. We extract MCF7's copy number information from Affymetrix SNP6.0 array, filtering out any altered segments lower than the MCF10A's Hi-C map resolution (1 Mb) and apply our simulation model on the normal-like data, thus obtaining a simulation of MCF7's abnormal karyotype. We then compare our simulated results with the real MCF7 Hi-C data set. As expected, the contact counts (for both the simulated data and the real data) are correlated with the copy number (Figure 1e). Both our simulations and the real data show blocks of higher/lower contact frequencies in regions affected by large copy number variants. In average, the intra-chromosomal maps of simulated and real MCF7 data have a Spearman correlation of 0.70 (0.54-0.84). We then summarize both data in 1D by summing the contact frequencies over each row. Overall, the simulated MCF7 profile is well correlated with the profile of real MCF7 Hi-C data (Spearman cor=0.88, Figure S3). We can observe that the sum of interactions for a genomic window is proportional to the copy number. This is expected, as a genomic region in multiple copies have a greater chance of being seen interacting with another region. Interestingly, for highest copy number, both profiles increase concurrently, but not at the same rate (Figure 1e). One explanation would be that these regions of very high copy number correspond to tandem focal amplifications. Their linear proximity on the genome would therefore explain the massive increase of contact frequencies that we observed in real data, and which are not modeled by our simulation.

Overall, those observations leads us to believe our simulation method appropriately models the effect of copy number variations on Hi-C data.

The ICE normalization is not suitable for cancer Hi-C data

We then apply this simulation model to assess the ability of the ICE normalization method to correct for copy number variations. We simulate two data sets with different properties from the publicly available Human IMR90 Hi-C data [Rao et al., 2014]. First, we generate a highly rearranged data set, with segmental gains, losses and a focal amplification up to 10 copies (Figure 1c). Then, we simulate a case of aneuploidy with gain or loss of entire chromosomes (Figure S2a). While the simulation is performed genome-wide, we restrict the CNVs to the first chromosomes to ease the results interpretation and visualization.

Several methods have been proposed to remove unwanted technical and biological variations from Hi-C data, none of which are adequate for abnormal karyotype data. These methods fall into two groups. The first explicitly models sources of biases, and cast the normalization procedure as a regression problem [Yaffe and Tanay, 2011, Hu et al., 2012]. The second group leverages a small number of hypothesis on the

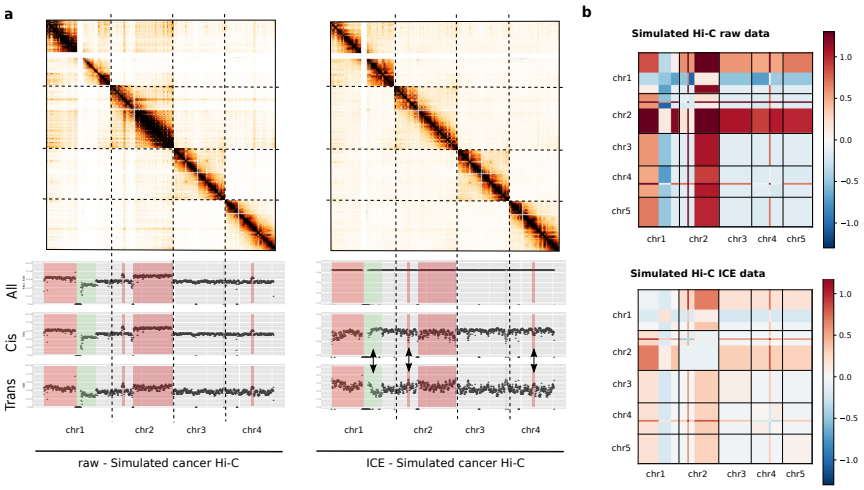


Figure 2: **Impact of matrix balancing normalization on simulated cancer Hi-C data.**
a. Simulated Hi-C contact maps (500 kb resolution) of the first four chromosomes and contact frequencies presented as the sum of genome-wide contacts per locus, using either all (inter and intra-chromosomal), *cis* (intra-chromosomal) or *trans* (inter-chromosomal) contacts. Rearranged regions are highlighted in red (gain) or green (loss). The 1D profile of ICE data is constant genome-wide as expected under the assumption of equal visibility. However, the iterative correction on simulated cancer data results in an shift of contacts between altered regions (arrows). **b.** Block-average error matrix of simulated raw and ICE cancer data (150 Kb resolution) (See supp method 1.4). The iterative correction does not allow to correct for segmental copy number bias.

bias and on the properties of Hi-C data to formulate the normalization procedure as matrix-balancing problems: these do not assume any specific sources of biases, and are (as long as the hypothesis are fulfilled) able to correct for any factors affecting contact frequencies [Imakaev et al., 2012, Cournac et al., 2012]. Among those methods, the iterative correction method (ICE, Imakaev et al. [2012]) has been successfully applied to many diploid Hi-C data sets. ICE relies on two assumptions: (1) the bias between two regions i and j can be represented as the product of individual biases of these regions : $N_{ij}^{\mathrm{ICE}} = \beta_i \beta_j C_{ij}$; (2) each bin should interact approximately the same number of times: $\sum_i N_{ij}^{\mathrm{ICE}} = k$, where C represents the raw count matrix, N^{ICE} the ICE normalized count matrix, β the bias vectors and k a constant.

In this section, we explore the effects of ICE on Hi-C data in the presence of copy-number variations. To do this, we leverage the simulated data set previously described (Figure 1c, Supp. Figure S2a) for which the ground-truth normalized data is found by applying ICE to the original diploid data. We are thus able to assess the performance of ICE to correct for unwanted sources of variation, including the copy number, by comparing the obtained matrices to the ground-truth.

Before normalizing the data with ICE, we first represent the data in 1D, by summing each row of the matrix. As previously mentionned, the sum of genome-wide interactions per bin is proportional to the copy number (Figure 2a). After applying ICE, each genomic region now interacts the same number of times genome-wide, as expected. However, ICE leads to an imbalance between *cis* and *trans* contact counts. As shown on Figure 2b, we observe that the *cis* contact counts are now depleted for regions with high copy number, and *trans* contact counts are enriched. On the other hand, lost regions now present higher contact probabilities than gain regions in *cis*. The same conclusions can be made in the context of aneuploidy (Figure S2). However, we notice that in this case, ICE could yield to the expected results if the analysis is restricted to intra-chromosomal contacts. We thus conclude that ICE is not adapted to correct for segmental copy number effect, and that, more importantly, it can lead to a misinterpretation of the contact probability between rearranged regions.

If the downstream analysis is restricted to intra-chromosomal interactions, one may ask whether applying ICE independently to each intra-chromosomal maps could mitigate the introduction of biases. We therefore independantly normalized by ICE all intra-chromosomal maps. However, although the effects are less strong, we observe the same phenomenon in complex rearrangements (Figure S4).

Altogether, these results demonstrate that ICE does not properly normalize data with abnormal karyotype and that its use is therefore not recommended in the context of cancer Hi-C data.

LOIC: a novel normalization strategy for cancer Hi-C data

As discussed above, ICE relies on the assumption of equal visibility of each genomic bin. In the presence of copy number variations, this assumption does not hold: genomic bins with higher copy number variations will interact overall more than genomic bins of lower CNVs. In addition, the copy number effect between loci i and j (B_{ij}), cannot be decomposed as the product of an effect in loci i and an effect in loci j, therefy also violating the ICE hypothesis. Instead, we propose to extend the ICE model, saying that the assumption of equal visibility remains true across regions of identical copy number. In addition, biases associated to fragments (such as fragment length, GC-content, or mappability) are still decomposable into the product of two region specific biases.

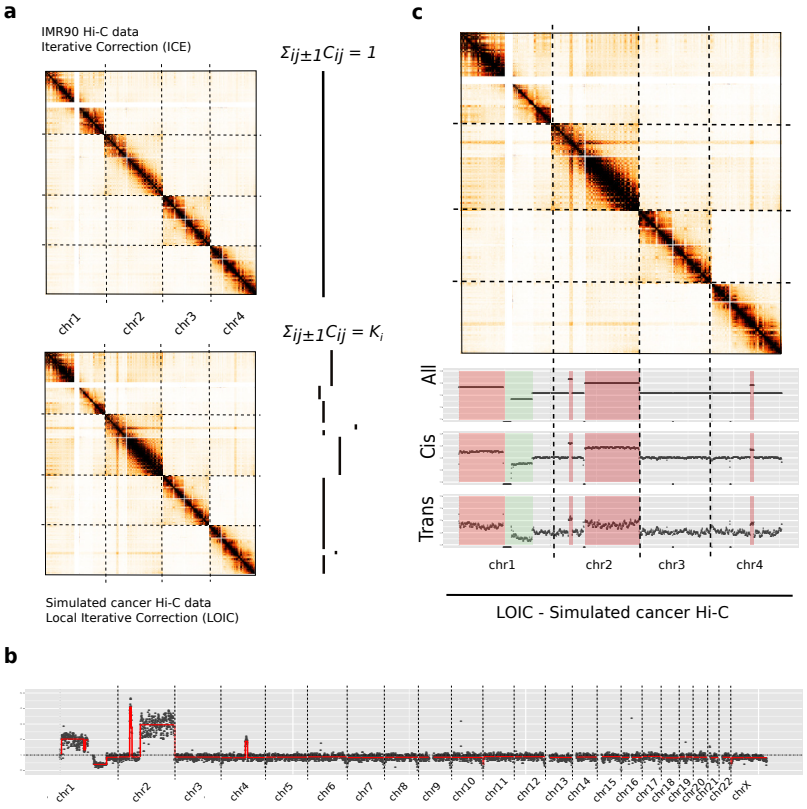


Figure 3: **Generalization of matrix balancing algorithms for cancer Hi-C data.**
a. Rationale of LOIC method versus standard ICE method. The LOIC method extends the ICE normalization by constraining the genome-wide Hi-C 1D profile to follow the copy number signal. **b.** Segmentation of the Hi-C 1D genome-wide profile of simulated cancer data. The red line represents the smoothing line that estimate the copy number level. **c.** LOIC normalized Hi-C contact maps of simulated data on the first four chromosomes. The 1D profiles are represented by the sum of genome-wide contacts at each locus using either all (inter and intra-chromosomal), *cis* (intra-chromosomal) or *trans* (inter-chromosomal) contacts. As a results, we can see that the LOIC method allows to normalize cancer Hi-C data keeping into account the copy number information.

We thus first propose to extend ICE by assuming that the sum of contacts for a given genomic bin is constant across genomic bins of identical copy number (Figure 3a) : $\sum_i C_{ij}^{LOIC} = k_j$, where k_j is the interaction profile associated to the copy number of j (see Methods). We refer to this method as a local iterative correction (LOIC). When there is no copy number aberration, LOIC solves exactly the same problem as ICE.

The LOIC procedure requires to identify DNA segments of equal copy number. External sources of data (such as genome sequencing or microarray data) can be used to infer DNA breakpoints along the genome, and thus to define the DNA segments of equal copy number. If such data is not available, we propose to directly infer the copy number from the Hi-C data (see Methods). Applying our method to the simulated data yields a copy number estimation well correlated with the profile used for simulation (Spearman cor=0.64) (Figure 3b). To further validate this step, we apply the segmentation procedure to the IMR90 diploid data set. We obtain a nearly uniform copy number profile (Figure S6c). Interestingly, we frequently observe a decrease of contacts at telomeric regions which can therefore leads to a breakpoint in the segmentation. This telomeric pattern is expected, even in a diploid sample, as the assumption of equal visibility in these regions can be discussed.

We then apply the LOIC procedure to our highly rearranged simulated Hi-C data set using the breakpoint positions estimated by our segmentation procedure (Figure 3c). As expected, we observe that the genome-wide sum of contacts of each bin is proportional to the copy number, and that bins within a DNA segment are normalized to the same level of interactions. The previous effects observed on *cis* and *trans* sum of contacts with the standard ICE strategy no longer hold true. In addition, we calculate the effective fragment length, the GC content and the mappability features for each 500 Kb bin as already proposed [Hu et al., 2012], and then represent the average contact frequencies among those genomic features. Despite a few local enrichment due to CNVs, we observe that the LOIC normalization is as effective as the ICE normalization to correct for GC content, effective fragment length and mappability (Figure S5). We then turn to the aneuploid simulated data set. In the latter case, the LOIC allows to conserve the inter-chromosomal scaling factor due to CNVs. Differences in intra-chromosomal maps between ICE and LOIC remains negligible in this case and are related to the segmentation profile (Figure S6a,b).

To conclude, the LOIC strategy can be seen as a generalization of the ICE method. In that sense, applying both methods to a diploid data set leads to identical results.

CAIC: estimating and removing the copy-number effect on cancer Hi-C data

In addition, we also propose to estimate and to correct the effect introduced by copy number changes. We assume that the copy number effect can be represented as a block-constant matrix where each block is delimited by a copy number change (see Methods). In addition, we assume that, on average, each pair of loci interacts the same way as any pair of loci at the same genomic distance s. In summary, the raw interaction count C_{ij} is roughly equal to the product of a CNV bias B_{ij} and the expected contact count at genomic distance s: $C_{ij} \simeq B_{ij} e_{s(i,j)}$. We thus cast an optimization problem to find the CNV block biases B and the expected contact count at genomic distance s (see Methods). We refer to this method as CNV-Adjusted Iterative Correction (CAIC).
We apply the CAIC normalization to the two simulated data sets. Looking at the 1D

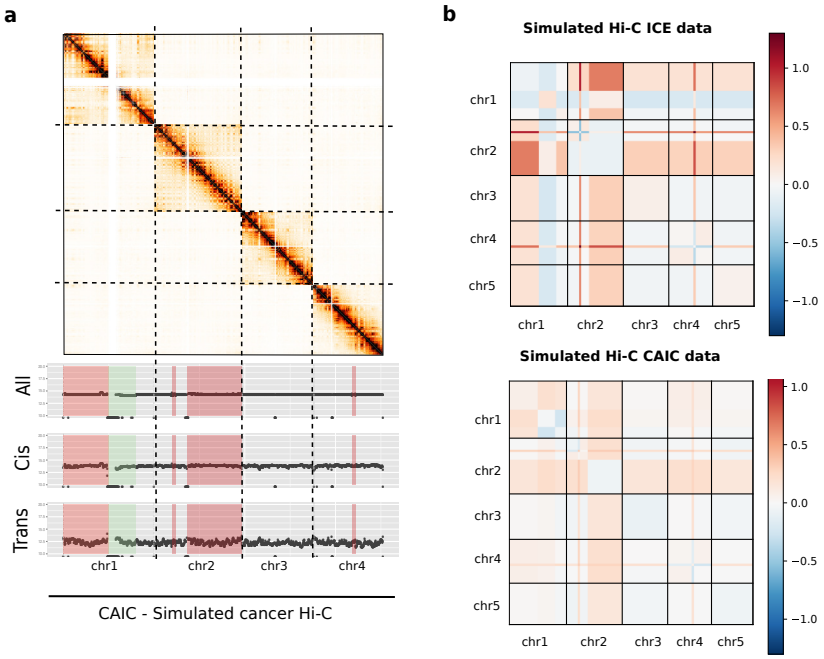


Figure 4: **CNV-adjusted normalization of cancer Hi-C data.** **a.** Hi-C contact maps of the four first chromosomes of our highly rearranged simulated data, together with the 1D signal of all, cis and trans data. Regions in red and green correspond to simulated gain and loss. **b.** Block-average error matrix of simulated ICE and CAIC Hi-C data. The CAIC efficiently removed the CNV effect, whereas the ICE normalization does not allow to correct for its effect.

signal of the CAIC normalized data using the *cis* and *trans* data validates that the method tends to remove the CNV effect (Figure 4a). In addition, the unbalanced effect that we previously observed with the ICE normalized data disappears. We then divide the normalized contact matrices by the expected count matrices, thus removing the structure due to genomic proximity. Taking the average per block, we observe that the expected CAIC matrices are much more uniform than the ICE normalized matrices (Figure S7 and S8). On the aneuploid simulated data set, it is worth noting that ICE and CAIC yield very close results. We then compare the normalized contact maps to the ground-truth by computing the error matrix as well as three additional error measures (see Figure 4b and Supp Methods 1.4). We observe that the copy number effect is well removed both on the aneuploid ($\ell_1 = 2.291 \times 10^6$, $\ell_2 = 3.532 \times 10^5$ and $\ell_{\max} = 0.185$) and highly rearranged data set ($\ell_1 = 2.055 \times 10^6$, $\ell_2 = 2.536 \times 10^5$ and $\ell_{\max} = 0.360$).
Altogether, those results demonstrate the CAIC normalization procedure effectively removes copy-number effect.

Application to breast cancer Hi-C data

A number of studies performed Hi-C experiments on cancer samples or cell lines [Barutcu et al., 2015, Taberlay et al., 2016, Le Dily et al., 2014]. We further explore the effect of our normalization procedures on two previously published Hi-C data from breast cancer cell lines: T47D [Le Dily et al., 2014] and MCF7 [Barutcu et al., 2015]. We process the T47D and MCF7 samples from raw data files to raw contact maps using the HiC-Pro pipeline [Servant et al., 2015]. As already seen in our simulation data (Figure 2a), we observe a strong copy number effect on the raw contact maps with respectively higher/lower contact frequency on gained/lost DNA regions in the T47D and MCF7 samples (Figure 5, Figure S9). Applying ICE on these data set does not remove entirely the copy number effect, and tends to flip the coverage profile between gained/lost regions in *cis*, therefore validating our previous observations on simulated data.

In order to estimate the copy number signal from these cell lines, we segment the 1D Hi-C profile as previously described. Interestingly, on both T47D and MCF7 data we observe a very good correlation between the copy number signal extracted directly from the Hi-C data and the copy number profile extracted from SNP6 Affymetrix array (Figure 5b, Figure S9b, Spearman cor=0.87 for both MCF7 and T47D data) We then apply the LOIC strategy as presented above so that the sum of each column/row follows the segmentation profile extracted from the data. As expected, we observe that the LOIC normalized contact maps conserves the copy number properties, and that the biases introduced by the ICE normalization no longer hold true. In addition, we also applied the CAIC normalization to correct of the copy number signal. Looking at the correlation between Hi-C counts and the copy number signal validates the efficiency of the methods (Figure 5c, Figure S10). In conclusion, applying the LOIC and CAIC methods on both cancer data set allows us to correct for systematic bias while conserving or removing the copy number structure.

Normalization of capture-Hi-C data with genomic duplication

In addition to cancer data, we also investigate the relevance of LOIC to normalize Hi-C data in the presence of local structural events. Recently, Franke et al. [2016]

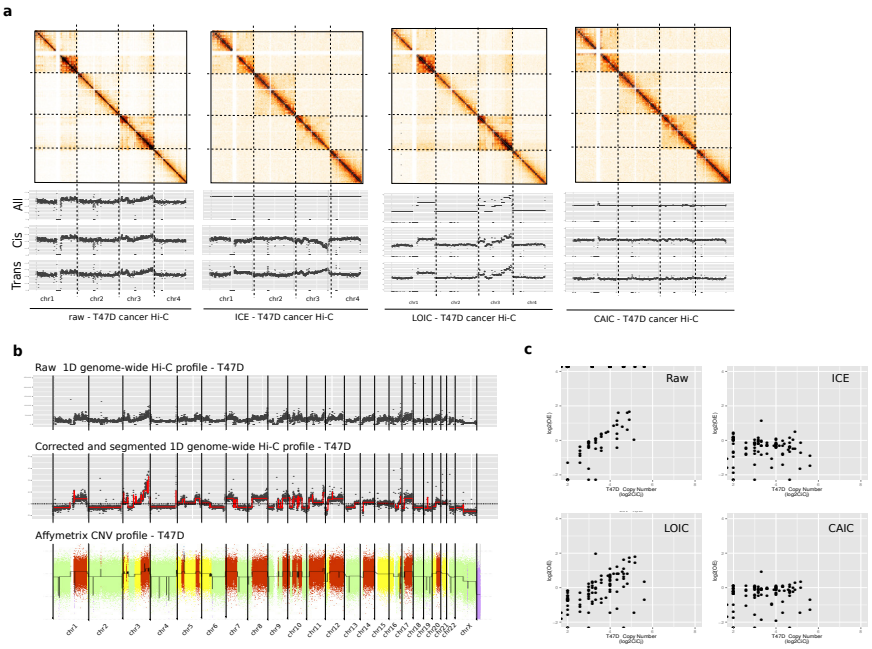


Figure 5: **Normalization of T47D Hi-C data.**
a. Hi-C contact maps (250 Kb resolution) of the first four chromosomes of T47D cancer Hi-C sample. When looking at the 1D *cis* and *trans* profiles, we observed that ICE introduces a bias in the normalized data, therefore validating the observation made on the simulated data. We then applied the LOIC and CAIC normalizations in order to efficiently correct the data from systematic bias, while removing or keeping the CNVs effect.**b.** Estimate the copy number signal from the Hi-C data after correction and segmentation of the 1D profile. The inferred copy number signal from the Hi-C data are highly correlated with the copy number profile from Affymetrix SNP6.0 array. **c.** Correlation of raw and normalized contact frequencies with the copy number.

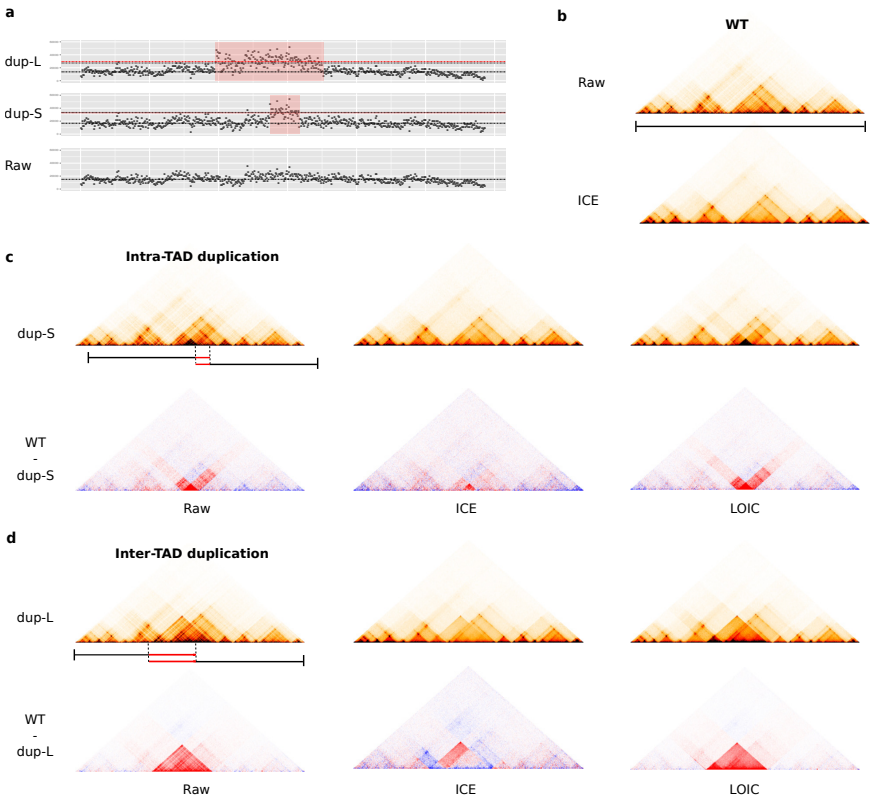


Figure 6: **Duplication in capture-Hi-C and normalization.** **a.** 1D profiles of capture-Hi-C wild-type sample (WT), with intra-TAD duplication (dup-S) or with inter-TAD (dup-L) duplication Franke et al.. As expected, the duplication samples are characterized by twice more contacts at the duplicated sites. **b.** Raw and ICE normalized contact maps of WT sample. **c.** Normalization of the dup-S sample with the ICE and LOICE methods. The duplication effect is visualized by subtracting the normalized WT and dup-S contact maps. **c.** Same approach applied to the dup-L sample.

investigated the effect of local duplications on chromatin structure and, in particular, on the formation of new topological domain. We therefore process the capture Hi-C data of the Sox9 locus and generate the raw and ICE contact maps at 10kb resolution. We focuse our analysis on samples with inter-TAD (dup-S) and intra-TAD (dup-L) duplications. In the context of inter-TAD duplication, Franke et al. [2016] described the formation of a new domain in the duplicated region by comparing the raw contact maps of wild type (WT) and dup-L samples (Figure 6c-d). Interestingly, when we compare the WT sample and the samples with the duplication events normalized by the ICE method, we observe that the ICE normalized maps do not allow duplication effects to be observed clearly. This is in agreement with our previous observations on cancer Hi-C data. We then apply our LOIC strategy following the observed 1D coverage profiles. As illustrated in Figure 6c, the LOIC normalization allows to remove systematic biases from the contact maps while keeping the copy number effect. Normalized maps allow to clearly observe the effect of both intra and inter-TAD duplication, validating the interest of the method to study local structural rearrangements.

Removing the CNVs signal avoids misinterpretation of the chromosome compartment calling of cancer Hi-C data

We then further explore the impact of CNVs and normalization on the chromosome compartment calling. Looking at intra-chromosomal contact maps, the chromosome compartment profile appears as a checker-board-like interaction pattern, shifting from blocks with either high and low interaction frequency. Thus, chromosome compartments are usually detected using a Principal Component Analysis (PCA) on the correlation matrix of the distance-corrected intra-chromosomal contact maps. The first principal component then distinguishes the active (A) from inactive (B) compartments [Lieberman-Aiden et al., 2009] (see Supp Methods).

We perform this compartment calling analysis on the MCF7 Hi-C data normalized by ICE, LOIC, or CAIC methods and integrate the results with the histone marks data obtained from the ENCODE project [Dunham et al., 2012]. Surprisingly, we observe that, for most chromosomes, the compartment calling is not affected by the CNVs (Figure S12, Figure S13, Figure S14).

We then assess how the A/B compartments correlate with the active and repressive histone marks genome-wide (see Supp Methods). We observe that active compartments are associated with open-chromatin marks such as H3K27ac, H3K36me3 and H3K4me, and this, whatever the normalization method used. Respectively, inactive compartments are associated with repressive marks such as H3K27me3 or H3K9me3 (Figure 7a). In addition, we observed that the A-type and B-type compartments are very similar regardless the normalization used.

However, looking at each chromosome independantly shows that, in the MCF7 data, the chromosome 8 harbours distinct compartment patterns according to the normalization method used. In this case, it is clear that the copy number affects the PCA analysis and the compartment calling (Figure 7b,c). We therefore conclude that it is important to correct for the copy number effect before running such analysis. Applying the CAIC normalization method outperforms the other methods, allowing to efficiently detect the A/B compartments on chromosome 8. The compartments pattern of the chromosome 8 extracted from the ICE normalized data is concordant with our previous conclusion that ICE is not appropriate to correct for CNVs, potentially leading to a wrong interpretation of the compartment profile. However, we also notice

that, in this case, the A/B compartments can be rescued by looking at the second principal component of the PCA.

Altogether, these results demonstrate that, although the compartment calling seems globally not affected by the copy number effects, applying the CAIC strategy to normalize the data improves the detection of A/B compartments and avoids potential issues in their interpretation.

Discussion

Chromosome conformation techniques are promising exploration tools to investigate links between the three-dimensional organization of the genome and functional and phenotypical effects in diseases. However, in the context of genomic rearrangements, dedicated methods should be applied as a novel source of signal related to copy number can arise. Predicting the exact effect of copy number variations on chromosome structure is a challenging task, as many factors influence the 3D structure of the genome and the resulting contact frequencies.

In this paper, we propose a simulation model to explore the effect of large copy number on Hi-C data. Starting from a diploid data set, our model is able to predict the effect of large copy number changes on interaction patterns. Although it represents a powerful tool to assess the ability of a method to deal with the copy number changes, it also demonstrates that predicting the real effect of copy number changes is very challenging. As an example, our model considers the duplicated regions as non-tandemly rearranged events, which, in some cases, certainly underestimates the intra-chromosomal effect of copy number. In addition, it does not integrate any biological knowledge and is therefore not designed to simulate changes due to the alteration of regulatory elements such as insulator regions. We also note that the sub-sampling strategy that we applied requires a high resolution diploid Hi-C data set in input. However, as the Hi-C sequencing depth is still increasing, this should no longer be a limitation.

Using our simulation model, we then demonstrate that applying ICE to data sets with abnormal karyotypes leads to unbalance corrections between amplified and lost regions. We further validate this observation using two Hi-C breast cancer data set.

We therefore propose two new normalization methods, respectively able to conserve or to remove the copy number, while correcting the data from other systematic biases. In addition, we also propose a segmentation procedure to directly extract the copy number signal and the breakpoints from the Hi-C contact probabilities and to use this information in the normalization process. Although this step is crucial for the normalization and can be challenging for noisy samples, our procedure performs well on all the data set used in this study, including the normal diploid samples.

The choice of the normalization method mainly depends on the context and biological questions. As an example, we demonstrate than, although the chromosome compartments calling and the PCA analysis are not dramatically affected by the copy number changes on MCF7 data, it can also lead to wrong interpretation if the data are not properly normalized. From our experience, this effect can be more important on other cancer types and mainly depends on the copy number profile of the tumor (data not shown). We therefore demonstrate the interest if our CAIC normalization method to remove the copy number biases, and to improve the reliability of the compartment calling.

In addition, we also illustrate the interest of keeping the CNVs information when

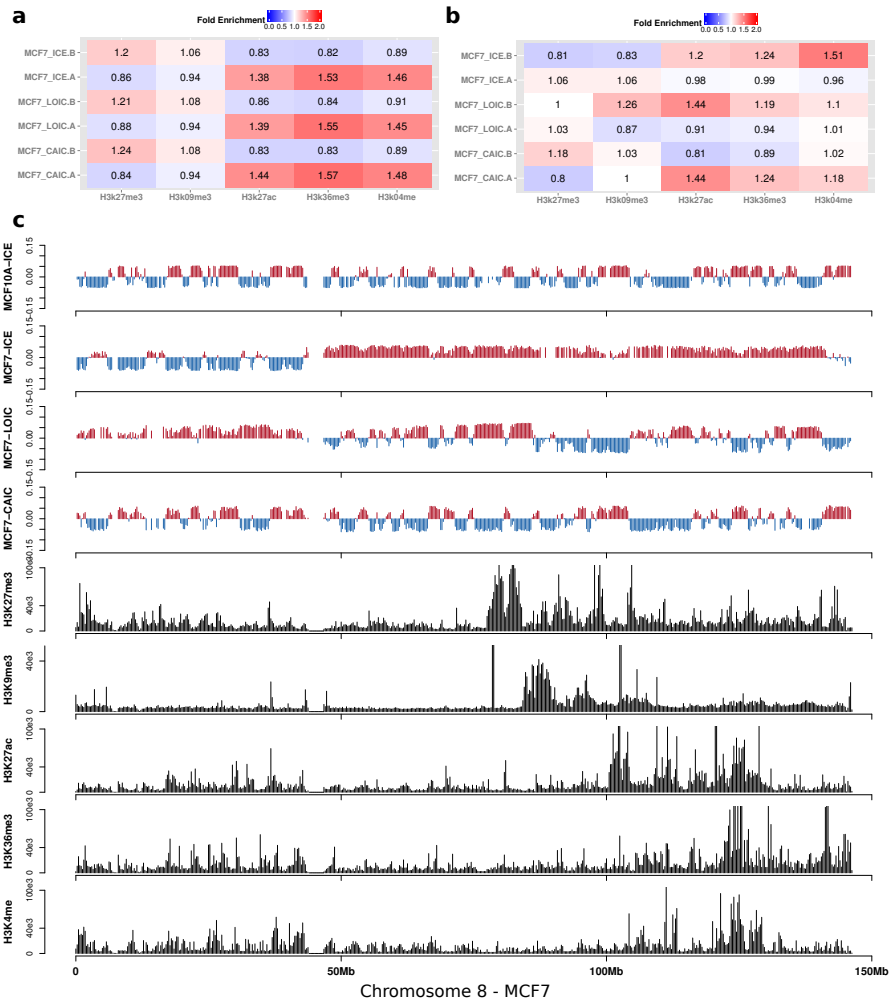


Figure 7: **Detection of chromosome compartments.**
a. Genome-wide enrichment of ChIP-seq histone marks in active (A) or inactive (B) compartments. Active compartments are enriched in open-chromatin marks, whereas inactive compartments are enriched in repressive marks. The results are concordant genome-wide, whatever the normalization method applied. **b.** Histone marks enrichment on chromosome 8 of MCF7 sample. On this chromosome, the copy number has a strong impact on the compartment calling. **c.** Results of the compartment calling for the chromosome 8 of MCF7 sample (first principal component), together with normalized ChIP-seq tracks. Active domains are in red. Inactive domains in blue.

looking at local structural rearrangements. In this context, we demonstrate that our LOIC method can easily be applied to properly remove the other systematic biases while keeping the copy number structure.

Taken together, our analysis highlights the importance of using dedicated methods for the analysis of Hi-C cancer data. It therefore paves the way to further explorations of the three-dimensional architecture of cancer genomes.

Methods

Let us first introduce some notations. Given a segmentation of the genome into n genomic windows (or bins), Hi-C data can be summarized by a n-by-n symmetric matrix C, in which each row and column corresponds to a specific genomic loci and each entry C_{ij} the number of times loci i and j have been observed in contact. Let $K \in \mathbb{R}^n$ the copy number profile of the sample of interest, which we represent as a piecewise constant vector.

We denote by $s(i,j)$ the genomic distance between the loci, defined as the number of base pairs between the center of the two loci; if i and j are not part of the same chromosome, we extend this definition by setting $s(i,j) = \infty$. In this paper, we derive different ways to normalize the raw count matrix C: we denote by N^y the contact count matrix normalized with method y (*e.g.* N^{ICE} represents the ICE normalized count matrix).

Simulation of cancer Hi-C data

Before we turn to how to appropriately model cancer Hi-C data, let us first review some terminology. In the literature, *cis*-contact counts refers to the contact counts between two loci of the same chromosomes: this includes intra-chromosomal contact counts but also inter-chromosomal contact counts of homologous chromosomes. In this paper, we restrict the use of *cis*-contact counts to contact counts issued from the same DNA fragment, and we denote by "*trans*-homologous" (*transH*) interactions, the interactions between homologous chromosomes. Note that *cis* and *transH* contact counts are mostly indistinguishable (with the exception of allele-specific Hi-C) in Hi-C data hence the simplification of terminology usually used.

We now return to the problem at hand: how to simulate a contact count matrix C^{sim} of a cancer genome with abnormal copy number from a raw diploid contact count matrix C. In order to model the change in contact count abundances due to copy number variation, we first need to understand precisely which interactions are observed in the case of a simple diploid genome. For that purpose, we denote by E_{ij} the expected contact count between loci i and j, and E_{ij}^{cis}, E_{ij}^{transH} and E_{ij}^{trans} the expected *cis*, *trans* and *transH*-contact counts between i and j.

- if loci i and j belong to the same chromosome, the expected contact count E_{ij} is the sum of (1) *cis*-counts from either of the homologous chromosomes; (2) the *transH*-counts between the two homologous chromosomes:

$$E_{ij} = 2E_{ij}^{cis} + 2E_{ij}^{transH}$$

- if loci i and j belong to different chromosomes, then the observed contact counts E_{ij} is the sum of either of the four possible *trans* interactions:

$$E_{ij} = 4E_{ij}^{trans}$$

This can be generalized to polyploid genome or to the context of chromosomal abnormalities (Figure S1).

- if loci i and j belong to the same chromosome, let k be the number of *cis* interactions. If i and j belong to the same DNA segment, $k = K_i = K_j$. When i and j belong to different DNA segments, k could in theory take values between $[0 - min(K_i, K_j)]$. Here, we simulated the data with $k = 2$, or $k = min(K_i, K_j)$ if $K_i < 2$ or $K_j < 2$. Then

$$E_{ij} = kE_{ij}^{cis} + (K_iK_j - k)E_{ij}^{transH}$$

  .

- if loci i and j belong to different chromosomes, then

$$E_{ij} = K_iK_jE_{ij}^{trans}$$

Now that we have derived how contact counts are decomposed in terms of *cis*, *transH* and *trans* contact counts, we can leverage those relationships to simulate the effect of copy number variations on contact count matrices.

In order to derive a scaling factor p_{ij} which incorporates the copy number effect, we need to estimate E_{ij}^{cis}, E_{ij}^{trans} and E_{ij}^{transH}, which is impossible without further assumptions. In practice, little is known about the probability of contact between homologous chromosomes, which is therefore difficult to estimate. However, we know that the chromosomes usually occupy their own space (chromosome territories) within the nucleus. We therefore make the assumption that all chromosomes are independent, and that the contact probability between homologous chromosomes can be estimated using the *trans* interaction between non-homologous chromosomes. We thus consider that $E_{ij}^{trans} = E_{ij}^{transH} = E^{trans}$.

From these relationships, we then calculate the scaling factor p_{ij} as following (recall that E_{ij} is the expected copy number between i and j on genome with abnormal chromosomal interactions):

- we estimate E^{trans} as the median *trans*-contact count;
- $E_{ij}^{cis} = C_{ij} - E^{trans}$;
- E_{ij} is estimated using the equations derived above;
- if loci i and j belong to the same chromosome, $p_{ij} = \frac{E_{ij}}{2E_{ij}^{cis}+2E^{trans}}$
- if loci i and j belong to different chromosomes, $p_{ij} = \frac{K_iK_j}{4}$

We thus obtain, for each entry of the contact count matrix C_{ij}, a ratio p_{ij} corresponding to the expected factor of enrichment or depletion of interactions for the loci i and j. In order to make the estimation of p_{ij} more robust, we estimate it constant per blocks of identical copy numbers by taking the median of the empirical values in each block. (See Figure S11). Thus, the factor matrix p can be assumed to be block constant between regions of identical copy number variations. We thereby smooth p by computing the median scaling factor of block of similar copy number.

Finally, the simulated contact counts C_{ij}^{sim} are generated by a binomial subsampling strategy of C_{ij} by a probability equal to $\frac{p_{ij}}{\max(p_{ij})}$ [Wiuf and Stumpf, 2006]:

$$C_{ij}^{sim} \sim B(C_{ij}, p_{ij}) \tag{1}$$

The reason for choosing a binomial subsampling as opposed to a simpler multiplication of the original Hi-C counts by a CNV-dependent factor, is that if C_{ij} follows a Poisson or Negative Binomial distribution, then C_{ij}^{sim} follows the same distribution with modified expectation [C.Wiuf and P.H.Stumpf]. One limitation of this model is that the simulated counts can only be smaller than the original counts, which may be problematic if we start from small counts. It thus requires a diploid Hi-C data set with a sufficient sequencing depth to apply the downsampling strategy.

LOIC: Correcting technical biases of Hi-C cancer data

To normalize the contact count matrix, we adapt the ICE method proposed by Imakaev et al. [2012] to incorporate the copy number effect. In particular, we use similar assumptions. First, the bias between two regions i and j can be decomposed as the product of two region-specific biases $\beta_{ij} = \beta_i \beta_j$.

$$\mathbf{E}C_{ij} = \beta_i \beta_j N_{ij}^{\mathrm{LOIC}}, \tag{2}$$

where $\beta \in \mathbb{R}^n$ is a vector of bin-specific *biases*, such as gc-content, fragment lengths, mappability, etc.

Second, all copy-number identical regions interact as much: $\sum_{i=1}^{n} N_{ij}^{LOIC} = \frac{1}{|\{l|K_l=K_i\}|} \sum_{l|K_l=K_i} C_{lm}$. We refer to the second hypothesis as the "local equal-visibility assumption" to contrast it with ICE's "equal-visibility assumption": instead of enforcing an interaction profile constant across all the genome, we enforce an interaction profile constant for regions of identical copy number number.

Similarly to Imakaev et al. [2012], this problem can be solved exactly using matrix-balancing algorithms (under the assumption that the matrix is full decomposable [Sinkhorn and Knopp, 1967]). Note that if there is no copy number variations, this boils down to solving exactly the same problem as ICE. On the other hand, in the presence of copy number variation, the resulting interaction profile will be a constant piecewise function, whose value depends on the copy number of the two loci.

In order to apply the proposed method, one needs to know *a priori* the set of bins with a given copy number or the copy number breakpoints. It can either be found via probing the samples to estimate it using specific technologies or through prior knowledge on the cell-line or sample studied. When none of these options are available, we can leverage the information provided by Hi-C data directly to estimate it.

Estimation of copy number from the contact count matrix

The copy number signal can be directly inferred from the Hi-C data in two steps. We first calculate the one-dimensional (1D) signal as the sum of genome-wide contact per bin, assuming that this signal reflects the true contact frequencies including the systematic Hi-C biases and the CNVs signal. We further calculate the GC content, the mappability and the effective fragment length of each bin end as already proposed [Hu et al., 2012]. The local genomic features of all chromosome bins are defined as the average of the corresponding features among all overlapping fragment ends. We then apply a Poisson regression model to correct the signal from GC content, mappability and fragment length, using the model proposed by Hu et al.[Hu et al., 2012]. The corrected profile is obtained by subtracting the fitted values to the observed data, and rescaled to be centered on 1. The normalized 1D data are then segmented using a pruned dynamic programming algorithm [Picard et al., 2011]. The segmented profile

is smoothed with the GLAD package [Hupé et al., 2011] in order to optimize the breakpoint locations and to remove false positives events. The segmentation is an important step of the method which may need to be adjusted according to the signal-to-noise ratio of the data. In this study, we apply the same parameters to all data sets and we consider the smoothed line after the segmentation as the Hi-C derived copy number profile.

CAIC: Removing the copy number effect

The previous section describes how to normalize the raw contact counts matrix C to adjust for unwanted variations such as GC-content, mappability, fragment lengths, while keeping the copy number information. We now propose to estimate the effect of copy number variations on the contact count matrix to offer the possibility of removing it. We denote by N^{CAIC} the normalized contact count matrix where the CNV effect has been removed.

We assume that the copy number effect for each pair of loci is identical for element with identical copy-number variations. This reflects that the copy-number effect between loci i and j is related to the amount of genetic material of those two regions, and thus identical between all pairs with similar copy-number variations. We can thus model the normalized contact count matrix as the product of a block-constant matrix B and corrected matrix N^{CAIC}: $N^{CAIC}_{ij} = B_{ij} N^{\mathrm{LOIC}}_{ij}$, where each block is a function of the copy number in i and in j.

In addition, we assume that, on average, each pair of loci interacts roughly the same way as any pair of loci at the same genomic distance s:

$$N^{CAIC}_{l,m} \simeq e(s(l,m))\,, \tag{3}$$

where $e(s)$ is the expected contact count at genomic distance s. We leverage this assumption to cast an optimization problem:

$$\begin{aligned} \min_{e,B} \quad & \sum_{i,j} (N^{LOIC}_{ij} - B_{ij}\, e(s(i,j)))^2 \\ \text{subject to} \quad & B \quad \text{is block-constant} \\ & e \quad \text{decreasing} \end{aligned}$$

We solve this optimization problem genome-wide by iteratively estimating the block constant matrix B and the expected counts function e using an isotonic regression. Note that the *trans* estimation can be done jointly on the whole genome independently from the *cis* estimation.

Availability of supporting data and code

All codes and data are available at https://github.com/nservant/cancer-hic-norm

Authors' contributions

NS, NV, EH, JV and EB designed the project. NS, NV and JV designed the simulation model, and the concepts of the normalization methods. NV implemented the

normalization approaches. NS performed Hi-C data analysis. All authors read and approved the final manuscript.

Fundings

This work was supported by the Labex Deep, the Ligue Contre le Cancer, the European Research Coucil (SMAC-ERC-280032), the ERC Advanced Investigator award (ERC-250367), the ABS4NGS project (ANR-11-BINF-0001), the Gordon and Betty Moore Foundation (Grant GBMF3834) and the Alfred P. Sloan Foundation (Grant 2013-10-27).

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We would like to thank D. Gentien and the Institut Curie genomic platform for providing the MCF7 and T47D Affymetrix data. Many thanks to P. Gestraud, P. Hupe, F. Picard, G. Rigaill for their help in defining the segmentation strategy, and to J. van Bemmel for her feedbacks about the manuscript.

References

Ferhat Ay and William S Noble. Analysis methods for studying the 3D architecture of the genome. *Genome biology*, 16:183, September 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0745-7.

A. R. Barutcu, A. J. Fritz, S. K. Zaidi, A. J. vanWijnen, J. B. Lian, J. L. Stein, J. A. Nickerson, A. N. Imbalzano, and G. S. Stein. C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization. *J. Cell. Physiol.*, Jun 2015.

Rameen Beroukhim, Xiaoyang Zhang, and Matthew Meyerson. Copy number alterations unmasked as enhancer hijackers. *Nature genetics*, 49:5–6, December 2016. ISSN 1546-1718. doi: 10.1038/ng.3754.

Boyan Bonev and Giacomo Cavalli. Organization and function of the 3D genome. *Nature reviews. Genetics*, 17:661–678, October 2016. ISSN 1471-0064. doi: 10.1038/nrg.2016.112.

Britta A M Bouwman and Wouter de Laat. Getting the genome in shape: the formation of loops, domains and compartments. *Genome biology*, 16:154, August 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0730-1.

Giovanni Ciriello, Martin L Miller, Blent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45:1127–1133, October 2013. ISSN 1546-1718. doi: 10.1038/ng.2762.

Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC genomics*, 13:436, August 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-436.

C.Wiuf and P.H.Stumpf. Binomial subsampling.

Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485:376–380, April 2012. ISSN 1476-4687. doi: 10.1038/nature11082.

Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V Lobanenkov, Joseph R Ecker, James A Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518:331–336, February 2015. ISSN 1476-4687. doi: 10.1038/nature14222.

I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li,

X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanari, M. L. Tress, M. J. van Baren, N. Walters, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang,

M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Reymond, R. Guigo, J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

William A Flavahan, Yotam Drier, Brian B Liau, Shawn M Gillespie, Andrew S Venteicher, Anat O Stemmer-Rachamimov, Mario L Suv, and Bradley E Bernstein. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, 529: 110–114, January 2016. ISSN 1476-4687. doi: 10.1038/nature16490.

Martin Franke, Daniel M Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerkovi, Wing-Lee Chan, Malte Spielmann, Bernd Timmermann, Lars Wittler, Ingo Kurth, Paola Cambiaso, Orsetta Zuffardi, Gunnar Houge, Lindsay Lambie, Francesco Brancati, Ana Pombo, Martin Vingron, Francois Spitz, and Stefan Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538:265–269, October 2016. ISSN 1476-4687. doi: 10.1038/nature19800.

Stefan Gröschel, Mathijs A Sanders, Remco Hoogenboezem, Elzo de Wit, Britta A M Bouwman, Claudia Erpelinck, Vincent H J van der Velden, Marije Havermans, Roberto Avellino, Kirsten van Lom, Elwin J Rombouts, Mark van Duin, Konstanze Dhner, H Berna Beverloo, James E Bradner, Hartmut Dhner, Bob Lwenberg, Peter J M Valk, Eric M J Bindels, Wouter de Laat, and Ruud Delwel. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*, 157:369–381, April 2014. ISSN 1097-4172. doi: 10.1016/j.cell.2014.02.019.

Louise Harewood, Kamal Kishore, Matthew D Eldridge, Steven Wingett, Danita Pearson, Stefan Schoenfelder, V Peter Collins, and Peter Fraser. Hi-c as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome biology*, 18:125, June 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1253-8.

Denes Hnisz, Abraham S Weintraub, Daniel S Day, Anne-Laure Valton, Rasmus O Bak, Charles H Li, Johanna Goldmann, Bryan R Lajoie, Zi Peng Fan, Alla A Sigova, Jessica Reddy, Diego Borges-Rivera, Tong Ihn Lee, Rudolf Jaenisch, Matthew H Porteus, Job Dekker, and Richard A Young. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (New York, N.Y.)*, 351:1454–1458, March 2016. ISSN 1095-9203. doi: 10.1126/science.aad9024.

M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

Philippe Hupé, N Stransky, JP Thiery, F Radvanyl, and E Barillot. GLAD: Gain and Loss Analysis of DNA. *R package version*, 2(0), 2011.

M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, 9:999–1003, 2012.

Effective normalization for copy number variation in Hi-C data

N. Servant [1,2,3,*†], N. Varoquaux [4,5,†], E. Heard [1,6,7], JP. Vert [1,2,3,†], E. Barillot [1,2,3,†]

August 7, 2017

[1]Institut Curie, Paris, France, [2]INSERM, U900, Paris, France, [3]Mines ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, Fontainebleau, France, [4]Department of Statistics, University of California, Berkeley, USA, [5]Berkeley Institute for Data Science, Berkeley, USA, [6]CNRS UMR3215, Paris, France, [7]INSERM U934, Paris, France, [8]Ecole Normale Superieure, Department of Mathematics and Applications, Paris, France.

1 Supplementary methods

1.1 Data

We evaluate the copy number effect on Hi-C data using publicly available dataset (see Table 1).
Simulation data are generated using the IMR90_CCL6 data from Rao et al. [2014] (GSE63525). Real data from T47D and MCF7 breast cancer cell lines, as well capture Hi-C data, are used to further validate our results.

Datatype	Sample	ID	Reference	Description
Hi-C	IRM90_CCL6	GSE63525	[Rao et al., 2014]	Diploid IMR90 Hi-C data at high resolution
Hi-C	Simulation 1	-	-	Highly rearranged simulated data derived from IMR90 Hi-C data
Hi-C	Simulation 2	-	-	Aneuploidy simulated data derived from IMR90 Hi-C data
Hi-C	T47D	GSE53463	[Le Dily et al., 2014]	T47D breast cancer Hi-C data
Hi-C	MCF7	GSE66733	[Barutcu et al., 2015]	MCF7 breast cancer Hi-C data
Hi-C	MCF10-A	GSE6673	[Barutcu et al., 2015]	MCF10-A nearly-diploid Hi-C data
Capture Hi-C	Mouse E12.5 limb buds	GSE78072	[Franke et al., 2016]	Capture Hi-C at the Sox9/Kcnj locus
ChIP-seq	MCF7 H3K27me3	ENCODE	-	H3K27me3 signal track
ChIP-seq	MCF7 H3K09me3	ENCODE	-	H3K09me3 signal track
ChIP-seq	MCF7 H3K27ac	ENCODE	-	H3K27ac signal track
ChIP-seq	MCF7 H3K36me3	ENCODE	-	H3K36me3 signal track
ChIP-seq	MCF7 H3K04me	ENCODE	-	H3K27me3 signal track

Table 1: Description of data used in this study.

1.2 Data Processing

All raw Hi-C data are processed with the HiC-Pro pipeline v2.8.0 [Servant et al., 2015] up to normalized ICE contact maps. Iterative corrections are processed using the iced python package (v0.4.2), after removing the 2% low coverage

*To whom correspondence should be addressed
†Equally Contributed

rows and columns.
Intra-chromosomal contact maps are annotated as previously described, by summarizing for each bin the GC content, effective fragment lengths and mappability [Hu et al., 2012]. In order to explore the bias in *cis* contact maps, we first split each feature into 20 bins of equal size. The *cis* contact maps are then normalized by the expected counts based on genomic distance, in order to generate the observed/expected (O/E) maps (See Section 1.3. For each feature bin, we then calculate the mean contact frequency over all *cis*-O/E maps.

Affymetrix SNP6.0 analysis raw data are normalized by technology specific software to extract signal for each probe. The obtained copy number profile is smoothed by a segmentation algorithm to remove noise and detect breakpoints. A similar process is applied on the allelic frequency probe [Popova et al., 2009]. The combined type of probes allow getting an estimate of absolute copy number for each probe, sample cellularity and tumour ploidy.

1.3 Computing Observed/Expected (O/E) matrices

We estimate the "Expected" matrices as follow. First, for each genomic distance s, compute the mean contact count interaction. then, apply an isotonic regression to enforce that the expected count decease as a function of the genomic distance. We set the *trans* expected count to the average *trans* contact count. We thus obtain the function $e(s)$, described in the Methods. We can then compute the O/E matrices:

$$O/E_{ij} = \frac{C_{ij}}{e(s(i,j))} \tag{1}$$

1.4 Estimating the error

To assess the ability of different methods to normalize abnormal karyotype data, we need quantitative measures of similarities between contact maps. We propose to use three measures, with different properties. First, we compute the O/E matrices as described before: this normalizes both for different coverage and the structure induced by structural properties of the DNA. We then compute the "block-average" of the matrix: between each copy-number breakpoint, we compute the mean of the matrix. We denote by $\bar{C}$the block-average of C. We then compare this resulting matrix with the "block-average" of the ground-truth G in three different manners: the ℓ_1 error, ℓ_2 error and $\ell_{\max}$.

$$\ell_1(C) = \sum_{i,j}(|\bar{C}_{ij} - \bar{G}_{ij}|)$$

$$\ell_2(C) = \sum_{i,j}(|\bar{C}_{ij} - \bar{G}_{ij}|^2)$$

$$\ell_{\max}(C) = \max_{i,j}(|\bar{C}_{ij} - \bar{G}_{ij}|)$$

In addition, we refer to the difference between block-average matrix of the ground truth and the normalize count as the "block average error matrix".

1.5 Chromosome compartments calling

Chromosome compartments calling is performed as previously described on 250Kb contact maps [Barutcu et al., 2015] using the HiTC R package [Servant et al., 2012] and the *pca.hic* function. First, O/E maps are estimated, thus normalizing contact counts for the structural effect of the genomic distance. Then, Principal Component Analysis (PCA) on the pearson correlation of the O/E matrices is applied. Active and inactive compartments are respectively assigned based on genome-wide gene density.

We further assess the enrichment of chromatin marks within A/B chromosomal compartments. To do so, the ChIP-seq signal is binned into 100 kb bins and normalized by the copy number signal. We then calculate the enrichment fold as the median of the signal track in bins within the cluster of interest, divided by the median of the signal track across all bins. The fold enrichment is calculated genome-wide, or for each chromosome independantly.

1.6 Capture-C analysis

The raw sequencing data were downloaded from GEO (GSE78072) and processed using HiC-Pro (v2.8.0). In order to compare the different contact maps, we applied the strategy proposed by Franke et al. For each map, we excluded the regions involved in the duplications and calculated a scaling factor for each map (sum of contacts/10^6). Each contact matrix was then scaled by its scaling factor before subtraction.

2 Supplementary tables

	ICE			CAIC		
	ℓ_1	ℓ_2	$\ell_{\max}$	ℓ_1	ℓ_2	$\ell_{\max}$
Aneuploid data set	2.321×10^6	4.574×10^5	0.313	2.291×10^6	3.532×10^5	0.185
Highly rearranged data set	3.984×10^6	8.832×10^5	1.344	2.055×10^6	2.536×10^5	0.360

Table S1: **Errors on the simulated data set**

3 Supplementary figures

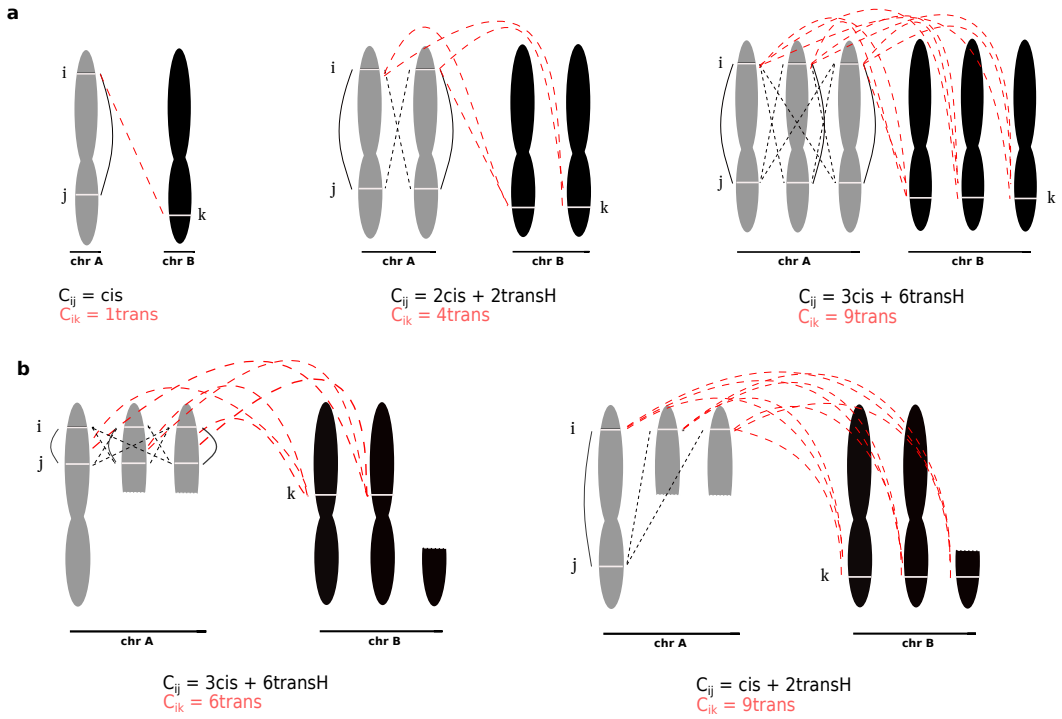


Figure S1: **Case study of simulated cancer Hi-C data.**
a. Example of contact frequency decomposition in the context of polyploid genome. **b.** Extension to local copy number rearrangement, when the two loci are on the same DNA segment (left) and on different segments (right).

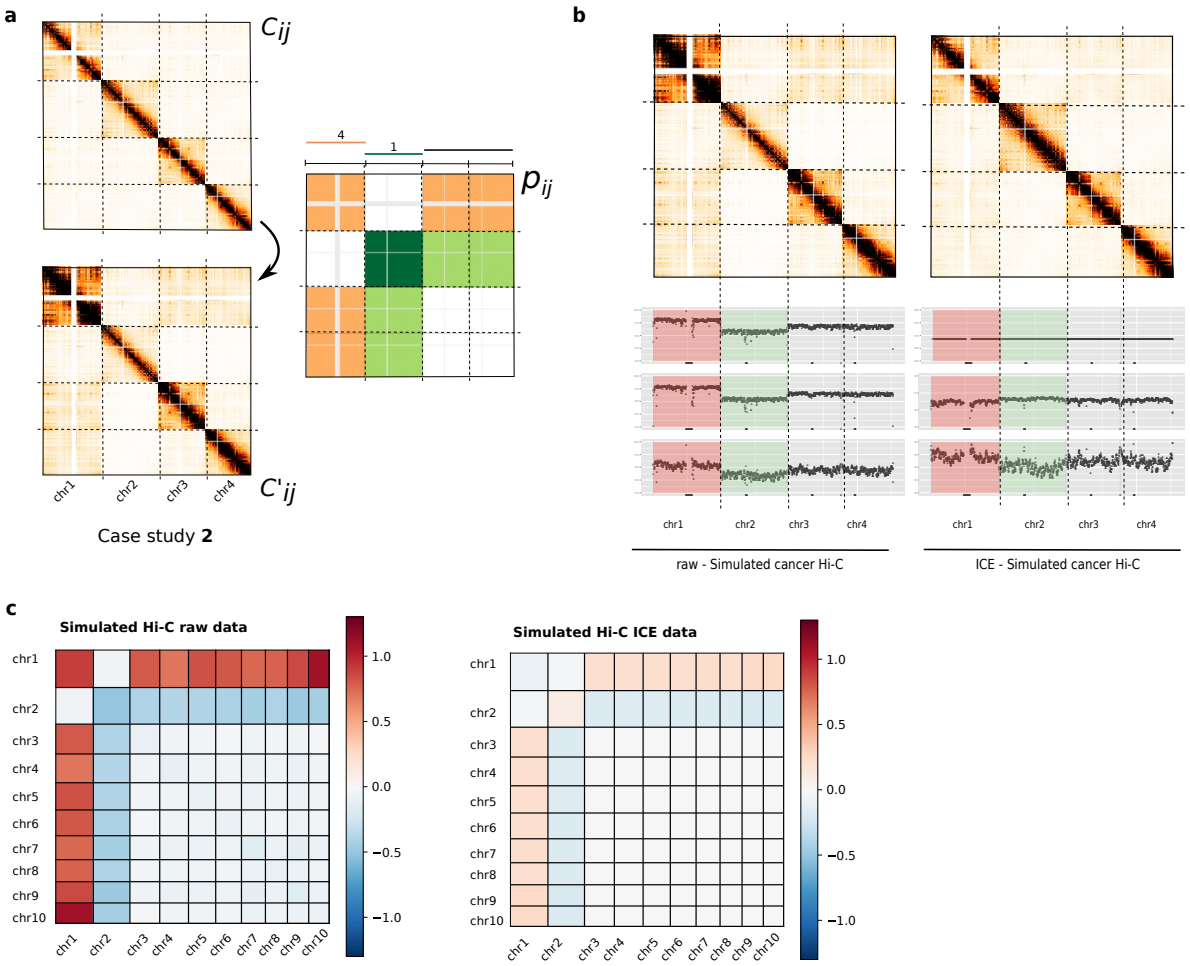


Figure S2: **Simulation of aneuploidy data.**
a. We explore different simulation scenarios based on different copy number profiles. This simulation is only based on gain or loss of complete chromosomes, therefore modeling aneuploidy effect. **b.** Applying the iterative correction on aneuploidy data should efficiently correct intra-chromosomal data from biases. However, the method does not properly rescale the inter-chromosomal contacts of different chromosomes.**b.** Block-average error matrix of simulated raw and ICE cancer data. As previously observed, the iterative correction does not allow to correct for segmental copy number bias.

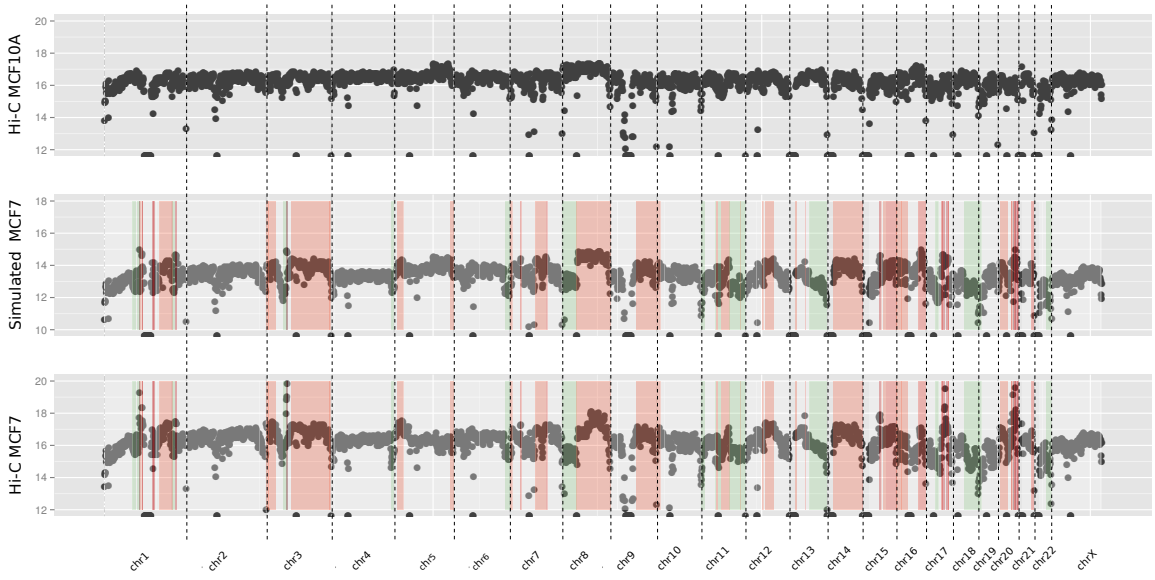


Figure S3: **Validation of the simulation model.**
1D genome-wide profiles of near-diploid MCF10A, simulated MCF7 and real MCF7 Hi-C data. The genome-wide profile of simulated MCF7 data is well correlated with the profile from real MCF7 Hi-C data, therefore validating the ability of the model to simulate large rearrangement events.

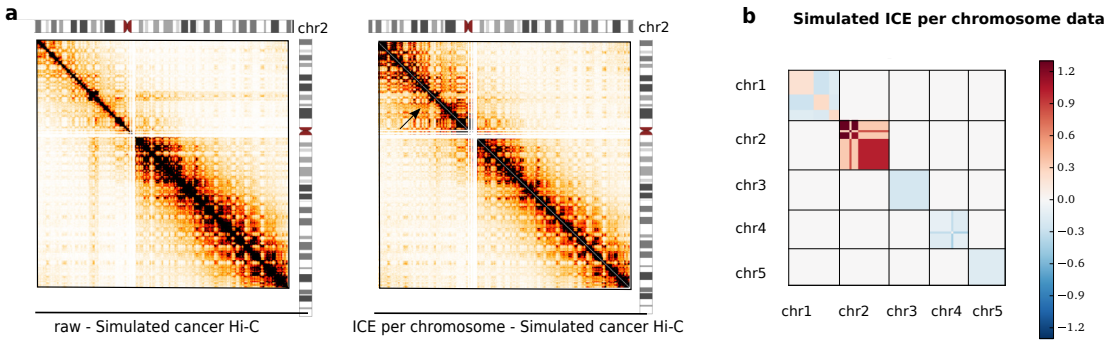


Figure S4: **ICE normalization per chromosome.**
a. Example of chromosome 2 contact maps before and after iterative correction per chromosome. **b.** Applying the iterative correction per chromosome on the intra-chromosomal data gives better results than the genome-wide approach. However, we can see that some biases still remains locally in the short-range contacts.

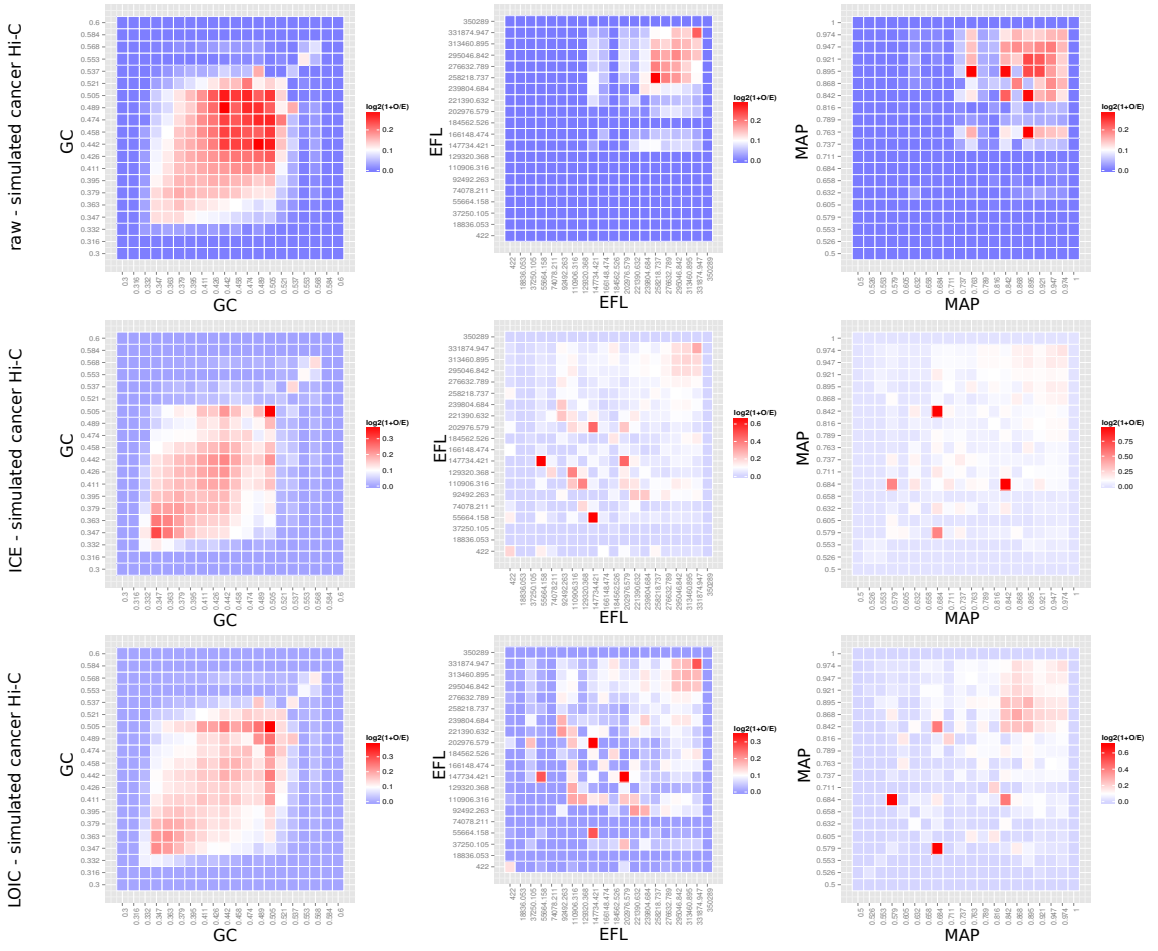


Figure S5: **Systematic biases in Hi-C *cis* experiment.**
The GC content, effective fragment length and mappability features were calculated for each 500 Kb bin. Each annotation feature was then splitted into 20 bins of equal size. The cis contact maps of raw, ICE and LOIC data were then normalized by the expected counts based on genomic distance, in order to generate the observed/expected (O/E) maps. For each feature bin, we then represented as a heatmap the mean contact frequency (log2).

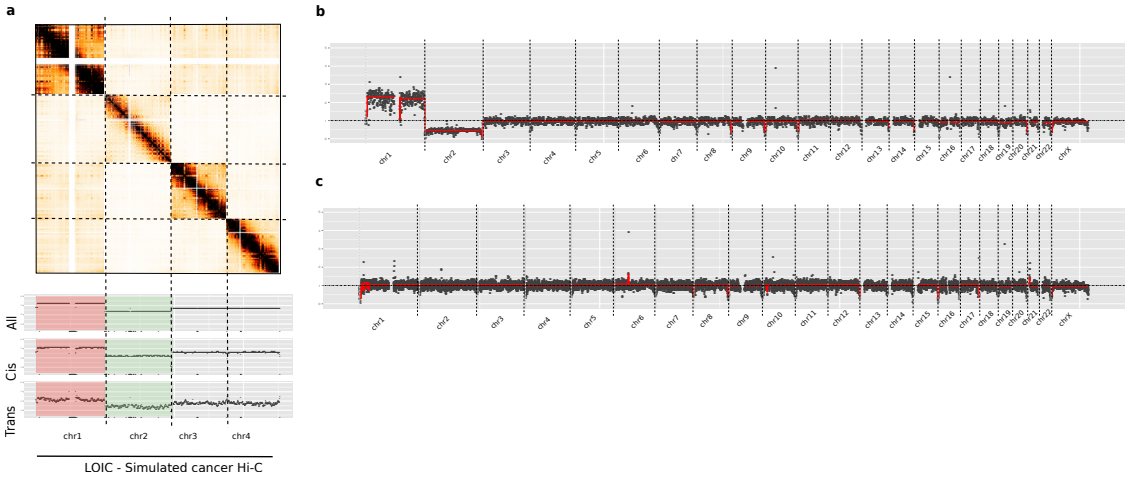


Figure S6: **Application of LOIC on simulated aneuploidy and diploid Hi-C data.** **a.** 1D profile of Hi-C data after LOIC normalization. As expected, the LOIC strategy allows to conserve the copy number information both in *cis* and trans contacts. **b.** Segmentation results of simulated aneuploid Hi-C data. **c.** Segmentation profile of diploid IMR90 Hi-C data. For diploid data, the segmentation profile is mainly flat along the genome. LOIC procedure should therefore gives very similar results as compared to the ICE method.

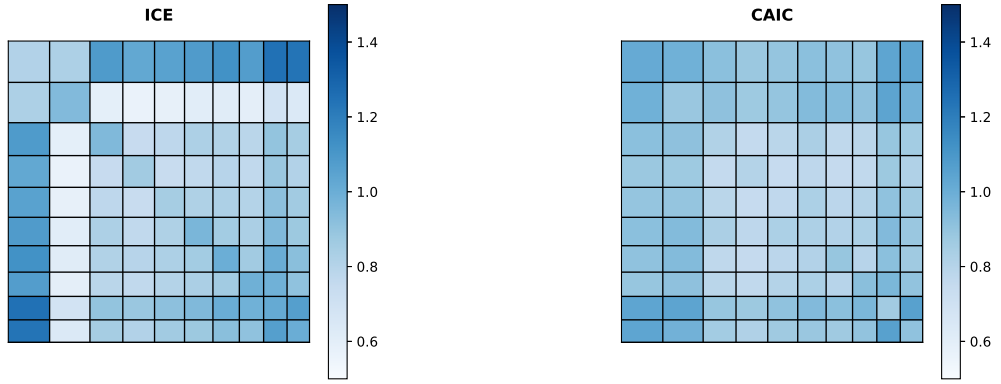


Figure S7: **"block-average" matrices for aneuploid simulated data set**
We observe that ICE introduces a bias when correcting aneuploid data, mostly by over-correcting depleted *trans* regions. CAIC yields as expected a nearly uniform matrix.

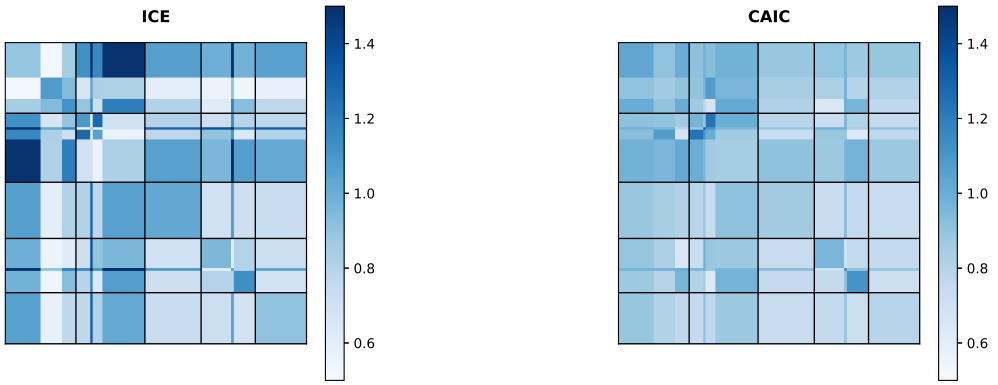


Figure S8: **"block-average" matrices for highly rearranged simulated data set** We observe that ICE introduces a bias when correcting abnormal karyotype data.

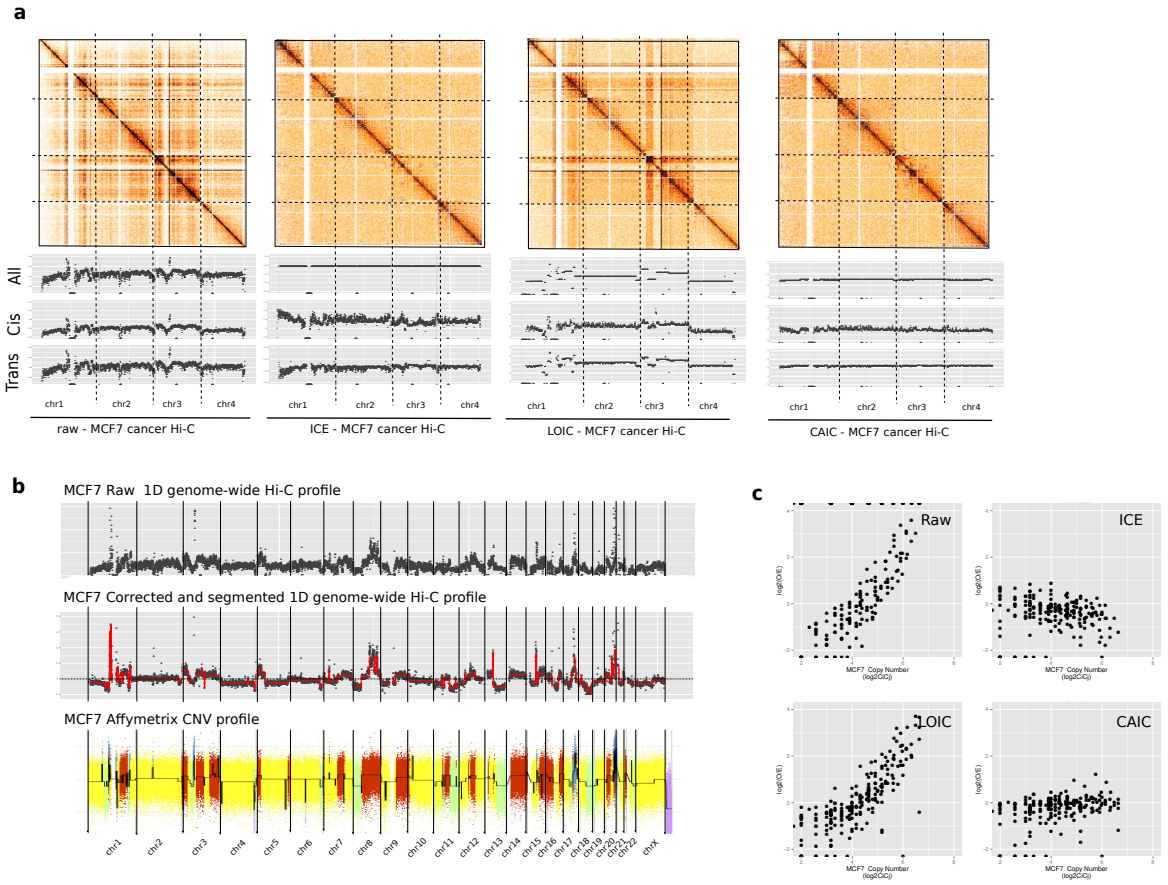


Figure S9: **Normalization of MCF7 Hi-C data. a.** Hi-C contact maps (250Kb resolution) of the first four chromosomes of MCF7 cancer Hi-C sample. As already shown on simulated and T47D data, we observed that ICE introduces a bias on normalized data. We then applied the LOIC and CAIC normalization approches to correct for systematic biases while removing or keeping the CNVs effect.**b.** In order to estimate the copy number signal from the Hi-C data, we applied a correction and segmentation method to the 1D profile. The inferred copy number signal from the Hi-C data are highly correlated with the copy number profile from Affymetrix SNP6.0 array. **c.** Relationship between contact frequencies with copy number on raw and normalized Hi-C data.

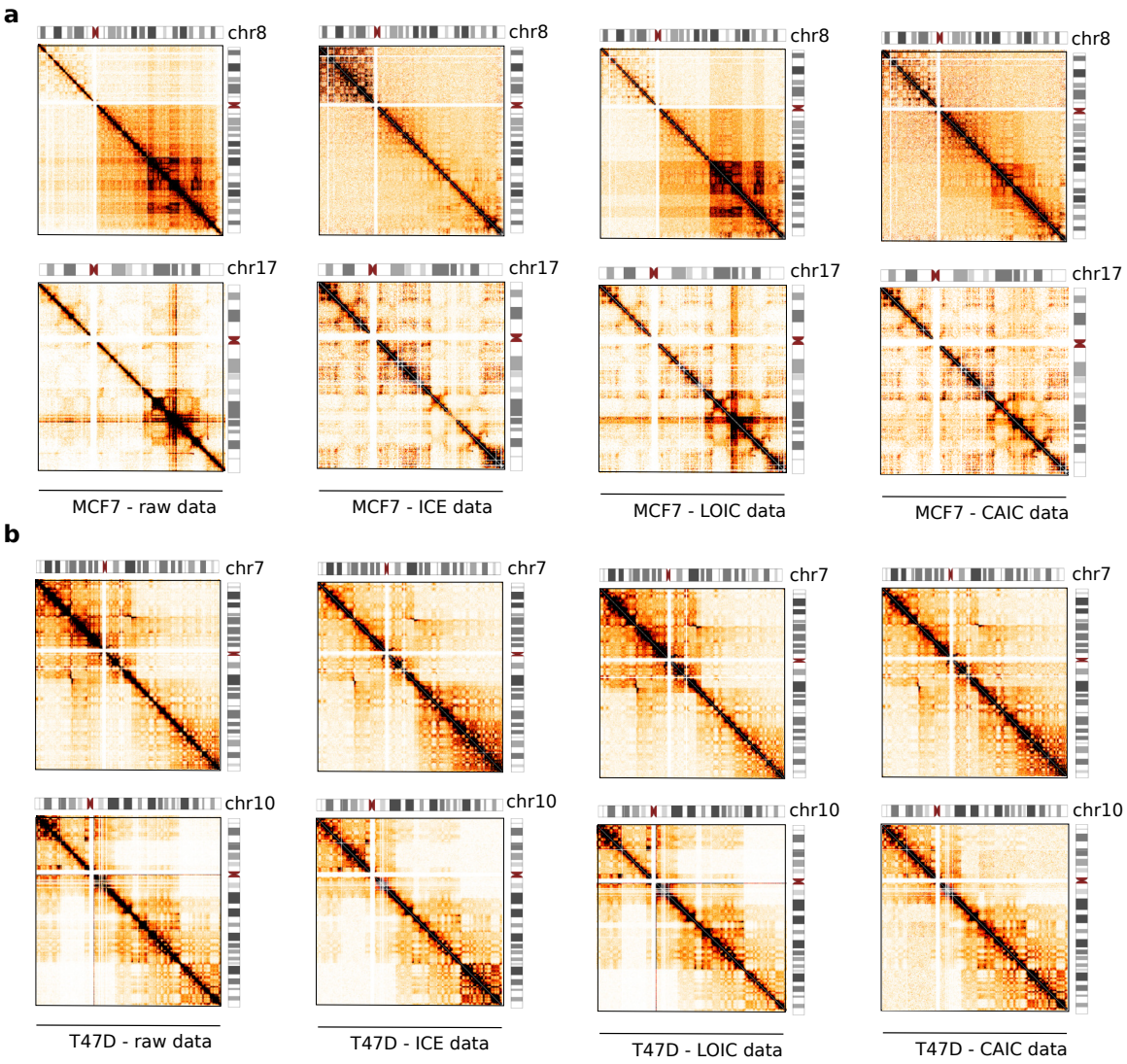


Figure S10: **Normalized intra-chromosomal contact maps. a.** Intra-chromosomal contact maps of chromosome 8 and 17 from MCF7 data, not normalized or normalized using the ICE, LOIC or CAIC method. **b.** Intra-chromosomal contact maps of chromosome 10 and 7 from T47D data, not normalized or normalized using ICE, LOIC or CAIC method.

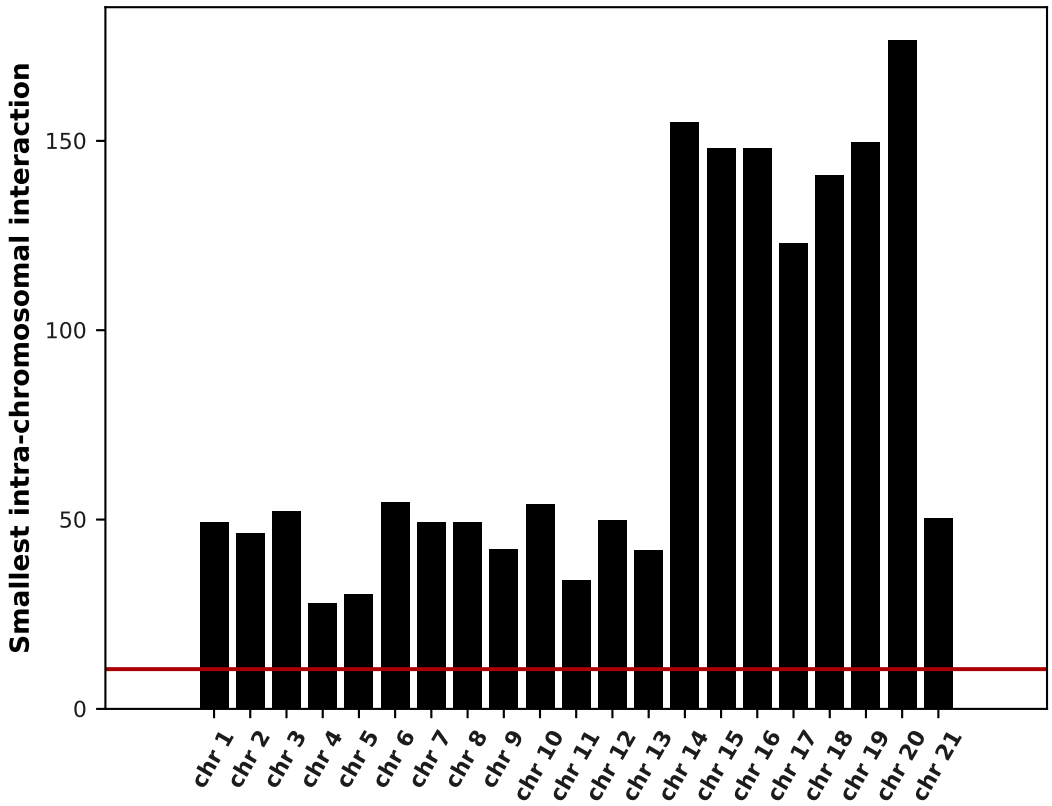


Figure S11: ***transH* versus *cis***
Estimated smallest *cis* interaction for each chromosome (See Section 1.3). The red horizontal line corresponds to the estimated *transH* interaction.

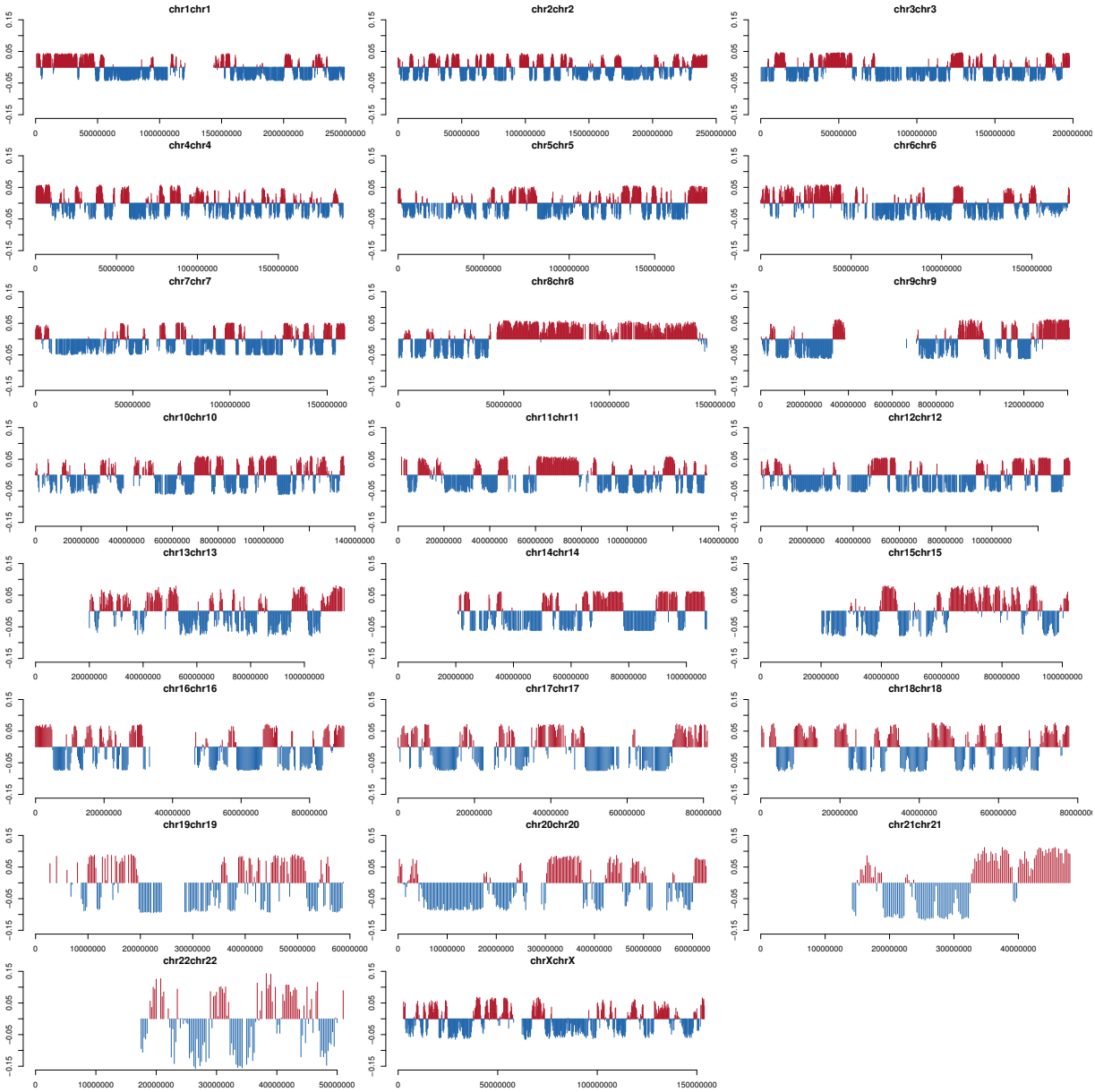


Figure S12: **MCF7 chromosome compartments - ICE data**
Results of the compartment calling using ICE normalized MCF7 data. First principal component representing active A-type (red) and inactive B-type (blue) chromosome compartments.

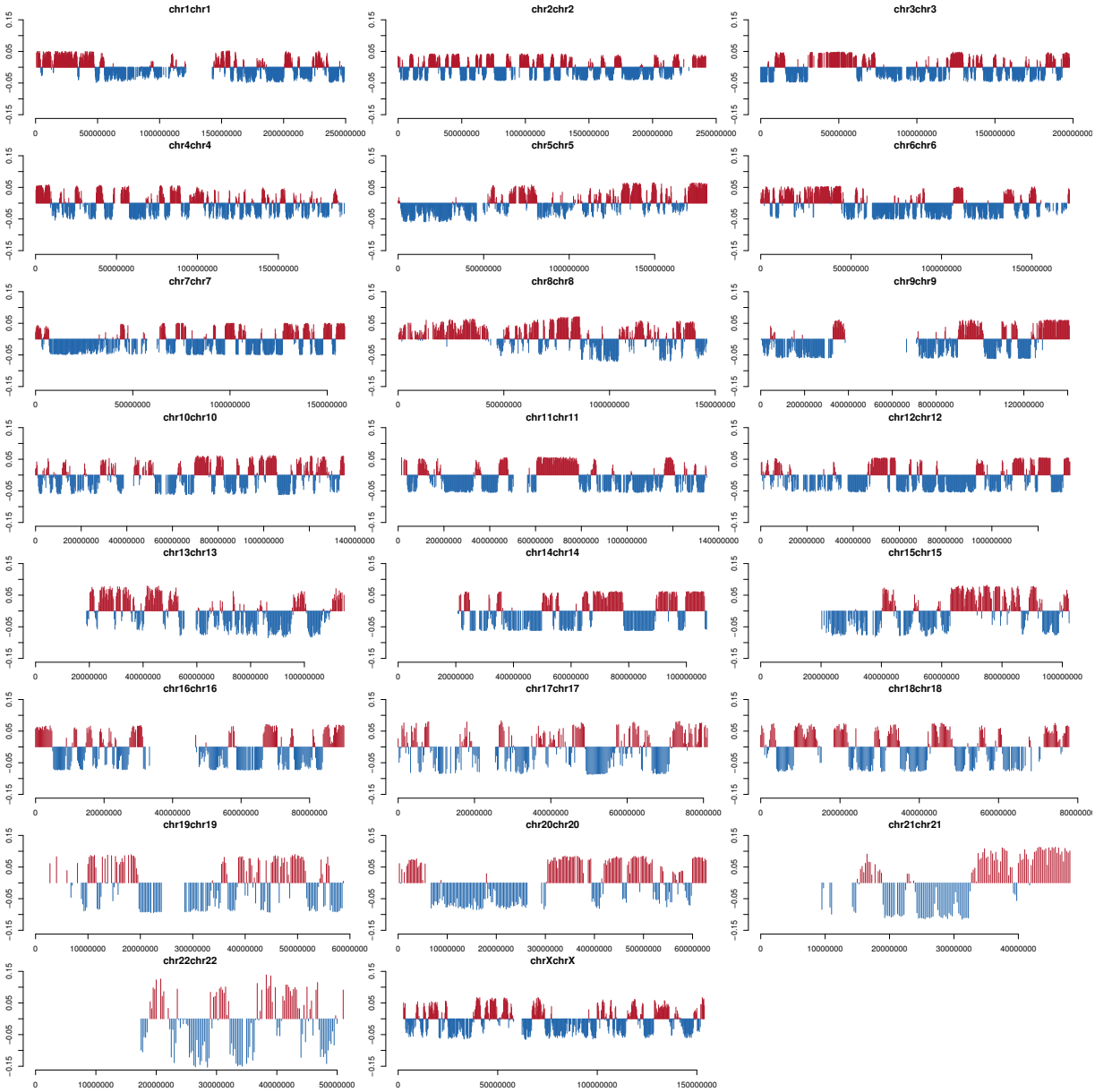


Figure S13: **MCF7 chromosome compartments - LOIC data**
Results of the compartment calling using LOIC normalized MCF7 data. First principal component representing active A-type (red) and inactive B-type (blue) chromosome compartments.

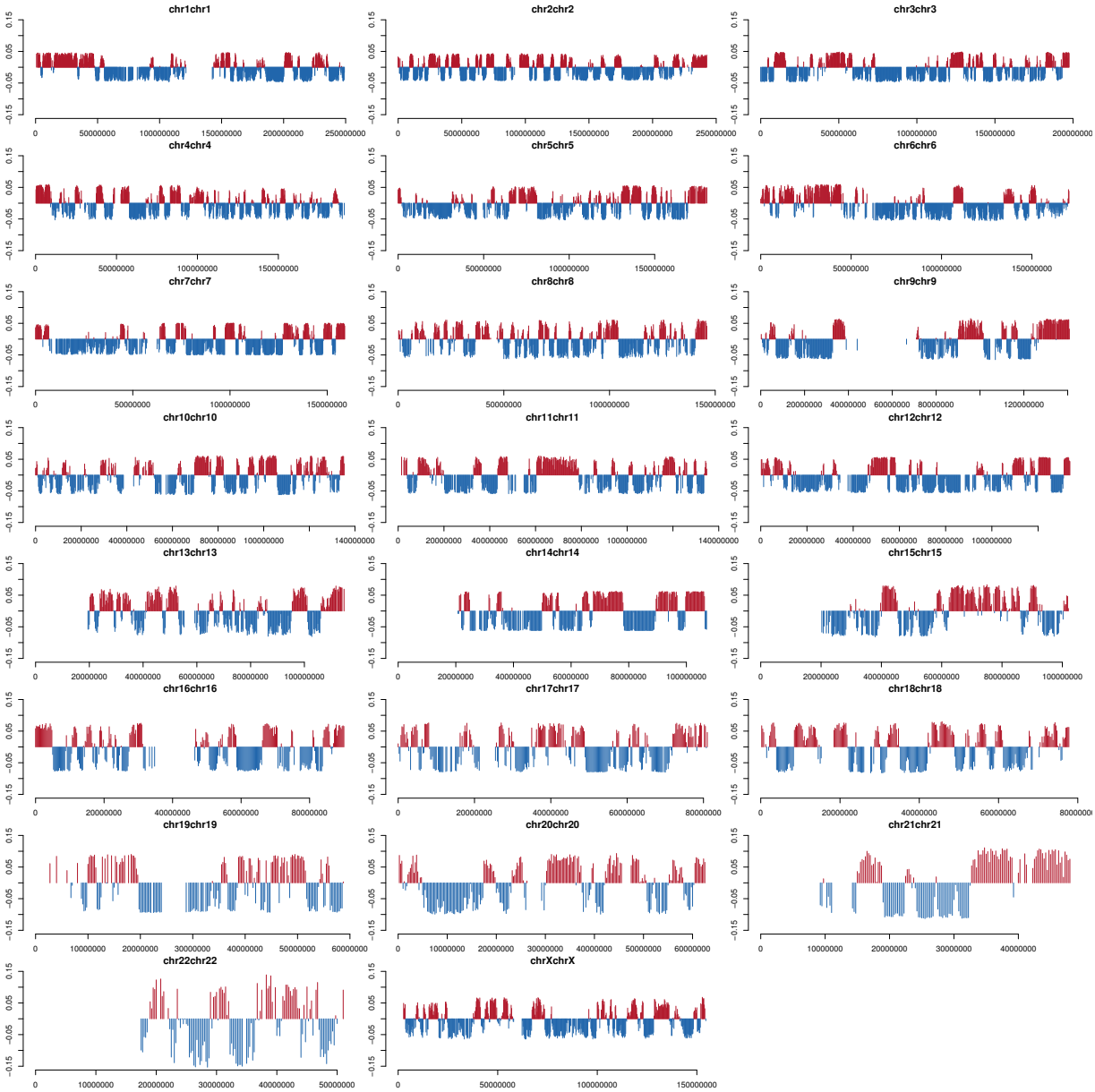


Figure S14: **MCF7 chromosome compartments - CAIC data**
Results of the compartment calling using CAIC normalized MCF7 data. First principal component representing active A-type (red) and inactive B-type (blue) chromosome compartments.

References

A. R. Barutcu, A. J. Fritz, S. K. Zaidi, A. J. vanWijnen, J. B. Lian, J. L. Stein, J. A. Nickerson, A. N. Imbalzano, and G. S. Stein. C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization. *J. Cell. Physiol.*, Jun 2015.

Martin Franke, Daniel M Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerkovi, Wing-Lee Chan, Malte Spielmann, Bernd Timmermann, Lars Wittler, Ingo Kurth, Paola Cambiaso, Orsetta Zuffardi, Gunnar Houge, Lindsay Lambie, Francesco Brancati, Ana Pombo, Martin Vingron, Francois Spitz, and Stefan Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538:265–269, October 2016. ISSN 1476-4687. doi: 10.1038/nature19800.

M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

Francois Le Dily, Davide Bau, Andy Pohl, Guillermo P Vicent, Franois Serra, Daniel Soronellas, Giancarlo Castellano, Roni H G Wright, Cecilia Ballare, Guillaume Filion, Marc A Marti-Renom, and Miguel Beato. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & development*, 28:2151–2162, October 2014. ISSN 1549-5477. doi: 10.1101/gad.241422.114.

Tatiana Popova, Elodie Mani, Dominique Stoppa-Lyonnet, Guillem Rigaill, Emmanuel Barillot, and Marc Henri Stern. Genome alteration print (gap): a tool to visualize and mine complex cancer genomic profiles obtained by snp arrays. *Genome biology*, 10:R128, 2009. ISSN 1474-760X. doi: 10.1186/gb-2009-10-11-r128.

S. S. P. Rao, M. H. Huntley, N. Durand, C. Neva, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin v looping. *Cell*, 59(7):1665–1680, 2014.

N. Servant, B. R. Lajoie, E. P. Nora, L. Giorgetti, C. J. Chen, E. Heard, J. Dekker, and E. Barillot. HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics*, 28(21):2843–2844, 2012.

N. Servant, N. Varoquaux, B. R. Lajoie, E. Viara, C. J. Chen, J. P. Vert, E. Heard, J. Dekker, and E. Barillot. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, 16:259, 2015.

3.3 Application to Uveal Melanoma Hi-C data

In addition to the analysis of public Hi-C data presented in the previous section, we also generated our own Hi-C data on uveal melanoma cell lines. The uveal melanoma is a good starting point to study the spatial organization of cancer genome, as it is well characterized and is fairly little rearranged compared to other types of cancer.

Although the ocular melanoma is the second most common type of melanoma after cutaneaous, it is still considered as a rare cancer type. In Europe, its incidence varies from 2 to 8 cases per million. Uvea is the most frequent site of origin of ocular melanomas, which mainly affect the choroid of the uveal tract leading to uveal melanoma, although other sites such as the iris or the ciliary body have also been reported (see Jovanovic et al. (2013) for a review). Thus, uveal melanoma is the most frequent primary tumor of the eye in adults. Conservative therapies of uveal melanoma are mainly based on radio or proton-therapy, and usually give effective local control of the tumor. Yet, the 5-year survival rates of uveal melanoma have not evolved for the last decades, and more than half of the patients succumb to metastasis within 10 years of diagnosis (see Carvajal et al. (2017) for a review). The liver is the most commonly affected site of metastasis (90% of patients), while other sites such as lung (24%) and bone (16%) have also been reported. Given the poor prognosis associated with the development of metastasis, there is a strong need for early diagnosis, as well as efficient treatment of the primary disease and control of the metastasis. Currently, the risk of developing a metastasis is mainly determined by the tumor size and its location, but recent advances in the field propose to use the genetic profile of tumors to classify them in sub-types associated with patient prognosis and survival.

On a genomic point of view, uveal melanomas are characterized by a low degree of genomic instability compared to other cancer types. The most common reported abnormalities include loss on chromosome 1p, 3, 6q, 8p, and 16q and gain on 1q, 6p, and 8q (Harbour (2012)). The loss of one copy of chromosome 3 arises in more than 50% of uveal melanoma, and is, so far, the main prognosis factor based on genomic profiling. Indeed, the chromosome 3 monosomy is associated with a poor prognosis, whereas a chromosome 3 disomy is rarely associated with metastasis. Uveal melanoma has therefore been first classified in two main groups, based on the chromosome 3 status. A normal status of chromosome 3 is significantly associated with a gain of chromosome 6p.

The gain of chromosome 6p has therefore frequently been reported as a good prognosis factor. Interestingly, the gain of 6p is frequently associated with a loss of 6q. However, no pathogenic mechanism implying these regions has been reported so far. On the other hand, the chromosome 3 monosomy has been described as frequently associated with a gain of the chromosome 8q. The association of both alterations leads to the most aggressive tumor type, and the gain of chromosome 8q is significantly associated with a higher risk of metastasis. Although the exact mechanisms involve in metastasis are not fully understood, the chromosome 8q contains many potential oncogenes which can be involved in this process, such as *Myc*, *Def1* or *Nbs1*. Additional genomic alterations have been reported on chromosome 1 or 16, but without clear association with prognosis and metastasis.

The monosomy of chromosome 3 has been largely explored with the aim of understanding the precise mechanism which can explain its assocation with bad prognosis and metastasis. Among the different studies, the BRCA1-associated protein-1 (*Bap1*) gene, located at chromosome 3p21.1, has been found mutated in more than 80% of poor prognosis tumor, and has been described as associated with metastasis (Harbour et al. (2010)). *Bap1* was identified as a tumor suppressor gene involved in DNA repair. In addition, *Bap1* seems to be able to remove monoubiquitin from histone H2A in association with the *Asxl1* gene. It could therefore play a criticial role in cancer tumorigeneis as a chromatin modifier. Recently, an additional sub-group associated with a chromosome 3 disomy and *Sf3b1* mutation has been described. While the chromosome 3 disomy is usually associated with a good prognosis, those patients have an increase risk of late metastasis compared to patients without mutations (Yavuzyigitoglu et al. (2016)). Finally, activating mutations in the MAP kinases pathway, mainly in *Gnaq* or *Gna11*, have also been reported in more than 80% of uveal melanomas. So far, it seems that a mutation in *Gnaq* or *Gna11* gene is not sufficient for malignant transformation to melanoma but could be an early event in uveal melanoma progression.

The work presented below is part of a collaborative project (R. Margueron, S. Roman-roman, M.H. Stern, E. Heard - Institut Curie). All experiments have been performed by D. Gentien. In the context of this project, we started to explore the chromatin organization of two uveal melanoma cell lines in comparison with normal Human melanocytes (Nmati et al. (2010)). The two cell lines, MP41 and MP46 are

derived from Mouse xenografts. The MP41 cell line is characterized by a loss of chromosome 3 and is not *Bap1* mutated. The MP46 model is characterized by an isodisomy of chromosome 3 (loss of one copy which is then duplicated) and is *Bap1* mutated (Figure 3.1). Both models are characterized by gain of chromosomes 6p and 8q and are *Gna11* mutated, and *Sf3b1* wildtype. The final goal of the project is therefore to compare the chromatin lanscape of *Bap1* wildtype and *Bap1* mutated tumor. In the context of the present work, the uveal melanoma Hi-C data were so far mainly used to validate our developments on data normalization. Additional experiments are in progress to further explore the epigenetic changes between *Bap1* wildtype and mutated tumors.

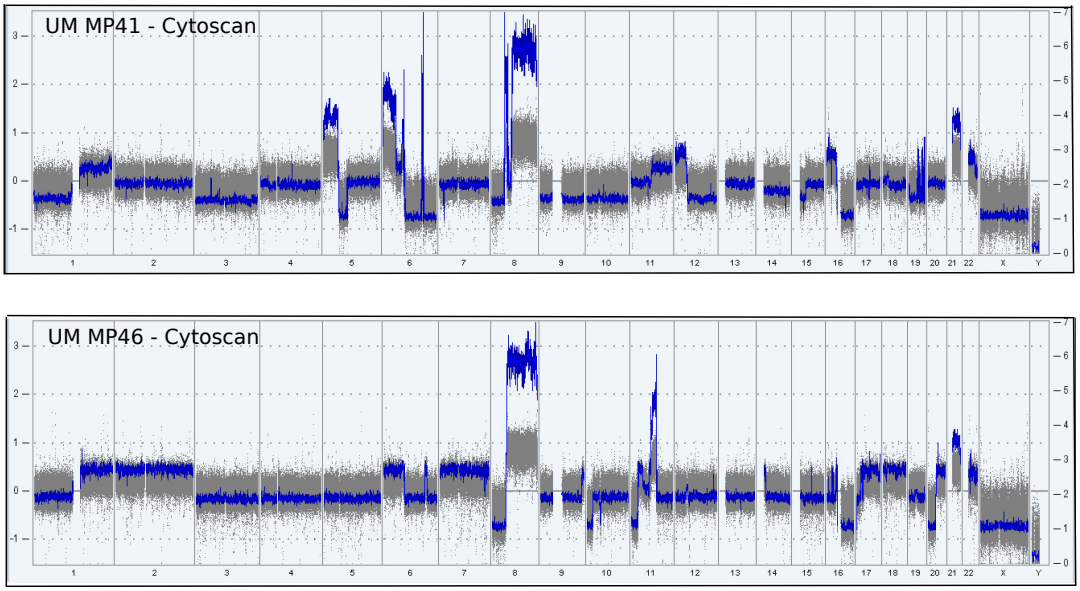


Figure 3.1: Genomic profiles of MP41 and MP46 uveal melanoma models - Copy number profiles from Affymetrix Cytoscan arrays. The microarray signal is represented in gray, and the estimated absolute copy number in blue.

3.3.1 Copy number profile inferred from Hi-C data

We applied the strategy previously described to infer the copy number profile from the Hi-C data. Briefly, we summarized the data by calculating the sum of the genome-wide contact map (1D Hi-C profile). We then applied the regression model proposed by Hu et al. to correct this profile from known biases such as GC content, mappability and fragment lengths (see section 3.2, Methods). We finally ran a segmentation algorithm

on the corrected Hi-C 1D profile to detect the breakpoint positions (Figure 3.2)).

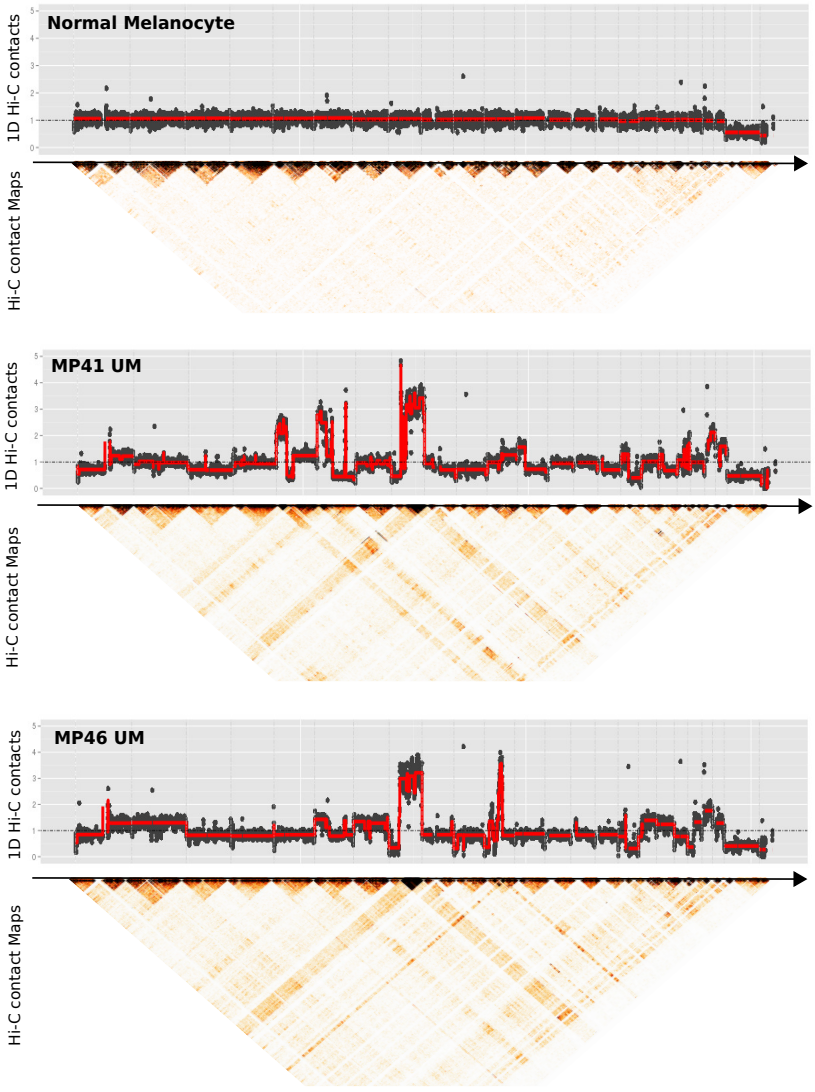


Figure 3.2: Estimation of copy number profiles from Hi-C data - 1D Hi-C profiles of normal melanocytes, and both MP41 and MP46 models. The red line represents the smoothing line after normalization and segmentation of the profile. The intra and inter-chromosomal contact maps are represented below each 1D Hi-C profile. As already observed, regions with higher/lower copy number are characterized by higher/lower contact frequencies.

As already observed on public breast cancer datasets and on our simulation model, the copy number extracted from the Hi-C data is very well correlated with the copy number profile from the Affymetrix Cytoscan data. We therefore retrieved the known alterations of chromosomes 3, 6 and 8 (Figure 3.2). Looking at the contact maps validates our previous observations. Regions with higher/lower copy number are characterized by blocks of higher/lower contact frequencies. Finally, we also applied the same segmentation procedure to the normal melanocyte sample. As expected, the result confirms that the genome of our control sample is fully diploid.

3.3.2 Normalization of Uveal Melanoma Hi-C data

We then leveraged these cancer Hi-C data in order to validate our previous observations on Hi-C data normalization. We therefore applied the ICE normalization, and the new methods previously described ; the LOcal Iterative Correction (LOIC), and the Copy number Adjusted Iterative Correction (CAIC). As already shown, the ICE normalization leads to an imbalance in the normalized contact frequency of intra-chromosomal maps, with a shift between gained and lost regions. After ICE normalization, the intra-chromosomal contact counts are depleted for regions with high copy number (Figure 3.3, MP41 chromosome 6). On the other hand, lost regions now present higher contact probabilities than gained regions in cis (Figure 3.3, MP46 chromosome 8).
We thus applied the LOIC and CAIC normalization methods on both MP41 and MP46 datasets. The LOIC normalization allows to correct for systematic bias while keeping the copy number information. We can therefore see that the effect observed on ICE normalized data is no longer true and that the LOIC normalized data well conserved the copy number structure. In addition, applying the CAIC normalization allows to efficiently remove the copy number biais. Blocks of higher/lower copy number now have similar behavior between normal and cancer samples. Interestingly, on MP46 chromosome 11 and MP41 chromosome 6, we observed that the CAIC normalization also corrects for copy number effect of blocks which are apart from the diagonal (Figure 3.3). These blocks are likely to be intra-chromosomal translocations. Of note, the CAIC method is not designed to correct for balanced translocations. The method works in this context only because the translocations are associated with alterations in the copy number status.

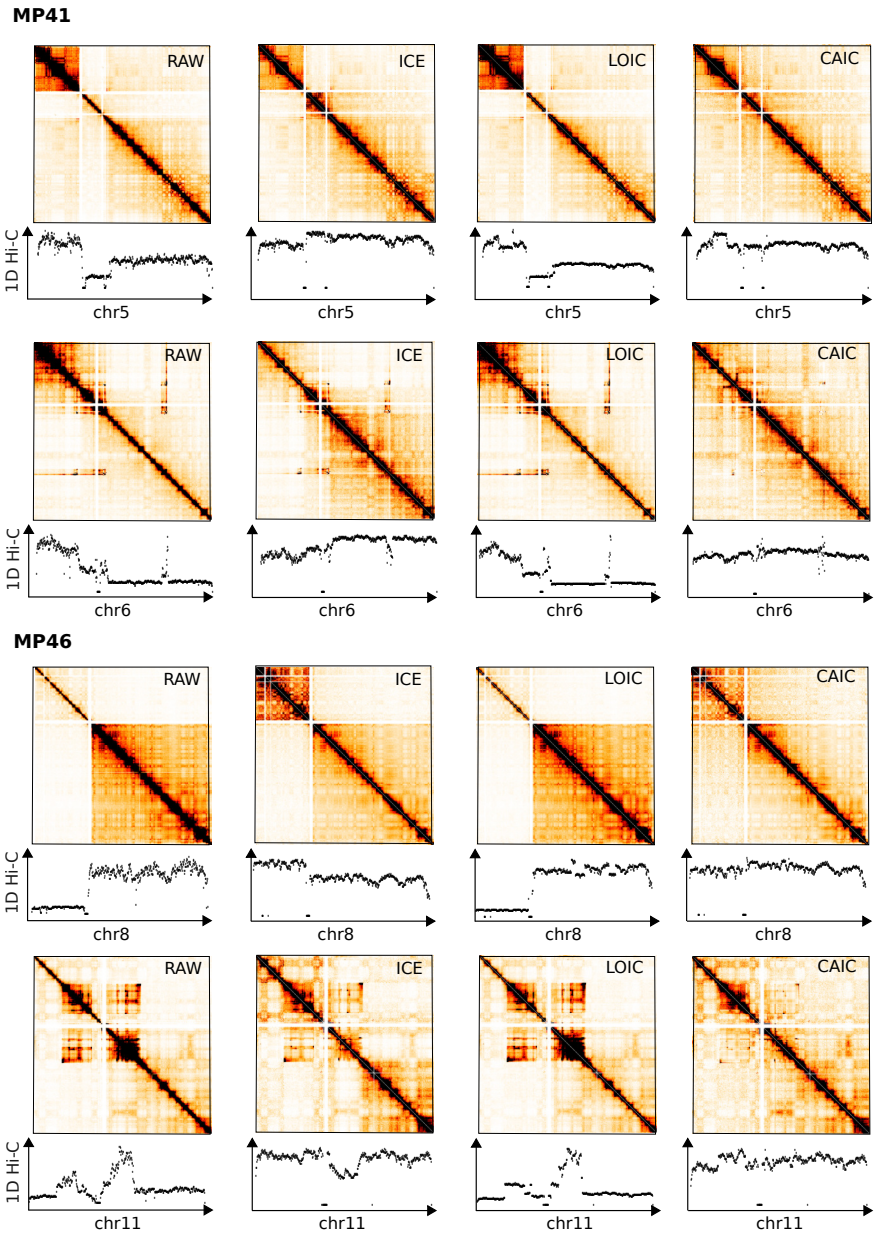


Figure 3.3: Normalization of Uveal Melanoma Hi-C data - Example of intra-chromosomal contact maps on both uveal melanoma models (MP41/MP46), using different normalization approaches (raw data, ICE, LOIC, or CAIC normalized data). The Hi-C 1D profile (sum of genome-wide contacts) is represented below each contact maps.

3.3.3 Detection of chromosome compartments

Finally, we ran the detection of A/B chromosome compartments on MP41 and MP46 Hi-C data normalized by the different methods. On the public MCF7 breast cancer data we previously investigated, we globally found very concordant results between all normalization methods. The chromosome 8 was the only one showing different A/B compartment patterns according to the normalization method, and we finally validated that the profile extracted with the CAIC normalized data was the only one correlated with the expected active and repressive histone mark patterns (see section 3.2).

In the context of uveal melanoma, we globally made the same observation, although more chromosomes seem impacted by the normalization strategy (Figure 3.4). As expected with the LOIC method, the first component of the principal component analysis (PCA) distinguishes regions of high/low copy number, validating the fact that removing the copy number effect is important before detecting the chromosomal compartments. In addition, our previous conclusions on the ICE normalization remain true here. The ICE normalization does not properly correct for the copy number, and this could affect the compartment calling as shown here on chromosomes 3, 6 and 8 (Figure 3.4a). Finally, only the comparment calling on CAIC normalized data seems not to be affected by the copy number. We made this conclusion by comparing the compartment calling on CAIC data with the one on normal melanocytes on chromosomes 3, 6 and 8. The average Spearman correlation between MP41 (resp. MP46) and the normal melanocyte increases from 0.49 (0.33) on ICE data to 0.74 (0.71) on CAIC data, which makes sense for data from the same cell type.

Finally, on the MCF7 breast cancer data, we also found that in some cases, the expected A/B compartments could be rescued by looking at the second component of the PCA. Interestingly, in the context of uveal melanoma, the second component still does not match to the profile obtained from the CAIC data (Figure 3.4b). This observation again highlights the importance of properly removing the copy number before compartment calling.

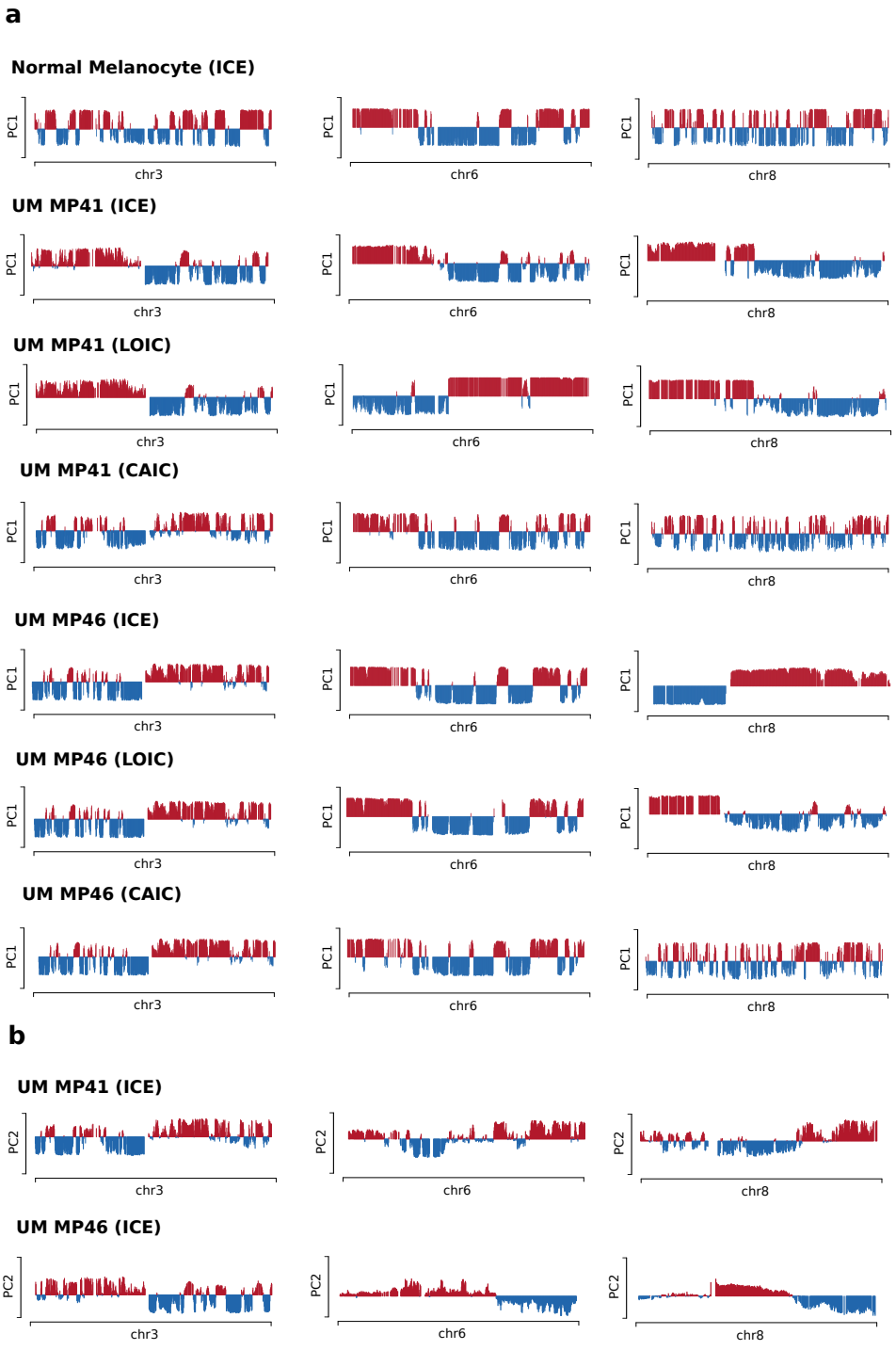


Figure 3.4: Chromosome compartments and Hi-C data normalization - a. Results of chromosome compartment calling for normal melanocytes, MP41 and MP46 data on chromosome 3, 6, and 8. The active compartments are in red and the inactive compartments in blue. The results are presented for each normalization method. **b.** Second principal component of PCA on ICE normalized data for MP41/MP46 on chromosomes 3, 6 and 8.

3.3.4 Changes in chromosome compartments between *Bap1* mutated *Bap1* wildtype tumors

We therefore used the CAIC normalized Hi-C data at a 250 kb resolution to detect open and close chromosome compartments on MP41 (*Bap1* wildtype) and MP46 (*Bap1* mutated) uveal melanoma tumors, and normal melanocytes, regardless their copy number profile.

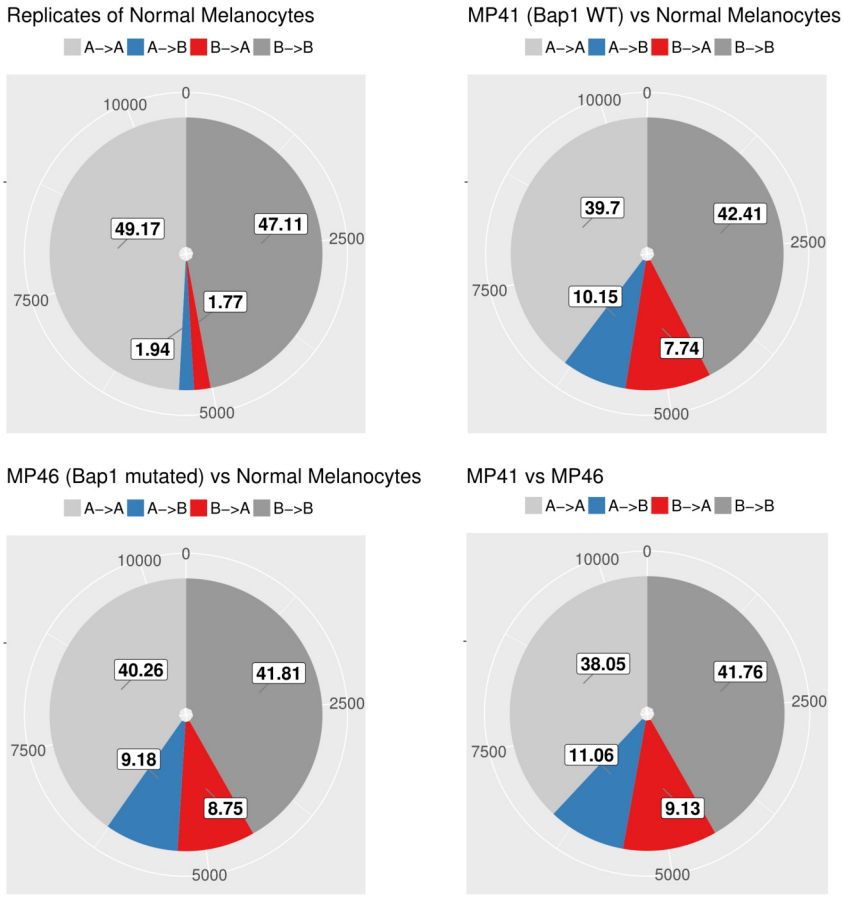


Figure 3.5: Chromosome compartment switches between *Bap1* wiltype and mutated tumors. - Pie chart showing the fraction of compartment changes between normal melanocytes, MP41 and MP46 tumors. 'A' and 'B' represent the open/closed compartments, respectively. 'A− >B' represents compartments that swith from an open to a close state. 'B− >A' represents compartments that swith from a close to an open state. 'A− >A' (or 'B− >B') represents compartments that do not change between samples, and stay open (or close).

We then further compared the chromosome compartments of both normal and tumor samples (Figure 3.5). As a control, we compared the compartments of two biological replicates of normal melanocytes. We observed that less than 2% of the compartments have a different states, therefore validating the reproductibility of the experiments and the analysis. We then compared each tumor samples (MP41 and MP46) against the normal melanocytes and between them. Interestingly, between 17 and 20% of all compartments moved to the opposite compartment (open to close and vice versa). In addition, we observed the same proportion of compartment switches between MP41 and MP46, suggesting that the two tumors have distinct epigenetic profiles, which might be driven by the *Bap1* mutation.
Additional experiments related to methylation (whole genome bisulfite sequencing), histone modifications (ChIP sequencing) and gene expression (RNA sequencing) are currently in progress. Integrating these information with the chromosome compartment profiles should allow us to validate the compartment switches that we observe and to better characterize the effect of the *Bap1* mutation on the epigenetic landscape of uveal melanoma tumors.

3.4 Discussion

The application of Hi-C technique to cancer genome is of growing interest to better characterize the impact on the chromatin structure on tumorigenesis. But analyzing cancer data usually requires dedicated strategies. This is one of the main conclusions of our work on Hi-C data normalization. Using both simulated and real data, we demonstrated that copy number variants can have an impact of Hi-C data normalization and interpretation. While several studies recently came out with cancer Hi-C data, we clearly showed that the popular iterative correction method cannot be applied in this context. In most cases, the normalization of cancer data consists in removing the copy number effect which can introduce unwanted effects on the data. In the context of Hi-C, this makes sense to detect significant interactions between regulatory elements and their targets, or to directly compare the contact maps between datasets. However, as recently discussed by Harewood et al., the Hi-C technique is also a powerful approach to detect genomic alterations. Other studies also used cancer Hi-C data to explore the impact of genome rearrangements on gene regulation (Taberlay et al. (2016)). It

therefore makes sense to efficiently normalize Hi-C data without removing the copy number effect, in order to better understand the consequences of such rearrangements on the cell regulation.

From this observation, we designed two new strategies for cancer Hi-C normalization. We first extended the ICE correction published by Imakaev et al. to propose a local version, extending the idea of equal visibility of all genomic loci, to the idea of equal visibility of loci from the same DNA segment (and therefore with the same copy number). This strategy thus allows us to correct the data for systematic bias such as mappability, fragment length and GC content and to conserve the copy number information. In terms of computation time, this method is as efficient as the original ICE normalization. However, by applying the method to several cancer datasets, we also observed that for some downstream analysis such as compartment calling, the copy number could mask the biological signal, leading to a misinterpretation of the results. We therefore designed another method to remove the copy number effect (CAIC - Copy number Adjusted Iterative Correction). The method first runs the LOIC correction to remove the systematic biases. It then iterates to minimize the differences between the observed counts per genomic distance per copy number block and the average, genome-wide, expected counts per genomic distance.

As demonstrated earlier, both methods can be of interest to normalize Hi-C cancer data, and currently provide good results. A couple of points can however be discussed and extended. First, we observed that, overall, the compartment calling is not really affected by the normalization method applied on the data. This remains true as long as the copy number effects is not too strong, and that there is no stricking rearrangements such as deletion of a large region followed by an amplication of the neighboring region. However, so far, we did not investigate the effect of the normalization on other downstream analyses, such as the TADs calling, or the detection of significant contacts. Going further in this way would be interesting to validate the interest of our new normalization methods.

Regarding the method itself, the CAIC procedure relies on the assumption that the copy number effect is constant per block. This make sense if we believe that the copy-number effect between two loci is related to the amount of genetic material of those two regions, and thus identical for all genomic loci belonging to the same block. However, this assumption is valid only if we hypothesize that the cis interactions are always

stronger than the trans interactions, which might not be always true. For instance, let us consider a large duplicated genomic region. As usual, we expect that the contact frequencies decrease with the genomic distance. It means that for short distances, we can estimate a strong copy number effect in cis. But for large distances, it is likely that the two genomic loci, although they belong to the same block, tend to interact as if they were on two different chromosomes (i.e. in trans). In this case, the copy number effect might not be constant per block but will rather depend on the genomic distance between the two interaction loci. One idea would therefore be to adapt our current simulation and normalization models to include the distance between genomic loci in the calculation of the copy number effect. Therefore, we could imagine to model the count C_{ij} between two loci i and j, as a Poisson random variable with mean :

$$\mathbf{E}C_{ij} = \beta_i \beta_j T_{ij} U_{ij}$$

where $\beta_i \beta_j$ accounts for technical bias such as fragment length, GC content or mappability, U_{ij} is a structural bias which captures both the influence of copy number and of genomic proximity, and T_{ij} accounts for the signal which is not explained by the other terms, in particular biological effects. Being able to estimate the β, T, and U parameters would therefore allow to normalize both diploid and cancer Hi-C data. Of note, as the U term is function of the genomic distance, correcting this effect could be a way to generate contact maps normalized by the expected genomic distances as required by several statistical methods used to call significant contacts between loci. Although everything remains to do, this could be an interesting extention to our work. Then, going back to more practical aspects, the version of the CAIC method currently provided is still time and memory consuming. Indeed, the current implementation is based on a dense representation of the Hi-C data and requires a significant amount of memory. In addition, the CAIC procedure is highly dependent on the number of breakpoints, as it iterates over all copy number blocks. For highly rearranged profile or high resolution data, it is likely that the method runs in several hours, or even days.

Finally, we presented here an application of our methods on uveal melanoma Hi-C data, on normal samples as well as on tumor samples with or without *Bap1* mutation. We demonstrated that both LOIC and CAIC methods allow to efficiently normalize

these data. Using CAIC normalized data, we then extracted the chromosome compartments and identified genomic regions that switch from an open to a close chromatin states (and vice verca) between normal and tumors, and according to the *Bap1* status. In addition, generation of whole genome sequencing, ChIP-seq, RNA-seq and whole-genome bisulfite data is in progress. This dataset is therefore a unique opportunity to better characterize the uveal melanoma tumors and to better understand the role of *Bap1* and how its mutation could impact the epigenome. The loss of *Bap1* is expected to be associated with an enrichment in H3K27me3 (LaFave et al. (2015)). However, its impact on the DNA methylation, on the chromatin conformation or on the gene expression remains to be fully elucidated. The developments that we have done to normalize cancer Hi-C data will allow us to precisely explore the chromatin organization of these samples.

4

c-myc oncogene expression and inactive X chromosome organization in mouse liver tumors

4.1 Background

The *c-myc* proto-oncogene (herein termed *Myc*) is one of the most studied genes in oncogenetics. *Myc* has been shown to be involved in many mechanisms of tumorigenese, including cellular proliferation, apoptosis, DNA replication or transcription (Gabay et al. (2014)). One particularity of *Myc* is that its sole over-expression is sufficient to induce tumorigenesis. On the other hand, returning to physiological level of expression can result in a significant tumor regression. *Myc* is a transcription factor that belongs to the bHLH-LZ (basic helix-loop-helix leucine zipper) family and dimerizes with the Max protein to bind DNA. *Myc* has been shown to bind many different sites on the genome, mainly in interaction with other factors. The *Myc-MAX* dimer is expected to preferentially bind E-box motifs (CANNTG) and to promote gene transcription (Zeller et al. (2006)). Although the exact mechanisms of *Myc* function are still under investigation, the *Myc-MAX* dimer has been described to mediate histone acetylation, thus leading to active gene transcription (Dang et al. (1999)). In a tumoral context, it has been shown that a high level of *Myc* is frequently associated with a global transcrip-

tional effect. In addition of accumulating at the promoters of hundred of genes, *Myc* is also able to bind regulatory regions, and is thus though to act as an amplifier of existing gene expression (Lin et al. (2012)).

In order to better understand the role of *Myc*, and more globally the epigenetic alterations that occur when it is induced in the context of liver tumors, we used an hybrid Mouse Cast/FVB model with a liver-specific tetracyclin transactivator regulating *c-myc* transgene (Tet-myc/Lap-tTA model). Mice that over-express *Myc* rapidly develop liver tumors, that can regress when the *Myc* expression is switched off by doxycycline (Kress et al. (2016)). We thus set out to explore the gene expression, histone modifications and chromatin conformation changes that can occur following *Myc* induction in liver tumors, particulary in the context of the inactive X chromosome which represents a powerful model for epigenetics and heterochromatin. The use of highly polymorphic mouse allowed us to efficiently distinguish the genetic and epigenetic profiles of homologous chromosomes, and especially of active and inactive X chromosomes.

Alterations of the X chromosome have been reported in several cancer types including breast cancer, ovarian cancer or medulloblastoma (Chaligné et al. (2015); Jger et al. (2013). Indeed, the X chromosome contains many genes that have been shown to be involved in tumorigenesis through various mechanisms such as point mutation, copy number variation, or loss of heterozygosity. While most of these alterations are at the level of the active X (Xa) chromosome, a few recent studies have started to investigate the potential role of the inactive X (Xi) in cancer. Recent studies have shown that the Xi chromosome can be unstable with a significantly higher rate of mutations compared to autosomes (Jger et al. (2013)). In addition, Chaligné et al. recently demonstrated that several genes could be abnormally reactivated in breast cancer, leading to a labile status of Xi chromosome which is associated with regional changes in H3K27me3 chromatin state. However, so far, the genetic and epigenetic modifcations able to affect the inactive X chromosome and promote cancer progression, remain largely unknown. There is therefore a growing interest in investigating the X chromosme, and the Xi chromosome in particular, in cancer. The Tet-myc model based on hybrid mice that we used here is a unique opportunity to adress this question in the context of the liver and *Myc*-induced tumors.

The work presented below is part of the collaborative project MODHEP (FP7 project coordinated by Bruno Amati) with several partners (Valerio Orlando, Yu Wei, Peter

Fraser, Edith Heard and Bruno Amati). The project is still ongoing and further validations especially on the exact genotype of the mice and its impact on downstream analysis, are still in progress. Data from the project were generated by Ronan Chaligné (Heard lab), Mayra Furlan (Fraser lab) and Juliette Gimenez (Orlando lab). First level data processing have been achieved by Felix Krueger, Aurélie Bousard and myself.

4.2 Results

4.2.1 Characterization of Tet-myc tumor samples

In order to assess X-inactivation status, Tet-myc x Lap-tTA mice were crossed to Cast (Cast-Eij) mice (Y. Wei, Institut Pasteur). The SNPs information between the two strains can then be used as a read-out to obtain allelic expression/enrichment information. Of the 11 liver samples initially collected, 9 had developed tumor nodules within 6 weeks as already reported (Kress et al. (2016), Figure 4.1). Due to the randomness of the X chromosome inactivation, this leads to mixed populations of cells with either one or the other X chromosome inactive. However clonal populations of cells are required to investigate the status of the Xi chromosome. Clonality assessment was thus performed on pieces of all tumors based on allelic expression of *Atrx*, *Xist* and *Rnf12* X-linked genes. Samples were considered clonal when average expression of the three genes was higher than 80% from one allele. Based on these results, we selected 4 tumors for downstream analysis (RC3, RC5, RC10, RC11). Of note, the RC10 and RC11 tumors were extracted from different nodules of the same mouse. In addition, the Tet-myc x Lap-tTA mice were considered as FVB_NJ mice, but further investigations are currently ongoing to confirm their exact genotype.

As a control to evaluate the status of the Xi chromosome in normal liver, we generated Xist -/+ hybrid mice (C57BL6/Cast-Eij), for which the Xi chromosome is always the Cast allele due to the fact that Xist is deleted on the Bl6 allele and thus only the Cast allele will inactive. We then extracted normal liver cells from these XistKO mice and used them as a wild-type control to compare the epigenetic status of the Xi in normal and tumors.

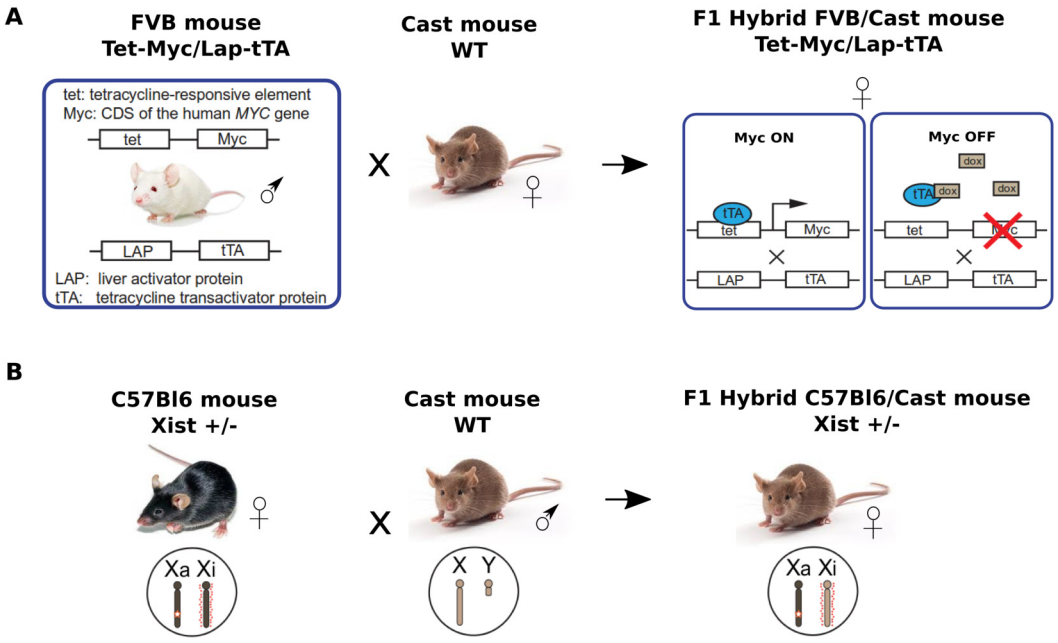


Figure 4.1: Tet-Myc and Xist +/- mouse models - **A.** Schematic representation of the Tet-myc/Lap-tTA mouse model. **B.** Schematic representation of the XistKO mouse. Adapted from Kress et al. and L. Marion-Poll.

DNA copy number variations (CNVs) are an important component of genetic variation in tumors. We thus first generated Hi-C data on liver normal and tumor data, and inferred the copy number status from the observed genome-wide contact maps (Figure 4.8, see Methods and Chapter 3.2). As expected the XistKO sample is fully diploid, while the tumors harbor distinct CNVs profiles. The RC10/RC11 tumors are near-diploid tumors, whereas RC3 and RC5 are karyotypically more rearranged. Most of CNVs changes involve loss or gain of complete chromosomes.

4.2.2 *Myc* enhances global gene transcription in tumors

Two different models have been proposed to explain how *Myc* binds to DNA. *Myc* has been shown to directly affect the expression of a subset of genes. In this case, *Myc* specifically binds the promoter of genes with a preference for the E-box consensus (CANNTG) motifs. In addition, in a tumoral context, *Myc* overexpression leads to an invasion of almost all active promoters, suggesting a role for *Myc* in promoting

transcriptional elongation in interaction with other protein complexes (Sab and Amati (2014)).

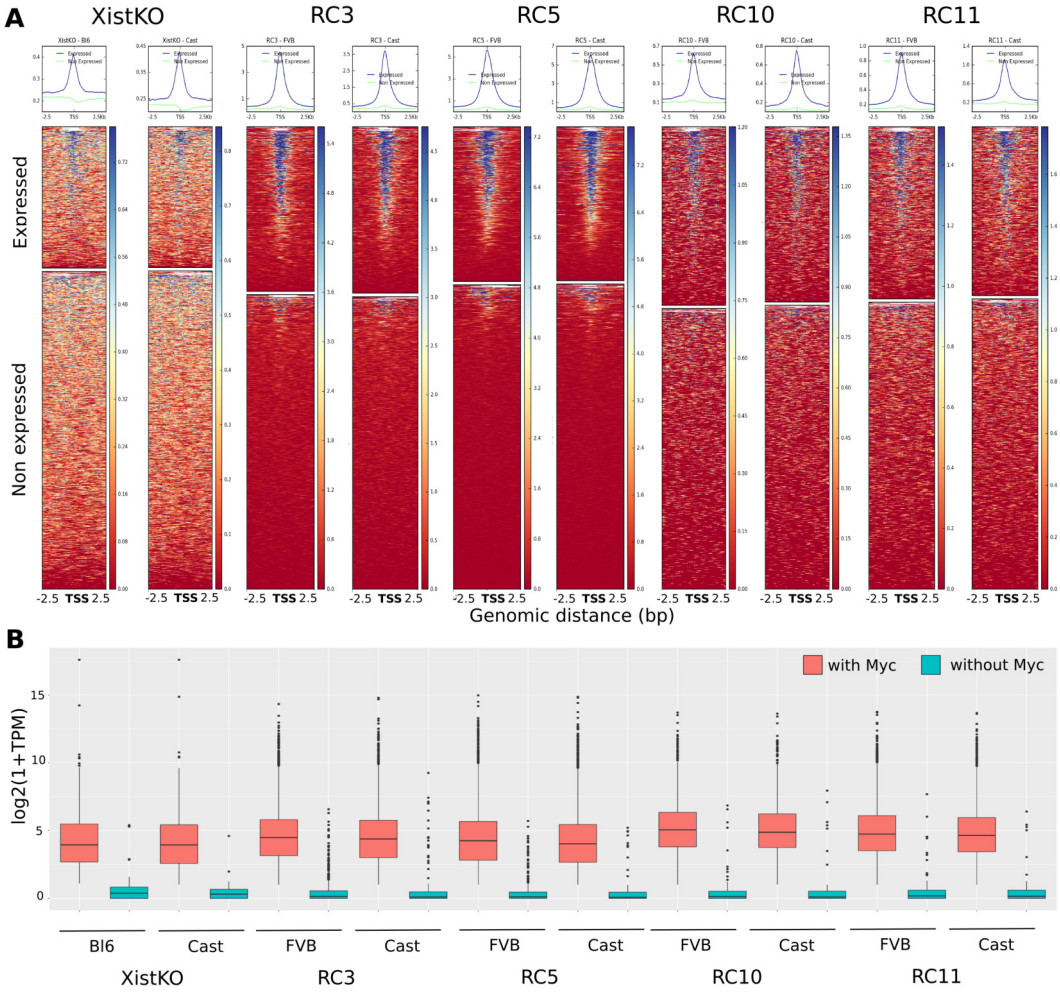


Figure 4.2: *Myc* binding at the gene promoters - A. Enrichment profile of *Myc* ChIP-seq samples around transcription start sites (TSS) of expressed and non expressed genes for both parental alleles. Most of autosomal expressed genes are associated with *Myc* binding on both alleles. **B.** Genes associated with *Myc* binding have a higher expression level than genes without *Myc*.

In order to explore the impact of *Myc* expression at the transcriptional level, we performed allele-specific RNA sequencing and chromatin immunoprecipitation sequencing (ChIP-Seq) analysis of the *c-myc* protein on the normal XistKO and tumor samples. Examination of at *Myc* binding sites on the XistKO samples confirms that *Myc* oc-

cupies the promoter of a few hundred expressed genes (TPM > 1, Figure 4.2A). As expected, a large fraction of *Myc* peak contains at least one E-box consensus motif (CACGTG, 46%). Motifs associated with *Myc* co-factors such as *Yy1*, *Sp1* or *Srefb1* were also detected (Figure 4.9). Increasing the *Myc* expression level in tumors has a significant effect on the total number of gene promoters bound by *Myc*.

As expected, *Myc* is associated with the most highly expressed genes in all samples, therefore supporting its role in transcription enhancement (Figure 4.2B). On tumor samples, the binding of *Myc* at promoters looks less specific with only 14% of CACGTG motif detected in RC10/RC11 tumors, and no significant E-box motif found in RC3/RC5 tumors. Interestingly, other binding motifs for proteins associated with *Myc* oncogenic activity such as *Nrf1* or *Nfat* binding sites were also detected in tumors (Morrish et al. (2003), Kenig et al. (2010)). Finally, we found that *Myc* induction has globally the same effect on both parental sets of autosomes. Indeed, in average 83% of *Myc* peaks detected on automosomes were found to bind the two homologuous chromosomes without obvious distinction at the promoter level (ratio ≥ 0.2 and ≤ 0.8).

4.2.3 *Myc* induction does not lead to global reactivation of the inactive X chromosome

We then explored whether *Myc* over-expression could have an impact on the global Xi chromosome status. Using allele-specific RNA expression, we first validated that most autosomal genes have a bi-allelic expression, compared to X-linked genes that are mostly expressed from the active X (Xa) chromosome (Figure 4.3A). Among the different epigenetic changes that occur during X inactivation, the acquisition of global H3K27me3 histone mark has been extensively studied (Marks et al. (2009)). We therefore analyzed H3K27me3 ChIP-seq data from both XistKO normal liver and *Myc*-induced liver tumor samples and assigned allele specific reads to both parental alleles. We then calculated the global H3K27me3 enrichment on automosomes and X chromosomes, and observed that the Xi is enriched in H3K27me3 marks in both normal (log2 enrichment of 1.11) and tumor samples (mean log2 enrichment of 0.89). Intriguingly, we also observed a slight H3K27me3 enrichment on the active X of some tumors which might be indicative of changes in chromatin organization of Xa chromosome. All together, these analyses suggest that the global H3K27me3 and gene activity status of the Xi chromosome in tumor samples is not affected by *Myc* over-expression.

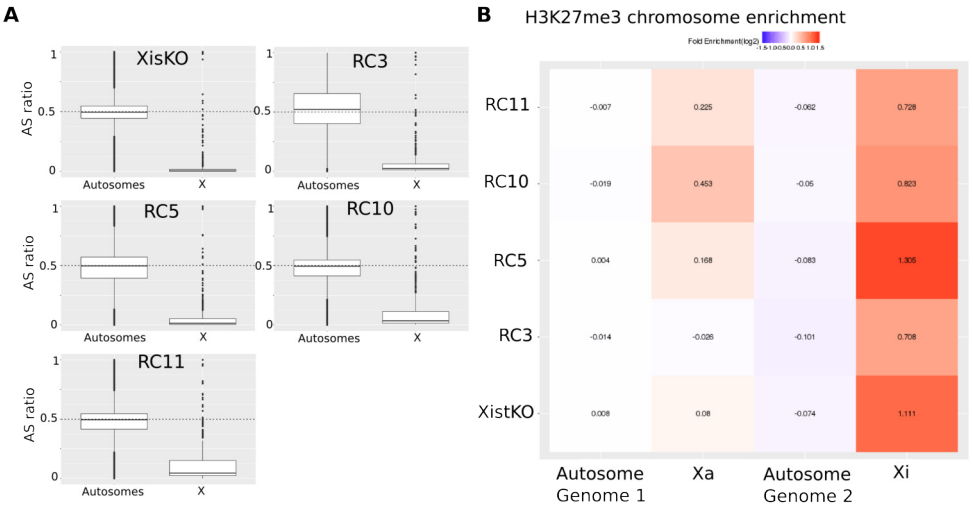


Figure 4.3: Global active and inactive X status in normal XistKO and liver tumors - A. Allelic ratio of expressed genes on autosomes and X chromosomes. XistKO are normal liver samples (n=2). RC3, RC5, RC10, RC11 are *Myc*-induced liver tumor samples (n=1).**B.** H3K27me ChIP-seq enrichment on both autosomes and X chromosomes compared to the input signal. Median signal of both H3K27me3 and input ChIP-seq were calculated per chromosome. The enrichment was then calculated as the log2 fold-change of H3K27me3 over input signal.

4.2.4 New genes escape X inactivation in *Myc*-induced liver tumors

Although the inactive X appears to remain globally silent in the tumor analyzed, we wished to characterize gene expression status more specifically on the Xi chromosome. We analysed the allele-specific ratios of expressed genes on the X chromosomes. In addition, to improve the accuracy of our allele-specific analysis, we performed allele-specific mapping of ChIP-seq experiments targeting H3K9ac and H3K27ac histone marks, and classified genes that showed partial or no X-chromosome inactivation as escapees using both expression and K27ac or K9ac histone marks (see Methods for details, Figure4.10). Interestingly, we observed that the RC10 and RC11 tumors have a higher fraction of escapees (8%) compared to normal XistKO or to RC3/RC5 tumor samples (3%, Figure 4.4A). We then defined as constitutive escapees all genes that escape X inactivation in the XistKO sample and were already reported in the literature (Li et al. (2016)).

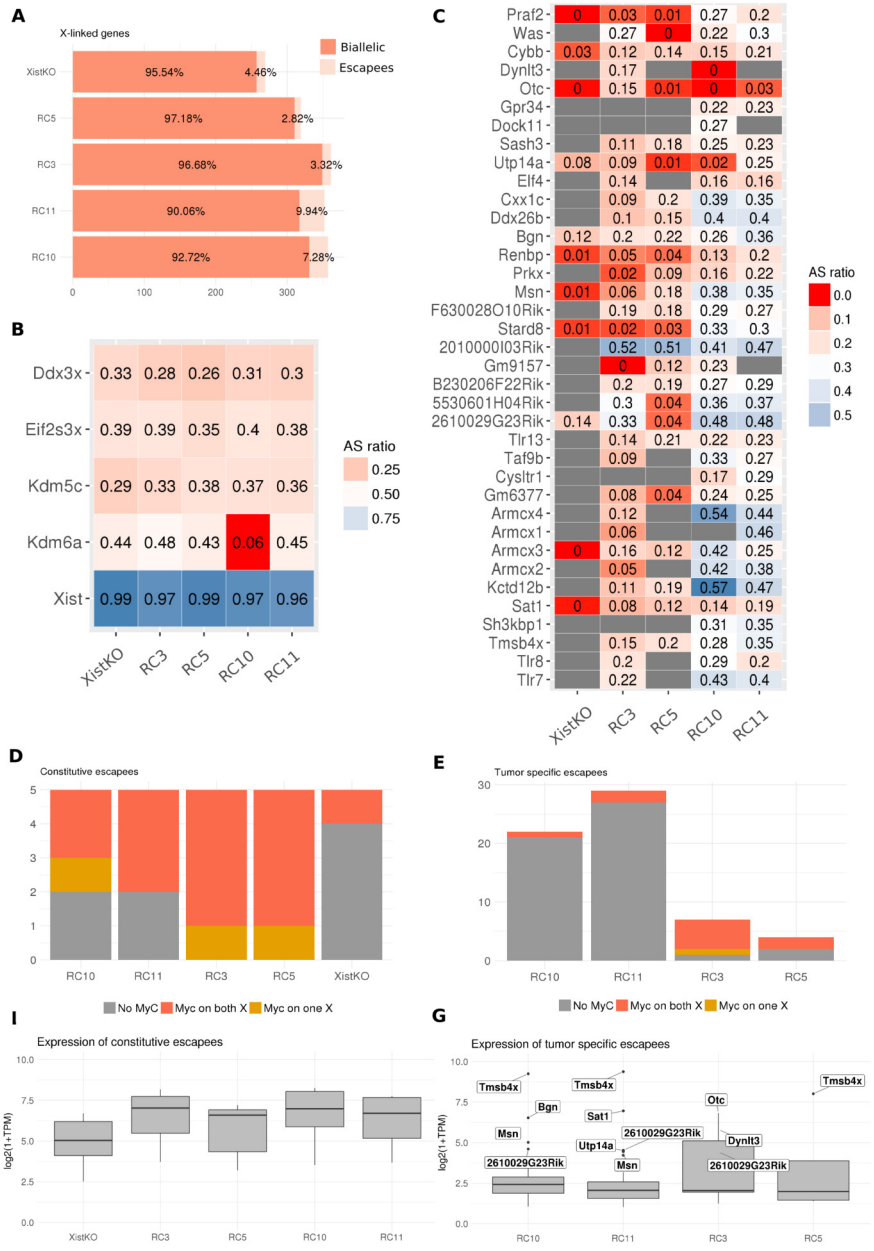


Figure 4.4: Constitutive and tumor-specific Xi escapees - **A.** Fraction of genes called as escapees in XistKO and tumor samples based on allele-specific expression ratios and allele-specific H3K27ac/H3K9ac histone marks.**B.** Heatmap of allele-specific expression ratios of constitutive escapees, with strictly Xi expression in blue and strictly Xa expression in red. **C.** Heatmap of allele-specific expression ratios of tumor specific escapees with strictly Xi expression in red and bi-allelic expression in blue. Genes in gray are not expressed (TPM < 1). **D.** Status of *Myc* binding on the promoter of constitutive escapees (gray ; no *Myc* binding detected, red ; *Myc* binds both alleles, yellow ; *Myc* binding was detected in only one allele.) **E.** Status of *Myc* binding on the promoters of tumor specific escapees. **F.** Expression level of constitutive escapees. **G.** Expression level of tumor specific escapees. Genes with an expression value higher than 20 TPM are indicated.

Constitutive escapees (*Xist*, *Kdm5C*, *Kdm6a*, *Eif2s3x*,*Ddx3x*) were globally highly expressed in all samples (Figure 4.4F), and are associated with *Myc* binding at their promoters on both alleles in liver tumors samples (Figure 4.4D). The *Jpx* (*2010000i03rik*),*Ftx* (*B230206F22Rik*) or *5530601H04Rik* genes, which are frequently described as escapees, are not expressed in the normal XistKO sample (<1 TPM), whereas they are found lowly expressed in tumor samples (2.6 TPM in average). Interestingly, the *Kdm6a* gene (*Utx*, a histone demethylase) was found to no longer escape in the RC10 tumor sample. We defined as tumor-specific escapees those genes found as escapees in the Xi chromosome of tumor samples but not in normal XistKO samples (Figure 4.4B,C). We thus detected 37 tumor-specific escapees, mainly in the RC10/RC11 samples. Interestingly, 25 of these genes are not reported as known escapees (Li et al. (2016)). The majority of these tumor-specific escapees are expressed at lower levels in the Xi (range of median expression in tumors : [2.21 - 2.81] TPM) compared to Xa chromosome (range of median expression in tumors : [3.72 - 6.04] TPM), and were not associated with *Myc* binding at their promoter (Figure 4.4E,G), therefore suggesting that different mechanisms could be involved in constitutive or tumor-specific escapees.

4.2.5 Chromosome organization in normal and *Myc*-induced tumor liver cells

To further investigate the structural changes in Xi organization and how this could potentially relate to increased escape from X chromosome inactivation, we explored the chromatin structure of the Xa and Xi chromosomes using allele-specific Hi-C data from tissue samples of the same source as above for ChIP-seq and RNA-seq (normal XistKO and *Myc*-induced tumors). First, using Hi-C data at 40Kb resolution, we investigated how TADs are organized in normal (XistKO) and *Myc*-induced tumor liver cells (Figure 4.5). In other mouse cell type, such as Neural Progenitro Cells (NPC), the Xa chromosome is structured in sub-megabase TADs and compartiments whereas the Xi chromosome globally presents no TADs or compartiments (Giorgetti et al. (2016)). In the following analysis, we focused on TADs. In addition, to avoid any technical bias in Hi-C experiments, we generated new Hi-C data from NPC using exactly the same experimental protocol than for normal XistKO and *Myc*-induced tumors.

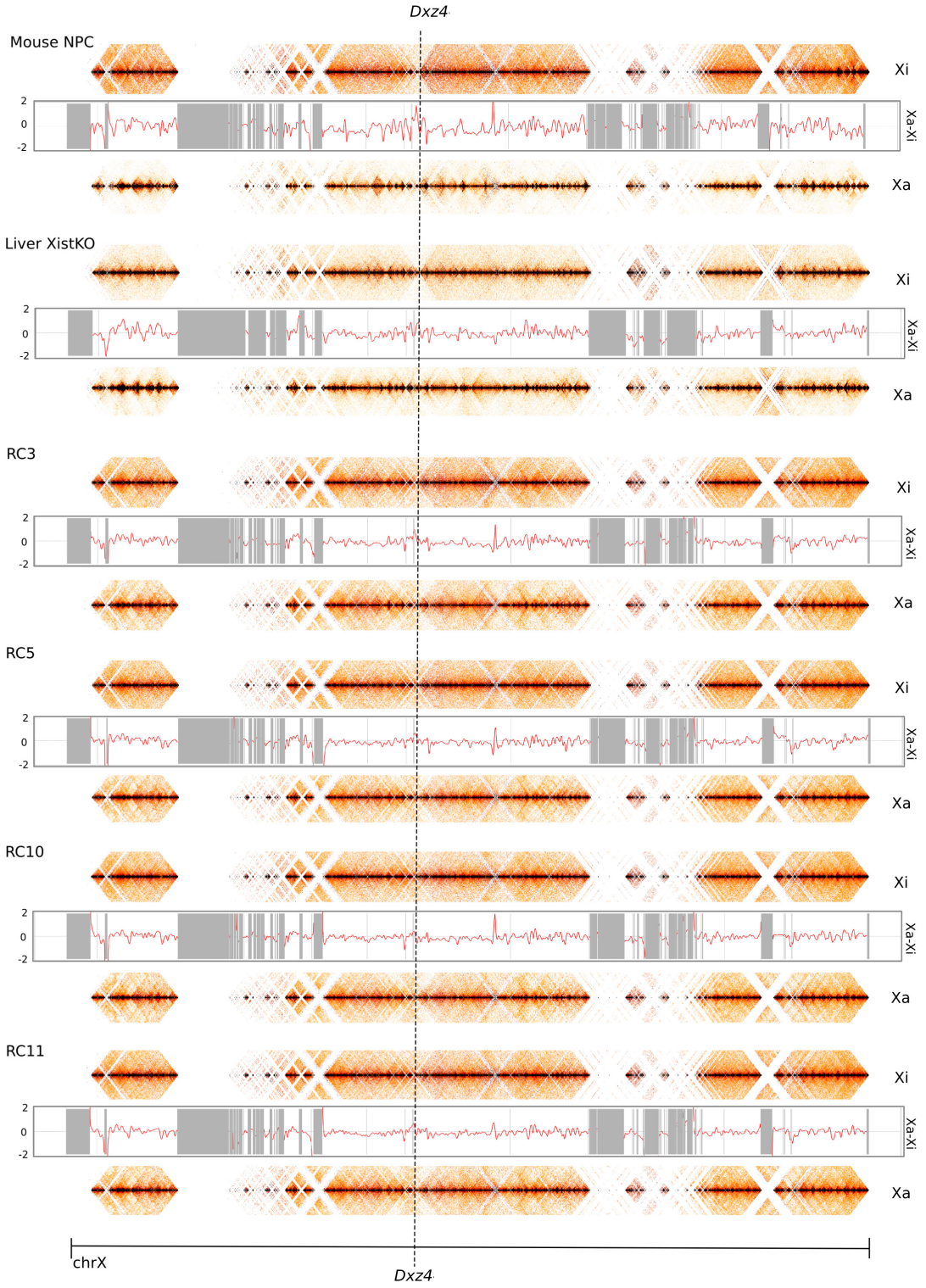


Figure 4.5: Insulation profiles of liver XistKO and tumor samples - Allele-specific Hi-C contact maps of short range contacts along the diagonal (40kb resolution). Active and inactive X chromosomes of each sample are represented, as well as the difference of insulation scores between both chromosomes (see Methods for details.)

Surprisingly, in XistKO liver cells, differences between Xa and Xi chromosomes appeared to be less striking than in NPC cells (Figure 4.5). Although, as expected, the Xa chromosome is more structured than the Xi chromosome. However, on the Xi chromosome of the XistKO sample, more TAD-like structures could be observed compared to NPC cells, suggesting a partial organization of the Xi in liver cells at the sub-megabase level. This observation raises the question of inactive X chromosome organization between different cell types, i.e NPC versus liver cells. One of the major difference between these cell types is their proliferation status. While NPC cells are higly proliferative, liver cells are more quiescent and are maintain in a poised state with minimal basal activity. One explanation would therefore be that the specific Xi and Xa conformations that we observed on liver cells is driven by the cell states and by their proliferative status. However, in the liver tumors (which are more proliferative), the differences in insulation profiles between the Xa and the Xi become even weaker (Figure 4.5). This suggests hat the differences in Xi structure between liver (normal and tumor) samples and NPCs could be at a different level. For example some of the architectural proteins that are thought to participate in TAD organisation may be more differently expressed, or else the epigenetic status of the Xi means that it is less inert and more prone to showing TAD-like organisation. This could be due to low-level expression that we were not able to pick up in our assays but that are sufficient to participate in enabling some degree of TAD organisation.

To validate our observations based on insulation profiles, we also converted the insulation scores into discrete TAD regions on Xa and Xi chromosomes (Crane et al. (2015)). As above, we observed that the normal liver cells (XistKO) have less TADs on their Xa chromosome compared to NPC cells, and more TADs on their Xi chromosome (Figure 4.6B). In NPCs it has been shown that all regions showing TAD-like structures on the Xi correspond to regions that are expressed (ie escape) from the Xi. This observation is suprising as very few number escapees were detected in the XistKO liver sample. In the tumor samples, the number of TADs was more variable, and correlated better with the number of escapees detected from RNA-seq experiments. Thus the status of the Xa and Xi may be different in the normal liver and tumor samples, and the reasons for TAD-like structures on the Xi may also be different.

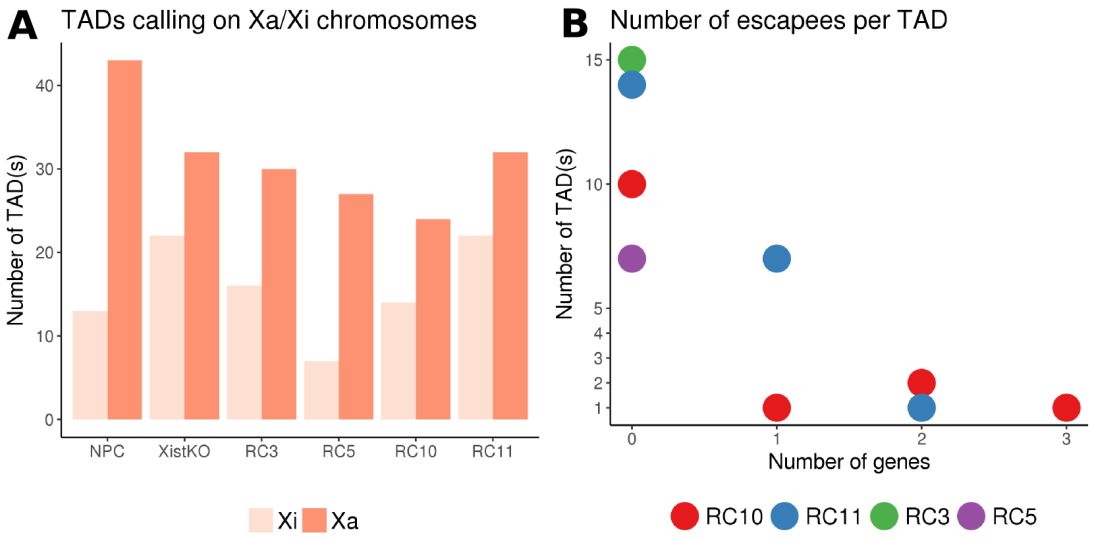


Figure 4.6: TADs in liver XistKO and tumor samples - **A.**Number TADs detected on Xa and Xi chromosomes. **B.** Number of tumor specific escapees per TAD.

We then assessed whether genes that escape X inactivation are clustered in TADs along the genome. We indeed detected a few genes belonging to the same TAD such as the Tmsb4x, Tlr8, Tlr7 or the 5530601H04Rik and pbdc1 genes. Otherwise, others escapees are either found alone in a TAD, or no TADs structure was detected by the insulation score (Figure 4.6B). This is similar to what was previously reported for NPCs. Overall, when we looked at the Hi-C and ChIP-seq data, we observed that constitutive escapees are characterized by a strong insulation, a clear enrichment in active marks at the promoter level (H3K9ac, H3K27ac), and a depletion of repressive mark (H3K27me3). On the other hand, tumor specific escapees present a weaker ChIP-seq signal which correlates with a lower gene expression level (Figure 4.11 for examples).

4.2.6 The *DXZ4* mega-domains boundary of liver tumors is weaker in liver tumors

A bipartite structure of the inactive X chromosome was recently described in human and mouse cells using Hi-C profiles from lymphoblastoid cell lines (Rao et al. (2014)), brain and Patski fibroblasts (Deng et al. (2015)), as well as NPC (Giorgetti et al. (2016)).

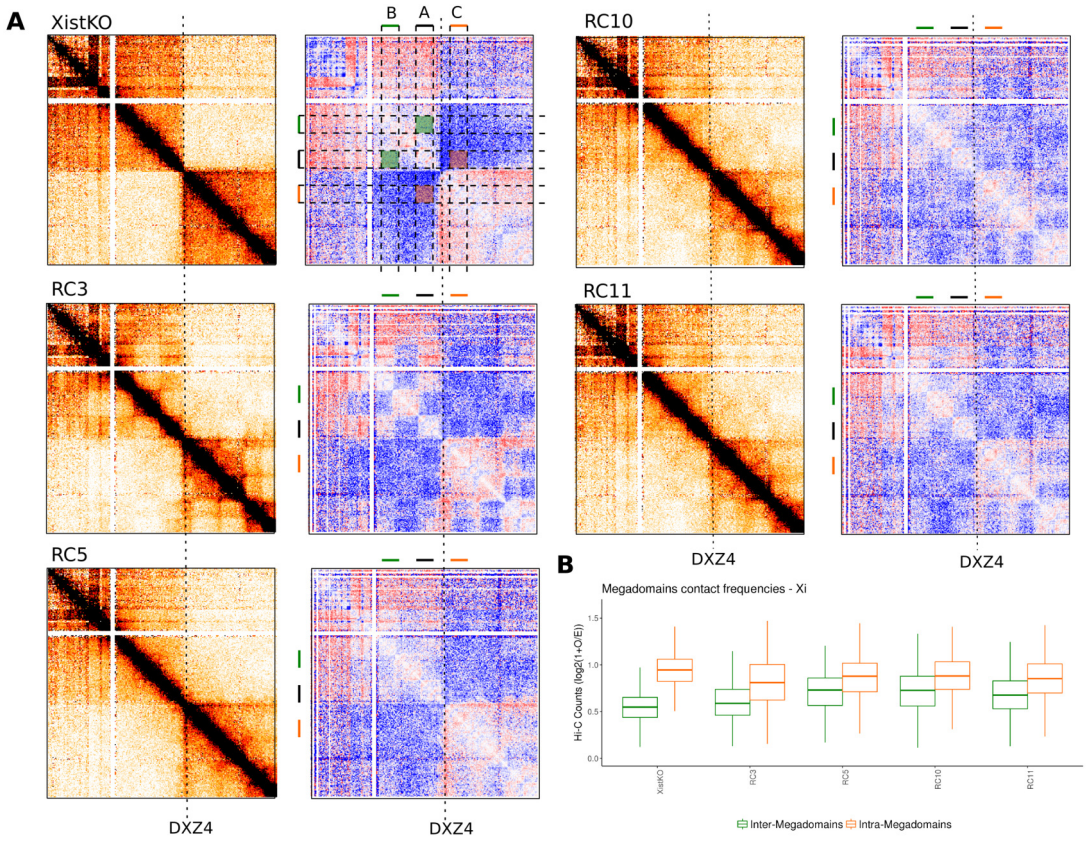


Figure 4.7: Inactive X chromosome mega-domains in XistKO and tumor samples - **A.** Allele specific Hi-C contact maps, and observed/expected (O/E) maps of Xi chromosome near the *DXZ4* mega-domain boundary (see Methods). The three regions (A, B, C) are equally distributed over the boundary (5Mb), and can therefore be used to compare the contact counts within the mega-domains and spanning the boundary **B.** Boxplots of O/E contact frequencies between A-B probes (intra-domain contact, green) and A-C probes (inter-domain contact, orange), as illustrated on XistKO contact maps. Differences between A-B and A-C contact frequencies are much stronger in XistKO samples compared to liver tumor samples, suggesting a partial loss of megadomains in a subset of cells.

These studies have shown that the inactive X is partitioned into two large interaction domains (mega-domains) separated by a boundary region near the *DXZ4* macrosatellite region, previously shown to have an unusual epigenetic status and to be highly enriched in CTCF binding (Horakova et al. (2012)).
Zooming into the *DXZ4* region in the XistKO liver sample revealed a strong boundary that separates the Xi chromosome into two large mega-domains as already described (Figure 4.7A). Interestingly, in liver tumor samples, the Hi-C signal at the bound-

ary appears to be weaker, suggesting a potential loss of mega-domain segregation in a subset of cells. The comparison of contact frequencies between regions within the same mega-domain, and regions spanning the boundary clearly shows a decrease in the mega-domain strength, in agreement with our observation on the normalized contact maps (Figure 4.7B). The RC10/RC11 tumors show the most striking loss. Although the precise relationship between the presence of mega-domains and expression on the Xi chromosome remains to be explored, it is tempting to speculate that the higher number of abnormal escapees found in RC10/RC11 tumors and Xi chromosomal organization changes such as the loss of megadomains are somehow related.

4.3 Discussion

We have conducted an in-depth analysis of allele-specific gene expression, histone modifications and chromatin structure of the inactive X chromosome in normal liver and tumors induced by *c-myc* over-expression. Our results confirm the previous conclusions made on breast cancer data (Chaligné et al. (2015), suggesting a potential disruption of the inactive X chromosome in cancer and different degrees of aberrant expression from the Xi, although global silencing is retained. The changes in Xi organization may be related to more labile transcriptional states, with new genes found to escape from X inactivation. Notably, in the tumors subject to permissive transcription, we also observed a decrease in the intensity of mega-domain boundary suggesting a link between transcription and the Xi chromatin structure. The decrease in mega-domains boundary formation could be more a consequence of a generally more active state on the Xi chromosome, rather than a cause of the transcriptional changes detected.

At the sub-megabase scale, our Hi-C allele-specific analysis reveals that active and inactive X chromosomes in liver cells tend to have less strinking different structures compared to the insulation profiles on NPC cells. One hypothesis would be that these observations could be related to the state of liver cells, which are more quiescent and less proliferative. This could also explain why the active X chromosome of liver cells harbors less TADs compared to NPC cells, if TADs appearance and maintenance is linked to cycling cells as recently proposed (Nagano et al. (2017)). Thus, so far, the precise nature of link between *Myc* induction and the organisational and transcriptional status of the Xi chromosome in tumors remains elusive. While most expressed genes

are bound by *Myc* at their promoter, we did not find a significant *Myc* enrichment in tumor-specific escapees. This therefore begs the question as to the mechanisms underlying the expression of these escapees, which does not seem to be directly mediated by *Myc*.

Finally, this project is still ongoing and several technical and biological questions still need to be adressed. The final conclusions of our analysis depend on two important piece of information: the mouse genotypes of each sample and the clonality of the tumor samples. Regarding the mouse genotype, the genotype of the active X chromosome in XistKO sample is not of a pure genetic background, and the mixture of strains means that accurate SNP detection is an issue. While the Xi chromosome is of pure Cast origin, the mixed genotype of the Xa chromosome can lead to innapropriate assignment of some reads to the Xi chromosome. Although the XistKO data has now been corrected, additional validations must be performed. A similar situation is found for the tumor samples, for which the exact mouse FVB genotype has to be confirmed. One way to validate the mice genotype would be to sequence their entire genome, and to directly call the heterozygous SNPs from these data. Having the genome sequence would be the best way to discard any ambiguous SNPs information. Alternatively, as ChIP-seq experiments were already performed, another idea would be to use the input samples to call for heterozygous SNPs. However, this approach requires a sufficient sequencing depth (at least 30X) to perform the variant calling. The question of the tumor clonality also turned out to be a challenge. For the ChIP-seq and RNA-seq samples, the clonality of all samples was assessed by analyzing the *Xist* gene which is expressed from only the inactive X. Samples showing too much reads assigned to the active allele should be discarded. For the Hi-C data, validating the clonality looks much more complicated if not impossible. Finally, the XistKO model also need to be further validated. Currently, the consequences of the Xist knock-out in one allele are unknown. Does it expain, for instance, the low number of escapees detected in the XistKO sample ? Being able to answer these questions would be of interest to confirm our current results.

4.4 Methods

4.4.1 Allele-specific mapping

In order to avoid biases in allele-specific analysis, all sequencing data (RNA-seq, ChIP-seq, Hi-C) were processed using the same strategy and aligned on a N-masked reference genome. SNPs position and genotype between Bl6/Cast (XistKO) and FVB/Cast genomes were gathered from the Mouse Genome Project. The SNPsplit software (Krueger and Andrews (2016)) was used to generate the N-masked reference genomes. All SNPs between parental genotypes were therefore replaced by a 'N' base on the reference genome in order to avoid mapping biases. All sequencing dataset were aligned on these N-masked genomes. RNA-seq reads were aligned with the Tophat software (v2.1.0, Kim et al. (2013)) and the Refseq annotation, while ChIP-seq and Hi-C reads were aligned with the Bowtie2 software (v2.2.9, Langmead and Salzberg (2012)). The SNPsplit tool was then used to assign allele-specific reads to either the Cast or the Bl6 (or FVB) genomes.

Of note, Hi-C data has been processed to take into account the potential mixed genotype of the Xa chromosome allele in the XistKO sample (Bl6/Cast). To avoid any bias, we discarded all SNPs positions where the genotype of the mixed strains (129S1, Balb, DBA2J) was equal to the Cast allele, thus avoiding potential wrong assignment to the Cast allele. This additional filtering removes 4 millions (over 20 millions) of usable SNPs.

4.4.2 Sequencing data processing

Once aligned and processed using the SNPsplit software, all aligned RNA-seq reads were used to quantify overall genes expression. Reads counts were estimated using the FeatureCounts software (subreads v15.1, Liao et al. (2014)) and the RefSeq gene annotation. Raw reads count were then transformed into Transcript Per Million (TPM) values to detect expressed (TPM>1) and non-expressed genes (TPM$\leq$1). Aligned reads were then splitted into allele-specific BAM files using the SNPsplit 'XX:G' flag, and allele-specific table counts were generated as previously. Allelic expression ratios were calculated from the allele specific counts using the formula $R = FVB/(Cast + FVB)$ for Tet-myc tumors or $R = Cast/(Cast + Bl6)$ for the XistKO samples.

ChIP-seq data were processed using the same strategy. After alignment on N-masked

mm9 genome, duplicated reads were removed using Picard tools (v1.65) and blacklisted regions from ENCODE consortium were discarded. Peak calling of *Myc*, H3K27ac and H3K9ac marks were performed with the MACS2 software (v2.0.10, Zhang et al. (2008)) using all aligned reads, regardless their allelic status. Allelic ratios were then calculated for each peak using only allele-specific reads as for RNA-seq. In addition, aligned data were splitted into allele-specific BAM files, and genome-wide coverage files (bigwig) were generated using the Deeptools software (v2.2.4). Bigwig files were normalized to 10 millions reads.

Hi-C data were aligned on the N-masked reference genome and processed by the SNPsplit and HICUP pipelines (v0.5.7, Wingett et al. (2015)) to generate the list of valid interaction products. HICUP output files were then converted and imported into the HiC-Pro pipeline (v2.8.0) to generates raw and normalized contact maps. Genome-wide contact maps were normalized using the ICE procedure (Imakaev et al. (2012)) implemented in HiC-Pro, as the CNV biases mainly involves aneuploidy with gain or loss of complete chromosomes.

4.4.2.1 Calling of escapees

Calling of escapees was performed using both allelic ratios of expression and H3K27ac/H3K9ac signals, which are hallmarks of active genes. Only expressed genes (TPM>1) were considered for the analysis. Genes with less than 10 allele-specific reads were also discarded to avoid false positives escapees. In addition, allelic ratio of H3K27ac/H3K9ac histone marks were calculated for each peak and assigned to a gene promoter (2Kb downstream the transcription start sites (TSS)). Finally, X-linked genes were called escapees if i/ their allelic expression ratio was higher than 0.15 ,ii/ there was at least one active mark (H3K27ac or H3K9ac, ratio > 0.2) within the promoter regions of the genes.

4.4.3 ChIP-seq data analysis

Myc enrichment plots presented in Figure 4.2 were generated using the *computeMatrix* Deeptools utility (v2.2.4).

4.4.3.1 Motifs discovery

Motifs detection has been performed with the RSAT online tool (http://rsat.sb-roscoff.fr/). The *peak motif* tool was run on the *Myc* peak sequences overlapping the promoter region of genes (Thomas-Chollier et al. (2012)). The conversion from peaks to genomic sequences was performed with the BEDTools suite (v2.17, Quinlan and Hall (2010)). Default parameters of *peak motif* have been used except the motif discovery which was restricted to over-represented words. Motifs annotation has been performed with the JASPAR core non-redundant vertebrate database. Then, the *matrix clustering* tool was used (Castro-Mondragon et al. (2017)) to clean the *peak motif* results and to define a set of consensus 'root' motifs (Figure 4.9). Finally, the *matrix scan* tool was used (Turatsinze et al. (2008)) with the sequence of *Myc* peaks and the 'root' motifs to precisely detect the number of motifs detected in each sequence.

4.4.3.2 H3K27me3 enrichment

To calculate the H3K27me3 enrichment at the chromosome level, we first summarized the ChIP-seq signal of both input and H3K27me3 marks into 100kb bins, and calculated the median counts of 100kb bins per chromosome. We then centered these values and calculated a chromosome enrichment as the log2 fold-change of H3K27me3 over input signal. Using the input allows to normalize the H3K27me3 signal by the copy number status of each chromosome. We finally represented the median of the enrichment score of all automosomes with the enrichment observed on the X chromosomes. This analysis has been performed using both ChIP-seq and input allele-specific signal.

4.4.4 Hi-C data analysis

To remove potential biases in the Hi-C data related to the SNPs density, we calculated the number of SNPs inside each genomic interval for all Hi-C bins. We then calculated the median number of SNPs per bin, and fixed a minimum required SNP density cutoff (median 1.5 IQR) as already proposed (Giorgetti et al. (2016)). Any bins with less SNPs than the cutoff were removed from all analyses. The SNP density cutoffs used for each bin size were: 40 kb, 43 SNPs; 250 kb, 776.5 SNPs; 500 kb, 1,767.25 SNPs. Insulation score has been calculated using the method described in Crane et al. with the following options : "–is 48000 –ids 320000 –im iqrMean –nt 0 –ss 160000 –yb 1.5

–nt 0 –bmoe 0". TADs have then been called from the insulation score using the *insulation2tads.pl* script, and the option "–mbs 0.2 –mts 0.5".
Observed over expected contact maps (O/E) has been generated using the BioConductor HiTC package (Servant et al. (2012)). Briefly, the expected count matrix was generated by calculating the mean of observed contact frequencies at a given genomic distance. Dividing the observed contact matrices by the expected ones allows to remove the counts $\sim$ distance relationship from the data.
CNV calling from Hi-C data have been performed at 250kb resolution data as already presented in Chapter 3.2.

4.5 Supplementary Figures

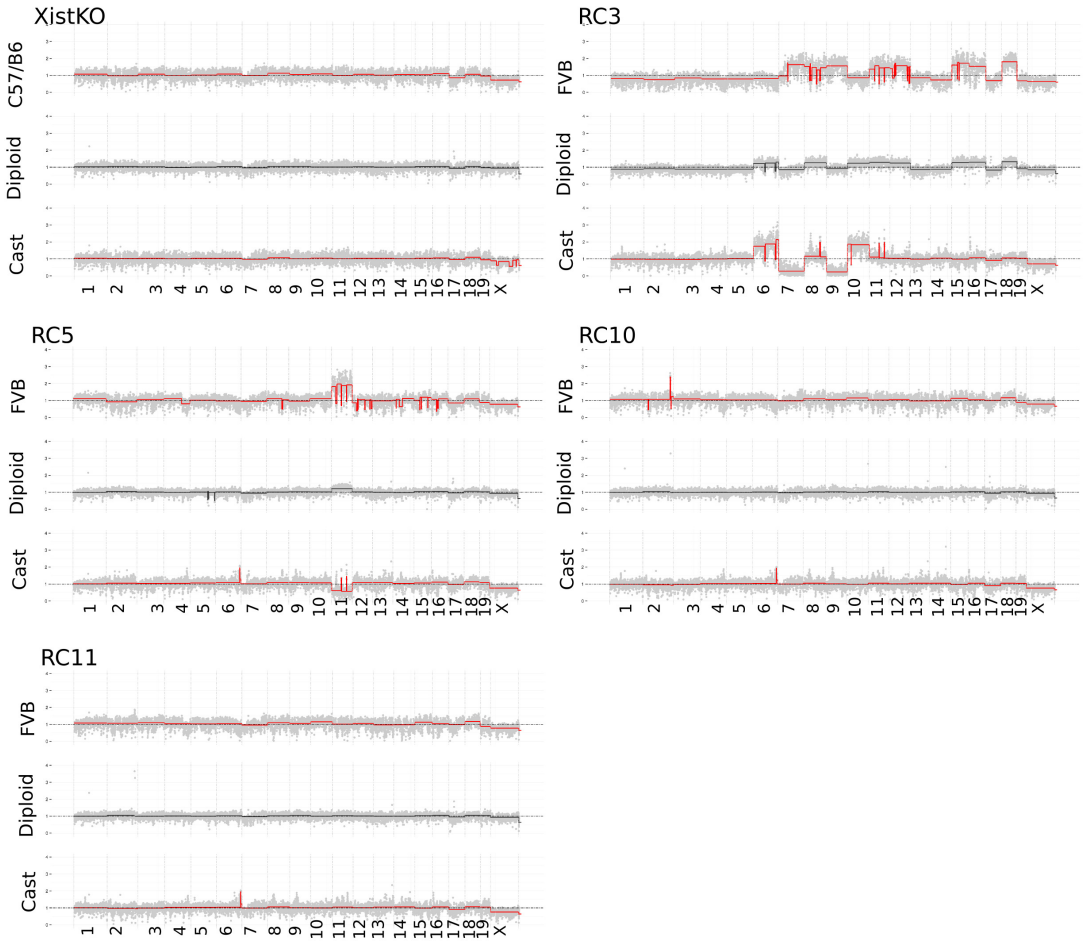
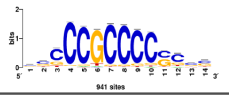
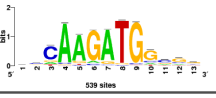
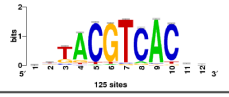
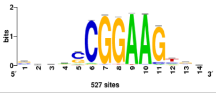
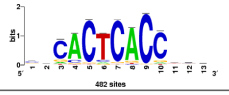


Figure 4.8: Copy number profile of Tet-myc and XistKO liver samples - Copy number profiles have been extracted from Hi-C data as previously discussed. Briefly, Hi-C contact maps were summarized in 1D by summing up the contact of each genomic locus (250Kb resolution). After GC content, mappability and fragment length correction, the 1D profile was segmented to estimate the copy number profile

Xist KO, 373 peaks			
	MNT, MAX, Npas2, ...	(46,65%)	 SP2, SP1, EGR1, ... (55,76%)
	YY1	(44,24%)	 CREB1, ATF1, Creb5, ... (43,16%)
	ELK3, ETV6, ETV5, ...	(51,47%)	 SREFBF1 (51,74%)

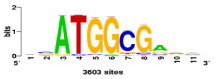
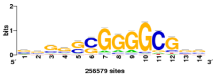
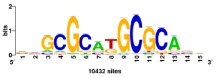
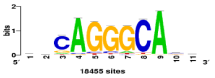
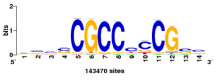
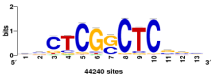
		RC3 16422 peaks	RC5 16791 peaks	RC10 4738 peaks	RC11 5943 peaks
	BLHLHE41, ARNTL, USF1, ...			(14,01%)	(14,15%)
	YY2, YY1			(15,07%)	(15,90%)
	SP2, KLF5 KLF4, ...	(42.12%)	(53,76%)		(26.03%)
	SP2, NRF1	(34,12%)		(21,93%)	(21,12%)
	ELK3, ETV6, ETV5, ...	(22,24%)		(14,52%)	(16,49%)
	NFATC2, NFAT5	(20,20%)	(23,97%)		
	TFAP2	(12,43%)			
	EGR1, SP2, SP1, ...		(48,35%)	(28,32%)	
	*no match*		(33,02%)		

Figure 4.9: *Myc* binding motifs on automosomes - Binding motifs detected on *Myc* ChIP-seq samples. The fraction of peaks with at least one motif is reported for each sample.

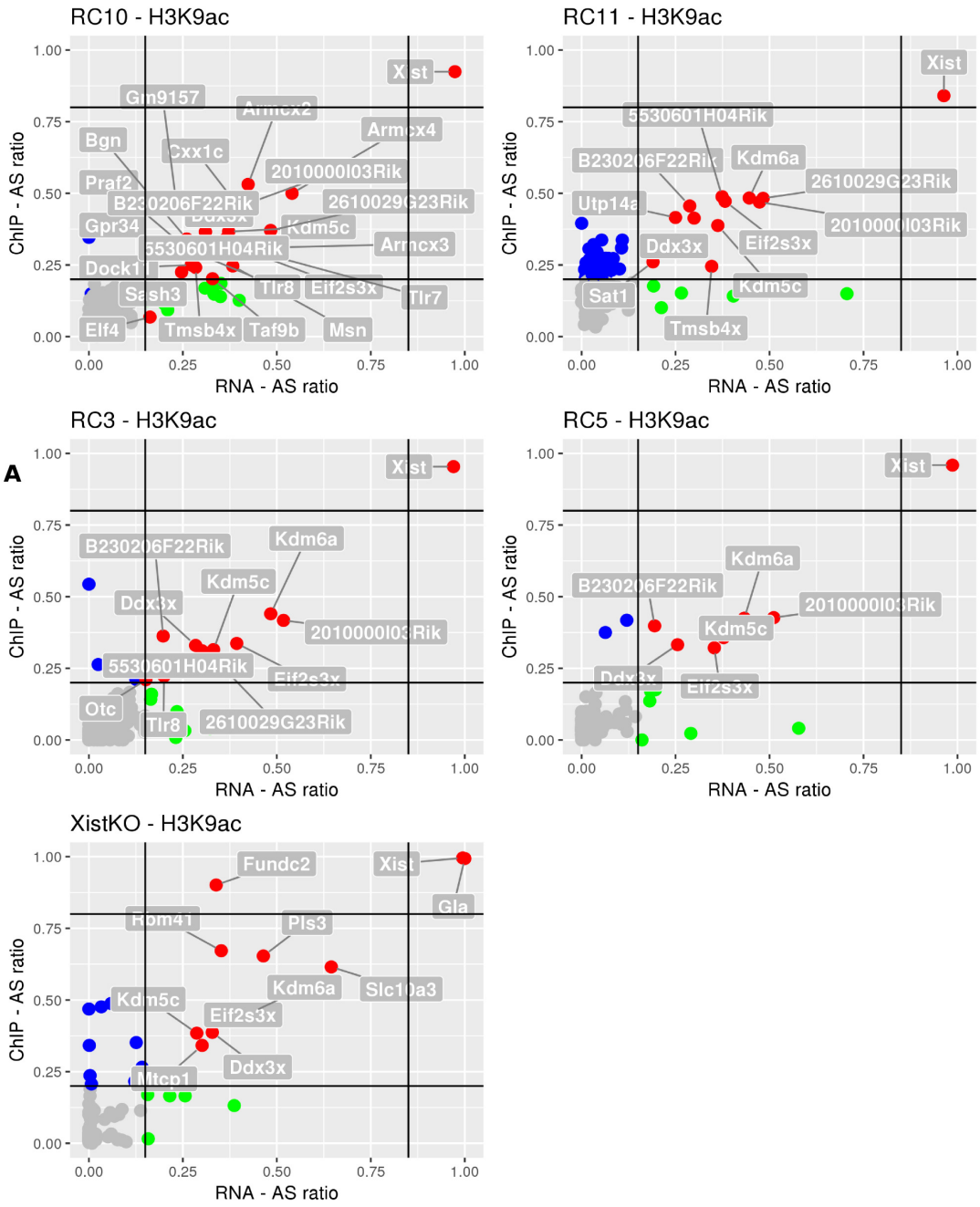


Figure 4.10: Comparison of allele-specific ratio from ChIP-seq and RNA-seq data - Allele-specific ratio of H3K9ac histone mark (y-axis) and RNA expression (x-axis). Red points represent X escapees detected in both RNA and ChIP-seq data. Blue (resp. green) points are ChIP-seq (resp. RNA) specific escapees.

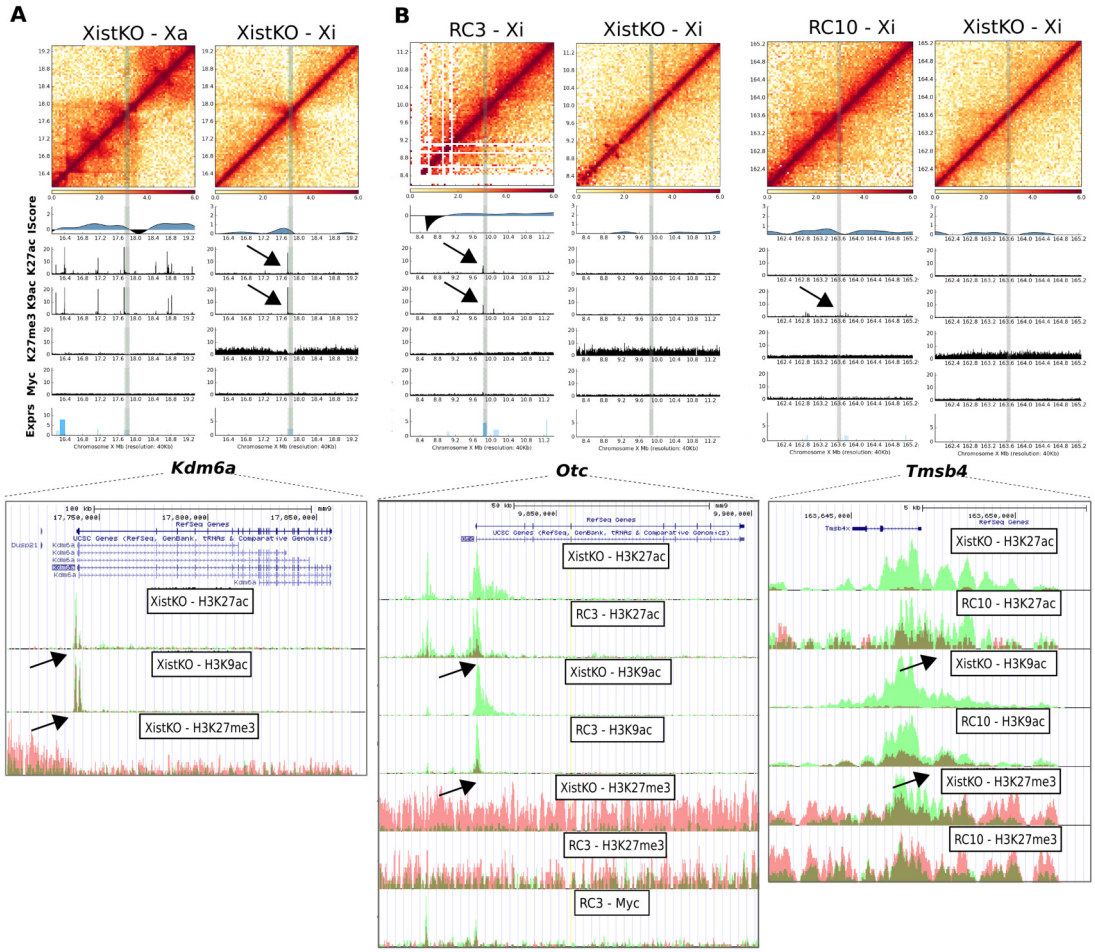


Figure 4.11: Chromatin changes at the escapees loci - A. Exemple of the Kdm6a constitutive escapee. Hi-C profile are represented with insulation score (Iscore), H3K27ac, H3K9ac, H3K27me3, *Myc*, and gene expression profiles. UCSC view centered on the promoter of the gene. Histone profile from the Xa (green) and Xi (red) are superimposed. **B**. Same representation for two tumor specific escapees (Otc in RC3, and Tmsb4 in RC10)

4. *C-MYC* ONCOGENE EXPRESSION AND INACTIVE X CHROMOSOME ORGANIZATION IN MOUSE LIVER TUMORS

5

Discussion & Perspectives

5.1 Chromatin organization and beyond

The development of chromosome conformation capture techniques such as Hi-C, have provided a new view of the genome organization in normal cells. In brief, these data have revealed a landscape of physically interacting regions along chromosomes, separated by hundred of kilobases up to several mega-bases. In addition, they have demonstrated that the chromatin organization follows a hierarchical model, from the formation of a single loop allowing enhancer-promoter interaction, to the formation of large active or repressive domains.

The relationship between this hierarchical structure and the epigenome is still under investigation. For instance, it is still unclear how the genome functions such as transcription, DNA repair, or replication are constrained by the the chromatin structure. In addition, some of these structures, like chromosome compartments, were described to be dynamic and to change from a cell state to another or during differentiation (Dixon et al. (2015)), while some others, like TADs, seem to be more stable. The precise molecular mechanisms involved in these different levels of organization, as well as their functional role, still need to be determined. However, we already have clear evidence of phenotypic consequences that the disruption of the chromatin organization can occur, leading in some cases to disease. These observations therefore demonstrate the importance of the chromatin structure for the cellular functions.

In the following section, we will discuss some of the current questions about chromatin organization that need to be addressed. We will also discuss the interest of looking at

the chromatin conformation in a disease context and in particular in cancer.

5.1.1 Chromatin organization and TADs : cause or effect ?

The discovery of the TADs structure and how they promote enhancer-promoter contacts is one of the most exciting discoveries of recent years. However, while TADs are still the subject of many studies, it still lacks a clear molecular definition of what is exactly a TAD ? and if they all rely on the same mechanisms and/or have all the same function ?

Almost all computational methods to identify TADs rely on visual inspection of contact maps, looking for boundaries splitting chromatin regions into domains. Thus, the detection of the boundaries highly depends on arbitrary thresholds and computational choices. As a consequence, the TADs number and size can vary from a study to another. The first studies reported TADs with an average size of 800 kb. Then, other studies reported the presence of sub-TADs below the mega-base resolution (Phillips-Cremins et al. (2013)), up to single chromatin loop (Rao et al. (2014)). From a molecular point of view, the difference between TADs and sub-TADs remains unclear, and their definition is so far, highly dependent on the resolution used for the data analysis. In this regard, it therefore raises the question of whether TADs, sub-TADs, or loops are really different ? They may simply represent the same features of the genome organization, representing different hierarchical levels with different dynamic states.

It remains to be elucidated how these domains are formed and maintained, and whether they represent the fondamental unit of the genome organization or if they are the consequences of its underlying activities. So far, the main indication that TADs and sub-TADs could represent different features of genome organization is that TADs have been described as relatively invariant and stable structures, whereas changes have been reported at the sub-TADs level during cell differentiation (Dixon et al. (2015); Phillips-Cremins et al. (2013)). In addition, TADs, but not sub-TADs, have been shown to be associated with replication domains, suggesting that they could play distinct roles in regulating DNA replication (Pope et al. (2014)).

Then, the precise mechanism leading to TADs formation is currently not fully understood. So far, the TAD insulation has been mainly associated to the presence of CTCF/cohesin factors at their boundaries. Recently, Nora et al. (2017) proposed a degron system in Mouse embryonic stem cell allowing to deplete (and restore) CTCF

from chromatin. They demonstrated that CTCF is absolutly essential for the insulation of TADs. Interestingly, when CTCF is depleted, Nora et al. reported that around 20% of TADs boundaries are still detected, suggesting that other mechanisms might also contribute to TADs insulation. In addition, the majority of CTCF sites are found within TADs, again suggesting that the CTCF binding cannot be considered as the sole determinant of TADs organization. Finally, CTCF alone is not sufficient to promote TADs formation, and other complexes as mediator or cohesin are also involved in chromatin folding and enhancer-promoter interactions. The molecular details of CTCF/cohesin bindings and how they impact the regulation of transcription as well as other epigenomic marks are still poorly understood. And the precise role of cohesin sub-units and other insulator proteins remain to be addressed. The current model is that cohesin is involved in the packaging of the chromatin fiber while CTCF locally defines the TADs boundaries. This so-called 'loop extrusion' model has been recently proposed based on polymer physics and modeling. It relies on the idea that a DNA loop can be generated dynamically by a pair of extruding factors (cohesin) that move along the DNA in opposite direction until they reach a boundary element (Fudenberg et al. (2016)). While this model is currently favored, it is mainly based on simulations and assumptions, and therefore requires further validation. For instance, if cohesin plays the role of extruding factor, it is unclear how the complexe is loaded and stably associated with the chromatin, or how it is maintained to translocate along hundreds of kilobases of chromatin. Finally, while it is now clear that CTCF and cohesin are involved in the TADs formation, how they precisely interact together is still an open question.

Then, it is also obscure whether all TADs have a functional role ? While some TADs explicitly require the presence of the CTCF/cohesin factors at their boundary, others can just be the result of neighboring structures, and might not have the same functional role. Many recent studies have focused on the TADs function. In summary, they suggest that TADs are chromatin units that constrain the DNA and promote enhancer-promoter interactions. However, it is still unclear whether TADs act as transcriptional units, with co-regulation of genes belonging to the same TAD.
At larger scale, there are several evidences that suggest a correlation between spatial genome organization and transcription. For instance, it is known that gene-rich

chromosomes are usually internally located in close proximity, whereas gene-poor chromosomes tend to be located at the nuclear periphery. It has also been shown in several cases that genes can relocate to different nuclear space according to their transcriptional status. But there are also number of studies with discordant results demonstrating that nuclear position and gene expression are uncoupled (see Meaburn (2016) for a review). While these conflicting conclusions on transcription and nuclear positioning would require further exploration, one explanation would be that some genes are more sensitive to their nuclear position than others. It is also likely that when a gene moves to another nuclear space, it also drives the relocalisation of neighboring regions (Meaburn (2016)). At the TAD level there are also several studies reporting that genes located within the same TADs display correlated expression (Le Dily et al. (2014); Nora et al. (2012)). However, the link between TAD and expression remains ambiguous. Are TADs formed according to their genes expression ? or are the genes expressed according to the TAD they belong to ? The first evidence that TADs tend to have a role in gene expression (and not the reverse) is the fact that TADs are well conserved between cell types and species, or during cell differentiation. Recently, Zhan et al. further explore the genes co-regulation during ES cells differentiation, and demonstrate that TADs correspond to the scale which maximize the co-regulation of genes. However, only 10% of TADs were detected as co-regulated among those exhibiting expression changes during differentiation. This result suggests that the co-regulation of genes within a TAD is not a general rule, but rather occurs on a subset of TADs and genes. Finally, the heterogeneity of changes in gene expression observed within TADs is not surprising, as the transcriptional activity also depends on many additional factors such as histone modifications and accessibility to transcription factors. This is also consistent with the fact that, in number of cases, enhancer-promoter interactions are already established prior to gene activation, supporting the idea that transcriptional activation and looping can be unrelated (Jin et al. (2013)).

The same question arises in the context of histone modifications. Dixon et al. first reported that TADs boundaries are enriched in active marks associated with promoters and gene bodies. In addition, TADs boundaries frequently occur simultaneously with changes in heterochromatin distribution. This therefore raises the question of whether TADs formation are mediated by chromatin modifications or, on the contrary, whether TADs constrain the spreading of these marks. So far, as for transcriptional activity,

there are several indications that TADs rather prevent the spreading of heterochromatin marks. First, in absence of H3K9me2 and H3K27me3 inactive marks, the TADs structure is not affected (Nora et al. (2012)) demonstrating that the establishment of TADs is not driven by these marks. Then, although chromatin marks can change during cell differentiation and between different tissues, TADs are usually conserved (Dixon et al. (2015)). However, the mechanism by which TADs would be able to constrain the heterochromatin spreading is unclear. Interestingly, Nora et al. recently described that, removing CTCF from chromatin did not activate the spreading of H3K27me3, meaning that other mechanisms should explain the overlap observed between heterochromatin and TADs.

Finally, Nora et al. (2017) also demonstrated that, while TADs are affected by the loss of CTCF, this is not the case of A/B chromosome compartments. This observation therefore suggests that the molecular components of these two levels of organization are uncoupled and do not rely on the same mechanisms. In addition, chromosome compartments are usually much more dynamic that TADs, as already reported during differentiation (Dixon et al. (2015)) or in cancer cell lines (Barutcu et al. (2015)). Changes in compartment status may therefore control the accessibility of the genome to regulatory elements and affect the expression of a subset of genes, but the precise underlying mechanisms allowing this level of reorganization also remain to be explored.

5.1.2 Spatial organization of cancer genomes : the missing piece of the puzzle ?

Genome alterations, including mutations, structural variations, or copy number variants are hallmarks of most cancer genomes. So far, most of the cancer genomic projects focus on the characterization of these alterations within large cohortes, and tend to detect recurrent alterations that can be associated with clinical features, such as response to treatment, survival or tumour progression. From a molecular point of view, the discovery of such events greatly improved our understanding of the process involved in the initiation and the progression of the tumour. Mutations in oncogene or tumour suppressor genes represent interesting targets for drug therapies and to improve patients outcome. Structural variants leading to fusion genes also provide clear diagnosis information in clinic. In these times of 'personalized medicine', most of the current projects

are based on the genome sequencing of patients with the hope to better characterize each tumour and to deliver the appropriate treatment to a patient in real time.

However, it is now clear that genetic alterations alone are not sufficient to explain the initiation or progression of all tumour types. Flavahan et al. recently reviewed the concept of epigenetic plasticity and its potential to give rise to all cancer hallmarks. Thus, epigenomic alterations in cancer are frequently explored, although their use is mainly restricted to research and are not frequently applied in clinics. Deregulation of DNA methylation or changes in histone modifications have been reported to be important oncogenic factors. The effect of DNA methylation is known to impact the constitutive expression of genes involved in cell cycle, DNA repair, apoptosis or differentiation. In addition, any changes in the activity of enzymes involved in post-translational modifications of histone tails can result in transcriptionnally active or repressed chromatin state. So finally, what would be the gain of exploring the chromatin structure, i.e. the interactome, of tumours ?

From a research point of view, information of the genome organization is one of the missing link between genetic and epigenetic alterations. So far, our view of the genome is mainly linear or limited to the description of single alteration. However, there is now clear evidence that the chromatin structure is highly dynamic, and that, for instance, many genes dynamically colocalize in the nucleus to favor genes activation or repression (Sexton et al. (2007)). This has been shown in prostate cell line which overexpress the ERG oncogenic transcription factor (Elemento et al. (2012)). The overexpression of the ERG factor leads to a global reorganization of the chromatin that allows ERG to activate many genes including those involved in prostate cancer.

In addition, a few recent studies have started to explore the genome organization of cancer genomes and most of them demonstrated that TAD disruption occurs in several cancers and are likely to be a common mechanism in tumorigenesis. Valton and Dekker recently reviewed these studies and proposed two different types of mechanisms in TAD disruption. The first one locally alters the TAD by mutating or deleting its boundary, leading in most cases to a fusion of adjacent TADs. The other mechanism involves genome rearrangements such as translocation or inversion which usually lead to the formation of new TADs. Both mechanisms can be associated to 'enhancer hajacking' as they allow enhancers to activate genes which are normally silenced.

At higher scale, structural variants could also have an impact on the organization of

lamina-associated domains (LADs). In normal cells, regions associated with LADs are usually silenced. But in cancer cells, because of potential genomic rearrangements, these regions can be reactivated and relocated away from the nuclear lamina. The opposite situation can also occur when active genes become associated to nuclear lamina and consequently silenced.
Therefore, the exploration of the chromatin topology could provide new insights into the mechanisms leading to tumorigenesis. Association between microdeletions and chromatin structure has also been reported in prostate cancer demonstrating a significant change in 3D structure of cancer prostate cells (Taberlay et al. (2016)). At short term, a systematic exploration of the chromatin organization of cancer genome could help in understanding the molecular mechanisms of many cancers.

5.1.3 Whole genome sequencing or Hi-C ?

So finally, as we have been witnessing for several years a race to the one that will sequence the higher number of human genomes, why not replacing genome sequencing by Hi-C ? Indeed, all information available from genome sequencing could also be detected in Hi-C. Recently, Harewood et al. demonstrated that copy number and chromosomal rearrangements can efficiently be detected in Hi-C. More importantly, Hi-C might even be more accurate to detect structural variants than genome sequencing as it does not rely on the presence of breakpoint spanning reads, and therefore generate much less false positives. In addition, the sequencing depth requires in Hi-C to detect such events is much lower than in genome sequencing, therefore also decreasing the cost of the experiment. Then, one question that remains to be validated is whether Hi-C can be used to detect somatic point mutations. One limitation of the standard protocols based on restriction enzymes is that the reads coverage is expected to be enriched near restriction sites, thus not ensuring a uniform genome-wide coverage. However, alternative digestion method, such as DNAse Hi-C, should overcome these limitations. Double digestion, or sequencing of longer reads could also be an alternative. To conclude, even if it remains to be validated, Hi-C could in theory be much more powerful than genome-wide sequencing to characterize cancer genome, as it should be able to give access to all genomic alterations in addition to the 3D structure of the genome. If the sequencing cost of such experiment is still expensive, it should no longer be an issue thanks to current advances in sequencing technologies. Large consortium,

like the ENCODE or the 4D nucleome projects should also rapidly start to deeply investigate the chromatin organization of normal and tumoral samples.

5.1.4 Treating the genome conformation

The ultimate goal of any research project on cancer is to improve diagnosis and patients care. While many clinical trials are currently set up, based on DNA (and/or RNA) sequencing of patients in real time, the question arises whether information about the spatial genome organization could emerge as a tool for diagnosis or drug design.
Currently, two main types of epigenetic drugs are being tested ; the HDAC inhibitors in the treatment of lymphomas and the DNMT inhibitors in the treatment of blood cancer. HDAC inhibitors act by relaxing the chromatin structure in order to restore the transcriptional activity of tumour suppressor genes. The DNMT inhibitors prevent abnormal methylation status of genes, therefore reducing the inhibition of genes involved in cell division.
In a treatment perspective, a better characterization of the 3D chromatin structure of cancer genomes could lead to the development to new therapeutic strategies. For instance, one can imagine to combine the CRISPR-Cas9 technology with the genome topology of a patient tumour to directly regulate the expression of driver genes. By inserting a TAD boundary between an oncogene and its enhancer, we should thus be able to inactivate this oncogene. The same mechanism could be used to inactivate genes involved in cell division or migration. The reverse could also be imagined to restore the function of a gene, by disrupting a TAD boundary in order to force new interactions between enhancers and tumour suppressor genes. Obviously, many technical and ethical questions will need to be addressed in the future before this type of TAD modification could be considered. However, there is no doubt that a better comprehension of the genome organization at the TAD level and at higher scale, will lead to new hypothesis and developments to better treat patients.

5.2 Future of Hi-C technique

In the last decade, several chromosome conformation technics have been proposed, each time improving previous limitations and increasing the final resolution. The Hi-C remains the method of choice to explore the chromatin organization genome-wide.

While the in-situ protocol based on restriction enzyme is still commonly used, other approaches like the DNase Hi-C offer a higher effective resolution by avoiding the dependence on a restriction enzyme. However, the cost of such assay remains extremely high, in particular to reach a resolution around the kilobase. As a reminder, Rao et al. generated around 5 billion pairwise contacts to reach the 1 kb resolution of the GM12878 contact map. This sequencing depth represents more than two flow cells on a Hiseq2500 sequencer. Of course, this resolution is not necessary for most of the analysis based on chromosome compartments or TADs calling, but it starts to be required to investigate single loop formation and molecular disruption of specific domains and genes.

Thus, alternative approaches based on targeted sequencing represent an interesting strategy, which should be increasingly used in the coming years. Promoter capture-C for instance, allows to enrich a 3C library in all interactions involving a list of known promoters. In the same way, capture Hi-C allows to explore, at very high resolution, the chromatin organization of a given genomic region. In a disease context such as cancer, it would be much more powerful (and cost effective) to design a set of capture probes spanning a few TADs around known oncogenes and to explore these regions in cohorts of patients. However, the current targeted strategies suffer from a very low specificity, and a large fraction of reads are actually off-targets or non-informative. As an example, from our experience, around 90% of reads are discarded in a capture-C experiment, a large majority of them representing unligated fragments. This is due to the fact that it seems challenging to combine, in the same experiment, both a biotin pull-down to enrich the library in valid interactions, and a capture enrichment. So far, the same issue seems to occur in capture Hi-C experiments. For instance, Franke et al. (2016) recently generated 10 kb resolution contact maps of the *Sox9* TAD (5 Mb) using capture Hi-C. To reach this resolution, more than 200 millions of paired-end reads were gathered, but less than 30 millions were informative valid 3C products (15%) used to build the maps. Thus, there is a real interest in developing these protocols to increase their efficiency. There is no doubt that these tools will be extensively used in a near future to discover new molecular mechanisms involved in diseases.

As many other sequencing applications, Hi-C is a population-based assay. It means that the spatial organization extracted from a Hi-C experiment is an average interaction profile from a population of cells. This has been intensively discussed, as for instance, it raises the question whether the observed average TAD structure exists within individual

cell, or if it just represents an average view of contacts occuring within a population of cells. This would have extreme consequences on the interpretation of TAD functions. The first single-cell Hi-C approach was proposed a couple of years ago (Nagano et al. (2013)), reporting that the TADs organization seems to be conserved between individual cells, but that the exact nuclear position and the inter-TAD contacts are highly variable between cells. This type of approach will be extremely useful to characterize structural dynamics across time and space. Recently, Ramani et al. proposed a new protocol based on multiplex DNA barcoding to explore the chromatin folding of thousands of cells in a single experiment. Such approach would be extremely powerful in the future, especially if it can be adapted to other epigenomic assays including DamID, ChIP or ATAC-seq to provide a complete view of the chromatin landscape of a cell. However, it should also be noted that the current single-cell Hi-C results suffer from very low resolution, and that improvements in the data resolution will be necessary to make reliable conclusions on these questions.

5.3 Bioinformatic challenges

In the past four years, many new bioinformatic methods have been proposed to analyse Hi-C data. However, going from the sequence to functional conclusions is still challenging, and as the field is moving very fast, the computational methods should also be adjusted accordingly.

With a step backwards, one realizes that actually many methods have been proposed without explicit knowledge on what they are looking for. For example, chromosome compartments usually appear on a contact map as alternating blocks of high and low interaction frequency, but they are currently identified using a PCA analysis, that does not explicitly search for this pattern. The same is true for TAD calling, where methods based on statistical algorithms such as segmentation, were designed to detect blocks along the diagonal, without any a priori knowledge of what exactly is a TAD. In parallel that our understanding of the chromatin structure is growing, it is important that statistical or bioinformatic algorithms rely on explicit models of our biological knowledge, in order to build robust, valid and reproducible methods.

In the following section, we present a few bioinformatic challenges that will have to be addressed in the coming years.

5.3.1 Data processing and normalization

The data processing is the first step of any analysis. It aims at extracting relevant information from raw sequencing reads. While many people consider that this part of the analysis is straightforward, this is usually not the case. An unefficient or inappropriate data processing can have strong impact on the downstream analysis and their biological conclusions. As already presented in Table 1.3, many tools are already available, but some parts of the analysis remain challenging and can still be improved.

5.3.1.1 Comparison of existing solutions

A few years ago, many biological studies were published with their own computational methods, making difficult the comparison of the results. Since then, several stand-alone tools and pipelines have been proposed, providing the basis for reproducibility and for a better comparison of the different studies. The main popular solutions for Hi-C data analysis have already been reviewed several times (see section 1.4.3), but to my knowledge, an in-depth comparison of the results generated from these tools in still lacking. The main difficulty in such comparison is to define appropriate data set and biological criteria to compare and evaluate them.

Coming back to the original question, the goal of the data processing is, starting from a set of sequencing reads, to be able to extract informative valid 3C products. Under the assumption that everybody agrees on the definition of a valid product, all these tools should give the same results, which is obviously not the case. One idea would be to run a comparison with a simulated dataset of valid and spurious ligation products, even if building such simulated dataset in an unbiased manner might be difficult. Highlighting differences between softwares, such as pairs of reads reported in one case and not in another, could be very informative to better understand the behavior of each algorithm and/or to detect differences and therefore improve these tools. A more descriptive comparison could also be of interest. For instance, the quality controls returned by each tool are usually different. As an example, the HiCUP pipeline (Wingett et al. (2015)) classifies non-informative 3C products in 6 different classes ('wrong size', 'contiguous sequence', 're-ligation', 'same fragment internal', 'same fragment dangling ends', 'same fragment circularized'), whereas HiC-Pro currently reports the same information organized in only four classes ('dangling-ends', 'self-circle', 're-ligation', 'dumped pairs'). It

would therefore be interesting to evaluate these differences and to see with biologists how helpfull they can be for the end-users. Finally, other parameters such as flexibility, speed, or simplicity to use would also be interesting feedbacks for the community.

5.3.1.2 Processing and repeated elements

The detection of valid interaction products follows a workflow which is now well defined and validated by the community (see section 1.4.1). All currently available tools are based on uniquely mapped reads that define the interacting loci used to call chromosome compartments or TADs. So far, only a very small number of studies have investigated the role of repeated elements in the chromatin organization. However, we have a few evidences supporting the idea that repeated elements could play a role in chromatin conformation. First, SINE elements have been found enriched at the TADs boundaries suggesting a potential role in stabilizing these structures (Dixon et al. (2012)). More recently, Cournac et al. deeply explored the relationship between repeats and chromatin folding, and proposed that the similarity in chromatin structure observed between homologous regions of Human and Mouse genomes could be explained by the presence of SINE elements. A systematic exploration of contact probabilities from repeated elements is thus required to better understand their role, as well as how regions enriched in repeated elements such as centromeres and telomeres are regulated.
In a computational point of view, the easiest (and most reliable) way to study the role of repeated elements in the chromatin conformation is to use uniquely aligned reads and to look for any enrichment in contacts involving repeated elements. This should be possible under the assumption that reads are long enough to partially overlap with a repeat element, and map unambiguously thanks to its neighboring regions. In addition, other mapping strategies could also be tested to rescue reads aligned at multiple positions. The first strategy would be to report all possible reads alignments, but this is almost impossible to use in practice because of the underlying huge computational complexity. A more sophisticated version is to report all alignments up to a given limit, and to weight each position by the number of time a read aligns on the genome. And the last strategy is to randomly report one alignment among all possible. Note that in practice the two last strategies should give very close results, and that the random assignment of reads is easier to put in practice as many mappers now propose this option by default. It would therefore be interesting to evaluate this strategy to align

Hi-C data, and to see whether it allows a better characterization of repeated elements. It will also be important to assess how reporting multi-mapped reads affects the current normalization strategies and downstream analysis.

5.3.1.3 Allele-specific analysis of Hi-C data

HiC-Pro is currently the only pipeline allowing, in a single command line, to build allele-specific contact maps from Hi-C sequencing reads. As explained in chapter 2, it currently relies on a N-masking strategy, where known heterozygote SNPs are replaced by a 'N' in the reference genome therefore avoiding mapping bias toward one of the parental allele.

However, other allele-specific strategies could also be of interest such as the 'parental genomes' or the 'diploid mapping' strategies. The 'parental genomes' strategy has been mainly used for RNA-seq data analysis (Borensztein et al. (2017); Gendrel et al. (2014)). The idea is to independently align the reads on the two reconstructed parental genomes, which are extrapolated from a reference genome (i.e. hg19, mm9, etc.) where all heterozygous SNPs positions have been replaced by the nucleotide matching the genotype of each parent. After mapping, the two alignment files are merged in order to identify the allele-specific reads that better align on one of the parental genome. The 'diploid mapping' strategy has been recently used by Giorgetti et al.. The idea is to directly align the sequencing reads on a 2N reference genome. Therefore, uniquely mapped reads correspond to sequences matching only one of the parental alleles. The main advantage of this method is that it offers a more 'natural' way of mapping the reads in an allele-specific manner as the assignment to one of two parental genomes is achieved by the mapper itself. However, this method can only be used with recent mappers able to build 2N chromosomes indexes and to load them into memory.

Finally, all these methods represent different ways of aligning reads in a allele-specific way. While they all have already been used in different biological contexts, a detailed comparison of their performance is still lacking. For instance, is there one method that significantly outperforms the others ? Is there any context (reads length, number of SNPs, sequencing application, etc.) that favours the use of one method ? All these questions could be addressed using a simulated dataset to control the number and the ratio of allele-specific reads. Of note, such framework could also be extended to the comparison of statistical methods developed to call allele-specific gene expression (see

Castel et al. (2015) for a review). As allele-specific analysis requires a high sequencing depth (and is costly), there is a real interest in using the best computational approach to classify the maximum number of allele-specific reads in an accurate manner, with the least possible bias.

5.3.1.4 Data normalization

As illustrated in this work, the Hi-C data normalization is still an important challenge of the data processing. Two main types of Hi-C data normalization have been proposed so far (see section 1.4.1.4). The first one relies on explicit statistical methods that model all known biases and remove them from the raw contact counts matrices. The other is based on matrix balancing algorithms and is expected to correct the data for all biases (known or unknown) under the assumption that the sum of the genome-wide contact for each locus is the same after data normalization (also called 'equal visibility' assumption).

Matrix balancing algorithms, as the ICE normalization (Imakaev et al. (2012)), are currently very popular. However, they also rely on strong assumptions that are important to keep in mind. For instance, the use of such technics to normalize Hi-C data enriched for specific targets, has to be carrefully assessed. HiChIP was recently proposed to enrich 3C products associated with specific DNA-binding proteins using chromatin immunoprecipitation (Mumbach et al. (2016)). While the protocol and data processing is close to standard in-situ Hi-C, the contact maps should be enriched in loci bound by the protein of interest. In this context, it is difficult to validate the assumption of equal visibility made by the matrix balancing algorithms. Alternative methods would need to be tested or developed to specifically normalize this type of data and remove systematic biases. The same question arises in the context of capture Hi-C data. Indeed, the assumption of equal visibility holds genome-wide. But making this assumption at the scale of a few mega-bases is questionable, especially because of TADs and sub-TADs structures. Other methods based on explicit models might be more appropriate in this case and have to be tested.

5.3.2 Single-cell Hi-C data

If until now, we mainly studied transcription and (epi)genomic variation at the level of cell population, recent single-cell sequencing technologies have emerged allowing to

explore differences between cells within a population. It is extremely likely that soon, the majority of sequencing applications will be available at the single-cell level. Since a couple of years, single-cell RNA sequencing has been successfully used in many context. In the meantime, we have now clear evidence that single-cell RNA-seq data analysis requires dedicated bioinformatic and statistical methods. It is likely that the same was also true for single-cell Hi-C data analysis. And as the technique will become more popular, there will be an increased need for methods to analyze this type of data.
At the data processing level, the workflow should not be too much different from standard Hi-C, and it is likely that current tools could be easily updated to handle these data. However, the challenge is more at the level of normalization and data interpretation. As an example, previous results of single-cell Hi-C data strongly rely on computational methods to interpret, extrapolate and model the observed signal (Nagano et al. (2013)). New methods will also need to be set up to identify group of cells sharing the same chromatin organization among all single-cells available. And as at the beginning of Hi-C data, standard methods will need to be defined and packaged to ensure reproducibility between studies.

5.3.3 Statistical comparison of contact maps

Until now, most of the published normalization methods address the question of within-sample normalization to remove the systematic biases but do not address the question of between-samples normalization which is necessary before samples comparison. The most obvious between-samples normalization method is based on the calculation of scaling factor to adjust for differences in sequencing depth. However, as it has already been demonstrated for RNA-seq data analysis, it is likely that more sophisticated methods, for instance based on the definition of an invariant set of features, give better results (Dillies et al. (2013)).
Then, a question that is currently of high interest is how to efficiently compare Hi-C contact maps ? How to find contacts that are specific to one condition and not to another ? So far, an efficient method to run differential analysis on contact maps is still lacking. One idea would be to extend methods dedicated to RNA-seq analysis, and to apply them on Hi-C contact maps, considering each pair of loci as a feature (Lun and Smyth (2015)). However, there is a strong fondamental difference between RNA-seq and Hi-C data. Indeed, contrary to RNA-seq data analysis which makes the

assumption of independence between features (which might be discussed), Hi-C data are strongly (spatially) dependent. Therefore, if two loci differentially interact between two conditions, it is likely that the neighboring loci also differentially interact. An ideal method for Hi-C data analysis would thus be able to detect regions (and not loci) that differentially interact, regardless the data resolution. In addition, this context of dependence also raises questions about multiple testing adjustment. Indeed, most of classical methods used to adjust for multiple testing, as the well-known Benjamini-Hochberg procedure (Benjamini (1995)), are based on the assumption of independence between features. Therefore, their application is not valid in the Hi-C context. In addition, while replicates are now common in RNA-seq experiments (and usually required for differential analysis), biological replicates of Hi-C experiments remain rare. One idea would be, for instance, to use local neighboring regions instead of replicates to calculate a variance. Finally, methods derived from the field of images comparison could be a good starting point to design a method dedicated to Hi-C data comparison.

5.3.4 SVs detection and genome assembly

In addition to their promises to detect enhancer hijacking in cancer, Hi-C has also recently been described as a method of choice to study genome rearrangements. As already discussed, structural variants have been shown to disrupt the canonical 3D structure of the genome. In this context, Harewood et al. recently demonstrated that Hi-C can be used to call copy number variants, but also more complex rearrangements such as balanced or unbalanced translocations, inversions, or amplifications. However, until now, detecting such events mainly relies on visual inspection of the contact maps. Consequently, there is a real interest in developing new computational methods to automatically detect these events. One idea would be first to normalize the contact maps by the counts $\sim$ distances relationship in order to bring out non-expected patterns. Indeed, it is likely that contact frequencies around chromosomal rearrangements do not match the expected distances from the reference genome. Then, applying a statistical model to call significant contacts as already proposed in the FitHiC (Ay et al. (2014)) or GOTHIC (Mifsud et al. (2017)) packages could help in detecting abnormal contacts and in merging nearby genomic bins that likely represent the same rearrangement event.

Moreover, an accurate detection of these structural events could also be useful to build

de-novo assembly of cancer genomes. There have been several promising studies that used the Hi-C distance signal for de-novo genome assembly (see Korbel and Lee (2013) for a review). These computational approaches have been used either to reorder DNA scaffolds build with standard genomic approaches, or to phase SNPs onto haplotypes at chromosome-scale.

Therefore, Hi-C seems to be a cost-effective approach for de-novo assembly, as opposed to more expensive approaches such as long reads sequencing. Compare to standard paired-end sequencing, Hi-C is more powerful is the sense that it does need to have a DNA fragment spanning the breakpoints to detect a rearrangement. In the future, Hi-C may enable the assembly of human cancer genomes, enhancing the detection of structural variants and their impact on tumorigenesis. Moreover, having a precise picture of the genomic rearrangements of a tumour could also pave the way to new applications in physics to better understand how these events occur in a cell, as well as new developments to infer the 3D structure of cancer genomes.

5.3.5 Data interpretation and integration

The 'Holy Grail' of many current projects in bioinformatics is the integrative analysis of heterogeneous profiles. While it seems to be one of the main current challenges in the field, understanding what it exactly means is not straightforward. Indeed, the naive approach to integrate multi-omics data would simply be to look at superimposed genomic tracks of different types, or to simply compare lists of genes (or any other features) detected from different datasets. In contrast, integrating heterogenous data can also rely on more complex mathematical methods designed to model the interactions among variables of different types (see Bersanelli et al. (2016) for a review).

Regarding the analysis of Hi-C data, it is clear that their comparison (or 'integration') with other data-set such as gene expression, histone modifications, methylation, replication timing, etc. will be an important challenge of the coming years. So far, it is difficult to say if dedicated new methods and algorithms have to be developed, in particular because the way of integrating the data highly depends on the biological questions to address. In addition, in most of the cases, simple descriptive statistics, as well as appropriate statistical methods based for instance, on classification or dimension reduction algorithms can be sufficient. Most often, the integration of heterogeneous data is based on the choice of a common feature, like for instance, genes. For example, a simple way

to integrate ChIP-seq and RNA-seq data is to compare the list of misregulated genes detected by RNA-seq with the list of genes/promoters for which a ChIP-seq peak has been found. This type of information can be 'integrated' with Hi-C data by looking whether misregulated genes belong to different chromosome compartments, or belong to the same TADs.

However, in a statiscal point of view, it is also true that being able to analyse multiple datasets in a single way could be more powerfull than running independent analysis. It could for instance be of interest to propose methods that predict gene expression based on histone modification or interaction profiles. It could also be interesting to compare groups of samples by combining information about gene expression, transcription factor binding sites and chromatin structure (see Angelini and Costa (2014) for a review).

To conclude, although it is clear that 'integrative analysis' can be a stimulating environment to develop new statistical methods, many biologically-driven and simple comparisons can already be achieved in a naive way, as this is already the case in most of the biological studies. Nevertheless, it is also true that good practices about statistical considerations or general strategies could be useful for the community.

6

Concluding Remarks

The field of chromosome conformation is full of promises and we are only just beginning to understand the functional implications of chromatin folding. The project that we presented here grew at the same time that knowledge in the field accumulated.
Our first aim was to build computational frameworks able to generalize the analysis of Hi-C data, by making it accessible to as many people as possible. We therefore started by the beginning, developing an efficient software to process any type of Hi-C-based data, independently from their biological protocol. HiC-Pro is freely available and is now widely used by the community. Since its publication, eleven versions were released, always improving the pipeline by adding new functionalities, most often proposed by the users themselves. As already discussed, we have attached the greatest importance to make HiC-Pro easy to use, documented and compatible with other software, thus encouraging other bioinformaticians in the field to contribute to the project.
We next focused our developments on the analysis of Hi-C data from cancer samples. Rapidly, we identified a few limitations mainly related to the normalization of these data. To better understand this issue, we designed a simulation model allowing to reproduce the effects that we observed on real data. Importantly, we demonstrated in details, that the current normalization methods cannot be applied in this context, therefore highlighting the importance of developing appropriate methods for cancer Hi-C analysis. Our simulation model allowed us to design new strategies that efficiently deal with the copy number effects to normalize Hi-C cancer data. Interestingly, we also demonstrated that the Hi-C is an efficient method to detect copy number variants, and observed a very strong correlation with copy number extracted from SNPs arrays.

In addition, we started to explore the chromatin organization of cancer samples using a Tet-Myc hybrid mouse model, and uveal melanoma cell lines. While these projects are still ongoing, we hope they will allow us to better understand the link between gene expression, histone modifications, and chromatin conformation in cancer.

To conclude, our work contributes to develop accurate computational methods to explore Hi-C data from normal and cancer samples. Many other challenges remain to be solved in the coming years. The application of Hi-C to cancer should provide significant insights into the role of the chromatin organization in tumorigenesis, with the hope of developing new clinical approaches for patients. To reach this goal, appropriate computational workflows will be necessary, some of them requiring new developments. At short term, we will continue to maintain and improve the tools we have developed, and to apply them to explore the impact of chromatin folding in cancer. I hope that our environment at the Institut Curie will provide plenty of opportunities for collaboration to study these questions, and to develop novel computational approaches.

The end.

References

Kadir Caner Akdemir and Lynda Chin. Hicplotter integrates genomic data with interaction matrices. *Genome biology*, 16:198, September 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0767-1. 98

Claudia Angelini and Valerio Costa. Understanding gene regulatory mechanisms by integrating chip-seq and rna-seq data: statistical solutions to biological problems. *Frontiers in cell and developmental biology*, 2:51, 2014. ISSN 2296-634X. doi: 10.3389/fcell.2014.00051. 200

A Annunziato. Dna packaging: Nucleosomes and chromatin. *Nature Education 1(1):26*, 2008. 5

Haitham Ashoor, Aurlie Hrault, Aurlie Kamoun, Franois Radvanyi, Vladimir B Bajic, Emmanuel Barillot, and Valentina Boeva. Hmcan: a method for detecting chromatin modifications in cancer samples using chip-seq data. *Bioinformatics (Oxford, England)*, 29:2979–2986, December 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt524. 104

Sandrine Augui, Elphge P Nora, and Edith Heard. Regulation of x-chromosome inactivation by the x-inactivation centre. *Nature reviews. Genetics*, 12:429–442, June 2011. ISSN 1471-0064. doi: 10.1038/nrg2987. 29

Ferhat Ay and William S Noble. Analysis methods for studying the 3d architecture of the genome. *Genome biology*, 16:183, September 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0745-7. 58, 61, 67, 101

Ferhat Ay, Timothy L Bailey, and William Stafford Noble. Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, 24:999–1011, June 2014. ISSN 1549-5469. doi: 10.1101/gr.160374.113. 63, 98, 198

Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837, May 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.05.009. 10

A Rasim Barutcu, Bryan R Lajoie, Rachel P McCord, Coralee E Tye, Deli Hong, Terri L Messier, Gillian Browne, Andre J van Wijnen, Jane B Lian, Janet L Stein, Job Dekker, Anthony N Imbalzano, and Gary S Stein. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome biology*, 16:214, September 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0768-0. 21, 40, 57, 187

A Rasim Barutcu, Andrew J Fritz, Sayyed K Zaidi, Andr J van Wijnen, Jane B Lian, Janet L Stein, Jeffrey A Nickerson, Anthony N Imbalzano, and Gary S Stein. C-ing the genome: A compendium of chromosome conformation capture methods to study higher-order chromatin organization. *Journal of cellular physiology*, 231:31–35, January 2016. ISSN 1097-4652. doi: 10.1002/jcp.25062. 43

Stephen B Baylin and Peter A Jones. A decade of exploring the cancer epigenome - biological and translational implications. *Nature reviews. Cancer*, 11:726–734, September 2011. ISSN 1474-1768. doi: 10.1038/nrc3130. 12

Yosef Benjamini, Yoav; Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1): 289300, 1995. 198

Joel B Berletch, Fan Yang, and Christine M Disteche. Escape from x inactivation in mice and humans. *Genome biology*, 11:213, 2010. ISSN 1474-760X. doi: 10.1186/gb-2010-11-6-213. 31

Joel B Berletch, Fan Yang, Jun Xu, Laura Carrel, and Christine M Disteche. Genes that escape from x inactivation. *Human genetics*, 130:237–245, August 2011. ISSN 1432-1203. doi: 10.1007/s00439-011-1011-z. 31

Rameen Beroukhim, Xiaoyang Zhang, and Matthew Meyerson. Copy number alterations unmasked as enhancer hijackers. *Nature genetics*, 49:5–6, December 2016. ISSN 1546-1718. doi: 10.1038/ng.3754. 38

Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17 Suppl 2:15, January 2016. ISSN 1471-2105. doi: 10.1186/s12859-015-0857-9. 199

Saul A Bert, Mark D Robinson, Dario Strbenac, Aaron L Statham, Jenny Z Song, Toby Hulf, Robert L Sutherland, Marcel W Coolen, Clare Stirzaker, and Susan J Clark. Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer cell*, 23:9–22, January 2013. ISSN 1878-3686. doi: 10.1016/j.ccr.2012.11.006. 15

Subhankar Biswas and C Mallikarjuna Rao. Epigenetics in cancer: Fundamentals and beyond. *Pharmacology & therapeutics*, February 2017. ISSN 1879-016X. doi: 10.1016/j.pharmthera.2017.02.011. 8

Valentina Boeva, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre, and Emmanuel Barillot. Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28:423–425, February 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr670. 104

Boyan Bonev and Giacomo Cavalli. Organization and function of the 3d genome. *Nature reviews. Genetics*, 17:661–678, October 2016. ISSN 1471-0064. doi: 10.1038/nrg.2016.112. 17, 18, 56

Maud Borensztein, Laurne Syx, Katia Ancelin, Patricia Diabangouaya, Christel Picard, Tao Liu, Jun-Bin Liang, Ivaylo Vassilev, Rafael Galupa, Nicolas Servant, Emmanuel Barillot, Azim Surani, Chong-Jian Chen, and

REFERENCES

Edith Heard. Xist-dependent imprinted x inactivation and the early developmental consequences of its failure. *Nature structural & molecular biology*, 24:226–233, March 2017. ISSN 1545-9985. doi: 10.1038/nsmb.3365. 195

Britta A M Bouwman and Wouter de Laat. Getting the genome in shape: the formation of loops, domains and compartments. *Genome biology*, 16:154, August 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0730-1. 26

S Boyle, S Gilchrist, J M Bridger, N L Mahy, J A Ellis, and W A Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*, 10:211–219, February 2001. ISSN 0964-6906. 19

Miguel R Branco and Ana Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS biology*, 4:e138, May 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040138. 19, 20

Richard D Carvajal, Gary K Schwartz, Tongalp Tezel, Brian Marr, Jasmine H Francis, and Paul D Nathan. Metastatic disease from uveal melanoma: treatment options and future prospects. *The British journal of ophthalmology*, 101: 38–44, January 2017. ISSN 1468-2079. doi: 10.1136/bjophthalmol-2016-309034. 145

Stephane E Castel, Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. Tools and best practices for data processing in allelic expression analysis. *Genome biology*, 16:195, September 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0762-6. 196

Jaime Abraham Castro-Mondragon, Sbastien Jaeger, Denis Thieffry, Morgane Thomas-Chollier, and Jacques van Helden. Rsat matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic acids research*, 45:e119, July 2017. ISSN 1362-4962. doi: 10.1093/nar/gkx314. 176

Ronan Chaligné, Tatiana Popova, Marco-Antonio Mendoza-Parra, Mohamed-Ashick M Saleem, David Gentien, Kristen Ban, Tristan Piolot, Olivier Leroy, Odette Mariani, Hinrich Gronemeyer, Anne Vincent-Salomon, Marc-Henri Stern, and Edith Heard. The inactive x chromosome is epigenetically unstable and transcriptionally labile in breast cancer. *Genome research*, 25:488–503, April 2015. ISSN 1549-5469. doi: 10.1101/gr.185926.114. 160, 172

Andrew Chase and Nicholas C P Cross. Aberrations of ezh2 in cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 17:2613–2618, May 2011. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-10-2156. 14

Chun-Kan Chen, Mario Blanco, Constanza Jackson, Erik Aznauryan, Noah Ollikainen, Christine Surka, Amy Chow, Andrea Cerase, Patrick McDonel, and Mitchell Guttman. Xist recruits the x chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science (New York, N.Y.)*, 354:468–472, October 2016. ISSN 1095-9203. doi: 10.1126/science.aae0047. 19

Zhao-xia Chen and Arthur D Riggs. Dna methylation and demethylation in mammals. *The Journal of biological chemistry*, 286:18347–18353, May 2011. ISSN 1083-351X. doi: 10.1074/jbc.R110.205286. 6

Susan J Clark. Action at a distance: epigenetic silencing of large chromosomal regions in carcinogenesis. *Human molecular genetics*, 16 Spec No 1:R88–R95, April 2007. ISSN 0964-6906. doi: 10.1093/hmg/ddm051. 15

M Ryan Corces and Victor G Corces. The three-dimensional cancer genome. *Current opinion in genetics & development*, 36:1–7, February 2016. ISSN 1879-0380. doi: 10.1016/j.gde.2016.01.002. 38, 39

Axel Cournac, Herv Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC genomics*, 13:436, August 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-436. 59

Axel Cournac, Romain Koszul, and Julien Mozziconacci. The 3d folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic acids research*, 44:245–255, January 2016. ISSN 1362-4962. doi: 10.1093/nar/gkv1292. 194

Emily Crane, Qian Bian, Rachel Patton McCord, Bryan R Lajoie, Bayly S Wheeler, Edward J Ralston, Satoru Uzawa, Job Dekker, and Barbara J Meyer. Condensin-driven remodelling of x chromosome topology during dosage compensation. *Nature*, 523:240–244, July 2015. ISSN 1476-4687. doi: 10.1038/nature14450. 65, 66, 169, 176

T Cremer and C Cremer. Rise, fall and resurrection of chromosome territories: a historical perspective. part ii. fall and resurrection of chromosome territories during the 1950s to 1980s. part iii. chromosome territories and the functional nuclear architecture: experiments and models from the 1990s to the present. *European journal of histochemistry : EJH*, 50:223–272, 2006a. ISSN 1121-760X. 16, 17

Thomas Cremer and C Cremer. Rise, fall and resurrection of chromosome territories: a historical perspective. part i. the rise of chromosome territories. *European journal of histochemistry : EJH*, 50:161–176, 2006b. ISSN 1121-760X. 16

Thomas Cremer and Marion Cremer. Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2:a003889, March 2010. ISSN 1943-0264. doi: 10.1101/cshperspect.a003889. 19

J A Croft, J M Bridger, S Boyle, P Perry, P Teague, and W A Bickmore. Differences in the localization and morphology of chromosomes in the human nucleus. *The Journal of cell biology*, 145:1119–1131, June 1999. ISSN 0021-9525. 19

Wu Ct and J R Morris. Genes, genetics, and epigenetics: a correspondence. *Science (New York, N.Y.)*, 293:1103–1105, August 2001. ISSN 0036-8075. doi: 10.1126/science.293.5532.1103. 3

Rola Dali and Mathieu Blanchette. A critical assessment of topologically associating domain prediction tools. *Nucleic acids research*, 45:2994–3005, April 2017. ISSN 1362-4962. doi: 10.1093/nar/gkx145. 67

C V Dang, L M Resar, E Emison, S Kim, Q Li, J E Prescott, D Wonsey, and K Zeller. Function of the c-myc oncogenic transcription factor. *Experimental cell research*, 253:63–77, November 1999. ISSN 0014-4827. doi: 10.1006/excr.1999.4686. 159

Emily M Darrow, Miriam H Huntley, Olga Dudchenko, Elena K Stamenova, Neva C Durand, Zhuo Sun, Su-Chen Huang, Adrian L Sanborn, Ido Machol, Muhammad Shamim, Andrew P Seberg, Eric S Lander, Brian P Chadwick, and Erez Lieberman Aiden. Deletion of dxz4 on the human inactive x chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences of the United States of America*, 113:E4504–E4512, August 2016. ISSN 1091-6490. doi: 10.1073/pnas.1609643113. 32

James O J Davies, Jelena M Telenius, Simon J McGowan, Nigel A Roberts, Stephen Taylor, Douglas R Higgs, and Jim R Hughes. Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature methods*, 13:74–80, January 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3664. 48, 50, 96, 97

James O J Davies, A Marieke Oudelaar, Douglas R Higgs, and Jim R Hughes. How best to identify chromosomal interactions: a comparison of approaches. *Nature methods*, 14:125–134, January 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4146. 80

Elzo de Wit and Wouter de Laat. A decade of 3c technologies: insights into nuclear organization. *Genes & development*, 26:11–24, January 2012. ISSN 1549-5477. doi: 10.1101/gad.179804.111. 43

Job Dekker. The three 'c' s of chromosome conformation capture: controls, controls, controls. *Nature methods*, 3:17–21, January 2006. ISSN 1548-7091. doi: 10.1038/nmeth823. 43

Job Dekker. Mapping the 3d genome: Aiming for consilience. *Nature reviews. Molecular cell biology*, 17:741–742, November 2016. ISSN 1471-0080. doi: 10.1038/nrm.2016.151. 42

Job Dekker and Edith Heard. Structural and functional diversity of topologically associating domains. *FEBS letters*, 589:2877–2884, Oct 2015. ISSN 1873-3468. doi: 10.1016/j.febslet.2015.08.044. 21, 22

Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science (New York, N.Y.)*, 295:1306–1311, February 2002. ISSN 1095-9203. doi: 10.1126/science.1067799. 42, 43, 49

Xinxian Deng, Wenxiu Ma, Vijay Ramani, Andrew Hill, Fan Yang, Ferhat Ay, Joel B Berletch, Carl Anthony Blau, Jay Shendure, Zhijun Duan, William S Noble, and Christine M Disteche. Bipartite structure of the inactive mouse x chromosome. *Genome biology*, 16:152, August 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0728-8. 32, 49, 170

Annette Denker and Wouter de Laat. The second decade of 3c technologies: detailed insights into nuclear organization. *Genes & development*, 30:1357–1382, June 2016. ISSN 1549-5477. doi: 10.1101/gad.281964.116. 43

Vishnu Dileep, Ferhat Ay, Jiao Sima, Daniel L Vera, William S Noble, and David M Gilbert. Topologically associating domains and their long-range contacts are established during early g1 coincident with the establishment of the replication-timing program. *Genome research*, 25:1104–1113, August 2015. ISSN 1549-5469. doi: 10.1101/gr.183699.114. 24

Marie-Agns Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Cline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Lalo, Caroline Le Gall, Brigitte Schaffer, Stphane Le Crom, Mickal Guedj, Florence Jaffrzic, and French StatOmique Consortium. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14:671–683, November 2013. ISSN 1477-4054. doi: 10.1093/bib/bbs046. 197

Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485:376–380, April 2012. ISSN 1476-4687. doi: 10.1038/nature11082. 21, 22, 24, 25, 65, 66, 101, 186, 194

Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V Lobanenkov, Joseph R Ecker, James A Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518:331–336, February 2015. ISSN 1476-4687. doi: 10.1038/nature14222. xxi, 21, 25, 33, 34, 183, 184, 187

Jose Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D Green, and Job Dekker. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16:1299–1309, October 2006. ISSN 1088-9051. doi: 10.1101/gr.5571506. 45, 49

Neva C Durand, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell systems*, 3:99–101, July 2016a. ISSN 2405-4712. doi: 10.1016/j.cels.2015.07.012. 98, 99

Neva C Durand, Muhammad S Shamim, Ido Machol, Suhas S P Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell systems*, 3:95–98, July 2016b. ISSN 2405-4712. doi: 10.1016/j.cels.2016.07.002. 61, 100

Olivier Elemento, Mark A Rubin, and David S Rickman. Oncogenic transcription factors as master regulators of chromatin topology: a new role for erg in prostate cancer. *Cell cycle (Georgetown, Tex.)*, 11:3380–3383, September 2012. ISSN 1551-4005. doi: 10.4161/cc.21401. 188

Jesse M Engreitz, Amy Pandya-Jones, Patrick McDonel, Alexander Shishkin, Klara Sirokman, Christine Surka, Sabah Kadri, Jeffrey Xing, Alon Goren, Eric S Lander, Kathrin Plath, and Mitchell Guttman. The xist lncrna exploits three-dimensional genome architecture to spread across the x chromosome. *Science (New York, N.Y.)*, 341:1237973, August 2013. ISSN 1095-9203. doi: 10.1126/science.1237973. 29

Manel Esteller. Epigenetics in cancer. *The New England journal of medicine*, 358:1148–1159, March 2008. ISSN 1533-4406. doi: 10.1056/NEJMra072067. 12

REFERENCES

Maria E Figueroa, Omar Abdel-Wahab, Chao Lu, Patrick S Ward, Jay Patel, Alan Shih, Yushan Li, Neha Bhagwat, Aparna Vasanthakumar, Hugo F Fernandez, Martin S Tallman, Zhuoxin Sun, Kristy Wolniak, Justine K Peeters, Wei Liu, Sung E Choe, Valeria R Fantin, Elisabeth Paietta, Bob Lwenberg, Jonathan D Licht, Lucy A Godley, Ruud Delwel, Peter J M Valk, Craig B Thompson, Ross L Levine, and Ari Melnick. Leukemic idh1 and idh2 mutations result in a hypermethylation phenotype, disrupt tet2 function, and impair hematopoietic differentiation. *Cancer cell*, 18:553–567, December 2010. ISSN 1878-3686. doi: 10.1016/j.ccr.2010.11.015. 13

Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Identification of alternative topological domains in chromatin. *Algorithms for molecular biology : AMB*, 9:14, 2014. doi: 10.1186/1748-7188-9-14. 21

Galina N Filippova, Mimi K Cheng, James M Moore, Jean-Pierre Truong, Ying J Hu, Di Kim Nguyen, Karen D Tsuchiya, and Christine M Disteche. Boundaries between chromosomal domains of x inactivation and escape bind ctcf and lack cpg methylation during early development. *Developmental cell*, 8:31–42, January 2005. ISSN 1534-5807. 31

Lee E Finlan, Duncan Sproul, Inga Thomson, Shelagh Boyle, Elizabeth Kerr, Paul Perry, Bauke Ylstra, Jonathan R Chubb, and Wendy A Bickmore. Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS genetics*, 4:e1000039, March 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000039. 20

William A Flavahan, Yotam Drier, Brian B Liau, Shawn M Gillespie, Andrew S Venteicher, Anat O Stemmer-Rachamimov, Mario L Suv, and Bradley E Bernstein. Insulator dysfunction and oncogene activation in idh mutant gliomas. *Nature*, 529:110–114, January 2016. ISSN 1476-4687. doi: 10.1038/nature16490. 39

William A Flavahan, Elizabeth Gaskell, and Bradley E Bernstein. Epigenetic plasticity and the hallmarks of cancer. *Science (New York, N.Y.)*, 357, July 2017. ISSN 1095-9203. doi: 10.1126/science.aal2380. 188

Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. Comparison of computational methods for hi-c data analysis. *Nature methods*, June 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4325. 65, 66

Jean-Philippe Fortin and Kasper D Hansen. Reconstructing a/b compartments as revealed by hi-c using long-range correlations in epigenetic data. *Genome biology*, 16:180, August 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0741-y. 20

Mario F Fraga, Esteban Ballestar, Ana Villar-Garea, Manuel Boix-Chornet, Jesus Espada, Gunnar Schotta, Tiziana Bonaldi, Claire Haydon, Santiago Ropero, Kevin Petrie, N Gopalakrishna Iyer, Alberto Prez-Rosado, Enrique Calvo, Juan A Lopez, Amparo Cano, Maria J Calasanz, Dolors Colomer, Miguel Angel Piris, Natalie Ahn, Axel Imhof, Carlos Caldas, Thomas Jenuwein, and Manel Esteller. Loss of acetylation at lys16 and trimethylation at lys20 of histone h4 is a common hallmark of human cancer. *Nature genetics*, 37:391–400, April 2005. ISSN 1061-4036. doi: 10.1038/ng1531. 14

Martin Franke, Daniel M Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schpflin, Katerina Kraft, Rieke Kempfer, Ivana Jerkovi, Wing-Lee Chan, Malte Spielmann, Bernd Timmermann, Lars Wittler, Ingo Kurth, Paola Cambiaso, Orsetta Zuffardi, Gunnar Houge, Lindsay Lambie, Francesco Brancati, Ana Pombo, Martin Vingron, Francois Spitz, and Stefan Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538:265–269, October 2016. ISSN 1476-4687. doi: 10.1038/nature19800. 38, 50, 97, 191

James Fraser, Carmelo Ferrai, Andrea M Chiariello, Markus Schueler, Tiago Rito, Giovanni Laudanno, Mariano Barbieri, Benjamin L Moore, Dorothee C A Kraemer, Stuart Aitken, Sheila Q Xie, Kelly J Morris, Masayoshi Itoh, Hideya Kawaji, Ines Jaeger, Yoshihide Hayashizaki, Piero Carninci, Alistair R R Forrest, FANTOM Consortium, Colin A Semple, Jose Dostie, Ana Pombo, and Mario Nicodemi. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular systems biology*, 11:852, December 2015. ISSN 1744-4292. doi: 10.15252/msb.20156492. 21

Peter Fraser and Wendy Bickmore. Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447:413–417, May 2007. ISSN 1476-4687. doi: 10.1038/nature05916. 20

Geoffrey Fudenberg and Leonid A Mirny. Higher-order chromatin structure: bridging physics and biology. *Current opinion in genetics & development*, 22:115–124, April 2012. ISSN 1879-0380. doi: 10.1016/j.gde.2012.01.006. 28, 63

Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A Mirny. Formation of chromosomal domains by loop extrusion. *Cell reports*, 15:2038–2049, May 2016. ISSN 2211-1247. doi: 10.1016/j.celrep.2016.04.085. 25, 185

Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G Y Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N Ariyaratne, Vinsensius B Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K D Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V Desai, Jane S Thomsen, Yew Kok Lee, R Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462:58–64, November 2009. ISSN 1476-4687. doi: 10.1038/nature08497. 46, 49

Meital Gabay, Yulin Li, and Dean W Felsher. Myc activation is a hallmark of cancer initiation and maintenance. *Cold Spring Harbor perspectives in medicine*, 4, June 2014. ISSN 2157-1422. doi: 10.1101/cshperspect.a014241. 159

Anne-Valerie Gendrel, Mikael Attia, Chong-Jian Chen, Patricia Diabangouaya, Nicolas Servant, Emmanuel Barillot, and Edith Heard. Developmental dynamics and disease potential of random monoallelic gene expression. *Developmental cell*, 28:366–380, February 2014. ISSN 1878-1551. doi: 10.1016/j.devcel.2014.01.016. 195

Johan H Gibcus and Job Dekker. The hierarchy of the 3d genome. *Molecular cell*, 49:773–782, Mar 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2013.02.011. 17

Luca Giorgetti and Edith Heard. Closing the loop: 3c versus dna fish. *Genome biology*, 17:215, October 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1081-2. 42

Luca Giorgetti, Rafael Galupa, Elphge P Nora, Tristan Piolot, France Lam, Job Dekker, Guido Tiana, and Edith Heard. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157:950–963, May 2014. ISSN 1097-4172. doi: 10.1016/j.cell.2014.03.025. 28

Luca Giorgetti, Bryan R Lajoie, Ava C Carter, Mikael Attia, Ye Zhan, Jin Xu, Chong Jian Chen, Noam Kaplan, Howard Y Chang, Edith Heard, and Job Dekker. Structural organization of the inactive x chromosome in the mouse. *Nature*, 535:575–579, July 2016. ISSN 1476-4687. doi: 10.1038/nature18589. 30, 32, 35, 167, 170, 176, 195

Sandra Goetze, Julio Mateos-Langerak, Hinco J Gierman, Wim de Leeuw, Osdilly Giromus, Mireille H G Indemans, Jan Koster, Vladan Ondrej, Rogier Versteeg, and Roel van Driel. The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. *Molecular and cellular biology*, 27:4475–4487, June 2007. ISSN 0270-7306. doi: 10.1128/MCB.00208-07. 28

M Grunstein. Histone acetylation in chromatin structure and transcription. *Nature*, 389:349–352, September 1997. ISSN 0028-0836. doi: 10.1038/38664. 10

J William Harbour. The genetics of uveal melanoma: an emerging framework for targeted therapy. *Pigment cell & melanoma research*, 25:171–181, March 2012. ISSN 1755-148X. doi: 10.1111/j.1755-148X.2012.00979.x. 145

J William Harbour, Michael D Onken, Elisha D O Roberson, Shenghui Duan, Li Cao, Lori A Worley, M Laurin Council, Katie A Matatall, Cynthia Helms, and Anne M Bowcock. Frequent mutation of bap1 in metastasizing uveal melanomas. *Science (New York, N.Y.)*, 330:1410–1413, December 2010. ISSN 1095-9203. doi: 10.1126/science.1194472. 146

Louise Harewood, Kamal Kishore, Matthew D Eldridge, Steven Wingett, Danita Pearson, Stefan Schoenfelder, V Peter Collins, and Peter Fraser. Hi-c as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome biology*, 18:125, June 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1253-8. 105, 154, 189, 198

R David Hawkins, Gary C Hon, Leonard K Lee, Queminh Ngo, Ryan Lister, Mattia Pelizzola, Lee E Edsall, Samantha Kuan, Ying Luu, Sarit Klugman, Jessica Antosiewicz-Bourget, Zhen Ye, Celso Espinoza, Saurabh Agarwahl, Li Shen, Victor Ruotti, Wei Wang, Ron Stewart, James A Thomson, Joseph R Ecker, and Bing Ren. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell*, 6:479–491, May 2010. ISSN 1875-9777. doi: 10.1016/j.stem.2010.03.018. 15

Edith Heard and Christine M Disteche. Dosage compensation in mammals: fine-tuning the expression of the x chromosome. *Genes & development*, 20:1848–1867, July 2006. ISSN 0890-9369. doi: 10.1101/gad.1422906. 31

Sarah Heerboth, Karolina Lapinska, Nicole Snyder, Meghan Leary, Sarah Rollinson, and Sibaji Sarkar. Use of epigenetic drugs in disease: an overview. *Genetics & epigenetics*, 6:9–19, 2014. doi: 10.4137/GEG.S12270. 10

Emil Heitz. as heterochromatin der moose. *Jb Wiss Bot*, 69: 762–818, 1928. 4

J G Herman, A Merlo, L Mao, R G Lapidus, J P Issa, N E Davidson, D Sidransky, and S B Baylin. Inactivation of the cdkn2/p16/mts1 gene is frequently associated with aberrant dna methylation in all common human cancers. *Cancer research*, 55:4525–4530, October 1995. ISSN 0008-5472. 12

Victoria K Hill, Jung-Sik Kim, and Todd Waldman. Cohesin mutations in human cancer. *Biochimica et biophysica acta*, 1866:1–11, August 2016. ISSN 0006-3002. doi: 10.1016/j.bbcan.2016.05.002. 39, 40

Denes Hnisz, Abraham S Weintraub, Daniel S Day, Anne-Laure Valton, Rasmus O Bak, Charles H Li, Johanna Goldmann, Bryan R Lajoie, Zi Peng Fan, Alla A Sigova, Jessica Reddy, Diego Borges-Rivera, Tong Ihn Lee, Rudolf Jaenisch, Matthew H Porteus, Job Dekker, and Richard A Young. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (New York, N.Y.)*, 351:1454–1458, March 2016. ISSN 1095-9203. doi: 10.1126/science.aad9024. 38

Andrea H Horakova, Shawn C Moseley, Christine R McLaughlin, Deanna C Tremblay, and Brian P Chadwick. The macrosatellite dxz4 mediates ctcf-dependent long-range intrachromosomal interactions on the human inactive x chromosome. *Human molecular genetics*, 21:4367–4377, October 2012. ISSN 1460-2083. doi: 10.1093/hmg/dds270. 171

Chunhui Hou, Li Li, Zhaohui S Qin, and Victor G Corces. Gene density, transcription, and insulators contribute to the partition of the drosophila genome into physical domains. *Molecular cell*, 48:471–484, November 2012. ISSN 1097-4164. doi: 10.1016/j.molcel.2012.08.031. 19

Tsung-Han S Hsieh, Assaf Weiner, Bryan Lajoie, Job Dekker, Nir Friedman, and Oliver J Rando. Mapping nucleosome resolution chromosome folding in yeast by micro-c. *Cell*, 162:108–119, July 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.05.048. 47, 49, 96

Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S Liu. Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics (Oxford, England)*, 28:3131–3133, December 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts570. 58, 59, 62, 147

Jialiang Huang, Eugenio Marco, Luca Pinello, and Guo-Cheng Yuan. Predicting chromatin organization using histone marks. *Genome biology*, 16:162, August 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0740-z. 22

Jim R Hughes, Nigel Roberts, Simon McGowan, Deborah Hay, Eleni Giannoulatou, Magnus Lynch, Marco De Gobbi, Stephen Taylor, Richard Gibbons, and Douglas R Higgs. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature genetics*, 46:205–212, February 2014. ISSN 1546-1718. doi: 10.1038/ng.2871. 48, 49

REFERENCES

Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9:999–1003, October 2012. ISSN 1548-7105. doi: 10.1038/nmeth.2148. xxi, 53, 54, 59, 60, 61, 69, 79, 104, 155, 175, 196

Maxim V Imakaev, Geoffrey Fudenberg, and Leonid A Mirny. Modeling chromosomes: Beyond pretty pictures. *FEBS letters*, 589:3031–3036, October 2015. ISSN 1873-3468. doi: 10.1016/j.febslet.2015.09.004. 25

Fulai Jin, Yan Li, Jesse R Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D Schmitt, Celso A Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503:290–294, November 2013. ISSN 1476-4687. doi: 10.1038/nature12644. 26, 186

Peter A Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13:484–492, May 2012. ISSN 1471-0064. doi: 10.1038/nrg3230. 8

Predrag Jovanovic, Marija Mihajlovic, Jasmina Djordjevic-Jocic, Slobodan Vlajkovic, Sonja Cekic, and Vladisav Stefanovic. Ocular melanoma: an overview of the current status. *International journal of clinical and experimental pathology*, 6:1230–1244, 2013. ISSN 1936-2625. 145

Natalie Jger, Matthias Schlesner, David T W Jones, Simon Raffel, Jan-Philipp Mallm, Kristin M Junge, Dieter Weichenhan, Tobias Bauer, Naveed Ishaque, Marcel Kool, Paul A Northcott, Andrey Korshunov, Ruben M Drews, Jan Koster, Rogier Versteeg, Julia Richter, Michael Hummel, Stephen C Mack, Michael D Taylor, Hendrik Witt, Benedict Swartman, Dietrich Schulte-Bockholt, Marc Sultan, Marie-Laure Yaspo, Hans Lehrach, Barbara Hutter, Benedikt Brors, Stephan Wolf, Christoph Plass, Reiner Siebert, Andreas Trumpp, Karsten Rippe, Irina Lehmann, Peter Lichter, Stefan M Pfister, and Roland Eils. Hypermutation of the inactive x chromosome is a frequent event in cancer. *Cell*, 155:567–581, October 2013. ISSN 1097-4172. doi: 10.1016/j.cell.2013.09.042. 160

Stephan Kadauke and Gerd A Blobel. Mitotic bookmarking by transcription factors. *Epigenetics & chromatin*, 6:6, April 2013. doi: 10.1186/1756-8935-6-6. 35

Michael H Kagey, Jamie J Newman, Steve Bilodeau, Ye Zhan, David A Orlando, Nynke L van Berkum, Christopher C Ebmeier, Jesse Goossens, Peter B Rahl, Stuart S Levine, Dylan J Taatjes, Job Dekker, and Richard A Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467:430–435, September 2010. ISSN 1476-4687. doi: 10.1038/nature09380. 26

Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, 30:90–98, December 2011. ISSN 1546-1696. doi: 10.1038/nbt.2057. 20, 49

Riku Katainen, Kashyap Dave, Esa Pitknen, Kimmo Palin, Teemu Kivioja, Niko Vlimki, Alexandra E Gylfe, Heikki Ristolainen, Ulrika A Hnninen, Tatiana Cajuso, Johanna Kondelin, Tomas Tanskanen, Jukka-Pekka Mecklin, Heikki Jrvinen, Laura Renkonen-Sinisalo, Anna Lepist, Eevi Kaasinen, Outi Kilpivaara, Sari Tuupanen, Martin Enge, Jussi Taipale, and Lauri A Aaltonen. Ctcf/cohesin-binding sites are frequently mutated in cancer. *Nature genetics*, 47:818–821, July 2015. ISSN 1546-1718. doi: 10.1038/ng.3335. 39

Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14:R36, April 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-4-r36. 174

Jan O Korbel and Charles Lee. Genome assembly and haplotyping with hi-c. *Nature biotechnology*, 31:1099–1101, December 2013. ISSN 1546-1696. doi: 10.1038/nbt.2764. 199

R D Kornberg and Y Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98:285–294, August 1999. ISSN 0092-8674. 4

Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128:693–705, February 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.02.005. 8, 10

Theresia R Kress, Paola Pellanda, Luca Pellegrinet, Valerio Bianchi, Paola Nicoli, Mirko Doni, Camilla Recordati, Salvatore Bianchi, Luca Rotta, Thelma Capra, Micol Rav, Alessandro Verrecchia, Enrico Radaelli, Trevor D Littlewood, Gerard I Evan, and Bruno Amati. Identification of myc-dependent transcriptional programs in oncogene-addicted liver tumors. *Cancer research*, 76:3463–3472, June 2016. ISSN 1538-7445. doi: 10.1158/0008-5472.CAN-16-0316. 160, 161, 162

Peter H L Krijger and Wouter de Laat. Identical cells with different 3d genomes; cause and consequences? *Current opinion in genetics & development*, 23:191–196, April 2013. ISSN 1879-0380. doi: 10.1016/j.gde.2012.12.010. 28

Peter Hugo Lodewijk Krijger and Wouter de Laat. Regulation of disease-associated gene expression in the 3d genome. *Nature reviews. Molecular cell biology*, 17:771–782, December 2016. ISSN 1471-0080. doi: 10.1038/nrm.2016.138. xxi, 36, 37, 38, 44

Felix Krueger and Simon R Andrews. Snpsplit: Allele-specific splitting of alignments between genomes with known snp genotypes. *F1000Research*, 5:1479, 2016. ISSN 2046-1402. doi: 10.12688/f1000research.9037.2. 174

Marta Kulis and Manel Esteller. Dna methylation and cancer. *Advances in genetics*, 70:27–56, 2010. ISSN 0065-2660. doi: 10.1016/B978-0-12-380866-0.60002-2. 12

Alexander Kenig, Thomas Linhart, Katrin Schlengemann, Kristina Reutlinger, Jessica Wegele, Guido Adler, Garima Singh, Leonie Hofmann, Steffen Kunsch, Thomas Bch, Eva Schfer, Thomas M Gress, Martin E Fernandez-Zapico, and Volker Ellenrieder. Nfat-induced histone acetylation relay switch promotes c-myc-dependent growth in pancreatic cancer cells. *Gastroenterology*, 138:1189–99.e1–2, March 2010. ISSN 1528-0012. doi: 10.1053/j.gastro.2009.10.045. 164

Lindsay M LaFave, Wendy Bguelin, Richard Koche, Matt Teater, Barbara Spitzer, Alan Chramiec, Efthymia Papalexi, Matthew D Keller, Todd Hricik, Katerina Konstantinoff, Jean-Baptiste Micol, Benjamin Durham, Sarah K Knutson, John E Campbell, Gil Blum, Xinxu Shi, Emma H Doud, Andrei V Krivtsov, Young Rock Chung, Inna Khodos, Elisa de Stanchina, Ouathek Ouerfelli, Prasad S Adusumilli, Paul M Thomas, Neil L Kelleher, Minkui Luo, Heike Keilhack, Omar Abdel-Wahab, Ari Melnick, Scott A Armstrong, and Ross L Levine. Loss of bap1 function leads to ezh2-dependent transformation. *Nature medicine*, 21:1344–1349, November 2015. ISSN 1546-170X. doi: 10.1038/nm.3947. 157

Bryan R Lajoie, Job Dekker, and Noam Kaplan. The hitchhiker's guide to hi-c analysis: practical guidelines. *Methods (San Diego, Calif.)*, 72:65–75, January 2015. ISSN 1095-9130. doi: 10.1016/j.ymeth.2014.10.031. 52, 66, 68

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9:357–359, March 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1923. 174

Franois Le Dily, Davide Ba, Andy Pohl, Guillermo P Vicent, Franois Serra, Daniel Soronellas, Giancarlo Castellano, Roni H G Wright, Cecilia Ballare, Guillaume Filion, Marc A Marti-Renom, and Miguel Beato. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & development*, 28:2151–2162, October 2014. ISSN 1549-5477. doi: 10.1101/gad.241422.114. 186

J T Lee and N Lu. Targeted mutagenesis of tsix leads to nonrandom x inactivation. *Cell*, 99:47–57, October 1999. ISSN 0092-8674. 29

Timothy J Ley, Li Ding, Matthew J Walter, Michael D McLellan, Tamara Lamprecht, David E Larson, Cyriac Kandoth, Jacqueline E Payton, Jack Baty, John Welch, Christopher C Harris, Cheryl F Lichti, R Reid Townsend, Robert S Fulton, David J Dooling, Daniel C Koboldt, Heather Schmidt, Qunyuan Zhang, John R Osborne, Ling Lin, Michelle O'Laughlin, Joshua F McMichael, Kim D Delehaunty, Sean D McGrath, Lucinda A Fulton, Vincent J Magrini, Tammi L Vickery, Jasreet Hundal, Lisa L Cook, Joshua J Conyers, Gary W Swift, Jerry P Reed, Patricia A Alldredge, Todd Wylie, Jason Walker, Joelle Kalicki, Mark A Watson, Sharon Heath, William D Shannon, Nobish Varghese, Rakesh Nagarajan, Peter Westervelt, Michael H Tomasson, Daniel C Link, Timothy A Graubert, John F DiPersio, Elaine R Mardis, and Richard K Wilson. Dnmt3a mutations in acute myeloid leukemia. *The New England journal of medicine*, 363:2424–2433, December 2010. ISSN 1533-4406. doi: 10.1056/NEJMoa1005143. 12

En Li and Yi Zhang. Dna methylation in mammals. *Cold Spring Harbor perspectives in biology*, 6:a019133, May 2014. ISSN 1943-0264. doi: 10.1101/cshperspect.a019133. 8

Guoliang Li, Liuyang Cai, Huidan Chang, Ping Hong, Qiangwei Zhou, Ekaterina V Kulakova, Nikolay A Kolchanov, and Yijun Ruan. Chromatin interaction analysis with paired-end tag (chia-pet) sequencing technology and application. *BMC genomics*, 15 Suppl 12:S11, 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-S12-S11. 46

Xin Li, Xiao-Long Cui, Jia-Qiang Wang, Yu-Kai Wang, Yu-Fei Li, Le-Yun Wang, Hai-Feng Wan, Tian-Da Li, Gui-Hai Feng, Ling Shuai, Zhi-Kun Li, Qi Gu, Jie Hao, Liu Wang, Xiao-Yang Zhao, Zhong-Hua Liu, Xiu-Jie Wang, Wei Li, and Qi Zhou. Generation and application of mouse-rat allodiploid embryonic stem cells. *Cell*, 164:279–292, January 2016. ISSN 1097-4172. doi: 10.1016/j.cell.2015.11.035. 165, 167

Xingwang Li, Oscar Junhong Luo, Ping Wang, Meizhen Zheng, Danjuan Wang, Emaly Piecuch, Jacqueline Jufen Zhu, Simon Zhongyuan Tian, Zhonghui Tang, Guoliang Li, and Yijun Ruan. Long-read chia-pet for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nature protocols*, 12:899–915, May 2017. ISSN 1750-2799. doi: 10.1038/nprot.2017.012. 52

Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30:923–930, April 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt656. 174

Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326:289–293, October 2009. ISSN 1095-9203. doi: 10.1126/science.1181369. 19, 20, 45, 46, 47, 49, 64, 65, 79

Charles Y Lin, Jakob Lovn, Peter B Rahl, Ronald M Paranal, Christopher B Burge, James E Bradner, Tong Ihn Lee, and Richard A Young. Transcriptional amplification in tumor cells with elevated c-myc. *Cell*, 151:56–67, September 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.08.026. 160

Aaron T L Lun and Gordon K Smyth. diffhic: a bioconductor package to detect differential genomic interactions in hi-c data. *BMC bioinformatics*, 16:258, August 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0683-0. 197

M F LYON. Gene action in the x-chromosome of the mouse (mus musculus l.). *Nature*, 190:372–373, April 1961. ISSN 0028-0836. 28

Mary F Lyon. The lyon and the line hypothesis. *Seminars in cell & developmental biology*, 14:313–318, December 2003. ISSN 1084-9521. 29

Wenxiu Ma, Ferhat Ay, Choli Lee, Gunhan Gulsoy, Xinxian Deng, Savannah Cook, Jennifer Hesson, Christopher Cavanaugh, Carol B Ware, Anton Krumm, Jay Shendure, Carl Anthony Blau, Christine M Disteche, William S Noble, and Zhijun Duan. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincrna genes. *Nature methods*, 12:71–78, January 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3205. 47, 48, 49, 55, 96

Emily Maclary, Michael Hinten, Clair Harris, and Sundeep Kalantry. Long nonoding rnas in the x-inactivation center. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 21:601–614, December 2013. ISSN 1573-6849. doi: 10.1007/s10577-013-9396-2. 29

REFERENCES

Winifred Mak, Tatyana B Nesterova, Mariana de Napoles, Ruth Appanah, Shinya Yamanaka, Arie P Otte, and Neil Brockdorff. Reactivation of the paternal x chromosome in early mouse embryos. *Science (New York, N.Y.)*, 303:666–669, January 2004. ISSN 1095-9203. doi: 10.1126/science.1092674. 28

Raphal Margueron and Danny Reinberg. The polycomb complex prc2 and its mark in life. *Nature*, 469:343–349, January 2011. ISSN 1476-4687. doi: 10.1038/nature09784. 31

Hendrik Marks, Jennifer C Chow, Sergei Denissov, Kees-Jan Franoijs, Neil Brockdorff, Edith Heard, and Hendrik G Stunnenberg. High-resolution analysis of epigenetic changes associated with x inactivation. *Genome research*, 19:1361–1373, August 2009. ISSN 1088-9051. doi: 10.1101/gr.092643.109. 164

Karen J Meaburn. Spatial genome organization and its emerging role as a potential diagnosis tool. *Frontiers in genetics*, 7:134, 2016. ISSN 1664-8021. doi: 10.3389/fgene.2016.00134. 186

Karen J Meaburn, Tom Misteli, and Evi Soutoglou. Spatial genome organization in the formation of chromosomal translocations. *Seminars in cancer biology*, 17:80–90, February 2007. ISSN 1044-579X. doi: 10.1016/j.semcancer.2006.10.008. 20

Matthias Merkenschlager and Elphge P Nora. Ctcf and cohesin in genome folding and transcriptional gene regulation. *Annual review of genomics and human genetics*, 17:17–43, August 2016. ISSN 1545-293X. doi: 10.1146/annurev-genom-083115-022339. 27

Eran Meshorer, Dhananjay Yellajoshula, Eric George, Peter J Scambler, David T Brown, and Tom Misteli. Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Developmental cell*, 10:105–116, January 2006. ISSN 1534-5807. doi: 10.1016/j.devcel.2005.10.017. 33

Borbala Mifsud, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, Simon Andrews, William Grey, Philip A Ewels, Bram Herman, Scott Happe, Andy Higgs, Emily LeProust, George A Follows, Peter Fraser, Nicholas M Luscombe, and Cameron S Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. *Nature genetics*, 47:598–606, June 2015. ISSN 1546-1718. doi: 10.1038/ng.3286. 48, 49

Borbala Mifsud, Inigo Martincorena, Elodie Darbo, Robert Sugar, Stefan Schoenfelder, Peter Fraser, and Nicholas M Luscombe. Gothic, a probabilistic model to resolve complex biases and to identify real interactions in hi-c data. *PloS one*, 12:e0174744, 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0174744. 198

Fionnuala Morrish, Christopher Giedt, and David Hockenbery. c-myc apoptotic function is mediated by nrf-1 target genes. *Genes & development*, 17:240–255, January 2003. ISSN 0890-9369. doi: 10.1101/gad.1032503. 164

Raphal Mourad and Olivier Cuvier. Predicting the spatial organization of chromosomes using epigenetic data. *Genome biology*, 16:182, August 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0752-8. 20

Maxwell R Mumbach, Adam J Rubin, Ryan A Flynn, Chao Dai, Paul A Khavari, William J Greenleaf, and Howard Y Chang. Hichip: efficient and sensitive analysis of protein-directed genome architecture. *Nature methods*, 13:919–922, November 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3999. 46, 98, 196

Catherine A Musselman, Marie-Eve Lalonde, Jacques Ct, and Tatiana G Kutateladze. Perceiving the epigenetic landscape through histone readers. *Nature structural & molecular biology*, 19:1218–1227, December 2012. ISSN 1545-9985. doi: 10.1038/nsmb.2436. 8

Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502:59–64, October 2013. ISSN 1476-4687. doi: 10.1038/nature12593. 25, 27, 28, 49, 192, 197

Takashi Nagano, Csilla Vrnai, Stefan Schoenfelder, Biola-Maria Javierre, Steven W Wingett, and Peter Fraser. Comparison of hi-c results using in-solution versus in-nucleus ligation. *Genome biology*, 16:175, August 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0753-7. 47

Takashi Nagano, Yaniv Lubling, Csilla Vrnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547:61–67, July 2017. ISSN 1476-4687. doi: 10.1038/nature23001. 172

Varun Narendra, Pedro P Rocha, Disi An, Ramya Raviram, Jane A Skok, Esteban O Mazzoni, and Danny Reinberg. Ctcf establishes discrete functional chromatin domains at the hox clusters during differentiation. *Science (New York, N.Y.)*, 347:1017–1021, February 2015. ISSN 1095-9203. doi: 10.1126/science.1262088. 27

Natalia Naumova and Job Dekker. Integrating one-dimensional and three-dimensional maps of genomes. *Journal of cell science*, 123:1979–1988, June 2010. ISSN 1477-9137. doi: 10.1242/jcs.051631. 20

Natalia Naumova, Maxim Imakaev, Geoffrey Fudenberg, Ye Zhan, Bryan R Lajoie, Leonid A Mirny, and Job Dekker. Organization of the mitotic chromosome. *Science (New York, N.Y.)*, 342:948–953, November 2013. ISSN 1095-9203. doi: 10.1126/science.1236083. 24, 35, 36, 63

Elphge P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485:381–385, April 2012. ISSN 1476-4687. doi: 10.1038/nature11049. 21, 22, 23, 27, 30, 31, 32, 63, 186, 187

Elphge P. Nora, Anton Goloborodko, AnneLaure Valton, Johan Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A. Mirny, and Benoit G. Bruneau. Targeted degradation of ctcf decouples local insulation of chromosome domains from higher order genomic compartmentalization. *BioRxiv*, 2016. doi: http://dx.doi.org/10.1101/095802. 27

Elphge P Nora, Anton Goloborodko, Anne-Laure Valton, Johan H Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A Mirny, and Benoit G Bruneau. Targeted degradation of ctcf decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, 169:930–944.e22, May 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.05.004. 184, 185, 187

Paul A Northcott, Catherine Lee, Thomas Zichner, Adrian M Sttz, Serap Erkek, Daisuke Kawauchi, David J H Shih, Volker Hovestadt, Marc Zapatka, Dominik Sturm, David T W Jones, Marcel Kool, Marc Remke, Florence M G Cavalli, Scott Zuyderduyn, Gary D Bader, Scott VandenBerg, Lourdes Adriana Esparza, Marina Ryzhova, Wei Wang, Andrea Wittmann, Sebastian Stark, Laura Sieber, Huriye Seker-Cin, Linda Linke, Fabian Kratochwil, Natalie Jger, Ivo Buchhalter, Charles D Imbusch, Gideon Zipprich, Benjamin Raeder, Sabine Schmidt, Nicolle Diessl, Stephan Wolf, Stefan Wiemann, Benedikt Brors, Chris Lawerenz, Jrgen Eils, Hans-Jrg Warnatz, Thomas Risch, Marie-Laure Yaspo, Ursula D Weber, Cynthia C Bartholomae, Christof von Kalle, Eszter Turnyi, Peter Hauser, Emma Sanden, Anna Darabi, Peter Siesj, Jaroslav Sterba, Karel Zitterbart, David Sumerauer, Peter van Sluis, Rogier Versteeg, Richard Volckmann, Jan Koster, Martin U Schuhmann, Martin Ebinger, H Leighton Grimes, Giles W Robinson, Amar Gajjar, Martin Mynarek, Katja von Hoff, Stefan Rutkowski, Torsten Pietsch, Wolfram Scheurlen, Jrg Felsberg, Guido Reifenberger, Andreas E Kulozik, Andreas von Deimling, Olaf Witt, Roland Eils, Richard J Gilbertson, Andrey Korshunov, Michael D Taylor, Peter Lichter, Jan O Korbel, Robert J Wechsler-Reya, and Stefan M Pfister. Enhancer hijacking activates gfi1 family oncogenes in medulloblastoma. *Nature*, 511:428–434, July 2014. ISSN 1476-4687. doi: 10.1038/nature13379. 38

Fariba Nmati, Xavier Sastre-Garau, Ccile Laurent, Jrme Couturier, Pascale Mariani, Laurence Desjardins, Sophie Piperno-Neumann, Olivier Lantz, Bernard Asselain, Corine Plancher, Delphine Robert, Isabelle Pguillet, Marie-Hlne Donnadieu, Ahmed Dahmani, Marie-Andre Bessard, David Gentien, Ccile Reyes, Simon Saule, Emmanuel Barillot, Sergio Roman-Roman, and Didier Decaudin. Establishment and characterization of a panel of human uveal melanoma xenografts derived from primary and/or metastatic tumors. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 16:2352–2362, April 2010. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-09-3066. 146

Attila Nmeth, Ana Conesa, Javier Santoyo-Lopez, Ignacio Medina, David Montaner, Blint Pterfia, Irina Solovei, Thomas Cremer, Joaquin Dopazo, and Gernot Lngst. Initial genomics of the human nucleolus. *PLoS genetics*, 6: e1000889, March 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000889. 19

Chin-Tong Ong and Victor G Corces. Ctcf: an architectural protein bridging genome topology and function. *Nature reviews. Genetics*, 15:234–246, April 2014. ISSN 1471-0064. doi: 10.1038/nrg3663. 26

P Oudet, M Gross-Bellard, and P Chambon. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell*, 4:281–300, April 1975. ISSN 0092-8674. 4

Luis A Parada, Philip G McQueen, and Tom Misteli. Tissue-specific spatial organization of genomes. *Genome biology*, 5:R44, 2004. ISSN 1474-760X. doi: 10.1186/gb-2004-5-7-r44. 19

Jennifer E Phillips-Cremins, Michael E G Sauria, Amartya Sanyal, Tatiana I Gerasimova, Bryan R Lajoie, Joshua S K Bell, Chin-Tong Ong, Tracy A Hookway, Changying Guo, Yuhua Sun, Michael J Bland, William Wagstaff, Stephen Dalton, Todd C McDevitt, Ranjan Sen, Job Dekker, James Taylor, and Victor G Corces. Architectural protein subclasses shape 3d organization of genomes during lineage commitment. *Cell*, 153:1281–1295, June 2013. ISSN 1097-4172. doi: 10.1016/j.cell.2013.04.053. 21, 184

Benjamin D Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L Vera, Yanli Wang, R Scott Hansen, Theresa K Canfield, Robert E Thurman, Yong Cheng, Gnhan Glsoy, Jonathan H Dennis, Michael P Snyder, John A Stamatoyannopoulos, James Taylor, Ross C Hardison, Tamer Kahveci, Bing Ren, and David M Gilbert. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515: 402–405, November 2014. ISSN 1476-4687. doi: 10.1038/nature13986. 22, 24, 184

Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature biotechnology*, 28:1057–1068, October 2010. ISSN 1546-1696. doi: 10.1038/nbt.1685. 12, 14

Rebecca C Poulos, Julie A I Thoms, Yi Fang Guan, Ashwin Unnikrishnan, John E Pimanda, and Jason W H Wong. Functional mutations form at ctcf-cohesin binding sites in melanoma due to uneven nucleotide excision repair across the motif. *Cell reports*, 17:2865–2872, December 2016. ISSN 2211-1247. doi: 10.1016/j.celrep.2016.11.055. 39

Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26:841–842, March 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq033. 176

Vijay Ramani, Jay Shendure, and Zhijun Duan. Understanding spatial genome organization: Methods and insights. *Genomics, proteomics & bioinformatics*, 14:7–20, February 2016. ISSN 2210-3244. doi: 10.1016/j.gpb.2016.01.002. 17, 43, 49

Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell hi-c. *Nature methods*, 14:263–266, March 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4155. 192

Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159:1665–1680, December 2014. ISSN 1097-4172. doi: 10.1016/j.cell.2014.11.021. 20, 24, 25, 26, 27, 32, 46, 47, 49, 56, 59, 66, 67, 170, 184, 191

Keith D Robertson. Dna methylation and human disease. *Nature reviews. Genetics*, 6:597–610, August 2005. ISSN 1471-0056. doi: 10.1038/nrg1655. 12

Santiago Ropero and Manel Esteller. The role of histone deacetylases (hdacs) in human cancer. *Molecular oncology*, 1:19–25, June 2007. ISSN 1878-0261. doi: 10.1016/j.molonc.2007.01.001. 14

REFERENCES

David M Roy, Logan A Walsh, and Timothy A Chan. Driver mutations of cancer epigenomes. *Protein & cell*, 5:265–296, April 2014. ISSN 1674-8018. doi: 10.1007/s13238-014-0031-6. 13, 14

Tyrone Ryba, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C Schulz, Allan J Robins, Stephen Dalton, and David M Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*, 20:761–770, June 2010. ISSN 1549-5469. doi: 10.1101/gr.099655.109. 24

Arianna Sab and Bruno Amati. Genome recognition by myc. *Cold Spring Harbor perspectives in medicine*, 4, February 2014. ISSN 2157-1422. doi: 10.1101/cshperspect.a014191. 163

Anthony D Schmitt, Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L Tan, Yun Li, Shin Lin, Yiing Lin, Cathy L Barr, and Bing Ren. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports*, 17:2042–2059, November 2016. ISSN 2211-1247. doi: 10.1016/j.celrep.2016.10.061. 28, 67

Yuri B Schwartz and Vincenzo Pirrotta. A new world of polycombs: unexpected partnerships and emerging functions. *Nature reviews. Genetics*, 14:853–864, December 2013. ISSN 1471-0064. doi: 10.1038/nrg3603. 31

Dirk Schbeler. Function and information content of dna methylation. *Nature*, 517:321–326, January 2015. ISSN 1476-4687. doi: 10.1038/nature14192. 6

Vlad C Seitan, Andre J Faure, Ye Zhan, Rachel Patton McCord, Bryan R Lajoie, Elizabeth Ing-Simmons, Boris Lenhard, Luca Giorgetti, Edith Heard, Amanda G Fisher, Paul Flicek, Job Dekker, and Matthias Merkenschlager. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome research*, 23:2066–2077, December 2013. ISSN 1549-5469. doi: 10.1101/gr.161620.113. 27

Nicolas Servant, Bryan R Lajoie, Elphge P Nora, Luca Giorgetti, Chong-Jian Chen, Edith Heard, Job Dekker, and Emmanuel Barillot. Hitc: exploration of high-throughput 'c' experiments. *Bioinformatics (Oxford, England)*, 28: 2843–2844, November 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts521. 62, 101, 177

Tom Sexton, Heiko Schober, Peter Fraser, and Susan M Gasser. Gene regulation through nuclear organization. *Nature structural & molecular biology*, 14:1049–1055, November 2007. ISSN 1545-9985. doi: 10.1038/nsmb1324. 188

Tom Sexton, Sreenivasulu Kurukuti, Jennifer A Mitchell, David Umlauf, Takashi Nagano, and Peter Fraser. Sensitive detection of chromatin coassociations using enhanced chromosome conformation capture on chip. *Nature protocols*, 7:1335–1350, June 2012. ISSN 1750-2799. doi: 10.1038/nprot.2012.071. 44

Sigal Shachar, Gianluca Pegoraro, and Tom Misteli. Hipmap: A high-throughput imaging method for mapping spatial gene positions. *Cold Spring Harbor symposia on quantitative biology*, 80:73–81, 2015. ISSN 1943-4456. doi: 10.1101/sqb.2015.80.027417. 17

Yoli Shavit, Ivan Merelli, Luciano Milanesi, and Pietro Lio'. How computer science can help in understanding the 3d genome architecture. *Briefings in bioinformatics*, 17:733–744, September 2016. ISSN 1477-4054. doi: 10.1093/bib/bbv085. 51

Hui Shen and Peter W Laird. Interplay between the cancer genome and epigenome. *Cell*, 153:38–55, March 2013. ISSN 1097-4172. doi: 10.1016/j.cell.2013.03.008. 11

Matthew D Simon, Stefan F Pinter, Rui Fang, Kavitha Sarma, Michael Rutenberg-Schoenberg, Sarah K Bowman, Barry A Kesner, Verena K Maier, Robert E Kingston, and Jeannie T Lee. High-resolution xist binding maps reveal two-step spreading during x-chromosome inactivation. *Nature*, 504:465–469, December 2013. ISSN 1476-4687. doi: 10.1038/nature12719. 29

Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nature genetics*, 38:1348–1354, November 2006. ISSN 1061-4036. doi: 10.1038/ng1896. 44, 49

Peter Staller. Genetic heterogeneity and chromatin modifiers in renal clear cell carcinoma. *Future oncology (London, England)*, 6:897–900, June 2010. ISSN 1744-8301. doi: 10.2217/fon.10.50. 14

Orsolya Symmons, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nassari, Wibke Schwarzer, Laurence Ettwiller, and Franois Spitz. Functional and topological characteristics of mammalian regulatory domains. *Genome research*, 24:390–400, March 2014. ISSN 1549-5469. doi: 10.1101/gr.163519.113. 23

Phillippa C Taberlay, Aaron L Statham, Theresa K Kelly, Susan J Clark, and Peter A Jones. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies dna methylation of enhancers and insulators in cancer. *Genome research*, 24:1421–1432, September 2014. ISSN 1549-5469. doi: 10.1101/gr.163485.113. 36

Phillippa C Taberlay, Joanna Achinger-Kawecka, Aaron T L Lun, Fabian A Buske, Kenneth Sabir, Cathryn M Gould, Elena Zotenko, Saul A Bert, Katherine A Giles, Denis C Bauer, Gordon K Smyth, Clare Stirzaker, Sean I O'Donoghue, and Susan J Clark. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome research*, 26:719–731, June 2016. ISSN 1549-5469. doi: 10.1101/gr.201517.115. 15, 38, 40, 41, 154, 189

Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M Iyer, David R Liu, L Aravind, and Anjana Rao. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science (New York, N.Y.)*, 324:930–935, May 2009. ISSN 1095-9203. doi: 10.1126/science.1170116. 6

Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, Paul Michalski, Emaly Piecuch, Ping Wang, Danjuan Wang, Simon Zhongyuan Tian, May Penrad-Mobayed, Laurent M Sachs, Xiaoan Ruan, Chia-Lin Wei,

Edison T Liu, Grzegorz M Wilczynski, Dariusz Plewczynski, Guoliang Li, and Yijun Ruan. Ctcf-mediated human 3d genome architecture reveals chromatin topology for transcription. *Cell*, 163:1611–1627, December 2015. ISSN 1097-4172. doi: 10.1016/j.cell.2015.11.024. 49

Morgane Thomas-Chollier, Carl Herrmann, Matthieu Defrance, Olivier Sand, Denis Thieffry, and Jacques van Helden. Rsat peak-motifs: motif analysis in full-size chip-seq datasets. *Nucleic acids research*, 40:e31, February 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr1104. 176

Winston Timp and Andrew P Feinberg. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nature reviews. Cancer*, 13:497–510, July 2013. ISSN 1474-1768. doi: 10.1038/nrc3486. 15

Jean-Valery Turatsinze, Morgane Thomas-Chollier, Matthieu Defrance, and Jacques van Helden. Using rsat to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature protocols*, 3:1578–1588, 2008. ISSN 1750-2799. doi: 10.1038/nprot.2008.97. 176

Anne-Laure Valton and Job Dekker. Tad disruption as oncogenic driver. *Current opinion in genetics & development*, 36:34–40, February 2016. ISSN 1879-0380. doi: 10.1016/j.gde.2016.03.008. 188

Harmen J G van de Werken, Gilad Landan, Sjoerd J B Holwerda, Michael Hoichman, Petra Klous, Ran Chachik, Erik Splinter, Christian Valdes-Quezada, Yuva Oz, Britta A M Bouwman, Marjon J A M Verstegen, Elzo de Wit, Amos Tanay, and Wouter de Laat. Robust 4c-seq data analysis to screen for regulatory dna interactions. *Nature methods*, 9:969–972, October 2012. ISSN 1548-7105. doi: 10.1038/nmeth.2173. 45

Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics (Oxford, England)*, 30:i26–i33, June 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu268. 63

Nelle Varoquaux, Ivan Liachko, Ferhat Ay, Joshua N Burton, Jay Shendure, Maitreya J Dunham, Jean-Philippe Vert, and William S Noble. Accurate identification of centromere locations in yeast genomes using hi-c. *Nucleic acids research*, 43:5331–5339, June 2015. ISSN 1362-4962. doi: 10.1093/nar/gkv424. 19, 57

Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science (New York, N.Y.)*, 339:1546–1558, March 2013. ISSN 1095-9203. doi: 10.1126/science.1235122. 36

Emanuela V Volpi and Joanna M Bridger. Fish glossary: an overview of the fluorescence in situ hybridization technique. *BioTechniques*, 45:385–6, 388, 390 passim, October 2008. ISSN 0736-6205. doi: 10.2144/000112811. 17, 42

Siyuan Wang, Jun-Han Su, Brian J Beliveau, Bogdan Bintu, Jeffrey R Moffitt, Chao-ting Wu, and Xiaowei Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Science (New York, N.Y.)*, 353:598–602, August 2016. ISSN 1095-9203. doi: 10.1126/science.aaf8084. 17

Michel Wassef, Veronica Rodilla, Aurlie Teissandier, Bruno Zeitouni, Nadege Gruel, Benjamin Sadacca, Marie Irondelle, Margaux Charruel, Bertrand Ducos, Audrey Michaud, Matthieu Caron, Elisabetta Marangoni, Philippe Chavrier, Christophe Le Tourneau, Maud Kamal, Eric Pasmant, Michel Vidaud, Nicolas Servant, Fabien Reyal, Dider Meseure, Anne Vincent-Salomon, Silvia Fre, and Raphal Margueron. Impaired prc2 activity promotes transcriptional instability and favors breast tumorigenesis. *Genes & development*, 29:2547–2562, December 2015. ISSN 1549-5477. doi: 10.1101/gad.269522.115. 15

Kerstin S Wendt, Keisuke Yoshida, Takehiko Itoh, Masashige Bando, Birgit Koch, Erika Schirghuber, Shuichi Tsutsumi, Genta Nagae, Ko Ishihara, Tsuyoshi Mishiro, Kazuhide Yahata, Fumio Imamoto, Hiroyuki Aburatani, Mitsuyoshi Nakao, Naoko Imamoto, Kazuhiro Maeshima, Katsuhiko Shirahige, and Jan-Michael Peters. Cohesin mediates transcriptional insulation by ccctc-binding factor. *Nature*, 451:796–801, February 2008. ISSN 1476-4687. doi: 10.1038/nature06634. 26

Iain Williamson, Soizik Berlivet, Ragnhild Eskeland, Shelagh Boyle, Robert S Illingworth, Denis Paquette, Jose Dostie, and Wendy A Bickmore. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes & development*, 28:2778–2791, December 2014. ISSN 1549-5477. doi: 10.1101/gad.251694.114. 42

Steven Wingett, Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews. Hicup: pipeline for mapping and processing hi-c data. *F1000Research*, 4:1310, 2015. doi: 10.12688/f1000research.7334.1. 61, 175, 193

Anton Wutz. Gene silencing in x-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature reviews. Genetics*, 12:542–553, July 2011. ISSN 1471-0064. doi: 10.1038/nrg3035. 4

Hugo Wrtele and Pierre Chartrand. Genome-wide scanning of hoxb1-associated loci in mouse es cells using an open-ended chromosome conformation capture methodology. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 14:477–495, 2006. ISSN 0967-3849. doi: 10.1007/s10577-006-1075-0. 45

Eitan Yaffe and Amos Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43:1059–1065, October 2011. ISSN 1546-1718. doi: 10.1038/ng.947. 57, 58, 59, 103, 104

Eitan Yaffe, Shlomit Farkash-Amar, Andreas Polten, Zohar Yakhini, Amos Tanay, and Itamar Simon. Comparative analysis of dna replication timing reveals conserved large-scale chromosomal architecture. *PLoS genetics*, 6:e1001011, July 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001011. 24

Serdar Yavuzyigitoglu, Anna E Koopmans, Robert M Verdijk, Jolanda Vaarwater, Bert Eussen, Alice van Bodegom, Dion Paridaens, Emine Kili, Annelies de Klein, and Rotterdam Ocular Melanoma Study Group. Uveal melanomas with sf3b1 mutations: A distinct subclass associated with late-onset metastases. *Ophthalmology*, 123:1118–1128, May 2016. ISSN 1549-4713. doi: 10.1016/j.ophtha.2016.01.023. 146

REFERENCES

Karen I Zeller, XiaoDong Zhao, Charlie W H Lee, Kuo Ping Chiu, Fei Yao, Jason T Yustein, Hong Sain Ooi, Yuriy L Orlov, Atif Shahab, How Choong Yong, Yutao Fu, Zhiping Weng, Vladimir A Kuznetsov, Wing-Kin Sung, Yijun Ruan, Chi V Dang, and Chia-Lin Wei. Global mapping of c-myc binding sites and target gene networks in human b cells. *Proceedings of the National Academy of Sciences of the United States of America*, 103:17834–17839, November 2006. ISSN 0027-8424. doi: 10.1073/pnas.0604129103. 159

Yinxiu Zhan, Luca Mariani, Iros Barozzi, Edda G Schulz, Nils Bluthgen, Michael Stadler, Guido Tiana, and Luca Giorgetti. 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*, January 2017. ISSN 1549-5469. doi: 10.1101/gr.212803.116. 67, 186

Li-Feng Zhang, Khanh D Huynh, and Jeannie T Lee. Perinucleolar targeting of the inactive x during s phase: evidence for a role in the maintenance of silencing. *Cell*, 129:693–706, May 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.03.036. 32

Tianyi Zhang, Sarah Cooper, and Neil Brockdorff. The interplay of histone modifications - writers that read. *EMBO reports*, 16:1467–1481, November 2015. ISSN 1469-3178. doi: 10.15252/embr.201540945. 8

Yong Zhang, Tao Liu, Clifford A Meyer, Jrme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of chip-seq (macs). *Genome biology*, 9:R137, 2008. ISSN 1474-760X. doi: 10.1186/gb-2008-9-9-r137. 175

Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjlinder, Anita Gndr, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh Sandhu, Umashankar Singh, Vinod Pant, Vijay Tiwari, Sreenivasulu Kurukuti, and Rolf Ohlsson. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics*, 38:1341–1347, November 2006. ISSN 1061-4036. doi: 10.1038/ng1891. 45, 49

Vicky W Zhou, Alon Goren, and Bradley E Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews. Genetics*, 12:7–18, January 2011. ISSN 1471-0064. doi: 10.1038/nrg2905. 6

Yun Zhu, Zhao Chen, Kai Zhang, Mengchi Wang, David Medovoy, John W Whitaker, Bo Ding, Nan Li, Lina Zheng, and Wei Wang. Constructing 3d interaction maps from 1d epigenomes. *Nature communications*, 7:10812, March 2016. ISSN 2041-1723. doi: 10.1038/ncomms10812. 22

Huda Y Zoghbi and Arthur L Beaudet. Epigenetics and human disease. *Cold Spring Harbor perspectives in biology*, 8:a019497, February 2016. ISSN 1943-0264. doi: 10.1101/cshperspect.a019497. 12