Human genome segmentation into structural domains : from chromatin conformation data to nuclear functions

Rasha Boulos

▶ **To cite this version:**

Rasha Boulos. Human genome segmentation into structural domains : from chromatin conformation data to nuclear functions. Signal and Image Processing. Ecole normale supérieure de lyon - ENS LYON, 2015. English. NNT : 2015ENSL1024 . tel-01273972

HAL Id: tel-01273972
https://theses.hal.science/tel-01273972

Submitted on 15 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ENS DE LYON

Lab PHYS ENS de LYON

Thèse

en vue de l'obtention du grade de:

Docteur de l'Université de Lyon, délivré par l'École Normale Supérieure de Lyon

spécialité: Physique

Laboratoire de Physique
École Doctorale de Physique et d'Astrophysique de Lyon

présentée et soutenue publiquement le 21 Octobre 2015 par:

Rasha E. BOULOS

Human genome segmentation into structural domains: from chromatin conformation data to nuclear functions

---

Directeur de thèse:

Benjamin AUDIT

Après l'avis de:

Romain KOSZUL Stéphane ROBIN

Devant la comission d'examen formée par:

Alain ARNEODO	DR CNRS	Membre
Benjamin AUDIT	DR CNRS	Directeur
Pierre BORGNAT	CR CNRS	Membre
Peter R. COOK	Professeur, Université d'Oxford	Membre
Arach GOLDAR	Ingénieur CEA	Membre
Romain KOSZUL	CR CNRS	Rapporteur
Stéphane ROBIN	DR CNRS	Rapporteur

UNIVERSITÉ DE LYON

Acknowledgements

Je commence maintenant à rédiger ce qui me semble la partie la plus compliquée de ce manuscrit : les remerciements. Comme Bouddha l'a dit : *Sois reconnaissant envers tous, tous t'enseignent.* J'espère alors relever le défi et n'oublier personne.

Tout d'abord je tiens à remercier Benjamin Audit mon directeur de thèse pour son encadrement et son soutien durant ces trois années. Merci Benjamin pour ta présence et ton temps, merci pour ton encouragement et ta motivation. Merci d'avoir pu guider et diriger ce travail tout en me laissant la liberté de faire mes choix.

Un grand merci à Alain Arneodo qui a fortement collaboré à ce travail. Merci Alain pour toutes les discussions que nous avons eues, merci pour les explications que tu nous apportais. Merci d'avoir toujours donner du sens au travail et d'avoir apporter ce grain et cette motivation pour donner plus.

Merci à Pierre Borgnat et Nicolas Tremblay pour la belle collaboration. Merci à tous les deux pour le temps que vous avez mis et toutes les discussions enrichissantes.

Merci aux membres de mon jury pour l'intérêt qu'ils ont porté à ce travail : Stéphane Robin, Romain Koszul, Peter Cook, Arach Goldar, vos remarques et questions ont enrichies cette thèse.

Je remercie mes collègues de bureau qui ont fait que le travail s'est déroulé dans une ambiance formidable. Merci Elise d'avoir toujours répondu à mes questions et pour avoir enrichi ma culture sur la France, merci pour tous les mots et expressions que tu m'as appris. Laura merci pour ton amour de la vie et ta joie de vivre, merci pour toutes les bêtises qu'on a fait ensemble. Cristina pas seulement ma co-bureau mais ma chère coloc aussi! Merci pour tous les bons moments qu'on a passé ensemble au bureau ou ailleurs, pour les soirées à se plaindre ou à rigoler en inventant notre nouveau langage regrouppant plusieurs langues. Merci aussi à Guénola pour toutes nos discussions et son soutien quand j'avais des doutes.

Merci aux membres des équipes SISYPHE et IXXI : Patrice, Patrick, Pablo, Stéphane, Ronan, Roberto, Gabriel, Rodrigo, Benjamin, Jordan, Jinane...

Merci aux membres actuels et passés du LJC : Françoise, Lotfi, Hanna, Simona, Qiong-Xu, Pascale, Elodie, Johan, Xavier, Julien, les chimistes...

Merci à tous mes amis qui m'ont soutenu en ésperant n'oublier personne, merci Larry, Reina, Sirena, Bassem, Rudy, Elias, Emil, Elena, Mabelle. Merci à mes amis au Liban qui malgré la distance ont toujours été là pour me soutenir durant les jours les plus durs : Elia, Christian, Roland, Marc.

Une pensée à la personne qui n'a pas voulu être cité dans cette page, mais à qui je dois beaucoup pour le soutien et la motivation lors de la rédaction du manuscrit. Merci à toi qui a su être là et me pousser à travailler, malgré tous les *makhassak!* Tu étais patient et tu as trouvé les bons mots pour m'encourager.

Pour finir, un grand merci à ma famille qui m'a appris que tous les problèmes dans la vie ne se résument pas à de simples équations et ne se résolvent pas par des algorithmes. Merci de m'avoir épaulé tout le long, votre confiance et amour malgré la distance ont fait de moi ce que je suis aujourd'hui. Vos questions bien qu'angoissantes en période de doute "quand est ce que tu soutiens cette thèse? C'est pas fini encore?..." m'ont permis de ne jamais dévier de mon objectif final.

A ma famille,

Abstract

The replication program of about one half of mammalian genomes is characterized by megabase-sized replication U/N-domains. These domains are bordered by master replication origins (MaOris) corresponding to $\sim$ 200 kb regions of open chromatin favorable for early initiation of replication and transcription. Thanks to recent high-throughput chromosome conformation capture technologies (Hi-C), 3D co-localisation frequency matrices between all genome loci are now experimentally determined. It appeared that U/N-domains were related to the organization of the genome into structural units. In this thesis, we performed a combined analysis of human Hi-C data and replication timing profiles to further explore the structure/function relationships in the nucleus. This led us to describe novel large (>3 Mb) replication timing split-U-domains also bordered by MaOris, to demonstrate that the replication wave initiated at MaOris only depends of the time during S phase and to show that chromatin folding is compatible with a 3D equilibrium in early-replicating euchromatin regions turning to a 2D equilibrium in the late-replicating heterochromatin regions associated to nuclear lamina. Representing Hi-C co-localisation matrices as structural networks and deploying graph theoretical tools, we also demonstrated that MaOris are long-range interconnected hubs in the structural network, central to the 3D organization of the genome and we developed a novel multi-scale methodology based on graph wavelets to objectively delineate structural units from Hi-C data. This work allows us to discuss the relationship between replication domains and structural units across different human cell lines.

Résumé

Le programme de réplication d'environ la moitié du génome des mammifères est caractérisé par des U/N-domaines de réplication de l'ordre du méga-base en taille. Ces domaines sont bordés par des origines de réplication maitresses (MaOris) correspondantes à des régions ($\sim$ 200 kb) de chromatine ouverte favorables à l'initiation précoce de la réplication et de la transcription. Grâce au développement récent de technologies à haut débit de capture de conformations des chromosomes (Hi-C), des matrices de fréquences de co-localisation 3D entre toutes les paires de loci sont deśormais déterminées expérimentalement. Il est apparu que les U/N-domaines sont reliés à l'organisation du génome en unités structurelles. Dans cette thèse, nous avons effectué une analyse combinée de données de Hi-C de lignées cellulaires humaines et de profils de temps de réplication pour explorer davantage les relations structure/fonction dans le noyau. Cela nous a conduit à décrire de nouvaux domaines de réplication de grande tailles (>3 Mb). Ces split-U-domaines aussi bordés par des MaOris, dèmontrer que la vague de réplication initiée aux MaOris ne dépend que du temps pendant la phase S et de montrer que le repliement de la chromatine est compatible avec un modèle d'équilibre 3D pour les régions euchromatiniennes à réplication précoces et un modèle d'équilibre 2D pour les régions heterochromatiniennes à réplication tardives associées à la lamina nucléaire. En représentant les matrices de co-localisation issues du Hi-Cen réseaux d'interactions structurelles et en déployant des outils de la théorie des graphes, nous avons aussi démontré que les MaOris sont des hubs interconnectés à longue portée dans le réseau structurel, fondamentaux pour l'organisation 3D du génome et nous avons développé une méthodologie

multi-échelle basée sur les ondelettes sur graphes pour délimiter objectivement des unités structurelles à partir des données Hi-C. Ce travail nous permet de discuter de la relation entre les domaines de réplication et les unités structurelles entre les différentes lignées cellulaires humaines.

Contents

Part I

Introduction and Background

Chapter 1

Introduction

The nucleus of human cells contains over six billion base pairs of DNA (lined up ~ 2 meters of DNA). Fitting that much DNA in a tiny cell nucleus (~ 5 micrometers) requires a high level of compaction. While complex and showing strong cell to cell variabilities, genome organisation is not random. Our current understanding of the nuclear architecture is rather precise at the two extremes of resolution. On the chromosomal level, during interphase, every chromosomes occupy a different nuclear region known as chromosome territories [1], while the metaphase chromosomes are more condensed forming the classic chromosome structure seen in karyotypes. On the scale of hundreds of base pairs, the DNA double helix composed by nucleotides, wraps around histone proteins forming nucleosomes connected by linkers leading to the "beads on a string" model [2]. The X-ray crystal structure of the nucleosome has been determined showing how the histone proteins are assembled and how the DNA is organised around them [3]. Between the atomic resolution and the chromosome resolution, our understanding of chromosome folding remains sparse even though many models have been proposed to describe chromatin folding into the so-called 30 nm fiber and then into higher-order folding (See Chapter 2, Section 2.3). Despite this huge resolution gap, it is increasingly recognised that the three-dimensional (3D) dynamical architecture of the genome plays an important role in the regulation of nuclear functions, including transcription and replication [1, 4–12]. In fact, over the last two decades, many studies have assessed the spatial proximity and nuclear organisation of specific genomic loci, using microscopic techniques such as fluorescent in situ hybridisation (FISH) or molecular biology techniques namely chromosome conformation capture (3C) demonstrating a correlation between chromatin topology and gene activity without understanding the structure/function causality effect [13]. Recent developments in genomic technologies have led to genome wide contact maps. Thanks to Hi-C technique, a high-throughput 3C based method [14], structural interaction frequencies between all genome pairs of loci is now achievable. In other words, Hi-C technique is particularly informative in deciphering genome-wide chromatin contacts on the mega-base or even tens of kilobase scales offering the unprecedented opportunity to close the resolution gap in understanding genome architecture. Pioneering Hi-C works using Principal Component Analysis (PCA) [15] allowed genome segmentation into two compartments (A/B compartments), one of gene-rich euchromatin and the other one of silent gene-poor heterochromatin [14]. Recently higher-resolution Hi-C interaction maps have revealed that the human genome is organised into distincts units, the so-called *Topologically Associated Domains* (TADs), where genomic interactions are strong within a domain and depleted between domains [16]. The organisation of these TADs relies on specific DNA contacts: DNA-DNA loops or DNA fragments tethered to nuclear features such as the nuclear lamina [17]. These processes are mediated by various architectural proteins that are important factors in the creation of these contacts. CTCF plays a key role in nuclear organisation [18]. It is able to bring together pieces of DNA, possibly through homodimerisation and appears to be a major player in chromatin structure formation in general. It also acts as an insulator between TADs, and is considered as a landmark feature of organising histones. TADs were found to be conserved in *Drosophila* [19, 20] and mammalian [16, 21]

genomes but they were less clearly apparent in *Arabidopsis thaliana* [22, 23], *Plasmodium falciparum* [24] and yeasts [25, 26] genomes.

Hi-C data have also been used to build 3D models of global or local chromatin folding allowing a visual representation of chromatin structure. Several approaches have been adopted such as the restraint-based models based on the assumption that interaction frequencies (measured by Hi-C) are inversely related to spatial distances [25, 27]. The exact relationship between distance and interaction frequency can potentially be calibrated measuring the distances of known loci using microscopy techniques (such as DNA-FISH). Additional constraints can be added by measuring the size and the geometry of the nucleus, or integrating knowledge about the physical properties of the DNA polymer. Once pairwise distances are estimated, the modelling problem can be turned into an optimisation problem by defining a score function. Standard optimisation packages based on non-linear programming can then be run to generate the lowest-scoring models. A different approach is to embrace the fact that for each loci, various dominant conformations are likely to co-exist in the compendium of cells. The Hi-C data reflect the frequency to which chromatin assumes these particular conformations. A prime example of this is based on the Integrated Modelling Platform (IMP) [28]. A large set of starting coordinates is initialised randomly, and potential structures are generated iteratively using simulated annealing. The resulting structures are subsequently clustered by spatial similarity with mirroring structures merged together. This process is able to effectively capture the variability of conformations by detecting local minima, while converging to dominant structures with high frequency given the right annealing parameters. Also linking Hi-C interaction frequencies to genomic distance, the authors in [29] have proposed an iterative algorithm to reassemble genomes. The algorithm generates a 1D genome structure that is consistent with the 3D contact data by applying virtual rearrangements to an initial ordering of DNA fragments. From the assumption that the interactions between chromosomes has enough information to distinguish between DNA segments coming from different species, the authors have generalised their method to allow the de novo assembly and scaffolding of the various genomes present in a mix of species without prior knowledge of their genome sequences [30].

Another line of research is based on polymer physics. The models estimate local characteristics of DNA polymer fibre in order to recapitulate higher order structures such as DNA supercoiling or chromosomes territories [31, 32]. The two most popular models are the *Equilibrium Globule* [33] and the *Fractal Globule* [34] models. Experiments show that the interaction frequency steeply decreases with genomic distance until $\sim$ 10 Mb followed by a more gradual decrease. This characteristic is in line with equilibrium states of a homopolymer which exerts a local repulsive force between the chains to avoid extensive clumps or knots. External confinement of polymer chains by the nuclear membrane results in a formation called equilibrium globule, which is consistent with chromatin characteristics elucidated from FISH data [33]. Grosberg *et al.* [34] proposed an alternative polymer model, the *Fractal Globule*, that accounts for the condensation of the chromatin fiber into a long-lived, non-equilibrium unknotted conformation. The model has been found to closely match the contact probabilities of the original Hi-C data [14]: the exponent -1.08 of the power-law decrease of the contact probability as a function of the genomic distance of the Hi-C data lies very near the exponent -1 predicted by the model. In the fractal globule model, nearby regions tend to form a hierarchy of clusters resulting

in topological domains on the small-scale, and chromosome territories on the large scale. The knot-free structure allows for dense packing while keeping the chromatin accessible to various proteins. However, as the Hi-C data generally stem from multiple cells the resulting models reflect the dominant "mean" structures, averaging out the more variable features of the genome organisation.

This thesis takes advantage of the availability of the Hi-C data to follow on the research programme of the host team and others that aim at understanding the large-scale organisation of mammalian genomes and, in particular, the role of chromatin structure and dynamics in the regulation of replication and transcription. Previous analyses of the replication timing profiles have shown that the genome can be segmented into megabase-sized domains that replicate relatively synchronously during S-phase called constant timing regions (CTRs) [35–37]. Early replicating CTRs are generally transcriptionally active gene-rich regions. Whereas, late CTRs have low gene density and are enriched in repressive epigenetic marks. Late CTRs are also located in lamin-associated domains. These early and late replication timing plateaus have been shown to be separated by rather steep timing transition regions (TTRs) of size ranging from 0.1 to 0.6 Mb and presumed to be replicated unidirectionally by a single fork coming from the early domain [36]. Hi-C data have also been compared to genome-wide mean replication timing (MRT) data of related cell lines [37, 38] and consistent with the original Hi-C analysis of the human genome [14], some dichotomic picture has been proposed where early and late replication loci occur in the separated A/B compartments of open and closed chromatin respectively. Previous team work on the analysis of DNA compositional asymmetry profiles led to the segmentation of the human genome into replication domains exhibiting a N-shaped nucleotide compositional skew profile called N-domains and overall bordered by $\sim$ 1000 putative replication origins [39–41]. This segmentation was further confirmed with the emergence of replication timing data (the timing at which each locus is replicated during the S-phase). Interestingly, when using a wavelet-based multi-scale analysis of MRT data [42, 43] for seven human cell types, about half of the genome was shown to be divided in megabase-sized U-shaped MRT domains [44]. Significant overlap was observed between the MRT U-domains of different cell types and also with germline N-domains [39–41, 45–47]. From the demonstration that the average fork polarity is directly reflected by both the compositional skew and the derivative of the MRT profile [44, 48, 49], it has been argued that the fact that the MRT derivative displays a N-shape in MRT U-domains sustains the existence of megabase-sized gradients of replication fork polarity in somatic and germline human cells [39, 43, 44]. When investigating the large scale organisation of human genes inside these replication domains, a remarkable organisation has been revealed [41, 46, 50]; in particular highly expressed genes in a given cell type are over-represented close to the corresponding U/N-domain borders [50]. When further mapping experimental and numerical chromatin mark data in these domains, the "master" replication origins at U/N-domain borders were shown to be specified by a region ($\sim$ 200 kb) of open and transcriptionally active chromatin significantly enriched in insulator DNA-binding proteins [44, 51]. More recently, availability of a large number of epigenetic mark profiles, allowed to characterise more precisely the link between the mean replication timing and the epigenetic status. Using principal component analysis [15] and classical clustering [52] on a set of 13 epigenetic marks at a spatial resolution of 100 kb (corresponding to the MRT profile resolution), the apparent complexity of the data set was reduced to four major groups of chromatin marks that share fea-

tures [53, 54]. These states have specific replication timing: an early transcriptionally active euchromatin state (C1), a mid-early repressive type of chromatin (C2) associated with polycomb complexes, and two late replicating states; a silent state (C3) not enriched in any available marks, and a gene-poor HP1-associated heterochromatin state (C4). When mapping these chromatin states inside the megabase-sized U/N-domains, it appears that the associated replication-fork polarity gradient corresponds to a directional path across three chromatin states, from C1 at the U/N-domain borders followed by C2, C3 and C4 at the centers [53, 54]. A recent high-resolution 4C study dedicated to the analysis of the interaction of some selected U/N-domain borders with the rest of the human genome [55] confirmed that these early-initiation zones play a major role in the chromatin tertiary structure. Moreover, the additional comparative analysis of replication U/N-domains and Hi-C data revealed that these functional domains are intimately linked to the genome 3D architecture [44]. In this sense, the relationships between the distribution of chromatin states, the spatio-temporal replication programme and the 3D genome architecture could be the key to understanding the chromatin complex organisation and its role in regulating nuclear functions. Our aim in this thesis is to objectively quantify the importance of the "master" replication origins (MaOris) at U/N-domain borders and to discuss the existence of structural domains as a counterpart to the replication domains.

Graphs [56, 57] have become extremely useful as a representation of a wide variety of complex systems in social sciences [58, 59], biology [60–62], computer sciences [62, 63], engineering and many other area of fundamental and applied sciences [64–66]. A graph consists in a set of nodes connected to each other by edges, for example a group of people can form a set of nodes for a social graph where acquaintances are the edges. Graph theory provides centrality measures to identify and quantify the nodes that occupy critical positions in a network. Another major property of graphs is their community structure, a community in a graph is a group of nodes highly connected in between them and less connected to the rest of the graph. Here, we explore the use of graph theory on the *chromatin interaction network* where nodes are the DNA loci and edges correspond to Hi-C interactions. On the one hand, using centrality measures, we quantify that MaOris are indeed key organisational features of the genome. On the other hand, we detect structural communities in the chromatin interaction network.

This manuscript is organised in seven chapters. After this short chapter of introduction, we recapitulate in Chapter 2 previous results of the host team in their biological context. In Chapter 3, we highlight key results from graph theory that are used in this thesis. The results part starts with Chapter 4, where we extend MRT profiles analysis to regions depleted in MaOris which leads to the identification of split-U-domains (bordered by two MaOris far from each other). Split-U-domains complement the U-domains, their borders present similar characteristics in terms of epigenetic marks; nevertheless split-U-domains specifically contain a central late replicating heterochromatin region. Study of the MRT derivative in those domains leads us to propose a *universal cascade model* for the replication programme that only depends of the timing during the S-phase. Taking into account the long range interactions, we discuss in Chapter 5 the "fractal globule" and the "equilibrium" model interpretations of Hi-C data. In fact, the chromatin fiber is not a homopolymer but a heteropolymer that accounts for the spatial compartmentalisation of the epigenome in the four prevalent chromatin states likely corresponding to different structural and mechanical properties. We show that only the heterochromatin

regions of differentiated human cell types are compatible with the fractal globule model while the euchromatin regions are rather compatible with the equilibrium globule model. We propose a unified understanding of these results in the framework of the equilibrium globule model where the differentiated cell types heterochromatin regions dynamics correspond to a transition to a confinement in 2D. In Chapter 5, we use centrality measures to show that MaOris are vertices of high centralities in the chromatin interaction network. In K562, they are found to form long-distance interconnected hubs of DNA interactions. We then test how this property extend to other cell types. We show that cell type specificities are reflected in the computation of centrality measures. Moreover depending on their status (peaks or asymmetric borders) MaOris do not play the same role. In Chapter 6, we reformulate the question of finding topological domains in Hi-C data into a question of community detection in the Hi-C interaction network. Using a multi-scale community detection method based on graph wavelets, we delineate at each scale, a partition of the nodes into structural communities. Importantly, the observed communities are genomic intervals, even though the community detection method does not assume that the communities should group adjacent loci of the genome. Analysis of those communities across scales provides a hierarchical organisation of the genome allowing to address the relationships between structure and function in a cell type and structural domains conservation between different cell types. Finally, Chapter 7 provides a summary of the main results of this thesis with concluding remarks. Altogether, this thesis explores the relationships between structure, function and epigenetics as a useful tool to fully apprehend genome regulation, in particular during differentiation. It reports already published results (See the communication list page 185) and some ongoing work.

Chapter 2

From DNA spatio-temporal replication programme to chromatin conformation capture data

In this chapter, we review previous analyses of mean replication timing profiles and DNA compositional asymmetry profiles that demonstrated the existence of Mb-sized replication domains with "master" replication initiation zones at the borders and late replicating zones in the center. The replication domain organisation was shown to present a strong relationship with the local chromatin state. Indeed, the study of tens of chromatin mark profiles from different cell types led to reducing the chromatin complexity to 4 prevalent chromatin states. The spatial repartition of those chromatin states along the genome is structured especially along the replication domains where there exists a gradient of chromatin from border to center. Besides this 1D description along the chromosomes, DNA replication can also be analysed in relation to the 3D organisation in the nucleus. Recently, developments in chromatin conformation technologies have provided new insights to the structural organisation of the nucleus at meso-scales. Interestingly, structural units seem to co-localise with replication domains,providing a unifying view of chromatin structure, nuclear functions and organisation. In this thesis, we try to address some issues raised by the structure/function relationships.

2.1 Organisation of the human genome into replication domains

DNA replication is a fundamental process of cell life. Dysregulation of DNA replication can challenge genome stability and lead to mutations, cancer and other genetic diseases [67, 68]. Yet this process remains not fully understood: mapping of DNA replication origins and the regulation of fork progression and termination are still open questions in molecular biology. In this section, we present the analysis of replication timing profiles and nucleotide compositional skew revealing the existence of chromosomal units for replication called U/N-domains. These domains are characterized by a U-shaped replication timing and N-shaped compositional skew and correspond to large scale gradient of the replication fork polarity [41, 43, 44, 69]. We finally present a "cascade model" that was proposed for the replication origin activation along replication U/N-domains [44, 69, 70].

2.1.1 Strand compositional asymmetry as the signature of genomic activity

According to the second parity rule [71, 72], under no strand-bias conditions, each genomic DNA (Fig. 2.1) strand should present equimolarities [73, 74] of A and T and of G and C. Actually, during genome evolution, mutations do not occur at random as illustrated by the diversity of the nucleotide substitution rate values [75–78]. This non-randomness is considered as a by-product of the various DNA mutation and repair processes that can affect each of the two DNA strands differently. Deviations from intrastrand equimolarities have been extensively studied and the observed skews have been mainly attributed to asymmetries intrinsic to the transcription (Fig. 2.2) or to the replication (Fig. 2.3) processes. As illustrated in Figure 2.3, during replication one strand is replicated continuously (leading strand), while the other strand is replicated in discrete steps towards the origin (lagging strand). Hence, mutational effects can affect the leading and lagging strands differently with different repair mechanisms leading to compositional asymmetry between the two DNA strands reflecting the directionality of the replication fork progression.

2.1.1.1 DNA double helix structure and Chargaff compositional symmetry rules

Deoxyribonucleic acid (DNA) encodes the genetic information used in the development and functioning of free living organisms. DNA molecules consist of two polynucleotides chains called DNA strands, held together by hydrogen bonds and resulting in a double helix structure (Fig. 2.1) [79, 80]. The nucleotide chain is composed of a backbone of alternating sugars and phosphates, and the nucleotides differ by their base: Thymine (T), Adenine (A), Guanine (G) or Cytosine (C). A fundamental property of DNA double helix is *base pairing* [79]: a guanine on one strand is always paired to a cytosine on the complementary strand by three hydrogen bonds and a thymine on one strand is always paired to an adenine on the complementary strand by two hydrogen bonds. The polynucleotide chain has also an orientation. A sugar in the backbone has its 5' phosphate group attached to the 3' hydroxyl group of the next sugar: this gives the polarity to the DNA strands. In the DNA double-helix, the two strands have opposite polarities *i.e.* they run anti-parallel to each other. Conventionally, the nucleotide sequence of one DNA

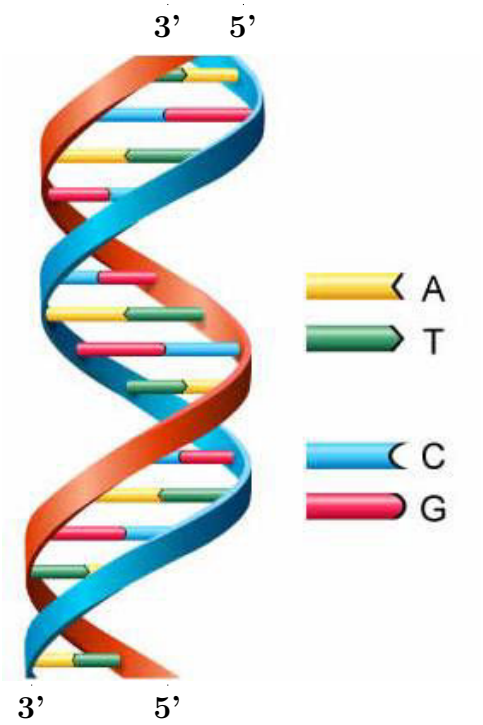


Figure 2.1. DNA double helix structure. Two helical nucleotide chains twisted around each other like a right handed spiral staircase, where the Adenine (A) (resp. Cytosine (C)) on one chain is always paired to a Thymine (T) (resp. Guanine (G)) on the other. The DNA helix has a pitch of 34 Å and a radius of 10 Å. The largest human chromosome is chromosome number 1 and consists of approximately 220 million base pairs and is 85 mm long.

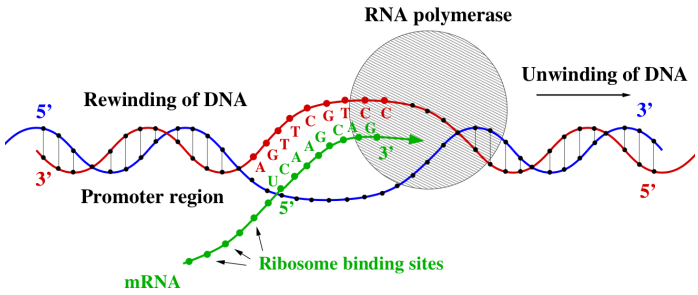


Figure 2.2. Schematic view of transcription. RNA polymerase traverses the template strand and uses base pairing complementarity with the DNA template to create an RNA copy.

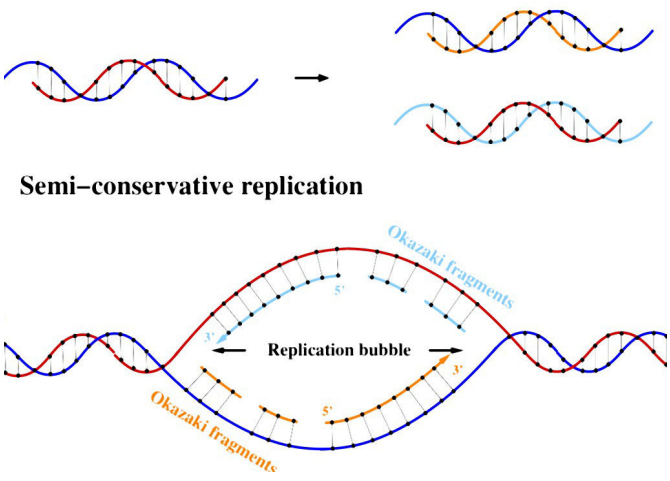


Figure 2.3. Schematic view of replication. The double helix is unwound and each strand acts as a template for a new strand. Bases are matched to synthesize the new partner strands.

strand is read in 5' → 3' direction. The polarity of the DNA strand has great biological importance. For instance, the DNA polymerase always synthesises the newly replicated strand in the 5' → 3' direction (Fig. 2.3).

DNA strands complementarity (base-pairing) imposes that the number of adenine and thymine (resp. guanine and cytosine) are identical when considering the two DNA strands, which is also known as Chargaff first parity rule [71]. Now supposing a model of mutations where we have the same mutational and repair mechanisms on the two strands (no strand bias conditions) it results that the substitutional rates are equal on the two strands and in turn that the asymptotic compositions of the two DNA strands are on average equal [73, 74, 81]. For example, noting (1) and (2) the two DNA strands, having the same substitutional rate thus implies that $[\mathrm{T}]_{(1)} = [\mathrm{T}]_{(2)}$. From the first parity rule $[\mathrm{T}]_{(2)} = [\mathrm{A}]_{(1)}$, which result in $[\mathrm{T}]_{(1)} = [\mathrm{A}]_{(1)}$ (parity rule 2 for A/T composition) [71, 72]. In the same manner, parity rule 2 for G and C gives $[\mathrm{G}]_{(1)} = [\mathrm{C}]_{(1)}$, under no strand bias conditions. This parity rule 2 is valid when considering long genomic portions (complete chromosome) (data not shown). In order to analyse possible deviations to parity rule 2 along small regions, compositional skews S_{TA} and S_{GC} have been defined as:

$$S_{\mathrm{TA}} = \frac{n_{\mathrm{T}} - n_{\mathrm{A}}}{n_{\mathrm{T}} + n_{\mathrm{A}}}, \qquad S_{\mathrm{GC}} = \frac{n_{\mathrm{G}} - n_{\mathrm{C}}}{n_{\mathrm{G}} + n_{\mathrm{C}}}, \tag{2.1}$$

where n_{A}, n_{C}, n_{G} and n_{T} are respectively the numbers of A, C, G and T in the analysed sequence window. Because of the observed correlation between the TA and GC skews [39, 40, 82], the total skew S defined as:

$$S = S_{\mathrm{TA}} + S_{\mathrm{GC}}, \tag{2.2}$$

was mainly considered. From the skew profiles $S_{\mathrm{TA}}(n)$, $S_{\mathrm{GC}}(n)$ and $S(n)$, obtained along the sequences in 1 kb windows, where n is the position (in kb units) from the origin, the skew DNA walks are computed as the cumulative skew profiles:

$$\Sigma_{\mathrm{TA}}(n) = \sum_{j=1}^{n} S_{\mathrm{TA}}(j), \qquad \Sigma_{\mathrm{GC}}(n) = \sum_{j=1}^{n} S_{\mathrm{GC}}(j), \tag{2.3}$$

and

$$\Sigma(n) = \sum_{j=1}^{n} S(j). \tag{2.4}$$

If locally the numbers of As and Ts (or of Gs and Cs) are equal, we have $S_{TA} = 0$ ($S_{GC} = 0$) and $S = 0$. However, nuclear processes break the no strand bias hypothesis leading to asymmetries in the nucleotides composition ($S_{TA} \neq 0$, $S_{GC} \neq 0$, $S \neq 0$). Note that, in pluricellular organisms, mutations responsible for the observed biases occur in germline cells only.

2.1.1.2 Transcription-associated strand compositional asymmetry

Transcription (Fig. 2.2) is a biological process important for the usage of the genetic information, it is the first step of gene expression. It consists in copying a particular segment of DNA into RNA by the RNA polymerase enzyme. RNA is similar to DNA with the replacement of thymines T by uracils U and usage of alternative sugar in the

backbone. Transcription proceeds in the following steps: RNA polymerase binds to DNA promoter and creates a transcription bubble, by separating the two DNA strands. The RNA polymerase synthesises the messenger RNA in the 5' → 3' direction by base-pairing using the complementary strand as a template. The RNA polymerase therefore progresses in the 3' → 5' direction on the template strand [83]. RNA sugar-phosphate backbone forms with the assistance from DNA polymerase. Hydrogen bonds of the untwisted RNA-DNA helix break, freeing the newly synthesised RNA strand. The coding strand and the transcribed strand undergo different mutational and repair events that generate strand asymmetry [84–89]. During transcription, the coding strand is transiently in a single-stranded state (ssDNA), while the transcribed strand is protected by the RNA polymerase. The coding strand is possibly more exposed to mutagenic lesions [84, 88–90] than the transcribed strand. It has also been proposed that repair mechanisms [84, 87, 89, 91] could generate strand asymmetries. A mechanism known as transcription-coupled repair (TCR) [83], associated with the passage of RNA polymerase, preferentially repairs towards the coding strand [92].

Asymmetries of substitution rates coupled to transcription have been first observed in prokaryotes [86, 93, 94]. In the human genome, excess of T was observed in a set of gene introns [95] and some large-scale asymmetry was observed in human sequences but they were attributed to replication [96]. A comparative analysis of mammalian sequences demonstrated a transcription-coupled excess of G+T over A+C in the coding strand [87]. In contrast to the substitution biases observed in bacteria presenting an excess of (C→T) transitions, these asymmetries are characterised by an excess of purine (A→G) transitions relatively to pyrimidine (T→C) transitions. These might be a by-product of the transcription-coupled repair mechanism acting on uncorrected substitution errors during replication [92]. Genome-scale analyses of gene sequences have definitely established the existence of transcription-coupled nucleotide biases in human and other eukaryotes [82, 97]. The comparison of the overall S_{TA} and S_{GC} skew profiles (Equation (2.1)) [82, 97] in transcribed regions to those in the neighbouring intergenic sequences showed at the 5' gene extremities, a sharp transition of both skews from about zero values in the intergenic regions to finite positive values in transcribed regions ranging between 4 and 6% for S_{TA} and between 3 and 5% for S_{GC}. At the gene 3'- extremities, the TA and GC skews also exhibit transitions from significantly large values in transcribed regions to very small values in untranscribed regions. However, in comparison to the steep transitions observed at 5'- ends, the 3'- end profiles present a slightly smoother transition pattern extending over ~ 5 kb and including regions downstream of the 3'- end likely reflecting the fact that transcription continues to some extent downstream of the polyadenylation site [82, 97] .

☞ A gene is defined as *sense* (+) *gene* if its coding sequence is on the reference strand, and as *antisense* (−) *gene* if its coding sequence is on the complementary strand. For a sense (+) gene the reference strand is the coding strand and the complementary strand is the transcribed strand. For an antisense (−) gene we have the opposite situation. Therefore the *gene orientation* (±) is a crucial parameter of transcription-associated strand asymmetry. Another crucial parameter is the *transcription rate*, which reflects how many times a gene has been transcribed during a cell cycle. Indeed, in mouse it was shown that transcription-associated compositional asymmetry was specifically correlated to the transcription rate in the germline [42].

2.1.1.3 Replication-coupled strand compositional asymmetry

In their 1953 announcement of the DNA double helix structure, Watson and Crick [79] stated "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material". Indeed when separated, each strand of the DNA can be a template to assemble a new complementary strand, thus producing two identical DNA molecules. This is how replication proceeds: when a cell divides, the genome of the mother cell is duplicated and one copy is transmitted to each of the two daughter cells. The DNA replication is semi-conservative (Fig. 2.3): each daughter cell inherits a DNA strand of the mother cell, which has served as a template for the DNA polymerase to synthesise the complementary strand [83]. At a replication origin (Fig. 2.4 (a)), the DNA double helix is opened, and two divergent replication forks replicate the DNA one on each side of the replication origin, creating a "replication bubble". Each replication fork is composed of two DNA polymerases that replicate separately the two parental strands. The DNA polymerases always synthesise the new strand in the $5' \rightarrow 3'$ direction progressing on the parental strand in the $3' \rightarrow 5'$ direction. Due to the anti-parallel polarities of the parental strands, one strand is synthesised continuously (the *leading strand*) and the other one discontinuously (the *lagging strand*). The parental strand oriented in the $3' \rightarrow 5'$ direction as seen by the replication fork (the leading strand template) is replicated continuously by the DNA polymerase, producing continuously the synthesised leading strand in the $5' \rightarrow 3'$ direction. On the parental strand oriented in the $5' \rightarrow 3'$ direction as seen by the replication fork (the lagging strand template), the DNA polymerase synthesises discontinuously small nascent strands, called Okazaki fragments, in the $5' \rightarrow 3'$ direction, progressing in the $3' \rightarrow 5'$ direction on the parental strand, opposite to the global replication fork movement. Replication could induce strand asymmetries by several means [84, 85, 98–100]. For instance the leading strand, when it serves as a template for the lagging synthesis of the complementary strand, is transiently in ssDNA, where it could be more exposed to mutagenic lesions [84, 85].

We next briefly review the studies about replication-associated skews in prokaryotes before presenting the results for eukaryotes.

The analysis of replication-associated strand asymmetries was first performed in bacterial genomes [74, 85, 98, 101, 102]. In bacteria, the spatio-temporal replication programme is particularly simple (Fig. 2.4). The replication origin is defined by a consensus sequence, therefore, replication always initiates at the same genomic locus (O in Fig. 2.4 (b)), two divergent forks then replicate the DNA until they meet at the replication terminus (T in Fig. 2.4 (b)). As illustrated in Figure 2.4 (b and c) for the GC skew, the GC and TA skews abruptly switch sign (over few kb) from negative to positive values at the replication origin and in the opposite direction from positive to negative values at the replication terminus. This step-like profile is characteristic of the replicon model [103]. In *Bacillus subtilis*, as in most bacteria, the leading (resp. lagging) strand (Fig. 2.4(a)) is generally richer (resp. poorer) in G than in C (Fig. 2.4(b)), and to a lesser extent in T than in A. This typical pattern is particularly clear when plotting the cumulated skews Σ_{GC} (Fig. 2.4(c)) and Σ_{TA} (Equation (2.3)); both present decreasing (or increasing) profiles in regions situated 5' (or 3') to the origin, displaying a characteristic $\vee$-shape pointing to the replication origin position (similarly a characteristic $\wedge$-shape is observed

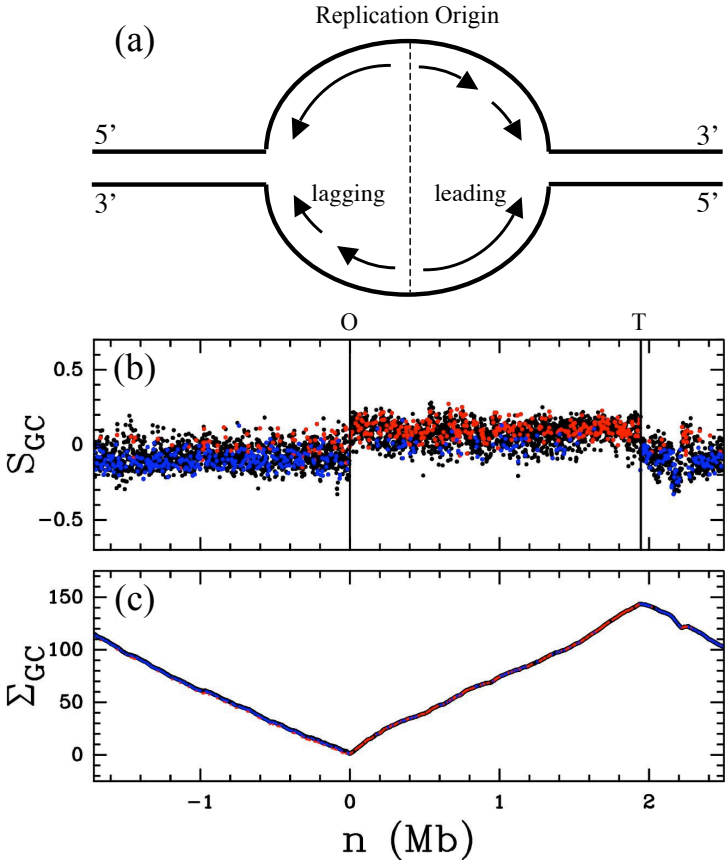


Figure 2.4. The replicon model. (a) Schematic representation of the divergent bi-directional progression of the two replication forks from the replication origin (Fig. 2.3). (b) S_{GC} calculated in 1 kb windows along the genomic sequence of *Bacillus subtilis*. (c) Cumulated skew Σ_{GC} (Equation (2.3)). The vertical lines correspond respectively to the replication origin (O) and termination (T) positions. In (b) and (c), red (resp. blue) points correspond to (+) (resp. (-)) genes that have the same (opposite) orientation than the sequence.

at the terminus position). The research of $\vee$ patterns in the cumulated skews has been extensively used as a strategy to detect the position of the (unique) replication origin in (generally circular) bacterial genomes [85, 98, 101, 102]. It is noteworthy that genes present a remarkable organisation around the replication origin of *Bacillus subtilis*. As shown in Figures 2.4(b) and 2.4(c), we observe that most of the (+) (resp. (-)) genes are preferentially on the right (resp. left) of the replication origin. This has suggested that the replication forks progression is co-oriented with transcription, as to minimize the risk of frontal collision between DNA and RNA polymerases [104–107].

The spatio-temporal replication programme in eukaryotes is much more complex. In eukaryotic cells, the cell cycle is divided into 4 consecutive phases. Cell division occurs during mitosis (M) phase. The rest of the cell cycle, called interphase is subdivided in G1, S and G2 phases. Cells grow continuously during the interphase, doubling in size and preparing for next division. Yet, DNA duplication occurs only during S phase. Replication is initiated at a number of replication origins ($\sim$ 50 000 in human genome [69]) and propagates until two converging forks collide at a terminus of replication (Fig. 2.3), initiation sites fire at different times during S-phase. The genomic positions and firing times of the initiation sites change from one cell cycle to another. During G1 phase the Origin Recognition Complex (ORC) binds to DNA. The binding of ORC is followed by the recruitment of several proteins including the helicase MCM (minichromosome maintenance). These proteins form the pre-replication complex (pre-RC) that constitutes a potential replication origin that may be activated during S-phase. In fact, there are more pre-RC deposited on DNA than actively needed during the S-phase. The subsequent activation of the pre-RC during S-phase leads to the recruitment of DNA polymerase

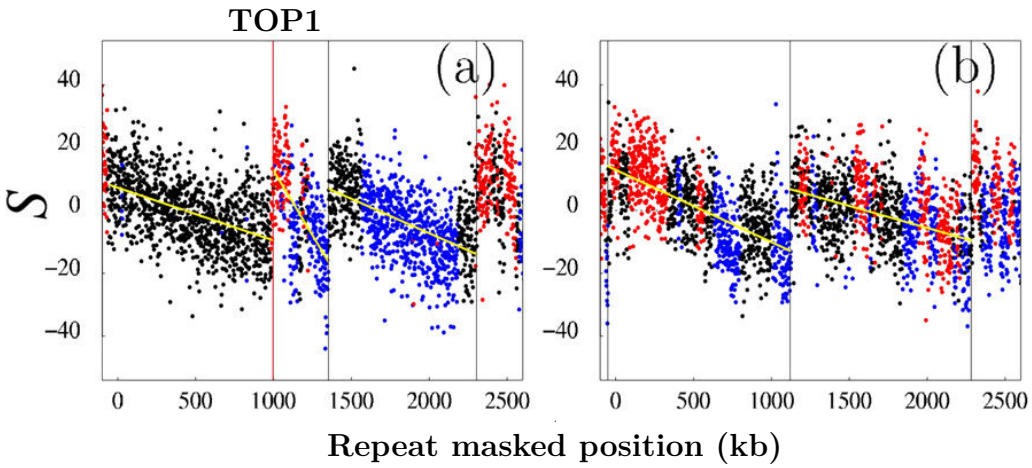


Figure 2.5. Skew S profiles along human genome fragments. (a) Fragment of human chromosome 20 including the TOP1 origin, a known origin (red vertical line). (b) Fragment of human chromosome 4. Vertical lines correspond to selected putative origins; yellow lines are linear fits of the S values between successive putative origins. Black, intergenic regions; red, (+) genes; blue, (-) genes. Note the fully intergenic regions upstream of TOP1. Skew profiles were calculated in 1 kb windows of the repeated masked sequence resulting in a shortening of the sequence length by a factor of ~ 2. Adapted from [40].

and other proteins necessary to DNA synthesis. The activation of different replication origins occurs at diverse moments of the S phase and is not deterministic [69, 108–112]. It was proposed that pre-RC activation could be triggered by the activation of neighbouring replication origins [69, 70]. Also, the activation depends on the neighbouring transcriptional activity and on the local chromatin structure [109–112]. In eukaryotes, the leading and lagging strands are presumably synthesised by two distinct DNA polymerases [113]. Strand asymmetries could result from the different error spectra of the two DNA polymerases [100]. Several studies have failed to show compositional biases related to replication, and analysis of nucleotide substitutions in the region of the *β-globin* replication origin in primates did not support the existence of mutational bias between the leading and the lagging strands [101, 114, 115]. Other studies have led to rather opposite results. For instance, strand asymmetries associated with replication have been observed in the subtelomeric regions of *S. cerevisiae* chromosomes, supporting the existence of replication-coupled asymmetric mutational pressure in this organism [116]. As illustrated in Figure 2.5 (a) for human TOP1 replication origin, it has been argued that most of the 9 human replication origins known at the time of the study correspond to rather sharp (over several kb) transitions from negative to positive S skew values [39, 40]. This is reminiscent of the behaviour observed in Figure 2.4 for *Bacillus subtilis*. According to the gene environment, the amplitude of the jump observed in the skew profiles can be more or less important and its position more or less localised (from a few kb to a few tens of kb). Indeed, it was previously mentioned that transcription generates positive TA and GC skews on the coding strand [41, 46, 82, 97, 117], which explains that larger jumps are observed when genes are on the leading strand so that replication and transcription biases add to each other. However, the observation that the skews in intergenic regions on both sides of the replication origins shift from negative to positive, suggests that there exists mutational pressure associated with replication, contributing to the mean compositional biases [40] (Fig. 2.5 (a)). Note that the value of the skew

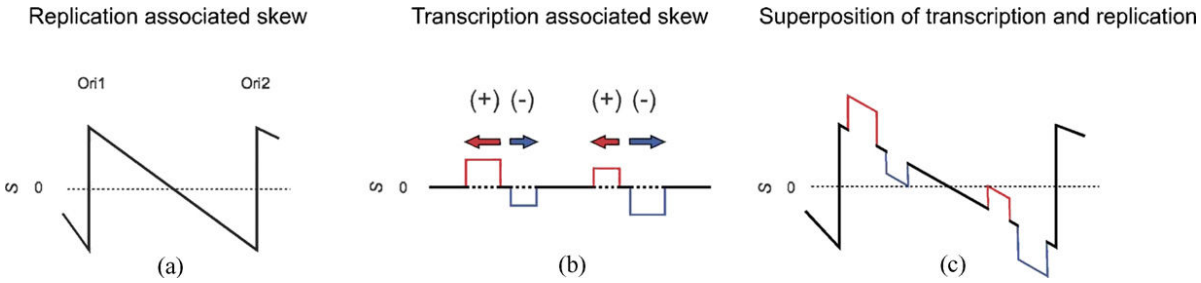


Figure 2.6. Model of "factory roof" skew profiles. (a) N-shaped replication-associated skew profile. (b) Transcription-associated skew profile showing positive square blocks at (+) gene positions and negative square blocks at (-) gene positions. (c) Superimposition of the replication- and transcription-associated skew profiles producing the final factory-roof pattern that defines "N-domains". Adapted from [40].

could vary from one origin to another, possibly reflecting different initiation efficiencies. As shown in Figure 2.5, sharp upward jumps, similar to the ones observed for the known replication origins, exist at many other locations along the human chromosomes. This observation led to the development of an upward jump detection methodology based on the wavelet-transform microscope [39, 40] (See Appendix A, Section A.1). When applying this wavelet-based method to the 22 human autosomes, retaining as putative replication origins upward jumps with an amplitude much larger than the one induced by transcription at the TSS [39, 40], a set of 1012 putative replication origins was detected.

☞ Overall, the detection of sharp upward jumps in the skew profiles with characteristics similar to those of experimentally determined replication origins and with no downward counterpart, provided further support for the existence of replication-associated strand asymmetries in human chromosomes, and led to the identification of numerous putative replication origins active in germline cells.

2.1.2 Skew N-domains as replication domains in the germline

Another striking feature observed along the skew profiles (Fig. 2.5), is the sharp linear decrease of the S profile between the putative replication origins. This pattern has been named "factory roof" profile (Fig. 2.6) [39–41, 46, 118]. Note that there exists extreme variability in the distance between two successive upward jumps, from spacing ∼100-200 kb up to ∼2-3 Mb [39, 40]. But what is important to notice is that some of these segments between two successive skew upward jumps are entirely intergenic (Fig. 2.5(a)), clearly suggesting that the observed peculiar N-shape skew profile is characteristic of a strand bias resulting solely from replication [40, 41, 46, 118].
Hence, it has been suggested that the overall factory roof profile observed in mammalian genomes actually results from the superposition of two patterns (Fig. 2.6) [41, 46]. One decreases steadily from the 5' to the 3' direction and would be attributable to replication in germline cells (Fig. 2.6(a)). To explain this replication-associated N-shaped pattern, a model in which replication first initiates at origins located at the borders of the N-shaped skew profile domain, followed by successive activations of secondary origins as replication progresses toward the center of this domain has been proposed [46]. It has been argued that the linear decline of the skew would reflect a progressive change in the proportion of center- and border-oriented forks that itself would reveal the dynamic pattern with which

secondary initiations would occur within the domain. The other pattern would result from transcription-associated strand asymmetries generating square-like profiles corresponding to (+) and (-) genes (Fig. 2.6(b)). When the two profiles are superimposed, this leads to the factory roof pattern (Fig. 2.6(c)) [41, 46]. Because the typical gene size (∼30 kb) is much smaller than the characteristic distance between two adjacent putative replication origins, these replication domains were qualified as "N-domains" [41, 46] in regards of their overall qualitative N-shape.

☞ The analysis of the skew profiles [46] with a wavelet-based method (Appendix A, Section A.3), led to a database of 663 mega-base sized N-domains* whose skew profile displays a N-shape (Fig. 2.5) bordered by 1062 putative replication origins spanning 33.8% of the genome. The size of these domains ranges from ∼200 kb to 2.8 Mb with a mean length $\bar{L}$ =1.9 Mb. 78% of the 1062 putative replication origins are intergenic and are located near to a gene promoter [46]. When investigating the large scale organisation of human genes with respect to replication, it has been shown that gene orientation and gene expression are not randomly distributed relatively to these domains [41, 50]. Around N-domain borders, genes are abundant and broadly expressed, and their transcription is co-oriented with replication fork progression. These features weaken progressively with the distance from N-domain borders. At the center of N-domains, genes are rare and expressed in few tissues. It was proposed that this specific organisation results from the constraints of accommodating the replication and transcription initiation processes at the chromatin level. Indeed, the mapping of experimental and numerical chromatin mark data in replication N-domains [51], showed that around most of the N-domain borders that replicate early in the S phase, there exist regions of a few hundred kb wide that are hypersensitive to DNase cleavage, that are hypomethylated and that present a significant enrichment in genomic energy barriers that impair nucleosome formation. This suggests that the replication initiation zones at N-domain borders are specified by an open chromatin structure favoured by the DNA sequence.

2.1.3 Experimental characterisation of the human replication programme

For years, progress in elucidating the mechanisms underlying replication initiation and its coupling to transcriptional activity and local chromatin structure has been hampered by the small number (approximately 30) of well-established origins in the human genome and more generally in mammalian genomes. In this section, we review recent experimental developments allowing better description of the replication programme in mammals.

2.1.3.1 Mapping replication origins

One might naturally think that it is easy to map replication origins because it could simply be specified by a sequence motif, allowing the initiator protein of replication to recognise it easily. However, as reviewed in [120], mapping replication origins is challenging in mammalian cells where the origin recognition complex (ORC) does not exhibit sequence specificity (as in budding yeast). Actually, ORCs may bound to different sites in

*The coordinates of the 678 human N-domains for assembly NCBI35/hg17 were obtained from the authors [46] and mapped using LiftOver to hg18 coordinates. We kept only the 663 N-domains that had the same size after conversion.

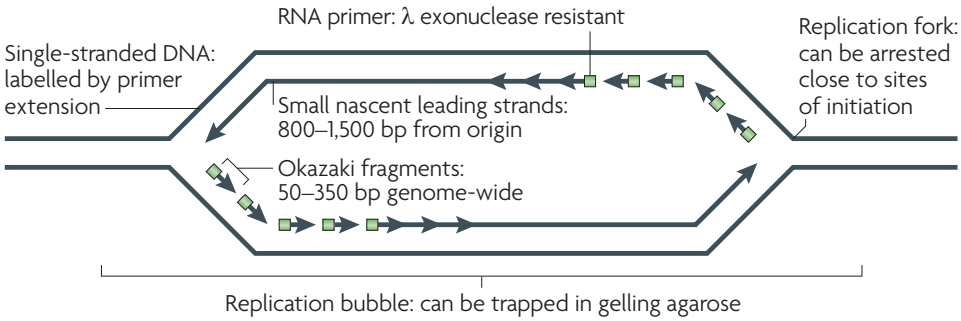


Figure 2.7. How can one find replication origins? Summary of the unique nucleic acid features found near origins of replication. When cells that have been synchronized before the onset of S phase initiate replication in the presence of replication fork inhibitors, replication forks are arrested close to sites of initiation so that any DNA synthesized must be close to origins. The sites where forks are arrested consist of primed templates that can be labelled at the sites of arrest by extension. The leading strands of DNA synthesis quickly become larger than Okazaki fragments and can be isolated as small single-stranded molecules that can be verified to be nascent either by metabolic labelling or by virtue of the fact that nascent strands have small stretches of RNA at their 5' ends that render them resistant to λ exonuclease. Finally, the physical structure of replication origins shortly after initiation is that of a bubble structure, which can be trapped in gelling agarose. Adapted from [119].

different cells, ORCs do not reveal origin efficiency or timing, some ORCs may not function as origins, and initiation may occur remotely from the ORC binding site. However, some hallmarks of replication initiation (summarised in Figure 2.7), have been exploited to map origins as reviewed in [119].

Cells can be synchronised and made to enter S-phase in the presence of a replication fork inhibitor resulting in the accumulation of nascent strands within a few kilobase of early firing origins. Replicated sequences can be detected by their twofold copy number [121, 122] or by labelling the nascent strands with tagged nucleotide precursors either before fork arrest [123–127] or after primer extension of the arrested forks [128, 129]. These methods of *trapping the earliest replicated DNA by replication fork arrest* has been applied at genome scale in budding [126–128] and fission [121, 124, 125] yeasts and recently in *D. melanogaster* [123]. Note that these methods are limited to mapping origins that fire very early in S-phase; actually replication fork arrest triggers a checkpoint response that inhibits origins that would normally fire later in S-phase [130].

Alternatively, in recent studies, *small nascent leading strands* (SNSs) were purified. In fact, there are two leading strands emanating bidirectionally from origins, and their 5' ends define the site of initiation (Fig. 2.7). Denaturing genomic DNA from proliferating cells releases single-stranded nascent DNA, which can be fractionated by size to identify strands closer to the origin. Moreover, nascent strands from asynchronous cells derive from all origins firing throughout S-phase, and their relative abundance should be a direct reflection of their efficiency of firing within a population of cells. This method has been applied to map replication origins genome-wide in different eukaryotic organisms including *Arabidopsis thaliana* [131], *Drosophila* [132], mouse [132, 133] and human [134–139].

Another approach to discover replication origins consist in *trapping replication bubbles* [140–142]. In this technique, restriction fragments containing replication bubbles are retarded in their mobility on a gel by allowing the gel matrix to form through the circular replication bubble. Trapped bubble-containing restriction fragments are recovered from the gel and can be coupled with DNA microarrays or genomic sequencing to identify bubble-containing restriction fragments genome wide.

Altogether, these studies led to identify from few tens of origins (32 in [134]) up to 59 185 in [138], 72 812 in [142] and recently around 100 000 in [139]. Despite some inconsistencies or poor concordance between certain of these studies [119, 143], some general trends have emerged confirming the correlation of origin specification with transcriptional organisation [119, 144]. The replication origins identified so far are strongly associated with annotated promoters and seem to be enriched in transcription factor binding sites [135, 136], in CpG islands [132, 133, 135, 139] and in G-quadruplex [132, 138, 139].

2.1.3.2 Mean Replication Timing

An alternative and more robust way to characterise the replication programme is to estimate the relative order in which DNA segments replicate in a cell population resulting in *replication timing profiles*. A wealth of genome-wide replication timing data is available for several eukaryotic organisms ranging from yeast [145], to plants [146], to *Drosophila* [110], to mouse [37, 38, 147, 148], and to human [35–38, 139, 149–151]. Recent genome-wide replication timing data have been collected in several human cell types [35–38, 139, 149–151], which enables to study changes in the replication programme across differentiation. Current technology is not able to measure the spatio-temporal replication programme in one cell. The characterisation of replication timing is done on a large population of cells (tens of millions). The availability of robust methodologies to analyse the DNA replication programme when averaging over large population of cells somewhat contrasts with the extreme difficulty to experimentally delineate individual replication origins [119, 152, 153]. Figure 2.8 illustrates the two most common classes of methods used to assess replication timing: *isolation of newly replicated DNA* (Fig. 2.8 a) and *DNA content approaches* (Fig. 2.8 b).

- **DNA content approaches:** This method (Fig 2.8 b) relies on analysing the number of copies of each locus to know if it has been replicated or not. For example, when comparing DNA content in S phase and G1, early replicating loci present a S/G1 DNA content ratio close to 2 whereas this ratio is close to 1 for late replicating loci, whatever the length of S phase. This method has been applied first, for human genome at a 1 Mb resolution [35], then at a 100 kb resolution for human chromosome 6 [150], and further extended genome-wide via high throughput sequencing in [36], and more recently in [139] allowing generation of genome-wide allele-specific timing profiles.

- **Isolation of newly synthesised DNA (Repli-seq):** In this method (Fig 2.8 a), a large population of cells, is temporarily cultivated in presence of BrdU which is a modified nucleotide. Cells in S-phase incorporate BrdU in place of thymine in newly synthesised DNA which becomes identifiable. In order to label cells at a specific time slot in the S-phase, the cell culture needs to be tightly synchronised. Synchronisation can be done by either whole culture synchronisation or isolation of

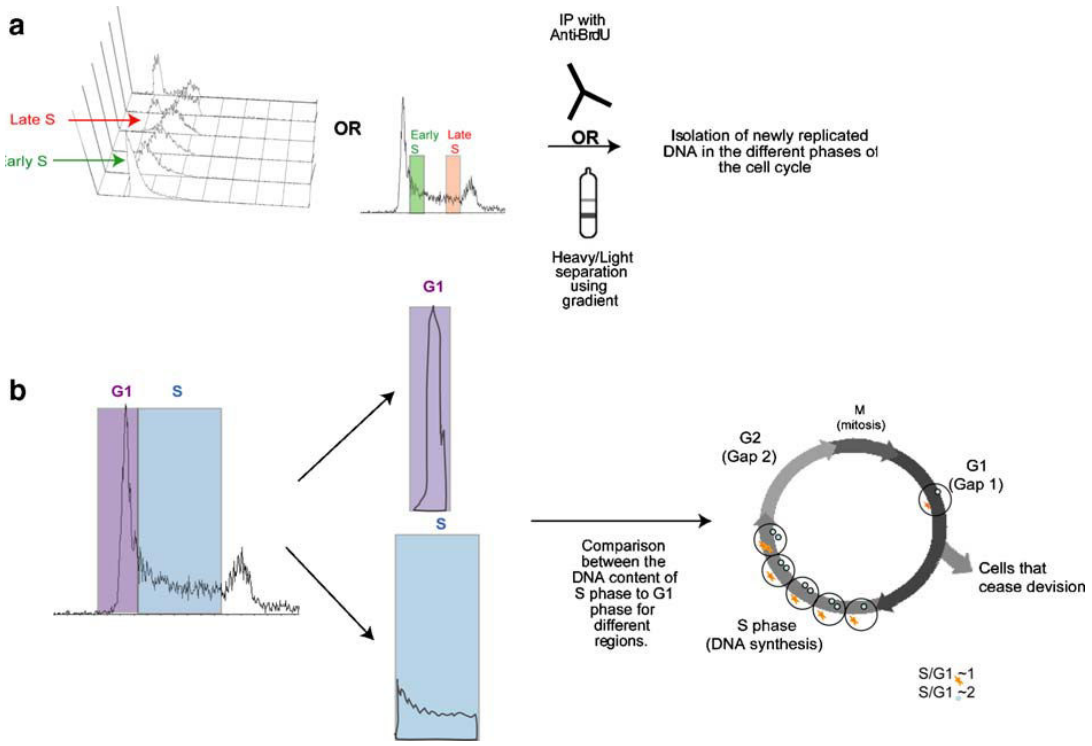


Figure 2.8. Two major technologies to measure replication timing. (a) Isolation of newly replicated DNA at various time points during S phase. The newly replicated DNA is labeled by BrdU and isolated by immunoprecipitation or density fractionation. (b) Replication timing is measured by the changes in DNA content. In an unsynchronized culture, the DNA content of a region decreases with its replication timing: early replicating regions are present in two copies in most of S phase cells, whereas late replicating regions are present in only one copy in most cells. DNA is harvested from S phase cells and from G1 phase cells and the DNA content of each region is compared. Adapted from [154].

cells at a certain stage of the cell cycle, usually by florescence-activated cell sorting (FACS). The cell population can be classified into 2, 4 or 6 bins (6 bins for the example in Figure 2.9), labelled G1b, S1, S2, S3, S4, G2, [151]). In the original study [155], cells were only classified in 2 classes: early or late replicating. This method with only 2 classes can not distinguish 2 early replicating regions. This limitation has been overcome by sorting additional fractions of the S-phase leading to 4 and 6 bins. After, this sorting, the newly synthesised DNA is sequenced and mapped on the genome. For each bin, the density of tags is computed genome wide. This method was first applied in *Drosophila* [110], in mouse [37, 38, 147, 148], and in human [37, 38, 149, 151].

In this manuscript, we use data from [149, 151]. To efficiently summarise the information contained in the six temporal bins, the data were normalised as described in [44]. For a given cell line and for each S-phase fraction the tag densities were computed in 100 kb windows and following the authors of [151] the tag densities were normalised to the same genome-wide sequence tag counts for each fraction, and a second normalisation was performed so that at each position, the sum over S-phase fractions be one. To filter out the noise that bias the mean replication timing profile estimate (Fig. 2.9 A), it was noticed that the genome-wide distribution of the normalised tag density (Fig. 2.9 D) presents a mode at $0.01 < m < 0.08$ and a long tail up to 1. For each S-phase fraction, $4m$

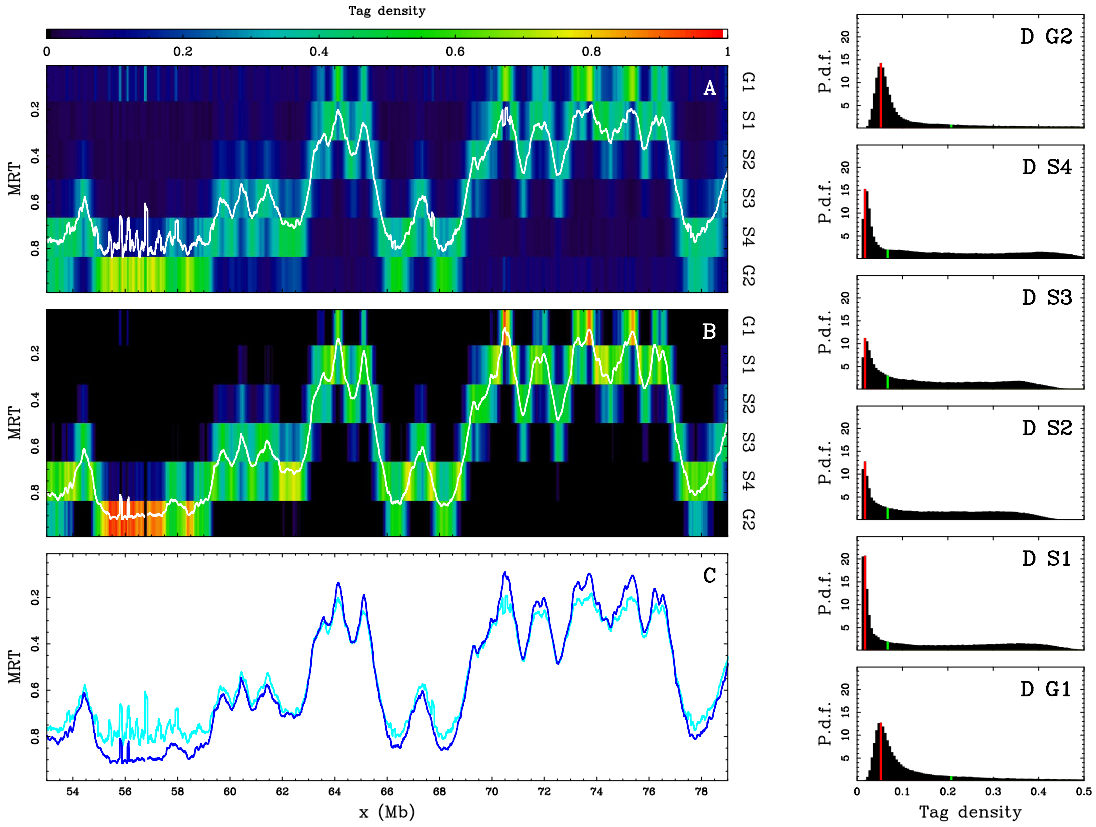


Figure 2.9. Establishment of the mean replication timing (MRT). (A) Normalized tag densities on a 25 Mb long fragment of chromosome 10, for the GM06990 cell line, and the corresponding computed MRT (white line). (B) "Denoised" normalized tag densities on the same genomic fragment and the corresponding MRT (white line). In (A) and (B) the tag densities for each S-phase fraction (G1-G2) are color coded using the color map situated at the top. (C) Comparison of the MRT computed on the normalized tag densities (cyan line) and the MRT computed on the "denoised" normalized tag densities (blue line), on the same genomic fragment. (D) Probability density function (P.d.f.) of the genome-wide distribution of the normalized tag densities for each S-phase fraction from G1 to G2 from bottom to top (black histogram). The mode m of the distribution is given by the red bar, the threshold $4m$ used for denoising is given by the green bar. Adapted from [44].

was subtracted from the tag density and resulting negative values were set to 0. Finally, each genomic position was re-normalised by the sum over S-phase fractions. The mean replication timing profile computed on these denoised tag densities superimposes on the original one but is much less noisy (Fig. 2.9 C).

☞ The MRT data have been extensively analysed in relation to transcriptional activity, chromatin state and genomic features like gene density and GC content. In metazoan, significant correlations have been observed between early replicating loci and regions presenting a strong transcriptional activity, an open chromatin state, a high gene density or a GC rich nucleotide composition [35, 36, 110, 150, 151, 156–158]. Replication timing profiles have also been used to delineate replication domains along chromosomes. The widely adopted point of view is to look for constant timing regions (CTRs) using segmentation algorithms like those developed to analyse variation in genome copy number (array CGH) [35, 147, 153, 157, 159, 160]. Alternatively, timing transition regions (TTRs) have been

extracted from replication timing profiles directly by looking for long regions of constant slope [36]. In both cases, these segmentations implicitly assume a crude dichotomous nature of replication domains. CTRs are interpreted as regions of coordinated origin firings and the intervening TTRs as origin-less regions replicated by the unidirectional progression of a single fork [36, 147, 157].

2.1.3.3 Experimental data corroborate replication origin predictions: N-domains are active replication domains in the germline

Examining the average timing profiles computed from Repli-Seq data [149, 151] for 5 cell types including one embryonic stem cell line (BG02), a fibroblast cell line (BJ), a lymphoblastoid cell line (GM06990), and an erythroid cell line (K562), it was observed that they all present numerous peaks pointing towards early replication timing (See Fig. 2.10 for BG02, GM06990 and K562 cell lines) [47]. The regions at the tip of the peaks are on average replicated earlier than their surrounding regions and thus harbor replication initiation zones highly active in the corresponding cell types. In the 11.4 Mb region analysed in Figure 2.10, a strong correspondence was observed between the germline putative replication origins at N-domain borders and the initiation zones pointed to by the timing peaks. Figure 2.10 shows a strong conservation of timing peak location between cell lines in this region. In order to perform a systematic comparison between these loci containing conserved replication origins and N-domain borders, a multi-scale methodology for the detection of peaks along timing profiles was developed (Appendix A, Section A.4). For the 5 considered cell lines, timing peaks were detected: 706 in GM06990, 795 in K562, 981 in BJ, 1556 in HeLa and 1690 in BG02 [161]. When comparing the observed distribution of distances of the N-domain borders to the closest timing peak to the expected distribution for uniformly distributed borders, a significant excess of short distances ($\lesssim$ 175 kb) was observed. Considering that N-domain borders and timing peaks coincide if they are within 100 kb from each other, 57% of N-domain borders were associated to an active replication region in at least one of the considered cell lines [161]. These results provided a quantitative evidence for the co-localisation of N-domain borders with the active initiation zones at MRT peaks in different cell lines and support the idea of the interpretation of skew N-domains (Fig. 2.10) as independent replication units in germline cells.

2.1.4 Megabase-sized gradients of replication fork-polarity: Towards a cascade model of replication initiations

Replication generates strand asymmetries by discriminating a leading and a lagging strand. Strand asymmetry is defined relatively to the reference strand for which we have the DNA sequence. A replication fork is defined as *sense* $(+)$ *fork* if it "moves" in the $5' \to 3'$ direction seen from the reference strand, and as *antisense* $(-)$ *fork* if it "moves" in the opposite $3' \to 5'$ direction. In other words, a sense $(+)$ fork comes from a replication origin that fired upstream ($5'$ direction of the reference strand), whereas an antisense $(-)$ fork comes from a replication origin that fired downstream ($3'$ direction of the reference strand). During the S-phase, each locus is replicated once and only once [162], and it is either replicated by a sense or an antisense fork. Over cell cycles, the locus x will be replicated by a proportion $p_{(\pm)}(x)$ of $(\pm)$ forks. As the proportions of sense and antisense forks always sum up to one, only the difference of proportions is

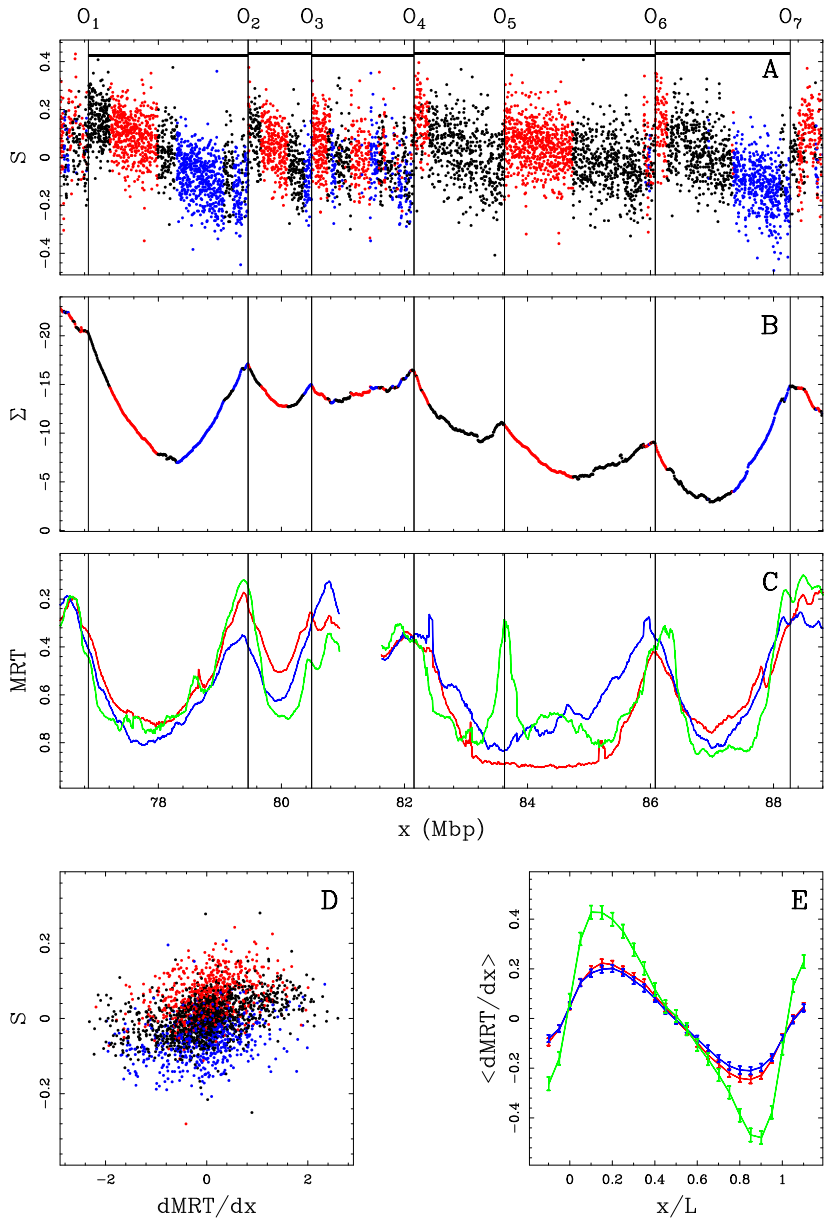


Figure 2.10. Comparing skew S and mean replication timing (MRT). (A) S profile (Equation (2.2)) along a 11.4 Mb long fragment of human chromosome 10 that contains 6 skew N-domains (horizontal black bars) bordered by 7 putative replication origins O_1 to O_7. Each dot corresponds to the skew calculated for a window of 1 kb of repeat-masked sequence. The colors correspond to intergenic (black), $(+)$ genes (red) and $(-)$ genes (blue). (B) Corresponding cumulative skew profile Σ obtained by cumulative addition of $S-$values along the sequence (Equation (2.4)). (C) MRT profiles from early, 0 to late, 1 for BG02 (green), K562 (red) and GM06990 (blue) cell lines. (D) Correlations between S and $d\mathrm{MRT}/dx$, in BG02 (100 kb windows) along the 22 human autosomes; colors as in (A). (E) Average $d\mathrm{MRT}/dx$ profiles ($\pm$ SEM) in the 663 skew N-domains after rescaling their length L to unity; colors as in (C). Adapted from [44].

relevant. This difference defines the *replication fork polarity*:

$$p(x) = p_{(+)}(x) - p_{(-)}(x). \tag{2.5}$$

We define the replication fork polarity for a locus x, but it can be equally defined for a genomic region. When the replication fork polarity $p = +1$ (resp. $p = -1$), the genomic region only undergoes leading (resp. lagging) strand synthesis, hence the strand asymmetry due to replication is maximal in such regions. Between these two extreme cases, the replication fork polarity can take values in the whole interval $[-1, 1]$. When the replication fork polarity $p = 0$, there is as many leading and lagging strand synthesis, and consequently there is no strand asymmetry due to replication in these regions. To establish the existence of replication domains associated with replication fork polarity gradients, we first recall the relations between replication fork polarity, nucleotide compositional skew and derivative of the replication timing profile.

2.1.4.1 Linking fork polarity to nucleotide compositional skew

In bacteria, the compositional skew is piecewise constant changing sign at the replication origin and termination positions This step-like skew profile mirrors the opposite orientations of the two divergent replication forks starting from the replication origin and meeting at the replication terminus (Section 2.1.1.3, Fig. 2.4). This property illustrates the existence of a general relationship between replication-associated compositional asymmetries and the average replication fork orientation. Using the formalism of Markov processes, it was demonstrated [44, 48, 49] that replication-associated asymmetries between the substitution rates of the two DNA strands induce, in the limit of small asymmetries, a nucleotide compositional skew S_R proportional to the replication fork polarity:

$$S_R(x) \sim p(x). \tag{2.6}$$

According to Equation (2.6), the observed linear decrease of the skew S in N-domains from positive (5' end) to negative (3' end) values likely reflects the progressive linear decrease of the replication fork polarity with a change of sign in the middle of the skew N-domains. These results provide strong support to the interpretation of skew N-domains (Fig. 2.6) as independent replication units in germline cells delimited by well positioned replication origins and presenting an overall gradient of the average replication fork orientation.

2.1.4.2 Replication fork polarity and replication timing data

In this section, linking replication fork polarity to MRT allows to bring another evidence of the organisation of DNA duplication in replication domains. In [44], it was shown that the replication fork polarity is related to MRT, under the central hypotheses that the replication fork velocity v is constant and that replication is bidirectional from each origin O_i. In fact, for a given cell cycle, let n be the number of activated origins, $x_1 < \cdots < x_n$ their positions along the genome and $t_1, \cdots, t_n$ their initiation times. Then the configuration $\mathcal{C} = O_1 \cdots O_n = (x_1, t_1) \cdots (x_n, t_n)$ (where and when the origins of replication fire during the S-phase) completely specifies the spatio-temporal replication programme (Fig. 2.11) [163, 164]. Let T_i be the termination locus where the fork coming from O_i meets the fork coming from O_{i+1} whose space-time coordinates (y_i, u_i) are:

$$\text{and} \quad \begin{array}{ll} y_i & = \frac{1}{2}[(x_{i+1} + x_i) + v(t_{i+1} - t_i)], \\ u_i & = \frac{1}{2}[(t_{i+1} + t_i) + (x_{i+1} - x_i)/v], \end{array} \tag{2.7}$$

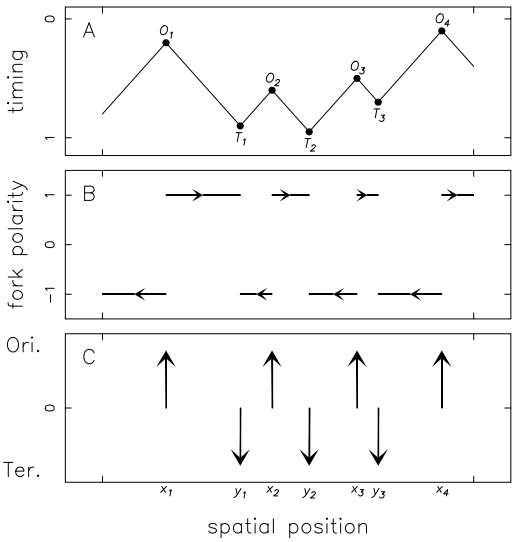


Figure 2.11. Modelling the spatio-temporal replication programme in a single cell. (A) Replication timing $r(x)$, (B) replication fork polarity $p(x)$ (Eq. (2.5)) and (C) spatial location of replication origins (upward arrows) and termination sites (downward arrows). $O_i = (x_i, t_i)$ corresponds to the origin i positioned at location x_i and firing at time t_i. Fork coming from O_i meets the fork coming from O_{i+1} at termination site T_i with space-time coordinates (y_i, u_i) given in Eq. (2.7). Note that one can deduce the fork polarity in B (resp. origin and termination site locations in C) by simply taking successive derivatives of the timing profile in A (Eqs. (2.9) and (2.10)). The fundamental hypothesis is that the replication fork velocity v is constant. Adapted from [43].

then the replication timing profile $r(x)$ and replication fork polarity $p(x)$ (Eq. (2.5)) around origin O_i ($x \in [y_{i-1}, y_i]$) are given by (Fig. 2.11):

$$r(x) = t_i + |x - x_i|/v \quad \text{and} \quad p(x) = \text{sign}(x - x_i). \tag{2.8}$$

Finally, using the Dirac function δ to represent origin locations $\delta(x - x_i)$ and termination sites $\delta(x - y_i)$ (Fig. 2.11(c)), the following fundamental relationships were obtained:

$$v \, \frac{d}{dx} r(x) = p(x), \tag{2.9}$$

$$v \, \frac{d^2}{dx^2} r(x) = \frac{d}{dx} p(x) = 2 \left(\sum_i \delta(x - x_i) - \sum_i \delta(x - y_i) \right). \tag{2.10}$$

In other words, we can extract, up to a multiplicative constant, the fork polarity $p(x)$ (Fig. 2.11(b)) and the location of origins and termination sites (Fig. 2.11(c)) by simply taking successive derivative (Equations (2.9) and (2.10)) of the timing profile $r(x)$ (Fig. 2.11(a)). Finally, the above results can be rewritten for application to experimental replication data obtained from a large number of cells (millions), from which only population statistics can be derived with a finite spatial resolution of tens of kb or more [36–38, 149–151]. Since taking the spatial derivative commutes with statistical and spatial average, one gets:

$$\frac{d}{dx} \text{MRT}(x) = \frac{1}{v} \langle p(x) \rangle_{\text{Cells},\Delta x} \, , \tag{2.11}$$

$$\frac{d^2}{dx^2} \text{MRT}(x) = \frac{2}{v} \left(N^{\text{Ori}}_{\text{Cells},\Delta x}(x) - N^{\text{Ter}}_{\text{Cells},\Delta x}(x) \right) \, , \tag{2.12}$$

where $\text{MRT}(x) = \langle r(x) \rangle_{\text{Cells},\Delta x}$, $\langle . \rangle_{\text{Cells},\Delta x}$ stands for the average over many cells and over the spatial resolution Δx and $N^{\text{Ori}}_{\text{Cells},\Delta x}(x)$ (resp. $N^{\text{Ter}}_{\text{Cells},\Delta x}(x)$) is the number of origins (resp. termination sites) per unit length averaged over many cells and the spatial

coordinate. Note that when replication fork speed is inhomogeneous with fork speed fluctuations that do not depend on the replication timing nor on the spatial coordinate then Equations (2.11) and (2.12) remain valid with v standing for the average replication speed.

Replication fork polarity provides a direct link between the skew S and the derivative of the MRT (Equations (2.6) and (2.9)). This link between skew profiles and the MRT data raised the issue whether the replication domains observed in germline cells as skew N-domains correspond to a mode of replication that also exists in different cell types.

To test this relationship, the MRT profiles of seven somatic cell lines (one embryonic stem cell, three lymphoblastoid, a fibroblast, an erythroid and HeLa cell lines) were used as a substitute to germline MRT [149, 151]. First the skew S was correlated with dMRT/dx, in the BG02 embryonic stem cells, over the 22 human autosomes (Fig. 2.10 D). The significant correlations observed in intergenic ($R = 0.40$, $P < 10^{-16}$), genic (+) ($R = 0.34$, $P < 10^{-16}$) and genic (−) ($R = 0.33$, $P < 10^{-16}$) regions are representative of the correlations observed in the other 6 cell lines [44]. These correlations are as important as those obtained between the dMRT/dx profiles in different cell lines [44], as well as those previously reported between the replication timing data themselves [37, 38, 151]. The correlations between S and dMRT/dx are even stronger when focusing on the 663 skew N-domains. The correlations obtained in intergenic regions ($R = 0.45 \pm 0.06$) are recovered to a large extent in genic regions ($R = 0.34 \pm 0.03$) where the transcription-associated skew S_T was hypothesised to superimpose to the replication-associated skew S_R [41, 45, 46]. Further evidence of this link between S and dMRT/dx was obtained when averaging, for the different cell lines, the dMRT/dx profiles inside the 663 skew N-domains after rescaling their length to unity (Fig. 2.10 E). These mean profiles are shaped as a N, suggesting that some properties of the germline replication programme associated with the pattern of replication fork polarity are shared by somatic cells [44].

☞**Apparent speed of replication**: If from cell to cell, a region R is systematically replicated by one fork moving across the region with a specified direction then $|\langle p(x)\rangle_{Cells,R}| = 1$ and the derivative of the timing profile along the region (Eq. (2.11)) is an estimate of the inverse of the fork speed. We can thus define the local apparent speed of replication v_{app} *i.e.*, the speed of a single replication fork reproducing the same slope of the timing profile, as the inverse of $\frac{d}{dx}\mathrm{MRT}(x)$. This definition clarifies the status of the MRT gradients. Rewriting Eq. (2.11), we get $v_{app,R} = v/\langle p(x)\rangle_{\mathrm{Cells},R}$ showing that the sign of the apparent replication speed indicates the predominant direction of replication progression and that, in flat domains of uniform MRT (infinite apparent replication speed), forks move equally in both directions.

2.1.4.3 Replication timing U-domains are replication domains robustly observed in different cell types

According to Equations (2.6) and (2.11), the integration of the skew S is expected to generate a profile rather similar to the MRT profile. In segments of linearly changing skew, the integrated S function is thus expected to show a parabolic profile. The integrated S function when estimated by the cumulative skew Σ (Equation (2.4), Fig. 2.10 B) along N-domains of a 11.4 Mb long fragment of human chromosome 10, indeed displays a U-shaped (parabolic) profile likely corresponding the MRT profile in the germline. Remarkably, the

6 N-domains effectively correspond to successive genome regions where the MRT in the BG02 embryonic stem cells is U-shaped (Fig. 2.10 C). The 7 putative initiation zones (O_1 to O_7) corresponding to upward S-jumps (Fig. 2.10 A), co-locate (up to the $\sim$ 100 kb resolution) with MRT local extrema which indicates that they are also highly active in BG02. These initiation zones can present cell specificity as exemplified by the putative replication origin O_5 which is inactive (or late) in both the K562 erythroid and GM06990 lymphoblastoid cell lines (Fig. 2.10 C) resulting in domain "consolidation" [148]. Two neighbouring U-domains ($[O_4, O_5]$ and $[O_5, O_6]$) in BG02 merged into a larger U-domain in the K562 and GM06990 cell lines. Note that the other 3 N-domains ($[O_1, O_2]$,$[O_2, O_3]$, and $[O_6, O_7]$) are MRT U-domains common to BG02, K562 and GM06990. To detect U-domains in MRT profiles at genome scale, a wavelet-based method (Appendix A, Section A.5) was developed which allowed to identify in 10 human cell lines (Table 2.1 show the analyzed data) from 664 (TL010) up to 1534 (BG02) U-domains of mean size ranging from 0.97 Mb (HeLa R2) up to 1.62 Mb (TL010) and covering from 40.40% (TL010) to 63.06% (BG02) of the genome (Table 2.2). Interestingly, for each cell line, the average MRT profile of U-domains has an expected parabolic shape representative of individual U-domains. Inside the U-domains, the derivative $d\text{MRT}/dx$ is N-shaped like the skew profile inside N-domains. When rescaling the size of each U-domains to unity for a given cell line, these profiles superimpose onto a common N-shaped curve well approximated by the average $d\text{MRT}/dx$ profile (Fig. 2.10 E).

Overall it was observed that (i) MRT U-domains are robustly observed in all cell lines, covering $\sim$ 50% of the human genome, (ii) about half of the U-domains in one cell line is shared by at least another cell line (from 38.4% to 61%) and (iii) this is also true for the skew N-domains (50.2%) that likely correspond to MRT U-domains in the germline. However about half of the genome that is covered by U-domains in fact corresponds to regions of high replication timing plasticity where replication domains may (i) reorganise according to the so-called "consolidation" scenario (Fig. 2.10), (ii) experience some boundary shift and (iii) emerge in a late replicating region as previously observed in the mouse genome during differentiation [148].

☞**A cascade model of initiations along replication domains**: To sum up, it appears that the organisation of replication into mega-based sized replication fork polarity gradients along about half of the human genome is a general characteristic of the average spatio-temporal replication programme in the germline (N-domains) as well as in different cell lines where MRT profiles were analysed (U-domains) (7 cell lines were analysed in [44]). Moreover, the analysis of replication kinetics questions the reported absence of replication origins within timing transition regions [36, 147, 157] and rather suggests a mode of replication by sequential activation of origins along these regions [70]. Taking into account that replication domain borders correspond to $\sim$ 200 kb wide regions of open chromatin (See [51] for N-domain borders and the corresponding analysis for U-domain borders in [44]), a model of the spatio-temporal replication programme was proposed where replication first initiates at "master" origins in the open chromatin environment at MRT domain borders, followed by successive activations of secondary origins within replication domains according to the observed U-shaped timing profile [42, 44, 51, 69, 70]. One possible mechanism for this is that secondary origins are remotely activated by the approach of a center oriented fork. This "saltatory transmission" of origin activation could explain why replication progress from replication domain borders much faster (3-5

Cell line	Cell Type	Replicates	Accession	Reference
BG02	Embryonic stem cell	1	NCBI SRA website SPR 0013933	[151]
IMR90	Foetal lung fibroblast	1	ENCODE wgEncodeUWRepliSeq	[165]
BJ	Skin fibroblast	2	NCBI SRA website SPR 0013933	[151]
GM06990	Lymphoblastoid	1	NCBI SRA website SPR 0013933	[151]
GM12878	Lymphoblastoid	1	ENCODE wgEncodeUWRepliSeq	[165]
TL010	Lymphoblastoid	1	NCBI SRA website SPR 0013933	[151]
H0287	Lymphoblastoid	1	NCBI SRA website SPR 0013933	[151]
K562	Myelogenous leukemia (blood cancer)	1	NCBI SRA website SPR 0013933	[151]
HeLa	Cervical carcinoma (cervical cancer)	2	NCBI SRA website SPR 0013933	[151]
HeLaS3	Derived from HeLa	1	ENCODE wgEncodeUWRepliSeq	[165]
MCF7	Mammary gland (breast cancer)	1	ENCODE wgEncodeUWRepliSeq	[165]

Table 2.1. Mean replication timing analysed data. First column indicates the cell line in wich the MRT is available, the second column is the cell type to which a cell line belongs, the third column is the number of available replicates, the forth column is the accession number under wich the data are available and the last column is the reference to the original paper.

	BG02	IMR90	BJ (R1)	BJ (R2)	GM06990	GM12878	TL010	H0287	K562	HeLaS3	HeLa (R1)	HeLa (R2)	MCF7
N	1534	1135	1150	1247	882	825	664	828	876	1388	1422	1498	891
L	1.09	1.28	1.19	1.15	1.52	1.58	1.62	1.57	1.42	1.13	1.06	0.97	1.38
C	63.06	54.81	51.52	54.18	50.41	48.40	40.40	48.80	46.96	58.89	56.75	54.52	46.16

Table 2.2. Characteristics of replication domains. Number (N) of detected MRT U-domains, their mean length (L) and their genome coverage (C) in each of the analysed cell lines.

times) than the known speed of single fork obtained by DNA combing technique [70]. The dynamic pattern with which secondary initiations occur governs the progressive change in the proportion of center- and border- oriented forks that is revealed by skew N-domains and replication U-domains. These results question to which extent chromatin state influences fork progression and secondary initiations and whether outside of replication domains, the genome replicates according to a similar or completely different set of rules.

2.2 Human genome replication proceeds through 4 chromatin states

2.2.1 Genomic DNA codes for open chromatin around replication skew domain borders

The complex formed by DNA and the proteins attached to it is named *chromatin*. DNA wraps around a bead-like structure formed by an octamer of proteins called histones. DNA turns about twice around each octamer [3, 166, 167]. The complex formed by the DNA and the eight histones is called nucleosome. Each canonical nucleosome contains 147 bp of DNA. Histones are the most prevalent proteins in chromatin. There are four types of histones: H3, H4, H2A, and H2B. H3 and H4 associate together to form a dimer; two dimers of H3-H4 associate to form a tetramer which is, in turn, surrounded by two dimers of H2A and H2B. Histones carry chromatin marks that convey a lot of functional

information by two mechanisms. First, one of the canonical histones can be replaced by a histone variant[†]. This replacement slightly modifies the nucleosome structure and its function. For instance, transcriptionally active regions are enriched in H3.3 and H2AZ variants instead of H3 and H2A, respectively. Second, histones are formed of a globular part that constitues the nucleosome core and a flexible "tail" that reaches outside toward the nuclear enviroment. These tails carry diverse covalent modifications that have a functional meaning. There is a specific notation to indicate modifications: H3K9ac means that a histone H3 carries a modification on its ninth amino-acid and that the modification is acetylation. Histone modifications can make chromatin looser (acetylation modifications) or can serve as an anchor to regulatory proteins. For instance, H3K27me3 is used as a docking station by the Polycomb Repressive Complex PRC1 that silences developmental genes[‡]. Other diverse proteins are associated to DNA. Some are needed to transcibe active genes (transcription factors), some repair and replicate DNA, some repress genes by compacting the chromatin and others modify histones. For example, the families of proteins that add and remove acetyl groups to histone tails are named HAT (Histone Acetyl Transferase) and HDAC (Histone DeACetylase). There is also a class of proteins that move nucleosomes along DNA or eject them called chromatin remodellers. Finally, there is a class of proteins that organise the fourth layer of chromatin, namely its three dimensional folding (cohesin and CTCF for example) [169].

Recently, the development of new techniques, in particular chromatin immunoprecipitation (ChIP) followed by massive parallel sequencing (ChIP-seq) [170], has enabled genome-wide analysis of many chromatin marks such as histone modifications, histone variant incorporation as well as of various DNA-binding proteins [171]. These techniques have been extensively applied to various eukaryotic genomes, from budding yeast [172], to plants [173, 174], fly [175, 176], mouse [171, 177] and human [171, 177, 178], and have led to significant progress in our understanding of the chromatin landscape and of its impact on gene regulation, replication origin specification and cell differentiation.

Genome-wide investigation of chromatin architecture has revealed that, at large scales (from 100 kb to 1 Mb), regions enriched in open chromatin fibers correlate with regions of high gene density [179]; whereas, at small scales (<1 kb), DNA accessibility, nucleosome distribution and modifications are important determinant for transcriptional activity [180–184]. Moreover, there is a growing body of evidence that transcription factors are regulators of origin activation (reviewed in [185]). In this context, it has been asked to which extent the remarkable genome organisation observed around N-domain borders is mediated by particular chromatin structure favorable to specification of early replication origins [41]. Historically, genome-wide availibility of DNase I hypersensitive (HS) sites [184] provided the opportunity to study open chromatin in relation to the observed nucleosome-depleted regions [180–184] that look very similar to the nucleosome-free regions (NFRs) previously observed at yeast promoters [186, 187]. Mapping chromatin mark data in the replication N-domains showed a significant enrichment around the borders with a decrease towards the domain center [51]. A significant subset of N-domain borders were shown to correspond to particular open chromatin regions, permissive to transcription. From the demonstration that there is an excess of NFRs at N-domain bor-

[†]Proteins are formed by a chain of molecular units called amino-acids. Histone variants have the same amino-acid sequence as the canonical histone up to a few substitutions.

[‡]For a complete summary on diverse histone modifications see [168].

ders, it has been suggested that these putative replication origins have been imprinted in the DNA sequence during evolution [51]. Similarly when mapping open chromatin marks inside replication U-domains, it has been recently shown that these domains are bordered by early replication initiation zones likely specified by $\sim$ 200 kb wide region of accessible, open chromatin permissive to transcription [44].

2.2.2 Few chromatin states sum up chromatin complexity

Statistical analyses of multivariate epigenetic data sets have shown that this huge combinatorial complexity can be reduced to a surprisingly small number of predominant chromatin states with shared features namely four in *Arabidopsis thaliana* [188], five in *Caenorhabditis elegans* [189] and four [20] or five [190] in *Drosophila*. In human, the application of a Hidden Markov Model (HMM) [191] as well as the implementation of adapted pattern-finding algorithm, have confirmed that distinct epigenetic modifications often exist in well-defined combinations corresponding to different genomic elements like promoters, enhancers, exons, repeated sequences and/or to distinct modes of regulation of gene expression (transcribed, silenced and poised) [191–193]. A recent study [194] of chromatin mark maps across nine different human cell types has ultimately identified fifteen main chromatin types which is a relatively limited number of epigenetic states but probably not the optimal complexity reduction one may achieve in human and more generally in mammalian genomes. The analysis of a wide set of chromatin regulators that add, remove or bind histone modifications reported in [195], is a very encouraging step in this direction since six major groups or modules of chromatin regulators were shown to encompass the combinatorial complexity and to be associated with distinct genomic features and chromatin environments.

In the following, we review the result of a recent study of the host team [53, 54], based on an integrative analysis of eleven genome-wide chromatin marks profiles at 100 kb resolution of mean replication timing (MRT) data. This study identified four major groups of chromatin marks with shared features. This study used principal component analysis (PCA)[§] [15] and classical clustering [52] to investigate relationships between genome-wide distributions of nine histone modifications, one histone variant and one DNA binding protein at a 100 kb resolution in five sommatic cell types including an immature myeloid cell line (K562), a monocyte cell line (Monocd14ro1746), a lymphoblastoid cell line (GM12878), a mammary epithelial cell line (Hmec), an adult dermal fibroblast cell line (Nhdfad) and an ESC cell line (H1 ES). Note that when comparing chromatin state organisation to MRT data, MRT profiles in the human embryonic stem cell line BG02 (resp. fibroblast IMR90) were used as surrogate profiles in H1 ES (resp. Nhdfad).

The matrix resulting from the computation of the Spearman correlation coefficient[¶] of each mark with the others is reproduced in Figure 2.12 as a heatmap after having reorganised rows and columns with a hierarchical clustering based on Spearman correlation distance. The correlation matrices obtained for the five somatic cell lines strongly ressemble to each other (data only shown for Nhdfad bottom pannel of Figure 2.12) while the

[§]PCA [15] is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called principal components.

[¶]Spearman correlation coefficient [15] is a nonparametric measure of statistical dependence between two variables whose relationship can be described by a monotonic function. A perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other.

pluripotent H1 ES cell line showed a different correlation structure between epigenetic marks (Fig. 2.12, top pannel) [54]. In the epigenetic mark matrices obtained for Nhdfad (Figure 2.12, bottom panel), all histone modifications that are known to be involved in transcription positive regulation, namely H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K36me3, H3k79me2 and H4K20me1, form a block that also includes the histone variant H2AZ and the transcription factor CTCF, meaning that all these marks are all correlated with each other and are likely to occupy similar regions in the genome [183, 194]. In fact, two lines are clearly apart in all correlation matrices as illustrated on the hierarchical clustering dendrogram (Figure 2.12, bottom panel). They correspond to the repressive chromatin marks H3K27me3 and H3K9me3 that are respectively associated with the so-called facultative and constitutive heterochromatins [53]. These two marks are recognized by the chromodomains of polycomb (Pc) proteins and heterochromatin protein 1 (HP1) respectively, components of distinct gene silencing mechanisms which may explain that they are anti-correlated with each other. While H3K9me3 behaves quite independently if not anticorrelated with most of the active chromatin marks (except for GM12878 where some positive correlations were observed), H3K27me3 correlates to some of them in a cell line dependent fashion but more systematically to CTCF and H4K20me1 (Figure 2.12, bottom panel). Taking advantage of the consistency between the epigenetic mark correlations in the five differentiated cell lines and to reduce the dimensionally of the data, principal component analysis (PCA) was applied on the "shared" epigenetic space [‖]. Four principal components accounted for 86% of the total set variance and the data showed meanigful patterns along them. Thus four chromatin states were identified and labeled with a color [53, 54]. The correlation matrix obtained for the same 11 epigenetic mark profiles of the pluripotent H1 ES cell line (Figure 2.12, top panel) displays important differences from the ones previously obtained for differentiated cell lines. Among others, let us mention the repressive polycomb-associated mark H3K27me3 which now strongly correlates with most of the active marks including H3K4me3 as the probable signature of bivalent ESC chromatin [183, 196–200]. Also the histone variant H2AZ that now correlates as much with both the repressive marks H3K27me3 and H3K9me3 as with some of the active marks, which is likely an indication of the specific highly dynamic and accessible chromatin of pluripotent cells [171, 183, 196–198, 201]. When reproducing PCA and clustering analysis on the H1 ES epigenetic data, again four PCs were enough to account for 86% of the total variance, and thus one could still reduce the ESC epigenetic complexity to four chromatin states but, as described below, these chromatin states are distinct from the ones delineated in sommatic cells confirming that ESCs and differentiated cells have different epigenomes [171, 177, 183, 196, 197, 201].

2.2.3 Epigenetic content of prevalent states in ESCs *vs* differentiated cells

The four prevalent chromatin states so identified in the five differentiated cell lines [54] are quite similar to the ones found in K562 in a preliminary study [53] (see also [198]). C1 is a gene rich transcriptionally active euchromatin state enriched in the histone modifications involved in transcription positive regulation, namely H3K4me1, H1K4me2, H3K4me3, H3K27ac, H3K36me3, H3K79me2 and H4K20me1, as well as in the histone variant H2AZ whose binding level was shown to correlate with gene activity in human [183]. C2 is a

[‖]For the five differentiated cell lines a shared epigenetic space was created by concatenating all the 100 kb epigenetic profiles of the same mark [54].

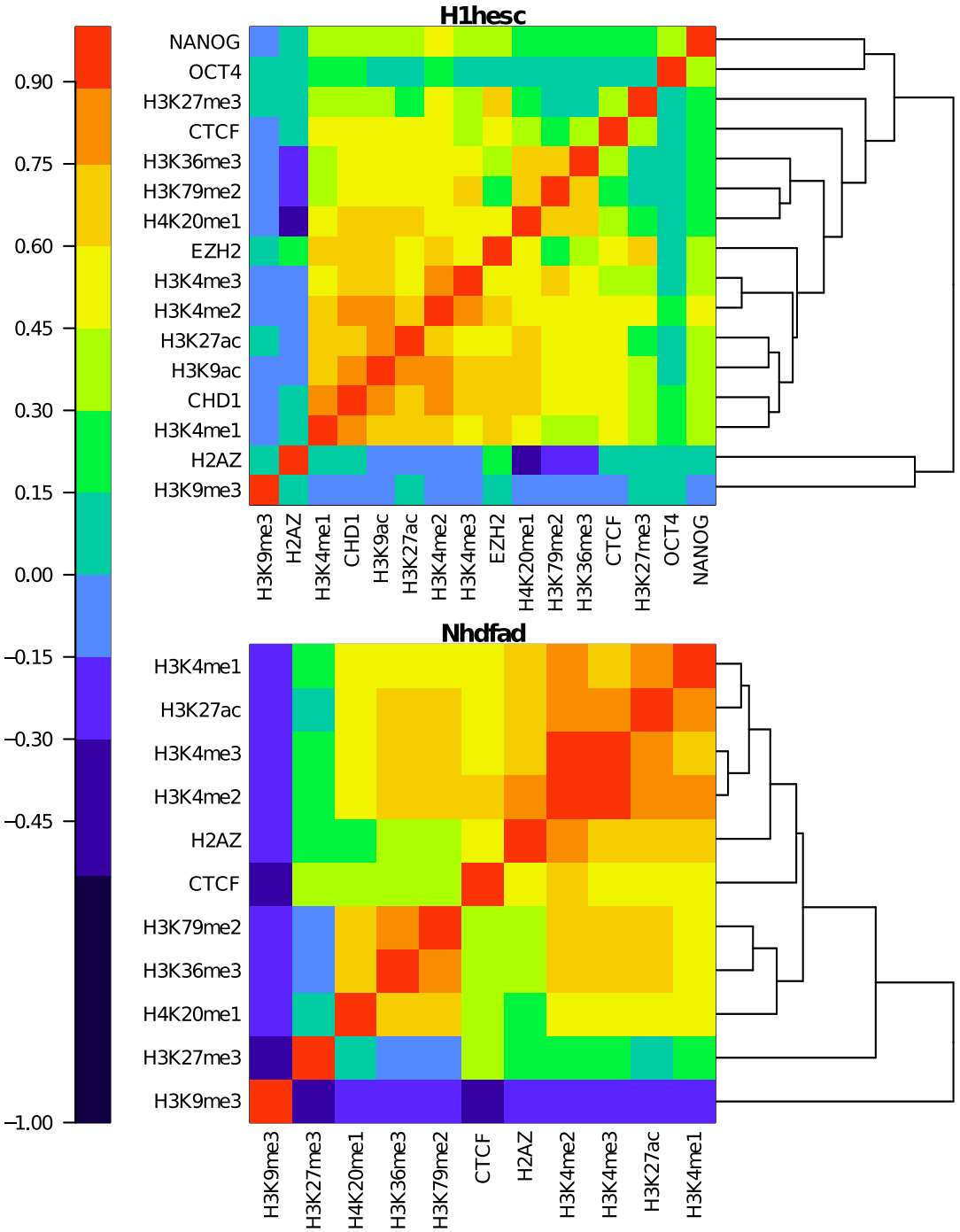


Figure 2.12. Spearman correlation matrix between epigenetic marks in H1 ES (top) and Nhdfad (bottom) cell lines. For each cell line, the Spearman correlation is computed over all 100 kb non overlapping windows with a valid score. Spearman correlation value is color coded using the colormap shown on the left. Lines for the epigenetic marks were reorganised by a hierarchical ordering using Spearman correlation distance as illustrated by the dendograms on the right of the corresponding matrices. This ordering implies that highly correlated epigenetic marks are close to each other. Reproduced from [54].

polycomb (Pc) repressed chromatin state [183, 202] notably associated with the histone modification H3K27me3. This epigenetic mark is recognised by the chromodomains of Pc that is known to induce gene silencing in the so called facultative heterochromatin [171, 183, 197, 198]. C3 can be compared to the "null" or "black" silent heterochromatin regions devoid of chromatin marks previously found in *Arabidopsis* [188] and *Drosophila* [20, 190]. C4 corresponds to a gene-poor constitutive heterochromatin state [183, 202] with all C4 100 kb loci containing the repressive mark H3K9me3 associated with the heterochromatin protein 1 (HP1). Note that the transcription factor CTCF that is known to establish chromatin boundaries to prevent the spreading of heterochromatin into transcriptionally active regions [169, 203] was found in C1 and to a lesser extent in C2. Prevalent chromatin states in pluripotent H1 ES cell line (EC1, EC2, EC3, EC4) are different even though they display some similarities with the above described differentiated chromatin states (C1, C2, C3, C4) [54]. Again among these four prevalent states, only one is transcriptionally active and three are silent. The first one is a gene rich euchromatin that contains all the active modification marks considered and is shared by pluripotent (EC1) and differentiated (C1) cells as well as the "unmarked" states EC3 and C3 that correspond to a silent state not enriched in any available epigenetic mark. The two other chromatin states bear more differences than similarities as the signature of the global accessible character of pluripotent chromatin [197, 198]. Almost all EC2 loci were found, like C2 loci, to be marked by H3K27me3 which is deposited by polycomb complex PRC2 and then enhanced PRC1 targeting [204, 205]. Consistently, EC2 is enriched in a subunit EZH2 of PRC2 containing a SET domain that acts on H3K27 as a methyltransferase, confirming the polycomb activity of this state. The additional observation that, relatively to EC1, EC2 contains more active mark H3K4me3 than C2 relatively to C1, is an indication of bivalent heterochromatin associated with bivalent genes [183, 196–200]. EC1, EC2 being the most genic chromatin states in ESCs, they both contain CTCF, as previously observed in differentiated chromatin states C1, C2, but EC2 is more enriched (via the bivalent genes) than C2 and vice versa for EC1 and C1. But, the most striking difference concerns the pluripotent state EC4 whose epigenetic content is qualitatively and quantitatively different from the one of C4. As compared to C4, EC4 contains significantly less HP1-associated heterochromatin mark H3K9me3 concomitant with an important excess in the histone variant H2AZ. In contrast to its local positioning, mainly at gene promoters, in EC1, EC2 and C1, C2, and its scarsity in C4, H2AZ known to be associated with nucleosome exchange and remodeling [172, 183, 206, 207] is broadly distributed in EC4 likely contributing to the highly dynamic properties of pluripotent chromatin and its refractory character to HP1-associated constitutive heterochromatin extension [171, 183, 197, 198, 207]. This interpretation was further strengthened by the observation that unlike C4, EC4 is enriched in CTCF which besides its insulator properties [169, 203], is also known to mediate long-range intra- and inter- chromosomal interactions [18, 208–212]. The fact that H2AZ was also found to be broadly distributed in the bivalent state EC2 containing bivalent genes confirmed that the polycomb repressed state C2 resulted from the spreading of H3K27me3 in differentiated cells [171, 183, 197, 198, 207].

2.2.4 Replication timing of chromatin states

As compared to previous integrative analyses of epigenetic data mainly performed at a few kb scale characteristic of gene promoters [20, 188–190], the results reported in this section were obtained, on purpose, at a much larger scale 100 kb allowing a direct com-

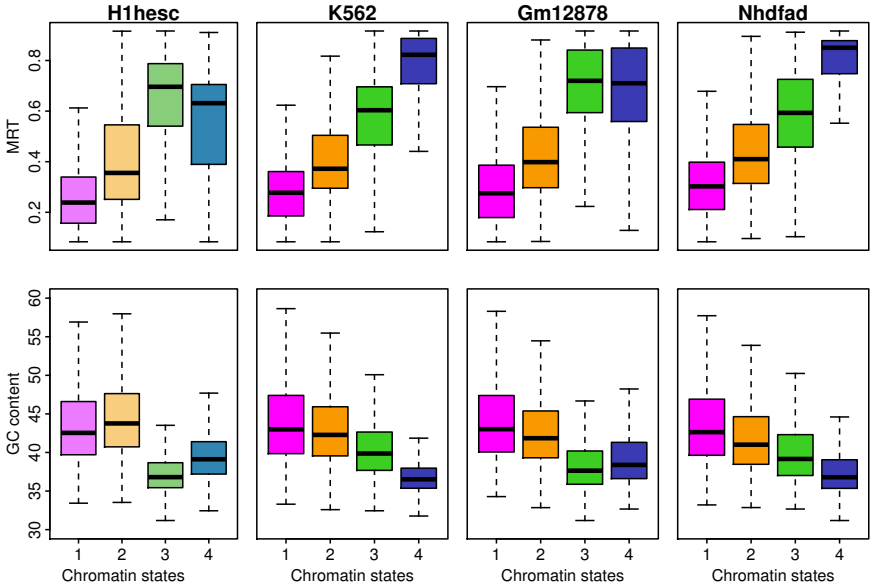


Figure 2.13. MRT and GC distributions in the four prevalent chromatin states. For pluripotent H1 ES cell line: EC1 (light pink), EC2 (light orange), EC3 (light green), EC4 (light blue), and for three differentiated cell lines (K562, GM12878, Nhdfad): C1 (pink), C2 (orange), C3 (green), C4 (blue). First row: Boxplots of MRT computed in 100 kb non-overlapping windows per chromatin state. Replication data in BG02, GM06990 and BJ were used as surrogates of replication data in H1 ES, GM12878 and Nhdfad, respectively. Second row: Boxplots of GC content computed in 100 kb non-overlapping windows per chromatin state. Adapted from Julienne *et al.* [54].

parison with MRT data [53, 54]. (We refer the reader to Ref. [213] for a complementary study of the coherence between promoter activity and large-scale chromatin enviroment.) This comparison was very instructive since it revealed the existence of a strong correlation between the four prevalent chromatin states and the MRT, and this for both the pluripotent (H1 ES) and the differentiated (K562, GM12878, Nhdfaf) cell lines (Fig. 2.13) [53, 54]. The transcriptionally active euchromatin states EC1 and C1 replicate early in the S-phase in agreement with previous studies of open chromatin marks in human and mouse [36, 38, 147, 151, 157, 214]. The pluripotent bivalent EC2 state and the differentiated polycomb repressed C2 facultative heterochromatin state both replicate slightly later in mid-S phase as recently confirmed by the sequencing of nascent DNA strands synthetized at replication origins in human [138]. Note that this result contrasts with previous observation that at a few kb scale, the repressive chromatin mark H3K27me3 indeed highly correlates to late replication [38]. The silenced unmarked EC3 and C3 states as well as the pluripotent chromatin states EC4 prepared to heterochromatization and the HP1-associated heterochromatin state C4 all replicate much latter up to the end of S-phase. Interestingly, whereas (EC1, C1) and (EC2, C2) have a clearly different MRT, they have almost the same high mean GC content as expected for gene-rich states in high GC isochores (Fig. 2.13) [215–219]. In contrast, a definite correlation between MRT and mean GC content was observed for the late replicating chromatin states: when C3 replicates before C4 (K562, Nhdfad), C3 has a higher GC content and vice-versa when C3 (EC3) replicates after C4 (EC4) (GM12878, H1 ES)) (Fig. 2.13). As emphasized by Julienne *et al.* [54], there is however a major difference between MRT of pluripotent and differentiated cell lines. EC4 exhibits a much wider MRT distribution than C4 with

a non-negligible proportion of early replicating (MRT<0.5) 100 kb loci, namely 35.7% (H1ES) as compared to 5.5% (K562), 19.2% (GM12878) and 4.2% (Nhdfad). This is the confirmation of the highly dynamic character of pluripotent chromatin states that are sufficiently accessible and open to enable origin firing and early replication. In that respect, the "master" replication origins firing early in EC4 chromatin state at U/N-domain borders specific to H1 ES, were shown to play a fundamental role in the loss of pluripotency and lineage commitment [54]. The discussion of chromatin states repartition inside replication domains is reported to Chapter 4 (Section 4.2.3).

2.3 A dichotomic view of the topological and functional nuclear organisation

The DNA of eukaryotic cells is enclosed in the cell nucleus. Generally, eukaryotes have their genome organised in several separated chromosomes. Yet, even in separated chromosomes, the length of the longest DNA molecule in eukaryotic genomes far exceeds the diameter of the cell nucleus which is on average 5 μm for mammalian cells. For instance the longest human chromosome** of 280 Mb, which is almost 8.5 cm in length [83]. The length of eukaryotic genomes implies two contradictory imperatives [42]. The genome must be condensed in such a way that it fits inside the nucleus while being highly organised and accessible so that every nuclear function (*e.g.* replication, transcription) can take place efficiently. This high degree of organisation coupled to a tight compaction is obtained by the association of DNA with proteins, *e.g.* CTCF proteins form DNA loops that have various regulatory effects [169].

The most obvious level of organisation is the so-called chromosomes territories [1]. Furthermore, chromatin has a specific spatial distribution in the nucleus: active regions are at the center and gene deserts are attached to the nucleus periphery with lamina fibers [171]. Moreover, besides linear features of chromatin such as chromatin states (discussed in the previous section), DNA replication has been related to the 3D organisation of the nucleus. It seems that replication starts at the center of the nucleus and goes towards the periphery. During this progression, it can reorganise the 3D distribution of histone marks [202]. The replication occurs at discrete foci in the nucleus according to a specific spatio-temporal organisation [220, 221]. Throughout the S phase, new replication foci de novo assemble immediately next to the previously active ones [222–225], and a substantial segregation between early- and late-replicated loci has been observed [226–229]. In addition, from the observation that thousands of replication forks are only found in hundreds of discrete BrdU-labeled foci, it has been proposed that DNA sequences replicating at the same time could gather [5, 7, 226, 230–233], possibly through the anchoring of giant chromatin loops on the nuclear skeleton. Yet, details of chromatin folding was not accessible until recently with the emergence of chromatin conformation capture (3C) methodologies [234] as powerful strategies to analyze in detail the long-range folding of chromatin. In the following, we briefly recall the different levels of chromatin organisation and present the 3C methodologies allowing to assess the meso-scale structuration of the nucleus.

**The longest human chromosome is chromosome 1. Human chromosomes are numbered from the longest to the shortest.

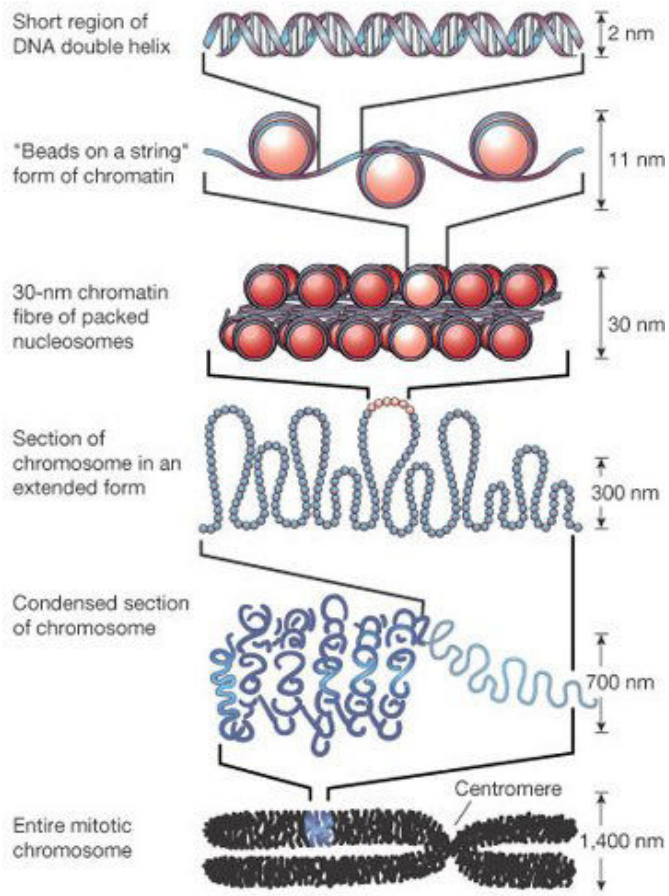


Figure 2.14. The organisation of DNA within the chromatin structure. The finest level of organisation is the nucleosome, in which two superhelical turns of DNA (a total of $\sim$ 150 base pairs) are wound around the outside of a histone octamer. Nucleosomes are connected to one another by short stretches of linker DNA. At the next level of organisation, the string of nucleosomes is folded into a fibre of about 30 nm in diameter, and this fibre is then further folded into higher-order structures. At structural levels larger than the nucleosome, and lower than the mitotic chromosomes, the details of folding are still uncertain (see section 2.4). (Redrawn from Ref. [235], and originally from Ref. [2]).

2.3.1 Hierarchical organisation of eukaryotic chromatin

Inside each human cell are two copies of the genome containing more than 3 billion base pairs. Fitting that much DNA in a tiny cell nucleus requires DNA compaction. The organised compaction must be highly dynamic. Indeed, the compaction fold is 10000 during mitosis (chromosome must be tightly condensed to enable their proper distribution between the two daughter cells) and "only" 300 in interphase[††]. This high degree of organisation coupled to a tight compaction is obtained by the association of DNA with proteins (chromatin). Chromatin is organised in successive layers of folding of increasing scale that are depicted in Fig. 2.14 [2]. Each layer has its functional relevance and carries regulatory informations [42]. As illustrated in Fig. 2.14, the packaging actually takes place at a number of scales:

- The DNA wraps around histone octamers to form the nucleosome (described in Section 2.2) leading to a "beads on string" fiber approximately 10 nanometers in width. The beads-on-string structure in turn coils into a 30-nm-diameter fiber that packs the nucleosomes more closely together.
- During cellular interphase, the 30-nm fibers is folded into higher-order loop structure to obtain a more compact structure that fits within the nucleus.
- During cell division, the chromatin is radically packed and condensed to form the metaphase chromosome as they are observed in karyotypes.

[††]The compaction fold is the ratio of actual end to end length of the chromosome in the nucleus over the length of the chromosome DNA laid out as a perfect DNA double helix.

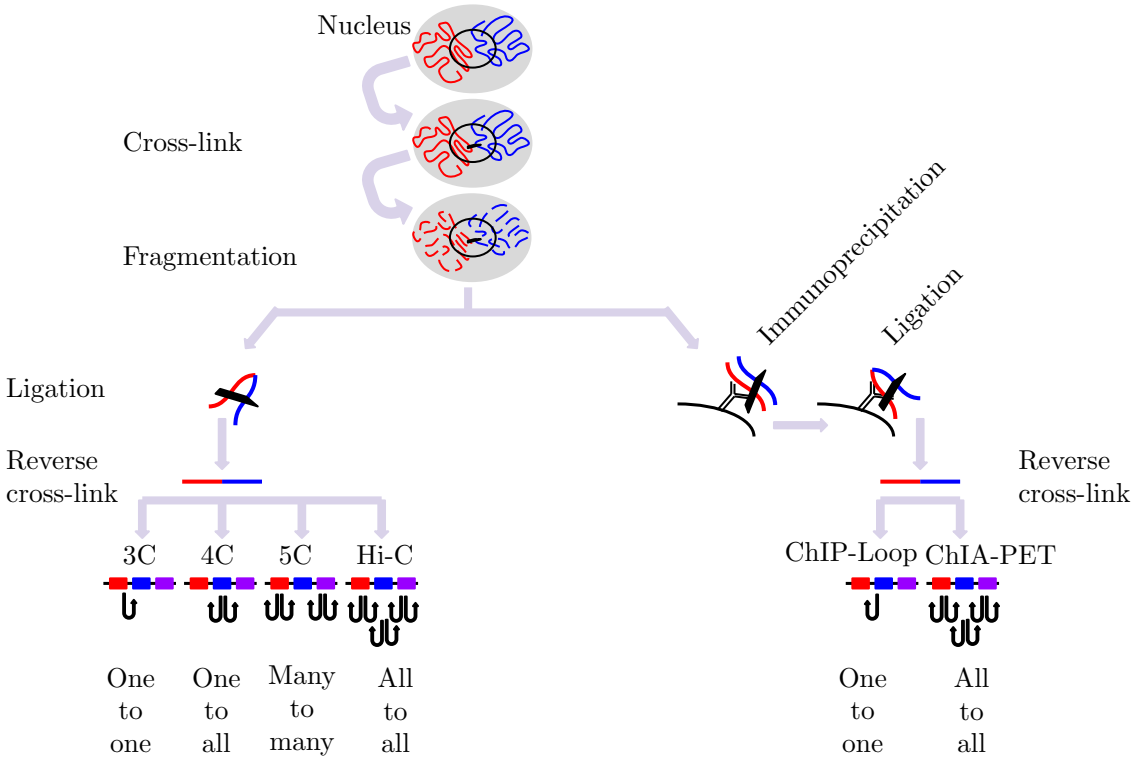


Figure 2.15. Schematic view of 3C and its derived methods. Initial common steps: cross-link and fragmentation followed by more specific manipulations depending on the method (see text for details) [236, 237].

Hence, DNA organisation in the nucleus is hierarchical. However, if the 10 nm fiber has been observed and well described, the 30 nm fiber and the higher-order loop structures are not well understood [236].

2.3.2 Probing the 3D nuclear meso-scale structuration using chromatin conformation capture methodologies

Recent development of chromosome conformation capture (3C) technology [234] and its high-throughput extensions, 4C (circular 3C) [238, 239], 5C (3C carbon copy) [240], and Hi-C (high-throuput 3C) [14], and methods combining 3C and ChIP (chromatin immunoprecipitation): ChIP-Loop (also called ChIP-3C) [241] and ChIA-PET (ChIP analyses by paired end tag sequencing) [241] has opened new perspectives in the study of DNA interactions. These techniques provide quantitative measurements of the interaction frequency between two selected loci (3C and ChIP-Loop), a selected loci and the rest of the genome (4C), multiple selected loci (5C) and genome-wide pairs of loci (Hi-C and ChIA-PET). These methods capture genome-wide contacts on the mega-base or even the tens of kilobase scales, and thus offer the opportunity to understand higher-order structures closing the gap between the atomic and chromosomal resolutions. In the following, we review briefly these experimental procedures. For more details on the 3C technologies, the interested reader is reffered to [236, 237].

The 3C technique consists in five steps (Fig. 2.15) [234]: The initial steps of 3C consist in fixing the DNA (cross-linking) using a fixative agent (formaldehyde is the most

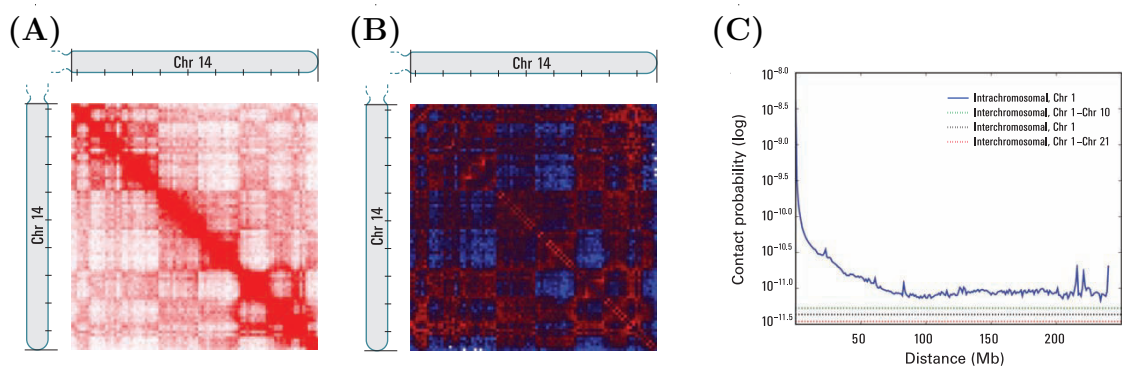


Figure 2.16. Hi-C data. (A) Hi-C contact map of the long arm of chromosome 14 showing intra-chromosome interactions. Each pixel represents all the interactions between two 1 Mb loci; intensity correspond to the total number of reads (Tick marks appear each 10 Mb). (B) The observed/expected ratio matrix shows loci with either more (red) or less (blue) interactions than would be expected, given their genomic distance (the ratio range from 0.2 to 5). (C) Probability of contact decreases as a function of genomic distance on chromosome 1 (reaching a plateau at $\sim$ 90 Mb). Interchromosomal contacts (dashed lines) differs for different pairs of chromosomes. Interchromosomal interactions are depleted relatively to intrachromosomal interactions. Adapted from [14].

common). This fixation gives a snapshot of the *in vivo* interactions within chromatin. This is followed by cutting the DNA with a restriction enzyme such as HindIII and NcoI used for most human datasets. These two enzymes recognise 6 bp which sets the intrinsic resolution of the method. The third step is to ligate in diluted conditions to promote ligation between cross-linked fragments instead of random fragments, so that two distal DNA loci brought together by chromatin looping are more likely to be ligated. With high temperature, the formed cross-links will reverse, resulting in linear chimeric DNA fragments with specific restriction ends (known as 3C library). To assess the frequencies of an interaction of interest, polymerase chain reaction (PCR)[‡‡] is used. Hence, for this method an *a priori* on the existence of the interaction is needed since their detection depends on a pair of primers, each corresponding to one partner. This technique is useful for specific loci of interest but has very limited throughput.

The 4C procedure allows to study the contact between a selected loci of interest and all the other regions of the genome (Fig. 2.15) [238, 239]. It enables to overcome the 3C limitation because it characterises all the interactions of one locus ("view-point") with all the others. For this, 4C has the same initial steps as the 3C until the reverse cross-link, then it needs a second restriction digest and circular ligation resulting in small DNA circles (4C library). The circularised DNA is used as template for inverse PCR. Using the "view-point" specific primers, result in the amplification of sequences containing this site, allowing to determine all genomic regions that have been ligated ligated to it.

5C technique is very similar to 3C, starting with cross-linking, fragmentation, ligation and reverse cross-link, followed by ligation mediated amplification and quantilation (Fig. 2.15) [240]. In fact, a set of multiplex primers connected to two universal primers are used for the PCR reaction. The PCR products are then sequenced to determine

[‡‡]PCR is a molecular biology technology used to amplify the quantity of a DNA fragment across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

Cell line	Cell Type	Enzyme	Accession	Reference
H1 ES	Embryonic stem cell	HindIII	GEO website GSE35156	[16]
IMR90	Foetal lung fibroblast	HindIII	GEO website GSE35156	[16]
GM06990	Lymphoblastoid	HindIII	GEO website GSE18199	[14]
GM06990	Lymphoblastoid	NcoI	GEO website GSE18199	[14]
K562	Myelogenous leukimia (blood cancer)	HindIII	GEO website GSE18199	[14]
HeLaS3	Derived from HeLa	HindIII	www.ebi.ac.uk E-MTAB-1948	[243]

Table 2.3. Hi-C analysed data. First column indicates the cell line, the second column the cell type to which it belongs, the third column the restriction enzyme used in the experiment, the fourth column is the accession number and the last column is the original publication reference number.

ligated sequences. Depending on the design of the multiplex primers, this method allow to capture all the pairwise interactions between many loci simultaneously.

The study of genome-wide chromatin interaction became possible with the Hi-C procedure (Fig. 2.15) [14]. Hi-C also starts with cross-linking and digestion, but before ligation, the ends are filled with biotine. The resulting ligation products consist of fragments marked with biotin at the junction which will enable selective purification of chimeric DNA ligation. Shearing the DNA and selecting the biotin containing fragments with streptavidin beads, result in the Hi-C library of all the interactions between all pairs of loci. Then the library is analysed by using massively parallel pair-end sequencing.

3C based-methods can be easily coupled with chromatin immunoprecipitation (ChIP) to build a chromatin. ChIP is carried out after cross-linking and sonification to enrich for fragments bound by a particular protein of interest. ChIp-Loop is the 3C version of methods based on ChIP to detect interactions between two selected loci (Fig. 2.15) [241]. ChIA-PET is the genome-wide version of ChIP-Loop, it consists in introducing a linker sequence in the junction of the two DNA fragments during nuclear proximity ligation to access the interactions of DNA fragments that are tethered together by protein factors (Fig. 2.15) [241].

Clearly, these various 3C based technologies offer the possibility to focus the available sequencing effort to the chromatin interactions of interest depending on the biological question asked.

In this thesis, we are interested in the characterisation of the genome structuration of DNA irrespectively to the presence of particular chromatin associated proteins. We thus take advantage of the availability of Hi-C data [14, 16, 242–245] (Table 2.3) for a global analysis of contact matrix between all genomic loci in order to extract information about the 3D structure of the chromatin in human. Figure 2.16 (A) shows an example of Hi-C data matrix representing the relative frequencies of intra-chromosomal physical interactions between pairs of loci. We mainly focused on intrachromosomal interactions that are by far the most frequent ones (Fig. 2.16 C) [14]. In this representation, the result of a 3C is a pixel in the Hi-C matrix, 4C results are lines of the Hi-C matrix and 5C corresponds to intersections of different lines corresponding to loci recognised by the multiplex.

Hi-C normalisation:

The complicated experimental procedures involved in Hi-C, can induce different biases and experimental artifacts [246]. Sources of biases in Hi-C interaction counts can be various [247, 248]:

- Read depth per region: One expects to observe equal read coverage across the genome since Hi-C is proned to be an unbiased assay of genomic structure. However, factors such as mappability (ability to map reads uniquely): regions with low level of unique sequences are usually unmappable and hence underrepresented in the final interaction reads [248].
- Nucleotide composition of DNA [249, 250]: Fragments with extreme GC content are underrepresented in the final interaction reads. PCR and/or deep sequencing introduce additional biases by favouring reads with GC content of about 45% [248].
- The length of restriction fragments: in theory, as the number of positions accessible to fixation along a restriction fragment increase with its size, the interaction probability should increase linearly with restriction fragment size [248].
- The size of the cross-linked fragment ends: long and short fragments may have variable ligation efficiencies. Optimal ligation takes place when both cross-linked fragments have intermediate length [247].

In many cases the effects of such biases may decrease correlation between replicate experiments. Several techniques were developed to limit the introduction of unwanted biases [247, 248, 251, 252]. After a step of filtering low and high interacting fragments (relatively to the mean) the different proposed methods consist in the following procedures:

- *A probabilistic model for Hi-C contact maps normalization* [247]: this integrated multiplicative probabilistic model amounts to compute the prior probability of contact between two fragments taking into account their mappability, length and GC content. The algorithm estimates maximum-likelihood model parameters given the empirical raw contact maps (Fig. 2.16). This approach has been presented as capable to remove the majority of systematic biases. However, its computation cost is high.
- *HiCNorm via Poisson regression* [251]: this method consists in estimating the bias effects due to length and GC content while fixing the mappability as a Poisson offset. This approach is similar to the previous one, but with fewer parameters. It requires less computation power.
- *Iterative algorithms* [248, 252]: these algorithms consist in normalising rows and columns iteratively until the Hi-C matrix is symmetric again. For example, *Iterative Correction* is a simple parameter free method using the expected frequency of every pair of regions. It ensures a uniform coverage profile *i.e.* "equal visibility" of each locus in the corrected contact map. This approach is widely used because it is assumed to account for unknown biases.

☞ Conclusions we draw from the normalised data in this manuscript do not differ significantly from the ones drawn with raw data (different results are discussed in the second

Cell line	Not sequenced	Low interacting	High interacting	Total
IMR90	1792	1100	97	2989
H1 ES	1734	1286	20	3040
K562	1732	565	584	2881
GM06990	1730	990	212	2932
HeLaS3	1836	986	130	2952

Table 2.4. Masked data. We remove from the original data (28688 loci) low and high interacting fragments along with fragments corresponding to not sequenced regions of the genome. Fixing the thresholds to $low = max(0, \bar{c} - 2\sigma)$, and $high = min(0.99L, \bar{c} + 2\sigma)$, where $\bar{c}$ is the mean number of interactions in a matrix and L the chromosome size.

part of this thesis). Along this manuscript, we filter the intrachromosomic data before any analysis and we remove low and high interacting fragments that are likely to introduce noise. For each considered Hi-C dataset (Table 2.3) and for each chromosome, we compute the mean $\bar{c}$ and the standard deviation σ of the total intrachromosomal interaction count per loci n_i (sum over the Hi-C matrix line (Fig. 2.16 A)). Using the thresholds to $low = max(0, \bar{c} - 2\sigma)$, and $high = min(0.99L, \bar{c} + 2\sigma)$, where L is the chromosome size (in pixel), we only retain loci where $n_i \in [low, high]$, removing 10% of the data (6% correspond to unsequenced fragments, ~ 2 to 4% correspond to low interacting fragments and $\sim 2\%$ correspond to high interacting fragments Table 2.4).

2.3.3 Dichotomous compartmentalisation of the nucleus

Meaningful interpretation of this huge amount of Hi-C data depends on effective and robust statistical analysis. The first step is to create a matrix comparing the observed number of reads between two loci or bins, to the expected number of reads between these two bins. The observed interaction matrix is the collection of the numbers of interactions between each pair of bins from the mapped data (Fig. 2.16 A). The expected number of interactions between two bins is derived from the experimental data by taking the total number of observed interactions at a genomic distance s divided by the total number of possible interaction at distance s across all the chromosomes. The observed interaction matrix can then be normalised by the expected interaction matrix to generate an Observed/Expected matrix (Fig. 2.16 B).

Examination of Hi-C matrices shows that high interactions occur between close neighbours (high values along the diagonal in Figure 2.16 A). In fact, the frequency of intrachromosome interactions decays as a power-law of the distance (Fig. 2.16 C) [14]. Interchromosomes interactions are depleted relatively to intrachromosome contacts; depending on the pair of chromosomes, interactions can be more or less frequent consistently with the chromosomes territories observed by FISH [14]. In the pioneering paper on the Hi-C data [14], the authors looked at the correlation matrix between the i^{th} column and the j^{th} row of the matrix. When normalising the data for the distance, they saw that loci that are nearby in space have correlated interaction profiles sharing interacting neighbours. From the sign of the first eigenvector of the distance normalised Hi-C correlation matrix, they revealed the existence of 2 compartments. Comparison of those compartments with

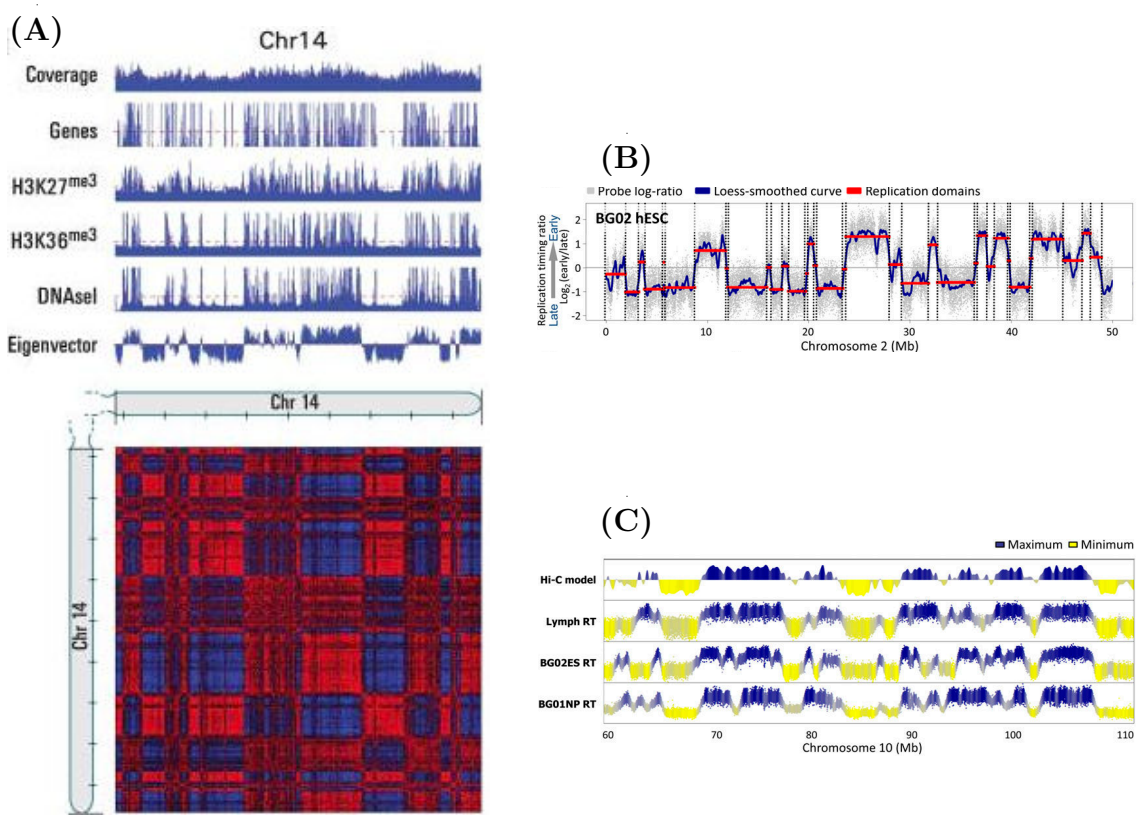


Figure 2.17. Dichotomic view of the topological and functional organisation of the nucleus. (A) Correlation map of chromosome 14 at a resolution of 100 kb. The principal component (eigenvector) of the distance normalised Hi-C correlation matrix (bottom) correlates with the distribution of genes and with features of open chromatin DNAseI, active H3K36me3 and repressive H3K27me3 marks. The correlation ranges from -1 (blue) to +1 (red). Adapted from [14]. (B) MRT profile along a 50 Mb fragment of human chromosome 2. Data shown are average of two replicate hybridisations (dye-swap) for human embryonic stem cell line BG02. DNA synthesised early *vs* late during S-phase was hybridised to an oligonucleotide microarray and the $\log_2$ ratio of early/late signal for each probe (probe spacing 1.1 kb) across the genome was plotted *vs* the map position (Section 2.1.3.2). Grey dots represent raw data, blue line represent loess-smoothed data, red lines correspond to identified early and late CTRs with their boundaries (dotted vertical lines). (C) First eigenvector of the distance normalised K562 Hi-C correlation matrix (Hi-C model) and replication timing in different cell lines. Negative values are represented in yellow and positive values are in blue. (B) and (C) are adapted from [38].

genetic and epigenetic features showed that one compartment correlates with high gene density, polycomb repression, gene activity and open chromatin while the second one corresponds to heterochromatic gene deserts (Fig. 2.17 A).

Along side, initial studies of MRT profiles in mouse [147, 148] and human [35, 36] have revealed the presence of mega-base scale regions with similar timing, called constant timing regions (CTRs) (Section 2.1.3.2), replicating either early or late in the S-phase by coordinated activation of multiple origins (Fig. 2.17 B). In good agreement with previous studies in *Drosophila* [190, 253], the CTRs present some correlation with epigenetic modifications [154]. Early CTRs are gene-rich, high GC isochore like regions that tend to be enriched in open chromatin marks [147, 254]. In contrast late CTRs are gene desert, low GC isochore-like regions that are mostly associated with repressive heterochromatin

marks [38, 147].
The comparison of MRT and Hi-C data (sign of the eigenvector of the distance normalised correlation matrix) [37, 38] (Fig. 2.17 C) suggested a dichotomic picture where early and late initiation of replication occur in separated nuclear compartments. This suggests the existence of long-range chromatin interactions between early CTRs and late CTRs but not between early and late CTRs which appear to be segregated in separated nuclear compartments of early replicating open chromatin and late replicating closed chromatin.

Note that the representation of intra-chromosomal interaction frequencies in a matrix form displays a finer level of structuration, characterized by diagonal blocks of length 10^5-10^6 bp: the interaction between regions of a same block is of high frequency relatively to the weaker frequency between regions of different blocks [16]. These blocks are linked to the functional organisation of the genome. To carry on this research and assess objectively the structure, various approaches have been developed (Chapter 6) [16, 20, 255–259]. However, the genome structuration expands over a wide range of scales [260] and is likely to involve nested structures. Only the method proposed in [258] is built in order to identify areas at different scales of observation. These methods will be detailed in Chapter 6.

2.3.4 Chromatin states, nuclear compartimentalisation and constant replication timing regions

In this section, we provide the link between the prevalent chromatin states discussed in Section 2.2.2 and the dichotomic description of the nucleus: early replicating regions are confined in open and active chromatin while late replicating regions are in close and inactive chromatin. In fact, once mapped to the genome (Fig. 2.18), the organisation of the four prevalent chromatin states EC1, EC2, EC3 and EC4 in the pluripotent H1 ES cell line does not differ so much in their genome coverage as also observed for the chromatin states C1, C2, C3 and C4 in the considered differentiated cell lines (see Table 1 in Ref. [54]). However, when looking at the length distribution of blocks of adjacent 100 kb loci in the same chromatin state, whereas EC1, EC2, EC3 and EC4 blocks have similar length distributions, the HP1-associated heterochromatin state C4 has a block length distribution that displays a fat tail not observed in the C1, C2 and C3 block length distributions which explains that, for example, in K562, the mean C4 block length ($\bar{L}$ = 882kb) is significantly larger than the mean block length of C1 ($\bar{L}$ = 327kb), C2 ($\bar{L}$ = 191kb), C3 ($\bar{L}$ = 438kb) [53, 54]. This peculiar length property of C4 blocks is shared by all differentiated cell lines except GM12878 where C3 blocks are larger ($\bar{L}$ = 576kb) as compared to C4 blocks ($\bar{L}$ = 276kb) (see Table 3 in Ref. [54]). Interestingly, as pointed out in Refs. [53, 54], for all differentiated cell lines as well as for the ESC line H1 ES, the association of C1+C2 (resp. EC1+EC2) on one side and of C3+C4 (resp. EC3+EC4) on the other side, results in Mb scale blocks of similar length distributions [53, 54]. These large blocks ($L \geq 1.5$ Mb) of active and inactive chromatins respectively correspond to early and late CTRs that are well conserved between pluripotent and differentiated cell lines (Fig. 2.18), the larger the size of the block, the higher the conservation level [54]. The gene rich, high GC C1+C2 (EC1+EC2) chromatin blocks (e.g. from $\sim$ 72 to 78 Mb in Fig.2.18), replicate very early in the S phase by the coordinated activation of multiple origins mainly located in C1 (resp. EC1) active loci whereas C2 (resp. EC2) loci are more likely replicated passively from forks coming from neighboring C1 (resp. EC1) loci. The

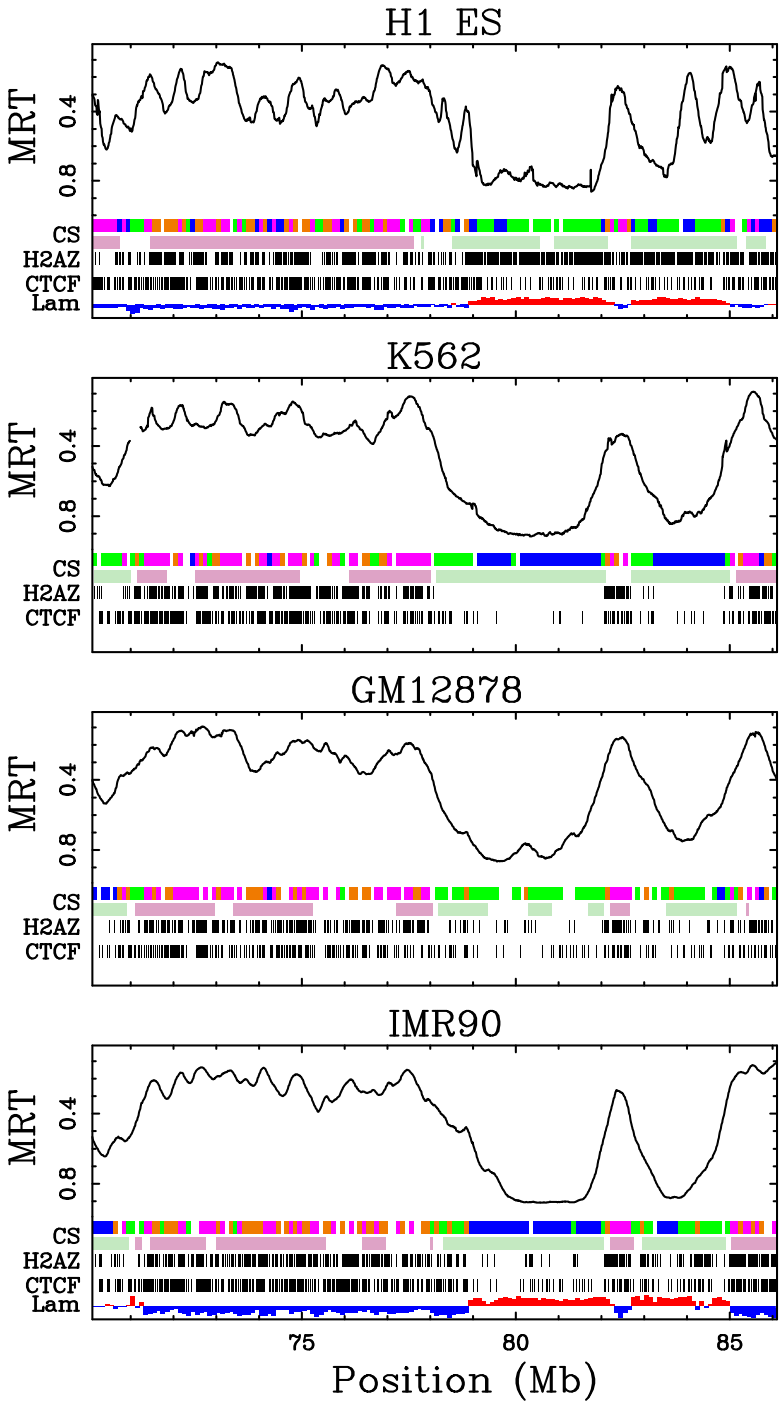


Figure 2.18. Chromatin state organisation and the dichotomic view. MRT profiles along a 16 Mb long fragment of human chromosome 11 from top to bottom in H1 ES, K562, GM12878 and IMR90. Below the MRT profiles are shown the spatial distribution of EC1, EC2, EC3 and EC4 chromatin state loci in H1 ES (top panel) and of C1, C2, C3 and C4 chromatin state loci in K562, GM12878 and Nhdfad (bottom panels). The chromatin state of each 100 kb window is represented using the same color coding as in Fig. 2.13. EC1+EC2 (resp. C1+C2) blocks are also represented in light pink, and EC3+EC4 (resp. C3+C4) in light green. At the bottom of the plot, intervals significantly enriched in H2AZ and CTCF are represented in black; in red ($\log_2$(binding ratio)> 0) and blue ($\log_2$(binding ratio)< 0) is also reported, when available, the lamina B1 binding profile in SHEF-2 (surrogate for H1 ES) and TIG3 (surrogate for IMR90). Chromatin states, H2AZ and CTCF data in Nhdfad are used as surrogates for IMR90 .

gene-poor, low GC C3+C4 (EC3+EC4) chromatin blocks (e.g. form $\sim$ 79 to 82 Mb in Figs 2.18), on the contrary replicate very late by the almost synchronous firing of multiple origins. Note that these results are quite consistent with the statistical model of Desprat *et al.* [36] where MRT is predicted from the distance to the nearest active promoter. Let us also emphasize that the largest EC3+EC4 chromatin blocks in H1 ES ($L_{max} \sim 5$ Mb) turns out to be significantly shorter than in differentiated cells ($L_{max} \sim 12$ Mb) [53, 54]. This replication domain consolidation induced by differentiation [44, 147, 148] results from an early replication initiation zone in ESCs that no longer fires early in somatic cells leading to the merging of the two neighboring EC3+EC4 chromatin blocks into a larger C3+C4 chromatin block in the differentiated cell lines (see for instance MRT peak specific to H1 ES at position $\sim$ 84 Mb in Figure 2.18).

☞ Hence, the combined analysis of genome-wide replication timing data and epigenetic data has revealed some 1D organisation of the genome into functional domains: early replicating open and active chromatin *vs* late replicating close inactive chromatin. This dichotomic picture correlated with the nucleus compartimentalisation into the structural A/B compartments.

2.4 From 1D chromatin state organisation in MRT U-domains to 3D chromatin folding

Even though the DNA sequence may play a role in the positioning of origins [137, 179], there is no consensus sequence for replication origins in metazoan. Origin positioning is cell line specific: even if there was a consensus sequence, an additional regulation mechanism would be needed [179]. Therefore, mechanisms that position and control the time of firing of origins must be epigenetic and linked to chromatin structure [144, 261–265]. In other words, there is a clear interest to study the relationship between DNA replication, the 3D chromatin interaction patterns, and genome-wide epigenetic mark distributions. The organisation of the human genome into replication U/N-domains covering about half of the genome provides us with an original point of view to unify the interplay between replication, transcription, epigenetic modifications and nuclear organisation (Section 2.1). "Master" replication initiation zones (MaOris) at U/N-domains border are major actors in the spatio-temporal replication programme [51, 69, 70], acting as insulating regions between chromatin domains of independent expression and duplication [44, 51]. Moreover the U/N-domain borders are significantly enriched in the insulator binding protein CTCF [44] which has been suggested to contribute to the formation of chromosomal hubs of interactions across chromosomes [210]; they thus play a central role in the genome's 3D architecture. Therefore, it is interesting to address to which extent a tertiary chromatin-structure counterpart to these functional replication U/N-domains exists.

The original structural partitioning of the chromosomes into A/B compartments derived from intra-chromosomal Hi-C interactions used distance normalised Hi-C data (Section 2.3.3) [14] giving as much weight to the very-distant interactions (over tens of mega-bases) as to the more local contacts (below few mega-bases), despite the fact that the latter are the most frequent ones (Fig. 2.16 C). Analysis of Hi-C data at higher resolution allowed the identification of blocks of high interactions along the diagonal, the so-called *Topologically Associated Domains* (TADs; discussed in Chapter 6) [16, 21].

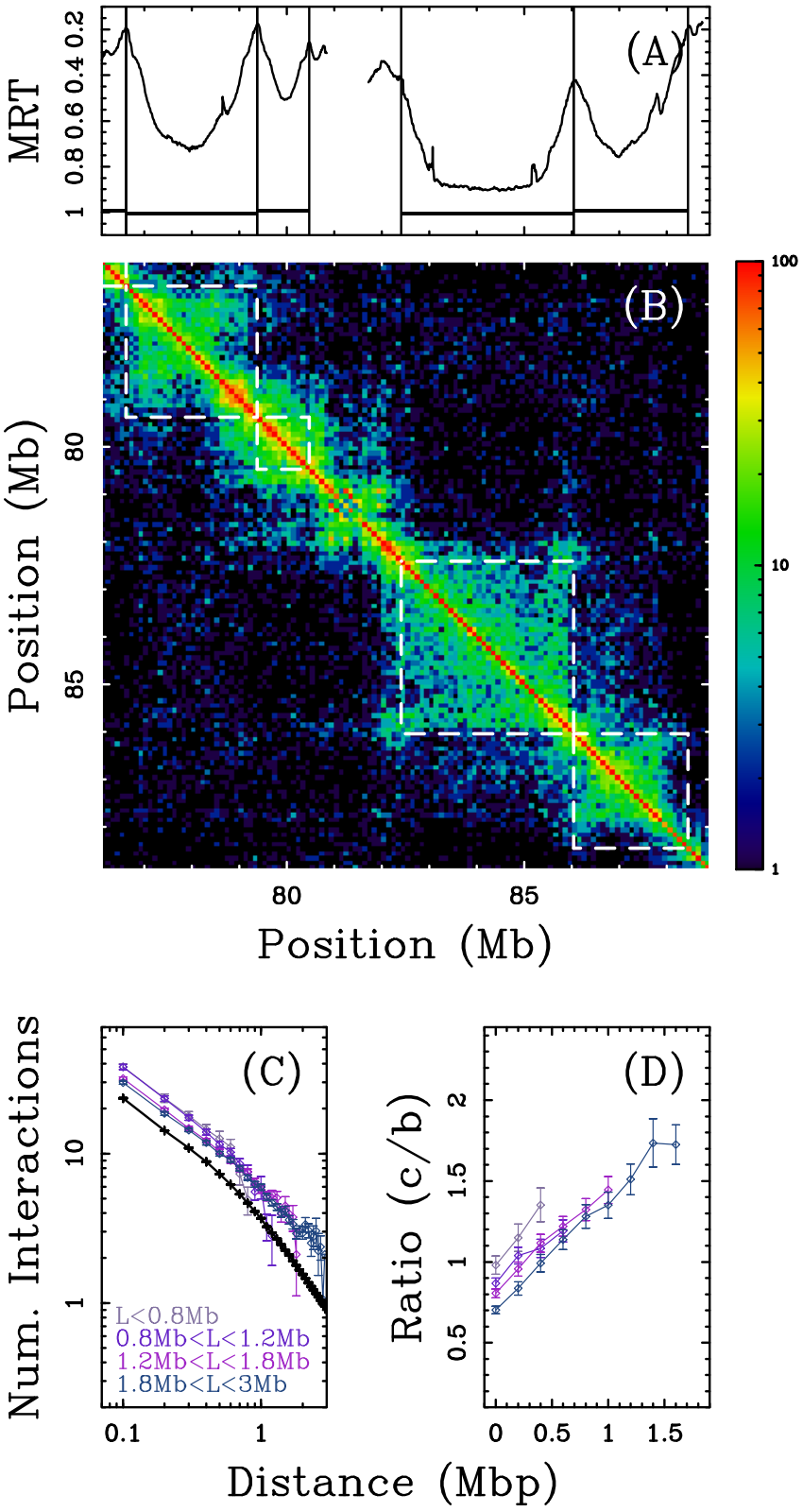


Figure 2.19. Are replication U-domains structural units of Hi-C matrix? See next page.

Figure 2.19. Are replication U-domains structural units of Hi-C matrix? (A) MRT profile from K562 cell line along a 12.8 Mb fragment of human chromosome 10; the vertical lines correspond to the borders of 4 detected replication timing U-domains (horizontal bars). (B) Hi-C interaction matrix corresponding to intrachromosome interactions on the corresponding 12.8 Mb long fragment of human chromosome 10, as measured in the K562 cell line. Each pixel represents all interactions between all pairs of 100 kb loci; intensity corresponding to the total number of reads is color coded according to the colormap (right) after normalisation to a constant number of reads (set to 10 millions). The dashed squares correspond to the 4 detected U-domains. (C) Number of interactions between two 100 kb loci versus the distance separating them (logarithmic scales) as computed genome wide (black) or in replication U-domains only, for four U-domain size categories: L<0.8Mb, 0.8Mb≤L<1.2Mb, 1.2Mb≤L<1.8Mb and 1.8Mb≤L<3Mb. (D) Ratio of the number of interactions between two 100 kb loci that are inside the same U-domain at equal distance from its center (c) and the number of interactions between loci in different U-domains at equal distance from a U-domain border (b), versus the distance between them; the color coding is the same as in (C).

Complementarily, as illustrated in Figure. 2.19, when focusing on interactions between loci separated by short genomic distances ($\lesssim$ 10 Mb) over which the contact probabilities are the highest (Fig. 2.16 C) [14], it was shown that replication U-domains correspond to matrix square-blocks of enriched interactions (Fig. 2.19 B) [44]. This clearly suggests that the segmentation of the genome into replication U-domains coincides to some extent with the segmentation into TADs. Consistently with previous results, the matrix square-block structure underlines that early replicating zones that border a U-domain have a high contact probability as the signature of 3D spatial proximity. However, the matrix square-block structure also signifies a high contact probability of the two early replicating borders with the late replicating U-domain center and sparse interactions between loci in separate U-domains. In other words, locally we do not observe a structural segregation between early- and late-replicating loci and U-domain borders appear to correspond to structural insulating barriers. Further examination of the average behaviour of the intrachromosomal contact probability as a function of genomic distance for the complete genome corroborated these observations [44]: the mean number of interactions between two 100 kb loci of the same U-domain decays when increasing their distance as observed genome-wide (Fig. 2.19 C) (further discussed in Chapter 5). However, when comparing the contact probability between loci inside a U-domain to the contact probability between loci lying in neighboring U-domains (Fig. 2.19 D), it was observed in K562 cell line that the latter is higher than the former for distances smaller than the characteristic size ($\sim$ 300 kb) of the open chromatin structure at U-domain borders [44, 51]. Above this characteristic distance, the tendency is reversed and the ratio significantly increases above 1 (Fig. 2.19 D), suggesting a correspondence between structural barriers and U-domain borders. This observation is strengthened by the observed enrichment in CTCF [44] at U/N-domain borders, that is known to be involved in chromatin loop formation conditioning communication between transcriptional regulatory elements [18, 169, 208, 209]. The Hi-C data do not always allow the direct visualisation of the chromatin contacts for chosen loci, principally because of insufficient sequencing depth to cover all the combinatorial possibilities (Section 2.3.2). This led the host team to explore the structural interaction partners of 10 *viewpoints* distributed along a large region (20 Mb) of the human chromosome 5 using the high resolution 4C methodology [55]. The MRT profile of

the region of interest presents large fluctuations with well defined U-domains that are very well conserved across 7 cell lines for which replication timing profiles were available (as illustrated Fig. 2.10 (C) for MRT profile along a fragment of chromosome 10). The 10 viewpoints were chosen to be located either in the middle of a U-domain or on a well-defined timing peak a the border between two successive U-domains. This high-resolution analysis nicely confirmed the results obtained with Hi-C data and emphasized the existence of specific interaction between U-domains borders even when separated by more than 10 Mb [55].

☞ These initial results prompted us to go further into the characterisation of replication domains as fundamental units of chromatin tertiary structure, in different cell lines where Hi-C data and replication timing data are available (Tables 2.1 and 2.3[§§]). This constitutes the main objective of this thesis. From the analysis of genome-wide MRT data, chromatin marks profiles and chromatin conformation capture data, we try to establish a link between the 1D organisation (Section 2.2) and its 3D structure (Section 2.3) of chromatin in relation with the spatio-temporal programme of DNA replication (Section 2.1). We use a unified classification of the genome that combines the replication U-domains for $\sim 50\%$ of the genome and early and late CTRs for the other half [53], we show that when the distance between two MaOris exceeds 3 Mb, a late CTR emerges in the central region of the replication domain whose length increases with inter-origin distance giving rise to what we will call replication split-U-domains (Chapter 4). We then discuss their role in the 3D organisation of chromosomes (Chapter 5). Finally, we explore methodologies allowing to systematically extract TAD-like structural domains from chromatin conformation data alone. In particular, we use graph theory (Chapter 3) to develop an approach that does not use the prior knowledge of the chromosome linear structure. Given this robust methodology to delineate structural domains, we compare replication and structural domains in different cell lines and we evaluate their stability across cell lines (Chapter 6).

[§§]Note that in order to include human embryonic stem cells in our analysis, we choose to jointly analyse H1 ES Hi-C data [16] and BG02 MRT data [151].

Chapter 3

Graph theory background and signal processing on graphs

Graphs have become extremely useful as a representation of a wide variety of complex systems in social sciences, biology, computer sciences and many other area of fundamental and applied sciences. Along this chapter, we recapitulate the basics of graph theory and the definitions of four centrality measures that will be useful in the context of this thesis. Then we present some recent developments in the area of signal processing on graphs providing us with generalised tools to identify and exploit the community structure of the underlying graph.

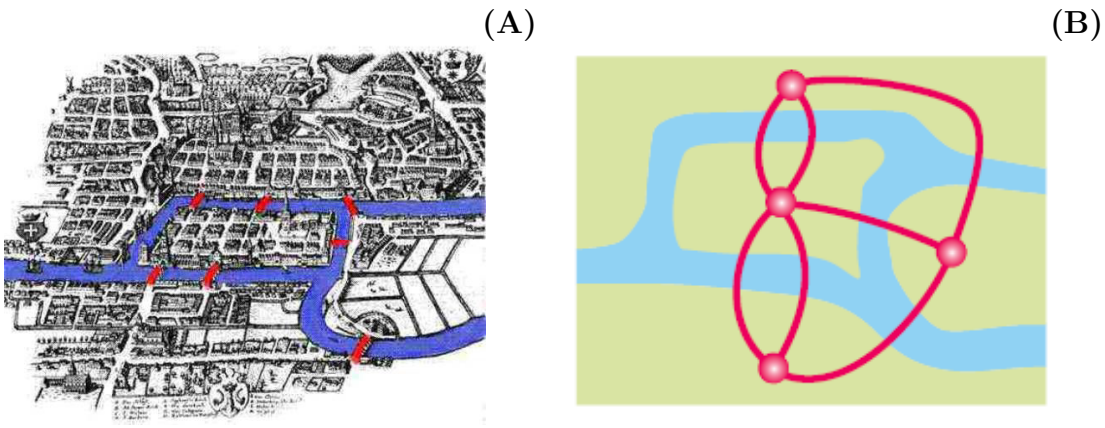


Figure 3.1. Konigsberg bridge problem. (A) The map of Konigsberg during Euler's time showing the actual layout of the seven bridges (red) over the Preger river (blue). (B) Illustration of Euler's model of the seven bridges of Konigsberg with four nodes for the land masses and seven edges for the bridges.

3.1 Origins of graph theory

The history of graph theory may be specifically traced back to 1736 [266, 267], when the Swiss mathematician Leonhard Euler solved the *Konigsberg bridge problem**. The Konigsberg bridge problem originated in the city of Konigsberg, formerly in Germany but now known as Kaliningrad and part of Russia, located on the Preger river. The city had seven bridges, which connected two islands with the main-land (Fig. 3.1 A). People staying there always wondered whether there was any way to walk over all the bridges once and only once.

In 1736 Euler came out with a solution to this problem by formulating the problem in terms of graph theory. He proved that it was not possible to walk through the seven bridges exactly one time. He abstracted the case of Konigsberg by eliminating all unnecessary features. He drew a picture consisting of "dots" that represented the land-masses (nodes) and the line-segments representing the bridges that connected those land-masses (edges) (Fig. 3.1 B). This simplified the problem to great a extent. Indeed, the problem can be merely seen as the way of tracing the graph with a pencil without actually lifting it. Euler not only proved that it is not possible to cross all the bridges once and only once but he came up with a theorem that explained why it was not possible and what should be the characteristic of the graph so that its edges could be traversed exactly once. The theorem was based on the *degree of nodes* (See section 3.2.1). Euler proposed that any given graph can be traversed with each edge traversed exactly once if and only if it has zero or two nodes with odd degrees. This must be so because otherwise, you must cross the same bridge twice to walk farther. The graphs following the above condition are called, *Eulerian circuit or path.*

*Konigsberg bridge problem was originally published in *Commentarii Academiae Scientiarum Petropolitanae* **8**, 1741, pp. 128-140. English translations are available in [266, 267].

3.2 Graph theory

Since the original work of Euler, networks have been widely used as a representation of complex systems[†] in social sciences [58, 59], biology [60–62], computer sciences [62, 63], engineering and many other area of fundamental and applied sciences [64–66]. Every person that has friends, acquaintances or any social ties is a part of a social network [58, 59]. Every person that has a mobile phone or an email adress is a part of a communication network [268, 269]. Every metro or train station or airport is a part of a transportation network [270]. Every animal is a part of the food web network [271]. Every neurone is a part of a neuronal network [272]. These networks can be represented as graphs [56, 57, 273, 274], mathematical objects where the elements of study (persons, metro stations, animals, neurones...) are represented as nodes and the connections between them constitute the links of the graph. Once the graph is formed, the analysis consists in extracting "useful" information from this representation. In a social network, a "useful" information can be "which person is the most influencial/important in the network". This notion of influence or importance can be objectively quantified using centrality measures [58, 59, 64–66] (Section 3.3). Another interesting information one can ask about a social network is if there exist groups of persons that share the same interests; this can be mathematically addressed as a problem of community detection [275] (Section 3.6). In that context, this chapter is intended to summarize a set of graph theory methods of particular relevance for the analysis of chromatin conformation data. Along with the definitions and methods selected to be presented in this chapter, illustrative examples will be used to facilitate the understanding. Some elements are presented on a toy network of 128 nodes (Fig. 3.8) that is built with a hierarchical structure at 4 scales. At the smallest scale, 16 groups of 8 nodes are fully linked together. Then pairs of groups are linked together by 4 edges, resulting in a structure of 8 groups of 16 nodes. Again each pair of groups is connected with 4 edges, resulting in a structure of 4 groups. Each pair of these groups of 32 nodes is connected by 2 edges, resulting in 2 groups of 64 nodes, at the largest scale, that are connected together with only one edge. This way of construction ensures that, at each scale, there is more edges inside each group than in between the groups. The general form of this graph with n nodes is a hierarchical benchmark introduced in [276] and it is widely used to test the multi-scale community mining tools.

3.2.1 The basics of graph theory

This first section gives classical definitions of the graph theory that can be found for example in [56, 57, 273, 274].

A *graph* $G = (V, E)$ is defined by a set of *nodes* (or *vertices*) V and a set of *edges* $E \subset V^2$ that links the nodes to each other. In other words, a graph is a set of nodes connected to each other by links. The number of nodes is noted $n = |V|$ and the number of edges is noted $m = |E|$. In principle, two nodes can be connected via many links called *parallel links*. A node can also be connected to itself forming a *loop*. Note that the two vertices connected by an edge are said to be *incident* with the edge, and *vice versa*. Two vertices which are incident with a common edge are *adjacent*, as are two edges which are incident with a common vertex, and two distinct adjacent vertices are *neighbours*. The

[†]A complex system is a system with non-trivial topological features, modelling real life systems.

set of neighbours of a vertex a in a graph G is denoted by $N_G(a)$. An *isolated* node in a graph, is a vertex with no incident edge to it. A graph is *connected* if it has no isolated nodes, otherwise it is *disconnected*.

An edge e_{ab} can be oriented from an origin node a to a target node b. The graph is said to be *directed* in that case. When "node a is connected to node b" implies that "node b is also connected to node a" the graph is called *not directed*. For example, let us consider a graph of two persons (nodes a and b). If we suppose that the set of edges reflect the fact that two persons know each others: if a knows b, then b knows a, and it is a not directed graph. However if we consider that nodes are animals in a food web network, and that edges represent the fact that an animal feeds on an other. Then if a feeds on b, b does not always feeds on a and the graph is hence oriented.

3.2.1.1 Drawings of graphs

Graphs are named so because they can be conventionally represented graphically (Section 3.1). Indeed, the graphical representation of graphs may help us to understand them better. Each vertex is indicated by a point and each edge by a line joining the points corresponding to its ends (Fig. 3.2 A). There is no single correct way to draw a graph [277]; the relative positions of points representing vertices and the shapes of lines representing edges can be chosen arbitrarily. Note that in some cases there exists a "natural" layout for nodes for example when nodes are geographical locations.

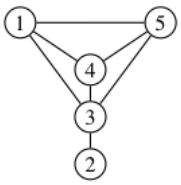
One objective in graph drawing is to find representations that convey specific information. One possible solution consists in grouping strongly connected nodes next to each other while "pushing away" the weakly connected ones in order that the graph layout enlightens the connectivity structure. One strategy to automatically obtain such graph layouts rely on "dynamical" approach where the graph drawing software modifies an initial vertex placement by iteratively moving the vertices according to a system of force based on physical metaphors related to systems of springs or to molecular mechanics. In other words this strategy relies on the notions of attraction and repulsion by analogy with the dynamics of interacting particles. In this analogy, nodes act as magnets that repulse each other while the edges behave as springs attracting vertices they connect. This will cause relative displacements making the system evolves dynamically during a transitional phase to eventually stabilize reaching a steady state. The search of this balance has been the subject of several studies leading to dynamical models known as *Force Directed Placement (FDP)* [278, 279]. All these models adopt the same basic principle but differ in the way of defining the forces. Note that the obtained stationnary state is not unique but as we will see in Chapter 5 (Section 5.2.2) this way to represent the graph indeed conveys meaningful information about the underlying structure.

3.2.1.2 Adjacency and incidence matrices

Although drawings are a convenient way of representing graphs, they are not suitable for storing graphs in a computer or for applying and developing mathematical methods to study their properties. In that respect, matrix representations of graphs are useful tools [56, 57, 273, 274].

The adjacency matrix A_G (or simply A) is a $n \times n$ matrix where the entries $a_{ij} = 1$

(A)



(B)

$$A = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix}$$

(C)

$$U = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Figure 3.2. Simple graph example with corresponding adjacency and incidence matrices. A graph of 5 nodes and 7 edges (A), and corresponding adjacency 5×5 matrix A (B) and incidence 5×7 matrix U (C).

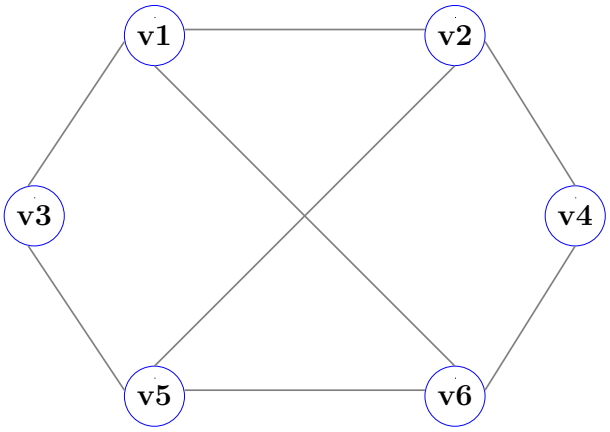


Figure 3.3. A hexagone graph with paths and cycles. Between the nodes v_1 and v_6 there are several paths, for instance $\{v_1, v_2, v_4, v_6\}$ and $\{v_1, v_3, v_5, v_6\}$ (odd paths of length 3), or simply $\{v_1, v_6\}$ (odd path of length 1). Between the nodes v_1 and v_4 there is no 1-path, however a shortest path is $\{v_1, v_2, v_4\}$ (even path of length 2), this is not the unique shortest path because another 2-path exists $\{v_1, v_6, v_4\}$. The distance between v_1 and v_6 is 1 and the distance between v_1 and v_4 is 2. A cycle in this graph is $\{v_1, v_2, v_4, v_6, v_5, v_3, v_1\}$. The diameter of this graph is 3, and its radius is 5.

if the nodes x_i and x_j are connected, and $a_{ij} = 0$ otherwise (Fig. 3.2 B); *i.e.* $a_{ij} = 1$ if the nodes x_i and x_j are adjacent, hence the name adjacency matrix. Another type of matrix can represent a graph: the incidence matrix U_G (or simply U), a $n \times m$ matrix where $u_{ik} = 1$ if the node x_i is incident to the edge e_k, and $u_{ik} = 0$ otherwise (Fig. 3.2 C).

In the following, we will be using the commonly used adjacency matrix because it is usually much smaller than the incidence matrix (in most cases the number of nodes is normally smaller than the number of edges) which makes it more convenient.

Note that the adjacency matrix of a non directed (resp. directed) graph is always symmetric (resp. not symmetric).

☞ This definition of graphs with either existent or abscent links are *non weighted* or binary (as their adjacency matrix is formed by 1 or 0). However, when assigning a weight to each edge the graph becomes *weighted* and the weighted adjacency matrix $W = (w_{ij})$ has not null values when x_i and x_j are connected and 0 otherwise: the higher the values of w_{ij}, the stronger the links between them. For example, in a social network, where nodes are people and edges represent phone calls between those people, a weight can be the duration of the call or the number of times people call each other.

3.2.2 General definitions

A graph is *finite* if both its vertex set and edges set are finite. The *null graph* is the graph with no vertices (and of course no edges in this case).

A *complete graph* is a simple graph in which any two vertices are connected. A graph is *bipartite* if its vertex set can be partitioned into two subsets X and Y so that every edge has one end in X and one end in Y; such a partition (X,Y) is called *bipartition* of the graph and X and Y are its parts.

A *path* is a simple (sub-)graph whose vertices can be arranged in a linear sequence in such a way that two vertices are adjacent if they are consecutive in the sequence, and are non adjacent otherwise (Fig. 3.3). A *cycle* or a *circuit* is a simple graph whose vertices can be arranged in a cyclic sequence in such a way that two vertices are adjacent if they are consecutive in the sequence, and are non adjacent otherwise (Fig. 3.3). Note that the vertices constituting a path or a cycle are distinct. However, in the definition of the Eulerian path or cycle, vertices can be visited more than once. The *length* of a path or of a cycle is the number of its edges. Paths and cycles of length k are called *k-path* or *k-cycle*, respectively; the path or cycle is *odd* or *even* according to the parity of k (Fig. 3.3). A *shortest path* or *geodesic path* is a path between two nodes of a graph with the minimal number of edges (Fig. 3.3). When there is not any path between two nodes, the distance between them is $+\infty$. The generalisation of path length in the case of weighted graphs requires to define the length l of an edge depending on its weight w. Interpreting an edge weight as the length ($l = w$) implies that higher weights are interpreted as weaker ties. If higher weight values indicate stronger ties, for instance in a communication network with frequency of contact as edge weights, it is necessary to define the edge length as a decreasing function of the weight. For example, an edge length can then be defined by subtracting the weight from an upper bound b ($l = b - w$ where b can be chosen as b= $\max_{i,j}(w_{ij}) + 1$), or by taking the inverse ($l = 1/w$) or an exponentially decreasing function ($l = e^{-w/w_o}$, $w_0 > 0$). A 3-cycle is often called *triangle*, a 4-cycle a *quadrilateral*, a 5-cycle a *pentagone*, a 6-cycle a *hexagone* and so on. The *diameter* of a graph is the longest distance between any two nodes of the graph, and its *radius* is the minimum of the longest path length between all pairs of nodes of the graph (Fig. 3.3). Note that in disconnected graphs the radius and the diameter are infinite.

☞ In this thesis, we only consider finite, not null, connected, not directed and *simple graphs*, *i.e.* graphs with a finite number of nodes ($n > 0$), where there is no isolated nodes, the edges are not oriented and with no parallel links nor loops.

3.3 Several faces of power in a network: centrality measures

As previously mentioned, the main point of representing complex systems in term of graphs, is to extract useful information from that representation such as "which vertex is the most *important* in the network". Obviously, the answer to this question depends on the meaning behind the word "important". Centrality measures [58, 59, 64–66, 280] are real-valued functions assigning to each vertex in a network some value capturing the *importance* of the nodes.

In the following, we introduce some selected centrality measures (this list is far from being exhaustive[‡]) that we will be using in Chapter 5. The definitions based on the adjacency matrix A of the unweighted graph, are immediatly extended for the weighted graphs using the weighted adjacency matrix W instead of A. Hence, we consider $G = (V, E)$, a graph with $|V| = n$, $|E| = m$, M its adjacency (A) or weighted adjacency (W) matrix. We note x_i, $i = 1, ..., n$, the nodes of the graph and e_{ij} the edge between the nodes x_i and x_j.

3.3.1 Degree centrality

Consider for example, a social network. Obviously, a person who knows a lot of people has a privileged position in the network. This is the simplest definition one can give to a centrality, *an important node is a node with a lot of connections*. The *degree centrality* is formally defined as follows [289]:

$$C_D(x_i) = \sum_{j=1}^{n} m_{ij}. \tag{3.1}$$

Note that C_D is also known as the *degree* of a node for unweighted graphs, and the *strength* of a node for weighted graphs. For the sake of comparision between different graphs, the degree centrality of a node can be normalized by $n-1$, the maximum possible number of connections for a node, or the maximum weight for weighted graphs. A graph where all the nodes have the same degree is said to be *regular*.

Degree centrality C_D measures the opportunity to receive information flowing through the network. Let us point out that it only takes into consideration the local structure of the graph by looking at the direct neighbours of a node, which can make it in some cases it is not really informative.

3.3.2 Closeness centrality

What if someone in a society knows a lot of people that are from his/her family. This person has many connections but is "far" from what's happening beyond the circle of his/her family. When we think of *an important node as one that is "close" to and can communicate quickly with the other nodes in the network*, we use the *closeness centrality* defined as follows [289]:

$$C_C(x_i) = \frac{1}{\sum_{j=1}^{n} \delta(x_i, x_j)}, \tag{3.2}$$

where $\delta(x_i, x_j)$ is the shortest path length between x_i and x_j. The inverse of C_C is called the *farness* of a node.

The most central nodes according to closeness centrality can "quickly" interact with other nodes because their average distance to other nodes is small. This measure is preferable to degree centrality because it does take into account not only the direct connections to

[‡]In fact, there are several other centrality measures based on the ones we define in this thesis [58, 59, 64–66, 280]. The difference between these variants are generally weak [281–287]. A critical review of this domain is presented in [288].

the reference node but also all the shortest paths leading to it. It results in taking into account a larger part of the graph topology.

3.3.3 Betweenness centrality

In a social network a person can be a link between people that don't know each others. So if we think that *an important node lies on a high proportion of shortest paths between other nodes, playing the role of a bridge in the network*, we get the definition of *betweenness centrality* [289]:

$$C_B(x) = \sum_{y \neq x \neq z} \frac{\sigma_{yz}(x)}{\sigma_{yz}}, \tag{3.3}$$

where $\sigma_{yz}(x)$ is the number of shortest paths from node y to node z through node x and σ_{yz} is the number of shortest paths from y to z, where x, y, $z \in V$. It is also common to use the normalized betweenness centrality given by:

$$\tilde{C_B}(x) = \frac{C_B(x) - \min_{x \in V}(C_B(x))}{\max_{x \in V}(C_B(x)) - \min_{x \in V}(C_B(x))}. \tag{3.4}$$

Nodes with high betweenness centrality have control over the flow of information in the network by facilitating, hindering, or even altering the communication between other nodes.

3.3.4 Eigenvector centrality

Finally, let us consider two nodes in a social network, Paul and John having both 5 friends, but where Paul's friends have only few friends whereas John's friends have a lot of friends. In that case, John is likely in a "better", more central position than Paul. This idea, *an important node is connected to other important nodes*, leads to define the *eigenvector centrality* of a node x_i, also known as *spectral centrality* to be proportional to the sum of the eigenvector centralities of its neighbors in a self-consistent manner [290]:

$$C_S(x_i) = \frac{1}{\lambda} \sum_{j=1}^{n} m_{ij} C_S(x_j). \tag{3.5}$$

Defining the vector $C_S = (C_s(x_1), ..., C_s(x_n))$, the above equation can be rewritten in a matrix form as:

$$\lambda C_S = M C_S. \tag{3.6}$$

Hence, C_S is an eigenvector of the (weighted) adjacency matrix associated with the eigenvalue λ. In general many vectors can be a solution of Equation (3.6). However, the additional requirement that all the entries in the eigenvector must be positive, implies the use of the eigenvector of the largest eigenvalue (Perron-Frobenius theorem).

The eigenvector centrality can be understood as a refined version of the degree centrality in the sense that it self-consistently takes into account the centrality of neighbour vertices. As stated in the example above, the degree centrality will rank two nodes (Paul and John) equally based on their immediate neighborhood but the eigenvector centrality will favor the nodes with more "global" effect (John in this example).

3.3.5 Correlations between centrality measures

Centrality measures are often classified according to the type of "influence" a vertex with high centrality has on the other vertices: immediate effects, mediative effects, and total effects [291]. The closeness and the degree centralities capture immediate effects, the betweenness centrality captures mediative effects and eigenvector centrality captures total effects. In that sense, these measures can be seen to be complementary rather than competitive because each one captures a different characteristic of the graph. Nevertheless, a question remains open: how correlated are the centrality measures? Many studies have examined this question and depending on the data studied, different results have been reported.

Degree, closeness, betweenness and eigenvector centralities were empirically compared in [284, 285]. Data sets in [284] were the most various: film actors network, scientist collaborations network, the internet of both autonomous system and routers, and protein interaction networks. Data sets in [285] were mostly sociometric diffusion networks in a variety of settings such as *(i)* diffusion of family planning practices in Korea where women in rural villages were asked to nominate five other village residents from whom they asked advice about family planning, and *(ii)* or diffusion of farming practices in Brazil where farmers were asked to name their three best friends, the three most influential people in their community and the three most influential farmers in their community... The first study [284] found the heighest correlation between the degree and the betweenness centralities while the second one [285] found the degree and the eigenvector centralities to be the most correlated followed by the degree and the betweenness centralities. Batool *et al.* [287] also compared the above four centralities with the eccentricity[§] on standard data sets such as the Zachary's Karaté club and the neural network of *C. elegans* where closeness centrality and eccentricity had the heighest correlation followed by the degree and the eigenvector centralities.

Other studies compared two or three of the above centralities to other definitions of centrality [281–283, 286]. Under conditions of random error, systematic error and incomplete data, comparison of degree, closeness and betweenness centralities showed a high correlation of the degree and closeness centralities while betweenness centrality remained relatively uncorrelated with all the other ones [281]. Interestingly, degree, betweenness and closeness among others were shown to be highly correlated on a subset regarding relationships between CEOs, clubs and boards [283]. Degree and betweenness were also found to be correlated when compared with six other centralities in a network of individuals connected through participation in a HIV risk behavior [282]. Finally, a disease spreading network analysis [286] found that closeness and betweenness centralities were uncorrelated.

☞ The four centrality measures described in this section: degree, closeness, betweenness and eigenvector centralities capture different properties of the graph and provide complementary informations about the node "role" in the graph as recapitulated in Table 3.1, where we also provide a tentative interpretation for situations where two centrality measures provide an opposite diagnostic about the importance of a node.

[§]Vertices with high eccentricity C_E are at short maximum distances to every other reachable vertices, $C_E(x_i) = 1/d_i$, where d_i is the maximum distance of the node x_i to all its neighbours.

	High Degree	High Closenness	High Betweenness	High Eigenvector
Low Degree	Number of connections	Key player tied to few important/active players	Ego's few ties are crucial for network flow	Node has few ties but to important players
Low Closeness	Node embedded in a cluster that is far from the center	Proximity to the center	Very rare! Would mean that ego monopolises the ties from a small number of people to many others	Node tied with key players far from others
Low betweenness	Ego's connections are redundant, communication bypasses him/her	Probably multiple paths in the network, ego is near many people, but so are many others	Lies between other on theirs geodesic paths	Ego's connections can be bypassed
Low Eigenvector	Connected to "isolated" nodes	Key player in the middle of the graph with no immediate link to important players	Node lies on different paths crucial for network flow but with no immediate connections to key players	Connected to key players

Table 3.1. Interpreting centrality measures. Recapitulation of the centrality measures meaning (diagonal) and signification of combination of low and high values for pairs of measures.

3.4 Signal processing on graphs

In applications such as the ones cited at the beginning of this chapter (social networks, electricity networks, transportation networks...), one may be interested in analysing the distribution of data values residing on the vertices of the graph, such data sets have been described as *signals on graphs*. Naturally, one can wonder what are the best strategies to characterise and to extract efficiently the information from these signals on graphs. Motivated by the impact of Fourier transform and wavelet transform in classical signal analysis, recent developments in the area of signal processing on graphs provide us with operators to process signals on graphs and generalize classical transforms to the graph setting. Taking advantage of these developments, we use in this thesis, a multiscale community detection method (See section 3.6) anchored in signal processing on graph. An introduction to this area is thus required. Moreover, signal processing on graphs uses spectral graph theory as a tool to define frequency spectra and expansion bases for graph Fourier transforms. In this section, we start with a short introduction about basic definitions and notations from spectral graph theory (Section 3.4.2). We then extend classical signal processing techniques to the graph setting via the definition of graph operators (Section 3.4.3) in the context of our application of interest: community detection (Section 3.6).

3.4.1 Defining a signal on a graph

A signal (or a function) $F : V \longrightarrow \mathbb{R}$ on the vertices of a graph can be seen as a vector $F \in \mathbb{R}^n$, where the i^{th} component $f(i)$ of the vector F represents the signal value at vertex x_i. A signal on a graph represents some data related to the network. For example, the gender in a social network can be seen as a signal on the graph where the signal can either takes the value 1 if the person is a female or -1 if the person is a male (Fig. 3.4).

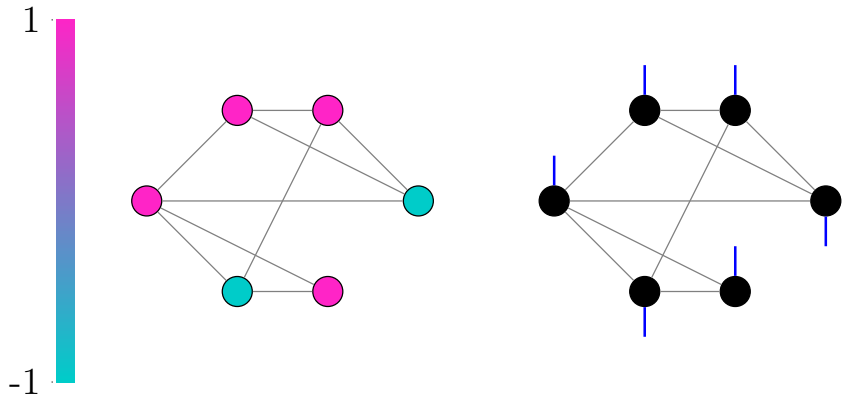


Figure 3.4. Signal on graphs. A graph representing friendship relationships between 6 persons. We assign a simple signal on the graph corresponding to the gender of the person: $\pm$ 1 if the person is a female or a male, respectively. Left pannel, the signal is color coded according to the colormap on the left. Right pannel, the signal is visualized by the bars oriented up for positive values and down for negative values.

Such signals can be directly visualized on the graph by coloring each node x_i according to f_i, or by drawing over each node a line proportional to the signal value (Fig. 3.4).

3.4.2 Graph spectral domains and Graph Fourier Transform

Since the beginning of graph theory, matrix theory and linear algebra were used to analyse the adjacency matrices of a graph. In the past few years, many developments in spectral graph theory have emerged especially with geometric interpretations. Historically, spectral graph theory focused on constructing, analysing and manipulating graphs. It has been used for the construction of expander graphs [292], graph visualisation [293], spectral clustering [294], graph coloring [293] and numerous other applications [295]. In the area of signal processing on graphs, spectral graph theory has become a tool to define frequency spectra and expansion bases for graph Fourier transforms.

Two visions have been adopted to construct the theory of graph signal processing. In both cases, when considering the circular graph we recover some of the well known classical results. The first vision adopts the oriented circular graph [296] as a starting point while the second one [297] uses the non oriented circular graph. In the first approach [296], the Fourier basis is constructed from the calculus of the generalized eigenvectors of the Jordan decomposition of the adjacency or weighted adjacency matrix. In the second approach [297], the Fourier basis is built from the eigenvectors of the Laplacian matrix of the graph as explained below. In that case, we will see that the eigenvalues correspond to the square of the Fourier mode, and in contrast with the first vision, they can be directly associated with the notion of frequency. In the following, we only present the construction on the Laplacian matrix [297] from wich spectral graph wavelet are constructed.

We first need to define some spectral graph theory elements that can be found in [298] for example. Let $G = (V, E)$ be an undirected and connected graph with $M = A$ or W, the adjacency or weighted adjacency matrix corresponding to the graph. The graph *Laplacian matrix* L is defined as:

$$L = D - M, \tag{3.7}$$

where D is a diagonal matrix whose element $d_{ii} = \Sigma_j m_{ij}$, is the degree (or weighted degree *i.e.* strength) of the node.

☞ Note that the Laplacian matrix can also be found under it normalized form:

$$\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}. \tag{3.8}$$

☞ In the case of a weighted graph, the non-normalised graph Laplacian is also known as the combinatorial graph Laplacian.

In all cases, L is a real symmetric matrix (because M is symmetric), and hence, it has a complete set of orthonormal column eigenvectors, denoted $\{\chi_l\}_{l=0,..,n-1}$. These eigenvectors have associated positive eigenvalues $\{\lambda_l\}_{l=0,..,n-1}$. Zero appears as an eigenvalue with multiplicity equal to the number of connected components. Since we are only interested in connected graphs, we consider the Laplacian eigenvalues ordered as $0 = \lambda_0 < \lambda_1 \leq \lambda_2 ... \leq \lambda_{n-1} := \lambda_{max}$. The set of λ_i's is called the *spectrum* of L (or spectrum of the associated graph G). We note $\chi = (\chi_0|\chi_1|\cdots|\chi_{n-1})$.

Note that on the circular graph of n nodes (numbered 0,...,$n-1$) the node x_i ($i \in [\![1,n-2]\!]$,) is connected to the nodes x_{i-1} and x_{i+1} and the node x_0 is connected to the nodes x_1 and x_{n-1}; in other words, the circular graph is nothing but the regular line where the first and the last nodes are connected to ensure periodic boundary conditions. L is the classical laplacian operator *i.e.* the discrete second derivative operator, and its eigenvectors are the usual discrete Fourier vectors. This constitutes the fundamental analogy between the classical case and the graphs. Actually, the main intuition behind signal processing on graphs is that it is considered as a generalization of the "classical" discrete signal processing. In fact, a classical discrete signal f of size n is nothing but a signal defined on a circular graph of size n.

Let us recall that the *classical Fourier transform*

$$\hat{f}(\xi) = \int_R f(t) e^{-2\pi i \xi t} dt, \tag{3.9}$$

is the projection of a function f on the complex exponentials, which are the eigenfunctions of the one dimensional Laplace operator:$-\Delta(e^{2\pi i \xi t}) = -\frac{d^2}{dt^2} e^{2\pi i \xi t} = (2\pi\xi)^2 e^{2\pi \xi t}$.
Similarly, we define the *Graph Fourier Transform*, (GFT), $\hat{F}$ of any signal $F \in \mathbb{R}^n$ on the vertices of G, as the projection of F on the eigenvectors ($\{\chi_l\}_{l=0,..,n-1}$) of the graph Laplacian:

$$\hat{f}(l) = \sum_{i=1}^{n-1} f(i) \chi_l(i). \tag{3.10}$$

Equation (3.10) can be written as:

$$\hat{F} = \chi^\top F, \tag{3.11}$$

where $\chi^\top$ is the transpose of χ.

In the same manner, the *classical inverse Fourier transform*:

$$f(t) = \int_R \hat{f}(\xi) e^{2\pi i \xi t} dt, \tag{3.12}$$

is the reconstruction of f as a weighted sum of the Fourier vectors, which is mimicked in the graph setting as:

$$f(i) = \sum_{l=0}^{n-1} \hat{f}(l)\chi_l(i), \tag{3.13}$$

that can be written in matrix notation as:

$$F = \chi \hat{F}. \tag{3.14}$$

Application to the circular graph:
As mentioned previously this construction of graph Fourier and inverse Fourier transforms relies on the observation that for the circular graph this construction leads to the classical Discrete Fourier Transform (DFT). Indeed, the adjacency and Laplacian matrices associated with the circular graph are:

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ 1 & 0 & \cdots & 1 & 0 \end{pmatrix}, \quad L = \begin{pmatrix} 2 & -1 & 0 & \cdots & -1 \\ -1 & 2 & -1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ -1 & 0 & \cdots & -1 & 2 \end{pmatrix}.$$

L is a special case of circulant matrix [299], allowing to easily compute its eigenvectors and associated eigenvalues (that we denote respectively μ and α to differentiate them from the eigenvectors and eigenvalues of any graph). A straightforward calculus leads to:

$$\alpha_m = 2(1 - \cos(\frac{2\pi m}{n})), \quad m = 0, ..., n-1. \tag{3.15}$$

Notice that $\alpha_j = \alpha_{n-j}$, $\forall j \in [1, n-1]$. If n is odd, there is only 1 eigenspace of size 1, the one associated with $\alpha_0 = 0$, and the constant eigenvector all its components equal to $\frac{1}{\sqrt{n}}$. If n is even, there is two eigenspaces of size 1, besides the one associated with $\alpha_0 = 0$, there is also the one associated with $\alpha_{n/2} = 4$[¶], and to the real eigenvector with all its components equal to $\frac{(-1)^m}{\sqrt{n}}$. All the other eigenvalues are of multiplicity 2 and the associated eigenspaces are defined by two vectors that are not unique (one can choose the cosine and sine). These eigenvectors corresponds to the columns of the DFT. Note that each eigenvalue α_m correspond to the square of the Fourier frequency (column m).

☞ A key point of this analogy is that it provides an equivalence between a signal defined on the nodes of the graph (which have no natural order and non-trivial links between them) and a signal defined on the spectrum graph, which is ordered ($\lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \lambda_{n-1}$).

To sum up, the fundamental analogy to define the graph Fourier operators (Equations (3.11) and (3.14)) resides on the fact that on the circular graph, the graph Laplacian is the classical discrete Laplacian operator and its eigenvectors are the Fourier vectors. Hence, on any graph, the eigenvectors χ_l of the Laplacian matrix L will be used as the Fourier vectors. Importantly, the graph Fourier modes χ_l have to be recomputed for

[¶]The greatest eigenvalue of the not normalised Laplacian is 4 and it is equal to 2 for the normalised one [298].

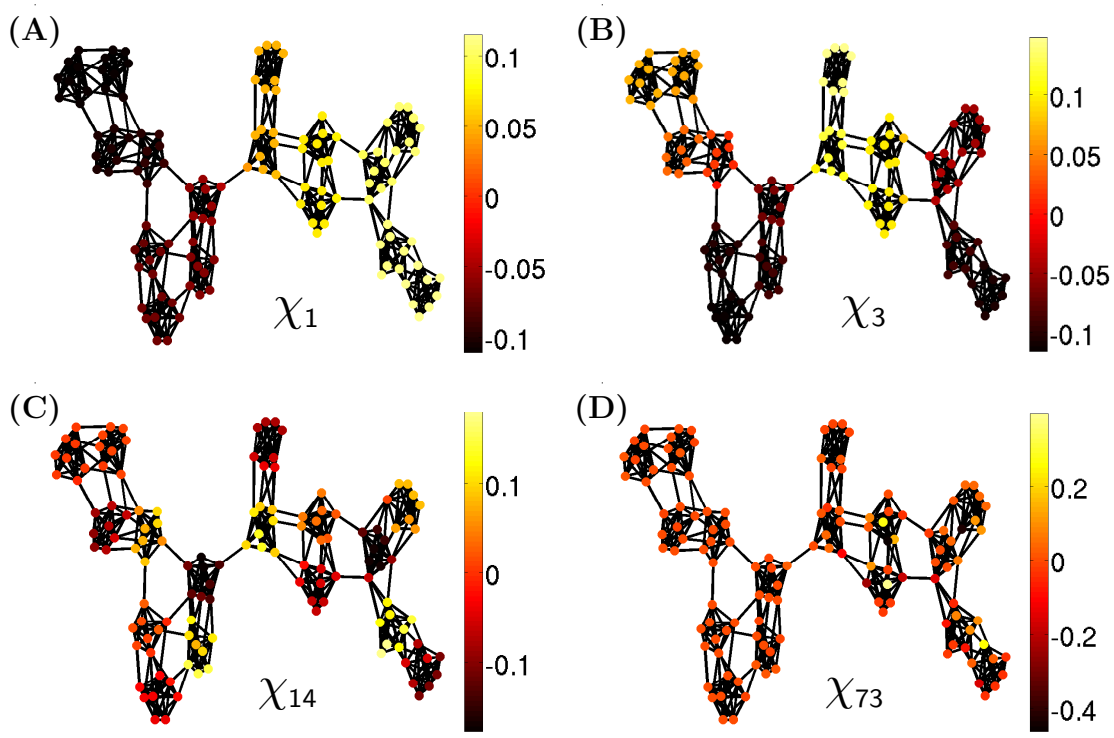


Figure 3.5. Graph Fourier modes. A toy graph of 128 nodes (See Fig. 3.8) with different Fourier modes color coded using the color map at the right of the graph.

each graph so that they in fact convey information about the graph topology. Figure 3.5 illustrates some Fourier modes on our toy graph (See page 59). One can already remark that the first few eigenvectors are very informative for community detection as we will discuss further in Section 3.6. For instance, partitioning the graph presented in Figure 3.5 according to positive values and negative values of χ_1, leads to 2 meaningful communities (Fig. 3.5 A).

3.4.3 Generalised operators for signals on graphs

One important point in this chapter is to define the graph wavelet transform for community detection on graphs. In this section, we present a way to generalise the classical signal processing operators to the graph settings. A detailed description of the classical operators can be found for example in [300], and their generalisation can be found in [297].

Classical filtering of a function f by a filter h is defined as a convolution in the direct space:

$$(f * h)(t) = \int_{-\infty}^{+\infty} f(\tau)h(t-\tau)d\tau. \tag{3.16}$$

Convolution (Equation (3.16)) relies on the translation and time reversal of the filter h which are difficult to define directly in the graph setting because of the lack of natural ordering of the nodes. However, convolution in the direct space corresponds to multiplication in the Fourier space:

$$\widehat{(f * h)}(\xi) = \hat{f}(\xi)\hat{h}(\xi). \tag{3.17}$$

Mimicking this property, convolution of the graph signal F by the graph filter H can be defined as the inverse Fourier transform (Equation (3.14)) of the product of the graph

Fourier transform of F and H:

$$F * H = \chi(\hat{H}.\hat{F}), \tag{3.18}$$

where $\hat{H}$ and $\hat{F}$ are the graph Fourier transform of H and F respectively and . stand for the component wise column multiplication.

Defining the diagonal matrix $\hat{H}_D$ as:

$$\hat{H}_D = \begin{bmatrix} \hat{h}(1) & 0 & \cdots & 0 \\ 0 & \hat{h}(2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{h}(n) \end{bmatrix}, \tag{3.19}$$

and using (Equation (3.11)), Equation (3.18) can be written as:

$$H * F = \chi \hat{H}_D \chi^\top F. \tag{3.20}$$

Equation (3.20) illustrates how one can generalise the notion of linear filtering for signals on graphs using graph Fourier transform.

Translation of a signal f can also be written using the convolution operator:

$$T_{t_0} f(t) = f(t - t_0) = (\delta_{t_0} * f)(t), \tag{3.21}$$

where δ_{t_0} is the Dirac function centered on t_0. Let Δ_k be the Dirac function on a graph centered on the node x_k: $\Delta_k(i) = 1$ on node x_k and $\Delta_k(i) = 0$ otherwise. Following Equation (3.18), we can then define a translation operator of graph signal F to a node x_k as:

$$F^{(k)} = \chi(\hat{F}.\hat{\Delta}_k). \tag{3.22}$$

Noting that $\hat{\Delta}_k = \chi^\top \Delta_k$ (Equation (3.11)), straightforward calculations leads to:

$$F^{(k)} = \chi \hat{F}_D \chi^\top \Delta_k. \tag{3.23}$$

Since the sequence matrix $(\Delta_1|...|\Delta_n)$ is the identity matrix, we can obtain the translations of signal F to all nodes using the following matrix operation:

$$(F^{(1)}|...|F^{(n)}) = \chi \hat{F}_D \chi^\top. \tag{3.24}$$

Equation (3.24) provides another example of the power of using Fourier space to transpose classical signal processing operators to the graph setting.

3.5 Spectral graph wavelets and the graph wavelet transform

We introduce in Appendix A, the continuous wavelet transform (WT) for applications to genomic data. In this section, we review the construction of graph wavelets and of graph wavelet transform with application to Hi-C data analysis. In fact, different approaches have been adopted to define wavelets on graphs. Wavelet definition can be based on the notion of diffusion on the node neighbourhood [301, 302], or using lifting [303, 304], or

filter banks [305, 306]. Following the logic of Section 3.4, we present here the construction of graph wavelets using the graph spectral domain [307]. Spectral graph wavelets [307] have been shown to be well adapted to the problem of community mining [308, 309]. They rely on the diagonalisation of the graph Laplacian as the spectral graph partitioning methods (see Section 3.6). As discussed in Section A.1, the definition of the classical WT given by Equation (A.1) can be interpreted as a space-scale expansion of a signal in terms of a family of *daughter wavelets* which are constructed from a single function, the "analysing wavelet" ψ, by means of translations and dilations. However, at a fixed scale s, Equation (A.1) can also be interpreted as the convolution of the signal with the wavelet ψ_s centered on 0 and derived from the analysing wavelet ψ as:

$$\psi_s(t) = \frac{1}{\sqrt{s}}\psi(\frac{t}{s}). \tag{3.25}$$

Assuming that the L_2 norm of ψ is 1, the $1/\sqrt{s}$ normalisation guaranties that the same is true for ψ_s. Importantly, given the spectral perspective of our construction, the Fourier transform of ψ_s can be written as follows:

$$\begin{aligned}\hat{\psi}_s(\xi) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{s}}\psi(\frac{t}{s})e^{-2i\pi\xi t}dt,\\ &= \sqrt{s}\int_{-\infty}^{+\infty} \psi(u)e^{-2i\pi s\xi u}du,\\ &= \sqrt{s}\hat{\psi}(s\xi).\end{aligned} \tag{3.26}$$

Spectral graph wavelets

In order to mimic Equation (3.26) to construct the graph Fourier transform $\hat{\varphi}_s$ of the graph wavelet φ_s centered on a node at a scale s, we introduce a continuous function $g : \mathbb{R}^+ \to \mathbb{R}$ equivalent to the pass-band filter $\hat{\psi}$, and define $\hat{\varphi}_s$ as follows:

$$\hat{\varphi}_s(k+1) = \sqrt{s}g(s\lambda_k), \qquad \forall k \in [0, n-1]. \tag{3.27}$$

Then denoting G_s the filter matrix at scale s defined by:

$$G_s = \begin{bmatrix} g(s\lambda_0) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & g(s\lambda_{n-1}) \end{bmatrix}, \tag{3.28}$$

we can use the translation operator (Equation 3.23) to build the graph wavelets $\varphi_{s,j}$ centered on any node j:

$$\varphi_{s,j} = \sqrt{s}\chi G_s \chi^\top \Delta_k. \tag{3.29}$$

Following Equation (3.24), we can write the matrix Φ_s of all the wavelets at scale s as the following simple matrix product:

$$\Phi_s = (\Phi_{s,1}|\Phi_{s,1}|\cdots|\Phi_{s,n}) = \sqrt{s}\chi G_s \chi^\top. \tag{3.30}$$

Equations (3.11) and (3.29) lead to $\hat{\varphi}_{s,j} = \sqrt{s}G_s\chi^\top\Delta_k$, allowing to calculate the L_2-norm of $\varphi_{s,j}$:

$$||\varphi_{s,j}||_2^2 = \sum_{i=1}^{n} \hat{\varphi}_{s,j}(i)^2 = s\sum_{k=0}^{n-1} \chi_k(j)^2 g(s\lambda_k)^2. \tag{3.31}$$

It thus appears that, in general, the proposed construction leads to graph wavelets that are not normalised, so that the $\sqrt{s}$ normalisation factor can be dropped. This can be remedied by a posteriori normalisation of the wavelets:

$$\tilde{\varphi}_{s,j} = \frac{1}{||\varphi_{s,j}||_2}\varphi_{s,j} \qquad \forall (s,j) \in \mathbb{R}^+ \times V. \tag{3.32}$$

The normalised wavelet matrix at scale s will be denoted:

$$\tilde{\Phi}_s = (\tilde{\varphi}_{s,1}|\cdots|\tilde{\varphi}_{s,n}). \tag{3.33}$$

The scale parameter is continuous and hence there exists an infinity of different wavelet matrices. In practice, one selects n_s scale parameters $(s_1, s_2, ..., s_{n_s})$ depending on the application. The family of all wavelets on graph is the union of all the wavelets at all the scales:

$$\Phi = \bigcup_{j=1}^{n_s} \Phi_{s_j}. \tag{3.34}$$

There is n wavelets at each scale s_j, Φ is hence composed by nn_s wavelets.

Considering a signal F on a graph. Its wavelet coefficient $W_F(s,j)$ on a node j at scale s is given by:

$$W_F(s,j) = \varphi_{s,j}^{\top} F. \tag{3.35}$$

The union of all the coefficients is the wavelet transform of F at a scale s. We denote WF_s the vector of size n grouping all the coefficients:

$$WF_s = (WF_{s,0}|WF_{s,1}|\cdots|WF_{s,n-1})^{\top} = \Phi_s F = \chi G_s \chi^{\top} F, \tag{3.36}$$

which could have been directly deduced from Equation (3.20).

☞ Wavelets on graph give an idea on how the nodes where they are centered sees the graph. As illustrated in Figure 3.6, on our toy graph, at small scales, the wavelet coefficients are higher in the "close" neighbourhood of the nodes and for larger scales the wavelet is more extended over a larger neighbourhood. At small scale, the wavelet centred on a node has an "ego-centered" view of the graph, it takes the value 1 at that node and 0 elsewhere (like a Dirac) (Fig. 3.6 A). On larger scales, the wavelet coefficients expand on the neighbourhood of the node (Fig. 3.6 B-D). The wavelet neighbourhood at a scale reflect the graph topology. In fact, this graph consists in a hierarchical structure (see page 59). In Figure 3.6 (B), the wavelet coefficients are positive on the selected node and 15 other nodes of its neighbourhood, reflecting the second level of organisation of the graph in 8 groups of 16 nodes. In Figure 3.6 (C) the positive values of the wavelet coefficients extend to 32 nodes, reflecting the third level of organisation in 4 groups. Finally in Figure 3.6 (D), the wavelet positive coefficients extend to 64 nodes, reflecting the organisation in 2 groups.

3.5.1 Fast graph wavelet transform

One of the computational difficulties encountered in graph signal processing is the dependence of all operators on the graph topology: one cannot calculate, as in the classical case, wavelets once and for all because these wavelets depend on the analyzed graph. In fact,

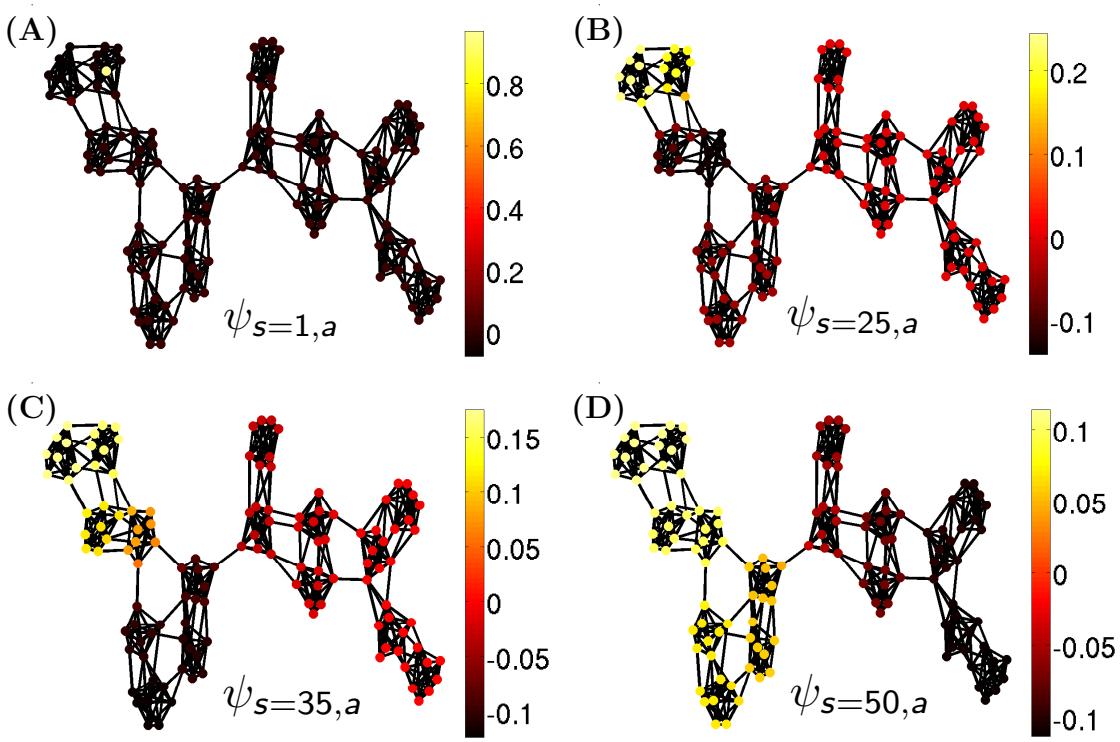


Figure 3.6. Wavelets on graph. A toy graph of 128 nodes (See Fig. 3.8) with wavelets centered at the yellow node visible in (A) across scales, from $s = 1$ in (A), to $s = 25$ in (B), to $s = 35$ in (C) and to $s = 50$ in (D).

Equation (3.30) shows that the calculation of the wavelet matrix at a scale s requires the knowledge of the Fourier matrix χ, itself calculated by the diagonalisation of the graph Laplacian. However, the diagonalisation of a matrix of size n needs at best a calculation time cubic in the number of nodes n, which makes it unpractical to use for graphs with more than a few thousands of nodes. To overcome this difficulty and to calculate the wavelet transform of a signal F quickly, it is in fact possible, using an approximated algorithm, to avoid the explicit calculation of the wavelets. As explained in [307], this approach consists in approximating each filter $g(s_j.)$ into a truncated Chebyshev polynomial of degree p:

$$g(s\lambda) \simeq \sum_{i=1}^{p} \alpha_i^{(s)} \lambda^i, \qquad \forall \lambda \in \mathbb{R}^+. \tag{3.37}$$

From Equation (3.28) and (3.37), it results the following approximation of the matrix G_s

$$G_s \simeq \sum_{i=1}^{p} \alpha_i^{(s)} \Lambda^i, \tag{3.38}$$

where Λ is the diagonal matrix of eigenvalues λ_k of the Laplacian matrix L.
Observing that $\chi \Lambda^i \chi^\top = L^i$, we can write the following approximations for the construction of graph wavelets and the computation of the graph wavelet transform:

$$\Phi_s \simeq \sum_{i=1}^{p} \alpha_i^{(s)} L^i, \tag{3.39}$$

and

$$W F_s \simeq \sum_{i=1}^{p} \alpha_i^{(s)} L^i F. \tag{3.40}$$

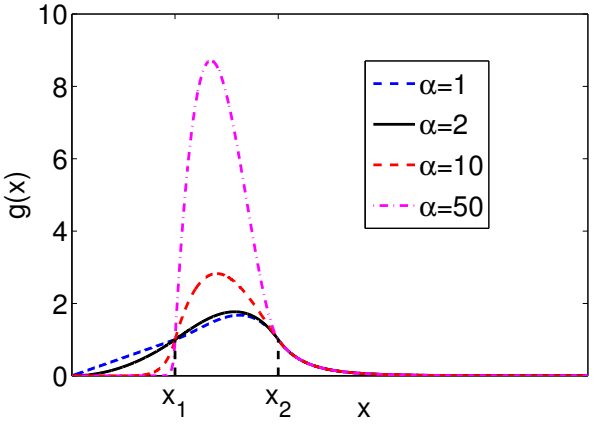


Figure 3.7. The wavelet filter. Shape of the filter function $g(x)$ (Equation (3.42)) for four different values of α. Adapted from [309].

Hence, instead of having to calculate the diagonalisation of L, one can compute wavelet coefficients Wf_s only via matrix-vector multiplications, given that L^i, $i = 1, ..., n$ have to be computed only once. For reasons of calculation speed and optimization of the approximation discussed in [307], the polynomial used in the approximation of Equation (3.37) is a truncation to degree p of the Chebyshev polynomial development of the filter $g(s.)$. The computation time for nn_s coefficients falls to $O(n_s m + nn_s p)$ where m is the number of links. For hollow graphs, *i.e.* graphs whose number of links is of the order of the number of nodes, the computation time of the algorithm is linear in n. Obviously, with this consideration one can only get an approximation of the graph wavelets Φ_s and graph wavelet coefficients WF_s, whose precision depends on p, the larger p, the better the approximation.

In the following we note $\mathcal{FWT}_{s,p}$ the fast wavelet transform operator at a scale s with polynomial approximation parameter p:

$$WF_s \simeq \mathcal{FWT}_{s,p}F. \tag{3.41}$$

3.5.2 Defining wavelet filter kernel

For a given graph, the wavelet family Φ is entirely determined by its kernel filter g and the choice of the n_s considered scales. Let us begin by discussing the choice of g. In classical signal processing, building its equivalent $\hat{\psi}$ has been the subject of many research efforts during the past thirty years and wavelets of many forms now exist, that are adapted to specific classes of problems (Appendix A) [300]. For graph wavelets, to ensure their localization, the only constraint that exists on g is about its behaviour at the origin which must behave like x^α ($\alpha > 1$). For now, a generic used form is a band-pass filter (like a *bell* Fig.3.7) with a behaviour at the origin like x^α and like $x^{-\beta}$ at large values. Other more sophisticated forms have been proposed [310, 311], which take into account the particular spectrum of each graph; we will not explore these variants that are more complicated and need the calculation the entire graph spectrum, thereby excluding large graphs due to computational cost.

We consider g of the following form (Fig. 3.7):

$$g(x;\alpha,\beta,x_1,x_2) = \begin{cases} x_1^{-\alpha}x^{\alpha} & x < x_1, \\ P(x) & x_1 \leq x \leq x_2, \\ x_2^{\beta}x^{-\beta} & x > x_2, \end{cases} \tag{3.42}$$

where $P(x)$ is the unique cubic polynomial interpolation which maintains continuity of g and its derivative g'. α, β, x_1 and x_2 are the filter parameters. They can be adjusted according to what one seeks to analyze [309]. In fact, the largest "interesting" scale is encoded in χ_1, the Fiedler vector that cuts the graph in 2 (Fig. 3.5) [312]: nodes x_i with negative $\chi_1(x_i)$ on the one hand and nodes x_j with positive $\chi_1(x_j)$ on the other hand (Fig. 3.5 A). In this thesis, as we use the graph wavelets for community detection, the construction of the filter g obeys the "constraints" as introduced in [309].

The parameter x_1 has the single effect of translating the range of scales along the real line and has no consequence on the filter; hence its value can be fixed to 1 without loss of generality. The effect of α is indirect and condition the maximum reached by $g(x)$ [309] (Fig. 3.7); and hence α is constrained to be small. In this thesis, we use $\alpha = 2$ as previously done in [307, 309]. The maximum scale s_{max} is set so that the filter function $g(s_{max}x)$ starts decaying as a power law for $x \geq \lambda_1$, hence $s_{max} = x_2/\lambda_1$. Requiring that the filter at the maximum scale is highly selective around λ_1, all the eigenmodes (especially λ_2) have to be attenuated. Choosing an attenuation by a factor 10 leads to $g(s_{max}\lambda_1) = 10g(s_{max}\lambda_2)$, and in turn $\beta = 1/log_{10}(\lambda_2/\lambda_1)$. To make sure at all scales the wavelets are sensitive to the lowest frequency (hence to the large scale community structure), we choose $s_{min} = x_1/\lambda_1$ so that $g(s_{min}\lambda_1) = 1$. Furthermore, imposing that $g(s_{min}.)$ spans the whole range of eigenvalues between 0 and 1 imposes that $s_{min} \times 1 = x_2$. Altogether these constraints lead to a set of parameters:

$$\begin{aligned} &\alpha = 2, && \beta = \frac{1}{log_{10}(\frac{\lambda_2}{\lambda_1})}, \\ &x_1 = 1, && x_2 = \frac{1}{\lambda_1}, \\ &s_{min} = \frac{1}{\lambda_1}, && s_{max} = \frac{1}{\lambda_1^2}. \end{aligned} \tag{3.43}$$

☞ As illustrated in Figures 3.5 and 3.6, the graph Fourier modes and the wavelet coefficients respectively encode information about the graph topology. Hence, it is enough to build community mining methods to construct the eigenvectors χ_l or $\varphi_i$in order to capture information on the graph structure.

3.6 Graph community mining

Consider for a moment, a society represented as a social network of friendships and acquaintances. It can be organized in different groups: families, people working together or going to school together, groups of friends... This feature is known as *community structure* [275, 313–315]. Formally, a community is a group of nodes highly interconnected to each other and less connected to the rest of the graph (Fig. 3.8). Identifying communities in a network helps us to understand the network's structure. For instance, communities

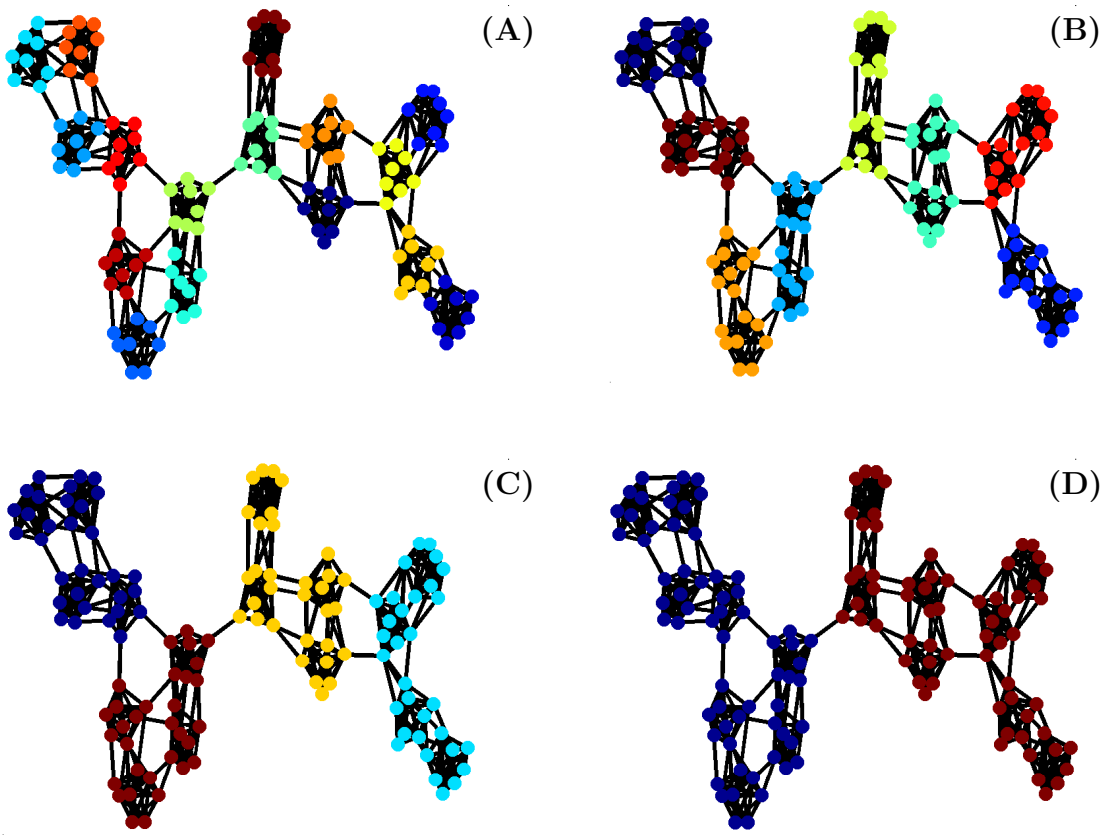


Figure 3.8. Multi-scale community structure. A toy graph of 128 nodes built with a hierarchical structure at 4 scales. The graph can be easily partitioned into 16 (A), 8 (B), 4 (C) and 2 (D) communities: from small (A) to large (D) scales. At each scale, each color corresponds to a community.

in a social network can reflect social groupings [309], communities in a citation network group papers with similar topics [316], communities in protein network group proteins that have similar functions [317]... Another example of application of community detection is to cluster together web customers that have similar interests in order to be able to make efficient recommandations [318].

It is also possible that communities are themselves grouped together to form bigger communities and so on as illustrated by our hierarchical toy graph on Figure 3.8. This *hierarchical* organisation is displayed by many real world systems. For instance, the human body is composed of cells that are the building blocks of tissues that form organs. Another example of hierarchical organisation is the educational system: students are grouped in classes that form a grade-level, grade-levels are then grouped into elementary, middle and high school levels. The aim of community detection in networks is to identify the modules and their hierarchical organisation using only the topology of the graph.

The problem of community detection started in sociology. In 1927, Rice [319] looked for groups of people in political bodies based on the similarity of their voting patterns. Weiss and Jacobson [320] also looked for work groups in a government agency based on the relationships/interviews they used to have. Homans [321] also tried to find group of people with similar interest. A group of informaticians and mathematicians also started

to formalise the problem into *graph partitioning* and developed algorithms for community detection [322, 323]. Later in the years 2000, Girvan et Neuman [313] defined a quality measure to assess the community structure, the well known *modularity*. This method (see Section 3.6.2) coupled with the fast *Louvain* algorithm [324], is widely used nowadays. Community detection is still the subject of interdisciplinary studies between mathematics, physics, informatics and social sciences [275]. In the following, we mainly focus on the most important aspects of this domain. We refer the reader to the work of Fortunato [275] for a comprehensive review of this field of research.

3.6.1 Traditional community detection methods

As reported in [275], traditional methods for community detection can be grouped into 4 main categories: *graph partitioning*, *hierarchical clustering*, *partitional clustering* and *spectral clustering*.

Graph partitioning consists in dividing the vertices of the graph in k groups of predefined size such that the number of edges lying between the groups is minimal. Many graph partitioning algorithms have been developed [325], but the inconvenient of this approach is that one needs to fix a priori the number of groups and sometimes even their sizes.

Hierarchical clustering [326] is a clustering technique that reveals the multilevel structure of the graph. By defining a measure of *similarity* (such as the distance between the nodes when they can be embedded in space [275]), and computing the similarity for each pair of nodes, hierarchical clustering groups together nodes of high similarity. To access a multi-level partitioning, either clusters of high similarity are grouped together or cluster are split by removing edges connecting vertices with low similarity. However most of the time, these techniques do not provide a way to discriminate between the many partitions of the graph and the computational complexity is high (for instance, when the nodes can be easily embedded in space, using the distance as a similarity measure leads to $O(n^2 \log n)$ complexity).

Partitional clustering [327–330] consists in defining a preassigned number k of clusters and a cost function to optimize. For instance, for the *minimum k-clustering* technique, the cost function is the diameter (which is the largest minimal distance between two nodes) of the cluster. The goal is to associate the nodes in k clusters in such a way that the largest diameter of the k clusters is the smallest possible.

Spectral clustering [294, 312] consists in using the eigenvectors of the adjacency matrix of the graph or the Laplacian or matrices derived from them. For instance, as illustrated for our hierarchical toy graph in Figure 3.5 A, the method in [312] will find two communities based on the sign of the Laplacian eigenvector associated with the first not null eigenvalue. However, the computation cost for large graphs is high because one needs to compute the first k eigenvectors of the chosen matrix.

Spectral clustering techniques have been recently revisited with *spectral algorithms* [331]. The intuition behind these algorithms is that close nodes of the network have similar values in the eigenvector. The nodes are embedded in a d dimensional space with coordi-

nates defined by the components of the d first eigenvectors. Using hierarchical clustering, nodes with similar positioning in the new space are grouped together. This algorithm can be repeated for many values of d.

3.6.2 Modern community detection methods

Community detection continues to be a very active field of research. Some "modern methods" can be classified in the following categories:

Divise algorithms consist in recursively identifying the edges that connect vertices of different communities and in removing them so that the clusters gradually appear as disconnected components. The point is to find a property of intercommunity edges allowing their identification. The most well-known algorithm in this category is the one of Girvan et Newman [313, 315]. After the computation of the edge betweenness centrality (the definitions of centralities introduced in Section 3.3 can be directly extended to the edges by reversing the role of edges and vertices), the edge with the largest centrality is removed. Then the centralities are again calculated on the new graph and so on until the largest centrality is below some predefined threshold.

Quality function based methods rely on the definition of a global function that quantifies the *quality* of a partition of the nodes into communities. Defining such functions depends on the intuition we have about what a community should be. The critical point is then to divise an algorithm allowing to find the best scoring partition given this quality function. The most popular quality function is the modularity [313] which compares the number of edges inside the communities to the number of edges outside the communities. Many efforts has been made to construct heuristics allowing to find the "best" partition in a reasonable time. Very good results in terms of speed and scalability are obtained with the Louvain method [324].

Many other methods are also being developed but much less used such as *dynamic algorithms* either based on *spin model* [332, 333] or *random walks* [334–336] (some of those methods can produce hierarchical communities and will be discussed along with the multiscale methods in the next section).

The above methods aim at detecting standard partitions where each vertex can be assigned to one community only. However, in some real life networks, a node can be shared by different communities, so that, many efforts are now made to develop methods allowing to detect overlapping communities such as the *Clique Percolation Method (CPM)* [337]. The method consists in building community starting from dividing the graph in cliques[‖] of k nodes. The authors consider that two k-cliques are adjacent when they share $k-1$ nodes. Then a community is defined as the maximal union of k-cliques that are adjacent. This definition allow communities to share vertices.

Finally, it is worth mentioning the recent interest in methods allowing to detect "dynamic" communities that follow the evolution in time of real systems [338, 339]. Basically at each time step, graph communities are detected and then the relationships between the partitions at successive times are inferred.

[‖] A clique is a group of nodes *fully connected* to each others.

3.6.3 Multi-scale community detection

As discussed previously and illustrated by our hierarchical toy graph (Fig. 3.8), the community structure of many real world systems cannot be fully captured by a single partition of the nodes as this notion may depend on the resolution at which we observe the system. The issue of scale was implicitly discarded in the above methods. In fact, the user does not choose the scale, and the algorithm outputs one partition arbitrarily. For instance, the algorithms based on modularity optimisation have been shown to favor an intrinsic scale of description [340, 341]. On the example of Figure 3.8, the modularity optimisation algorithms will pick the partition in 8 communities (Fig. 3.8 B) because it has the maximum modularity 0.83 while the partition in 16 communities has a modularity of 0.80, the partition into 4 communities has a modularity of 0.74 and the partition into 2 communities has a modularity of 0.50, However, all the partitions can be meaningful and one should be able to recover them all. Multi-scale community mining methods have been introduced as a response to the possible existence of a hierarchy of relevant community descriptions. These methods have, in general, a freely tunable parameter allowing to detect different partitions on a "interesting" range of scales. However, an important question remains: how to choose "interesting" partitions or scales? In this section, we briefly recall existent multi-scale community mining methods either based on random walk processes [342, 343] or on definitions of parametric modularities [344, 345]. We then discuss a recently developed method deeply rooted in signal processing on graphs and that uses graph wavelets [309]. This is the method we choose in this thesis to analyse Hi-C data (Chapter 6).

Inspired by statistical physics, the multiscale method proposed by *Reichardt and Bornholdt* [344], relies on the analogy between community mining and Potts model. The Potts model is a generalisation of the Ising model which is a collection of "up" and "down" spins. For each spin i (node x_i in the analogy), there is an associated spin state σ_i (corresponding to the community in the analogy). Each spin interacts with the adjacent spins in the graph with a non null interaction energy when spin states are different. Summing the energy associated to each edge results in the total energy of the spin configuration. The final state is obtained by minimising the total energy over the possible spin configurations. The Ising model, in a graph, corresponds to splitting the graph in two communities. The Potts spin model identify the optimal number of communities.

The method proposed by *Arenas et al.* [345] relies on adding loops of weight r in the graph. The weighted adjacency matrix W is replaced by $W_r = W + rI$, where I is the identity matrix, and the modularity Q is replaced by $Q_r = Q(W_r)$ where r is the scale parameter. High values of r allow detection of small communities, while small values of r result in larger communities.

Methods based on *random walk* [342, 343] use the interpretation of W in terms of transition matrix of a Markov chain. The intuition is that a flow on the graph will be trapped for a longer "time" inside a community before being able to escape, suggesting that the quality of a partition can be measured in terms of persistence of flows taking place on the graph. The authors define a quality function that correspond to the probability of a random walker to start in a community i and finishes in community j after a time t, from which is substracted the probability that two independant random walkers are in i

and j at stationnary state. t plays the role of the scale, and for $t = 1$ the method coincides with the method of maximizing the modularity. The authors have also established the link between their quality function and the ones defined in the Reichardt [344] and Arenas [345] methods.

3.6.3.1 Community mining using wavelets on graphs

The intuition behind this wavelet based method is to consider that the wavelet centered around the node is an "ego-centered" vision of the graph, and that two nodes in the same community must have pretty much the same view of the graph (discussed above). In other words, and this is the central idea of the algorithm, all the nodes are classified together if their neighbourhood as defined by graph wavelets are similar. The method consists at each scale s in three steps [309]:

- For each node a, one defines its feature vector as the coefficients of the wavelet $\psi_{s,a}$ that encodes local information on the graph topology seen by the node a (Fig. 3.6.)
- To compare the nodes that is to define to which extent two nodes a and b have a similar environment, a distance matrix $\mathcal{D}_s$ is created where the correlation distance $\mathcal{D}_s(a, b)$ between nodes a and b is one minus the correlation between the wavelets $\varphi_{s,a}$ and $\varphi_{s,b}$:
$$\mathcal{D}_s(a, b) = 1 - \frac{\varphi_{s,a}^{\top}\varphi_{s,b}}{||\varphi_{s,a}||_2||\varphi_{s,b}||_2} \qquad \forall (a, b) \in V^2. \tag{3.44}$$
Note that this distance measure is independent of graph wavelet normalisation.
- A hierarchical clustering algorithm is used to classify the nodes. The hierarchical algorithm outputs a dendogram that needs to be "cut" to obtain a partition P_s. To cut the dendogram, the method defines a criterion based on averaging the maximal gaps of all the root leaf paths of the dendogram. For each node a, one computes the gap function Γ_a as the path length between the leaf corresponding to node a and the begining of the dendogram. Then after averaging all gaps functions into a global gap function, the best cut corresponds to the maximum of this global gap function.

Repeating these three steps for each scale s, one obtains a multi-scale set of partitions P_s of the nodes [309].

The major inconvenient in this procedure is the computation cost. In fact, to calculate the wavelets one needs to diagonalise the graph Laplacian which is a problem for large graphs as discussed previously. However, we saw that it is meaningful to approximate the wavelet transform using the $\mathcal{FWT}_{s,p}$ (Equation (3.41)). Hence, it is possible to approximate each wavelet by:

$$\varphi_{s,a} = \mathcal{FWT}_{s,p}\Delta_a, \tag{3.45}$$

and then to calculate the correlations to estimate the distance matrix. In [309], the authors demonstrated that it is enough to actually compute the $\mathcal{FWT}_{s,p}$ of a "few" random gaussian vectors to estimate $\mathcal{D}_s$ which makes the method suitable for the analysis of large graphs.

3.6.3.2 Evaluating the robustness of communities

In summary, all the multiscale methods discussed above define a function of similarity between the nodes or a quality function, parameterised by a notion of scale dependent on the method. Given that one can access to the partitions P_s for all the s values within some prescribed scale range, a question arises: which is the best partition? Hence, it is essential to combine these methods with a measurement of relevance of the considered scales. This relevance can be estimated in terms of *stability* of the associated partitions. There are several types of stability.

Stability can be measured directly by disrupting the graph [343, 346, 347]. It consists in creating many graphs G_1, G_2,... (sometimes called *bootstraps*) of the initial graph G by perturbing the edges (by changing their weights or removing them with a certain probability...), and calculating the associated partitions P_1, P_2,... Identifying stable partitions from the initial graph is done by measuring similarity between the P_1, P_2,... partitions. Similarity between partitions can be measured in different ways. For example, considering 2 partitions, one can look at the number of nodes that are in the same community for the 2 partitions n_{11}, the number of nodes that are in different communities in the 2 partitions n_{00}, the number of nodes that are in the same community in one partition and in different communities in the second one n_{10} and n_{01}. Using these four numbers (that sum always to n(n-1)/2) similarity coefficients can be defined [348–352]. A common similarity index based on the above numbers is the *Rand index* $R = \frac{n_{11} + n_{00}}{n(n-1)/2}$ [351]. Another way of defining stability is by looking at the variation of information between two partitions [353].

Another method to measure stability uses the advantage of stochasticity of some optimization algorithms [343, 354]. Launching the algorithms many times can lead to different partitions and by measuring similarities between the partitions one gets the stability.

Finally a way of getting the stability of the partitions in a multi-scale method consists in looking at the size of the range of scales for which a given partition is found [343, 345, 355]. This relies on the fact that a partition is stable if it is conserved across different scales.

☞ As explained in Section 3.6.3.1, we can compute an approximation of the graph wavelets distance matrix D_S (Equation (3.44)) using a stochastic approach based on a set of Gaussian random vectors. One can use this stochasticity to estimate the stability of the partition P_s obtained. The intuition is as follows: repeating the algorithm J times, if the partitions P_s^j are different, the partition is considered unstable and if the partitions are quite similar, the original partition is considered as stable. Formally, stability γ at scale s is defined as the mean of the similarities between the different partitions P_s^j:

$$\gamma(s) = \frac{2}{J(J-1)} \sum_{i>j} simi(P_s^i, P_s^j), \tag{3.46}$$

where *simi* is an index of similarity between two partitions like the Rand index defined above. In practice, the choice of the index used to define the similarity has no or little effect on the result.

3.7 Conclusion

To sum up, we saw in this chapter that graphs are extremely useful as a representation of complex systems. On the one hand, there exist centrality measures (section 3.2) that capture different notions of "importance" of a node based on the diverse intuitions we can have behind the word "importance". These centrality measures allow to identify key players in the network (Table 3.1). On the other hand, community detection (Section 3.5), allows to understand the structure of the graph by identifying groups of highly interconnected nodes. As illustrated on Figure 3.8, a graph can be partitioned at different scales. We presented a multi-scale community detection method based on spectral graph wavelets. The wavelets introduce the notion of scale as they represent how a node "sees" the graph (Fig. 3.6) and this method groups together nodes that "see" the graph in a similar way.

This motivates the use of graph theory to decipher the DNA structural organisation in relation with nuclear functions such as replication. After defining the DNA network based on chromatin interaction data, we will see if key functional features of the genome such as the "master" replication origins (MaOris) correspond to key players in the DNA network. Our strategy will be to study their centrality scores relatively to the other genome fragments (Chapter 5). Moreover, using multi-scale community detection, we will identify structural compartments of the DNA at different resolutions and see to which extent they correspond to domain of functional organisation (Chapter 6).

Part II

Results and Discussion

Chapter 4

Towards a unified classification of the genome

About half of the human genome was shown to be divided in domains that display a characteristic U-shaped replication timing profile with early initiation at the borders and late replication in the center. Significant overlap was observed between U-domains of different cell lines and also with germline replication domains exhibiting a N-shaped nucleotide compositional skew. A recent study described another type of skew structures that cover $\sim 13\%$ of the human genome and are bordered by putative master replication origins similar to the ones flanking skew-N-domains. These skew-split-N-domains have a shape reminiscent of a N, but split in half, where the skew decrease over ~ 1.5 Mb leaving a large central region of null skew whose length increases with domain size. This prompts us, in this chapter, to ask the question of the existence of caracteristic size for MRT U-domains and/or the existence of MRT-split-U-domains as the counterpart of the skew-split-N-domains leading to a unified classification of the genome.

4.1 A universal cascade model at the heart of the replication programme

As discussed in Chapter 2, the human genome can be segmented based on strand composition asymmetry profiles (Equation (2.2)) resulting in the so-called skew (S) N-domains (Section 2.1.2) [39, 41, 46, 47, 97], and/or based on mean replication timing profiles resulting in the so-called MRT U-domains (Section 2.1.4.3) [43, 44, 161]. We showed that both the skew S and the MRT derivative (dMRT/dx) reflect replication fork polarity (Sections 2.1.4.1 and 2.1.4.2) [44, 48, 49] and thus that the cumulative skew Σ (Equation (2.4)) can be considered as a footprint of the MRT in germline. The N-shaped skew component was associated with replication domains in the germline (Section 2.1.2), with upward jumps bordering the skew N-domains qualified as putative "master" replication origins. A significant overlap was observed between the skew N-domains and the MRT U-domains (Section 2.1.4.3). In these megabase-sized domains, the MRT derivative is N-shaped like the nucleotide compositional skew in N-domains in the germline. These peculiar N-shaped patterns in the MRT derivative were observed in every cell type and were shown to be the signature of the existence of large-scale gradients of replication fork polarity (Section 2.1.4) [44, 69, 161] originating from early initiation zones Mb away from each other. The "master" replication origins at U/N-domain borders were found to be hypersensitive to DNase I cleavage, to be enriched in epigenetic marks involved in transcription regulation and to present some local excess of the insulator binding protein CTCF, the hallmarks of localized ($\sim$ 200-300 kb) open chromatin structures [44, 51, 161] (Section 2.1.2). The internal part of the U/N-domains actually corresponds to silent facultative (polycomb repressed) and constitutive (HP1-associated) heterochromatin regions that replicate later and later as we move from U/N-domains borders to center (discussed in Section 4.2.3) [53, 213]. The first model proposed to account for the skew N-shape simply considered that there were only two origins in a skew-N domain, one at each border, and that the forks emanating from each origin met and stopped at a termination site that was uniformly distributed in the domain [39, 40]. This model was not in agreement with commonly accepted inter-origin distances in mammals that are around 50 - 100 kb and not 1 - 2 Mb [69, 70, 230, 356, 357]. A cascade model of secondary origin firing was recently proposed (Section 2.1.4.3 page 34) to account for the gradient of replication fork polarity inside U/N-domains [69, 70]. The U-shape of MRT profiles indicates that the effective replication velocity (which equals the inverse of the MRT derivative [44, 48]; see page 33) increases from U-domain borders to center [44, 70] as the signature of an increasing origin firing frequency during S-phase [358]. This cascade model involves the superposition of specific and efficient initiations at domain borders with random and less efficient initiations elsewhere, in addition to firing stimulated by the propagating forks [69, 70]. However, an important aspect for any model is to account for the antisymmetric shape of skew-N domains is the fact that the slope is inversely proportional to the domain size [39–41] suggesting that skew N-domain borders are not independent objects. This led us to investigate what happens in between two successive and very distant upward jumps in human skew profiles and if they do not border a detected skew N-domain, whether they present the same genome organization around them.

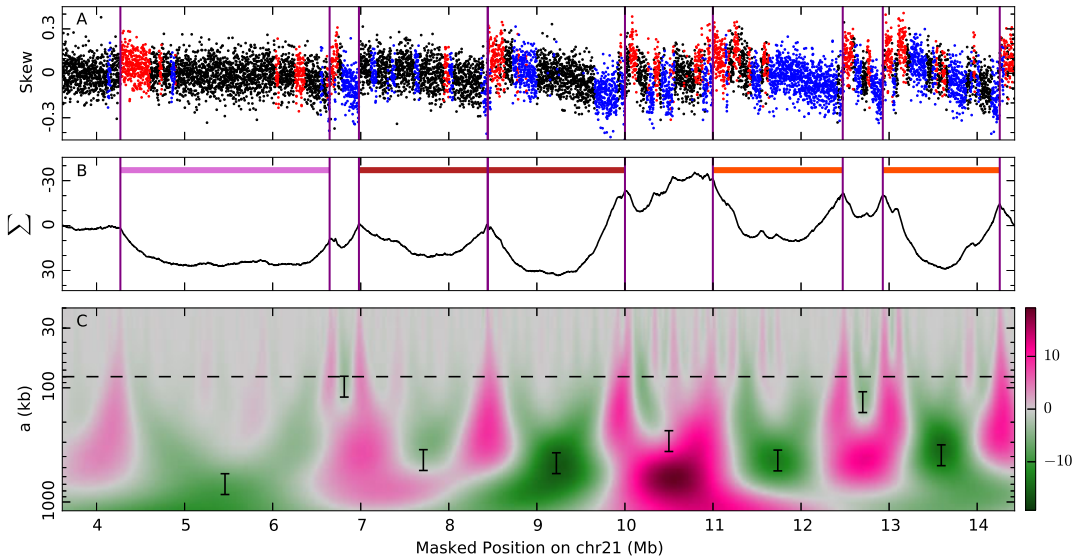


Figure 4.1. Detection of skew split N-domains. (A) Skew profiles calculated in non-overlapping 1kb windows with color dots corresponding to intergenic (black), (+) genes (red) and (-) genes (blue) plotted *vs* the native position along a 10.8 Mb fragment of human chromosome 21. (B) corresponding cumulative skew profile Σ obtained by the cumulative addition of S-values along the sequence (Equation (2.4)). For display purpose, average value of S over the region of interest was set to 0 prior to computing Σ and the ordinate axis goes downwards. The horizontal color bars correspond to the 5 Σ-U domains detected along the fragment. (C) Space-scale representation of second order variation of $\Sigma(x)$: $T^{(-1)}_{g^{(2)}}[\Sigma](x,a)$ (Appendix A, Section A.5) values are color coded using green (resp. pink) for negative (resp. positive) curvature. The horizontal dashed line marks the scale 80 kb used to detect regions of preferential replication initiation: $T^{(-1)}_{g^{(2)}} \geq 1.5$ (vertical lines in A and B). Black vertical bars delineate the scale range where strong negative curvature is expected for parabolic U-shaped Σ profile. Regions delineated by two successive regions of preferential replication initiation are kept as Σ-U domains if $T^{(-1)}_{g^{(2)}}[\Sigma] \leq -4$ at their midpoint for some scale value in this range [43, 359].

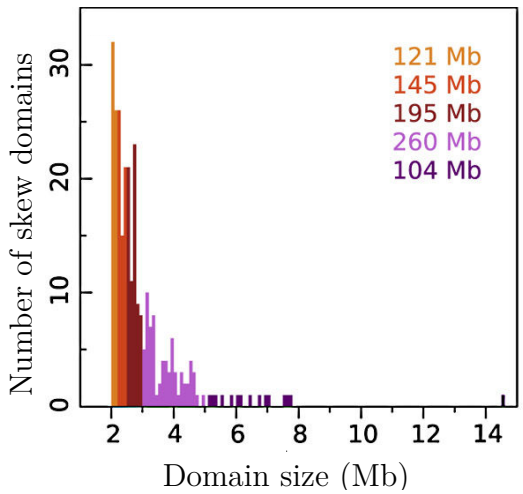


Figure 4.2. Histogram of detected large Σ-U domain sizes. Σ-U domains were grouped into five size categories: 2 Mb $\leq L <$ 2.2 Mb (light orange) covering 121 Mb (N $=$ 58); 2.2 Mb $\leq L <$ 2.5 Mb (orange) covering 145 Mb (N $=$ 62); 2.5 Mb $\leq L <$ 3 Mb (dark orange) covering 195 Mb (N $=$ 72); 3 Mb $\leq L <$ 5 Mb (light purple) covering 260 Mb (N $=$ 70) and 5 Mb $\leq L <$ 15 Mb (purple) covering 104 Mb (N $=$ 15).

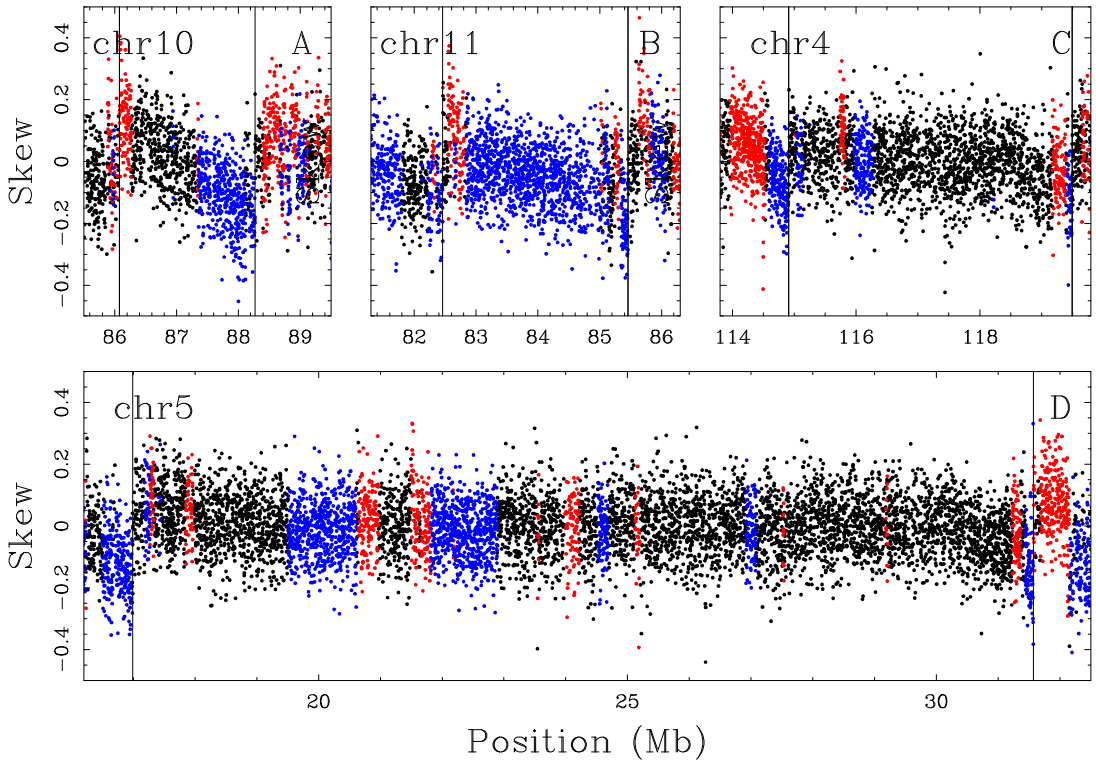


Figure 4.3. Skew profiles S. Skew profiles calculated in non-overlapping 1kb windows with color dots corresponding to intergenic (black), (+) genes (red) and (-) genes (blue) plotted *vs* the native position along different chromosomes. (A) One Σ-U domain of size 2 Mb $\leq L \leq$ 2.2 Mb; (B) one Σ-U domain of size 2.5 Mb $\leq L \leq$ 3 Mb; (C) one Σ-U domain of size 3 Mb $\leq L \leq$ 5 Mb; (D) one Σ-U domain of size 5 Mb $\leq L \leq$ 15 Mb.

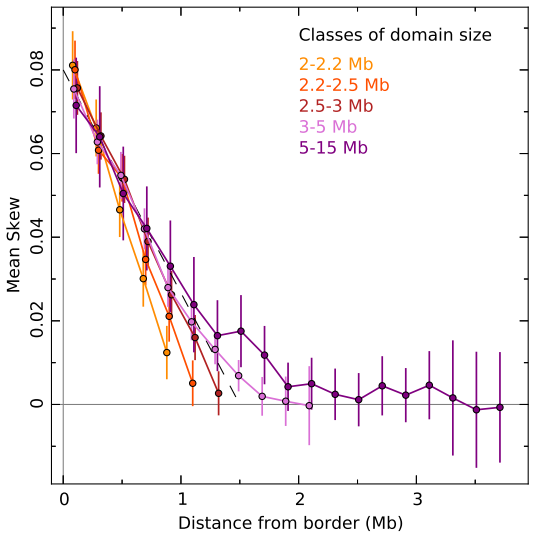


Figure 4.4. Mean skew. Mean (masked) skew S calculated in non-overlapping 200 kb windows vs distance to the closest Σ-U domain border. Σ-U domains were grouped into the same five size categories as in Fig. 4.2. The black dashed line originating at $\bar{S} = 0.08$ at domain border and crossing $\bar{S} = 0$ at 1.5 Mb is drawn to guide the eyes.

4.1.1 Skew linearly decreases up to 1.5 Mb from MaOris

As illustrated in Fig. 4.1, in segments of linearly changing skew, the integrated S function, when estimated by the cumulative skew Σ, displays a U-shaped profile (similar to the parabolic U-patterns observed in MRT profiles in Figure 2.10 C) [44]. Thus we can analyse the cumulative skew Σ instead of the skew S [359], to detect large skew domains using the same wavelet-based protocol described in Appendix A (Section A.5) to delineate U-patterns in MRT profiles [43, 44] (Fig. 4.1 B). Using this methodology, we delineated 284 cumulative-skew-U (Σ-U) domains of size $L \geq 2$ Mb in the human genome among which 7 were discarded by manual curation*. Five of these Σ-U domains lying on human chromosome 21 are shown in Figure 4.1. Overall, the data set of 277 Σ-U domains spans 826.0 Mb (Fig. 4.2) of the native sequence covering about 28.9% of the sequenced genome and involves 510 distinct domain borders likely corresponding to active replication origins in the germline. These domains were grouped into five categories (Fig. 4.2): 2 Mb $\leq L <$ 2.2 Mb covering 121 Mb (N = 58); 2.2 Mb $\leq L <$ 2.5 Mb covering 145 Mb (N = 62); 2.5 Mb $\leq L <$ 3 Mb covering 195 Mb (N = 72); 3 Mb $\leq L <$ 5 Mb covering 260 Mb (N = 70) and 5 Mb $\leq L <$ 15 Mb covering 104 Mb (N = 15). Some of these Σ-U domains correspond to skew-N domains previously detected with the N-let method originally used to detect N-domains (Appendix A, Section A.3): 56 out of the 86 N-domains ($\sim$ 65%) of size 2 Mb $\leq L \leq$ 2.8 Mb identified in [46]. Indeed, 123 out of the 165 different skew N-domain borders ($\sim$ 75%) were recovered at a 50 kb precision. As illustrated by the four large Σ-U domains shown in Fig. 4.3, for domains of size 2 Mb $\leq L \leq$ 3 Mb, the previously described N-shape with a linear decrease of the skew profile in between two upward jumps were recovered. However, the skew profile in the large Σ-U domains (> 3 Mb) does not correspond to an N-domain. The skew does not decrease on the whole domain length but only on $\sim$ 1.5 Mb, independently on the domain size, leaving a large central region of null skew (Fig. 4.3 C, D). This shape reminiscent of an N but split in half and with a central region of null skew led us to name them Split-N-Domains. When plotting the average skew $\overline{S}$ versus the distance to the border for different classes of domain sizes, we recovered for the first three classes that the slope of the linear decrease of S varies as the inverse of the domain size $1/L$ (Fig. 4.4), as originally observed for skew-N domains [39–41, 46]. However, in the last two classes of skew split-N-domains, the skew linearly decreases from the border in a similar fashion over $\sim$ 1.5 Mb irrespectively of the domain size (Fig. 4.4). Remarkably, the mean skew jump amplitude at the Σ-U domain borders does not vary too much with the type of the domains ($\overline{S} \sim 7.8\%$) except some systematic slight decrease with the domain size (Fig. 4.4). This might be some indication that very much like N-domain borders [39–41, 46, 47, 51], split N-domain borders also likely correspond to well positioned "master" replication origins firing early in S-phase.

☞ The characteristic of split-N-domains that sets them apart from N-domains are their central regions that have a null skew (Fig. 4.4). The link between the two borders of the domains seems to have been disrupted in split-N domains by the presence of this region whose length can reach several megabase. The fact that the skew profile does not decrease on the whole domain length but rather on 1.5 Mb independently of the domain size suggests that 1.5 Mb is a characteristic (limiting) length and/or time scale in the organisation and coordination of replication, transcription and chromatin [359].

*Because they do not display the characteristic central region of null skew described later on, unlike the others.

	BG02	IMR90	BJ (R1)	BJ (R2)	GM06990	GM12878	TL010	H0287	K562	HeLaS3	MCF7
N	30	62	76	67	105	107	150	122	107	38	84
L	3.59	3.86	4.21	4.10	4.13	4.13	4.24	4.27	4.28	3.97	4.13
C	4.05	9.02	12.04	10.34	16.32	14.78	23.96	19.63	17.24	5.68	13.06

Table 4.1. Characteristics of large detected MRT domains (L $\geq$ 3 Mb). Number (N) of detected domains, their mean length (L) in Mb and their genome coverage (C) in each of the analysed cell lines.

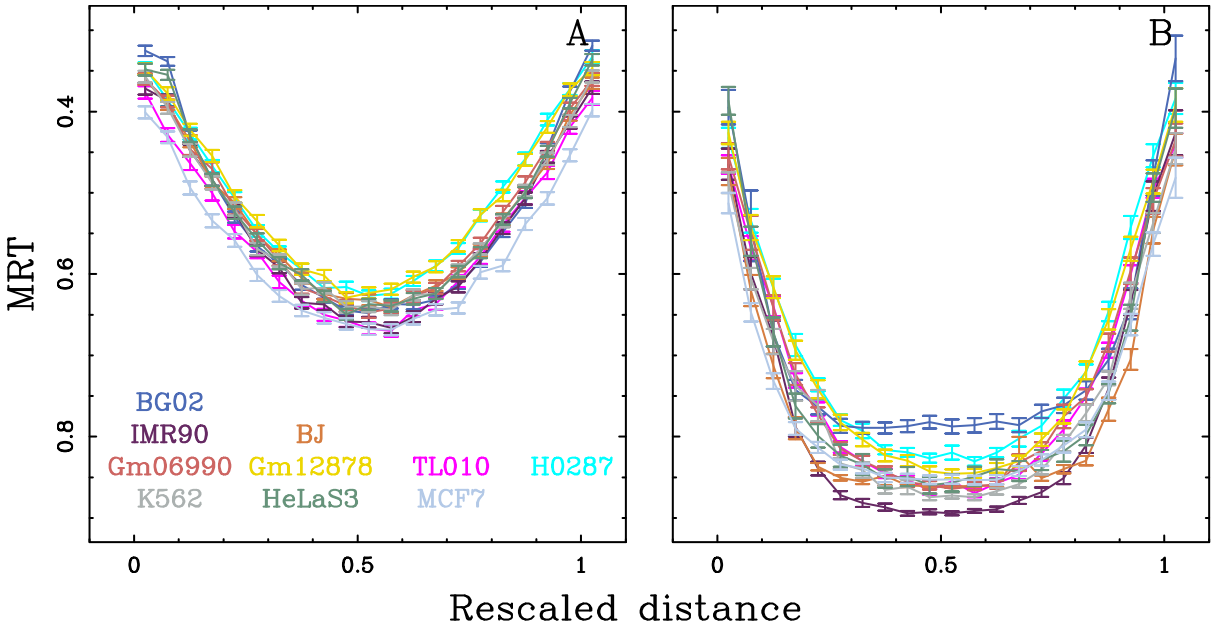


Figure 4.5. MRT U-domains and split U-domains in different human cell lines. (A) Average MRT profiles inside detected replication U-domains (L $<$ 3 Mb). (B) Average MRT profiles inside detected replication timing-domains larger than 3 Mb. The distance between domain borders was rescaled to 1. Each cell line is identified by a color: BG02 (light blue), IMR90 (pink), BJ (R1) (orange), GM06990 (pink), Gm12878 (yellow), TL010 (magenta), H0287 (cyan), K562 (grey), HeLaS3 (green), MCF7 (grey-blue).

4.1.2 MRT split-U-domains are robustly observed in different cell lines

To address the hypothesis of the existence of a limiting length and/or time scale organisation, we extend our study to the MRT profiles. In fact, as both the skew S and the MRT derivative reflect the replication fork polarity (Sections 2.1.4.1 and 2.1.4.2), we expect to obtain the same behaviour when analysing the MRT derivative. Historically, the wavelet-based methodology allowing to detect U-domains (L$\leq$3 Mb) [43, 44] (Appendix A, Section A.5), identified larger domains (from 30 in BG02 up to 150 in TL010, (Table 4.1)), of mean size ranging from 3.59 Mb in BG02 to 4.28 Mb in K562, covering around $\sim$ 10% of the genome (Table 4.1). Figure 4.5 illustrates the average MRT profile in different cell types along the so-called replication U-domains (Fig. 4.5 A), and along the larger domains (Fig. 4.5 B) when rescaling the domain size to 1. The average MRT profile in U-domains (Fig. 4.5 A) has the expected parabolic shape representative of individual U-domains (Fig. 4.6, first column), where the borders replicate early whereas the centers replicate later. Figure 4.5 B also shows domains with early replicating borders on average. However, we see the emergence of a flat base in the average profiles corre-

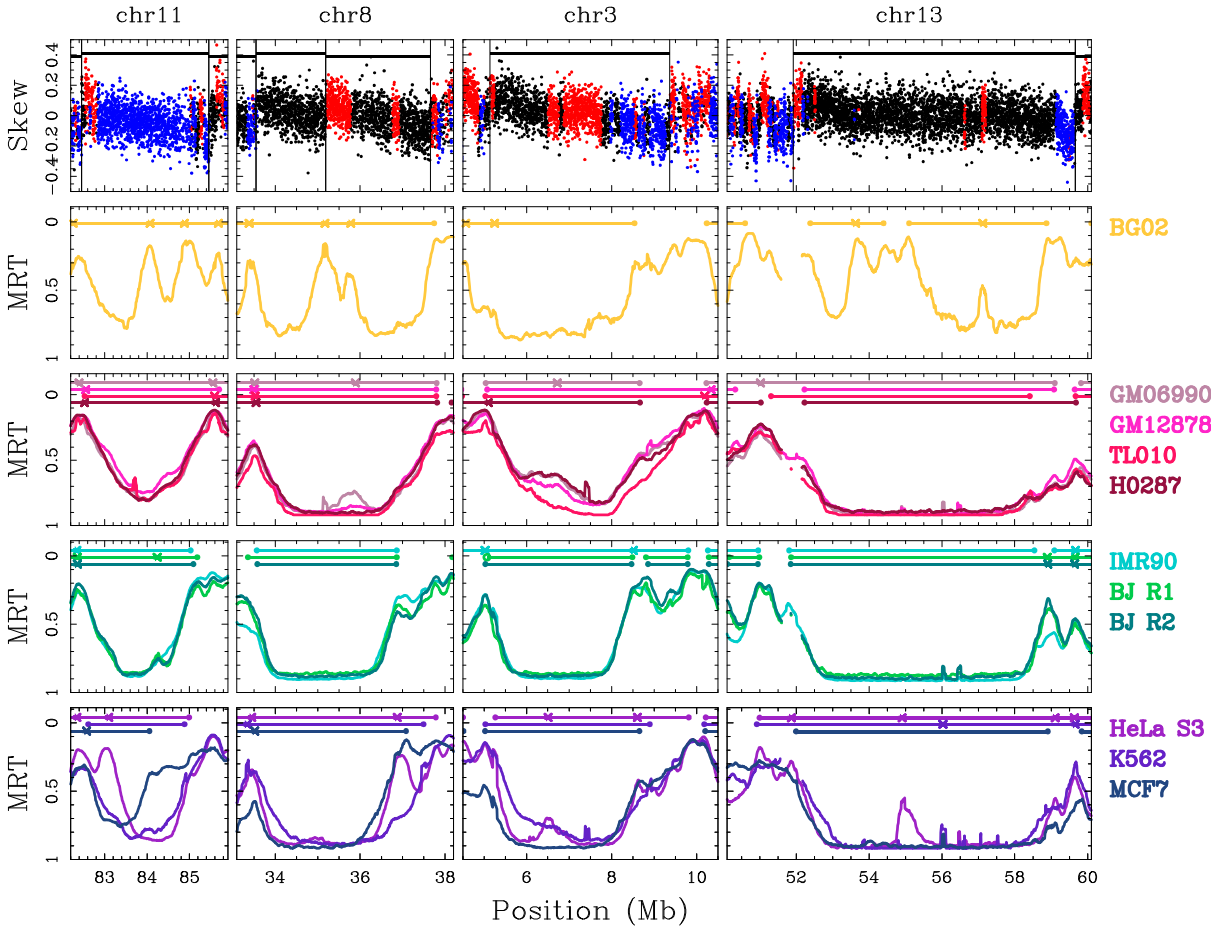


Figure 4.6. MRT U- and split-U-domains in different human cell lines. From top to bottom, skew profiles calculated in non-overlapping 1kb windows with color dots corresponding to intergenic (black), (+) genes (red) and (-) genes (blue), MRT profiles in ES (yellow), lymphoblastoid (pink), fibroblast (green), cancer (purple) cell lines, plotted *vs* the native position along different chromosomes. First column: 3.6 Mb of chromosome 11, second column: 5 Mb of chromosome 8, third column: 6 Mb of chromosome 3, fourth column: 10 Mb of chromosome 13. The horizontal bars correspond to N/split-N (black) or U/split-U (colors) domains, the vertical bars correspond to the N/split-N borders. The (x) on the horizontal colored bars correspond to a MRT domain peak common border of two juxtaposed domains and the (•) correspond to an asymetric border common to a U/split-U-domain and a MRT plateau.

sponding to the late replicating central region. This central region appears on average, to replicate later than the center of U-domains (Fig. 4.5). When looking at individual large domains (Fig. 4.6, columns 2 to 4), we see that the domain borders replicate early followed by a transition region (of few hundreds kb) before reaching a plateau of a late replicating central region. Interestingly, for a larger domain size, we observe a larger central region (see for example the difference between chromosomes 8 and 13 on Figure 4.6, columns 2 and 4). These domains were not further analysed in previous studies dedicated to the characterisation of replication U-domains [44]. In this thesis and in the light of the new skew split-N-domains, we analyse these large MRT domains that we will refer to as replication split-U-domains consistent with their shape of a U splited in half.

Note that in regions where the split-U-domains seem to be conserved in differentiated

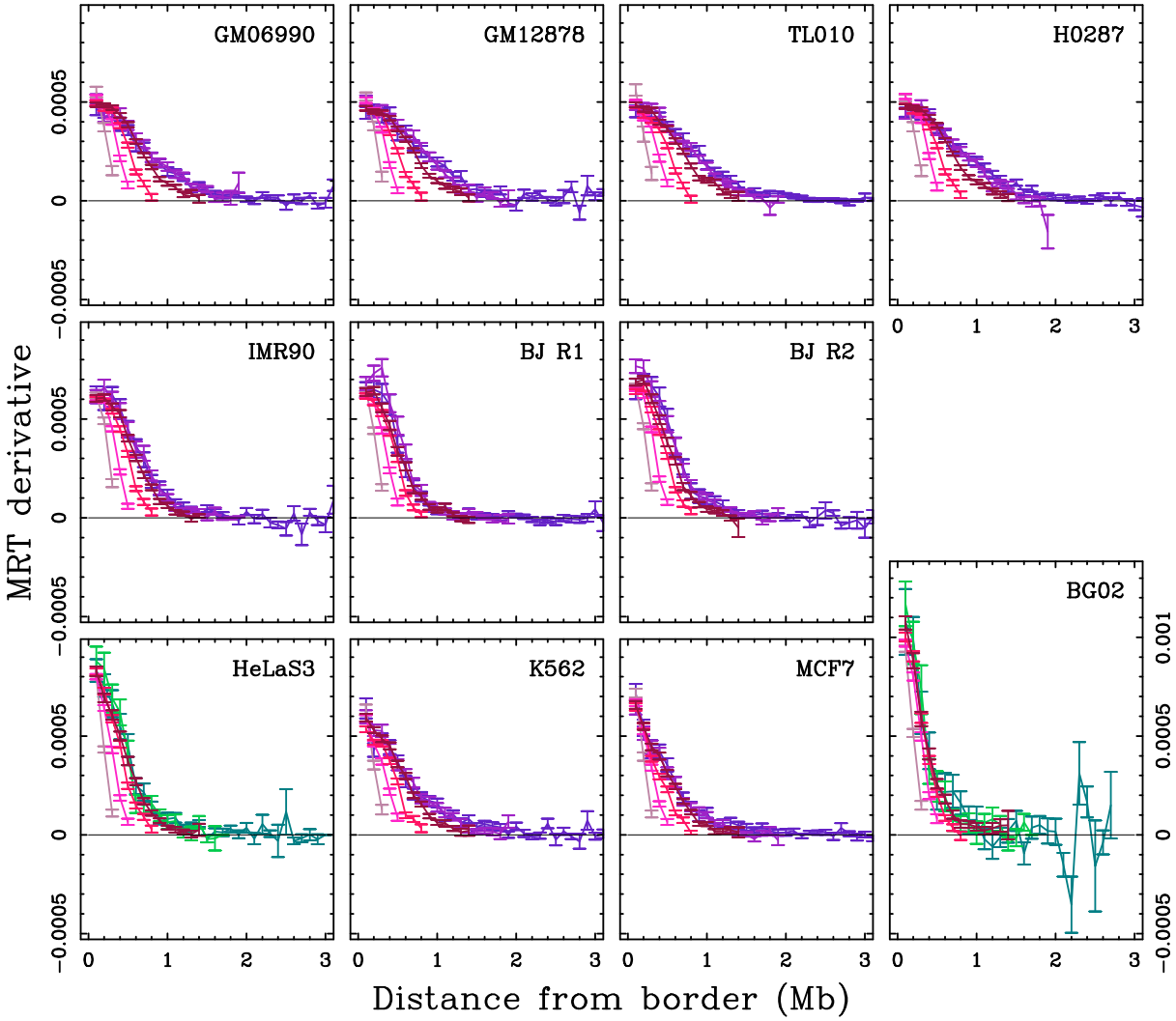


Figure 4.7. Mean MRT derivative. Mean derivative of the MRT *vs* the distance to the closest domain border in different cell lines. U-domains were grouped into 4 categories: $L < 0.8$ Mb (light pink), 0.8 Mb $\leq L <$ 1.2Mb (pink), 1.2 Mb $\leq L <$ 1.8 Mb (magenta) and 1.8 Mb $\leq L <$ 3 Mb (dark magenta). The split U-domains for lymphoblasts (GM06990, GM12878, TL010, H0287), fibroblast (IMR90, BJ (R1 and R2)) and cancer (K562, MCF7) cell lines were grouped in 2 categories: $3 \leq L < 4$ Mb (light purple) and $L \geq 4$ Mb (dark purple). For BG02 and HeLaS3 split-U-domains were grouped in 2 categories: $3 \leq L <$ 3.5 Mb (light green) and $L \geq$ 3.5 Mb (green). The MRT derivatives were computed from the MRT profiles using the wavelet-based methodology at 100 kb described in Appendix A (Section A.2)

cell types, we observe smaller domains for BG02 (Fig. 4.6, column 4). This property of BG02 embryonic stem cell line will be discussed along with the domains conservation in Section 4.2.1. For now, we investigate the existence of a characteristic limiting length and/or time in the MRT derivative.

Figure 4.7 shows for different domain size categories, the MRT derivative as a function of the distance to the borders, in different cell lines. The MRT derivative is computed from the MRT profiles using the wavelet-based methodology at scale $s = 100$ kb described in Appendix A (Section A.2). Independently of the cell line, for the smaller

domain size categories ($L < 0.8$ Mb, 0.8 Mb $\leq L <$ 1.2Mb, 1.2 Mb $\leq L <$ 1.8 Mb) the MRT derivative decreases linearly. However, for the larger domains, the curves seem to decrease linearly from the border over a certain distance depending on the cell line and then they become flat. For the embryonnic stem cell line (BG02), characterised by smaller U-domain sizes [44], the curves start to flatten before 1 Mb. This distance seems to be larger for all differentiated cell lines (except for HeLaS3 where it is also $\lesssim$ 1 Mb): from $\sim$ 1.2 Mb for fibroblasts and MCF7, to $\sim$ 1.4 Mb for lymphoblasts and K562.

Since the skew S and the derivative of the MRT profile both reflect the average fork polarity at a given locus (Section 2.1.4), these results confirm the existence of a certain limiting characteristic size in the behaviour of polarity *vs* distance to replication timing domain borders, as previously observed for replication skew domains (Fig. 4.4). However, it is not the same for all cell lines and we do not systematically recover the 1.5 Mb characteristic size found for split-N-domains. This raises the question of the origin of this limiting characteristic size: should it be interpreted as a characteristic length-scale or as a characteristic time-scale? Figure 4.7, shows that the limiting scale for BG02 seems to be smaller than the ones observed in the differentiated cell lines. The S-phase is shorter in BG02 than in the other cell lines which could be the reason why the characteristic scale is shorter. We thus explore the hypothesis that we are in fact facing a characteristic time scale.

4.1.3 Evidencing a characteristic time-scale

In order to test the above hypothesis, we look at how the MRT at the domain border affect the above curves. Figure 4.8 shows the distribution of MRT values at domain borders in the different size categories. For BG02, most of the replication domain borders replicate with a MRT<0.5: the histogram has a peak around MRT $\sim$ 0.25 for the U-domains and around MRT $\sim$ 0.35 for split-U-domains (Fig. 4.8). A smaller secondary peak appears also for the split-U-domains at MRT $\sim$ 0.7; corresponding to less than 10% of the borders as illustrated by the cumulative histogram where we clearly see that $\sim$ 80% of the borders replicate before the mid S-phase (Fig. 4.8). For the differentiated cell lines, we see two peaks in the domain border MRT histogram, a first one around $0.2 \leq$ MRT ≤ 0.3 and a second one around $0.5 \leq$ MRT ≤ 0.7, this second peak appears to be more pronounced for the larger domains, as illustrated by the shift to the right of the cumulative histogram (Fig. 4.8). Nevertheless, in general, for all the cell lines, the cumulative curves of the MRT distribution (black line in the top pannel of Fig. 4.8) seem to follow the same behaviour, many origins ($\sim$ 30%) fire early in the S-phase (MRT<0.25), followed by other origins ($\sim$ 30%) replicating in the mid S-phase ($0.25 \leq$ MRT < 0.5), and the remaining firingh late in the S-phase (MRT>0.5). Accordingly, we split the domains in three groups: early replicating borders (MRT<0.25), mid replicating borders ($0.25 \leq$ MRT <0.5) and late-replicating borders (MRT ≥ 0.5).

Interestingly, when we look at the MRT derivative as a function of the distance to the closest domain border in those timing groups, we clearly see that the curve corresponding to late MRT borders tend to flatten before the ones corresponding to the early MRT borders (Fig. 4.9 and Supplemental Figures in Annexe B: Figs B.1, B.2, B.3, B.4). For large split-U-domains (L $\geq$ 3.5 Mb for BG02 and HeLaS3, and L $\geq$ 4 Mb for the other cell lines), Figure 4.9 shows that the distance at which the curves change behaviour is

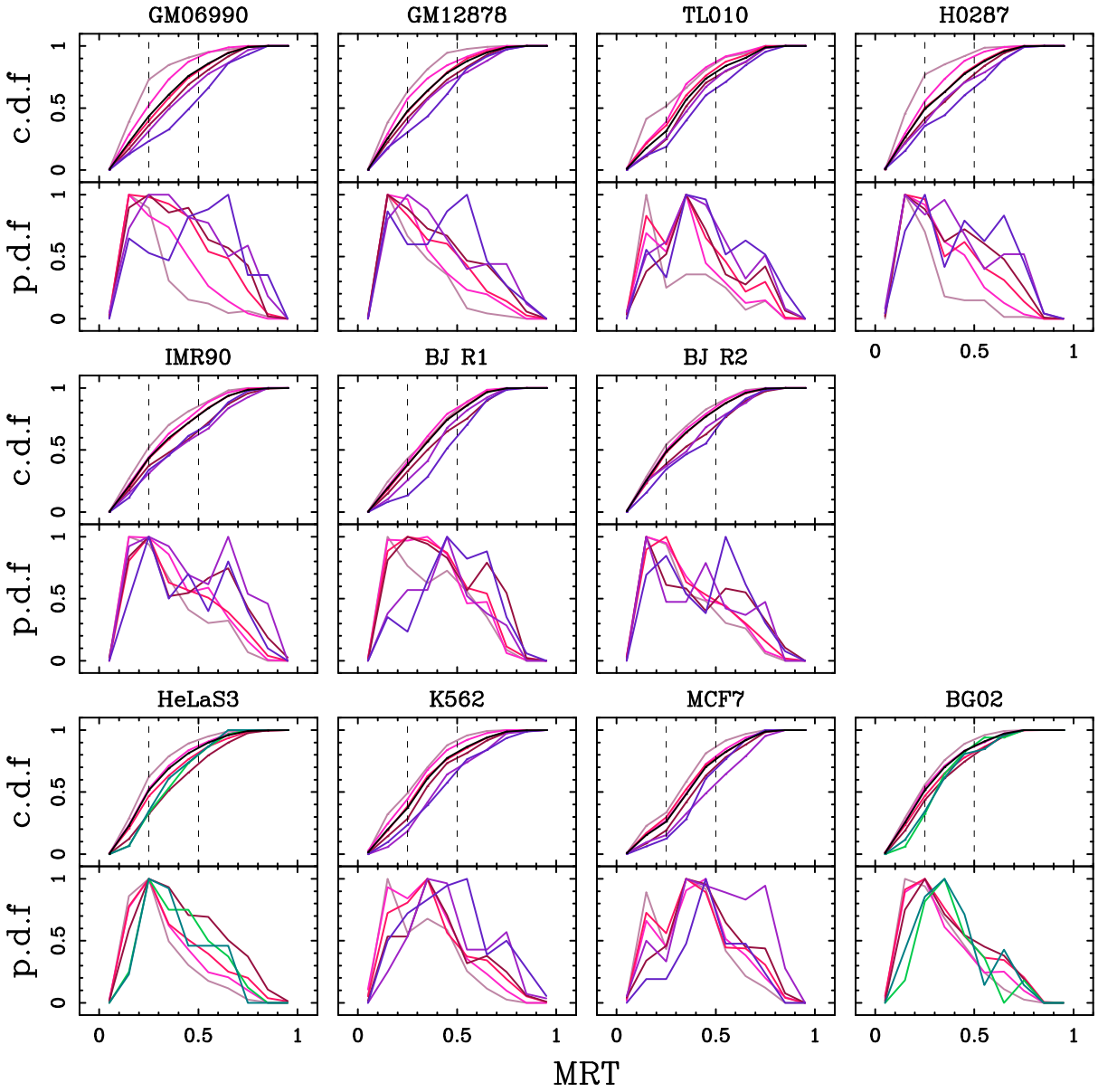


Figure 4.8. MRT at replication domain borders. For each cell line, and for each replication timing domain size categorie (same categories and color coding as in Fig. 4.7), the bottom pannel shows the histogram (p.d.f) of domain border MRT and the top pannel shows the cumulative histogram (c.d.f). In the top pannel, the black curve represent the mean over all the domains and the dashed black vertical lines at MRT equal to 0.25 and to 0.5 delimiting the three zones defining the borders' timing categories.

not the same for the three timing categories: the curves corresponding to late replicating borders flatten at smaller distances followed by the curves corresponding to mid replicating borders, in turn followed by the curves corresponding to early replicating borders. In BG02 and HeLaS3, the curves corresponding to late replicating borders change behaviour around $\sim$ 600 kb, around $\sim$ 900 kb in fibroblasts and MCF7 and around $\sim$ 1 Mb in lymphoblasts and K562. While for the early replicating borders, the curves flatten around $\sim$ 1 Mb in BG02 and HeLaS3, around $\sim$ 1.4 Mb in lymphoblasts, $\sim$ 1.6 Mb in fibroblasts and $\sim$ 1.8 Mb in K562 and MCF7. This favors the hypothesis of a characteristic time scale organisation rather than a characteristic length scale, because if we

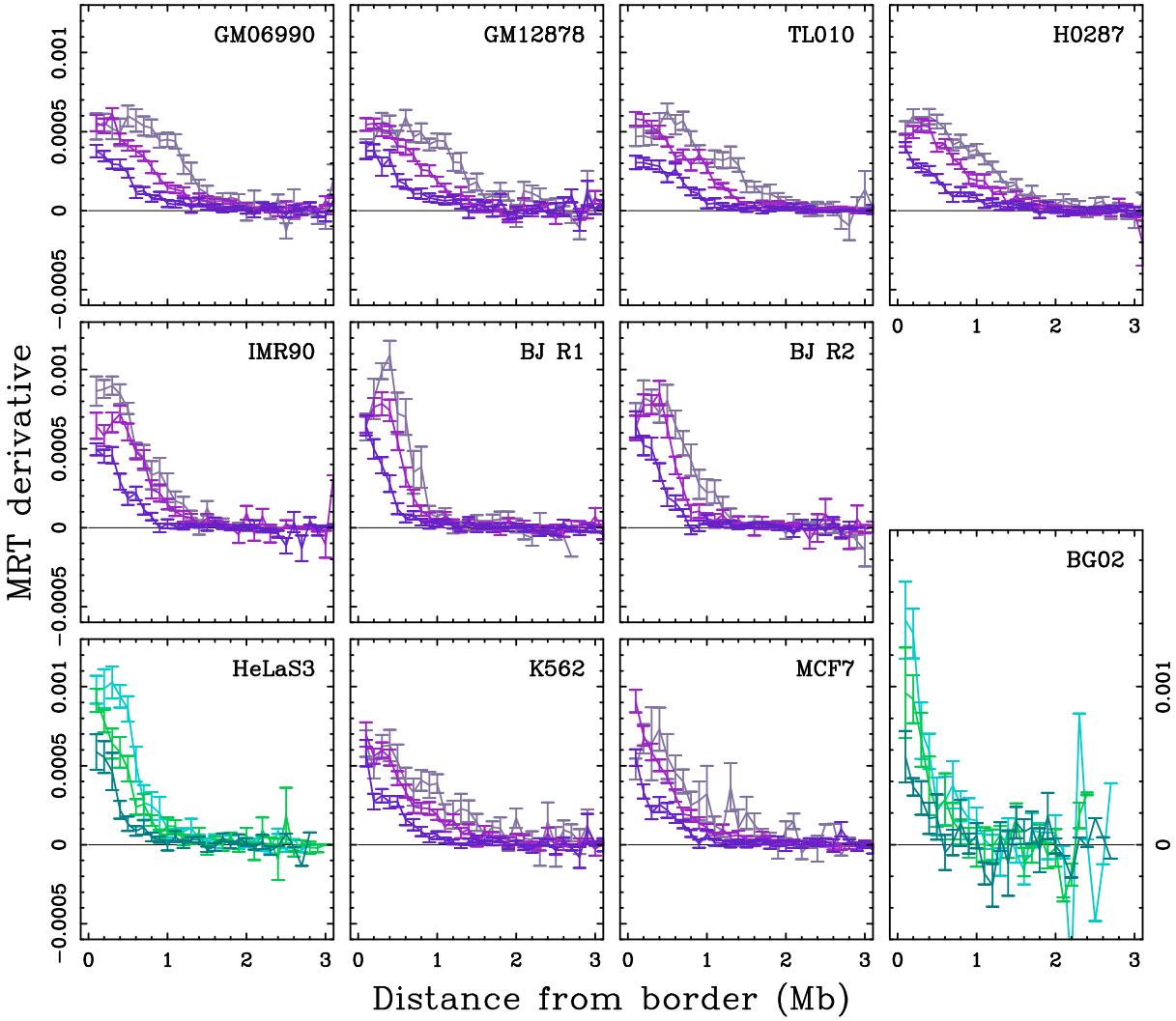


Figure 4.9. Mean MRT derivative. Mean derivative of the MRT *vs* the distance to the closest domain border in different cell lines for the largest split domains (L $\geq$ 4 Mb shades of purple and L $\geq$ 3.5 Mb shades of green) and the different timing categories: early timing (MRT<0.25) light purple and light green, mid timing (0.25 $\leq$ MRT <0.5) purple and green, and late timing (MRT >0.5) dark purple and dark green.

shift the curves of the late MRT borders (or the mid replicating borders) they will superimpose with the ones of the early replicating borders. In other words, the apparent velocity of replication (the inverse of the MRT derivative (page 33)) depends on the timing in the S-phase or the timing left to reach the end of the S-phase. As the end of the S-phase approaches, the apparent speed of replication increases, and diverges in the latest moments of the S-phase. For the late replicating borders, we observe a smaller MRT derivative that translates in term of apparent speed to a faster velocity: when the origins fire late, replication proceed faster at domain borders. The presence of the null plateau can be explained by the fact that there is no preferred directionality of the replication fork suggesting a random initiation of replication in these late timing regions in order to complete replication before the end of S-phase. Indeed, this is consistent with the vision of constant timing regions (CTRs) replicating late in the S-phase by multiple origins. Altogether, these observations are in agreement with the domino model of replication (page 34), where an accelerating wave of replication likely initiates at the domain

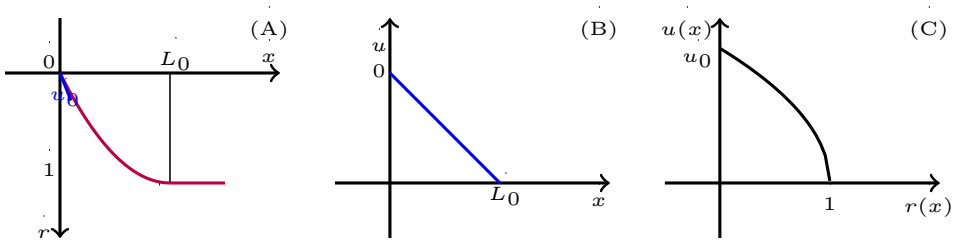


Figure 4.10. Linking the MRT derivative and the MRT. (A) $r(x) = ax^2 + bx + c$, the parabolic MRT as a function of the position x in the vicinity of a replication domain border. (B) $u(x) = 2ax + b$, the derivative of the MRT as a function of the position. (C) The resulting MRT derivative as a function of the MRT (Equation (4.1)).

borders and propagates towards the center via a cascade of origins firing [69, 70] and this independently of the domain size. Thus, replication would first initiate in efficient zones specified by open chromatin [51], followed by progressive activation of secondary origins in less open chromatin due to the approach of an incoming fork and the effective speed of this cascade process would only depend on the time left to complete the S-phase.

In the following, we further test this model with the specific goal to establish a link between the MRT derivative (reflecting the effective velocity of the replication progression) and the MRT.

4.1.4 Modeling replication inside MRT domains

We note $r(x)$ the MRT at a position x and $u(x)$ the MRT derivative at that position. In the proposed model, starting from a replication origin at a split-U domain border, $r(x)$ follows a parabola: $r(x) = ax^2 + bx + c$ (Fig. 4.10 A) and its derivative $u(x) = 2ax + b$ is linear (Fig. 4.10 B), up to the end of S-phase where the derivative u reach 0. Using the following boundary conditions: $r(0) = 0$, $r(L_0) = r_f = 1$, $u(0) = u_0$ and $u(L_0) = 0$, where $r = 0$ is the beginning of S-phase, $x = 0$ is the position at the domain border, r_f is the timing at the end of the S-phase and L_0 the characteristic length (position where the curve change behaviour), it is easy to obtain: $r(x) = -\frac{u_0^2}{4}x^2 + u_0 x$ and $u(x) = -\frac{u_0^2}{2}x + u_0$, leading to the following relation between $u(x)$ and $r(x)$ (Fig. 4.10 C):

$$u(x) = u_0(1 - r(x))^{1/2}. \tag{4.1}$$

Equation (4.1) represents the behaviour for $x \in [0, L_0]$ but also applies in the central region of split-U-domains where $u = 0$ and $r = 1$. Given a constant value to u_0 (slope at time $r = 0$), Equation (4.1) establishes that there is a unique relationship between MRT and its derivative within split-U-domains. To test this model, we start considering large MRT split-U-domains ($L \geq 4$ Mb) in GM06990 cell line, where the MaOris are "far" enough, so that the replication cascades originating from these MaOris do not influence each others. We consider all half split-U-domains *i.e.* from each border to the center, we sort them according to the MRT at the domain border: from the earliest replicating borders to the latest replicating borders, we calculate the mean of $u(x)$ and $r(x)$ over groups of 15 domains where the timing at domain borders are similar (Fig. 4.11 A). The data for the different groups of domains impressively fall on a unique curve, demonstrating that there indeed exists a unique relationship between MRT and its derivative along split-U-domains.

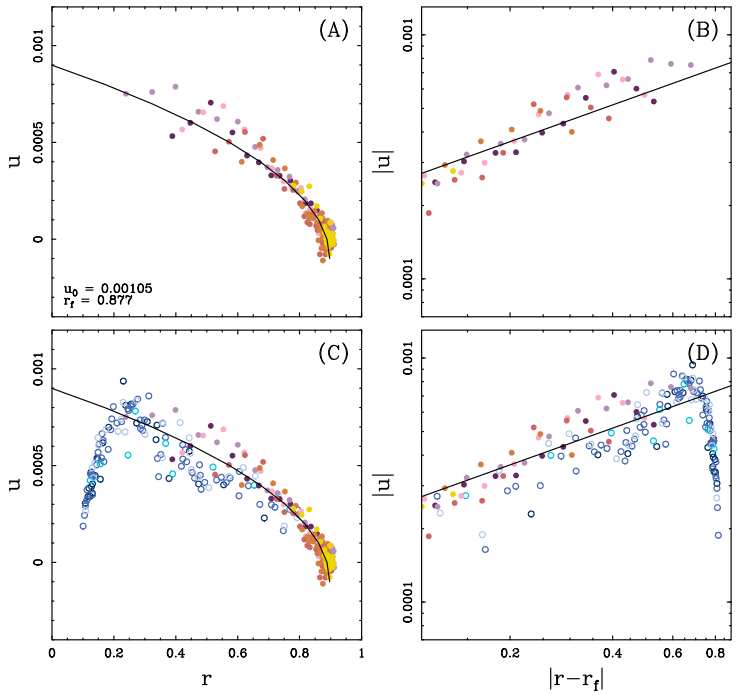


Figure 4.11. MRT derivative as a function of the MRT in GM06990 cell line. (A) Mean MRT derivative *vs* the MRT, the coloured bullets (•) corresponding to the mean values along large split U-domains ($L \geq 4$ Mb), each color correspond to the mean over 15 domains after sorting the domains relatively to the MRT value found at the border. The black curve represents the fit of the parabolic model (Equation (4.3)) to the presented data with $u > 0$. (B) The absolute value of the MRT derivative *vs* the absolute value of $(r_f - r)$, in a loglog plot where we used the data from (A) with same color coding as in (A). The black line is a linear function of slope (1/2) to guide the eyes. (C) Same as (A) where the supplementary blue (○) correspond to the mean MRT derivative values at the U-domain borders over groups of 15 domains sorted according to their border timing, in different size categories $L < 1$ Mb, 1 Mb $\leq L <$ 2Mb, 2 Mb $\leq L <$ 3 Mb and 3 Mb $\leq L <$ 4 Mb (from cyan to dark blue). (D) same as (B) with the data presented in (C).

We then proceed to fit these data by our model predictions (Equation (4.1)). Note that the 1 in Equation (4.1) comes from the fact that in the theory we considered $r_f = 1$. However, given the way the MRT is computed (Section 2.1.3.2), r is never equal to 1. Thus to fit the curves, we do not force r_f to be 1, we leave it as a free parameter and Equation (4.1) becomes:

$$u(x) = u_0(r_f - r(x))^{1/2} \qquad \forall x \in [0, L/2], \tag{4.2}$$

which is equivalent to:

$$r(x) = -(\frac{u}{u_0})^2 + r_f \qquad \forall x \in [0, L/2]. \tag{4.3}$$

Equation (4.3) is easy to fit using Matlab with a linear fit model of the form $ax^2 + b$. The deviation of the data to the model predictions can be further quantified in a logarithmic representation (Fig. 4.11 B), where the data present a slope slightly greater than the value 1/2 predicted by the model . The average MRT slope at U-domain borders closely match the one at split-U-domain borders (Fig. 4.7) suggesting that at U-domains borders,

the proposed cascading model process might not be influenced by the cascade emanating from the opposite borders so that the "universal" relationship between MRT and its derivative (Equation (4.1)) could also be observed at these loci. Indeed, averaging u and r in groups of 15 U-domain borders[†] sorted by their MRT value, we obtain data points that nicely follow the "universal" behaviour observed along split-U-domains (Fig. 4.11 C, D). The decrease observed at the earliest timing can be attributed to the impossibility to precisely compute timing change in this range of the RepliSeq data protocol (Section 2.1.3.2).

We reproduce this analysis for different cell lines (Fig. 4.12 and Supplementary Fig. B.5). Using the initial data sets for the MRT computation (Supplementary Fig. B.5), we notice, for some cell lines, problematic behavior on the u *vs* r representation, especially at early replicating borders. The most critical case is observed for MCF7 cell line where a sharp discontinuity is apparent between the very early U-domain borders and the other ones (Supplementary Fig. B.5). This is indeed due to the way the data were normalised. As explained in Chapter 2 (page 27), the normalisation procedure consists in normalising first the columns to ensure a constant rate of DNA synthesis along S-phase, followed by column normalisation to ensure that each locus is replicated once and only once per cell cycle. The second normalisation is performed a second time after denoising. This second round of locus normalisation can affect the first one and leads to not normalised lines meaning that constant rate of DNA synthesis is not guaranteed anymore. Therefore, we experiment an iterative normalisation method consisting in performing line and column normalisation iteratively until no change is observed in the results[‡]. Results obtained with this iterative normalisation are illustrated on Figure 4.12. Mostly, the changes are observed for early replicating borders, probably because as illustrated on Figure 2.9 (D G1), the noise in the data are more pronounced in G1.

Note that the iterative normalisation does not change drastically the results it only fixes the noise observed in G1. Indeed, when reproducing for example the analysis reported in Figure 4.9 with the data normalised iteratively (Supplementary Fig. B.6), the same results are observed starting with a higher MRT derivative.

For each cell line, the MRT derivative *vs* the MRT data points for the different groups of split-U-domains fall on a unique "universal" curve (Fig. 4.12). Hence, there exists a unique relationship between MRT end its derivative along split-U-domains, and this in all the analysed cell lines. Interestingly, this "universal" behaviour extends to U-domain borders. Finally it also appears that the relationship between MRT and its derivative is well captured by Equation (4.3) in all cell lines (Fig. 4.12). If we consider that r_f is determined from the computation of MRT, the only free parameter in the model is u_0. From the fit obtained in Figure 4.12 we can see that lymphoblasts have a smaller u_0 than fibroblasts, resulting in a higher apparent speed of the replication fork in lymphoblasts which is consistant with the results obtained in [70].

[†]As the MaOris are determined at $\sim$ 200 kb precision, we consider for the computation of the MRT derivative the third 100 kb window from each U domain border.

[‡]Typically the results are robust after 15 iterations.

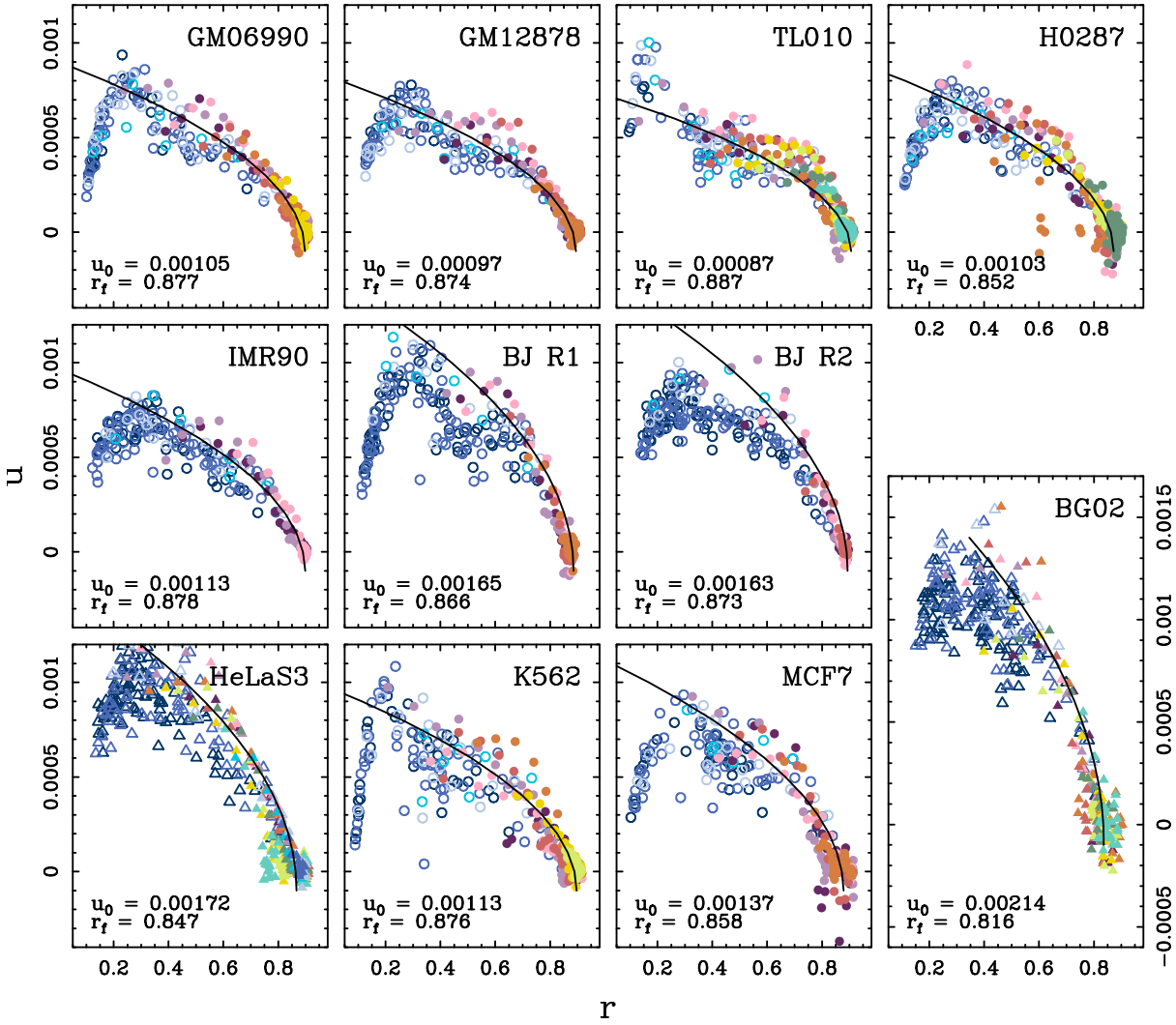


Figure 4.12. MRT derivative as a function of the MRT. Same as in Fig. 4.11 C in different cell lines, with the same colour coding. For HeLaS3 and BG02 different size categories were considered (i) for U-domains: $L < 1$ Mb, 1 Mb $\leq L <$ 2 Mb, 2 Mb $\leq L <$ 2.5 Mb from light to dark blue (△) and (ii) for split-U-domains: $L \geq 2.5$ Mb coloured (▲).

4.1.5 A universal cascade model of origin firing

We have identified a new type of replication domains, the split-U-domains covering ∼ 10% of the genome (Table 4.1). Like the U-domains, split-U-domains are characterised by early replicating regions around the borders. However, what distinguish them from the U-domains is their central region exhibiting a late replicating plateau in the MRT profiles which length increases with the domain size (Figs. 4.5 and 4.6). Thus, the MRT increases from the border to a certain distance reaching a relatively high value in the late central region. The observation that in large replication domains, the MRT derivative decreases to zero over a distance independent of the domain size (Fig. 4.7), suggests that there must exist of a limiting time or length scale. Interestingly, when sorting the domains according to the MRT at their borders and analysing separately the domains of early, mid and late replicating borders, we have observed that when the replication starts later, the plateau is reached faster (Fig. 4.9). This favours a model with a characteristic time scale, as

	BG02	IMR90	BJ (R1)	BJ (R2)	GM06990	GM12878	TL010	H0287	K562	HeLaS3	MCF7	N
BG02	30	7	8	10	6	6	10	8	10	3	8	12
IMR90	7	62	36	34	19	19	24	17	24	15	16	15
BJ (R1)	8	39	76	56	21	24	32	24	32	17	23	23
BJ (R2)	8	36	54	67	21	20	29	22	28	18	22	22
GM06990	6	19	20	22	105	58	63	65	30	14	30	19
GM12878	11	19	24	23	60	95	57	48	31	14	26	14
TL010	11	22	30	29	62	58	150	65	41	13	24	28
H0287	8	20	25	24	68	49	67	122	39	13	27	25
K562	11	19	32	28	27	31	40	35	107	15	30	23
HeLaS3	4	8	14	13	12	12	11	8	14	38	12	10
MCF7	7	15	22	19	27	27	23	22	29	14	84	21
N	16	16	24	22	19	15	26	25	23	11	26	97

Table 4.2. Number of matches between MRT (split-) U-domains of different cell lines. A split-U-domain (column) is considered to have a matching (split-) U-domain (row) when at least 80% of these (split-) U-domains were common to the two cell lines.

confirmed by the fact that the MRT derivative and thus the replication velocity are directly linked to the MRT (Equation (4.1) and Fig. 4.12). Altogether these results suggest that replication inside split-U- and U-domains (referred to as (split-) U-domains) follows a "universal cascade model" that only depends on the time left till the end of the S-phase.

Questions we now address are whether these split-U-domains are conserved across cell lines as the U-domains [44] and whether their borders present the same characteristics as MaOris as previously observed for split-N-domain and N-domain borders [359].

4.2 Segmentation of the genome in replication domains

Replication U-domains were shown to be at the heart of the organisation of the replication spatio-temporal programme [44]. In the previous sections, we described a new kind of megabase-sized MRT domains where the early replicating "MaOris" are far from each other. In this section, we discuss to which extent split-U-domains present similar properties as the U-domains. In fact, U-domains were robustly observed across cell lines [44]. First, we test to which extent split-U-domains are also conserved across cell lines (Section 4.2.1). Then, we ask whether the "MaOris" at split-U-domain borders have similar functional properties as the "MaOris" at U-domain borders (Section 4.2.2). Then we extend the chromatin state organisation study to these novel replication domains (Section 4.2.3). We will systematically compare the results in different cell lines where the data are available: H1 ES (with BG02 replication domains as surrogates), GM06990, IMR90 (with Nhdfad chromatin states), HeLaS3 and K562.

4.2.1 Split-U/N-domain conservation

Figure 4.6 illustrates the remarkable pattern of conservation of (split-) U-domains and (split-) N-domains between different cell lines. The relatively small (3 Mb) N-domain of chromosome 11 colocalises with U-domains in the four lymphoblast cell lines (GM06990, GM12878, TL010, H0287), and with the robustly conserved U-domain in the fibroblast cell lines (IMR90, BJ R1 and R2). However, we observe specific active origins in BG02 (that are absent in differentiated cell lines) resulting in domains consolidation [147]: three

	BG02	IMR90	GM06990	K562	HeLaS3	N
Specific	8 (26.7%)	17 (27.4%)	47 (44.8%)	45 (42.1%)	15 (39.5%)	42 (43.3%)
Shared by 1	11 (36.7%)	21 (33.9%)	34 (32.3%)	36 (33.6%)	5 (13.2%)	35 (36.1%)
Shared by 2	7 (23.3%)	17 (27.4%)	19 (18.1%)	20 (18.6%)	13 (34.2%)	13 (13.4%)
Shared by 3	3 (10.0%)	5 (8.1%)	4 (3.8%)	5 (4.7%)	5 (13.1%)	6 (6.1%)
Shared by 4	1 (3.3%)	2 (3.2%)	1 (1.0%)	1 (1.0%)	0	1 (1.0%)
Shared by 5	0	0	0	0	0	0
Total	30	62	105	107	38	97

Table 4.3. Domain conservation across cell lines. Using conservation data summarised in Table 4.2, for each cell line we computed the number of split-U/N-domains that are cell line specific, or have a matching (split-) U-domain in 1,...,5 of the considered cell lines.

	BG02	IMR90	GM06990	K562	HeLaS3	N
Specific n=1	26.28%	17.91%	15.59%	15.75%	18.34%	17.06%
Shared by 1 n=2	22.32%	22.61%	19.67%	20.53%	23.29%	21.50%
Shared by 2 n=3	20.97%	23.00%	21.97%	20.79%	22.91%	22.11%
Shared by 3 n=4	15.66%	16.92%	21.45%	22.01%	18.48%	18.21%
Shared by 4 n=5	10.62%	14.26%	15.20%	14.85%	12.50%	14.23%
Shared by 5 n=6	4.15%	5.31%	6.12%	6.07%	4.49%	6.89%

Table 4.4. MRT split-U and U-domain borders conservation across cell lines. A domain border is considered to be shared by n cell lines if the distance between the border of the considered cell line and the distance to the closest border in the $n-1$ cell lines is less than 100 kb.

U-domains of BG02 consolidate to form a single U/N-domain in differentiated cell lines and in the germline. We observe the same consolidation effect of BG02 domains along chromosome 8 showing a split-U-domain in all considered differentiated cells but not in the germline (Fig. 4.6, second column). We observe another consolidation of BG02 U-domains along chromosome 13 leading to a "well" conserved split-U/N-domain in differentiated cell lines and in the germline (Fig. 4.6, fourth column). Note in this region, the activation in HeLaS3 cancer cell line of an origin present in BG02. When we look at the chromosome 3 fragment, we clearly see a conserved split domain even if the domain borders do not match exactly (Fig. 4.6, third column). In fact, in lymphoblast cell lines, the inflection point around 8.7 Mb on chromosome 3 was detected as a domain border in GM06990 and H0287 but not in GM12878 and TL010 simply due to the thresholds used in the domain border detection method (Appendix A, Section A.5). For the same domain, the other border (around 5 Mb) is not exactly in the same position in the four cell lines: the border seems to be shifted. However, the conservation of replication domains in different cell lines seems to be important in this region.

Conservation of U/N-domains between cell lines was previously described in [44]. Here, to determine genome-wide the amount of split-U/N-domains that are conserved in different cell lines, we compute for each cell line pair, the mutual covering of the corresponding sets of (split-) U-domains: a split-U-domain of the reference cell line has a matching (split-) U-domain when the two domains cover more than 80% of each others, *i.e.* two (split-) U-domains are shared by two cell lines if the number of base-pairs they share relatively to the size of the largest one is more than 80% (Table 4.2). Systematically, all the cell lines share *(i)* the least of their domains with BG02, for instance 6 for GM06990 (5.7%) and GM12878 (6.3%), probably because BG02 has the smallest number of split-U-domains which are also of the shortest length (Table 4.1), and *(ii)* at least $\sim$ 10% of their domains with another differentiated cell line. Interestingly, cell lines of the same type share more domains between them than with other types. For example, out of 105 domains, GM06990 shares 60 with GM12878, 62 with TL010 and 68 with H0287. To compare the conservation of the domains between more than two cell lines, we take a "representative" of each cell type; we choose: BG02, IMR90 (for fibroblasts), GM06990 (for lymphoblasts), K562 and HeLaS3 (for cancer cells) and the germline (split-N-domains). For each split-domain, we calculate the number of cell lines in which a matching (split-) U-domain is found (Table 4.3). None of the domains is shared by all the considered cell lines. However, each cell line shares more than 50% of its domains with at least one other cell line.

Finally, to evaluate the "consolidation" and the "boundary shift" [147] phenomena observed in Figure 4.6 (for both U/N- and split-U/N- domains), we address the border conservation across cell lines taking into account the resolution limit of the domains detection method. We consider that a border is conserved between cell lines if it is shifted by less than 100 kb (Table 4.4). BG02 presents the highest proportion of specific borders (26.28%) probably because it presents the smallest domains and consolidation is more frequent for this cell line. For each cell line, around 80% of its borders are shared by at least one other cell line. Altogether, these results along with the previous observation that U-domains are conserved across cell lines [44] suggest that both U-domains and split-U-domains covering together more than 60% of the genome (BG02 67.11%, IMR90 63.83%, GM06990 66.73%, K562 64.20% , HeLaS3 64.57%, N 59.22%) are highly con-

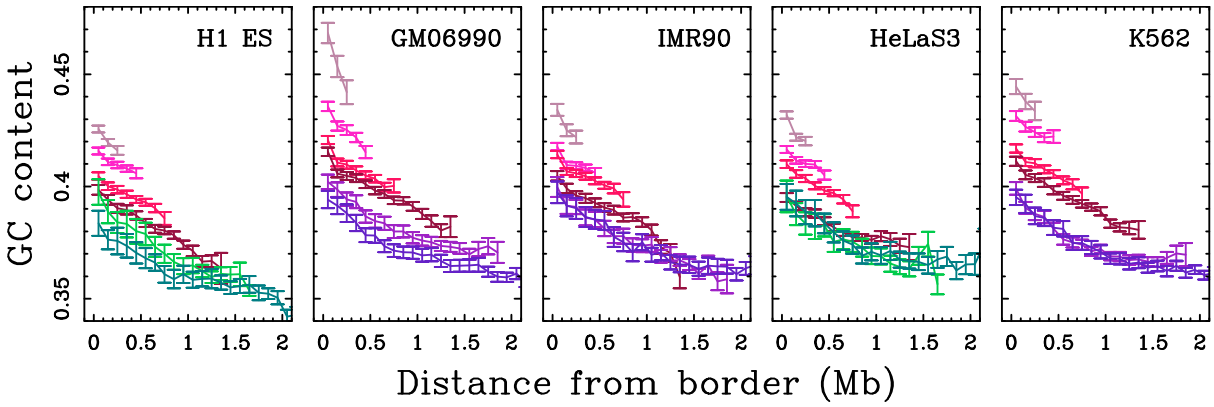


Figure 4.13. MaOris are GC rich. Mean (native) GC content (Section 4.4) calculated in 100 kb non overlapping windows *vs* the distance from the nearest (split-) U-domain border, in different cell lines and for different domain sizes (same color coding as in Fig. 4.7).

served between different cell lines and interestingly, their borders are shared by many cell lines.

In the following, we look at the organisation of the split-U/N-domains and address to which extent ubiquitous origins have a specific role in regulating the spatio-temporal replication programme.

4.2.2 Split-U-domain borders are "MaOris"

4.2.2.1 GC content

There is a definite evidence that the compositional heterogeneity in a DNA sequence correlates with its GC content [219] which is recognized as a fundamental property of the DNA and is likely to be one of the possible keys to understand the genome organisation [217–219, 360] . Moreover, it has been shown that GC-rich and GC-poor regions correlate well with early and late replicating domains. In this spirit, we look at the GC coverage along replication domains. Figure 4.13 shows for the different cell lines, and for different domain sizes, the GC content *vs* the distance to the nearest (split-) U-domain border. In all cell lines, we systematically see a decrease in the GC content from the replication domain borders to the centers. We also note that for the larger domains, the borders are less enriched in GC: for larger domains sizes, there is a systematic decrease in the position of the curves *i.e.* the curves are lower. Moreover, for the same size categories, the GC content is higher in GM06990 and K562 than in H1 ES. Hence, consistently with what has been shown for U-domains [44], GC content decreases from split-U-domain borders to center. However, GC content in split-U-domains is lower than in U-domains.

4.2.2.2 Gene organisation

Both N- and split-N- domain borders were shown to be active hypomethylated regions with a specific gene organisation around them [41, 359]: within domains, genes are primarily found around the borders. Similarly, when looking at the gene density along replication domains (Fig. 4.14), we clearly see a decrease of gene density from the domain borders towards the center. However, this sharp decrease of gene density takes place over a few 100 kb in contrast to the decrease of the GC content over the whole replication

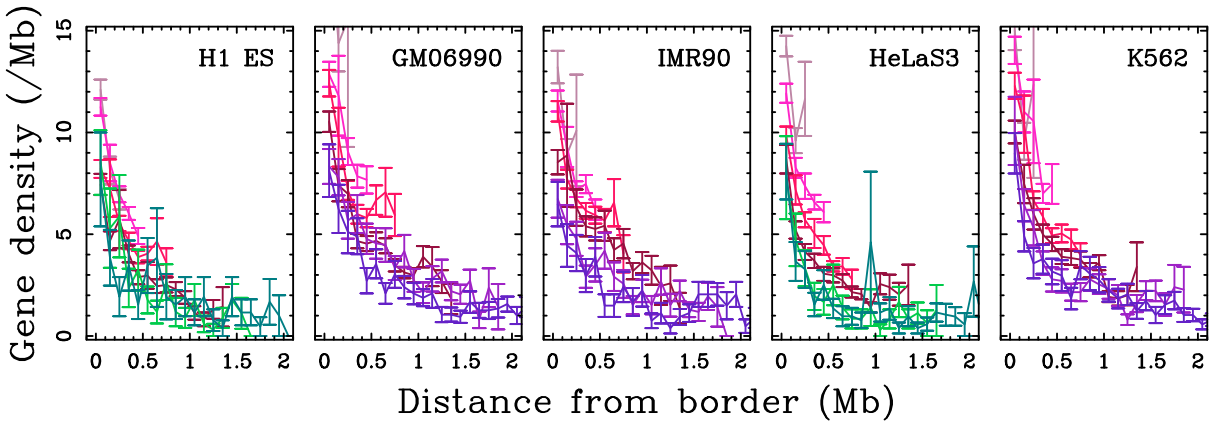


Figure 4.14. MaOris are gene rich. Gene density (Section 4.4) *vs* the distance from the nearest (split-) U-domain border, in different cell lines and for different domain sizes (same color coding as in Fig. 4.7).

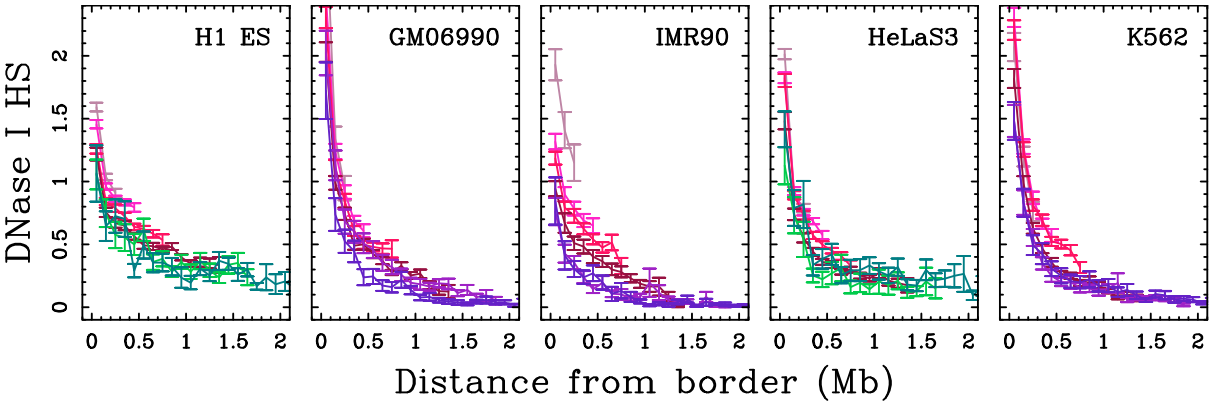


Figure 4.15. MaOris are open chromatin regions. Mean coverage by DNase I HS (Section 4.4) (relatively to the genome average) as a function of the distance the distance from the nearest (split-) U-domain border in different cell lines (same color coding as in Fig. 4.7).

domain. Interestingly, the central region of split-U-domains is found to be gene desert ~ 1 gene per Mb (Fig. 4.14), which is consistent with the fact that low GC regions are known to be poor in genes [217, 218, 360]. This is also consistent with previous observation of gene clustering at replication U-domain borders [44]. Thus, this is another evidence that split-U-domain borders have a similar role as the MaOris at U-domain borders.

4.2.2.3 Hypersensitivity to DNase I

High-throughput sequencing and whole-genome tiled strategies have been developed to identify DNase I hypersensitive sites (HS) as markers of open chromatin across the genome [184]. Mapping DNase I HS along MRT U-domains showed that domain borders are in ~ 200 kb regions enriched in open chromatin markers [44, 51]. In the same way, we look at the DNase I HS enrichment along split-U-domains (Fig. 4.15). Independently of the domain size, we observe that the mean coverage in DNase I HS is maximal at U- and split-U-domain borders and decreases significantly from the borders to the center. This means that, whatever the cell line, early replicating regions at split-U-domain borders as at U-domain borders [44], are at the center of a ~ 200 kb open chromatin region. Note

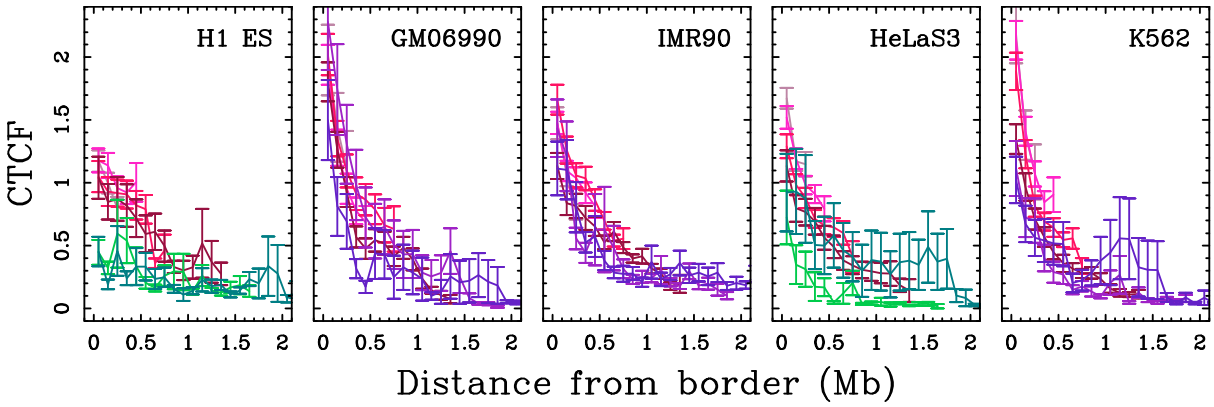


Figure 4.16. Enrichment in CTCF insulator-binding protein at MRT domains borders. Mean coverage by CTCF enriched peaks (relatively to the genome average); (Section 4.4) as a function of the distance from the nearest (split-) U-domain border in different cell lines (same color coding as in Fig. 4.7).

that we also observe a systematic decrease in the enrichment in DNase I HS with the domain size that is very clear for IMR90. Hence, we recover the previous observations along U-domains [44, 51], split-U-domain borders are open chromatin regions, whereas their central region is depleted in open chromatin marks consistently with the observation of very low gene density in these regions (Fig. 4.14).

4.2.2.4 CTCF Insulator-binding protein

When looking at CTCF insulator binding protein, that is known to be involved in chromatin loop formation conditioning communication between transcriptional regulatory elements [18, 169, 208, 209], we observe, for all differentiated cell lines (Fig 4.16) a decrease from the borders to the center of the MRT domains. Note that this decrease is not as stricking for H1 ES, specially for the split-U-domains where the CTCF coverage is more homogeneous. In fact, it has been shown that in pluripotent cells, CTCF distribution is different from its distribution in differentiated cells [54], combined to the presence of NANOG and OCT4 that were recently shown to contribute to the overall folding of ESCs genome via specific long-range contacts [361, 362] (Section 5.1). Hence, in differentiated cell lines, we recover for split-U-domains the decrease of CTCF abundance over a few 100 kb like previously observed for U-domain [44]. In H1 ES, the decrease in split-U-domains is much weaker than in U-domains.

☞ Altogether these results show that the "master" replication origins at split-U/N-domains borders are found in high GC, gene rich regions, enriched in DNase I HS and insulator binding protein CTCF as a signature of $\sim$ 200 kb of an open chromatin structure. The internal part of these domains are in low GC, gene poor regions, with low DNase I HS and CTCF coverages. This suggests that MaOris at split-U-domains borders are functionally equivalent to those bordering U-domain.

In the next section, we address chromatin states organisation (Section 2.2.2) in these replication domains and discuss the difference between ubiquitous and specific MaOris. We consider chromatin states in H1 ES (with BG02 replication domains as surrogates), GM12878 (with GM06990 replication domains as surrogates), Nhdfad (with IMR90 repli-

cation domains as surrogates) and K562. In the text, we refer respectively to those cell lines as H1 ES, GM06990, IMR90 and K562.

4.2.3 Chromatin state organisation inside replication domains

Replication U-domains were detected as regions having a U-shaped MRT profile bordered by two early replicating regions [43, 44]. It has been suggested that mapping the organisation of the four prevalent chromatin states within replication U-domains can provide complementary information on the genomic organisation of chromatin states and on the modifications of this organisation during cell differentiation. When concentrating our study on the replication U- and split-U-domains identified in H1 ES, GM06990, IMR90 and K562, the curves reported in Figure 4.17 superimpose very well onto each other so that the pattern previously observed in U-domains [53, 54] is likely to be conserved in split-U-domains. Some remarkable organisation of the four prevalent chromatin states is revealed with some notable differences that distinguish the global dynamical and accessible character of pluripotent chromatin from the expanding HP1-associated heterochromatin in differentiated cells. Consistently, for C1/EC1 and C2/EC2 the repartition inside MRT domains is similar between cell lines. This is no longer the case for C3/EC3 and C4/EC4. The highly expressed gene-rich open euchromatin state C1/EC1 is found to be confined in a closed ($\sim$ 200 kb) neighborhoud of the MaOris at replication domains borders and the polycomb repressed state C2 is mainly found occupying the mid S-phase 200-300 kb region away from U- and split-U-domain borders and this independently of the domain size. Interestingly, the C1/EC1 state is depleted in the central regions of the split-U-domains where the coverage in C1/EC1 reach a null plateau (Fig. 4.17). Unmarked C3 and constitutive C4 heterochromatin states homogeneously occupy large domain centers. However in IMR90 and K562, C3 is less abundant than C4 while in GM06990 C3 is more abundant relatively to C4 in large domain centers. Hence, C3/EC3 appears to be characteristic of split-U-domains in H1 ES and GM06990. In fact, the difference between C3 and C4 coverage is more pronounced for larger domains: we observe a plateau at 0.65 in C3 coverage for the split-U-domains of GM06990 whereas the C4 coverage reaches a plateau $\sim$ 0.75 in the central region of IMR90 and K562. In H1 ES, the distributions of EC3 and EC4 are more similar to the one of C3 and C4 in GM06990. EC3 is still depleted at domain borders and mainly covers the center of large domains. Importantly unlike C4, EC4 is now found at many domains borders as well as inside the domains. EC4 distribution in the split-U-domains of H1 ES is more homogeneous all along the domain. For IMR90 and K562 the results are consistent with the conclusion of [53, 54] suggesting that the replication "wave" starting from the early initiation zones at domain borders and propagating inside these domains via progressive activation of secondary origins [70, 363], actually progresses in a gradient of chromatin structures from openness (C1) to compactness (C3, C4) via the polycomb repressed state C2 [53, 54]. The homogeneous low distribution of EC4 in H1 ES reflects the fact that EC4 exhibits a much wider MRT distribution than C4 in differentiated cell lines. Note that as previously shown in Figure 2.13, C4 in GM06990 presents the wider MRT distribution between the differentiated cell lines concomitantly with its under representation in replication domains. For instance, 35.7% (H1 ES) as compared to 19.2% (GM06990), 5.5% (K562) and 4.2% (IMR90) C4 loci replicate early in the S-phase (MRT<0.5) [54].

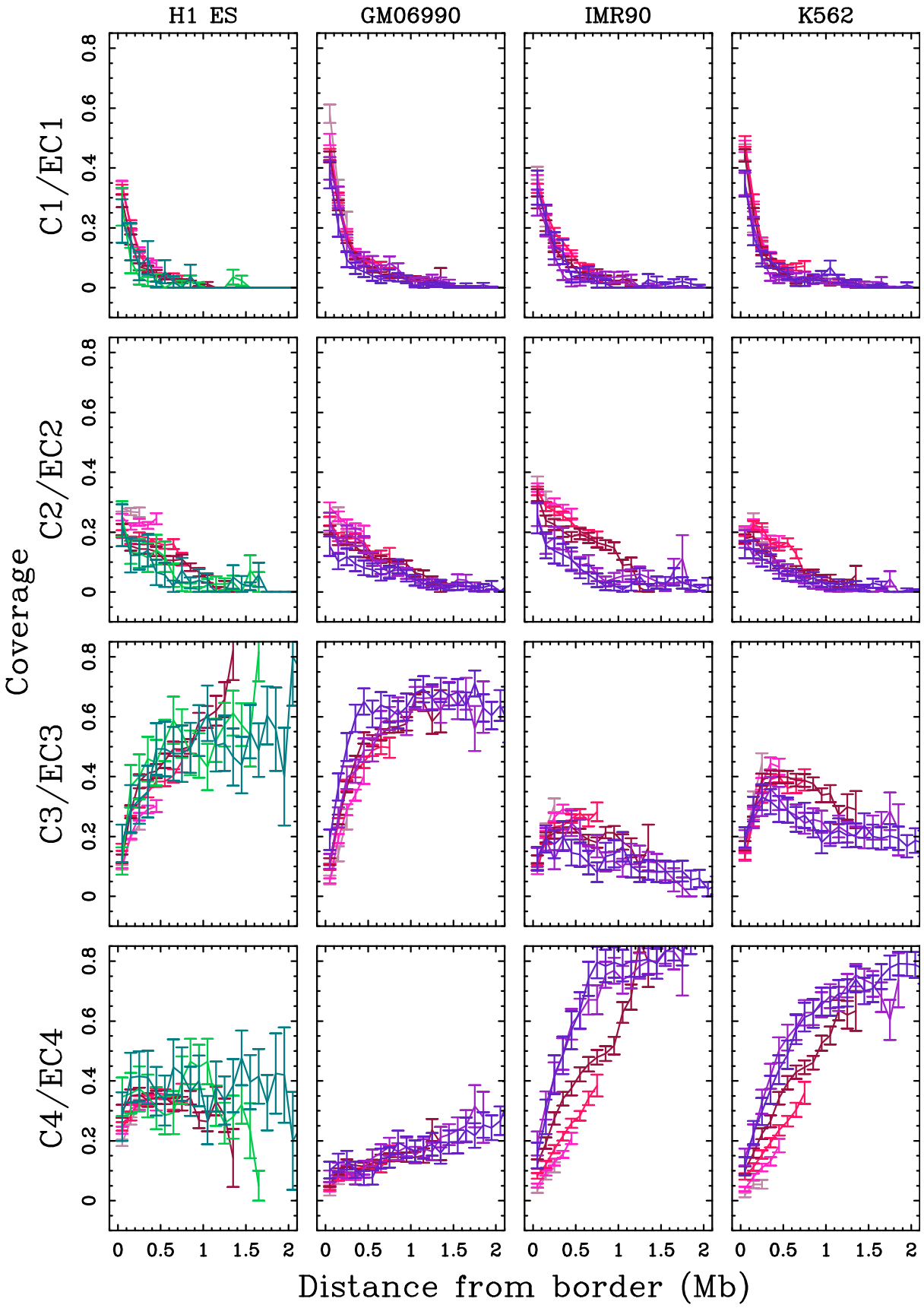


Figure 4.17. Chromatin state repartition inside replication domains. Mean coverage of chromatin states (Section 4.4) *vs* the distance from the nearest (split-) U-domain border, in different cell lines and for different domain sizes (same color coding as in Fig. 4.7).

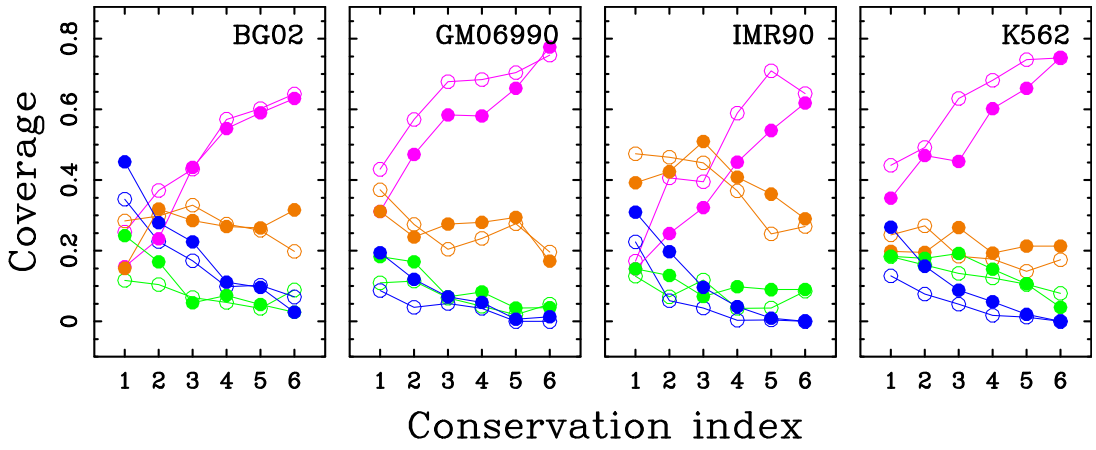


Figure 4.18. Chromatin state coverage at replication domain borders. Coverage of chromatin states at MRT domain borders as a function of the index of conservation (Table 4.4), in different cell lines. Colors correspond to chromatin states C1 and EC1 (pink), C2 and EC2 (orange), C3 and EC3 (green), C4 and EC4 (blue). Circles (○) represent U-domain borders and bullets (●) represent split-U-domain borders.

4.2.4 Ubiquitous *vs* specific MRT domain borders

MRT changes induced by differentiation result in an important changes in the number and size of replication U/N-domains [44] and split-U/N-domains (Table 4.1). Small neighbouring U/N-domains merged to become one large coordinately replicated domains (3 (resp. 2) domains merged to 1 in Fig. 4.6 column 1 (resp. 2)). This replication domain consolidation [38, 147, 148, 364] is thus the consequence of an active early replication initiation zone in stem cells that no longer fires early in somatic cells. We characterise this consolidation phenomenon from pluripotent to differentiated cell lines as well as between differentiated cell lines by defining an index of conservation n that quantifies the number of U- and split-U- domain borders in a given cell line that are shared by $n-1$ other cell lines (Table 4.4). When looking at the chromatin state coverage at replication domain borders according to the conservation index (Fig. 4.18) consistently with previous observation for U-domain borders [54], we see that ubiquitous origins at split-U-domain borders are always in a C1/EC1 open chromatin state. However, there is a striking difference for H1 ES specific borders that a significant proportion are found in EC4 in contrast to differentiated cell lines where specific borders are mainly in C1 or C2. Interestingly, regardless the cell line, split-U-domain borders are less in C1/EC1 and more in C4/EC4 than U-domain borders (Fig. 4.18).

Complementarily, we analyse the replication timing at U- and split-U-domain borders as a function of the conservation index. Ubiquitous borders always replicate the earliest. Interestingly, BG02 borders replicate earlier than in differentiated cell lines. Moreover, in all cell lines, MRT increases with the conservation index which is consistent with the increasing proportion of C1/EC1 borders (Figs. 4.18 and **??**).

The differences in chromatin state between cell line specific and ubiquitous domain borders raise the question: whether these two classes of domain borders present differences in the DNase I HS and CTCF coverage. In that respect, we look at the DNase I HS enrichment for ubiquitous and specific borders (Fig. 4.19). Both presente an enrichment in

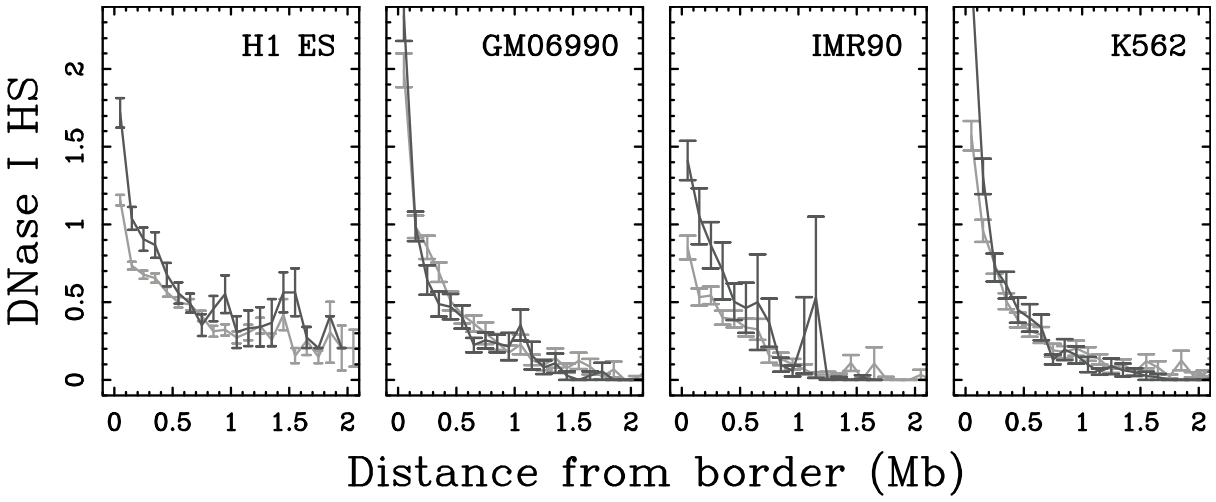


Figure 4.19. MaOris are open chromatin regions. Same as in Figure 4.15 for borders conserved in all considered cell lines (n=6, dark grey) and for specific borders (n=1, light grey).

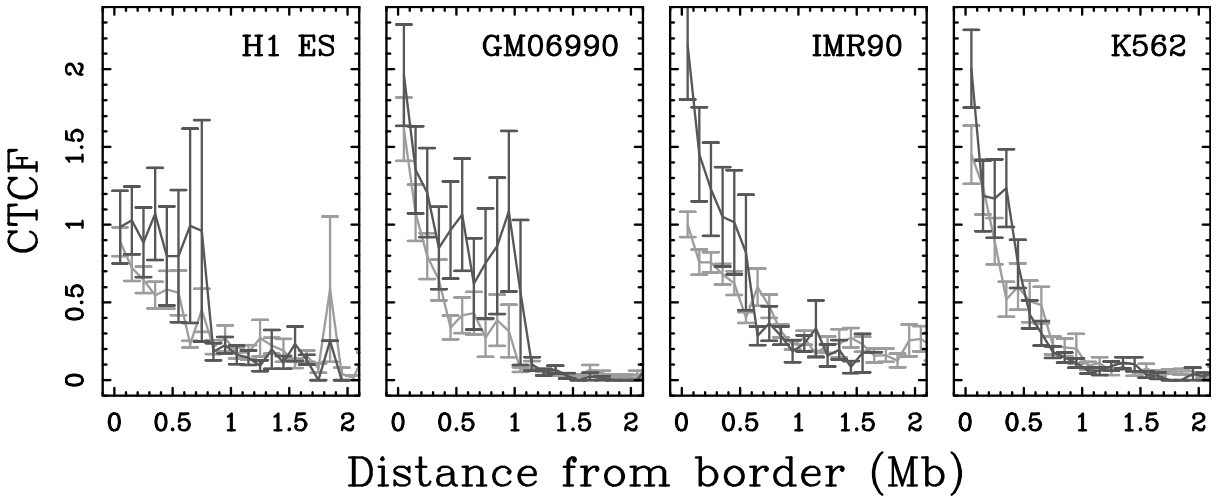


Figure 4.20. Enrichment in CTCF insulator-binding protein at MRT domain borders. Same as in Figure 4.16 for borders conserved in all selected cell lines (n=6, dark grey) and for specific borders (n=1, light grey).

DNase I HS relative to domains centers (the curves decrease from borders to the center). However, no clear difference is observed between the two types of borders. In the same manner, looking at CTCF enrichment (Fig. 4.20) does not reveal any difference between the ubiquitous and specific borders: both curves, in all considered cell lines, decrease from the borders towards the center of the domains. Only in IMR90 the ubiquitous borders seem to be more enriched in CTCF. This suggests that even if the ubiquitous borders are likely to be encoded in the sequence [365] they have the same functional role in terms of chromatin accessibility as the specific ones.

4.3 Towards a unified view of the replication spatio-temporal programme

The analysis of genome-wide MRT, and epigenetic data has revealed some 1D organisation of mammalian genomes into cell type dependent megabase-sized replication domains. In human, high (resp. low) GC, gene-rich (resp. poor), active early (resp. inactive late) CTRs covers about 25% (resp. 25%) of the genome that are replicated very early (resp. late) by the coordinated and almost synchronous activation of multiple origins more or less randomly spatially distributed [54, 69]. The larger these early and late CTRs, the higher the conservation level between pluripotent and differentiated cell lines [54]. The other half of the human genome is organised in tissue-specific U-shaped MRT domains bordered by "master" replication initiation zones enriched in open and transcriptionally active marks [43, 44, 53, 54, 161]. From those borders initiates a replication wave that further propagates and accelerates towards the domain center via the successive firing of secondary origins, more or less randomly dispersed, possibly by fork-simulated initiation [69]. When the distance between the two bordering MaOris exceeds $L \gtrsim 3$ Mb, some inactive late CTR emerges in the central region whose length increases with inter-origin distance. These domains identified as split-MRT U-domains are reminiscent of the skew-split-N-domains previously found in the germline [359]. The central region of split-U-domains consistently with what has been shown for late CTRs seem to be replicated late in the S-phase by random initiations of replication. This late replicating region at the centers of the split-U-domains is what set them apart from the U-domain. In fact, the early replicating regions at the split-U-domain borders exhibit similar properties as the ones observed around MaOris at U-domains borders. Interestingly, for all the analysed cell lines, the MRT derivative *vs* the MRT for different groups of split-U-domains follows a universal curve predicted by the model (Equation (4.3)). This led us to propose a "universal" cascade model for the replication process in human.

Interestingly, split-U- and U-domains present some conservation between different cell lines. This led us to define a conservation index for the domains borders. For all the cell lines, about 20% of the borders are cell line specific and about 5% are ubiquitously found in all the considered cell lines. From the high density of nucleosome free regions encoded in the DNA sequence around those ubiquitous borders, it was suggested that they could be specified by a genetic mechanism whereas epigenetic mechanisms would be responsible for the specification of cell line specific MaOris [365]. However, comparative analysis of DNase I HS and CTCF occupancy profiles around ubiquitous and cell line specific MaOris did not reveal different functional properties.

Embryonic stem cell line present some specificities related to the consolidation effect. The MaOris firing early in EC4 chromatin state at domain borders specific to H1 ES located in a low GC, gene-desert enviroment, were shown to play a fundamental role in the loss of pluripotency and lineage commitment [54]. As we will discuss in (Chapter 5, Section 5.1) [38, 147, 148, 364], the early to late transitions associated with the consolidation of pluripotent specific EC1, EC2 and EC4 MaOris to HP1-associated C4 heterochromatin likely coincide with the emergence of compact chromatin near the nuclear periphery and with a dramatic large-scale 3D genome organisation that may constitute an epigenetic barrier to cellular reprogramming.

4.4 Data materials

GC content

The GC-content is computed over the native sequence as:

$$GC = \frac{n_G + n_C}{n_a + n_T + n_G + n_C} \tag{4.4}$$

where n_T, n_A, n_G and n_C are the numbers of T, A, G and C counted along the genome.

Gene density

As human gene coordinates, we used the UCSC Known Genes table, data were downloaded from the Genome Browser of the University of California Santa Cruz (UCSC). When several genes presenting the same orientation overlapped, they were merged into one gene whose coordinates corresponded to the union of all the overlapping gene coordinates. This resulted in 23 818 genes. Gene density was computed as the number of gene promoter (5' (resp. 3') end or + (resp. -) genes) per length of DNA.

DNase I HS data

DNase I hypersensitive sites data were downloaded in the Encode standard format "narrowpeaks" (http://genome.ucsc.edu/FAQ/FAQformat.html). DNase I HS narrowpeaks are genomic intervals identified as hypersensitive zones to DNase I within a false discovery rate of 0.5% using the HotSpot algorithm. We downloaded the tables from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeUwDnaseSeq/ for H1 ES, GM06690, HeLaS3 and K562 and from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeOpenChromDnase for IMR90.

CTCF data

CTCF chromatin immunoprecipitation data were downloaded in the Encode standard format "broadpeaks" (http://genome.ucsc.edu/FAQ/FAQformat.html). Broadpeaks format is a table of significantly enriched genomic intervals. The signal value associated with each enriched intervals is the fold enrichment compared to a uniform distribution of reads. Data were downloaded for H1 ES, GM12878 (as surrogates for GM06990), Nhdfad (as surrogates for IMR90), HeLaS3 and K562.

Chromatin states

Chromatin states used in this thesis comes from the authors [54].

Chapter 5

3D structuration of the human genome, replication domains and chromatin states

The study of replication timing profiles (MRT) led to a unified view of the genome where the genome can be partitioned into 25% of early constant timing regions (CTRs), 25% of late CTRs and 50% of U-domains. While genome-wide replication timing profile and chromatin states repartition provide characterisation of the 1D spatio-temporal replication program, the recent development of chromatin conformation capture technique has led to rapid advances in the study of the so-called chromatin tertiary structure. In this chapter, we address the 3D structural organisation of CTRs on the one hand and the (split-) U-domains on the other hand.

In the previous chapter, we established that the human genome can be segmented into early CTRs (25%), late CTRs (25% that emerge between two MaOris far from each other leading to the formation of split-U-domains) and U-domains (50%) (Section 4.3). In this chapter, we start discussing the structural 3D organisation of the CTRs (Section 5.1) and of the U-domains and split-U-domains (Sections 5.2) as described by the Hi-C data. Let us recall that Hi-C technology was introduced to study genome-wide chromatin interactions resulting in Hi-C matrices (Fig. 5.1) representing the relative frequencies of colocalisations between all pairs of loci (Chapter 2, Section 2.3.2).

We will systematically consider Hi-C data in K562, GM06990, H1 ES*, and IMR90 obtained with HindIII restriction enzyme. Moreover, we use both raw and normalised data for IMR90 in order to see the effect of the normalisation on the results. Hi-C data normalised as in [247] (See page 47) should be free of biases related to GC content, fragment mappability and length. We discuss whether such biases really effect the results. We also consider data obtained for GM06990 with a different restriction enzyme (NcoI) to test to wich extent the results depend on the restriction enzyme used in the Hi-C protocol. In fact, the two restriction enzymes, HindIII and NcoI, used for the experiments, cleave DNA sequence for AAGCTT and CCATGG sequences respectively. HindIII is the most frequently used. We test to which extent the choice of this enzyme is crucial for the experiment by comparing the datasets obtained for GM06990 with the two different enzymes.

5.1 "Equilibrium" *vs* "fractal globule" interpretations of Hi-C data

In Chapter 2 (Section 2.3.4) we recalled that C1+C2 (resp. EC1+EC2) and C3+C4 (resp. EC3+EC4) regions correspond respectively to the early and late CTRs. To analyse Hi-C data in early and late CTRs, we first identify C1+C2 (resp. EC1+EC2) and C3+C4 (resp. EC3+EC4) blocks. To efficiently detect these blocks, we look at the coverage in C1+C2 and C3+C4, in windows of 500 kb; than the window is classified according to its highest coverage. The window will be considered C1+C2 (resp. C3+C4) if more than 60% of it is covered by C1 and C2 (resp. C3 and C4). When the the highest coverage does not reach 60%, the window is not classified (Fig. 5.1). To obtain larger blocks we look, in the same way, for more than 60% coverage by C1+C2 (resp. C3+C4) in 1 Mb windows (Fig. 5.1). The first set of blocks is used to compute the results described in this section.

5.1.1 Physical modelling of genome topology: "equilibrium" versus "fractal" globule descriptions

Analysis of the pioneering Hi-C data [14] have revealed that early active CTRs and late inactive CTRs do not significantly interact suggesting that they occupy different compartments of open and close chromatin inside eukaryote nuclei [14, 37, 38]. To provide some understanding of this reported compartmentalisation, some polymer-like modelling approaches have been recently developed [366–368] to account for the power-law depen-

*BG02 MRT split-U- and U-domains are used as surrogates for H1 ES.

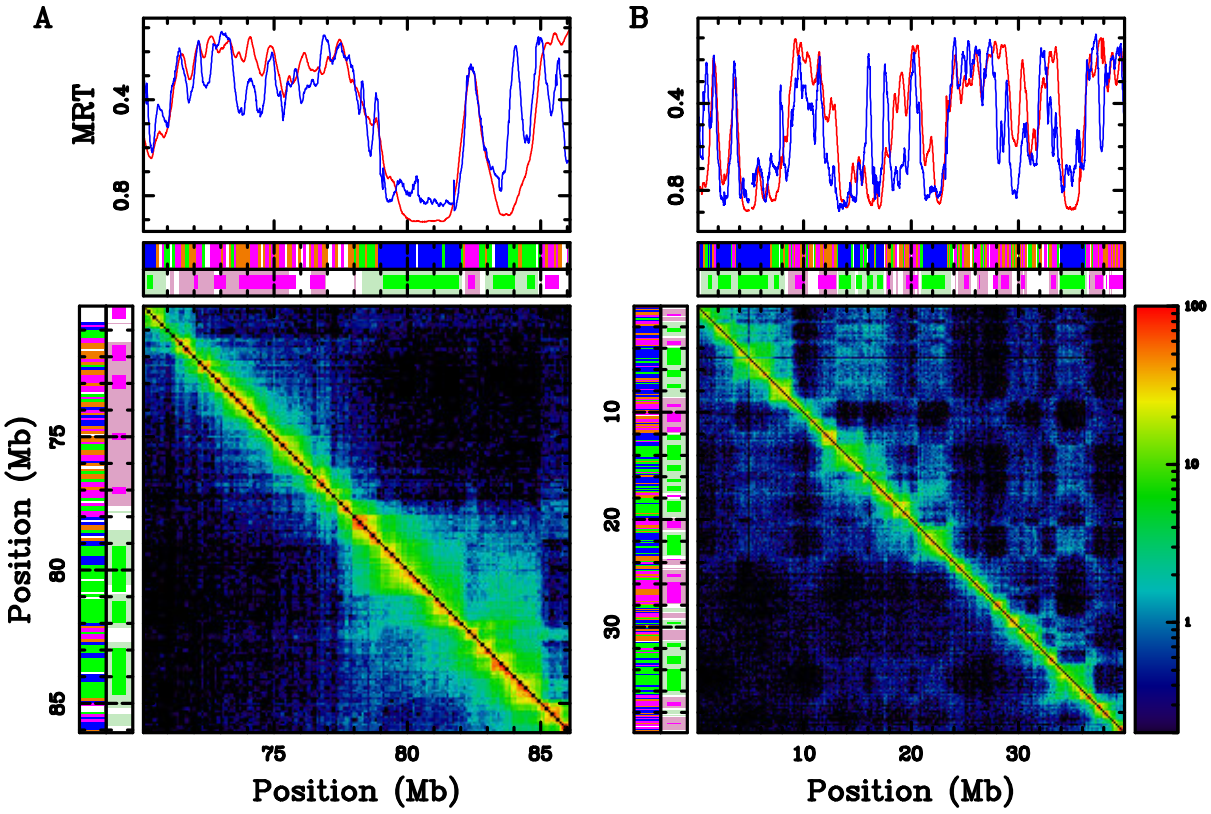


Figure 5.1. Chromatin state organisation and Hi-C data. (A) Hi-C contact map corresponding to intrachromosome interactions in a 16 Mb long fragment of human chromosome 11. Top panel: MRT profiles in H1 ES (blue) and IMR90 (red). Bottom panel: Hi-C interaction frequency in H1 ES (under the diagonal) and in IMR90 (above the diagonal); on the left of (resp. above) the interaction frequency map are represented 100 kb windows belonging to a 500 kb EC1+EC2 (resp C1+C2) block (pink) and to 1 Mb block(s) with a coverage in EC1+EC2 (resp. C1+C2) higher than 60% (light pink); similarly are also represented 100 kb windows belonging to a 500 kb EC3+EC4 (resp. C3+C4) block (dark green) and to 1 Mb block(s) with a coverage in EC3+EC4 (resp. C3+C4) higher than 60% (light green). The chromatin states (C1 and EC1 pink, C2 and EC2 orange, C3 and EC3 green, C4 and EC4 blue) are also represented. (B) Same as in (A) but for a longer 39.9 Mb fragment of human chromosome 2.

dencies observed over some range of scales in both fluorescent in situ hybridization (FISH) data [369] and chromosome capture data [14, 55, 234, 238–242, 370]. These various models predict a power-law behaviour of the end-to-end distance R of a subchain of length s:

$$R(s) \sim s^{\nu}, \tag{5.1}$$

and of the contact probability P_c between loci at genomic distance s:

$$P_c(s) \sim 1/s^{\alpha} \quad \text{with} \quad \alpha = d_s/2 = d_f/d_w, \tag{5.2}$$

where d_s is the spectral dimension, d_f the geometrical fractal dimension and d_w the dynamical fractal dimension [371–373]. When an ideal chain polymer is confined to a finite (rather small) volume, or when the attraction between monomers dominates over excluded volume repulsion ("poor solvent" conditions), then the polymer undergoes a transition from the 3D uncorrelated random walk coil into an equilibrium globule filled with random

walks that are uncorrelated to each other due to collisions with the globule boundary. In this equilibrium globule space-filling state, $d_f = 3$ and $d_w = 2$ as the characteristic of diffusion law. Hence, this model predicts the following scaling exponents $\nu = 1/2$ (Equation (5.1)) and $\alpha = 3/2$ (Equation (5.2)) over a range of distances smaller than the characteristic size of the globule ($R(s) \sim const$ and $P_c(s) \sim const$ for $s > N^{2/3}$, where N is the polymer total length) [368, 371–373]. These theoretical predictions were shown to be relevant to interpret FISH [366, 374] and Hi-C [25] data in *S.Cerivisiae*. Numerical simulations have confirmed that for small chromosomes like yeast chromosomes ($N \lesssim$ 1 Mb), the time to overcome hindering entanglements and to mix and reach equilibrium is comparable to the time duration of the cell cycle ($\sim$ 1 hour) [366]. But for larger chromosomes as mammalian chromosomes, experimental data have provided different estimates of the scaling exponents ν and α. In the range of scales from $\sim$ 0.7 Mb to $\sim$ 7 Mb, FISH [33, 366, 375] and Hi-C [14] experiments exhibit power-law scaling of $R(s)$ and $P_c(s)$ with exponents $\nu \simeq 1/3$ and $\alpha \simeq 1$. According to Equation (5.2), this is again consistent with a space-filling chromatin structure $d_f = 3$, but with a dynamical dimension $d_w = 3$ ($\geq$2) as the signature of anomalous diffusion (subdiffusion) [373]. To explain these experimental results and in particular the slower power-law decay of the contact probability $P(s) \sim s^{-1}$, pioneering authors [14, 368] have proposed as an alternative to the equilibrium globule model, the "crumple" or fractal globule model originally introduced by Grosberg *et al.* [34]. A fractal globule consists of crumples formed on all scales due to topological constraints: first small crumples are formed as the result of some local polymer collapses induced by the constraints imposed by other parts of the polymer; then the so-formed thicker polymer-of-crumbles experiences similar collapses into larger crumples and so on. Besides the original theoretical argumentation [34, 376], numerical simulations [14, 368] have confirmed that the fractal globule model predicts scaling exponent values $\nu = 1/3$ and $\alpha = 1$, in good agreement with FISH and Hi-C data. As compared to the highly knotted and slowly equilibrating "equilibrium" globule model [368, 371, 372], the fractal globule model [368] accounts for a self-organisation of the chromatin fiber into a long-lived, non-equilibrium unknotted conformation allowing easy opening and closing or translocation of chromosomal regions over large distances in the nucleus [377]. Besides facilitating chromatin loop folding and unfolding, possibly involved in the regulation of transcription and replication, the fractal globule model has another very attractive property as far as the observed compartmentalisation of the genome into mammalian nuclei [13, 14, 16, 370, 378, 379]. The fractal globule has a striking territorial organisation (continuous regions of the genome in the size range 0.7 Mb-7 Mb are compactly folded rather than being spread), which strongly contrasts with the mixing observed in the equilibrium globule. While being very appealing, the fractal globule is a long-live intermediate state on the way to becoming an equilibrium globule. This process is very slow (equilibration time $\sim N^3$) [366, 368] and depends on the stringency of the topological constraints. Simulations have shown that introducing some occasional DNA strand passing to mimic the role of DNA topoisomerase II, can significantly speed up equilibration of the fractal globule into an equilibrium one [368, 380]. Note that the $R(s) \sim s^{1/3}$ ($\nu = 1/3$) scaling observed for human chromosomes using FISH techniques has been recovered numerically in the simulation of equilibrated unknotted rings [368, 381]. Altogether these results enlighten the potential fundamental role of topological constraints in the segregation of chromosome territories observed by optical microscopy during the interphase [11, 369, 382, 383] as well as in the emergence of a compartmentalisation of the genome in individual chromosome as revealed by chromatin

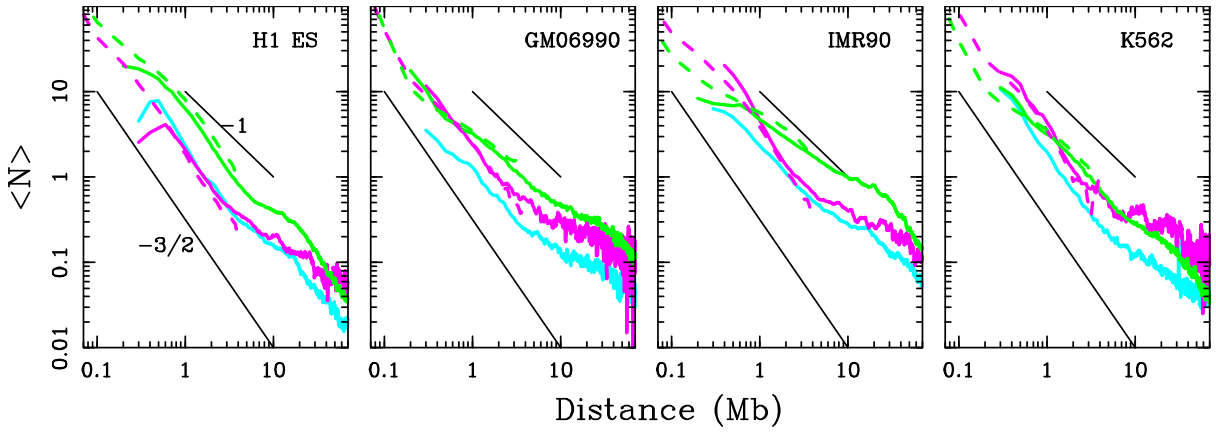


Figure 5.2. Structural organisation of CTRs. Mean intrachromosome Hi-C contacts *vs* genomic distance (logarithmic representation) between pairs of loci located in the same (dashed curve) or in different (solid curve) early active (pink) or late inactive (dark green) CTRs of length $L \geq 1$Mb in different cell lines. The black straight lines correspond to the power-law behaviour $P_c \sim s^{-\alpha}$ (Equation (5.2)) predicted by the "equilibrium" globule model ($\alpha = 3/2$) [368, 371–373] and the fractal globule model ($\alpha = 1$) [14, 34, 368].

capture experiments [13, 14, 16, 370, 378, 379]. Finally, let us mention that models of the high-order chromatin structure have been proposed that explicitly incorporate long-range looping phenomenon [384]. Their behaviour depends on the setting of a number of parameters which makes them difficult to use to interpret the experimental observations. We refer the reader to the review by [385] for a more general discussion on the physical modeling of higher-order chromatin structure.

5.1.2 Epigenomic folding of active early CTRs and inactive late CTRs

As discussed in Chapter 4 (Section 4.2.3), the chromatin fiber is not a homopolymer but a heteropolymer that accounts for the spatial compartmentalisation of the epigenome into four prevalent chromatin states likely corresponding to different structural and mechanical properties (*e.g.* different persistence lengths) of the chromatin fiber. As previously observed in *Drosophila* [20], the 3D folding of the epigenome is likely to be governed by the self-interactions between chromatin states that promote physical bridging, *e.g.* via the specific interactions of some architectural proteins (CTCF, Polycomb, lamina, ...) [13, 14, 370, 378, 379, 386–389]. Along that line, we reanalyse the Hi-C data (obtained with HindIII restriction enzyme) in various differentiated (IMR90[†], K562, GM06990) and pluripotent (H1 ES) human cell types with the specific purpose to investigate seperately the intra-chromosomal contact probability between pairs of loci in the active early replicating C1+C2 (resp. EC1+EC2) CTRs and between pairs of loci in the inactive late replicating C3+C4 (resp. EC3+EC4) CTRs (Fig. 5.2). When considering pairs of loci inside a C3+C4 CTR or in different distal C3+C4 CTRs, we consistently recover a slow power-law decay $P_c(s) \sim 1/s$ ($\alpha = 1$) over the range of scales from ~ 0.7 Mb to 7 Mb (Fig. 5.2), as previously obtained genome wide [14] and this for the three considered differentiated cell types. Interestingly, this signature of long-range interactions is no longer observed when considering pairs of loci inside a C1+C2 CTR or in different

[†]Chromatin states data C1-C4 of Nhdfad are used as surrogates for IMR90.

distal C1+C2 CTRs. The contact probability in these highly genic early CTRs enriched in the insulator protein CTCF decays much faster with genomic distance but is still describable by a power-law $P_c(s) \sim 1/s^{3/2}$ ($\alpha = 3/2$) in rather good agreement with the predictions of the "equilibrium" globule model [368, 371–373]. The additional observation of the scarsity of interactions between active early replicating CTR loci and inactive late replicating CTR loci strongly suggests some spatial segregation in the differentiated cell nuclei. As reported in previous works [197, 369, 378, 379, 383], active chromatin is positioned preferentially in the nuclear interior: small gene-rich chromosomes spatially cluster at the center of the nucleus together with the genic domains of longer chromosomes. The driving force (if any) bringing these active GC-rich genomic regions toward the nucleus center could be the colocalisation of distant genes into transcription factories [4–8, 369, 390] or a passive force resulting from the preferential spatial positioning of gene-poor silent AT-rich genomic regions at the nucleus periphery [379]. The inactive late-replicating C3+C4 CTRs are enriched in lamina proteins that are known to associate with the heterochromatin protein HP1 [391]. They likely correspond to lamina-associated heterochromatin domains (LADs) more or less confined to the nucleus periphery [17, 392–396]. The observed $P_c(s) \sim s^{-1}$ behaviour (Equation (5.2)) might also be explained in the framework of the "equilibrium" globule model $P_c(s) \sim s^{-d_f/d_w}$ with $d_f = 2$ (instead of 3) and $d_w = 2$ leading to $\alpha = 1$. In this interpretation, the structure of all chromatin states are expected to reach equilibrium in dividing cells and the different power-law exponents α (Equation (5.2)) underline the embedding of chromatin states in structural domains of different geometrical dimension. This interpretation is supported by the results of high-resolution confocal imaging and fluorescence correlation spectroscopy of mouse Swiss NIH embryonic fiblroblast (NIH 3T3) that provide the following estimates of the heterochromatin fractal dimension $d_f = 2.2 \pm 0.2$ and of the dynamical fractal dimension $d_w = 2.6 \pm 0.1$ as the signature of subdiffusion in the crowded heterochromatin layer at the nuclear envelop [397, 398]. These experimental estimates yield $\alpha \simeq 0.85$, *i.e.* an even slower power-law decay of the contact probability than predicted by the 2D "equilibrium" globule model ($\alpha = 1$). Note that this is what we observed in the mean number of interactions between inactive late C3+C4 CTR-loci in IMR90 ($\alpha \simeq 0.65 < 1$) as compared to K562 ($\alpha \simeq 1.1 \sim 1$) and GM06990 ($\alpha \simeq 0.87 \sim 1$).

Similar results are obtained when considering GM06990 Hi-C data obtained for a different restriction enzyme NcoI, and for IMR90 Hi-C data normalised as in [247] (Supplementary Fig. B.7). Over the range of scales of interest from ~ 0.7 Mb to 7 Mb, we observe a similar power-law decay $P_c(s) \sim 1/s$ ($\alpha = 1$). The only difference is the positions of the curves relative to each others. When considering loci belonging to C3+C4 CTRs $\alpha = 0.66$ for the normalised IMR90 data and $\alpha = 0.86$ for the GM06990/NcoI data.

Interestingly, when performing a similar analysis of Hi-C data in the pluripotent H1 ES cell line (Fig. 5.2), we got strikingly different results than in somatic cell lines as the signature of a unique higher-order genome structure possibly shaped by pluripotency factors [197, 198, 201, 202, 207, 361, 362]. Different to the long-distance interactions observed inside and between the nuclear lamina-associated inactive late replicating C3+C4 CTRs ($\alpha = 1$), the mean number of contacts between inactive late replicating EC3+EC4 CTR loci definitely decays much faster with genomic distance suggesting some loss of spatial organisation (Fig. 5.2). Importantly, the contact probability behaves as $P_c(s) \sim s^{-3/2}$ ($\alpha = 3/2$), very much like the power-law behaviour obtained between ac-

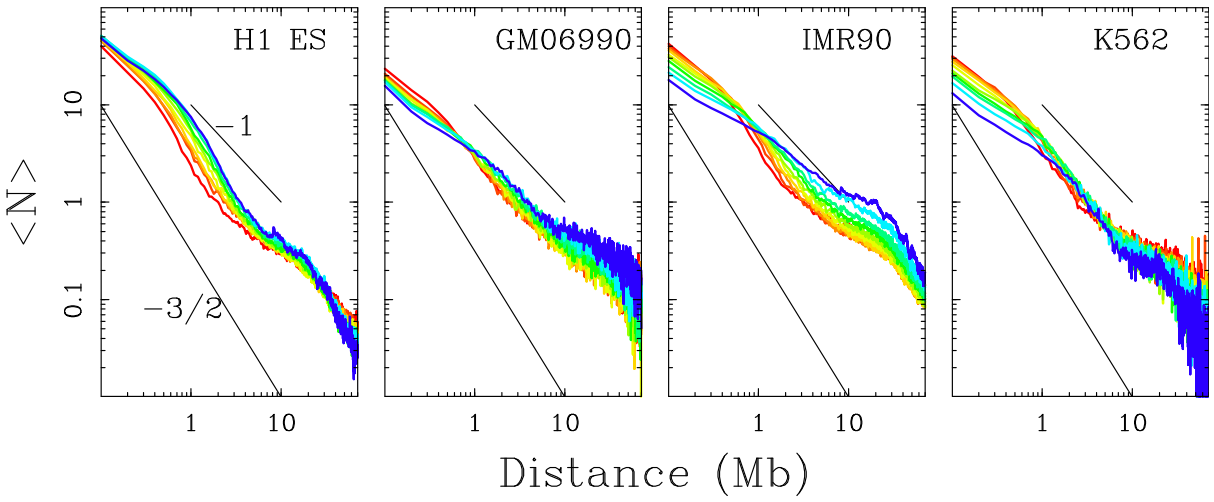


Figure 5.3. Structural organisation of the replication programme. Mean intrachromosome Hi-C contacts vs genomic distance (logarithmic representation) between pairs of loci classified according to timing deciles, from the earliest replicating first decile of loci (red), to the later replicating last decile of loci (blue), in different cell lines.

tive early replicating EC1+EC2 CTR loci (Fig. 5.2). The additional observation that qualitatively and quantitatively, a similar contact frequency distribution is also found between (inactive EC3+EC4/ active EC1+EC2) pairs of loci, is a strong indication that pluripotent chromatin does not display spatial segregation and is more randomly mixed and less engaged in specific long-range contacts [362], in consistency with the predictions of the 3D "equilibrium" globule model. The accessible and more relax EC3+EC4 CTRs might be more central in the nucleus ($d_f = 3$, $d_w = 2$, $\alpha = 3/2$) than the HP1-associated heterochromatin C3+C4 CTRs confined at the nuclear periphery ($d_f = 2$, $d_w = 2$, $\alpha = 1$). Altogether, these results confirm that during differentiation, chromatin structure switches from a highly dynamic, accessible and permissive euchromatin in ESCs to a spatially compartmentalised organisation with accumulating transcriptionally inactive and late replicating heterochromatin regions confined at the nuclear periphery [13, 38, 148, 197, 198, 201, 202, 361, 362, 378, 379].

5.1.3 Structural organisation and replication programme

A recent modified version of Hi-C [24] suggests that loci brought in contact by chromatin architecture exhibit similar MRT. To test this idea, we classify genome-wide 100 kb windows in each cell line according to their MRT. We then make different categories according to MRT deciles where a decile is any of the nine values that divide the sorted data into ten equal parts so that each category groups 10% of the data. We refer to these categories as MRT deciles where the first decile corresponds to the earliest replicating 10% of loci and the tenth or the last decile corresponds to the latest replicating 10% of loci. We then reproduce the above analysis for genomic loci classified according to these MRT deciles (Fig. 5.3). Interestingly, in the considered differentiated cell lines, the slope of the contact probability between pairs of loci in the different deciles progressively decreases from $\sim$ -1 to $\sim$ -3/2, when going from the latest to the earliest MRT decile (Fig. 5.3). This progressive decrease is also observed when reproducing the analysis for the GM06990 Hi-C data obtained with NcoI and the normalised IMR90 Hi-C data (Supplementary Fig. B.8). In GM06990, the slope transition is easier to observe

with the NcoI data (Supplementary Fig. B.8) thanks to the differential vertical shift of the contact *vs* the distance curve for the different MRT deciles. This can be interpreted as some evidence of radial nuclear organisation [8, 242, 369, 378, 399] consistent with the observation that the spatial distribution of replication foci [5, 6, 400] changes over the course of the S-phase from a central to a peripheral positioning in the cell nucleus [5–8, 228, 369, 390, 401, 402]. The results reported in this study strongly suggest that this radial nuclear organisation (in somatic cells) is a typical example of a transition between 3D and 2D equilibrium statistical physics with the MRT as the underlying key cell type dependent parameter. They are in good agreement with recent works showing that CTCF in concert with cohesine contribute to create a favorable chromatin architecture that promotes early replication [387, 403], whereas nuclear lamina interactions likely play a direct role in replication origin licensing and activation [387, 404].

The results obtained for H1 ES are different than the ones obtained with differentiated cell lines. H1 ES seems to be compatible with the equilibrium globule model. We do not observe the 3D-2D transition which can be a signature of late replicating EC4 spreading in the nucleus.

5.1.4 Transition from 3D to 2D equilibrium globule chromatin organisation in differentiated cell lines

The genome average of $P_c(s)$ for the pioneering data in K562 cell line results in a power law with exponent $\alpha \simeq -1$ in the 0.7-7 megabase scale range [14]. This scaling exponent was interpreted as the signature of a genome structuration into the so-called *fractal globule model*, an out of equilibrium, knot-free, polymer conformation optimising the compaction/accessibility tradeoff [368]. Here, taking into account that chromosomes are heterogeneous polymers constituted of a succession of megabase-sized regions of different chromatin states, we show that the exponent $\alpha = -1$ compatible with the fractal globule model is only observed for the heterochromatin regions of differentiated human cell types. For the corresponding euchromatin regions as well as all regions of an embryonic stem cell line, we observe an exponent $\alpha \simeq -3/2$ as predicted by the simple 3D equilibrium globule model. We propose that the observed $\alpha = -1$ in fact corresponds to a 2D equilibrium. These results suggest that the exponent $\alpha = -1$ observed in heterochromatin regions is the signature of a transition to 2D dynamics associated with the nuclear lamina. When compared to the human replication programme, this transition corresponds to a progressive segregation from 3D dynamic at the nucleus center of early replicating regions to 2D dynamic confined at the nuclear envelop of late replicating regions. These results shed a new light on replication foci formation and dynamics in human.

5.2 Master replication origins and long-range chromatin interactions in replication domains

In Chapter 2 (Section 2.4), we recalled the results of a preliminary study [44] suggesting that replication U-domains correspond to square blocks of enriched interactions of the K562 Hi-C matrix (Fig. 2.19). More quantitatively, it was shown in K562 [44] that *(i)* the mean number of pairwise interactions is higher inside the U-domains than genome-wide (Fig. 2.19 C) and *(ii)* the contact probability between loci inside a U-domain is higher

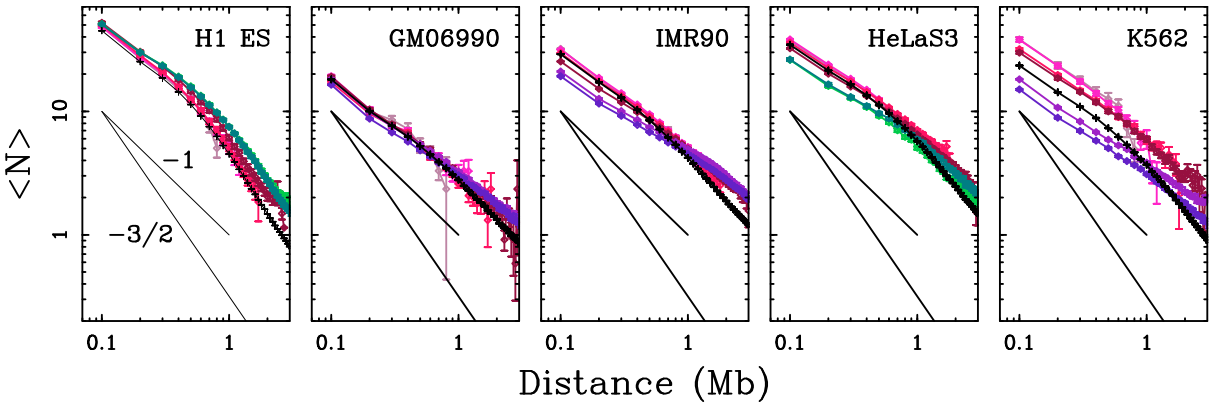


Figure 5.4. Are (split-) U-domains structural units of the genome? Mean number of interactions (<N>) between two 100 kb loci versus the distance separating them (logarithmic scales) as computed genome-wide (black curve) or in replication domains sorted by size categories. U-domains were grouped into 4 categories: $L < 0.8$ Mb (light pink), 0.8 Mb $\leq L <$ 1.2Mb (pink), 1.2 Mb $\leq L <$ 1.8 Mb (magenta) and 1.8 Mb $\leq L <$ 3 Mb (dark magenta). The split U-domains for lymphoblasts (GM06990), fibroblast (IMR90) and cancer (K562) cell lines were grouped in 2 categories: $3 \leq L < 4$ Mb (light purple) and $L \geq 4$ Mb (dark purple). For BG02 and HeLaS3 split-U-domains were grouped in 2 categories: $3 \leq L <$ 3.5 Mb (light green) and $L \geq$ 3.5 Mb (green).

than the one between loci lying in neighbouring U-domains (Fig. 2.19 D). Moreover, it was also shown, using 4C experiment in lymphoblastoid cell line, that the early MaOris localised in $\sim$ 200 kb region of open chromatin at U-domain borders preferentially interact [55]. This suggests that replication U-domains correspond to structural domains of self-interacting chromatin and that their borders act as insulating regions both in Hi-C and 4C experiments. Here, we aim at objectively quantifying the "importance" of replication split-U- and U-domains borders using centrality measures introduced in Chapter 3 (Section 3.3).

5.2.1 Replication (split-) U/N-domains and the 3D organisation into structural units

We first extend the analysis performed in [44] and [55] to *(i)* split-U-domains and *(ii)* other cell lines (described at the beginning of the Chapter), to test the generality of the results obtained for K562 and GM06990.

We first analyse the decay of the contact probability versus the genomic distance for loci inside replication split-U- and U-domains. Consistently with the observations for K562 and GM06990 replication U-domains, Figure 5.4 shows that the mean number of interactions between two 100 kb loci of the same replication domain decays when increasing their distance, in all cell lines and for all domain sizes, as observed genome-wide. Importantly for K562 the mean number of pairwise interactions is higher inside the U-domains (L<3 Mb) than genome-wide and this seems to depend on the U-domain length: the smaller the domain the higher the mean number of interactions. For split-U-domains, at small distances, the mean number of pairwise interactions is lower inside the split-U-domains than observed genome-wide. At large distance ($\gtrsim$ 1 Mb) the number of interactions inside split-U-domains is higher than genome-wide which can be the result

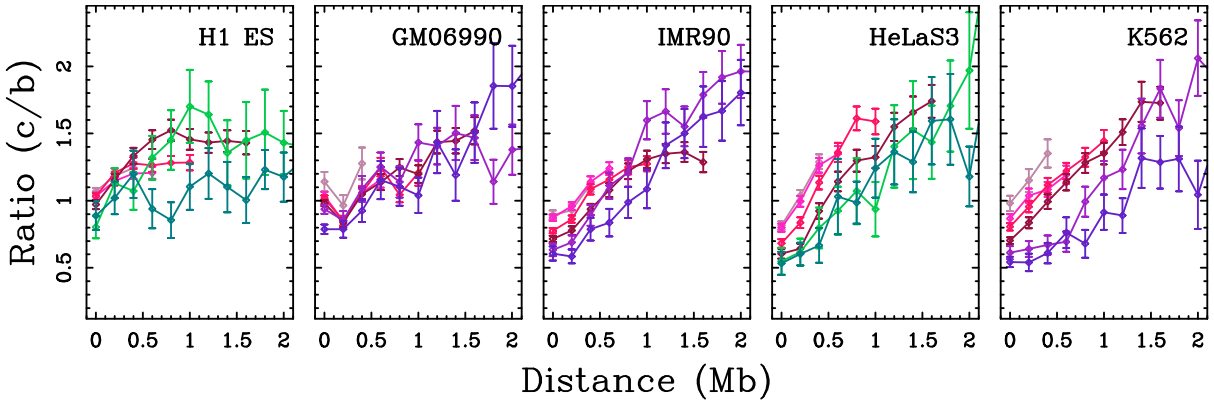


Figure 5.5. Are (split-) U-domains structural units of the genome? Ratios of the number of interactions between two 100 kb loci inside the same replication domain at equal distance from its center (c) and the number of interactions between loci on opposite sides and equal distance from replication domain borders (b), *vs* the distance between them, in different cell lines and for different domain sizes (colors as in Fig. 5.4).

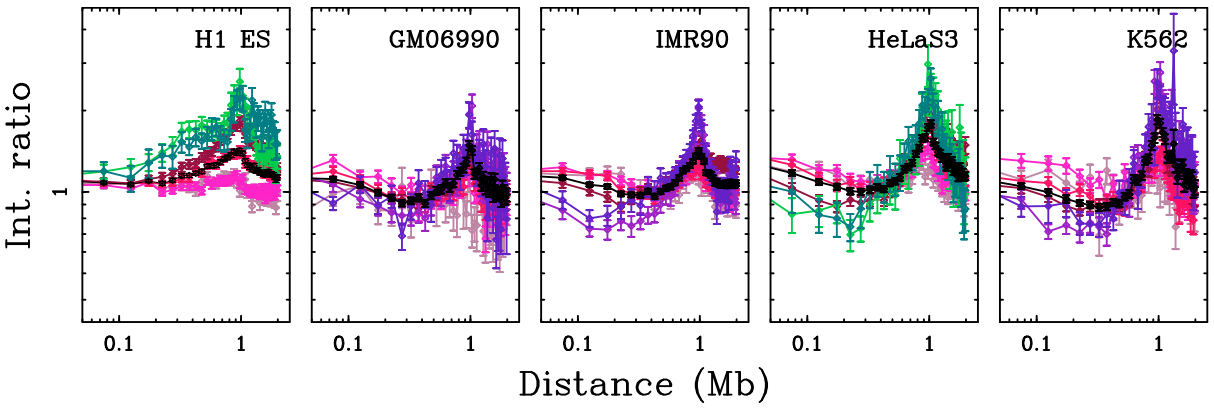


Figure 5.6. MRT peaks at the heart of genome organisation. 4C-like interaction profile of (split-) U-domain borders. We extract Hi-C interaction profiles between (split-) U-domain borders and all their neighbours in the direction of the opposite (split-) U-domain border as a function of the distance from the reference border. We only keep the 4C-like profiles where the border delimits two consecutive (split-) U-domains. Interaction counts were normalised by the mean interaction count averaged over all pairs of loci separated by the same genomic distance. All 4C-like interaction ratio profiles are finally averaged after rescaling the distances, so that all domain sizes are 1. The average ratio (Int. ratio) across MRT peaks in the different cell lines is computed genome-wide (black curve) and in the different domain categories (same color coding as in Fig. 5.4).

of an organisation into self-interacting structural domains. This is also observed for the considered differentiated cell lines GM06990, IMR90 and HeLaS3 even if the vertical offset between the curves is less pronounced. In H1 ES, the mean number of interactions between two 100 kb loci inside replication domains decreases when increasing the distance separating them similarly to the genome average. Note that neither normalising the data for IMR90 nor changing the restriction enzyme used to generate the Hi-C library for GM06990 drastically affect the results (Supplementary Fig. B.9). However, the relative positions of the curves are changed, we observe a vertical shift (Supplementary Fig. B.9). Consistently with the observation for CTRs (Fig. 5.2), above 0.7 Mb $P_c(s)$ decreases with a power law exponent $\sim$ -1 in differentiated cell lines while this exponent is $\sim$ -3/2 for H1 ES.

We now compare the contact probability between two loci inside a (split-) U-domain or lying in neighbouring (split-) U-domains (Fig. 5.5). We look at the ratio of the number of interactions between two 100 kb loci inside a replication domain at equal distance from the center and the number of interactions between two 100 kb loci on opposite sides and equal distance from the border. We recover for K562, as in the original work [44], that these ratios increase up to two fold for large distances. This is generally observed for (split-) U-domains in all considered cell lines (Fig. 5.5), and this even when the restriction enzyme is changed or when the data are normalised (Supplementary Fig. B.10). Note that for H1 ES (Fig. 5.5) the ratio increases to reach a plateau. These observations suggest that replication domain borders correspond to structural barriers consistently with the previous observation (Fig. 4.16) that replication domain borders are enriched in the insulator binding protein CTCF.

To further investigate the role of the domain borders, we compute the 4C-like interaction profiles of split-U- and U-domain borders (like originally proposed in [55]). Recall, that 4C procedure captures the interactions between one selected locus known as the "view-point" and all the other loci (page 45). In that sense, a line in the Hi-C matrix is equivalent to the 4C profile for the viewpoint situated on the diagonal. This allows to compute 4C-like profiles for viewpoints located at (split-) U-domain borders. We look at the interactions of these borders and all their neighbouring loci in the direction of the opposite domain border as a function of the distance to the reference border. To ensure that the opposite domain border is well defined, we limite the analysis to the borders that delimit two consecutive domains. As initially observed in [55] for GM06990 and K562 domains, we get for all the cell lines (Fig. 5.6 and Supplementary Fig. B.11) that on average there exists more interactions between domain borders than expected given their genomic distance. Note that the larger the domains, the higher the interactions between the borders. This observation suggests that replication domain borders loop out the intervening late-replicating regions to contact each other [55].

Altogether these results reflect a three-dimensional structural organisation where master replication origins at domain borders likely play a major role in the formation of the 3D structure.

5.2.2 From chromatin conformation capture data to chromatin interaction network

In our initial analysis [405], we were interested in Hi-C experimental data for erythroid human cell line (K562) [14] and we focused on both the intra- and interchromosomal contact maps. Hi-C contact maps are positively defined and symmetric and so are proned to be represented and analyzed using graph theory [210] (Chapter 3). A natural way is to use the Hi-C contact matrix as the adjacency matrix $\mathcal{A}$ of a graph $G = (V, E)$, where the vertices v_i are the 100 kb DNA loci and the edges $e(v_i, v_j)$ link connected vertices. In our initial work [405], we used two somehow complementary graph descriptions: (i) a weighted network where the weights assigned to each edge are the number of corresponding binary interactions and (ii) a non weighted network where the entry a_{ij} of $\mathcal{A}$ is 1 if v_i and v_j are connected and 0 otherwise. Because the number of interactions decreases very fast when increasing the seperation s between the loci (like s^{-1} or $s^{-3/2}$) [14, 210], the weighted network amounts to focus on interactions between loci seperated by short genomic distances ($\lesssim$ 10 Mb) over which contact probabilities are the highest. Alternatively, the non weighted network takes equally into account short-range ($\lesssim$ 10 Mb) and very long range interactions within a chromosome ($\geq$ 40 Mb).

To provide some perceptual illustration of the results of the centrality measure analysis, we use in this work the open-source Gephi software [406], a interactive visualisation tool to manipulate large graphs such as social or biological networks. This software implement graph layout algorithms. As discussed in Chapter 3 (Section 3.2.1.1), one way to represent the graph consists in grouping nodes strongly connected while pushing away the weakly connected ones using a dynamical approach. Here, we use the special force-directed algorithm called *Force Atlas 2* (of the Gephi software). This algorithm, such as the Force Directed Placement algorithms (Section 3.2.1.1), amounts to simulate a physical system of particles (vertices) distributed in a plane (2D): vertices repulse each other like magnets while edges attract the vertices they connect like springs. These forces create a dynamic that converges to a balanced final state that is expected to help the interpretation of the data (Fig. 5.7). In this representation, the information is conveyed by the relative positions of the vertices. Figure 5.7 illustrates the successive steps of the algorithm. The nodes are randomly put in a 2D space (Fig. 5.7 A), the system then evolves according to the repulsion and attraction forces (Fig. 5.7 B,C) to reach a worm-like steady state (Fig. 5.7 D). This state reflects the high level of interactions between close loci and follows the linear structure of the chromosomes. In fact, when projecting the genomic positions on the graph nodes, we check that they are sorted accordingly (data not shown). Hence, the worm-like layout reflects the quick decay of the contact frequency as a function of the genomic distance [14]. Interestingly, the projection of U-domains on the graph (Fig. 5.8) shows that borders are found closer to each other in the graph layout unlike centers that mainly interact at short distances and are found looping out at the periphery of the stationary, worm-like final configuration. This illustrates the central role of the MRT domain borders in the specification of the 3D architecture of the human chromosomes.

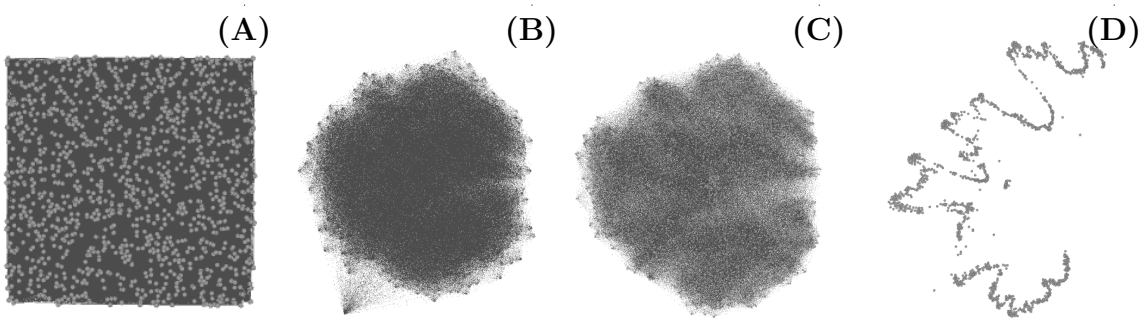


Figure 5.7. Worm-like graph representation of Hi-C data. (A) All the 100 kb loci of human chromosome 11 (using K562 Hi-C data [14]) are randomly put in 2D space as nodes (dots) with the interactions found between them as edges connecting the nodes (black lines on A-C). (B)-(D) The nodes repulse each others while the edges tend to bring them together. The system evolves reaching a stationary state (D).

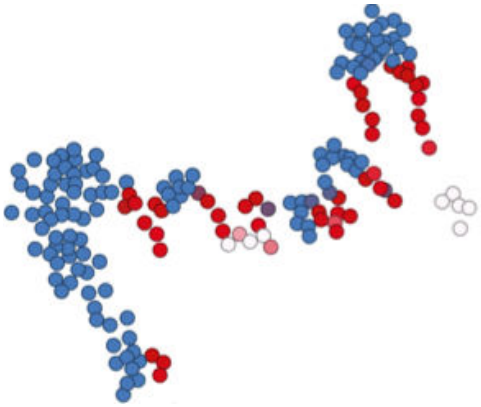


Figure 5.8. Replication domains and the worm-like graph representation. A 18 Mb long fragment of human chromosome 10 (using K562 Hi-C data [14]). The nodes at replication domain borders (400 kb on each side) are colored in red, the interior of the replication domains are represented in blue, and the other nodes outside replication domains are represented in white.

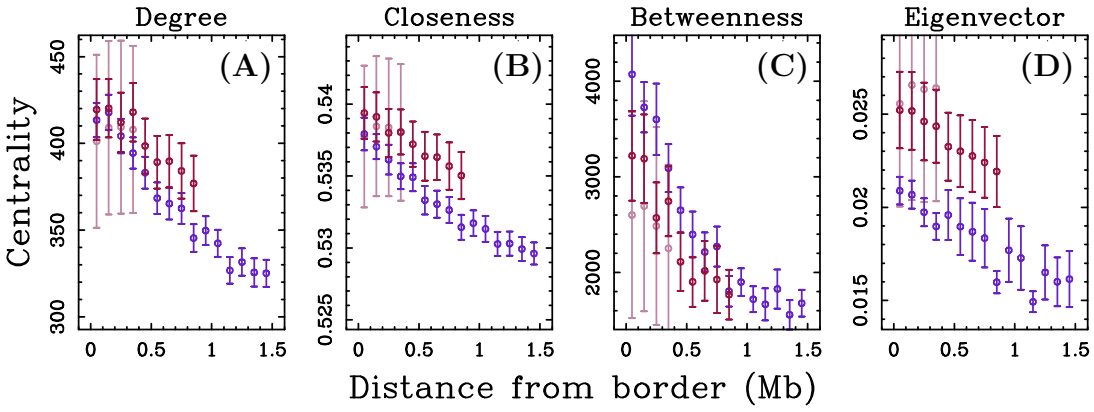


Figure 5.9. Replication domains borders are hubs in K562 DNA interactions network. (A) Degree, (B) Closeness, (C) Betweenness and (D) Eigenvector intrachromosomic centralities *vs* to the distance from the closest (split-) U-domain border in K562 cell line for weighted graphs in different size categories of replication domains: L $<$ 1.8 Mb (light pink), 1.8 $\leq$ L $<$ 3 Mb (pink) and L $\geq$ 3 Mb (purple).

5.2.3 Replication domain borders are hubs of the chromatin interaction network in the K562 cell line

To objectively quantify the above observations, we report hereafter the results corresponding to a statistical analysis performed over the U-domains (L $<$ 3 Mb) and split-U-domains (L $\geq$ 3 Mb) identified in the 22 human autosomes in K562 [405].

After filtering low and high interacting fragments from the Hi-C data, as discussed in Chapter 2 (page 48), we compute the four previously defined centralities: degree (C_D), closeness (C_C), betweenness (C_B) and eigenvector (C_S) centralities (Section 3.3), for the weighted graphs resulting from all the intra-chromosomal interactions. Note that to compute the shortest path involved in the calculus of C_B and C_C (Equations (3.3) and (3.2)), we take the inverse of the weight as the length of an edge. We classify the replication domains relatively to their size: L $<$ 1.8 Mb, 1.8 $\leq$ L $<$ 3 Mb and L $\geq$ 3 Mb, and then we look at the average centralities of vertices corresponding to loci of the replication domains as a function of the distance to the closest domain border (Fig. 5.9). The replication domain borders are local maxima of the degree, closeness, betweenness and eigenvector centralities suggesting that they correspond to critical vertices in the Hi-C interaction network (Fig. 5.9). The decrease from borders to center is much more pronounced for split-U-domains and U-domains larger than 1.8 Mb than for U-domains smaller than 1.8 Mb.

The decrease of the degree centrality C_D (Fig. 5.9 A) confirms that replication domain borders have a much higher intrachromosomal contact frequency than the centers in K562. The closeness centrality (Fig. 5.9 B) only slightly decreases when moving from the domain borders to center. Recall that a higher C_C reflects the "central" position of the node that can be caricatured by the fact that, in the stationary state (Fig. 5.7), the domain borders form somehow feet of loops "inside" the graph while the centers are pushed to the periphery. In Figure 5.9 C, the betweenness centrality C_B is also shown to decrease. The decrease of C_B quantifies the role of these replication domain borders as "hubs" in the chromatin interaction network. This enlightens the insulator properties of these replication domain borders that prevent cross-talk between neighbouring domains likely establishing self-interacting independent expression domains [44]. The eigenvector centrality C_S (Fig. 5.9 D) also decreases from replication domain borders to center. This illustrates the fact that replication domain borders are hubs that predominantly interact with other hubs.

These results (Fig. 5.9) suggest that replication domain borders are interconnected hubs at the heart of the chromatin interaction graph. In fact, the "early" replicating initiation zones of open and transcriptionally active chromatin that border replication U/N-domains in the genome wide intrachromosome Hi-C chromatin network are local maxima for the different centrality measures in K562.

5.2.4 Are replication domain borders hubs in different cell types?

To assess the robustness of the link between functional domains and genome 3D architecture, we reproduce this genome-wide analysis in different human cell lines where Hi-C data and replication timing profiles are available.

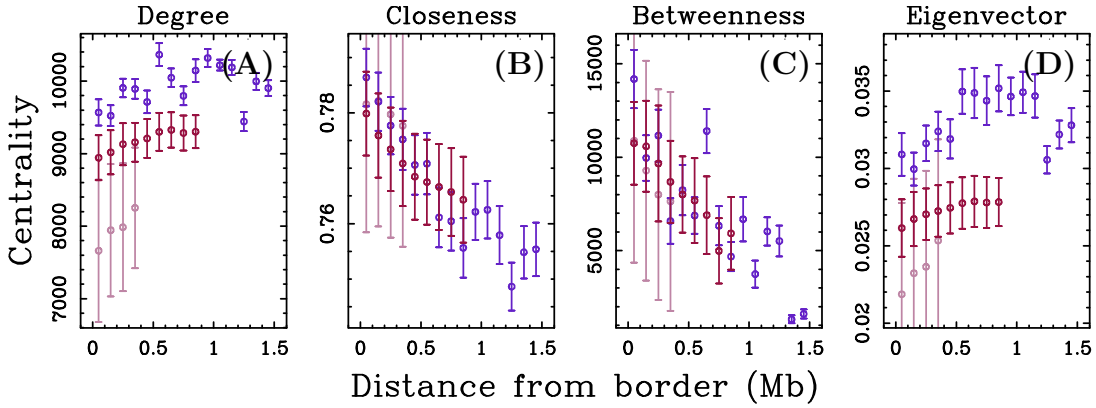


Figure 5.10. Same as in Figure 5.9 for H1 ES interaction network.

5.2.4.1 MaOris are hubs in H1 ES embryonic stem cell line

We reproduce the analysis described in Section 5.2.3 using H1 ES Hi-C data with BG02 replication domains. We proceed in the same manner to construct the chromatin interaction network where the 100 kb loci are the nodes of the graph and the Hi-C interactions constitute the weighted edges. We calculate C_D, C_C, C_B and C_S for all the graph vertices and we look at the centralities of the nodes inside replication split-U- and U-domains, classified as above in three size categories: L $<$ 1.8 Mb, 1.8 $\leq$ L $<$ 3 Mb and L $\geq$ 3 Mb (Fig 5.10).

In contrast to the results observed for K562, the replication domain borders are only local maxima of the betweenness and closeness centralities (Fig 5.10 B, C). C_D slightly increases from replication domain borders towards the center (Fig 5.10 A), suggesting that domain centers have more connections than the borders. C_S, that can be seen as a generalisation of C_D (Section 3.3.4), also increases (Fig 5.10 D). This is consistent with the fact that in H1 ES, EC3+EC4 CTRs have more interactions than EC1+EC2 CTRs (Fig. 5.2) and that replication domain borders are mainly EC1 and EC2 while the centers are EC3 and EC4. Nevertheless, the decrease of C_C and C_B from domain borders to center (Fig 5.10 B, C) suggests that replication domain borders are at the heart of the structural organisation and mediate interactions as in K562. In fact, as suggested in Chapter 3 (Table 3.1), the combination of low (degree and eigenvector) and high (closeness and betweenness) values for the replication domain borders means that they have few connections (low degree) and they are not connected to key player of the network (low eigenvector), however, they are indeed key players in the "center" of the network (high closeness) and they are crucial to the network flow (high betweenness).

These results confirm the "important" role of the replication domain borders relative to domain centers. They are "hubs" in the pluripotent chromatin interaction network in the sense that they are used as bridges for the graph connections such as the *airline hubs* that are airports which an airline uses as a transfer point to get passengers to their destinations when travelers want to move between airports with no direct flights.

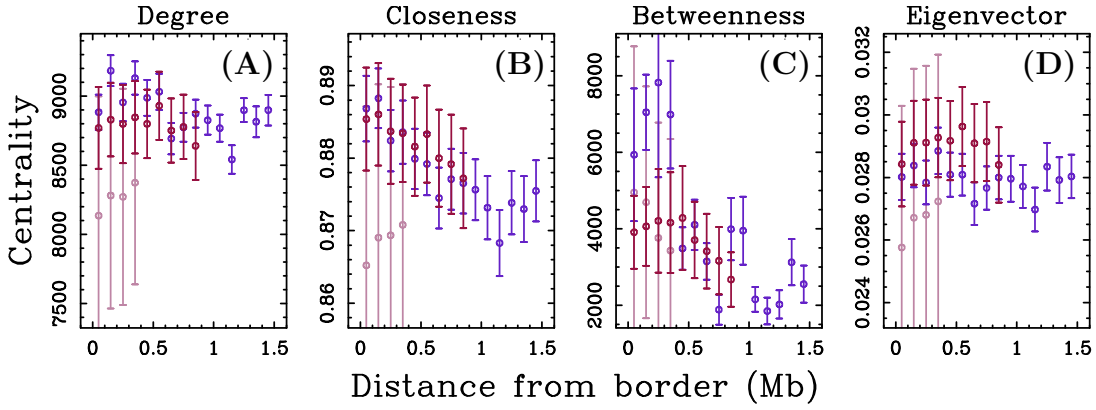


Figure 5.11. Same as in Figure 5.9 for IMR90 interaction network.

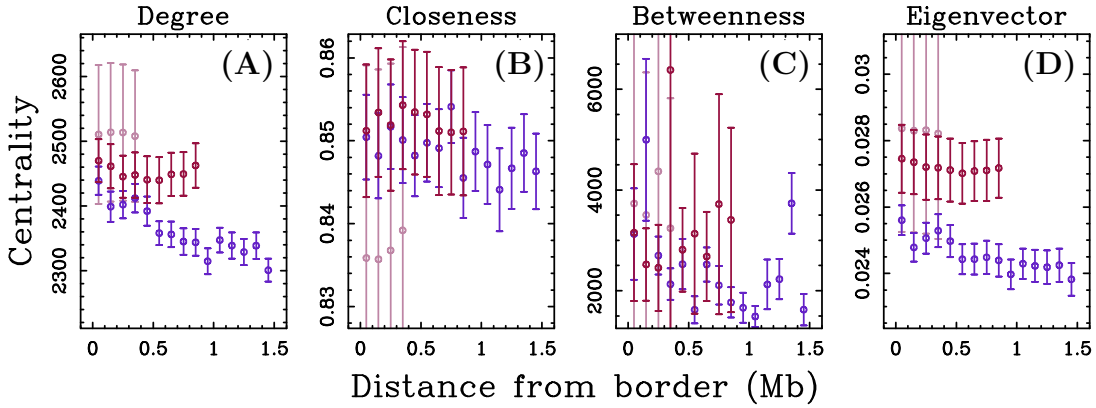


Figure 5.12. Same as in Figure 5.9 for IMR90 interaction network normalised as in [37].

5.2.4.2 Are MaOris hubs in the IMR90 interaction network?

We carry out similar analysis using IMR90 intra-chromosomal Hi-C data, both the raw data (Fig. 5.11) and the normalised data (Fig. 5.12) as in [37] (See Chapter 2 page 47), obtained from [16].

The results obtained with raw IMR90 data are more similar to the ones obtained with H1 ES than to the ones obtained for K562. When looking at the centrality measures inside replication domains, the replication domain borders are only local maxima of the betweenness centrality (Fig. 5.11 C), and to a lesser extent of the closeness centrality (Fig. 5.11 B). However, C_D and C_S are almost constant along the replication domain (Fig. 5.11 A, D). This can be related to previous observations on CTRs (Fig. 5.2) where C1+C2 CTRs had more connections than C3+C4 CTRs up to 1 Mb, whereas the tendency was reversed at larger distances, so that on average C1+C2 CTRs or C3+C4 CTRs have the same number of interactions but with partners distributed at different distances. Interestingly, when normalising the data, C_D for split-U-domains slightly decreases while C_C gets more homogenous in the domains. Note that C_B for large domains continues to decrease and C_S is not affected (Fig. 5.12). This raises the question about the normalisation effect on the data. The connections are not changed however the weights associated with the connections are modified as is reflected by the changes in the degree centrality computation (Figs. 5.11 and 5.12). Nevertheless, in IMR90 the MaOris at replication domain borders still mediate genomic interactions being local maxima of the betweenness centrality.

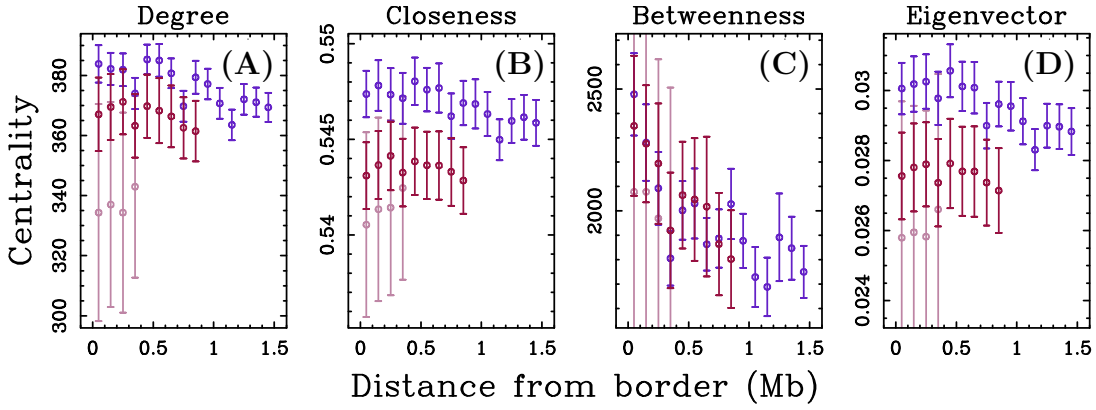


Figure 5.13. Same as in Figure 5.9 for GM06990/HindIII interaction network.

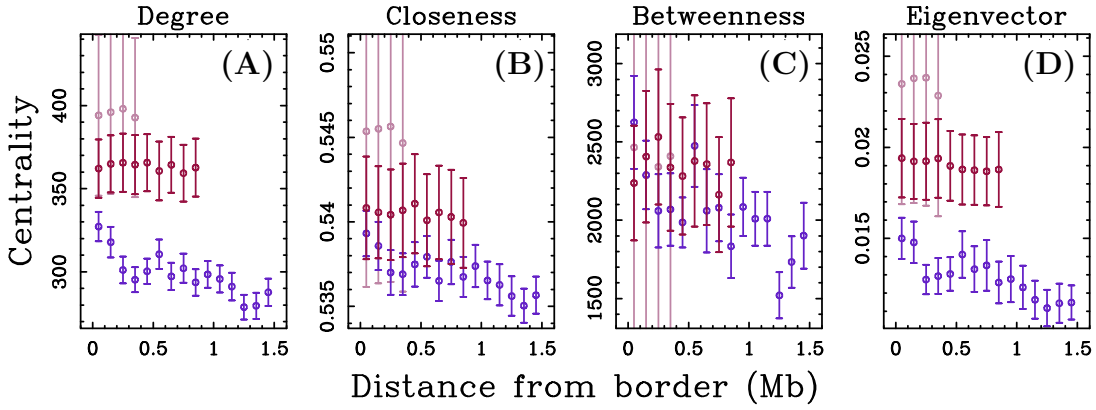


Figure 5.14. Same as in Figure 5.9 for GM06990/NcoI interaction network.

5.2.4.3 Does the choice of the restriction enzyme affect GM06990 hubs?

We reproduce the previous analysis on GM06990 intrachromosomal Hi-C interaction networks resulting from the two datasets obtained when cutting the ligated molecules (in the Hi-C experiment) with HindIII and with NcoI (Figs. 5.13 and 5.14). In both cases, replication domain borders are local maxima of the betweenness centrality (Figs. 5.13 C and 5.14 C) especially for the split-U-domains. The other centrality measures for these size categories present rather flat profiles (Figs. 5.13 A, B, D and 5.14 A, B, D). The fact that degree, closeness, and eigenvector centralities are flat means that the replication domain borders do not present more interactions as compared to loci inside the replication domains. However, the decrease of the betweenness centrality suggests that the replication domain borders are really important for the flow inside the network. In other words, the replication domain borders still play the role of hubs as communication bridges inside the graph that mediate other interactions.

It is worth to mention that for the degree, closeness and eigenvector centralities using HindIII as restriction enzyme, the values obtained for the smaller domain sizes are systematically lower than the ones obtained for the split-U-domains (Fig. 5.13). On the opposite, with restriction enzyme NcoI, the curves for smaller domain sizes are systematically higher than the ones obtained for split-U-domains (Fig. 5.14). This illustrates that the choice of the restriction enzyme can effect the results in unexpected and not understood ways. Full characterisation of the effect of the choice of the restriction enzyme

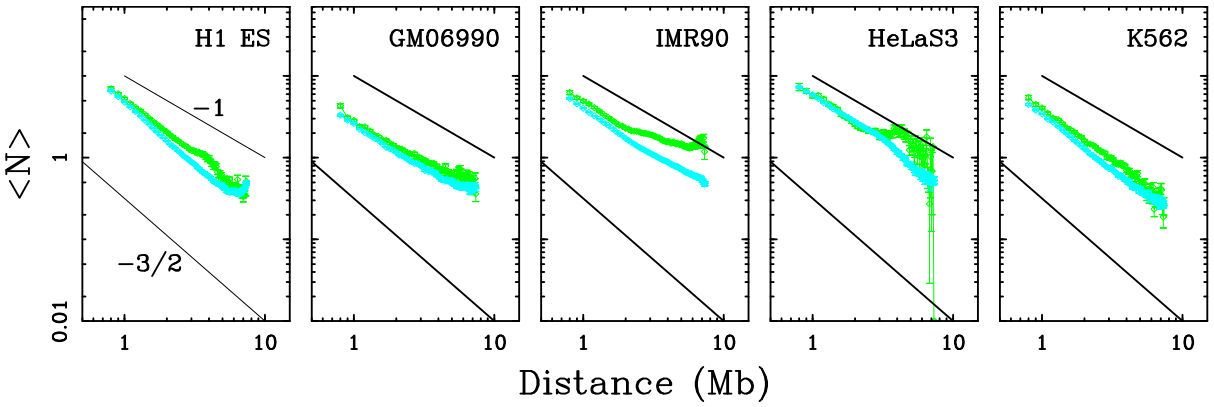


Figure 5.15. MaOris and the 3D genome organisation. Mean intrachromosome Hi-C contacts *vs* genomic distance (logarithmic representation) between two loci in the central region of two juxtaposed (split-) U-domain separated by a MRT peak (green), or on each side and equidistant to a MaOri bordering an early CTR and a (split-) U-domain (blue) in H1 ES, GM06990, IMR90, HeLaS3 and K562. The black straight lines correspond to the power-law behaviour $P_c(s) \sim s^{-\alpha}$ predicted by the equilibrium globule model in 3D ($\alpha = 3/2$) and in 2D ($\alpha = 1$). We considered only split-U- and U-domains of length L $\geq$ 1.2 Mb. To avoid MRT to be a confounding factor, we excluded late U-domain borders with MRT>0.5.

requires further Hi-C experiments.

☞ To summarise, using graph centrality measures, we have shown that the "early" replicating initiation zones of open and transcriptionally active chromatin that border replication split-U/N- and U/N-domains play a predominant role of hubs in the intra-chromosomes interactions network regardless the considered cell type, as local maxima of the betweenness centrality C_B. Nevertheless, cell type specific properties can be observed, for instance in K562 MaOris are highly interconnected to each other (local maxima for the degree and eigenvector centralities) and in H1 ES, MaOris are at the center of the graph (local maxima of the closeness centrality). The "master" replication origins are not only barrier elements that delimit self-interacting topological domains of independant expression and duplication, they also mediate long-range interactions among distant DNA elements within chromosomes. In contrast, even if the late replicating regions at the centers of the (split-) U/N-domains can be highly (or even more) connected than the borders, they can be by-passed.

5.3 MaOri plasticity and genome organisation

From the comparison of MRT profiles in different cell types in Chapter 4 (Section 4.2), we saw that replication split-U- and U- domains borders are dynamic. Each cell type was shown to share about half of their MaOris at domains borders with at least one other cell type (including the skew N-domain borders in the germline) but only a small proportion ($\lesssim 5\%$) are ubiquitous to all considered cell types (Table 4.4, page 103). Furthermore, the MaOris can correspond to either MRT peaks (common to two juxtaposed MRT U-domains) or to a dissymetric border (common to an early CTR and a MRT U-domain). Here, we ask whether all MaOris play the same role in the genome architecture or depending on their status, they can be more or less efficient. In fact, in differentiated

cell types, when investigating the contact probability between (C3,C4) loci at the centers of two juxtaposed (split-) U-domains, we recover the same long-range interactions over distances 0.7 Mb $\lesssim s \lesssim$ 7 Mb (Fig. 5.15) as previously observed between inactive late (C3+C4) CTR loci (Fig. 5.2). In the same way, the contact probability between two loci on each side of and equidistant to a MaOris bordering an early CTR and a MRT (split-) U-domain decays faster on the range of distances 0.7 Mb $\lesssim s \lesssim$ 7 Mb (Fig. 5.15), as previously observed for pairs of loci in separate early C1+C2 CTRs and late C3+C4 CTRs (Fig. 5.2). When analysing the H1 ES Hi-C contact probability between pairs of loci in the (EC3, EC4) central part of two juxtaposed (split-) U-domains, we found over the range of distances 0.7 Mb $\lesssim s \lesssim$ 7 Mb, a power-law decay $P_c(s) \sim s^{-3/2}$ ($\alpha = 3/2$) as between pairs of loci on either sides of a master replication initiation zones common to an active early (EC1+EC2) CTR and a (split-) U-domain (Fig. 5.15). This is consistent with the similar contact probability behaviour previously observed between active early (EC1+EC2) CTR loci, as well as between (early/late) pairs of loci (Fig. 5.2). This confirms the absence of spatial compartmentalisation in a more plastic and accessible pluripotent chromatin that statistically seems to be well described by the 3D "equilibrium" globule model.

To further quantify these observations, we compute the betweenness centrality across (split-) U-domain borders after sorting the borders according to their status (Fig. 5.16). We first consider the weighted intrachromosome Hi-C network with all the interactions, then we remove edges connecting loci distant by less than 1 Mb. MaOris (both MRT peaks and early CTR/U-domain common borders) are found to be "hubs" in the H1 ES Hi-C interaction network corresponding to local maxima of the centralities (Fig. 5.16 A, B) both when considering all the interactions and when considering only long-range interactions. When considering all the interactions, we notice that the randomly distributed early firing origins in active early CTRs sharing a common master replication origin with a (split-) U-domain, have a similar high centrality as the border, strongly suggesting that early replicating C1 100 kb loci are the dominating "hubs" in the H1 ES chromatin interaction network. This is consistent with the observed enrichment of these loci in CTCF [53, 54], which besides its insulator properties [169, 203] is also known to mediate long-range intra- and inter- chromosomal interactions in somatic cells [18, 208–212]. Moreover, these MaOris in H1 ES were shown to be highly enriched in the key pluripotency transcription factors NANOG and OCT4 that are known to have an important role in ESC-specific interactions and in the spatial clustering of pluripotency genes via the formation of (small) chromatin loops [361, 362]. These results shed a new light on these transcription factors that likely play also a role in the maintenance of the replication spatio-temporal programme in pluripotent cells [54]. In IMR90, MaOris that separate an early CTR and a replication domain, are local maxima of the betweenness centrality computed on the complete interaction graph (Fig. 5.16 C). Again, the vertices in C1+C2 CTRs have similar high centrality values (Fig. 5.16 C). However, in contrast to H1 ES (Fig. 5.16 A), the MRT peaks bordering two (split-) U-domains, are no longer local maxima of the betweenness centrality (Fig. 5.16 C). Also when concentrating on long range interactions, MRT peaks in IMR90 do not correspond to local betweenness centrality maxima (Fig. 5.16 D). When removing short range interactions, interestingly, but not surprisingly as regards to the slow power-law decay of the contact probability previously observed in IMR90 between HP1-associated heterochromatin C4 loci in inactive late CTRs (Fig. 5.2) and in the central regions of juxtaposed (split-) U-domains

(Fig. 5.15), the "hubs" of high betweenness centrality in the IMR90 chromatin interaction network are no longer the early replicating C1 loci but late replicating C4 loci (Fig. 5.16 D). This confirms the predominant role of the structural proteins that regulate the spreading of heterochromatin LADs in IMR90 [407].

☞ The 3D nuclear chromatin organisation differs between tissues and cell types as the signature of the chromatin folding induced by the self-interaction between chromatin states that promote physical bridging between distal elements, *e.g.*, via the specific interactions of some structural proteins. Thus, in the K562 cell line, the highly active early replicating euchromatin (100-kb) loci in early CTRs and in the master replication initiation zones at MRT U-domain borders were shown to be the main "hubs" in the chromatin interaction network [44, 408]. The observed enrichment of these loci in CTCF strongly suggests that CTCF is a key factor underlying long-distance intra- and inter chromosomal interactions in this cell line [18, 208–212]. In IMR90 cell line, as the signature of the important spreading of the HP1-assciated heterochromatin, the main "hubs" in the long range chromatin interaction network are instead the inactive late replicating heterochromatin loci in late CTRs and inside MRT U-domains. This suggests that, the structural proteins that regulate the anchoring of the Lamina B1 heterochromatin to the nuclear envelop [407] are determinant factors in the long-range interactions underlying, the high-order chromatin architecture in IMR90. In H1 ES the MaOris at replication domains borders appear to be fundamental determinants of pluripotency maintenance.

In all the cases, centrality measures from graph theory confirm the fundamental role played by the MaOris at replication domain borders in regulating the 3D organisation suggesting a compartmentalisation of the genome into structural units separated by these MaOris. In the next, chapter we will use multi-scale community detection to quantify the existence of structural domains as a counterpart of the replication domains.

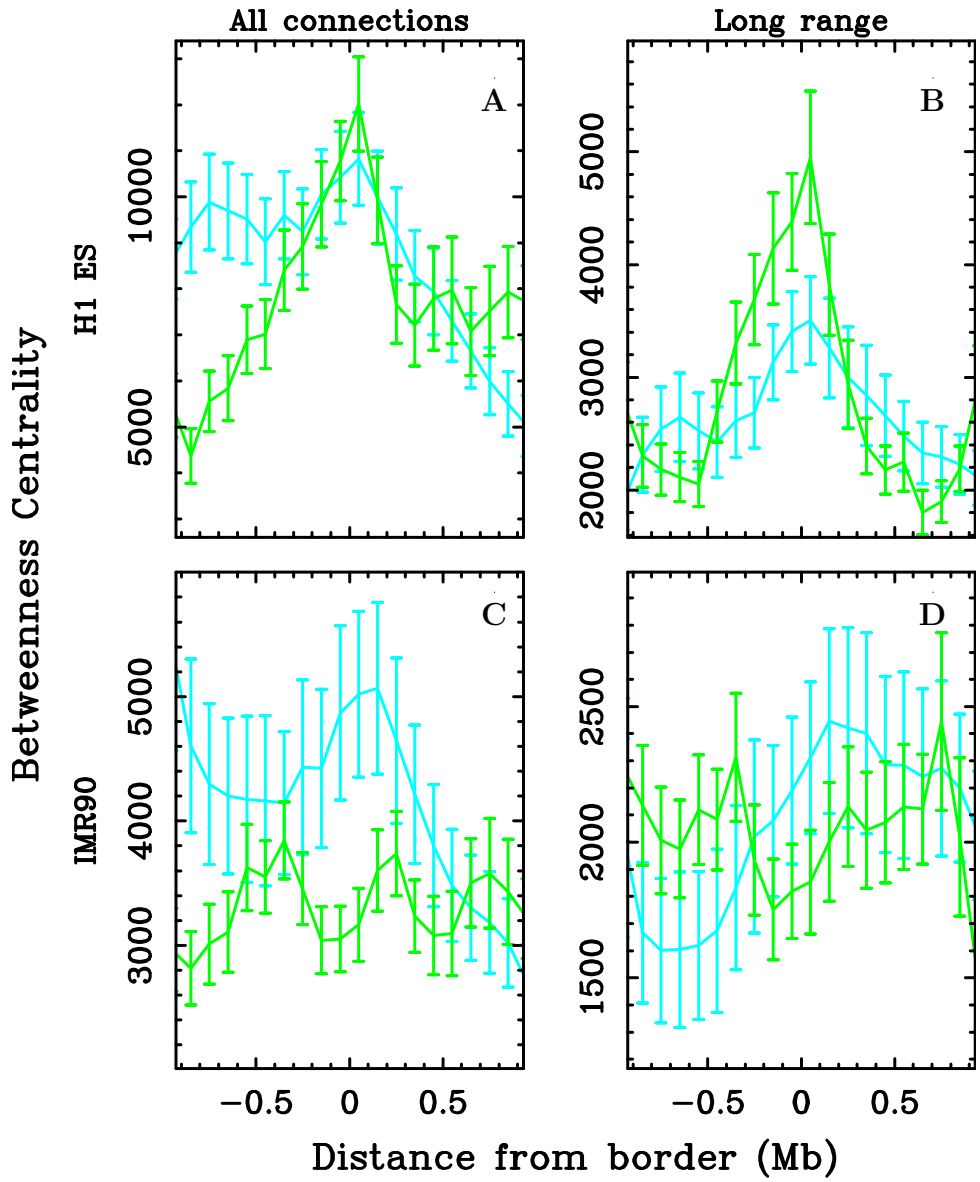


Figure 5.16. MaOri plasticity at the heart of the 3D genome organisation. Betweenness centrality across (split-) U-domain borders in H1 ES (A, B) and IMR90 (C,D): MRT peaks (green), dissymetric borders (blue). We only considered (split-) U-domains of length $L \geq 1$ Mb and we exclude late (split-) U-domain borders with MRT >0.5. We consider the complete weighted graph in A and C and the weighted graph with edges connecting loci distant by more than 1 Mb in B and D. The distance is relative to the borders: positive values indicate loci inside (split-) U-domain, the negative values can be either the loci in early CTRs (blue) or in a neighbouring U-domain (green).

Chapter 6

Delineating structural communities into the DNA interaction network

It's well recognised that the genome structure plays an important role in gene regulation, DNA replication and epigenetic modifications. Understanding genome structure is fundamental to study nuclear functions. In this chapter, we look for communities in the chromatin interaction network retreived from Hi-C data (known to probe the 3D structure). As these communities reflect groups of loci that present a high amount of interactions, we compare them to functional partitions of the genome to address the structure/function relationship. This analysis of structural networks obtained from different cell lines allow us to assess the level of structure conservation between cell types.

6.1 Hi-C data reveal chromatin organisation into structural domains

As discussed in Chapter 2, the spatial organisation of genome in the cell nucleus is highly linked to biological functions. Interestingly, 3D organisation and genome activity were shown to be linked at different scales in the nucleus. At the kilobase-scale, chromatin loops bring together distal regulatory elements such as enhancers and their target genes [409]. At the megabase scale genes co-occupy functional sites of the nucleus such as foci of Polycomb proteins [410], or of active RNA polymerase [411]. At the nucleus level, chromosomes cluster together forming chromosomes territories [1] that are organised in a way to put together gene-poor chromosomes in the predominantly heterochromatic periphery and gene rich regions in the euchromatic interior. Altogether these studies suggest a hierarchical multi-scale organisation of the genome. Until the emergence of 3C protocols (Chapter 2 Section 2.3.2), there was a lack of information allowing to precisely investigate the intermediary scales of genome 3D structuration. At the scale of $\sim$ 10 Mb, the analysis of Hi-C data [14] revealed the structuration of the genome into two compartments (A/B) (Chapter 2, Section 2.3.3) of loci sharing the same genome wide contact pattern. At the Mb scale, examination of Hi-C data (Figs. 2.16 and 6.1) shows the existence of diagonal blocks (squares) of enriched interactions [16]. Those squares arise when all the loci forming a genome interval preferentially interact with other loci in the same interval. Identification of those interaction blocks *i.e* extracting from the data some *structural motifs*, raises a methodological challenge for genomics in order to understand chromatin folding [16, 21, 244, 255–259, 412]. Many studies have been dedicated to extract structural motifs from Hi-C data based on looking for changes in a 1D profile derived from the Hi-C matrices [16, 412], on dynamic programming [258] or on a combination of both approaches [259].

The original study [16] describing diagonal blocks of Hi-C matrices as *topologically associated domains* (TADs) delineated block borders using the directionality index profiles. These genome 1D profiles take into account the upstream and downstream interactions of a genomic region, TAD borders corresponding to the transition from mainly upstream to mainly downstream interacting regions. The so identified kilobase-to-megabase sized TADs were found to correlate with many one dimensional features of the chromatin such as histone modifications [16], coordinated gene expression [21, 255], lamina association [16] and DNA replication timing [16, 256]. Moreover, TADs were found to be conserved in *Drosophila* [19, 20] and mammalian [16, 21] genomes but they were less clearly defined in *Arabidopsis* [22, 23], *Plasmodium falciparum* [24] and yeast [25, 26] genomes. Finally, TADs were suggested to be a stable property of the human genome as they appeared to be conserved between different cell lines [16]. Hence, topologically associated domains were included in many studies and are known to be at the heart of the genome 3D organisation although there is increasing evidence suggesting that TADs hierarchically co-associate to build up larger chromosomal structures [20, 413].

Another segmentation method for Hi-C data into TADs-like domains uses dynamic programming [259] that explicitly takes advantage of the genomic order. This method relies on the fact that Hi-C matrices are symmetric and instead of looking for squares of high interactions in the whole matrix, it is enough to look for triangle blocks along the di-

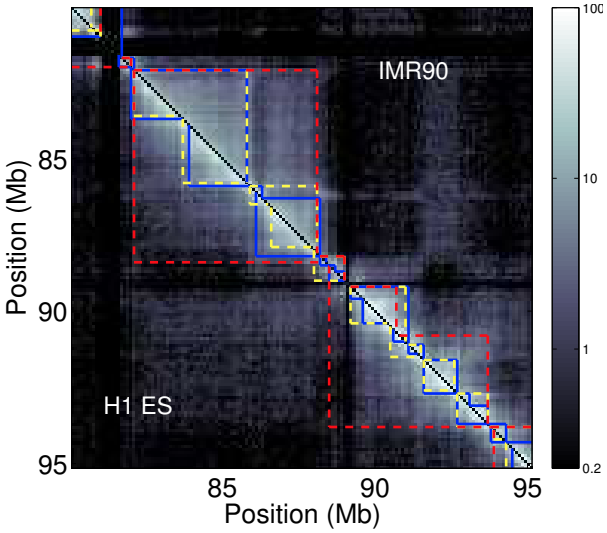


Figure 6.1. Hi-C contact maps reveal a multi-scale organisation. Hi-C contact map along a 15 Mb long fragment of human chromosome 10 in H1 ES (resp. IMR90) under (resp. above) the diagonal with intensity of interactions color coded according to the right colormap. Blue lines represent TADs [16] in the two cell lines. Colored dashed lines correspond to 2 partitions into communities obtained at small (yellow) and large (red) scales. Black columns and rows correspond to masked regions.

agonal of the matrix. The authors of [259] proved that maximisation of the likelihood with respect to the block boundaries can be reformulated in terms of a 1D segmentation problem. The blocks obtained with this method show good agreement with TADs. Block boundaries colocalise with TAD borders even though they are less numerous [259]. Also reformulating the question of decomposing Hi-C matrices in a 1D problem, the authors in [412] developed a wavelet-based change point method to detect changes in the total Hi-C interaction count profile allowing to segment the genome.

All these methods do not take into account the multi-scale organisation of the genome. In fact, looking at Figure 6.1, it is clear that TADs can be nested in bigger domains and sometimes subdivided in smaller domains.

Only the method proposed in [258] using dynamic programming developed a multi-scale segmentation algorithm of Hi-C data. Introducing a scale parameter γ to set the size of the intervals forming the interaction blocks, the authors identified a set of domains that show a high level of intra-domain interactions and a low level of between-domain interactions. Interestingly, comparison of the original TADs [16] with the domains obtained with this method at the resolution corresponding to the TADs, showes a good agreements between the two genome structural partitions. The aim of this work is to propose a novel method for Hi-C data segmentation that not only allows the multi-scale identification of structural domains but also does not rely on the 1D genome structure, in other words that does not assume that the structural domains should be contiguous genomic intervals.

6.2 Structural communities of DNA network form a hierarchy of genome intervals

6.2.1 Wavelet-based community detection in the DNA network

As discussed in Chapter 5 (Section 5.2.2), Hi-C matrices can be represented as graphs where nodes represent DNA loci and the edges connect interacting loci. Here, we reformulate the question of finding structural domains as a question of finding communities

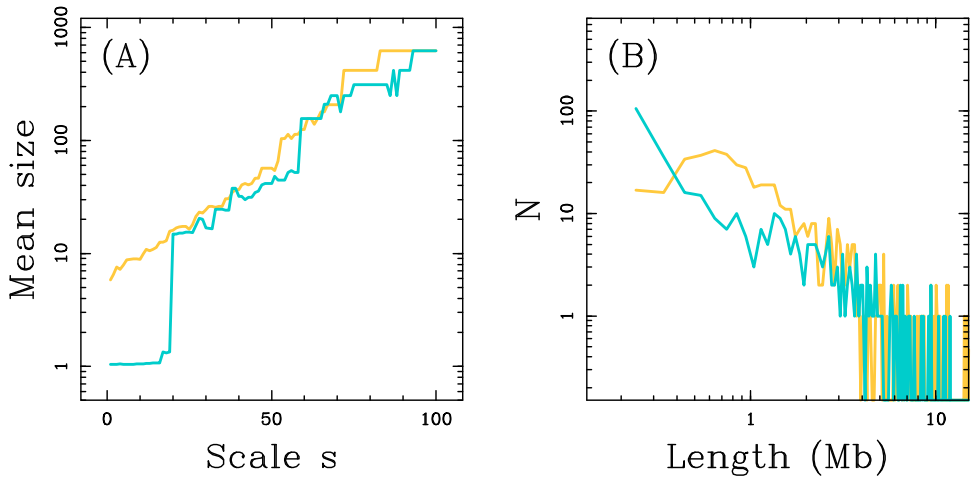


Figure 6.2. Community size across scales. (A) Mean community size for chromosome 12 as a function of the index of the scale in IMR90 (blue) and H1 ES (yellow) using the intra-chromosome Hi-C data. (B) Histogram of the length of the interval-communities that are left in the database after filtering trivial and redundant communities (see text).

in the DNA interaction network. The DNA interaction network depends on the genome structure only over the size (100 kb) of the regions used to define its nodes. As introduced in Chapter 3 (Section 3.6), a community in a graph is an ensemble of nodes that are more connected to each other than with the other nodes. Here, we use the wavelet-based multi-scale community detection method (Chapter 3, Section 3.6.3.1) in order to scan many scales without privileging any. We use the fast algorithm described in Section 3.5.1 (page 73) to compute the distance correlation matrix with $\eta = 200$ random vectors. We consider for each analysed cell line, the 22 intra-chromosomal matrices. In a first approach, we consider the Hi-C data from two human cell lines: IMR90 and H1 ES [16]. We filter the data as discussed in Chapter 2 (page 48), leaving $\sim 90\%$ of the data to study. We analyse for instance, 314 (resp. 326) 100 kb nodes for chromosome 21 in IMR90 (resp. H1 ES) and 2 179 (resp. 2 172) nodes for chromosome 1 in IMR90 (resp. H1 ES). The remaining 100 kb loci are concatenated which result in new *masked positions*. We systematically apply the wavelet-based multi-scale community detection method to all the connected interaction networks scanning 100 scales logarithmically distributed in the range of available scales (between $s_{min} = 1/\lambda_1$ and $s_{max} = 1/\lambda_1^2$ (page 76)).

The obtained results for the different chromosomes are quite similar. Here, we show the results obtained for human chromosome 12 in H1 ES and IMR90 as representative examples. This chromosome consists initially of 1 324 nodes. After the filtering procedure 1 250 nodes are left in IMR90 and 1 249 in H1 ES (Table 2.4). When applying the wavelet-based community detection method on the two networks, we obtain 100 partitions of the masked genome for each cell line, one at each scale. In total we obtain 23 927 (resp. 4 266) communities for IMR90 (resp. H1 ES). As expected, the size of the resulting communities increases with the scale parameter (Fig. 6.2 A). For H1 ES the increase of the mean community size with the scale is homogeneous suggesting that there is no characteristic size for the community structure. For IMR90, we observe a first range of scales where the communities reduce to singletons (mean size ~ 1) and followed by an abrupt transition to a community mean size ~ 17 (Fig. 6.2 A). The existence of singletons over a relatively large range of scales explains why the total the number of communities in IMR90 is larger

than in H1 ES. Note that when removing the trivial singletons communities we obtain 3 342 (resp. 4 266) community in IMR90 (resp. H1 ES).

6.2.2 Structural communities correspond to genome intervals

It is noticeable in Figure 6.1 that the interaction frequency outside the diagonal blocks characterising the structural compartimentalisation as described in Dixon *et al.* [16] is not negligible (look for instance at the region around [82,89] Mb in IMR90 that highly interacts with the region around [92,93] Mb (Fig. 6.1) or at the paving we observe away from the diagonal for IMR90 in Figure 5.1 B). This suggests that communities do not necessary reduce to intervals along the genome. Hence, for each non trivial community (community of size > 1), we calculate the proportion P_{int} of the largest set of successive 100 kb loci covered by the community over the size of the community: $P_{int} = 1$ when all the nodes of the community constitute an interval of the masked genome and $P_{int} = 1/N$ where N is the size of the community when the community do not contain any pair of consecutive loci of the genome. Considering $P_{int} \geq 0.95$ as a criterion for a community to constitute an interval along the genome, we observe for the 2 cell lines that more than 99% of the communities correspond to intervals of the genome. This property for the communities remains true for all the scales and whatever the size of the communities. This is consistent with the fact that at all scales, genomic neighbours tend to strongly co-localise, resulting in higher frequency of interactions. These results demonstrate that the strongest motifs of structural organisation involve contiguous genomic segments. We will refer to the communities forming a genomic interval as interval-communities.

In the following, we only keep the communities that correspond to an interval ($P_{int} \geq 0.95$) reducing them to their main interval. This allows us not only to adopt a simple representation for the obtained communities as illustrated on Figure 6.3 but also to construct our database (Section 6.2.4).

6.2.3 A hierarchical organisation of the genome

Figure 6.3 represents for a 20 Mb long fragment of chromosome 12, the obtained interval-communities across scales. The differences observed between the resulting community size distributions in IMR90 and H1 ES (Figure 6.2 A), are visible in this representation. We clearly see a first range of scales (≤ 20) where the interval-communities reduce to singletons in IMR90 (Fig. 6.3 A) and not in H1 ES (Fig. 6.3 B). At a larger critical scale, non trivial interval-communities appear in IMR90. Not that the mean size of the interval-communities for this first meaningful partitioning in IMR90 is larger than the ones observed in H1 ES for its first meaningful partitioning (smallest scale). This results in a lack of small non trivial interval-communities in IMR90.

A stricking property illustrated by this representation (Fig. 6.3) is the hierarchical organisation of the communities. Across scales, small communities merge together to form bigger communities at larger scales (Fig. 6.3). Hence, the community borders present at the smallest scale progressively disappear at some larger scale allowing the emergence of bigger communities. Importantly, the conservation of borders from large scales to small scales is very high as illustrated in Figure 6.4 for H1 ES. For each pair of scales $s2 > s1$, we look at the proportion of borders at the larger scale $s2$ that are also present at the

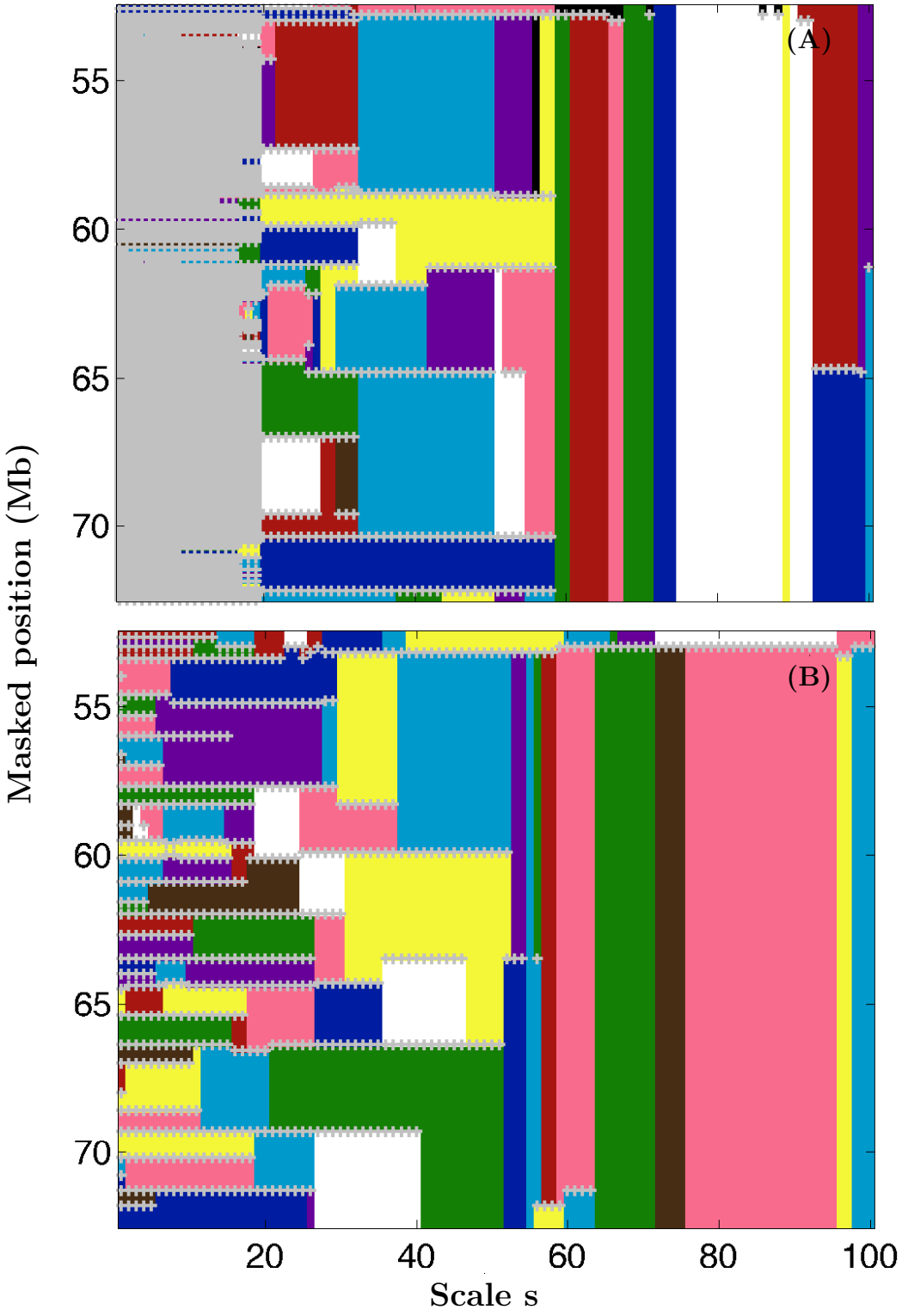


Figure 6.3. Multi-scale interval-communities. Multi-scale community structure along a 20 Mb long fragment of human chromosome 12 in IMR90 (A) and H1 ES (B) cell lines. At each scale the interval-communities are represented by a colored segment (colors were limited to 10 for readability) bordered by grey +. When a community is found at 2 consecutive scales the same color is used.

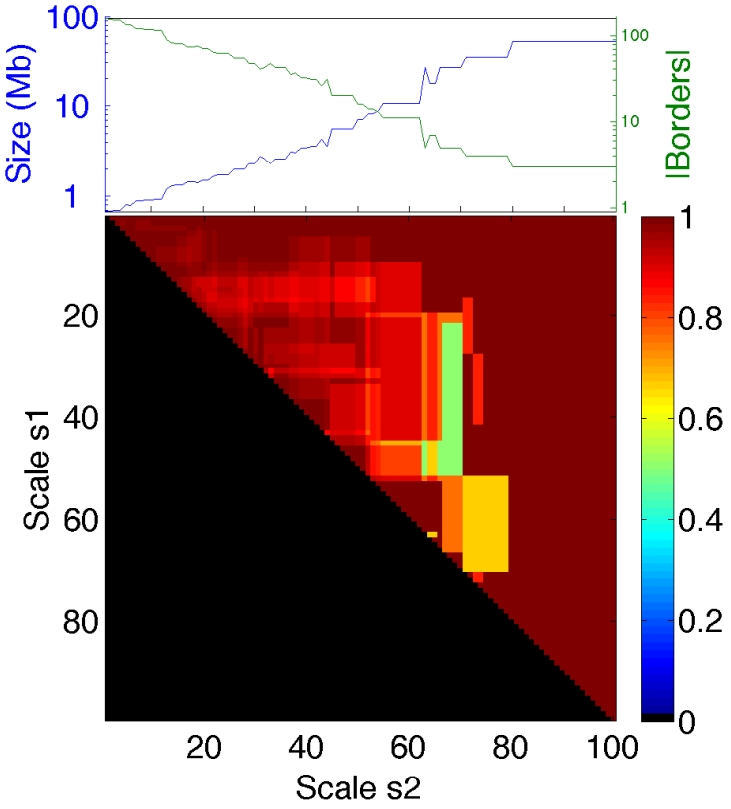


Figure 6.4. Conservation of community borders across scales. (Bottom) Proportion of borders at a scale $s2 > s1$ that are borders at scale s1, for human chromosome 12 in H1 ES cell line. (Top) The average size of communities and the number of borders as a function of s_2.

smaller scale $s1$. This proportion is close to 1 regardless of the scales (Fig. 6.4). Note that the smallest proportion of border recovery ($\sim$40%) is obtained at a scale $s2$ at which we have 5 borders and only two are recovered. At scale $s1$, the three missing borders being shifted few pixels away from a scale to another. The fact that the borders are conserved across scales means that there is no "new" structure that emerges and that only existent ones merged together, *i.e.* small structures are nested into bigger ones. This is consistent with the results of recent studies suggesting that TADs hierarchically co-associate to form larger structures [20, 413].

6.2.4 A hierarchical database of structural communities

Another important property illustrated on Figure 6.3 is the redundancy of the communities obtained across scales. Hence, we construct our database of communities keeping only once each non trivial interval-communities (size $\geq$ 2 nodes and $P_{int} \geq 0.95$). We also filter out the communities that at least double in size when reintegrating the masked regions of the genome, e.g. interval-communities spanning the centromers.

This leads to 386 (resp. 537) non trivial interval-communities in IMR90 (resp. H1 ES) for the chromosome 12. Interestingly, when looking at the genomic length distribution of those interval-communities (Fig. 6.2 B), we observe that IMR90 has more interval-communities involving only 2-3 nodes and a deficit in community size from $\sim$ 500 kb to $\sim$ 1.5 Mb relative to H1 ES. This is consistent with the structuration described previously (Figs. 6.2 A and 6.3). Moreover, for the communities of size $\gtrsim$ 2 Mb, we see that the length distribution is similar for the two considered cell lines suggesting that communities are conserved between cell lines in that scale range. We will further discuss the conservation of communities between different cell lines in Section 6.4.

We apply the wavelet-based community detection framework to the 22 human autosomes in different cell lines:

- H1 ES and IMR90 cell lines that we have used until now to illustrate our results and for which TADs [16] are available, allowing a direct comparison of our structural communities with what is considered as reference structural domains in the literature (Section 6.3.4).

- K562 and GM06990 cell lines that we will use further to discuss the conservation between cell lines (Section 6.4).

- HeLaS3 cell line where the data were obtained on synchronised cells during mitosis and G1 allowing a comparison of the community structure during the cell cycle (Section 6.3.5).

The results reported previously for the intra-chromosomal structural networks of chromosome 12 in IMR90 and H1 ES are representative of the results obtained for the 22 autosomes using the above mentioned Hi-C datasets. The genome-wide distributions of genomic length of the interval-communities in IMR90 and H1 ES (Fig. 6.5) are similar to the one observed for chromosome 12 (Fig. 6.2 B). When applied to different cell lines, the wavelet-based community detection method results in few thousands of interval-communities in each cell line as described in Table 6.1. Interestingly, the size distributions of the interval-communities obtained in GM06990, K562 and HeLaS3 cell lines are more similar to the one obtained in IMR90 than to the one of H1 ES (Fig. 6.5): there is more 200 kb communities and a deficit around $\sim$ 500 kb in differentiated cell lines relatively to H1 ES; for communities larger than $\sim$ 2 Mb the distributions are similar. An interpretation of the observed general excess of interval-communities of size $\sim$ 500 kb to $\sim$ 1.5 Mb in H1 ES compared to differentiated cell lines can be that cell differentiation is accompanied by the merging of the small structural communities in a structural consolidation scenario similar to the one described for replication timing domains (Section 4.2.4) [38, 147, 148, 364]. Interestingly, in a logarithmic representation, community size distributions (s $\gtrsim$ 2 Mb) follow a power-law behaviour s^{α} with $\alpha =\sim -1.3$ (Fig. 6.5). This suggests that there exists domains at all scales ($\gtrsim$ 2 Mb) without a characteristic size for the genome structuration. Note that if communities of size $\sim s$ form a partition of the genome of length L then the number of communities of this scale is equal to L/s leading to $\alpha = -1$ ($\gtrsim -1.3$).

☞ To sum up, the wavelet-based community detection method allowed us to identify structural-communities in the chromatin interaction networks that led to a database of several thousands of domains (Table 6.1). These communities form hierarchies of genome intervals which led us to call them interval-communities. Interestingly, it seems that these structural interval-communities may reflect some functional properties. In H1 ES cell line, characterised by smaller replication domains [44], we observe an excess of interval-communities in the range of scales from $\sim$ 500 kb to $\sim$ 1.5 Mb compared to the differentiated cell lines. Moreover, for interval-communities larger than 2 Mb, all the cell lines showed a similar length distribution suggesting that they are conserved from one cell line to another.

Cell line	N	N(filtered)	Remaining communities	Distinct borders
H1 ES	12 343	65	12 278	5 751
IMR90	8 852	25	8 827	6 824
GM06990	10 279	60	10 219	6 967
K562	13 383	30	13 353	8 273
HeLaS3 G1	6 752	36	6 716	4 108
HeLaS3 M	1 059	4	1 055	885

Table 6.1. Number of structural communities. For each cell line, N is the number of distinct non redundant and non trivial (size ≥ 2 *i.e.* 2 nodes) communities. N(filtered) is the number of communities filtered out because *(i)* they do not correspond to a genomic interval or *(ii)* they double in size when going back to the original (not masked) positions. The last two columns correspond to the number of communities and distinct borders in the database.

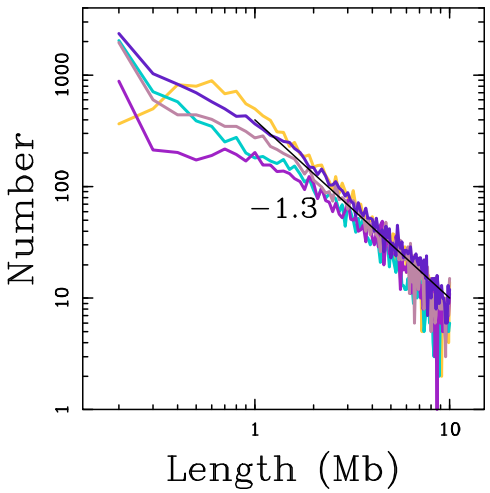


Figure 6.5. Genomic length distribution of non-redundant interval-communities. Histogram of interval-communities genomic length in a log-log representation and calculated in 100 kb bins for different cell lines IMR90 (blue), H1 ES (yellow), GM06990 (pink), K562 (purple) and HeLa (G1) (light purple). The black straight line correspond to the power-law behaviour with $\alpha = -1.3$.

6.3 Structural communities encompass genome segmentations at multiple scales

In this section, we enlighten properties confirming that our interval-communities indeed correspond to structural units. In particular, we explicitly compare the interval-communities to the TADs [16] that are considered as a reference for the structural description of Hi-C data. TADs [16] were identified at both 20 kb and 40 kb resolutions. However, our adopted resolution is 100 kb like the resolution of the MRT data. We thus, use the data obtained at the 40 kb resolution, and assign each TAD border to the corresponding 100 kb pixel and finally keep only TADs larger than 200 kb (3 pixels). This leads to a database of 2 993 (resp. 2 263) TADs in H1 ES (resp. IMR90), with 3 905 (resp. 3 096) distincts borders in H1 ES (resp. IMR90).

6.3.1 Are interval-communities structural domains?

A first question that one can ask is if the communities are "really" structural units of the 3D DNA organisation inside the nucleus. In fact, the community detection method will always output a partition at each scale whether or not the analysed network presents a "true" community structure. In order to verify that there are more interactions within interval-communities than between interval-communities, we compare the number of contacts between two 100 kb loci that are inside the same interval-community at equal distance from its center and the number of interactions between two loci at equal distance from one of the interval-community borders, as a function of the distance separating the pairs of loci. The obtained average ratio of different interval-communities is reported in Fig. 6.6. We classify our interval-communities in different size categories: $0.3 \leq L < 0.6$ Mb, $0.6 \leq L < 1$ Mb, $1 \leq L < 2$ Mb, $2 \leq L < 3$ Mb, $3 \leq L < 5$ Mb, $5 \leq L < 10$ Mb and $10 \leq L < 100$ Mb. We observe that on average there is more interactions within the communities than between communities, regardless of the cell line and the community size: the interaction ratio systematically increases to some maximal value at distances $\sim$ 1-2 Mb, from ~ 1.6 in GM06990 and K562, to ~ 2.2 in H1 ES and ~ 3 in IMR90. Over larger distances, the ratio remains rather constant in GM06990 and K562 and decreases to ~ 1.5 in H1 ES and IMR90 (Fig. 6.6). This property holds true, even for domains larger than 10 Mb, corroborating the correspondence between structural barriers and community borders as previously observed for (split-) U-domain borders (Fig. 5.5) [44]. As a comparison, we perform the same analysis at the TAD borders grouped in the following size categories: $0.3 \leq L < 0.6$ Mb, $0.6 \leq L < 1$ Mb, $1 \leq L < 2$ Mb and $2 \leq L < 3$ Mb. The interaction ratio *vs* distance curves present very similar shapes as observed for interval-communities (Supplementary Fig. B.12), reaching maximal values ~ 3 in both H1 ES and IMR90. These results provide evidence that interval-communities, very much like TADs, constitute units of 3D genome organisation bordered by structural barriers.

6.3.2 Are chromosomes structural communities in the full interaction network?

It has been suggested that Hi-C data contain enough information to allow us to distinguish between different chromosomes and, hence, can be used for genome assembly [29, 30]. This raises the question whether the wavelet-based community detection method can recover

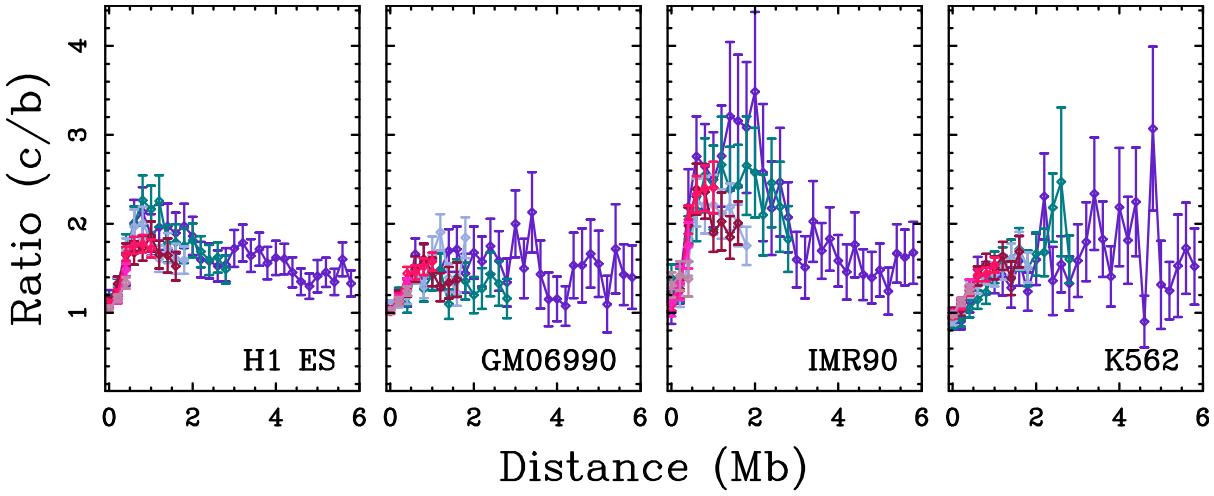


Figure 6.6. Are interval-communities structural domains? Ratio of the number of interactions between two 100 kb loci that are inside the same community at equal distance from its center (c) and the number of interactions between loci in different communities at equal distance from a community border (b), versus the distance between them. Different colours correspond to different community size categories: $0.3 \leq L < 0.6$ Mb (light pink), $0.6 \leq L < 1$ Mb (pink), $1 \leq L < 2$ Mb (magenta), $2 \leq L < 3$ Mb (dark pink), $3 \leq L < 5$ Mb (light blue), $5 \leq L < 10$ Mb (blue) and $10 \leq L < 100$ Mb (purple).

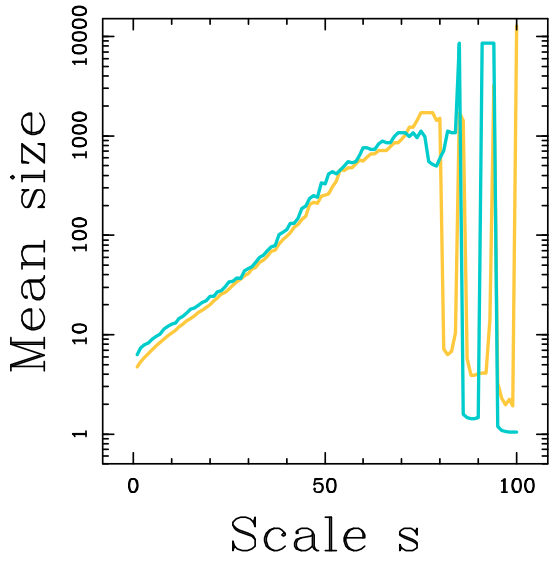


Figure 6.7. Mean community size across scales in the full interaction network. Mean community size (number of nodes) as a function of the index of the scale in IMR90 (blue), H1 ES (yellow) when using the full Hi-C network of both intra- and inter- chromosomal Hi-C interactions between the 22 autosomes.

chromosomes as communities in the chromatin interaction network. To adress this question, we consider the "full" interaction network in H1 ES and IMR90 where the nodes are the 100 kb loci coming from all the 22 autosomes and the edges are both intra- and inter-chromosomal Hi-C interactions. We then apply the community detection algorithm to these "full" networks.

Note that this way of constructing the network does not allow us to scan the same range of scales as in the chromosome by chromosome approach. As described in Chapter 3 (Section 3.5.2), the range of available scales depends on the first eigenvalue of the network Laplacian ($s_{min} = 1/\lambda_1$ and $s_{max} = 1/\lambda_1^2$) which depends on the graph structure. When the graph compartmentalisation is marked (*i.e.* there seems to be distinct connected components), λ_1 is small and close to 0 ($\lambda_1 = 0$ when the graph is disconnected). In the full interaction network, the long range interactions (interactions between different chromosomes) are less frequent [14] and hence the graph has 22 well connected components with few connections between them. Hence, λ_1 is smaller which results in larger values of s_{min} and s_{max}. Thus, the range of scales is shifted towards larger scales compared to the chromosome by chromosome approach.

At the largest scales previously unavailable, we observe instabilities (Fig. 6.7) of the community detection method. As can be seen in the mean size *vs* scale curve in Figure 6.7, for scales $\gtrsim 80$ we observe partitions made of large communities, as expected, alternating with partitions made of trivial singletons. At the smallest scales, contrary to what we observe on the intra-chromosome networks (Fig. 6.2 A), we no longer obtain many singletons because of the new scale boundaries. Between the two extremes, the mean size of the communities is similar between IMR90 and H1 ES (Fig. 6.7). Further examination of these communities reveals that they either correspond to "small" communities fully embedded within one chromosome or include one or more complete chromosomes. Hence, in both H1 ES and IMR90, the structural community detection respects the organisation of the genome into chromosomes. In H1 ES, all the chromosomes were identified as communities, at a certain scale relative to each chromosome. In IMR90, a few counter examples are observed either because at the largest scale available the chromosome still corresponds to few communities or because intra-chromosome community partitions abruptly switch to partitions where communities encompass 2 or more chromosomes.

To sum up, the wavelet-based community detection method successfully recovers chromosomes as communities in the full interaction networks. Intra-chromosomal communities detected in the full interaction network are consistent with the ones obtained with the chromosome by chromosome approach, with some noise in the position of interval-community borders across the scales (moving $\pm$ 1 pixel from a scale to another). But, as the range of scales depends on the matrix, to avoid loosing small scale details, it is better to concentrate on intra-chromosomes data. In the rest of this work, we only consider the interval-communities obtained when applying the wavelet-based method on individual chromosomes, as described in Table 6.1.

6.3.3 Comparing genomic domain distributions

In the rest of this chapter, we adopt the three following points of view for the comparison of genomic domains. Consider two sets of domains $\mathcal{D}_1$ and $\mathcal{D}_2$ with two sets of associated borders $\mathcal{B}_1$ and $\mathcal{B}_2$ that we want to compare. We consider:

- Mean best mutual coverage: We define the mutual coverage m_c between two domains $d_1 \in \mathcal{D}_1$ and $d_2 \in \mathcal{D}_2$ as their intersection length $L_{d_1 \cap d_2}$ divided by the maximum length of the two domains lengths L_{d_1} and L_{d_2}:

$$m_c(d_1, d_2) = \frac{L_{d_1 \cap d_2}}{\max(L_{d_1}, L_{d_2})}. \tag{6.1}$$

  The maximal value 1 of m_c is obtained when the two domains d_1 and d_2 are identical. Then, for each domain $d_1 \in \mathcal{D}_1$, we define its best mutual coverage with $\mathcal{D}_2$ domains ($bm_{c_{\mathcal{D}_2}}$) as its maximal mutual coverage with $\mathcal{D}_2$ domains:

$$bm_{c_{\mathcal{D}_2}}(d_1) = \max_{d_2 \in \mathcal{D}_2}(m_c(d_1, d_2)). \tag{6.2}$$

  Sorting the $\mathcal{D}_1$ domains by size, we compute the mean best mutual coverage with $\mathcal{D}_2$ of groups of 50 $\mathcal{D}_1$ domains that we plot as a function of the mean length of the domains in the group.

- We say that a domain d has a match in $\mathcal{D}_2$ if $bm_{c_{\mathcal{D}_2}}(d) \geq 0.8$. $P_{\mathcal{D}_2}(\mathcal{D})$ is then defined as the proportion of domains $d \in \mathcal{D}$ that have a match in $\mathcal{D}_2$. Sorting the $\mathcal{D}_1$ domains by size, we consider them in groups $\mathcal{D}$ of 50 domains and plot $P_{\mathcal{D}_2}(\mathcal{D})$ as a function of the mean length of the domains in $\mathcal{D}$.

- We say that a border b has a match in $\mathcal{B}_2$ when there is a border in $\mathcal{B}_2$ less than 100 kb away from b *i.e* $\pm$ 1 pixel away. $P_{\mathcal{B}_2}(\mathcal{B})$ is then defined as the proportion of borders $b \in \mathcal{B}$ that have a match in $\mathcal{B}_2$. Sorting the $\mathcal{B}_1$ borders according to their associated size (see below), we consider them in groups $\mathcal{B}$ of 100 and plot $P_{\mathcal{B}_2}(\mathcal{B})$ as a function of the average associated size of the borders in $\mathcal{B}$.

Interval community borders need to be classified according to the size of the interval-communities they border prior to being used as a reference borders. Indeed, it is clearly not as significant for a border to have a match with the borders of 200 kb interval-communities than with borders of 2 Mb interval-communities which are one order of magnitude less numerous than the former (Fig. 6.5). Since borders are conserved across the scales (Fig. 6.4) and, at each scale, they delimit two consecutive interval-communities, we assign a size to each interval-community border in the following way. At each scale, the border is associated with the minimum length of the two bordering communities. The largest of these lengths across the scales is finally retain as the "size" associated to the border. In this way, interval-community borders that exist over a large range of scales are likely to be associated to a large size than borders that only exist at small scales. As expected, the number of distinct borders as a function of their associated size decreases rapidly (Fig. 6.8). We compute the proportion of the genome covered by the interval-community border pixel and its two neighbouring pixels ($\pm$1 pixel resolution) for all borders of associated size greater than some minimal size (Fig. 6.9). As expected, this proportion decreases as a function of the minimal size (Fig. 6.9). Setting the proportion covered by the interval-community borders to 35%, we end up with different numbers of

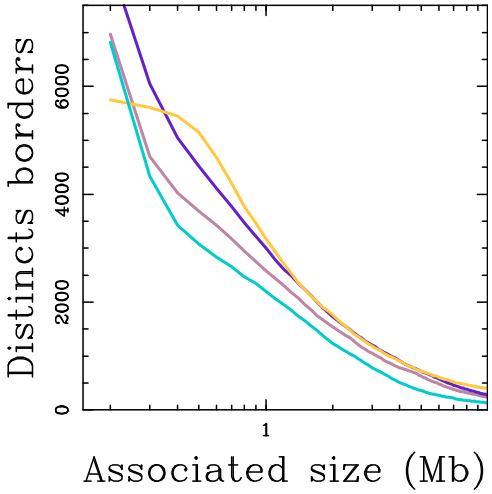


Figure 6.8. Distribution of the interval-community border associated size. Number of distinct borders when classified according to the associated size (see text) in different cell lines: IMR90 (blue), H1 ES (yellow), GM06990 (pink) and K562 (purple).

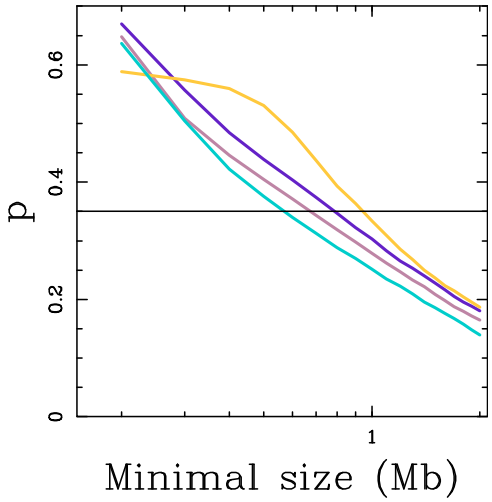


Figure 6.9. Genome proportion covered by interval-community borders. The genome fraction covered by the interval-community borders (borders pixels and their two neighbours) of associated size greater than a minimal size *vs* this minimal size in different cell lines: IMR90 (blue), H1 ES (yellow), GM06990 (pink) and K562 (purple). The horizontal black line corresponds to the 35% threshold we adopt to classify the borders.

distinct borders: 3 468 in H1 ES, 2 834 in IMR90, 3 171 in GM06990 and 3 478 in K562 (Figs. 6.8 and 6.9). In the analysis presented next, we use this set of conserved borders that cover 35% of the genome at ± 1 pixel resolution as interval-community borders reference set.

6.3.4 Are TADs structural communities?

We next compare our communities to the previously TADs [16] in H1 ES and IMR90, asking to which extent the TADs and TAD borders are recovered in our hierarchical database of interval-communities (Section 6.2.4). We follow the comparison framework described in Section 6.3.3 associating the size of the shortest neighbouring TAD to each TAD border.

Figure 6.10 (A) shows the behaviour of the mean best mutual coverage between TADs and interval-communities as a function of TADs size. Mean best mutual coverage is slightly higher in H1 ES as compared to IMR90 for all sizes: ranging from 62% (resp. 52%) at small size (300-500 kb) to 91% (resp. $\sim$ 89%) at larger size ($\sim$ 1-2 Mb) in H1 ES (resp. IMR90). This suggests a good recovery of the largest TADs by the interval-community classification. Given the 100 kb resolution used in this analysis, it is not surprising to observe lower mutual coverages at small sizes where 1 pixel error results in a dramatic

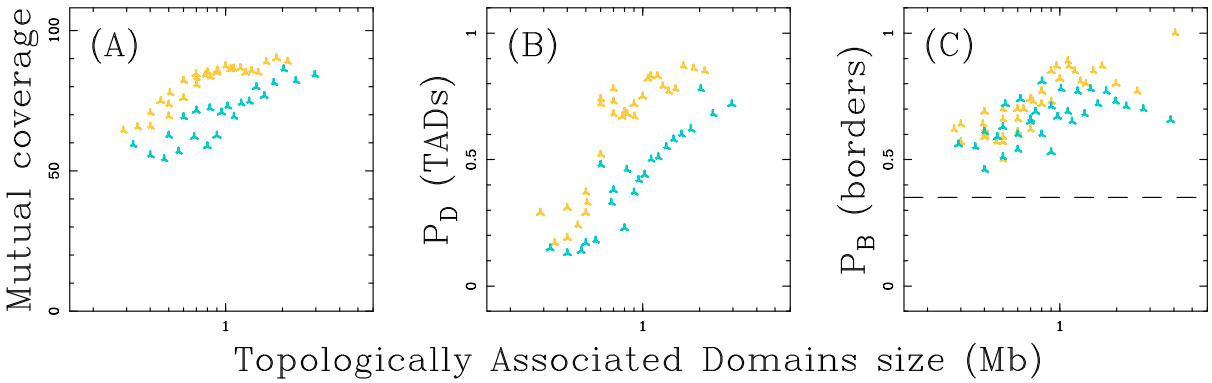


Figure 6.10. TADs are structural communities. Mean best mutual coverage (Equation 6.2, Section 6.3.3) of TADs with interval-communities (A); proportion of TADs that have a match in the interval-community database (Section 6.3.3) (B) and proportion of TAD boundaries that have a matching interval-community border (C) as a function of the TAD size (Section 6.3.3), in H1 ES (yellow) and IMR90 (blue). In (C) the TAD size associated with a border is the minimum of the size of the two bordering domains; only the set of interval-community borders covering 35% of the genome are used (Section 6.3.3). The black horizontal dashed line shows the expected border matching proportion.

lowering of mutual coverage. Now looking at the proportion of TADs that have a matching structural community, we see that, the proportion of recovered TADs increases with the domain size (Fig. 6.10 B). Only about 1/5 of the smallest TADs ($\lesssim$ 500 kb) are recovered consistently with the fact that in this scale range a match has to be exact. For TADs longer than 1 Mb the proportion of match is relatively high: in IMR90 it increase from 40% for TADs of 1 Mb up to 70% for TADs $\geq$ 2 Mb and in H1 ES from 70% for TADs $\sim$ 1 Mb up to 85% for TADs of $\sim$ 2 Mb (Fig. 6.10 B). Comparison of TAD borders to interval-community borders shows good concordance for the two datasets (Fig. 6.10 C). In fact, we classify the TAD borders according to the smallest TAD size they border and for the interval-community borders we restrict the analysis to the borders with associated size large enough so that at 100 kb resolution ($\pm$ 1 pixel) they collectively cover 35% of the genome (Section 6.3.3). We clearly see that H1 ES TAD borders are recovered from 50% up to $\sim$ 90% depending on the TAD border associated size, and that IMR90 TAD borders were recovered from 50% up to $\sim$80%, while the expected match by chance is 35% (Fig. 6.10 C). These results quantify the high level of TAD recovery by interval-communities for domain length $\gtrsim$ 1 Mb.

Altogether, these results show that there is a significant agreement between TADs and the interval-communities. This provides further evidence that the interval-communities are indeed structural domains of the human genome.

6.3.5 Structural communities during the cell cycle

To further test the robustness of the wavelet-based community detection method with respect to the possible absence of a community structure over some scale range, we apply the method to two Hi-C datasets obtained in synchronised HeLaS3 cells during G1 and M phase, respectively [243]. A recent study [243] showed that the highly compartmentalised organisation described before from non synchronous cells [14, 16, 242, 244, 245] is restricted to interphase and that during a cell cycle, chromosomes transit from a de-

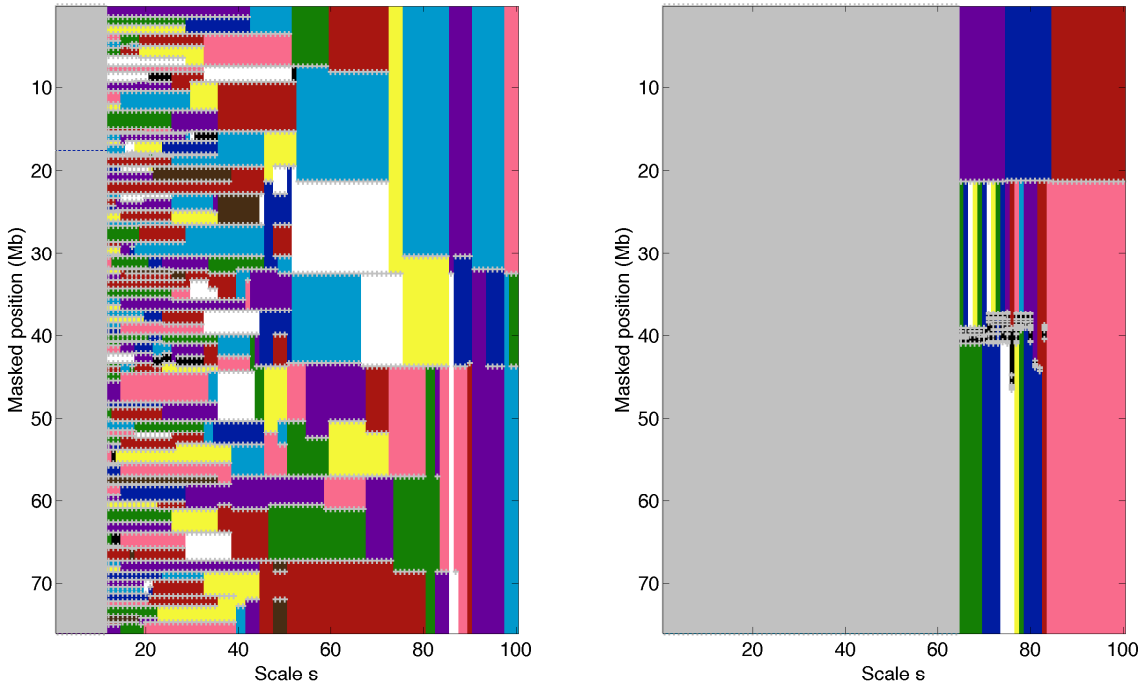


Figure 6.11. Structural communities during the cell cycle. Same as in Fig. 6.3 for HeLaS3 cells during G1 (left) and mitosis (right), along the complete masked chromosome 17.

condensed and spatially organised state during interphase to a highly condensed and morphologically reproducible metaphase chromosome state [243]. This study provides Hi-C data for HeLaS3 cells during mid-G1 and metaphase. In the former phase, the interaction maps display similar plaid patterns of regional enrichment or depletion of long range interactions (as the one shown in Figure. 6.1) while the maps in mitotic cells change and the plaid patterns disappear [243]. Here, we analyse the Hi-C matrices with our wavelet based multi-scale community detection method. For each chromosome, we construct two intra-chromosomal structural networks at 100 kb resolution: one with the mid-G1 data and the other one with the mitosis data, and we apply separately the community mining method on each network (Section 6.2.4). For mid-G1 HeLaS3 cells we obtain 6 752 non trivial communities from which we filter out 36 that do not correspond to genomic intervals resulting in 4 108 distincts borders (Table 6.1). For mitosis cells, we obtain 1 059 communities from which we filter out 4 resulting 885 distincts borders (Table 6.1). Figure 6.11 shows for chromosome 17 the communities at the two moments of the cell cycle. Consistently with non synchronous cells, G1 cells present a hierarchical structure into interval-communities that increase in size across scales (Fig. 6.11). At small scales, we observe singletons that are then grouped to form bigger communities. Note that the size distribution of the interval-communities in HeLaS3 (G1) is similar to the previously described interval-communities size distribution in IMR90 (Fig. 6.5). However, for metaphase cells, for more than half of the scale range ($s \lesssim 65$), each node is considered as a community. At larger scales, we observe a sharp discontinuity of the community sizes distribution: nodes are abruptly grouped in 3 then 2 communities (Fig. 6.11). Interestingly, the obtained two large scale communities in chromosome 17

correspond to the two chromosome arms. This result is representative of what we observe for all chromosomes. More specifically, chromosomes 16, 21 and 22 do not show any structure (each node constituted a community on the full available range of scales). The 18 other autosomes show similar metaphase structuration pattern as chromosome 17 (Fig. 6.11) where at small scales each node is considered as a community and then after an abrupt transition nodes are grouped in 2 to 5 communities. For 11 out of these 18 chromosomes, when divided in two communities, these communities correspond to the two chromosomes arms. These results demonstrate that the wavelet-based community detection method does not produce misleading intermediate scale communities when no structuration exists in that scale range.

☞ Altogether these results confirm that the wavelet-based community detection is able to capture meaningful structure from the Hi-C data. When considering intra-chromosomal Hi-C data, we recover the TADs (the reference domain-like organisation of the chromatin) for sizes significantly larger than the resolution of the analysis. When considering all Hi-C interactions inside chromosomes and in between chromosomes, we find that chromosomes are structural units. Finally, when considering data at two different moments of the cell cycle, one with a domain-like organisation and another one with no structural properties, the wavelet-based community detection method is able to robustly capture these two situations.

6.4 Structural communities are robustly observed across cell lines

In the pionnering study [16], TADs were described to be conserved between cell lines. We observed that interval-communities in different cell lines, present similar size distributions (Fig. 6.5). This led us to investigate to which extent they are conserved across cell lines. To compare the communities obtained in different cell lines, we use each of the database of interval-communities obtained in H1 ES, GM06990, IMR90, K562 as reference domain sets and compute the proportion of conserved interval-communities of the 3 other cell lines relative to this reference set. As explained in Section 6.3.3, an interval-community has a matching interval-community in the reference set if its best mutual coverage with the reference set (Equation 6.2) is greater than 0.8. Figure 6.12 shows that small interval-communities ($\lesssim$ 600 kb) are not well conserved between different cell lines. However, when considering interval-communities of larger sizes, higher conservation is observed (Fig. 6.12). For instance, more than 60% of intervals-communities of length $L \gtrsim 0.6$ Mb in the differentiated cell lines correspond to an interval-community in H1 ES (Fig. 6.12 A). H1 ES interval-community dataset thus contains a large proportion of the interval-communities observed in the differentiated cell lines above $\sim$ 600 kb; we recall that below this threshold our analysis is sensitive to the 100 kb resolution (Section 6.3.4). When using one differentiated cell line interval-community database as reference, we observe a maximal recovery rate that is similar for the 3 other cell lines: 45% for sizes $\gtrsim$ 2 Mb in IMR90 (Fig. 6.12 C), 65% for sizes $\gtrsim$ 1.5 Mb in GM06990 (Fig. 6.12 B) and 70% for sizes $\gtrsim$ 1.5 Mb in K562 (Fig. 6.12 D). The observed differences likely reflect the excess of interval-communities in the size range 0.5-1.5 Mb observed in H1 ES compared to the differentiated cell lines (Fig. 6.5). These domains not observed in differentiated cell lines might be subject to some structural consolidation scenario during cell differentiation. As

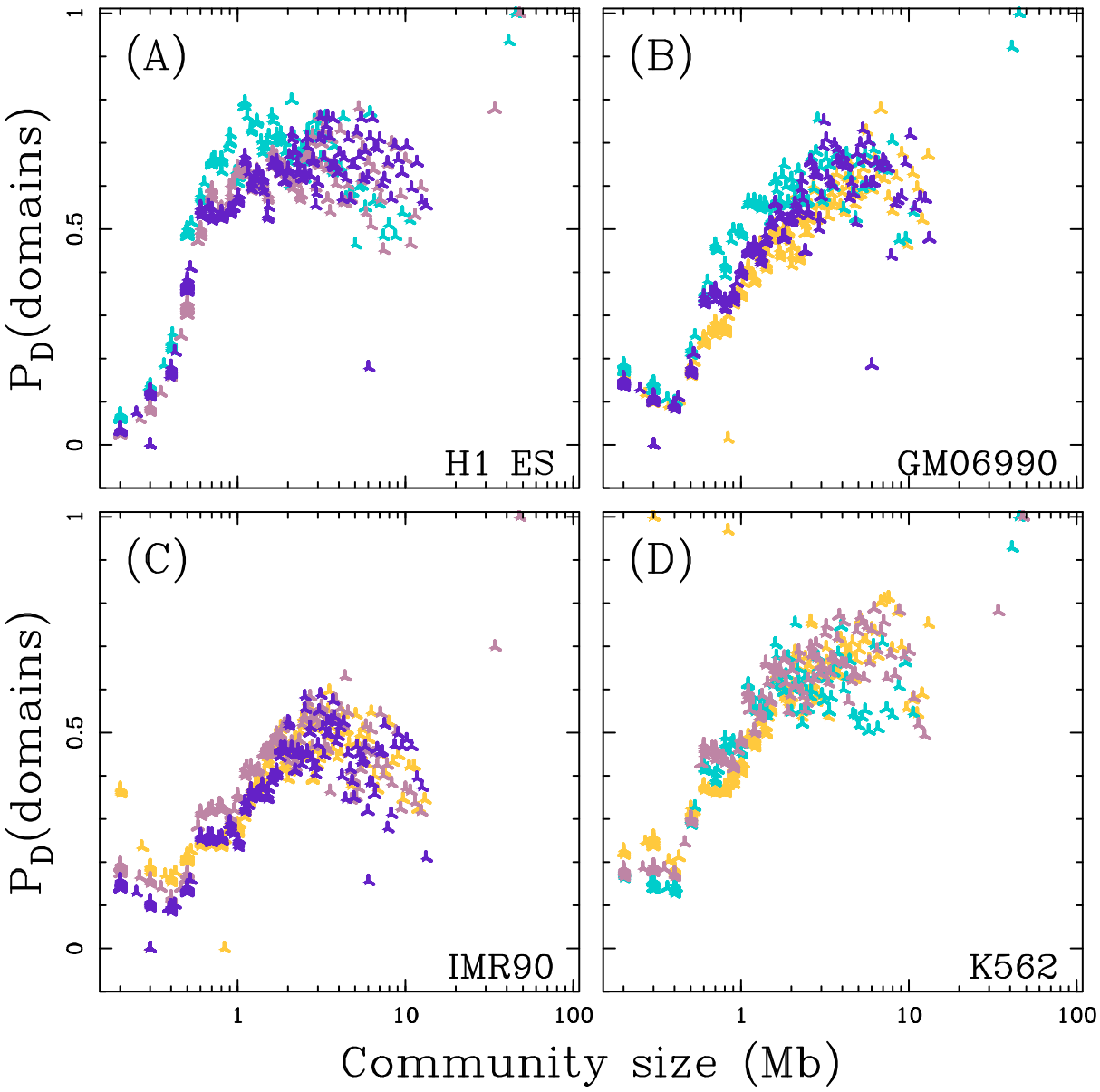


Figure 6.12. Matching interval-communities between cell lines. For each reference cell line (the different plots), we look at the proportion of interval-communities in the different query cell lines H1 ES (yellow), IMR90 (blue), GM06990 (pink) and K562 (purple). An interval-community of a cell line has a match in the reference cell line when there exists an interval-community in the reference cell line such that the two interval-communities have a mutual coverage larger than 0.8 (Section 6.3.3). Proportion of interval-community matches is computed over groups of 50 query interval-communities ordered by size (Section 6.3.3).

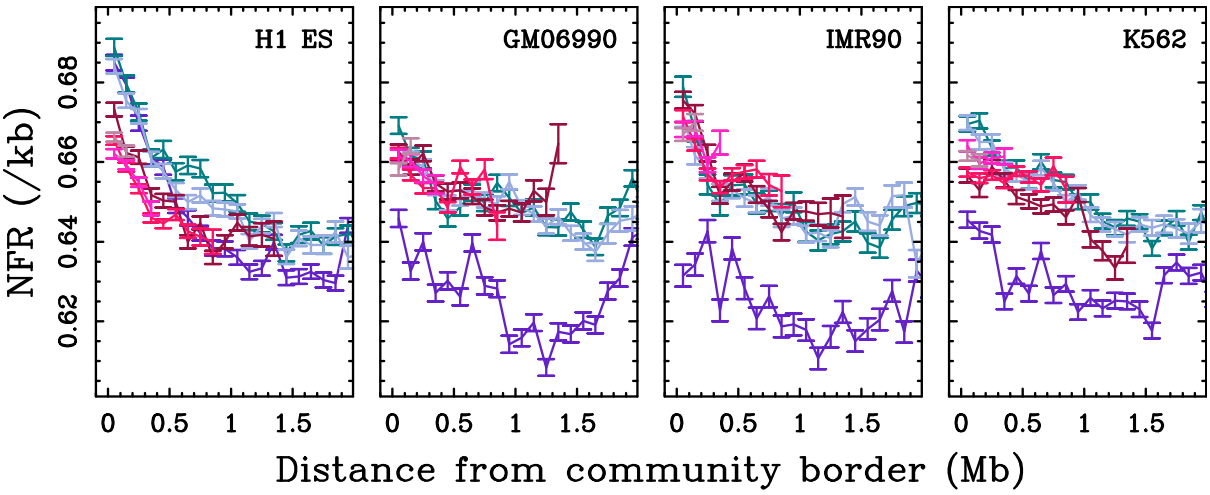


Figure 6.13. Sequence encoded NFR density around community borders. Mean excluding energy barrier density (per kb) as a function of the distance from the closest interval-community border in different cell lines and in different interval-community size categories (color coded like as in Fig. 6.6).

a comparison, we perform the same analysis over the TADs that were shown to be conserved between H1 ES and IMR90 cell lines [16](Supplementary Fig. B.13). As observed when comparing interval-communities, the correspondance between TADs in the two cell lines decreases for domain sizes $\lesssim 600$ kb. For larger domain sizes, we observe that H1 ES TAD dataset contains more (maximal value $\sim 60\%$) of the IMR90 TADs than the IMR90 TAD dataset contains H1 ES TADs ($\sim 45\%$). These results underline a conservation of structural domains between cell line in the 45%-70% range for both interval-communities and TADs of size ~ 1 Mb up to the largest interval-communities of size ~ 10 Mb.

Previous analyses of replication (split-) U-domain have shown that ubiquitous U-domain borders systematically found in 6 different cell lines are encoded in the DNA sequence via a local enrichment in nucleosome excluding energy barriers [365]. Here we ask to which extent this sequence-encoded chromatin property could explain structural domain conservation across cell lines. Previous work revealed that promoter regions for protein-coding genes are extremely hypersensitive to DNase I digestion [184]. These regions were shown to be nucleosome depleted [180–184], very much like the nucleosome free regions (NFRs) observed at yeast promoters [186, 187]. Numerical studies showed that, to a large extent, these NFRs are coded in the DNA sequence via high energy barriers that impair nucleosome formation [414–416]. Furthermore, these excluding genomic energy barriers were shown to play a fundamental role in the collective nucleosomal organisation observed over rather large distances along the chromatin fiber [414]. Using the same physical modeling of nucleosome formation energy based on sequence-dependent bending properties introduced for modeling nucleosome occupancy profiles in the yeast genome [414, 415] sequence encoded NFRs were identified in the human genome as the genomic energy barriers that are high enough to induce a nucleosome depleted region in the nucleosome occupancy profile [51, 365].

When mapping these intrinsic NFRs inside the interval-communities (Fig. 6.13), we observe an enrichment around the community borders for all the cell lines. Note that the decrease from the borders to the center of the interval-communities is sharper in H1

ES than in the other cell lines and this for all the interval-communities sizes. For the largest communities (size $\geq$ 10 Mb), the NFR density is significantly lower and almost flat in the differentiated cell types. As a control, the same analysis of TAD borders (data not shown), leads to the same behaviour. The increase in NFR density around interval-community and TAD borders suggests that these borders are at least partly encoded in the DNA sequence which may explain their conservation across cell lines.

☞ Hence, the interval-communities identified with the wavelet-based community detection method seem to be stable between different cell lines. The enrichment in sequence-encoded NFRs at the borders suggests that these conserved structural communities are specified by a genetic mechanism.

6.5 Structure-function relationships in the nucleus

TADs were previously shown to correlate with the DNA one dimensional features such as replication timing [16, 256] and chromatin states [244]. We investigate in this section the relationships between (i) replication split-U- and U- domains (Sections 2.1.4.3 and 4.1.2), (ii) chromatin states blocks (Section 2.2.2) and (iii) interval-communities (Section 6.2.4).

6.5.1 Are replication domains structural communities?

Replication domains appeared in a preliminary study to correspond to structural domains (Section 2.4, Fig. 2.19) [44] and their borders were shown to have a particular role in regulating the DNA structual network (Chapter 5). Here, we address whether there exist structural domains as counterpart of the replication (split-) U-domains or, in other words, if the replication (split-) U-domains constitute communities in the DNA interaction network. We sort replication (split-) U-domains in IMR90, GM06990, K562 and H1 ES (using BG02 domains as a surrogate) by size and we associate (split-) U-domain borders with the size of the smallest domain they border. We then query their presence in the interval-community database (Section 6.2.4) using the protocol described in Section 6.3.3. For groups of 50 (split-) U- domains, we compute *(i)* their mean best mutual coverage and *(ii)* the proportion of their matching by a structural community. For groups of 100 (split-) U-domain borders, we compute the proportion that have a matching interval-community borders (Fig. 6.14).

The mean best mutual coverage versus replication domain size curves are rather similar to what we obtained for the TADs in Figure 6.10. For all cell lines, the mean best mutual coverage slowly increases with (split-) U-domain size, from $\sim$47% (IMR90) up to $\sim$68% (H1 ES) for small replication domain sizes ($\lesssim$600 kb) to $\sim$70% (IMR90) up to 82% (K562) for large replication domain size ($\gtrsim$4 Mb) (Fig. 6.14 A). Note that the proportion of (split-) U-domains that have a counterpart in the interval-community database is rather low (10-30%) for small replication domains but reaches significant percentages for the largest replication domains (40% in IMR90 up to 65% in K562) (Fig. 6.14 B). The proportion of matching borders is almost constant: regardless of the replication domain size, between 40% to 50% of the borders are recovered as an interval-community border when the proportion expected by chance is 35% (Fig. 6.14 C). It thus appears that if the largest (split-) U-domains present some convincing concordance with interval-communities (Fig. 6.14 B),

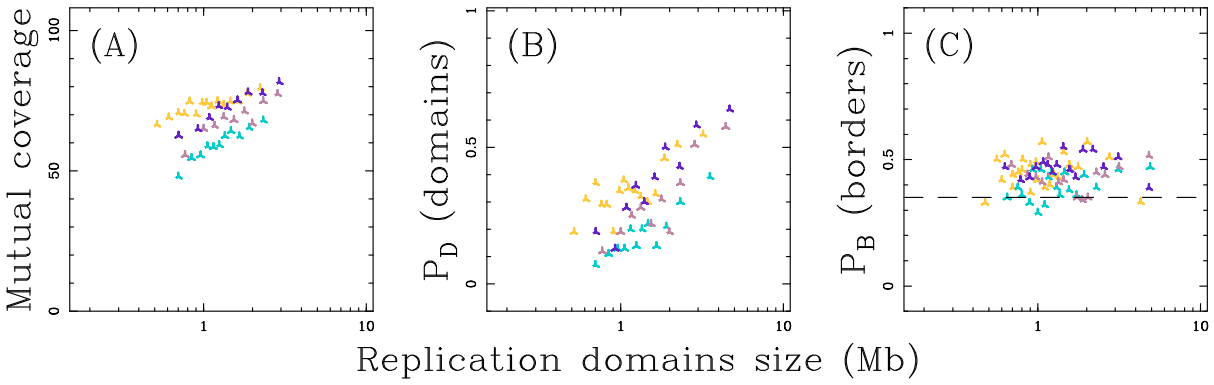


Figure 6.14. Are replication domains structural communities? Same as in Fig. 6.10 for replication domains in different cell lines: H1 ES (yellow), IMR90 (blue), GM06990 (pink) and K562 (purple).

the relationship between replication (split-) U-domains and interval-communities is more complex for smaller domains. As initially reported in *Drosophila* [20] and more recently in human using higher resolution (kb) Hi-C data [244], fine scale structural domains correspond to regions of homogeneous chromatin status. Given the open chromatin structure ($\pm$ 200 kb) around replication domain borders [44, 51] and the gradient of chromatin state from replication domain borders to center [53, 54], replication (split-) U-domais might in fact be related to structural domains in the following way. Replication domain borders would sit inside but close to the borders of an open chromatin structural- domain explaining their association to interval-community borders slightly above the random expectation (Fig. 6.14 C). Replication domain central regions would correspond to heterochromatin structural domains covering more than 80% of the domain length for the largest (split-) U-domain only, explaining the good concordance of these largest replication domain with interval-communities observed in (Fig. 6.14 B). This prompts us to test the correspondance between chromatin state domains (Section 6.5.2) and interval-communities.

6.5.2 Are chromatin states structural communities?

As mentioned above, recent results suggest that chromatin states correlate with the 3D structural organisation of the genome [20, 244]. Here, we ask whether chromatin state domains are better matched by interval-communities than replication (split-) U-domains. We reproduce the same analysis as in Section 6.5.1, replacing (split-) U-domains by the different chromatin blocks C1+C2 (resp. EC1+ EC2) and C3+C4 (resp. EC3+EC4) (Fig. 6.15).

The analysis of the mean best mutual coverage and proportion of domain recovery provides very similar results for chromatin blocks (Fig. 6.15 A12, A34, B12, B34) as those previously obtained for (split-) U-domains (Fig. 6.14). The mean best mutual coverage decreases to 30%-40% depending on block type and cell line, for the smallest chromatin block sizes ($\sim$200 kb) (Fig. 6.15 A12, A34). The proportion of chromatin blocks of size $\lesssim$ 500 kb with an interval-community counterpart is systematically below 10% (Fig. 6.15 B12, B34). However, for the 4 cell lines and the 2 chromatine block types, the proportion of chromatin block borders that match an interval-border at 100 kb resolution ($\pm$ 1pixel) is rather constant with chromatin block size and takes values significantly larger

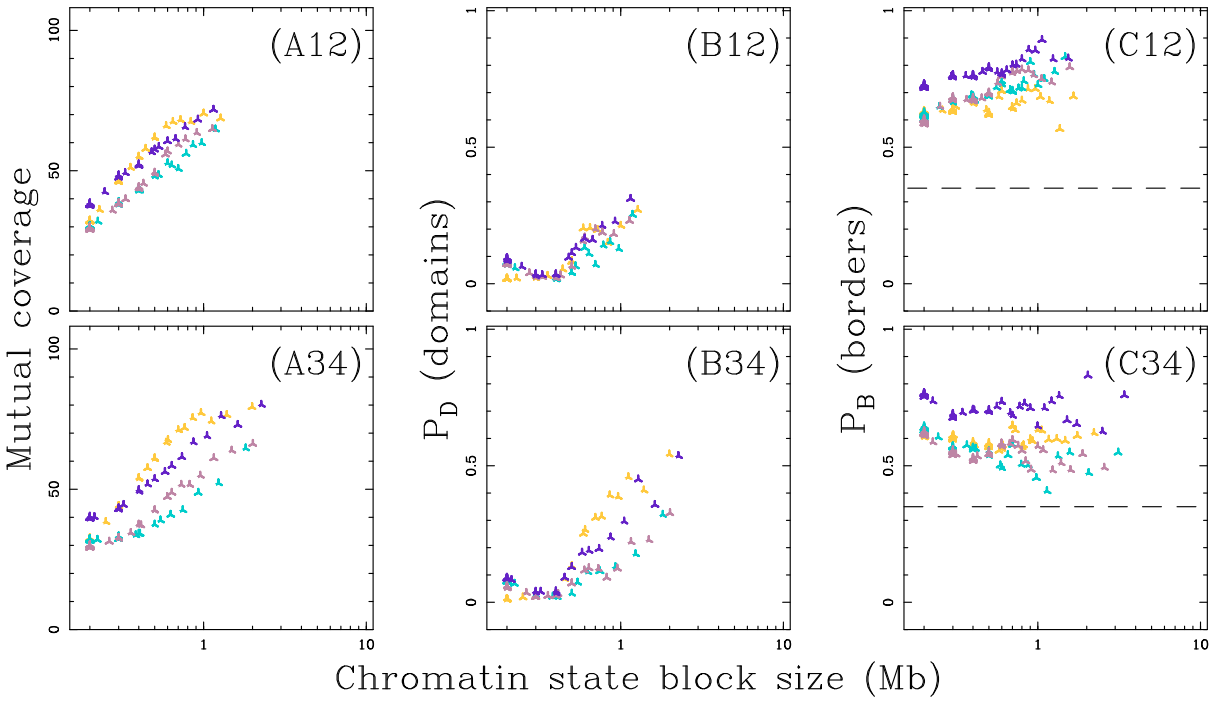


Figure 6.15. Are chromatin states structural communities? Same as in Fig. 6.10 for C1+C2 (resp. EC1+EC2) chromatin blocks (A12, B12, C12) and C3+C4 (resp.EC3+EC4) chromatin blocks (A34, B34, C34): in IMR90 (blue), GM06990 (pink) and K562 (purple) (resp. H1 ES (yellow)).

than the 35% random expectation level*. Border recovery is higher in K562 ($\sim$ 70% in C1+C2 and $\sim$ 80% in C3+C4 blocks) than in the other cell lines ($\sim$ 60% in C1+C2/EC1+EC2 and $\sim$ 60% in C3+C4/EC3+EC4 blocks) (Fig. 6.15 C12, C34). Repeating this analysis separately for C1/EC1, C2/EC2, C3/EC3 and C4/EC4 chromatin blocks, we obtain qualitatively and quantitatively the same results (Supplementary Fig. B.14).

We thus observe a significant localisation of chromatin block borders at interval-community borders concomitant with a rather low recovery of complete chromatin blocks as structural domains. A possible explanation of this situation is that there exists interval-community borders inside chromatin blocks that persists to larger scale than the interval-community borders colocalising with the chromatin block borders. These results provide evidence that as previously suggested [20, 244], there is some correspondence between structural domains and blocks of homogeneous chromatin states. They also question the rule that governs the hierarchical association of neighbouring structural domains at small scales to form the larger scale structural domains. It appears that neighbouring domains of similar chromatin states do not necessarily group to form the larger structural domains. The hierarchical segmentation of the genome in interval-communities performed at higher resolution is likely to provide new insights to this question.

Recovering chromatin block borders as structural community borders is consistant with the recent high resolution Hi-C study [244] suggesting the existence of small scale struc-

*Note that for a chromatin block of size comparable to the resolution used to match borders ($\pm$ 100 kb), border matching simply reflects colocalisation of the chromatin block with an interval-community border.

tural domains ($\sim$180 kb) that correlate with chromatin states.

6.6 Towards a multi-scale description of the genome organisation

The wavelet-based multi-scale community detection method allowed us to detect a set of structural interval-communities in the intra- and inter- chromosomal interactions networks. The set of structural communities was found to form a hierarchical ensemble of genomic interval, where, at large scales chromosomes were identified as communities. Recovering communities as genomic intervals suggest that chromosome folding is essentially due to interactions between "close" neighbours along the genome. This hierarchical description in terms of structural domains covers the initial proposed topologically associated domains [16] without favoring any scale. It allowed us us to compare the multi-scale structural domains between different cell lines. We observed a high level of structural conservation between cell lines up to the largest scales (for example $\sim$ 65% of the $\sim$ 1 Mb interval-communities in differentiated cell lines are also found in H1 ES cell line). The observation that sequence encoded nucleosome excluding energy barriers were enriched at structural-community borders suggests a possible genetic mechanism for this conservation. Segmentation of the genome in terms of structural communities led us to compare functional and structural organisations. If our results question the exact relationship between replication (split-) U-domains and structural domains (Section 6.5.1), they also provide interesting perspectives to further understand the relationship between chromatin states and structural domains across the scales (Section 6.5.2).

This novel methodology is actually independent of the exact genome assembly, avoiding the artefacts of the analysis of Hi-C data coming from cells exhibiting rearrangements relative to the reference genome. The obtained interval-communities respect the 1D order of the chromosomes and hence can be used in a genome reconstruction pipeline from Hi-C data [417]. In fact, Hi-C data has been argued to convey enough information to closing gaps in genome assemblies [29]. In this study, we mainly detailed the analysis of the intrachromosomic interaction networks, and only briefly discussed the analysis of the complete genome interaction network. It will be interesting to analyse structural communities obtained with the whole genome interaction network and to look at how the chromosomes are grouped together and whether we can detect the chromosomal rearrangements observed in cancer cells [418] for example.

On a different perspective, one can wonder how to link the communities to the dichotomic view and the segmentation into A/B compartments suggested in pioneering Hi-C study with distance normalised Hi-C intra-chromosomic data (Section 2.3.3) [14]. When applying the wavelet-based community detection method on the interaction network obtained after distance normalisation of the Hi-C intra-chromosomal matrix in K562, we succeed in grouping the nodes into 2 communities alternating along the chromosomes rather than interval-communities (data not shown). The A/B compartments were defined using the sign of the first eigenvector of the correlation matrix of distance normalised Hi-C data [14]. The correlation between the vector associating to each node its community at a certain scale and this first eigenvector found $\sim$80% in chromosome 17, for example. Further analysis may provide a link between the two types of segmentation. In fact, it has been

recently argued that the A/B mega-domains are likely to be sub-compartmentalised when considering data at higher resolution [244].

☞ Clearly wavelet-based multi-scale community detection provides a computational framework to analyse structural domains along the genome, to correlate these domains with functional domains and to compare genome organisation between different cell lines as well as, in the same cell line, between the chromosome folding at different moments of the cell cycle. Interval-communities form structural domains as the interactions inside them are greater than the interactions in between them. Further analysis are needed to better understand the communities and their borders relative to proteins that are known to be involved in chromatin looping. It will be interesting to investigate chromatin marks around community borders to see which different chromatin marks are involved in the DNA 3D architecture. Wavelet-based multi-scale community detection is a robust methodology to detect structural domains. The hierarchy of interval-communities encompass information previously described at single scale only. Using higher resolution map allowing the detection of the elementary structural units, will provide us with a comprehensive view of the multi-scale structural genome organisation.

Chapter 7

Conclusion and Perspectives

In human, high (resp. low) GC, gene-rich (resp. poor), active (resp. inactive) CTRs covers about 25% (resp. 25%) of the genome that are replicated very early (resp. late) by the coordinated and almost synchronous activation of multiple origins more or less randomly spatially distributed [54, 69]. The other half of the human genome is organised in tissue-dependent U-shaped mean replication timing (MRT) domains bordered by "master" replication initiation zones (MaOris) enriched in open and transcriptionally active marks [43, 44, 53, 54, 161]. It was proposed that from these borders initiates a replication wave that further propagates and accelerates towards the domain center via the successive firing of secondary origins, more or less randomly dispersed [69, 70]. This cascade model involves the superposition of specific and efficient initiations at domain borders with random and less efficient initiations elsewhere, in addition to firing stimulated by propagating forks [69, 70].

In this manuscript, investigating what happens in between two successive and very distant MaOris in human, we reported the discovery of MRT split-U-domains, bordered by putative replication origins with similar properties as the MaOris flanking MRT U-domains and harboring a large central region of late replication timing. These split-MRT U-domains are reminiscent of the skew split-N-domains previously described in the germline [359]. What distinguish split-U/N-domains from the U/N-domains is their central regions that, similarly to the late CTRs, are gene deserts of low and constant GC content. This demonstrates that when the two MaOris are far from each other, a late CTR emerges between them leading to the formation of a split-U-domain. The length of the late CTR in the central region of a split-U-domain increases with the size of the domain. In fact, the MRT increases from the split-U-domain borders up to a certain distance reaching a relatively high value in the central region. Complementarily, in these large replication domains, the MRT derivative decreases to zero over a distance independent of the domain size, suggesting the existence of a limiting time or length scale. Further analysis of replication timing inside these domains showed that, regardless the cell line, the MRT derivative *vs* the MRT follows a "universal" curve. This led us to propose that the cascade model of replication is a general property of the human replication programme that only depends on the time left till the end of the S-phase.

Replication split-U and U-domains cover altogether 60% of the genome and are highly conserved between different cell lines: each cell line shares about half of its domains with at least another cell line. Regardless the domain size, we showed that the replication wave initiating at replication domain borders corresponds to a directional path across the four chromatin states previously described in differentiated cell lines (C1, C2, C3, C4) and in embryonic cell line (EC1, EC2, EC3, EC4), from C1/EC1 at the replication domain borders followed by C2/EC2, C3/EC3 and C4/EC4 at the center. The novel split-U-domains enlighten the striking difference between the epigenetic coating in the large late replicating central region of these replication domains in differentiated cell lines. Blood cells (GM06990) do not endure mechanical constraints whereas fibroblast structural cells

(IMR90) belong to a mechanically constrained tissue. Concomitantly, we observed that split-U-domain central regions in GM06990 are characterised by chromatin state C3 corresponding to the absence of all chromatin marks whereas in IMR90 these regions are characterised by the chromatin state C4 corresponding to HP1 heterochromatin associated with the nuclear lamina. It is thus possible that the nature of chromatin state in these regions is a response to the mechanical properties of their environment.

When concentrating on the MaOris at replication domain borders, we saw that each cell line shares about 80% of its borders with at least another cell line. These borders were found to be in a GC rich, open chromatin region. Furthermore, besides their role in regulating nuclear functions, these borders likely prevent cross-talk between Hi-C topological domains. This observation was strengthened by the enrichment of MaOris in CTCF which besides its insulating property contributes to chromatin 3D folding. Using a graph theoretical approach, we were able to identify the MaOris as hubs in the chromatin interaction network (local maxima of betweenness centrality). In addition, using a multi-scale community detection method to analyse chromatin Hi-C interaction networks, we identified a hierarchy of structural domains. When projected along the genome, these structural-communities reduce to genomic intervals suggesting that chromatin folding is mainly driven by interactions between close neighbours. Interestingly, TADs that are described as the reference physical units of metazoan chromosomes characterised by high intra-domains contact frequencies with conserved borders enriched in insulator protein CTCF [16] were found in majority in our database of interval-communities. Comparison of interval-communities between different cell lines showed that these structural units are highly conserved, especially for larger community sizes ($\sim$ 2 Mb). However, at smaller sizes, pluripotent H1 ES showed an excess in interval-communities of sizes 500 kb-1.5 Mb, as compared to differentiated cell lines. This is consistent with H1 ES presenting smaller replication domains [44]. This suggests that cell differentiation could be responsible of the merging of small structural units in a scenario similar to the one observed for replication domains [38, 147, 148, 364].

A recent Hi-C experimental study at much higher (kb) resolution has provided some refined partitioning of the human genome by TADs of mean size $\sim$180 kb [244] much closer to the estimate $\sim$100-kb previously reported in *Drosophila* [20]. Interestingly, as in *Drosophila*, these refined TADs seem to have some specific epigenetic chromatin identity that can change dramatically their functional identity in different cell types [20, 244, 379]. Within the limit of our resolution (100 kb), we were able to identify chromatin block borders as structural-community borders. Our database of interval-communities consitute a hierarchical partitioning of the genome. The fact that we recovered the chromatin state block borders suggests that functional units hierarchically associate together. The wavelet-based community detection provides us with a tool to address this idea and investigate further the existence of some underlying rules for the association of structural/functional domains across scales.

The analysis of Hi-C data in the other half of the genome (not covered by replication U-domains) has provided compelling evidence for the existence of a 3D compartmentalisation of the genome in differentiated human cell types. Active early CTRs display long-distance interactions (0.7 Mb $\lesssim s \lesssim$ 7 Mb) similar to the ones predicted by the 3D "equilibrium" globule model [368, 371–373] as an indication of their central positioning

in the nuclear interior. Inactive late CTRs display significantly different long-range interactions similar to the ones predicted by the 2D "equilibrium" globule model strongly suggesting some segregation and confining of these lamina-associated heterochromatin domains to the nucleus periphery [17, 392–396]. The Hi-C interactions observed inside and in between (split-) U-domains confirm the existence of some radial nuclear organisation with replication waves initiating from master initiation zones at the nucleus center and further propagating towards a more peripheral heterochromatin positioning at the nuclear membrane. This provides a very attractive understanding of the experimental observation that the spatial distribution of replication foci changes over the course of the S-phase from a central to a more peripheral positioning in the cell nucleus [5–8, 228, 369, 390, 401, 402].

This 3D nuclear chromatin organisation differs between tissues and cell types as the signature of the chromatin folding induced by the self-interaction between chromatin states that promote physical bridging between distal elements, *e.g.*, via the specific interactions of some structural proteins. Thus, in the K562 cell line, the highly active early replicating euchromatin (100-kb) loci in early CTRs and in the master replication initiation zones at (split-) U-domain borders were shown to be the main "hubs" in the chromatin interaction network [44, 408]. The observed enrichment of these loci in CTCF strongly suggests that CTCF is a key factor underlying long-distance intra- and inter chromosomal interactions in this cell line [18, 208–212]. In IMR90 cell line, as the signature of the important spreading of the HP1-associated heterochromatin, the main "hubs" in the long-range chromatin interaction network are instead the inactive late replicating heterochromatin loci in late CTRs and at the center of (split-) U-domains. This is an indication that the structural proteins that regulate the anchoring of the Lamina B1 heterochromatin to the nuclear envelop [407] are likely determinant factors in the long-range interactions underlying the high-order chromatin architecture in IMR90. Specific properties of the H1 ES cycle such as a high proliferation rate and a shortened G1 phase that are necessary for self-renewal and the maintenance of pluripotency [419, 420], could explain the differences observed between chromatin landscapes, gene expression and MRT profiles in pluripotent embryonic stem cells and sommatic cells [197, 198, 201, 207]. In mammals, tens of thousand replication origins are prepared in G1-phase which is more than actively needed in S-phase [69, 262, 421]. Replicon size, which is dictated by the spacing between active origins, was shown to correlate to the length of chromatin loops [265] and to be smaller in H1 ES than in defferentiated cells [254], as confirmed by the smaller characteristic size of (split-) U-domains in H1 ES than in somatic cell types [44]. The shorter G1-phase and cell cycle duration may thus explain the highly dynamic plastic chromatin in pluripotent cells as a lack of time for trancrisptionally inactive heterochromatin to establish [197, 198, 201, 207]. This absence of genome compartmentalisation in pluripotent cells was confirmed by Hi-C data that revealed that the pluripotent chromatin architecture statistically ressembles to the one predicted by the 3D "equilibrium" globule model regardless the heteropolymer (epigenetic) nature of the chromatin fiber. As enriched in CTCF and pluripotent transcription factors NANOG and OCT4, that were recently shown to contribute to the overall folding of embryonic stem cells genome via specific long-range contacts [361, 362], the master replication initiation zones at MRT U-domain borders appear to be fundamental determinants of pluripotency maintenance. In particular they are at the heart of the so-called consolidation phenomenon [38, 147, 148, 364] corresponding to early to late transitions from embryonic stem cells to differentiated cells coinciding with the emergence of compact heterochromatin at the nuclear periphery. These results shed a new

light on the role of replication in the epigenetically regulated chromatin reorganisation that underlies the loss of pluripotency and lineage commitment [54].

To summarise, in this thesis manuscript we reported the discovery of novel replication split-U-domains where replication follows a universal origin firing cascade model. From the analysis of Hi-C data, we showed a segregation from a 3D to a 2D equilibrium chromatin organisation in differentiated cell lines that is not present in H1 ES. We used graph theory to analyse the chromatin interaction network. As compared to most complex networks in biology such as metabolic and gene regulation networks, edges in the chromatin interaction network link vertices that belong to a 1D physical object: the DNA heteropolymer. It is thus a rather specific situation where the graph properties are in direct relationship with the physics of the supporting polymer. We used centrality measures to identify key players in the network. We were able to identify structural communities using a wavelet-based community detection method. This allowed us to compare the structure between different cell lines and to investigate structure/function relationships. The main advantage of this method is that it does not depend on the exact genome assembly. The interval-communities were found to respect the underlying 1D structure allowing in principle to (re)construct genome from Hi-C data. Further analysis of chromatin marks, around interval-community borders will help to characterise them and understand better their nuclear functions.

We also applied our methodology to Hi-C data at 2 different moments of the cell cycle, one in G1 with domain-like organisation and the other one in mitosis with no apparent structure. Interestingly, we were able to capture interval-communities in the first case while no structure emerged in the second. This demonstrates that in the limit of the availability of the data we are in a position to follow the structural domains evolution during the cell cycle. In that perspective, it could be interesting to apply community detection using a dynamic graph approach. Dynamic graphs are graphs whose structure evolves in time. A recent study [255] showed distinct structural transitions of TADs with hormone-induced gene regulation. In that context, dynamical graph theoretical concepts look very promising to follow and to understand the evolution in time of structural units.

Some of our analyses were limited by the Hi-C data resolution. It will be interesting for example to detect interval-communities at higher resolution. In fact, as suggested by a recent study, TADs are subdivided into smaller domains of same epigenetic coating [244]. Looking at communities in these new data can provide better quantification of the chromatin state blocks as epigenetic communities. Moreover, as our methodology allows to detect hierarchical structural domains, it constitutes a way to understand how structural domains hierarchically co-associate.

In biology, it is always instructive to compare different cell lines. In that sense, it would be interesting to reproduce our analysis to more cell lines. From a more experimental point of view, analysis of two sets of GM06990 Hi-C data (obtained with two different restriction enzymes in the Hi-C protocol) presented some divergences relative to other considered cell lines. It could be interesting to see how those results compare to GM12878 Hi-C data from [242].

So far, genome-wide methodologies require thousands to millions of cells and thus only

provide population averages. Accordingly, our understanding and modelling of chromatin-mediated regulation of nuclear functions are simply a mean field view of the dynamic and stochastic nature of chromosomal structures. Although the genome is faithfully replicated each cell cycle, the epigenome coating of TADs could be in part variable between daughter cells [13, 378, 379, 386]. An important challenge for future research will be to devise single-cell experimental strategies to move from probabilistic chromosome conformations averaged over millions of cells towards the determination of chromosome and genome structure in individual cells. Very promising pioneering single-cell Hi-C [422] and nuclear lamina interactions [407, 423] experiments have confirmed that intra- and inter-chromosome contact structures are highly variable between individual cells. In particular, each cell cycle, a different subset of LADs contact the nuclear lamina in a rather stochastic manner and the chromosomes adopt different configurations. This emerging highly dynamic view of chromosomal organisation looks very attractive as far as progressing in our understanding of cell fate decisions of individual cells in different organisms.

Meta-analyses [385, 424] of replication timing profiles [151], Hi-C data [14] and somatic copy-number alterations (SCNA) observed in cancer samples from diverse cancer types [425] showed that SCNAs tend to fuse genomic regions that, prior to the rearrangement, spatially co-localised within the nucleus and have similar replication timing. This illustrates that combined structure/function characterisations of nuclear processes are likely required to fully investigate cancer progression. Thus, experimental protocols and computational tools allowing to fully apprehend this relationship will provide a framework for further studies in different cell types, in both health and disease.

Part III

Annexes

Appendix A

The continuous wavelet transform and applications for analysing genomic data

This appendix is dedicated to present the continuous wavelet transform (WT) and to summarize the main steps of the methodologies developed to extract objective information from strand compositional asymmetry (skew) profiles (Chapter 2, Section 2.1.2) and DNA replication timing profiles (Chapter 2, Section 2.1.4.3).

A.1 The continuous wavelet transform

The WT is a space-scale analysis which consists in expanding signals in terms of wavelets which are constructed from a single function, the "analysing wavelet" ψ, by means of translations and dilations. The WT of a real-valued function f is defined as [426–428]:

$$T_{\psi}^{(\alpha)}[f](x_0, s) = s^{\alpha} \int_{-\infty}^{+\infty} f(x)\psi(\frac{x_0 - x}{s})dx \ , \tag{A.1}$$

where x_0 and s (> 0) are the space and scale parameters respectively and α is the normalisation exponent. We assume $\alpha = -1$ when the value of α is not specified; it is the most convenient choice when using the WT for multifractal analysis. The analysing wavelet ψ is generally chosen to be well localized in both space and frequency. Usually ψ is required to be of zero mean for the WT to be invertible. It is in fact possible to further require ψ to be orthogonal to low-order polynomials [300]. The number of vanishing moment of a wavelet ψ is the largest integer n_{ψ} such that:

$$\int_{-\infty}^{+\infty} x^m \psi(x)dx \ , \quad 0 \leq m < n_{\psi} \ . \tag{A.2}$$

A.2 Defining robust scale-derivatives using wavelets

A commonly used class of analysing wavelets is defined by the successive derivatives of the Gaussian function:

$$g^{(n)}(x) = \frac{d^n}{dx^n} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \ , \tag{A.3}$$

for which $n_{\psi} = n$ and more specifically $g^{(1)}$ and $g^{(2)}$ that are illustrated in Figure A.1(a,b). Interestingly, the WT of a signal f with $g^{(n)}$ (Equation (A.3)) takes the following simple expression:

$$T_{g^{(n)}}^{(-(n+1))}[f](x, s) = \frac{d^n}{dx^n} \left(g_s^{(0)} * f \right)(x). \tag{A.4}$$

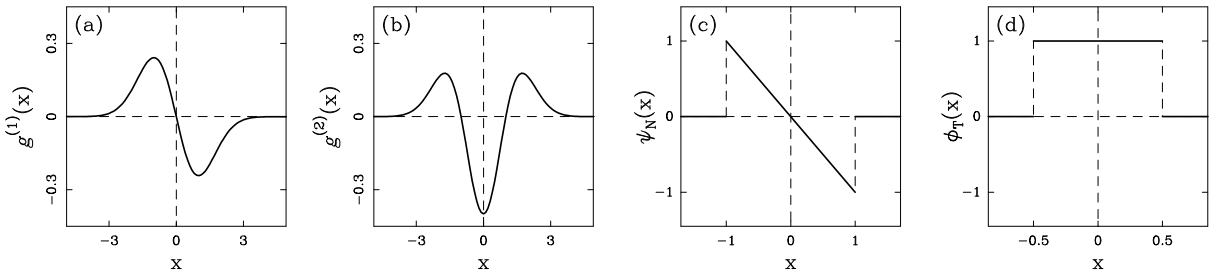


Figure A.1. Set of analysing wavelets $\psi(x)$ that can be used in Equation (A.1): (a) $g^{(1)}$ and (b) $g^{(2)}$ as defined in Equation (A.3); (c) ψ_N as defined in Equation (A.5) and used to detect replication N-domains (Fig. A.2). (d) Box function ϕ_T modelling step-like skew profiles induced by transcription.

Equation (A.4) shows that the WT computed at scale s with $g^{(n)}$ and normalisation exponent $\alpha = -(n+1)$ simply reduces to the n^{th} derivative of the signal f smoothed by a dilated version $g_s^{(0)}(x) = \frac{1}{s} g^{(0)}(x/s)$ of the Gaussian function. Note that the norm of $g_s^{(0)}$ is 1, so that the convolution $g_s^{(0)} * f$ is a moving average of the f profile. Equation (A.4) defines a robust framework for the estimation of noisy experimental profile variations over different length scales. Indeed, the derivative of a noisy profile is not defined and, correspondingly, the naive derivative of a noisy experimental profile based on the numerical difference between two successive samples is ill-defined and extremely unstable.

A.3 Delineating N-shaped replication domains using wavelets: Disentangling transcription- and replication- associated strand asymmetries

The WT can be used to perform multi-scale pattern recognition in the (space, scale) half-plane. Indeed, the wavelet coefficient $T_\psi[S](x_0, a)$ quantifies to which extent, around position x_0 over a distance a, S has a similar shape as the analysing wavelet ψ (Equation (A.1)). Hence, the first step to detect putative replication N-domains (Section 2.1.2) consists in computing the WT of strand compositional asymmetry profiles S using as analysing wavelet [45, 46, 429]:

$$\psi_N(x) = -x\chi_{[-1,1]}(x)\ , \tag{A.5}$$

where $\chi_{[-1,1]}$ is the characteristic function of the interval $[-1, 1]$ (Fig. A.1(c)). $\psi_N(x)$, called N-let because of its shape that looks like the letter N, is adapted to perform an objective segmentation of skew profiles into N-shaped domains. As illustrated in Figure A.2, the space-scale location of significant maxima values in the 2D WT decomposition (red areas in Fig. A.2(b)) indicates the middle position (spatial location) of candidate replication domains whose size is given by the scale location. In order to avoid false positives, we then check that there does exist a well-defined upward jump at each domain extremity. These jumps appear in Fig. A.2(b) as blue cone-shape areas pointing at small scale to the skew jump positions where are located the putative replication origins. Note that because the analysing wavelet is of zero mean (Equation (A.2)), the WT decomposition is insensitive to (global) asymmetry offset. However, the overall observed skew S also contains some contribution induced by transcription that generates step-like blocks (Fig. A.1(d))

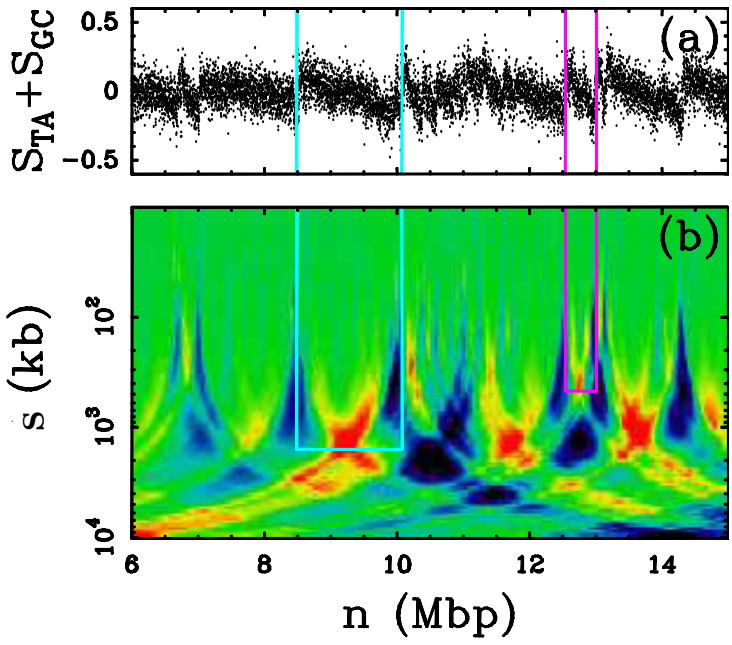


Figure A.2. Multi-scale pattern recognition of replication N-shaped skew profiles. (a) Skew profile S of a 9 Mb long repeat-masked fragment of human chromosome 21. (b) WT of S using ψ_N (Fig. A.1(c)); $T_{\psi_N}[S](n,s)$ is color-coded from dark-blue (min; negative values) to red (max; positive values) through green (null values). Light-blue and purple lines illustrate the detection of two replication domains of significantly different sizes. Note that in (b), blue cone-shape areas signing upward jumps point at small scale (top) towards the putative replication origins and that the vertical positions of the WT maxima (red areas) corresponding to the two indicated replication domains match the distance between the putative replication origins (1.6 Mb and 470 kb respectively).

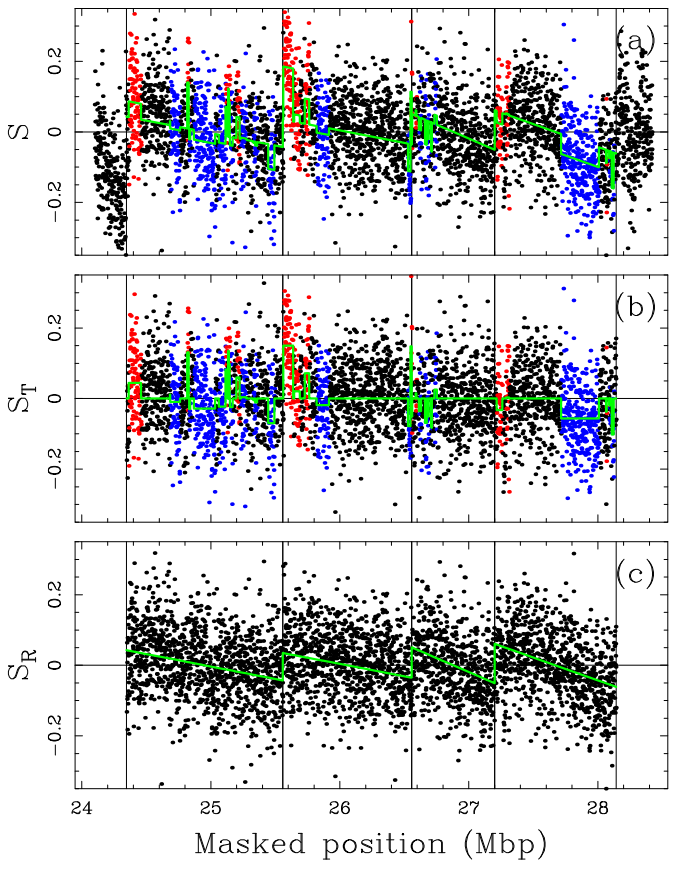


Figure A.3. (a) Skew profile S of a 4.3 Mb long repeat-masked fragment of human chromosome 6 [45]; each point corresponds to a 1 kbp window: red, sense $(+)$ genes; blue, antisense $(-)$ genes; black, intergenic regions (the color was defined by majority rule); the estimated skew profile (Equation (A.6)) is shown in green; vertical lines correspond to the locations of 5 putative replication origins that delimit 4 adjacent domains identified by the wavelet-based methodology. (b) Transcription-associated skew S_T obtained by subtracting the estimated replication-associated profile (green lines in (c)) from the original S profile in (a); the estimated transcription step-like profile (second term on the *rhs* of Equation (A.6)) is shown in green. (c) Replication-associated skew S_R obtained by subtracting the estimated transcription step-like profile (green lines in (b)) from the original S profile in (a); the estimated replication serrated profile (first term in the *rhs* of Equation (A.6)) is shown in green. Modified from [45].

corresponding to (+) and (−) genes [82, 97, 430] as illustrated in Figure A.3. Hence, when superimposing the replication serrated and transcription step-like skew profiles, we get the following theoretical skew profile in a replication domain [45, 46]:

$$S(x') = S_R(x') + S_T(x') = -2\delta \times (x' - 1/2) + \sum_{\text{gene}} c_g \chi_g(x'), \tag{A.6}$$

where position x' within the domain has been rescaled between 0 and 1, $\delta > 0$ is the replication bias, χ_g is the characteristic function for the g^{th} gene (1 when x' points within the gene and 0 elsewhere; see Fig. A.1(d)) and c_g is its transcriptional bias calculated on the forward strand (likely to be positive for (+) genes and negative for (−) genes). The objective is thus to detect human replication domains by delineating, in the noisy S profile obtained at 1 kbp resolution (Fig. A.3(a)), all chromosomal loci where S is well fitted by the theoretical skew profile Equation (A.6). We only retained the domains the most likely to be bordered by putative replication origins, namely those that are delimited by upward jumps corresponding to a transition from a negative S value $< -3\%$ to a positive S value $> +3\%$. Also, for each domain so-identified, we used a least-square fitting procedure to estimate the replication bias δ, and for each gene transcription bias c_g. The resulting χ^2 value was then used to select the candidate domains where the noisy S profile is well described by Equation (A.6). As illustrated in Figure A.3, this method provides a very efficient way of disentangling the step-like transcription skew component (Fig. A.3(b)) from the N-shaped component induced by replication (Fig. A.3(c)).

A.4 Multiscale detection of peaks in replication timing profiles

The simple intuitive idea allowing for effective detection of peaks in a noisy replication timing profile $\langle r(x) \rangle$ is to delineate positions x along the signal that are a local extrema ($\frac{d}{dx}\langle r(x) \rangle \sim 0$) and present a strong curvature ($\frac{d^2}{dx^2}\langle r(x) \rangle \gg 0$) as expected at the tip of a peak symmetrical about a vertical axis*. Within the framework of our mathematical modelling of replication timing profiles (Section 2.1.3), over such loci the average fork polarity is null (Equation (2.11)) and there is an excess of replication origins over termination sites (Equation (2.12)) as expected in a region containing an efficient replication origin. As previously mentioned, we used the WT based framework (Section A.2) to measure signal variation at different scales of observation. When a profile f is the graph of a Brownian motion *i.e.* the increments of f are independent, identically distributed Gaussian variables [431], then the WT coefficients are Gaussian with a standard deviation independent of the scale of analysis when choosing the normalisation exponent $\alpha = -3/2$: $T_{g^{(n)}}^{(-3/2)}[f](x,s) \sim \mathcal{N}(0, \sigma_o)$ [42, 432]. Hence, in order to threshold scale-derivatives in a uniform manner with respect to the fluctuations for a Brownian profile, the first and second order fluctuations of the timing profiles are estimated using this normalisation exponent instead of those prescribed in Equation (A.4). The basic principle of the detection of peaks in the replication timing profiles with the WT is illustrated in Figure A.4. In a first step, we determined (i) the regions of the space-scale half plane candidates to be a local MRT extrema by applying the following thresholding of the WT of $\langle r(x) \rangle$ using $g^{(1)}$:

*It is the common habit to plot replication timing profiles with the time axis oriented downward so that timing peaks are local minima and correspond to positive curvature.

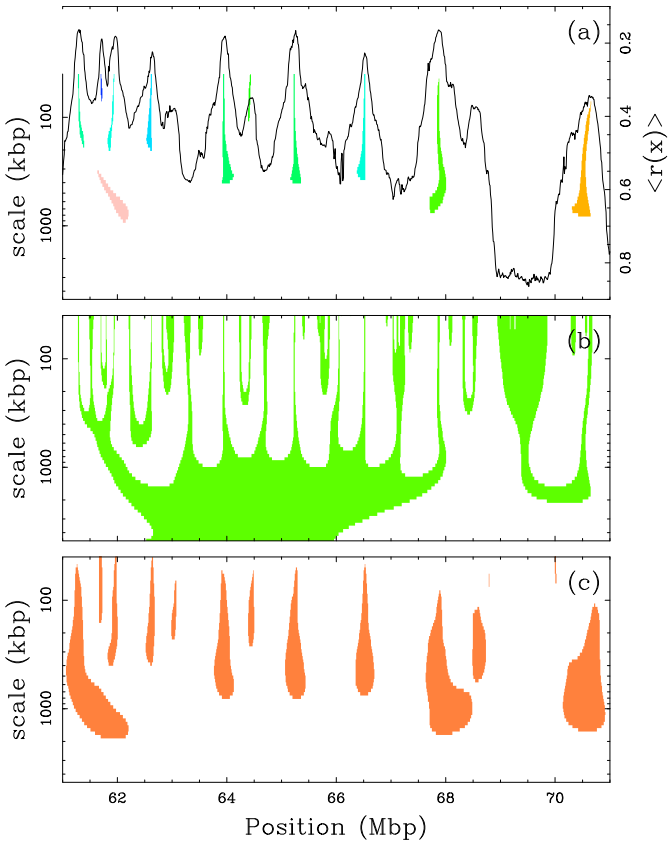


Figure A.4. (a) Mean replication timing profile $\langle r(x) \rangle$ normalised between 0 (start of S phase) and 1 (end of S phase). The color patches are space-scale representations of the detected peaks (see main text). (b) Regions of the space-scale half plane where the timing profile is flat according to: $\left| T_{g^{(1)}}^{(-3/2)}[\langle r \rangle](x,s) \right| < c_1$ with $c_1 = 0.01$. (c) Regions of the space-scale half plane where the timing profile presents a significant negative curvature according to $T_{g^{(2)}}^{(-3/2)}[\langle r \rangle](x,s) > c_2$ with $c_2 = 0.03$. Modified from [161].

$\left| T_{g^{(1)}}^{(-3/2)}[\langle r(x) \rangle](x,s) \right| < c_1$ (Fig. A.4(b)) and (ii) the regions of strong MRT concavity by applying the following thresholding of the WT of f using $g^{(2)}$: $T_{g^{(2)}}^{(-3/2)}[\langle r(x) \rangle](x,s) > c_2$ (Fig. A.4(c)). In a second step, we determined the connected regions of the space-scale half plane where both requirements are fulfilled (color regions in Fig. A.4(a)). Finally, connected regions that have a scale extension (ratio between the region largest and smallest scales) smaller than 1.74 are disregarded in order to guaranty the existence of a well defined peak robust with respect to the scale of observation [43, 47, 161].

A.5 Delineating U-shaped replication timing domains

To detect systematically U-domains along replication timing profiles, we developed a wavelet-based method (Equation A.1) that consists in looking for regions bordered by points that are local maxima of the curvature (corresponding to the location where the U-shape breaks) and presenting a significant negative curvature in their central regions, the hallmark of a parabolic curve (Fig. A.5) [43, 44]. If the speed of replication fork is constant (see Section 2.1.4.2) then these borders correspond to regions presenting an excess of replication origins over termination sites (Equation (2.12)) and the parabolic shape of the central region corresponds to a gradient of replication fork polarity (Equation (2.11)). As discussed in Section A.2, the wavelet transform using the second derivative of the Gaussian constitutes a robust methodology to measure the curvature of a noisy signal view at a certain scale of observation (Equation (A.4)). The convenient normalisation exponent α of the WT (Equation (A.1)) to estimate the threshold on the central curvature c depends on the characteristics of the U-shaped timing motifs. If the depth of the U-shapes is proportional to their width *i.e.*, if they are of the form $x^2/l_{1/2}$, where $l_{1/2}$ is

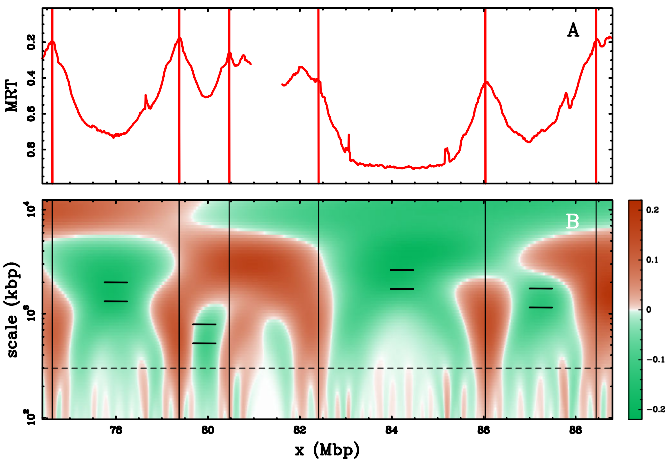


Figure A.5. A multi-scale method to delineate U-domains along replication timing profiles [43, 44]. (a) $\langle r(x)\rangle$ obtained in K562 cell line. (b) Space-scale representation of second-order variations of $\langle r(x)\rangle$; $T^{(-1)}_{g^{(2)}}[\langle r(x)\rangle]$ (Equation (A.1)) values are color coded using green (resp. orange) shades for negative (resp. positive) curvature (note that the timing axis is going downwards). The horizontal dashed line marks scale 300 kb used to detect regions of preferential replication initiation (vertical lines). Pairs of horizontal bars delineate the scale range where strong negative curvature is expected for parabolic U-shaped timing profile. Regions delineated by two successive regions of preferential replication initiation are kept as U-domain if $T^{(-1)}_{g^{(2)}}[\langle r(x)\rangle] \leq -0.03$ at their midpoint for some scale value in this range. Modified from [44].

half the width, then $c \propto 1/l_{1/2}$. If the depth of the U-shapes is constant *i.e.*, if they are of the form $x^2/l^2_{1/2}$, then $c \propto 1/l^2_{1/2}$. For a parabolic shape profile of size $2l_{1/2}$, the scale where extremal curvature is observed using the WT is proportional to $l_{1/2}$. Hence, in order to apply a constant threshold at each scale, the first scenario prescribes the usage of a normalisation exponent $\alpha = -2$, whereas the second scenario requires $\alpha = -1$. We put the emphasis on the detection of the largest U-domains which are constrained by the duration of S phase *i.e.*, we chosed the second scenario which is more permissive than the first one for large U-shapes. The basic principle of the detection of U-domains in the replication timing profiles with the WT is illustrated in Figure A.5. First, candidate U-domain borders are determined at scale 300 kb as the maxima location with a WT value $T^{(-1)}_{g^{(2)}}[\langle r(x)\rangle] \geq 0.15$. Then, regions encompassed between two successive border candidates are accepted as U-domains if they present a sufficiently negative curvature value $T^{(-1)}_{g^{(2)}}[\langle r(x)\rangle] \leq -0.3$ at their mid-point in the scale range $[0.24 \times 2l_{1/2}, 0.36 \times 2l_{1/2}]$. Otherwise, regions were rejected. Indeed, for a parabolic shape profile of finite size $2l_{1/2}$, the scale where extremal curvature is observed using $T^{(-1)}_{g^{(2)}}$ is proportional to $l_{1/2}$ but also depends on the shape of the profile at the border of the regions. The previous scale range has been estimated numerically and corresponds to the situations where both regions flanking the U-domain are either other U-domains (then the extremal value of $T^{(-1)}_{g^{(2)}}$ is observed at scale $0.24 \times 2l_{1/2}$) or flat timing profile regions (then the extremal value of $T^{(-1)}_{g^{(2)}}$ is observed at scale $0.36 \times 2l_{1/2}$).

Appendix B

Supplementary figures

B.1 Supplementary Figures for Chapter 4

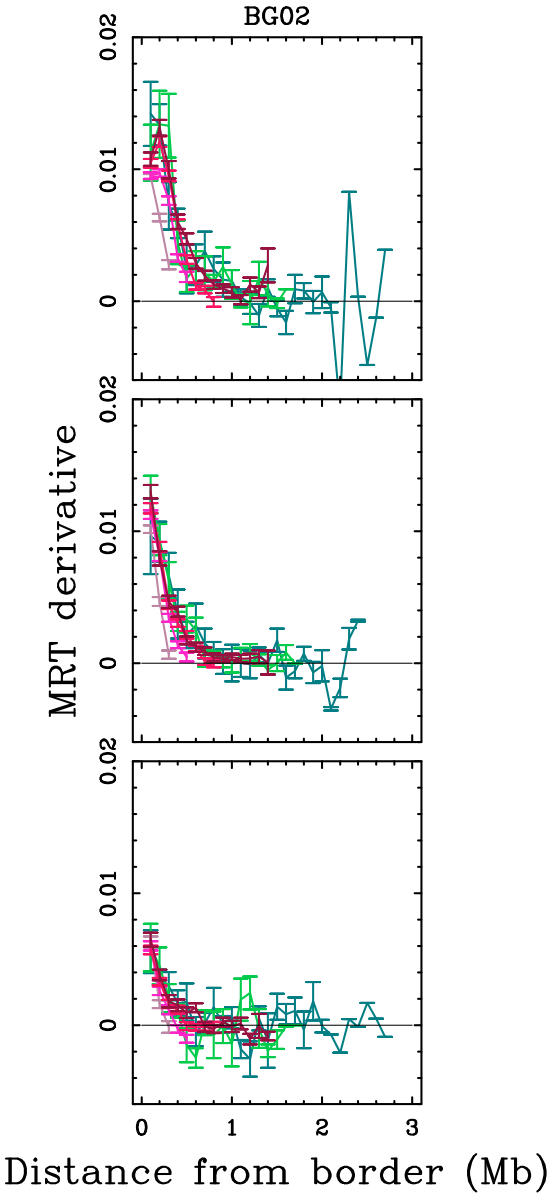


Figure B.1. Mean MRT derivative. Mean derivative of the MRT *vs* the distance from the closest domain border in BG02 for the different domain size categories (see Fig. 4.7) and the different timing categories: early timing (MRT<0.25) top pannel, mid timing (0.25 $\leq$ MRT <0.5) center pannel, and late timing (MRT >0.5) bottom pannel. As observed for the large domains (Fig. 4.7), the curves corresponding to late replicating borders seems to flatten before the one of early replicating borders for all the domain size categories larger than 1.8 Mb.

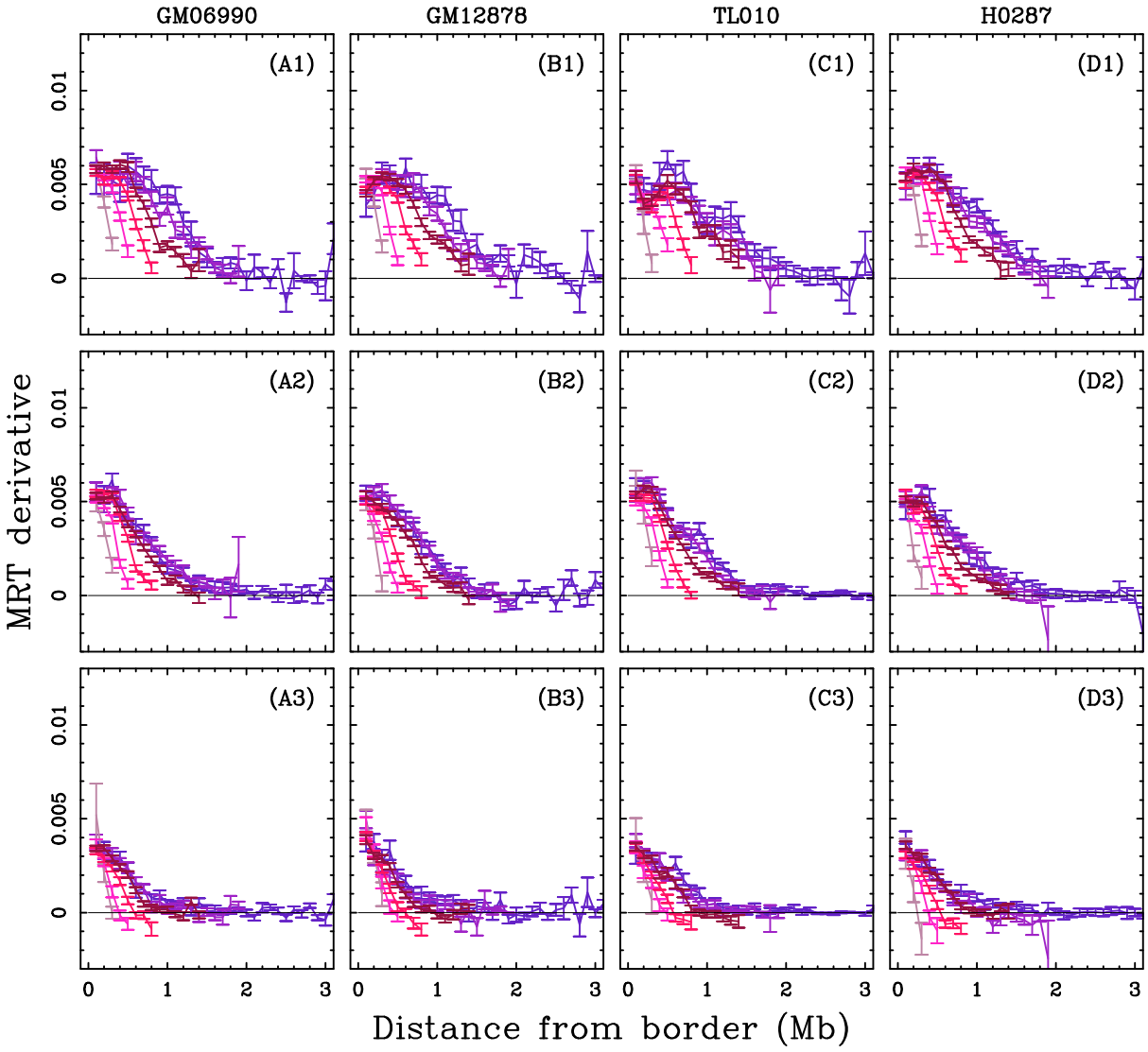


Figure B.2. Same as in Fig. B.1 for lymphoblasts cell lines. Consistently, the curves in A3-D3 corresponding to late replicating borders flatten before the ones of mid replicating borders (resp. early replicating borders) in A2-D2 (resp. A1-D1).

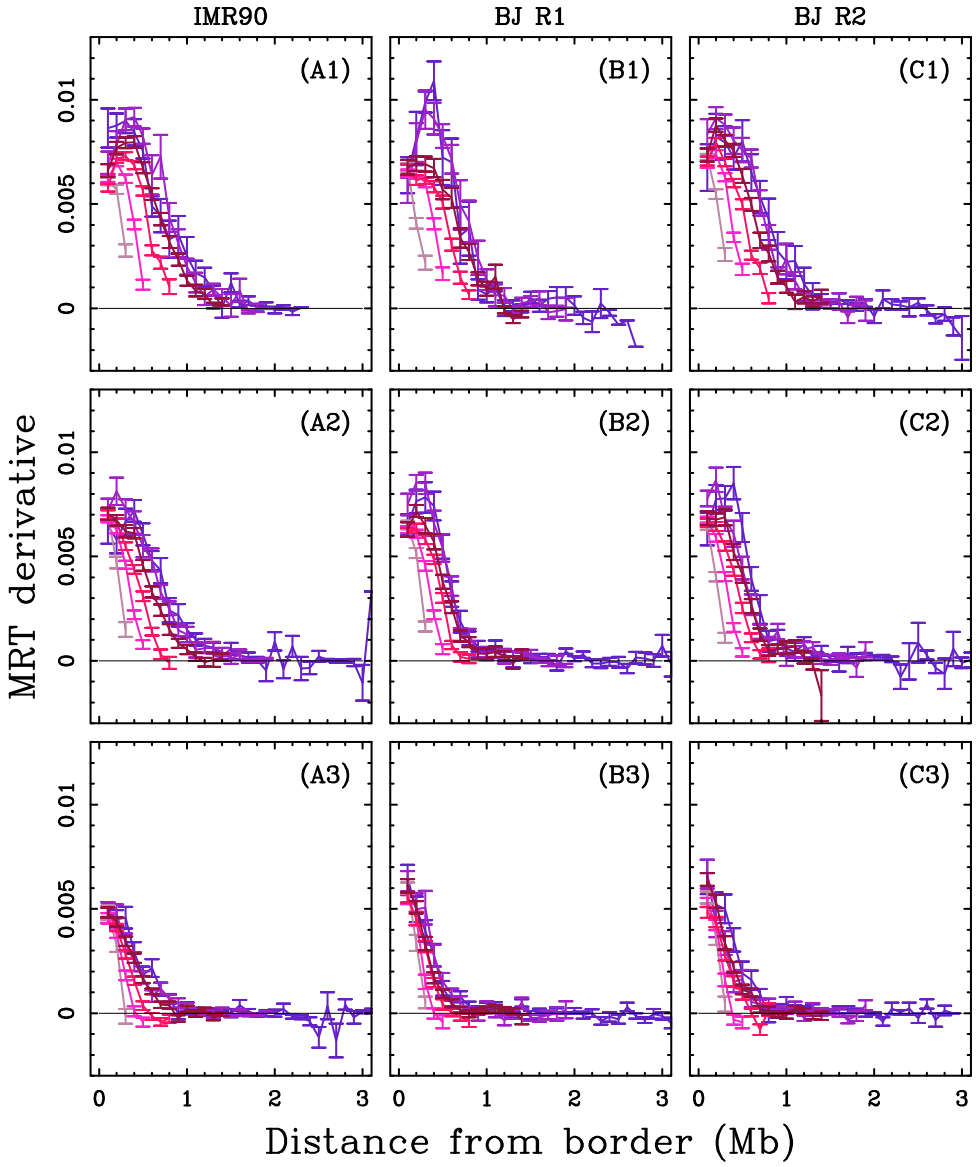


Figure B.3. Same as in Fig. B.1 for fibroblasts cell lines.

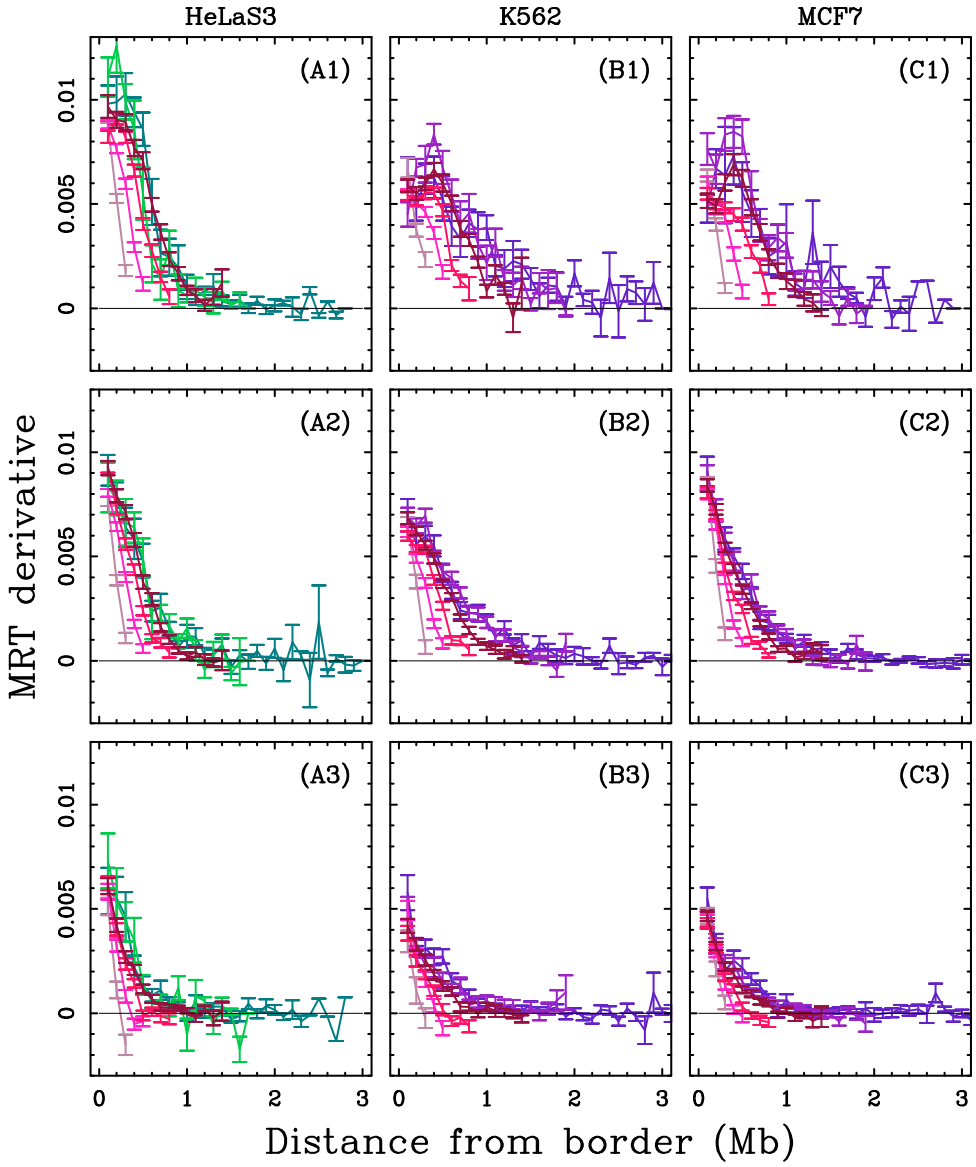


Figure B.4. Same as in Fig. B.1 for cancer cell lines.

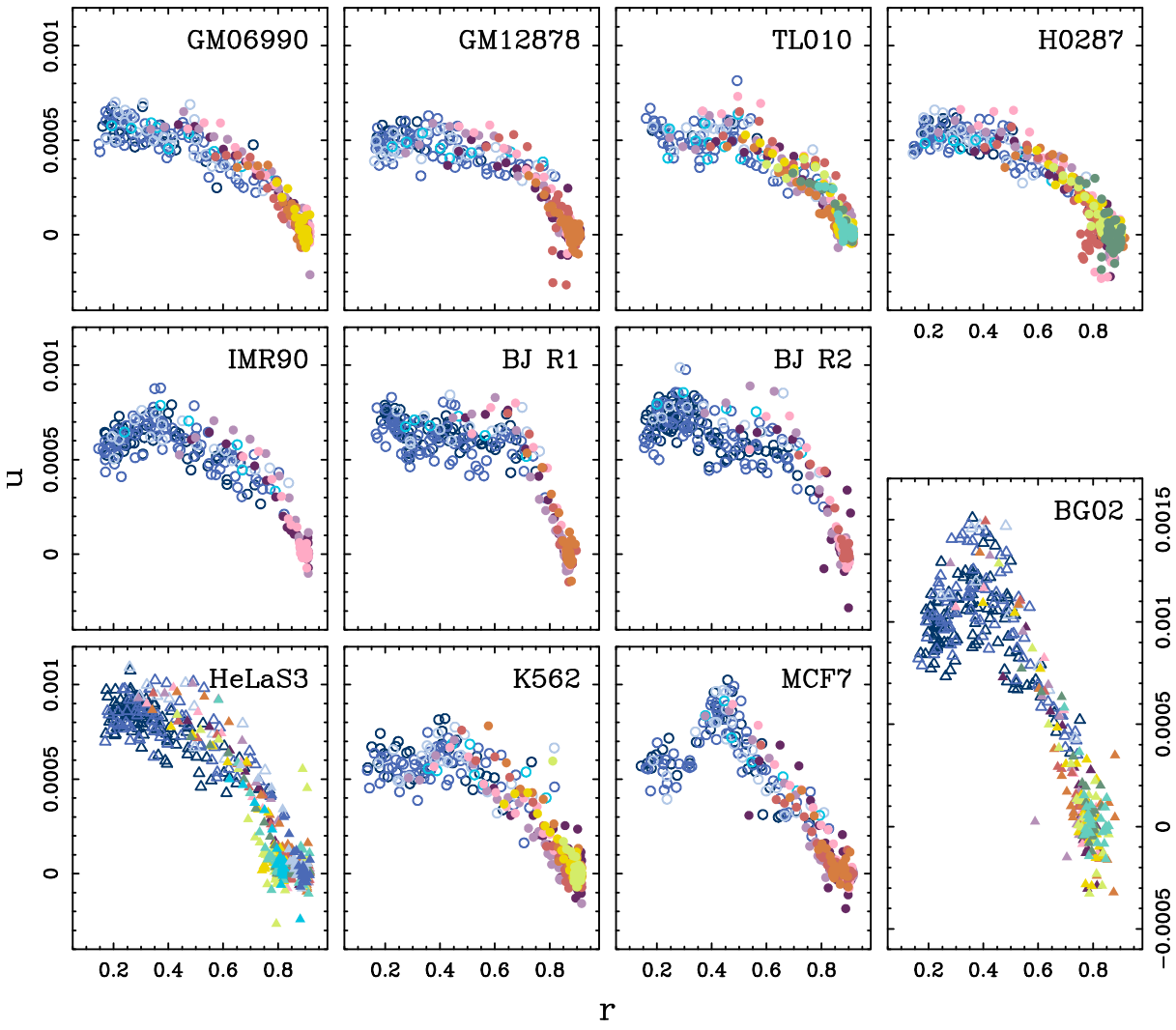


Figure B.5. MRT derivative as a function of the MRT. Same as in Fig. 4.12 with the original normalisation.

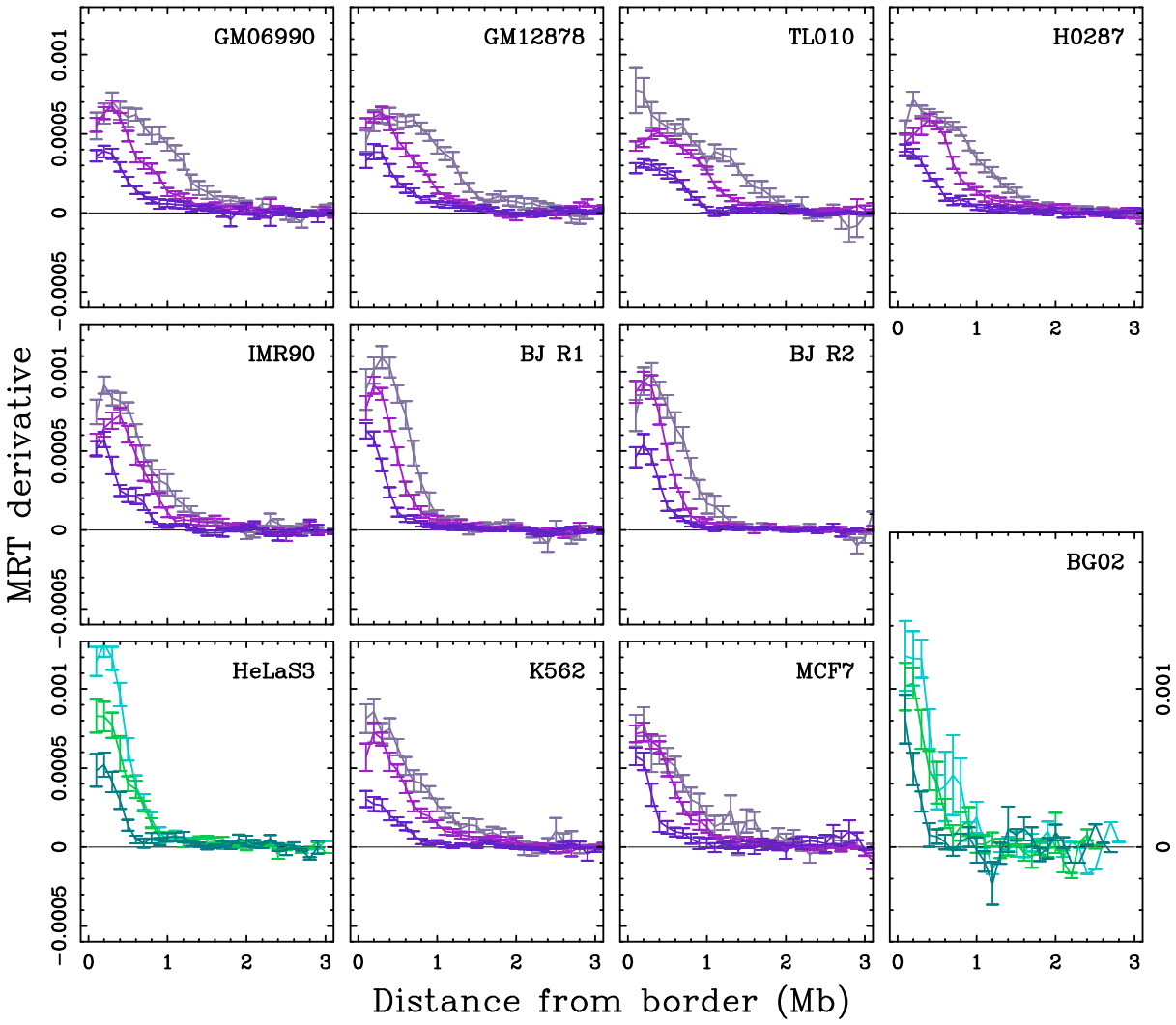


Figure B.6. Mean MRT derivative. Same as Figure 4.9 with the iteratively normalised data.

B.2 Supplementary Figures for Chapter 5

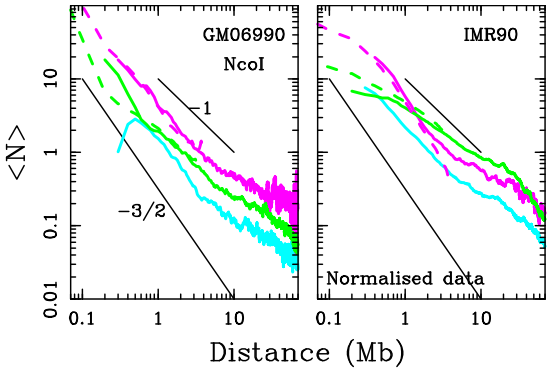


Figure B.7. Same as Fig. 5.2 for GM06990 using NcoI enzyme and for IMR90 normalised data.

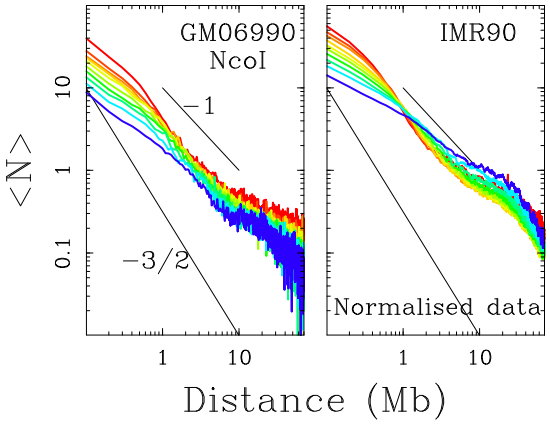


Figure B.8. Same as Fig. 5.3 for GM06990 using NcoI enzyme and for IMR90 normalised data.

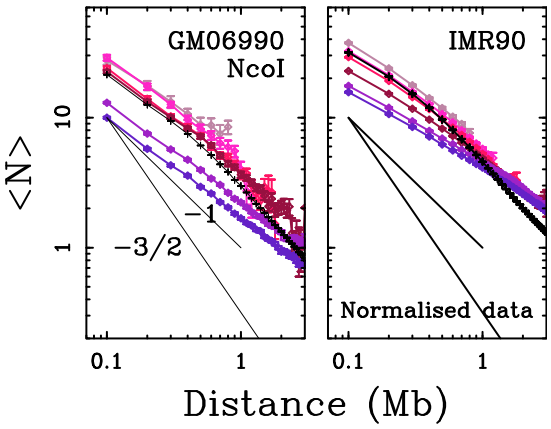


Figure B.9. Same as Fig. 5.4 for GM06990 using NcoI enzyme and for IMR90 normalised data.

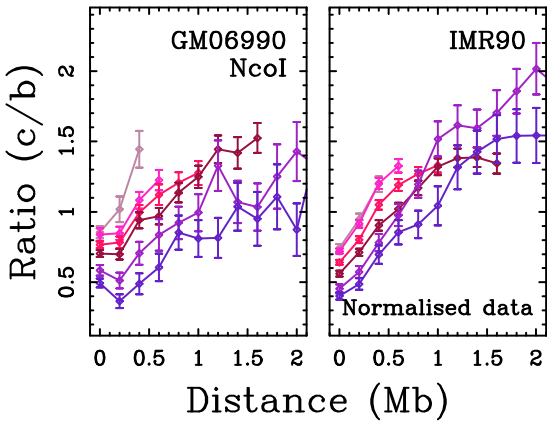


Figure B.10. Same as Fig. 5.5 for GM06990 using NcoI enzyme and for IMR90 normalised data.

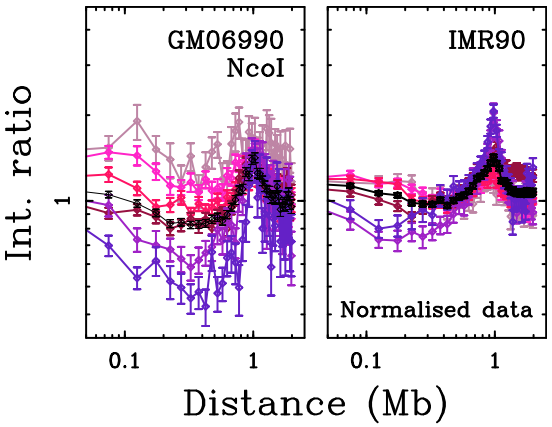


Figure B.11. Same as Fig. 5.6 for GM06990 using NcoI enzyme and for IMR90 normalised data.

B.3 Supplementary Figures for Chapter 6

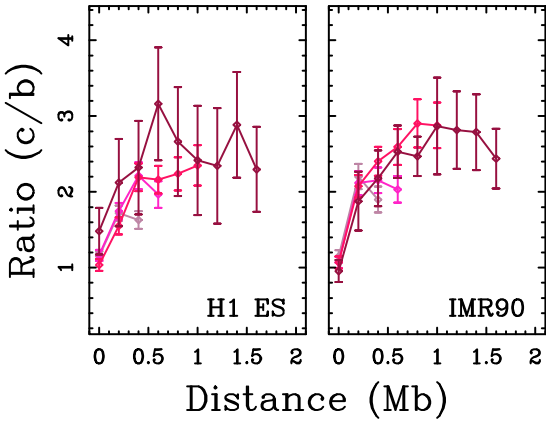


Figure B.12. Same as Figure. 6.6 for the TADs grouped in different size categories: 0.3 $\leq L <$ 0.6 Mb (light pink), 0.6 $\leq L <$ 1 Mb (pink), 1 $\leq L <$ 2 Mb (magenta), 2 $\leq L <$ 3 Mb (dark pink).

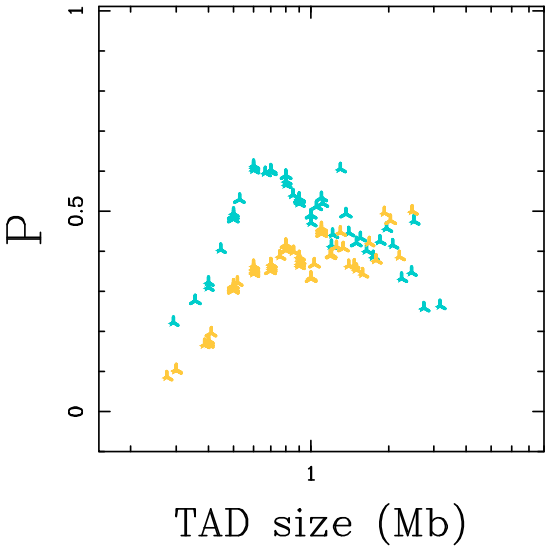


Figure B.13. Same as Figure 6.12 for the comparison of the TAD sets in H1 ES and IMR90 (blue) where IMR90 set is the query set and H1 ES the reference set. (yellow) corresponds to the reversed analysis.

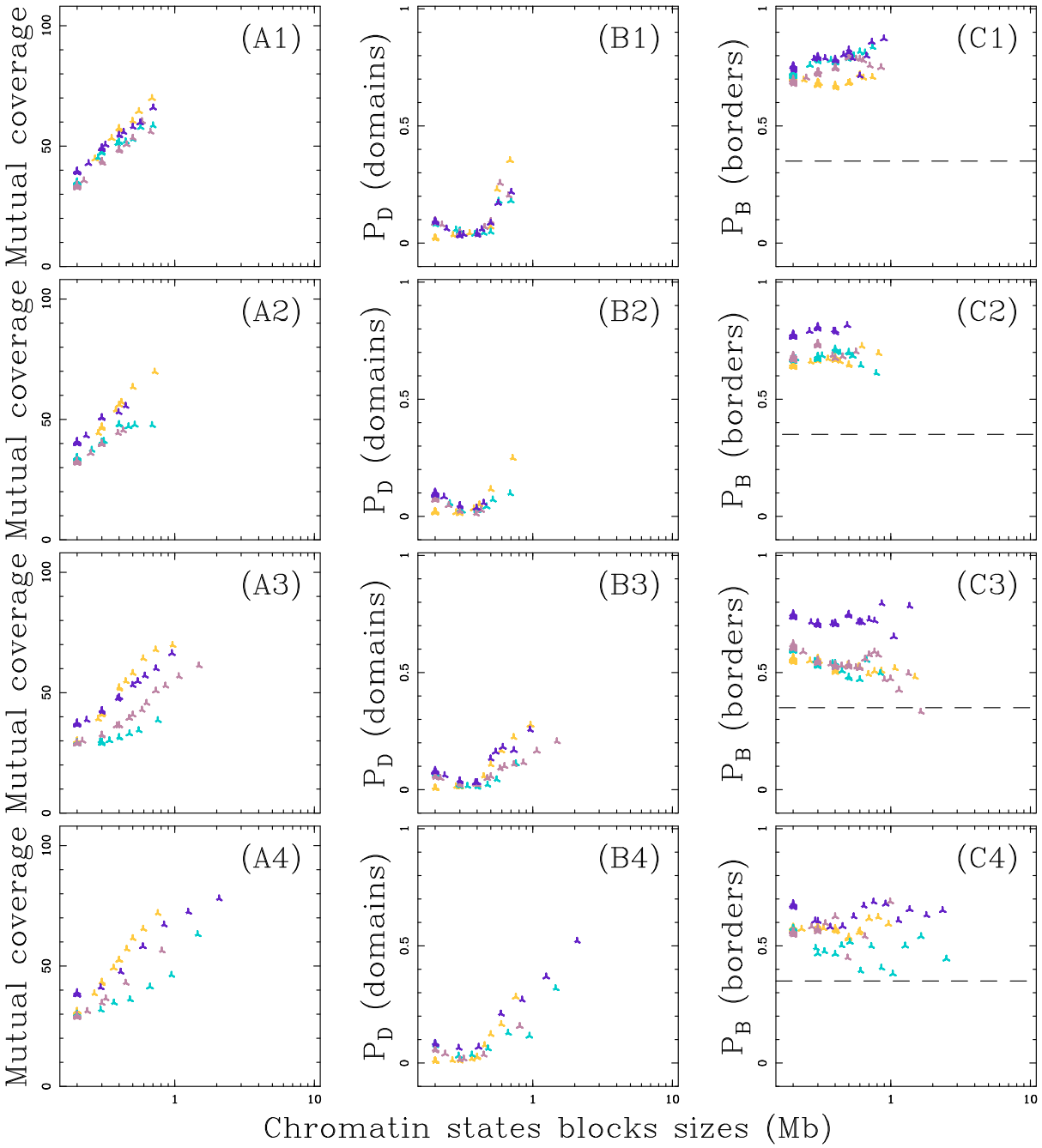


Figure B.14. Are chromatin states structural communities? Same as in Fig. 6.15 in C1/EC1 (first row), C2/EC2 (second row), C3/EC3 (third row) and C4/EC4 (forth row) chromatin blocks.

Scientific communications

Articles

- **R. E. Boulos**, A. Arneodo, P. Jensen & B.Audit, *Revealing long-range interconnected hubs in human chromatin interaction data using graph theory*, Physical Review Letters, 111, 118102 (2013).
- **R. E. Boulos**, H. Julienne, A. Baker, C. -L. Chen, N. Petryk, M. Kahli, Y. d'Aubenton-Carafa, A. Goldar, P. Jensen, O. Hyrien, C. Thermes, A. Arneodo & B.Audit, *From the chromatin interaction network to the organization of the human genome into replication N/U-domains*, New Journal of Physics, 16, 115014 (2014).
- L. Zaghloul, G. Drillon, **R. E. Boulos**, F. Argoul, C. Thermes, A. Arneodo & B. Audit, *Large replication skew domains delimit GC-poor gene deserts in human*, Computational Biology and Chemistry, 53, 153-165 (2014).
- **R. E. Boulos**, G. Drillon, F. Argoul, A. Arneodo & B. Audit, *Structural organization of human replication timing domains*, FEBS Letters, DOI: 10.1016 (2015).

Conference Proceedings

- **R. E. Boulos**, A. Arneodo, P. Jensen & B. Audit, *Graph analysis of chromatin conformation data in relation with the human replication program*, 14ème Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Toulouse (2013), p. 35–42.
- B. Audit, A. Baker, **R. E. Boulos**, H. Julienne, A. Arneodo, C. -L. Chen, Y. d'Aubenton-Carafa, C. Thermes, A. Goldar, G. Guilbaud, A. Rappailles & O. Hyrien, *Relating mammalian replication program to large-scale chromatin folding*, ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics (ACM BCB), Washington, D.C. (2013), p. 800–811.
- **R. E. Boulos**, N. Tremblay, A. Arneodo, P. Borgnat, & B. Audit, *Applications des ondelettes sur graphe en génomique*,Proceedings of Gretsi (2015).

Oral Presentations

- **R. E. Boulos**, A. Arneodo, P. Jensen & B. Audit, *Long-range interconnected hubs in human chromatin conformation graph*, 3rd Les Houches School in computational physics: DNA, from molecules to evolution, Les Houches (2013).

- **R. E. Boulos**, A. Arneodo, P. Jensen & B. Audit, *Graph analysis of chromatin conformation data in relation with the human replication program*, 14ème Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Toulouse (2013).
- **R. E. Boulos**, *Using graph signal processing to study DNA structure*, 9ème Ecole d'Été de Peyresq en Traitement du Signal et des Images "Signaux et Images en Grandes Dimensions", Peyresq (2014).
- **R. E. Boulos**, N. Tremblay, P. Borgnat, P. Jensen, A. Arneodo, & B. Audit, *Human genome organization into strucrural communities*, Theoretical Approaches for the Genome and the proteome, Bourget du Lac (2014).

Other Communications

- **R. E. Boulos**, A. Arneodo, P. Jensen & B.Audit, *Using graph theory to describe chromatin 3D organization*, Poster in the joint conference for GDR ISIS et GDR Phénix (November 2013).
- **R. E. Boulos**, N. Tremblay, P. Borgnat, P. Jensen, A. Arneodo & B. Audit, *Human genome organization into structural communities*, Poster in LyonSysBio (December 2014).

Bibliography

[1] T. Cremer & C. Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2**, 292–301 (2001).

[2] B. Alberts. *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell* (Garland Publishing, 1998).

[3] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent & T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).

[4] A. S. Belmont, S. Dietzel, A. C. Nye, Y. G. Strukov & T. Tumbar. Large-scale chromatin structure and function. *Curr. Opin. Cell Biol.* **11**, 307–311 (1999).

[5] P. R. Cook. The organization of replication and transcription. *Science* **284**, 1790–1795 (1999).

[6] P. R. Cook. *Principles of Nuclear Structure and Functions* (Wiley, New York, 2001).

[7] R. Berezney. Regulating the mammalian genome: the role of nuclear architecture. *Adv. Enzyme Regul.* **42**, 39–52 (2002).

[8] N. Gilbert, S. Gilchrist & W. A. Bickmore. Chromatin organization in the mammalian nucleus. *Int. Rev. Cytol.* **242**, 283–336 (2005).

[9] T. Misteli. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787–800 (2007).

[10] T. Sexton, H. Schober, P. Fraser & S. M. Gasser. Gene regulation through nuclear organization. *Nat. Struct. Mol. Biol.* **14**, 1049–1055 (2007).

[11] M. R. Branco & A. Pombo. Chromosome organization: new facts, new models. *Trends Cell. Biol.* **17**, 127–134 (2007).

[12] P. Fraser & W. Bickmore. Nuclear organization of the genome and the potential for gene regulation. *Nature* **447**, 413–417 (2007).

[13] G. Cavalli & T. Misteli. Functional implications of genome topology. *Nat. Struct. Mol. Biol.* **20**, 290–299 (2013).

[14] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander & J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

[15] A. J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning* (Springer, New York, 2008).

[16] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu & B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

[17] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat & B. van Steensel. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).

[18] L. Handoko, H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. H. Lee, C. Ye, J. L. H. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W.-K. Sung, Y. Ruan & C.-L. Wei. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.* **43**, 630–638 (2011).

[19] C. Hou, L. Li, Z. S. Qin & V. G. Corces. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol. Cell* **48**, 471–484 (2012).

[20] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay & G. Cavalli. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).

[21] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen, J. Dekker & E. Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).

[22] S. Feng, S. J. Cokus, V. Schubert, J. Zhai, M. Pellegrini & S. E. Jacobsen. Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in *Arabidopsis*. *Mol. Cell.* **55**, 694–707 (2014).

[23] S. Grob, M. W. Schmid & U. Grossniklaus. Hi-C analysis in *Arabidopsis* identifies the KNOT, a structure with similarities to the flamenco locus of *Drosophila*. *Mol. Cell.* **55**, 678–693 (2014).

[24] F. Ay, T. L. Bailey & W. S. Noble. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).

[25] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau & W. S. Noble. A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).

[26] H. Tanizawa, O. Iwasaki, A. Tanaka, J. R. Capizzi, P. Wickramasinghe, M. Lee, Z. Fu & K.-i. Noma. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.* **38**, 8164–8177 (2010).

[27] M. A. Marti-Renom & L. A. Mirny. Bridging the resolution gap in structural modeling of 3D genome organization. *PLoS Comput. Biol.* **7**, e1002125 (2011).

[28] M. A. Umbarger, E. Toro, M. A. Wright, G. J. Porreca, D. Bau, S.-H. Hong, M. J. Fero, L. J. Zhu, M. A. Marti-Renom, H. H. McAdams, L. Shapiro, J. Dekker & G. M. Church. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol. Cell* **44**, 252–264 (2011).

[29] H. Marie-Nelly, M. Marbouty, A. Cournac, J.-F. Flot, G. Liti, D. P. Parodi, S. Syan, N. Guillen, A. Margeot, C. Zimmer & R. Koszul. High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* **5**, 5695 (2014).

[30] M. Marbouty, A. Cournac, J.-F. Flot, H. Marie-Nelly, J. Mozziconacci & R. Koszul. Metagenomic chromosome conformation capture (meta3c) unveils the diversity of chromosome organization in microorganisms. *Elife* **3**, e03318 (2014).

[31] J. F. Marko & E. D. Siggia. Polymer models of meiotic and mitotic chromosomes. *Mol. Biol. Cell* **8**, 2217–2231 (1997).

[32] M. Bohn & D. W. Heermann. Repulsive forces between looping chromosomes induce entropy-driven segregation. *PLoS One* **6**, e14428 (2011).

[33] R. K. Sachs, G. van den Engh, B. Trask, H. Yokota & J. E. Hearst. A random-walk/giant-loop model for interphase chromosomes. *Proc. Natl. Acad. Sci. USA* **92**, 2710–2714 (1995).

[34] A. Y. Grossberg, S. K. Nechaer & E. Shakhnovich. The role of topological constraints in the kinetics of collapse of macromeleculs. *J. Physique* **49**, 2095–2100 (1988).

[35] K. Woodfine, H. Fiegler, D. M. Beare, J. E. Collins, O. T. McCann, B. D. Young, S. Debernardi, R. Mott, I. Dunham & N. P. Carter. Replication timing of the human genome. *Hum. Mol. Genet.* **13**, 191–202 (2004).

[36] R. Desprat, D. Thierry-Mieg, N. Lailler, J. Lajugie, C. Schildkraut, J. Thierry-Mieg & E. E. Bouhassira. Predictable dynamic program of timing of DNA replication in human cells. *Genome Res.* **19**, 2288–2299 (2009).

[37] E. Yaffe, S. Farkash-Amar, A. Polten, Z. Yakhini, A. Tanay & I. Simon. Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet.* **6**, e1001011 (2010).

[38] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton & D. M. Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–770 (2010).

[39] E.-B. Brodie of Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes & A. Arneodo. From DNA sequence analysis to modeling replication in the human genome. *Phys. Rev. Lett.* **94**, 248103 (2005).

[40] M. Touchon, S. Nicolay, B. Audit, E.-B. Brodie of Brodie, Y. d'Aubenton-Carafa, A. Arneodo & C. Thermes. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc. Natl. Acad. Sci. USA* **102**, 9836–9841 (2005).

[41] M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo & C. Thermes. Human gene organization driven by the coordination of replication and transcription. *Genome Res.* **17**, 1278–1285 (2007).

[42] A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d'Aubenton-Carafa & C. Thermes. Multi-scale coding of genomic information: From DNA sequence to genome structure and function. *Phys. Rep.* **498**, 45–188 (2011).

[43] B. Audit, A. Baker, C.-L. Chen, A. Rappailles, G. Guilbaud, H. Julienne, A. Goldar, Y. d'Aubenton Carafa, O. Hyrien, C. Thermes & A. Arneodo. Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. *Nat. Protoc.* **8**, 98–110 (2013).

[44] A. Baker, B. Audit, C.-L. Chen, B. Moindrot, A. Leleu, G. Guilbaud, A. Rappailles, C. Vaillant, A. Goldar, F. Mongelard, Y. d'Aubenton Carafa, O. Hyrien, C. Thermes & A. Arneodo. Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *PLoS Comput. Biol.* **8**, e1002443 (2012).

[45] B. Audit, S. Nicolay, M. Huvet, M. Touchon, Y. d'Aubenton Carafa, C. Thermes & A. Arneodo. DNA replication timing data corroborate in silico human replication origin predictions. *Phys. Rev. Lett.* **99**, 248102 (2007).

[46] A. Baker, S. Nicolay, L. Zaghloul, Y. d'Aubenton-Carafa, C. Thermes, B. Audit & A. Arneodo. Wavelet-based method to disentangle transcription- and replication-associated strand asymmetries in mammalian genomes. *Appl. Comput. Harmon. Anal.* **28**, 150–170 (2010).

[47] C.-L. Chen, L. Duquenne, B. Audit, G. Guilbaud, A. Rappailles, A. Baker, M. Huvet, Y. d'Aubenton Carafa, O. Hyrien, A. Arneodo & C. Thermes. Replication-associated mutational asymmetry in the human genome. *Mol. Biol. Evol.* **28**, 2327–2337 (2011).

[48] A. Baker, H. Julienne, C. L. Chen, B. Audit, Y. d'Aubenton Carafa, C. Thermes & A. Arneodo. Linking the DNA strand asymmetry to the spatio-temporal replication program. I. About the role of the replication fork polarity in genome evolution. *Eur. Phys. J. E* **35**, 92 (2012).

[49] A. Baker, C. L. Chen, H. Julienne, B. Audit, Y. d'Aubenton Carafa, C. Thermes & A. Arneodo. Linking the DNA strand asymmetry to the spatio-temporal replication program: II. Accounting for neighbor-dependent substitution rates. *Eur. Phys. J. E* **35**, 123 (2012).

[50] L. Zaghloul, A. Baker, B. Audit & A. Arneodo. Gene organization inside replication domains in mammalian genomes. *C. R. Mécanique* **340**, 745–757 (2012).

[51] B. Audit, L. Zaghloul, C. Vaillant, G. Chevereau, Y. d'Aubenton-Carafa, C. Thermes & A. Arneodo. Open chromatin encoded in DNA sequence is the signature of "master" replication origins in human cells. *Nucleic Acids Res.* **37**, 6064–6075 (2009).

[52] L. Kaufman & P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis* (John Wiley & Sons, New York, 1984).

[53] H. Julienne, A. Zoufir, B. Audit & A. Arneodo. Human genome replication proceeds through four chromatin states. *PLoS Comput. Biol.* **9**, e1003233 (2013).

[54] H. Julienne, B. Audit & A. Arneodo. Embryonic stem cell specific "master" replication origins at the heart of the loss of pluripotency. *PLoS Comput. Biol.* **11**, e1003969 (2015).

[55] B. Moindrot, B. Audit, P. Klous, A. Baker, C. Thermes, W. de Laat, P. Bouvet, F. Mongelard & A. Arneodo. 3D chromatin conformation correlates with replication timing and is conserved in resting cells. *Nucleic Acids Res.* **40**, 9470–9481 (2012).

[56] D. B. West. *Introduction to Graph Theory* (Prentice Hall, Englewood, Cliffs, NJ, 1995).

[57] B. Bollobas. *Modern Graph Theory* (Springer, New York, USA, 1998).

[58] S. Wasserman & K. Faust. *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).

[59] J. Scott. *Social Network Analysis: A Handbook* (Sage Publications, London, 2000).

[60] A. W. Rives & T. Galitski. Modular organization of cellular networks. *Proc. Natl. Acad. Sci. USA* **100**, 1128–1133 (2003).

[61] V. Spirin & L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA* **100**, 12123–12128 (2003).

[62] J. F. F. Mendes & S. N. Dorogovtsev. *Evolution Networks: From Biological Nets to the internet and WWW* (Oxford University Press, Oxford, 2003).

[63] R. Pastor-Satorras & A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approch* (Cambridge University Press, Cambridge, 2004).

[64] R. Albert & A.-L. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47 (2002).

[65] S. Boccaletti, V. Latora, Y. Moreno, M. Charez & D. Hwang. Complex networks: Structure and dynamics. *Phys. Rep.* **424**, 175 (2006).

[66] *Large Scale Structure and Dynamics of Complex Networks*, edited by G. Caldarelli & A. Vespignani (World Scientific, Singapore, 2007).

[67] C. E. Pearson, K. Nichol Edamura & J. D. Cleary. Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* **6**, 729–742 (2005).

[68] I. Hiratani & D. M. Gilbert. Replication timing as an epigenetic mark. *Epigenetics* **4**, 93–97 (2009).

[69] O. Hyrien, A. Rappailles, G. Guilbaud, A. Baker, C.-L. Chen, A. Goldar, N. Petryk, M. Kahli, E. Ma, Y. d'Aubenton Carafa, B. Audit, C. Thermes & A. Arneodo. From simple bacterial and archaeal replicons to replication N/U-domains. *J. Mol. Biol.* **425**, 4673–4689 (2013).

[70] G. Guilbaud, A. Rappailles, A. Baker, C.-L. Chen, A. Arneodo, A. Goldar, Y. d'Aubenton-Carafa, C. Thermes, B. Audit & O. Hyrien. Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. *PLoS Comput. Biol.* **7**, e1002322 (2011).

[71] E. Chargaff. Structure and function of nucleic acids as cell constituents. *Fed. Proc.* **10**, 654–659 (1951).

[72] R. Rudner, J. D. Karkas & E. Chargaff. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proc. Natl. Acad. Sci. USA* **60**, 921–922 (1968).

[73] J. W. Fickett, D. C. Torney & D. R. Wolf. Base compositional structure of genomes. *Genomics* **13**, 1056–1064 (1992).

[74] J. R. Lobry. Properties of a general model of DNA evolution under no-strand-bias conditions. *J. Mol. Evol.* **40**, 326–330 (1995).

[75] T. Gojobori, W. H. Li & D. Graur. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**, 360–369 (1982).

[76] W. H. Li, C. I. Wu & C. C. Luo. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**, 58–71 (1984).

[77] D. A. Petrov & D. L. Hartl. Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc. Natl. Acad. Sci. USA* **96**, 1475–1479 (1999).

[78] Z. Zhang & M. Gerstein. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.* **31**, 5338–5348 (2003).

[79] J. D. Watson & F. H. Crick. Molecular structure of nucleic acids; A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).

[80] J. D. Watson & F. H. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967 (1953).

[81] J. R. Lobry & C. Lobry. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol. Biol. Evol.* **16**, 719–723 (1999).

[82] M. Touchon, S. Nicolay, A. Arneodo, Y. d'Aubenton-Carafa & C. Thermes. Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett.* **555**, 579–582 (2003).

[83] B. Alberts, J. A., J. Lewis, M. Raff, K. Roberts & W. P. *Molecular Biology of the Cell, 5th ed* (Garland Publishing, 2008).

[84] M. P. Francino & H. Ochman. Strand asymmetries in DNA evolution. *Trends Genet* **13**, 240–245 (1997).

[85] A. C. Frank & J. R. Lobry. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**, 65–77 (1999).

[86] M. P. Francino & H. Ochman. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. *Mol. Biol. Evol.* **18**, 1147–1150 (2001).

[87] P. Green, B. Ewing, W. Miller, P. J. Thomas & E. D. Green. Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33**, 514–517 (2003).

[88] P. Polak & P. F. Arndt. Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res.* **18**, 1216–1223 (2008).

[89] C. F. Mugal, H.-H. von Gronberg & M. Peifer. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol. Biol. Evol.* **26**, 131–142 (2009).

[90] A. Beletskii & A. S. Bhagwat. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. *Proc Natl Acad Sci U S A* **93**, 13919–13924 (1996).

[91] M. P. Francino, L. Chao, M. A. Riley & H. Ochman. Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science* **272**, 107–109 (1996).

[92] J. Q. Svejstrup. Mechanisms of transcription-coupled DNA repair. *Nat. Rev. Mol. Cell Biol.* **3**, 21–29 (2002).

[93] J. M. Freeman, T. N. Plasterer, T. F. Smith & S. C. Mohr. Patterns of genome organization in bacteria. *Science* **279**, 1827 (1998).

[94] A. Beletskii, A. Grigoriev, S. Joyce & A. S. Bhagwat. Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. *J. Mol. Biol.* **300**, 1057–1065 (2000).

[95] L. Duret. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**, 640–649 (2002).

[96] C. Shioiri & N. Takahata. Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *J. Mol. Evol.* **53**, 364–376 (2001).

[97] M. Touchon, A. Arneodo, Y. d'Aubenton-Carafa & C. Thermes. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res.* **32**, 4969–4978 (2004).

[98] E. P. Rocha, A. Danchin & A. Viari. Universal replication biases in bacteria. *Mol. Microbiol.* **32**, 11–16 (1999).

[99] E. P. C. Rocha, M. Touchon & E. J. Feil. Similar compositional biases are caused by very different mutational effects. *Genome Res.* **16**, 1537–1547 (2006).

[100] P. Polak & P. F. Arndt. Long-range bidirectional strand asymmetries originate at cpg islands in the human genome. *Genome Biol. Evol.* **1**, 189–197 (2009).

[101] J. Mrázek & S. Karlin. Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA* **95**, 3720–3725 (1998).

[102] E. R. Tillier & R. A. Collins. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* **50**, 249–257 (2000).

[103] F. Jacob, S. Brenner & F. Cuzin. On the regulation of DNA replication in bacteria. *Cold Spring Harb. Symp. Quant. Biol.* **28**, 329–342 (1963).

[104] B. J. Brewer. When polymerases collide: replication and the transcriptional organization of the E. coli chromosome. *Cell* **53**, 679–686 (1988).

[105] E. P. C. Rocha, P. Guerdoux-Jamet, I. Moszer, A. Viari & A. Danchin. Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. *J. Biotech.* **78**, 209–219 (2000).

[106] P. Lopez & H. Philippe. Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation. *C. R. Acad. Sci. III* **324**, 201–208 (2001).

[107] E. P. C. Rocha. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes. *Trends Microbiol.* **10**, 393–395 (2002).

[108] O. Hyrien & M. Méchali. Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of Xenopus early embryos. *EMBO J.* **12**, 4511–4520 (1993).

[109] S. A. Gerbi & A. K. Bielinsky. DNA replication and chromatin. *Curr. Opin. Genet. Dev.* **12**, 243–248 (2002).

[110] D. Schübeler, D. Scalzo, C. Kooperberg, B. van Steensel, J. Delrow & M. Groudine. Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. *Nat. Genet.* **32**, 438–442 (2002).

[111] M. Anglana, F. Apiou, A. Bensimon & M. Debatisse. Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell* **114**, 385–394 (2003).

[112] D. Fisher & M. Méchali. Vertebrate HoxB gene expression requires DNA replication. *EMBO J.* **22**, 3737–3748 (2003).

[113] T. A. Kunkel & P. M. Burgers. Dividing the workload at a eukaryotic replication fork. *Trends Cell Biol.* **18**, 521–527 (2008).

[114] M. Bulmer. Strand symmetry of mutation rates in the beta-globin region. *J. Mol. Evol.* **33**, 305–310 (1991).

[115] M. P. Francino & H. Ochman. Strand symmetry around the beta-globin origin of replication in primates. *Mol. Biol. Evol.* **17**, 416–422 (2000).

[116] A. Gierlik, M. Kowalczuk, P. Mackiewicz, M. R. Dudek & S. Cebrat. Is there replication-associated mutational pressure in the Saccharomyces cerevisiae genome? *J. Theor. Biol.* **202**, 305–314 (2000).

[117] E. Louie, J. Ott & J. Majewski. Nucleotide frequency variation across human genes. *Genome Res.* **13**, 2594–2601 (2003).

[118] A. Arneodo, Y. d'Aubenton-Carafa, B. Audit, E.-B. Brodie of Brodie, S. Nicolay, P. St-Jean, C. Thermes, M. Touchon & C. Vaillant. DNA in chromatin: from genome-wide sequence analysis to the modeling of replication in mammals. *Adv. Chem. Phys.* **135**, 203–252 (2007).

[119] D. M. Gilbert. Evaluating genome-scale approaches to eukaryotic DNA replication. *Nat. Rev. Genet.* **11**, 673–684 (2010).

[120] A. Schepers & P. Papior. Why are we where we are? understanding replication origins and initiation sites in eukaryotes using ChIP-approaches. *Chromosome Res.* **18**, 63–77 (2010).

[121] C. Heichinger, C. J. Penkett, J. Bahler & P. Nurse. Genome-wide characterization of fission yeast DNA replication origins. *EMBO J.* **25**, 5171–5179 (2006).

[122] N. Yabuki, H. Terashima & K. Kitada. Mapping of early firing origins on a replication profile of budding yeast. *Genes Cells* **7**, 781–789 (2002).

[123] H. K. MacAlpine, R. Gordan, S. K. Powell, A. J. Hartemink & D. M. MacAlpine. *Drosophila* ORC localizes to open chromatin and marks sites of cohesin complex loading. *Genome Res.* **20**, 201–211 (2010).

[124] M. Hayashi, Y. Katou, T. Itoh, A. Tazumi, M. Tazumi, Y. Yamada, T. Takahashi, T. Nakagawa, K. Shirahige & H. Masukata. Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast. *EMBO J.* **26**, 1327–1339 (2007).

[125] P.-Y. J. Wu & P. Nurse. Establishing the program of origin firing during S phase in fission yeast. *Cell* **136**, 852–864 (2009).

[126] Y. Katou, Y. Kanoh, M. Bando, H. Noguchi, H. Tanaka, T. Ashikari, K. Sugimoto & K. Shirahige. S-phase checkpoint proteins Tof1 and Mrc1 form a stable replication-pausing complex. *Nature* **424**, 1078–1083 (2003).

[127] C. J. Viggiani, S. R. V. Knott & O. M. Aparicio. Genome-wide analysis of DNA synthesis by BrdU immunoprecipitation on tiling microarrays (BrdU-IP-chip) in *Saccharomyces cerevisiae*. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5385 (2010).

[128] W. Feng, D. Collingwood, M. E. Boeck, L. A. Fox, G. M. Alvino, W. L. Fangman, M. K. Raghuraman & B. J. Brewer. Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. *Nat. Cell Biol.* **8**, 148–155 (2006).

[129] T. Sasaki, S. Ramanathan, Y. Okuno, C. Kumagai, S. S. Shaikh & D. M. Gilbert. The Chinese hamster dihydrofolate reductase replication origin decision point follows activation of transcription and suppresses initiation of replication within transcription units. *Mol. Cell Biol.* **26**, 1051–1062 (2006).

[130] H. Masai, S. Matsumoto, Z. You, N. Yoshizawa-Sugata & M. Oda. Eukaryotic chromosome DNA replication: where, when, and how? *Annu. Rev. Biochem.* **79**, 89–130 (2010).

[131] C. Costas, M. de la Paz Sanchez, H. Stroud, Y. Yu, J. C. Oliveros, S. Feng, A. Benguria, I. Lopez-Vidriero, X. Zhang, R. Solano, S. E. Jacobsen & C. Gutierrez. Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks. *Nat. Struct. Mol. Biol.* **18**, 395–400 (2011).

[132] C. Cayrou, P. Coulombe, A. Vigneron, S. Stanojcic, O. Ganier, I. Peiffer, E. Rivals, A. Puy, S. Laurent-Chabalier, R. Desprat & M. Méchali. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res.* **21**, 1438–1449 (2011).

[133] J. Sequeira-Mendes, R. Diaz-Uriarte, A. Apedaile, D. Huntley, N. Brockdorff & M. Gomez. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.* **5**, e1000446 (2009).

[134] I. Lucas, A. Palakodeti, Y. Jiang, D. J. Young, N. Jiang, A. A. Fernald & M. M. Le Beau. High-throughput mapping of origins of replication in human cells. *EMBO Rep.* **8**, 770–777 (2007).

[135] J.-C. Cadoret, F. Meisch, V. Hassan-Zadeh, I. Luyten, C. Guillet, L. Duret, H. Quesneville & M.-N. Prioleau. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl. Acad. Sci. USA* **105**, 15837–15842 (2008).

[136] N. Karnani, C. M. Taylor, A. Malhotra & A. Dutta. Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. *Mol. Biol. Cell* **21**, 393–404 (2010).

[137] E. Besnard, A. Babled, L. Lapasset, O. Milhavet, H. Parrinello, C. Dantec, J.-M. Marin & J.-M. Lemaitre. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol.* **19**, 837–844 (2012).

[138] F. Picard, J.-C. Cadoret, B. Audit, A. Arneodo, A. Alberti, C. Battail, L. Duret & M.-N. Prioleau. The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells. *PLoS Genet.* **10**, e1004282 (2014).

[139] R. Mukhopadhyay, J. Lajugie, N. Fourel, A. Selzer, M. Schizas, B. Bartholdy, J. Mar, C. M. Lin, M. M. Martin, M. Ryan, M. I. Aladjem & E. E. Bouhassira. Allele-specific genome-wide profiling in human primary erythroblasts reveal replication program organization. *PLoS Genet.* **10**, e1004319 (2014).

[140] L. D. Mesner, V. Valsakumar, N. Karnani, A. Dutta, J. L. Hamlin & S. Bekiranov. Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription. *Genome Res.* **21**, 377–389 (2011).

[141] L. D. Mesner, E. L. Crawford & J. L. Hamlin. Isolating apparently pure libraries of replication origins from complex genomes. *Mol. Cell* **21**, 719–726 (2006).

[142] L. D. Mesner, V. Valsakumar, M. Cieslik, R. Pickin, J. L. Hamlin & S. Bekiranov. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome Res.* **23**, 1774–1788 (2013).

[143] J. L. Hamlin, L. D. Mesner & P. A. Dijkwel. A winding road to origin discovery. *Chromosome Res.* **18**, 45–61 (2010).

[144] M. Méchali. Eukaryotic DNA replication origins: many choices for appropriate answers. *Nat. Rev. Mol. Cell Biol.* **11**, 728–738 (2010).

[145] M. K. Raghuraman, E. A. Winzeler, D. Collingwood, S. Hunt, L. Wodicka, A. Conway, D. J. Lockhart, R. W. Davis, B. J. Brewer & W. L. Fangman. Replication dynamics of the yeast genome. *Science* **294**, 115–121 (2001).

[146] T.-J. Lee, P. E. Pascuzzi, S. B. Settlage, R. W. Shultz, M. Tanurdzic, P. D. Rabinowicz, M. Menges, P. Zheng, D. Main, J. A. H. Murray, B. Sosinski, G. C. Allen, R. A. Martienssen, L. Hanley-Bowdoin, M. W. Vaughn & W. F. Thompson. Arabidopsis thaliana chromosome 4 replicates in two phases that correlate with chromatin state. *PLoS Genet.* **6**, e1000982 (2010).

[147] I. Hiratani, T. Ryba, M. Itoh, T. Yokochi, M. Schwaiger, C.-W. Chang, Y. Lyou, T. M. Townes, D. Schubeler & D. M. Gilbert. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.* **6**, e245 (2008).

[148] I. Hiratani, T. Ryba, M. Itoh, J. Rathjen, M. Kulik, B. Papp, E. Fussner, D. P. Bazett-Jones, K. Plath, S. Dalton, P. D. Rathjen & D. M. Gilbert. Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res.* **20**, 155–169 (2010).

[149] C.-L. Chen, A. Rappailles, L. Duquenne, M. Huvet, G. Guilbaud, L. Farinelli, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, O. Hyrien & C. Thermes. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.* **20**, 447–457 (2010).

[150] K. Woodfine, D. M. Beare, K. Ichimura, S. Debernardi, A. J. Mungall, H. Fiegler, V. P. Collins, N. P. Carter & I. Dunham. Replication timing of human chromosome 6. *Cell Cycle* **4**, 172–176 (2005).

[151] R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler & J. A. Stamatoyannopoulos. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. USA* **107**, 139–144 (2010).

[152] D. M. MacAlpine & S. P. Bell. A genomic view of eukaryotic DNA replication. *Chromosome Res.* **13**, 309–326 (2005).

[153] T. Ryba, D. Battaglia, B. D. Pope, I. Hiratani & D. M. Gilbert. Genome-scale analysis of replication timing: from bench to bioinformatics. *Nat. Protoc.* **6**, 870–895 (2011).

[154] S. Farkash-Amar & I. Simon. Genome-wide analysis of the replication program in mammals. *Chromosome Res.* **18**, 115–125 (2010).

[155] D. M. Gilbert & S. N. Cohen. Bovine papilloma virus plasmids replicate randomly in mouse fibroblasts throughout s phase of the cell cycle. *Cell* **50**, 59–68 (1987).

[156] D. M. MacAlpine, H. K. Rodriguez & S. P. Bell. Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev.* **18**, 3094–3105 (2004).

[157] S. Farkash-Amar, D. Lipson, A. Polten, A. Goren, C. Helmstetter, Z. Yakhini & I. Simon. Global organization of replication time zones of the mouse genome. *Genome Res.* **18**, 1562–1570 (2008).

[158] M. L. Eaton, J. A. Prinz, H. K. MacAlpine, G. Tretyakov, P. V. Kharchenko & D. M. MacAlpine. Chromatin signatures of the drosophila replication program. *Genome Res.* **21**, 164–174 (2011).

[159] N. Karnani, C. Taylor, A. Malhotra & A. Dutta. Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res.* **17**, 865–876 (2007).

[160] N. Weddington, A. Stuy, I. Hiratani, T. Ryba, T. Yokochi & D. M. Gilbert. Replicationdomain: a visualization tool and comparative database for genome-wide replication timing data. *BMC Bioinformatics* **9**, 530 (2008).

[161] B. Audit, L. Zaghloul, A. Baker, A. Arneodo, C.-L. Chen, Y. d'Aubenton Carafa & C. Thermes. Megabase replication domains along the human genome: relation to chromatin structure and genome organisation. *Subcell. Biochem.* **61**, 57–80 (2012).

[162] J. F. X. Diffley. Regulation of early events in chromosome replication. *Curr. Biol.* **14**, R778–R786 (2004).

[163] S. C.-H. Yang, N. Rhind & J. Bechhoefer. Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Mol. Syst. Biol.* **6**, 404 (2010).

[164] A. P. S. de Moura, R. Retkute, M. Hawkins & C. A. Nieduszynski. Mathematical modelling of whole chromosome replication. *Nucleic Acids Res.* **38**, 5623–5633 (2010).

[165] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

[166] L. Chantalat, J. M. Nicholson, S. J. Lambert, A. J. Reid, M. J. Donovan, C. D. Reynolds, C. M. Wood & J. P. Baldwin. Structure of the histone-core octamer in KCl/phosphate crystals at 2.15 Å resolution. *Acta Crystallogr. D Biol. Crystallogr.* **59**, 1395–1407 (2003).

[167] T. J. Richmond & C. A. Davey. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).

[168] T. Kouzarides. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).

[169] J. E. Phillips & V. G. Corces. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).

[170] D. E. Schones & K. Zhao. Genome-wide approaches to studying chromatin modifications. *Nat. Rev. Genet.* **9**, 179–191 (2008).

[171] V. W. Zhou, A. Goren & B. E. Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* **12**, 7–18 (2011).

[172] O. J. Rando & H. Y. Chang. Genome-wide views of chromatin structure. *Annu. Rev. Biochem.* **78**, 245–271 (2009).

[173] F. Roudier, F. K. Teixeira & V. Colot. Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. *Trends Genet.* **25**, 511–517 (2009).

[174] S. Feng & S. E. Jacobsen. Epigenetic modifications in plants: an evolutionary perspective. *Curr. Opin. Plant Biol.* **14**, 179–186 (2011).

[175] modENCODE Consortium, S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. R. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White & M. Kellis. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**, 1787–1797 (2010).

[176] P. V. Kharchenko, A. A. Alekseyenko, Y. B. Schwartz, A. Minoda, N. C. Riddle, J. Ernst, P. J. Sabo, E. Larschan, A. A. Gorchakov, T. Gu, D. Linder-Basso, A. Plachetka, G. Shanower, M. Y. Tolstorukov, L. J. Luquette, R. Xi, Y. L. Jung, R. W. Park, E. P. Bishop, T. K. Canfield, R. Sandstrom, R. E. Thurman, D. M. MacAlpine, J. A. Stamatoyannopoulos, M. Kellis, S. C. R. Elgin, M. I. Kuroda, V. Pirrotta, G. H. Karpen & P. J. Park. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**, 480–485 (2011).

[177] B. E. Bernstein, A. Meissner & E. S. Lander. The mammalian epigenome. *Cell* **128**, 669–681 (2007).

[178] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

[179] D. M. Gilbert. In search of the holy replicator. *Nat. Rev. Mol. Cell Biol.* **5**, 848–855 (2004).

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

[181] F. Ozsolak, J. S. Song, X. S. Liu & D. E. Fisher. High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.* **25**, 244–248 (2007).

[182] D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei & K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).

[183] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev & K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).

[184] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey & G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

[185] H. Kohzaki & Y. Murakami. Transcription factors and DNA replication origin selection. *Bioessays* **27**, 1107–1116 (2005).

[186] W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes & C. Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* **39**, 1235–1244 (2007).

[187] G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler & O. J. Rando. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**, 626–630 (2005).

[188] F. Roudier, I. Ahmed, C. Bérard, A. Sarazin, T. Mary-Huard, S. Cortijo, D. Bouyer, E. Caillieux, E. Duvernois-Berthet, L. Al-Shikhley, L. Giraut, B. Després, S. Drevensek, F. Barneche, S. Dèrozier, V. Brunaud, S. Aubourg, A. Schnittger, C. Bowler, M.-L. Martin-Magniette, S. Robin, M. Caboche & V. Colot. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *EMBO J.* **30**, 1928–1938 (2011).

[189] T. Liu, A. Rechtsteiner, T. A. Egelhofer, A. Vielle, I. Latorre, M.-S. Cheung, S. Ercan, K. Ikegami, M. Jensen, P. Kolasinska-Zwierz, H. Rosenbaum, H. Shin, S. Taing, T. Takasaki, A. L. Iniguez, A. Desai, A. F. Dernburg, H. Kimura, J. D. Lieb, J. Ahringer, S. Strome & X. S. Liu. Broad chromosomal domains of histone modification patterns in *C. elegans*. *Genome Res.* **21**, 227–236 (2011).

[190] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker & B. van Steensel. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell* **143**, 212–224 (2010).

[191] J. Ernst & M. Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).

[192] Z. Wang, D. E. Schones & K. Zhao. Characterization of human epigenomes. *Curr. Opin. Genet. Dev.* **19**, 127–134 (2009).

[193] B.-K. Lee, A. A. Bhinge, A. Battenhouse, R. M. McDaniell, Z. Liu, L. Song, Y. Ni, E. Birney, J. D. Lieb, T. S. Furey, G. E. Crawford & V. R. Iyer. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Res.* **22**, 9–24 (2012).

[194] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis & B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

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

[196] R. D. Hawkins, G. C. Hon, L. K. Lee, Q. Ngo, R. Lister, M. Pelizzola, L. E. Edsall, S. Kuan, Y. Luu, S. Klugman, J. Antosiewicz-Bourget, Z. Ye, C. Espinoza, S. Agarwahl, L. Shen, V. Ruotti, W. Wang, R. Stewart, J. A. Thomson, J. R. Ecker & B. Ren. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).

[197] E. Meshorer & T. Misteli. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat. Rev. Mol. Cell Biol.* **7**, 540–546 (2006).

[198] J. Zhu, M. Adli, J. Y. Zou, G. Verstappen, M. Coyne, X. Zhang, T. Durham, M. Miri, V. Deshpande, P. L. De Jager, D. A. Bennett, J. A. Houmard, D. M. Muoio, T. T. Onder, R. Camahort, C. A. Cowan, A. Meissner, C. B. Epstein, N. Shoresh & B. E. Bernstein. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).

[199] V. Azuara, P. Perry, S. Sauer, M. Spivakov, H. F. Jrgensen, R. M. John, M. Gouti, M. Casanova, G. Warnes, M. Merkenschlager & A. G. Fisher. Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8**, 532–538 (2006).

[200] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber & E. S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).

[201] T. Prikrylova, J. Pachernik, S. Kozubek & E. Bartova. Epigenetics and chromatin plasticity in embryonic stem cells. *World J. Stem Cells* **5**, 73–85 (2013).

[202] T. Chandra, K. Kirschner, J.-Y. Thuret, B. D. Pope, T. Ryba, S. Newman, K. Ahmed, S. A. Samarajiwa, R. Salama, T. Carroll, R. Stark, R. Janky, M. Narita, L. Xue, A. Chicas, S. Ni, Y.nez, R. Janknecht, Y. Hayashi-Takanaka, M. D. Wilson, A. Marshall, D. T. Odom, M. M. Babu, D. P. Bazett-Jones, S. Tavare, P. A. W.

Edwards, S. W. Lowe, H. Kimura, D. M. Gilbert & M. Narita. Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. *Mol. Cell* **47**, 203–214 (2012).

[203] T. H. Kim, Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenkov & B. Ren. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).

[204] J. A. Simon & R. E. Kingston. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.* **10**, 697–708 (2009).

[205] V. Pirrotta & H.-B. Li. A view of nuclear polycomb bodies. *Curr. Opin. Genet. Dev.* **22**, 101–109 (2012).

[206] G. Mizuguchi, X. Shen, J. Landry, W.-H. Wu, S. Sen & C. Wu. ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303**, 343–348 (2004).

[207] E. Meshorer, D. Yellajoshula, E. George, P. J. Scambler, D. T. Brown & T. Misteli. Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev. Cell* **10**, 105–116 (2006).

[208] C. Hou, R. Dale & A. Dean. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc. Natl. Acad. Sci. USA* **107**, 3651–3656 (2010).

[209] R. Ohlsson, V. Lobanenkov & E. Klenova. Does CTCF mediate between nuclear organization and gene expression? *Bioessays* **32**, 37–50 (2010).

[210] M. Botta, S. Haider, I. X. Y. Leung, P. Lio & J. Mozziconacci. Intra- and interchromosomal interactions correlate with CTCF binding genome wide. *Mol. Syst. Biol.* **6**, 426 (2010).

[211] M. Merkenschlager & D. T. Odom. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152**, 1285–1297 (2013).

[212] Y. Li, W. Huang, L. Niu, D. M. Umbach, S. Covo & L. Li. Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. *BMC Genomics* **14**, 553 (2013).

[213] H. Julienne, A. Zoufir, B. Audit & A. Arneodo. Epigenetic regulation of the human genome: coherence between promoter activity and large-scale chromatin environment. *Front. Life Sci.* **7**, 44–62 (2013).

[214] I. Hiratani, S.-I. Takebayashi, J. Lu & D. M. Gilbert. Replication timing and transcriptional control: beyond cause and effect–part II. *Curr. Opin. Genet. Dev.* **19**, 142–149 (2009).

[215] E. S. Lander *et al.* Initial sequencing and analysis of the human genomes. *Nature* **409**, 860–921 (2001).

[216] E. S. Lander. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).

[217] G. Bernardi. The human genome: organization and evolutionary history. *Annu. Rev. Genet.* **29**, 445–476 (1995).

[218] G. Bernardi. Misunderstandings about isochores. Part 1. *Gene* **276**, 3–13 (2001).

[219] A. Eyre-Walker & L. D. Hurst. The evolution of isochores. *Nat. Rev. Genet.* **2**, 549–555 (2001).

[220] H. Nakayasu & R. Berezney. Mapping replicational sites in the eucaryotic cell nucleus. *J. Cell Biol.* **108**, 1–11 (1989).

[221] R. T. O'Keefe, S. C. Henderson & D. L. Spector. Dynamic organization of DNA replication in mammalian cell nuclei: spatially and temporally defined replication of chromosome-specific alpha-satellite DNA sequences. *J. Cell Biol.* **116**, 1095–1110 (1992).

[222] H. Leonhardt, H. P. Rahn, P. Weinzierl, A. Sporbert, T. Cremer, D. Zink & M. C. Cardoso. Dynamics of DNA replication factories in living cells. *J. Cell Biol.* **149**, 271–280 (2000).

[223] A. Sporbert, A. Gahl, R. Ankerhold, H. Leonhardt & M. C. Cardoso. DNA polymerase clamp shows little turnover at established replication sites but sequential de novo assembly at adjacent origin clusters. *Mol. Cell* **10**, 1355–1365 (2002).

[224] N. Sadoni, M. C. Cardoso, E. H. K. Stelzer, H. Leonhardt & D. Zink. Stable chromosomal units determine the spatial and temporal organization of DNA replication. *J. Cell Sci.* **117**, 5353–5365 (2004).

[225] A. Maya-Mendoza, P. Olivares-Chauvet, A. Shaw & D. A. Jackson. S phase progression in human cells is dictated by the genetic continuity of DNA foci. *PLoS Genet.* **6**, e1000900 (2010).

[226] H. Ma, J. Samarabandu, R. S. Devdhar, R. Acharya, P. C. Cheng, C. Meng & R. Berezney. Spatial and temporal dynamics of DNA replication sites in mammalian cells. *J. Cell Biol.* **143**, 1415–1425 (1998).

[227] D. Zink, T. Cremer, R. Saffrich, R. Fischer, M. F. Trendelenburg, W. Ansorge & E. H. Stelzer. Structure and dynamics of human interphase chromosome territories in vivo. *Hum. Genet.* **102**, 241–251 (1998).

[228] D. Zink, H. Bornfleth, A. Visser, C. Cremer & T. Cremer. Organization of early and late replicating DNA in human chromosome territories. *Exp. Cell Res.* **247**, 176–188 (1999).

[229] F. Grasser, M. Neusser, H. Fiegler, T. Thormeyer, M. Cremer, N. P. Carter, T. Cremer & S. Müller. Replication-timing-correlated spatial chromatin arrangements in cancer and in primate interphase nuclei. *J. Cell Sci.* **121**, 1876–1886 (2008).

[230] D. A. Jackson & A. Pombo. Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. *J. Cell Biol.* **140**, 1285–1295 (1998).

[231] K. Koberna, A. Ligasova, J. Malinsky, A. Pliss, A. J. Siegel, Z. Cvackova, H. Fidlerova, M. Masata, M. Fialova, I. Raska & R. Berezney. Electron microscopy of DNA replication in 3-D: evidence for similar-sized replication foci throughout S-phase. *J. Cell. Biochem.* **94**, 126–138 (2005).

[232] S. M. Gasser & U. K. Laemmli. A glimpse at chromosomal order. *Trends Genet.* **3**, 16–22 (1987).

[233] U. K. Laemmli, E. Käs, L. Poljak & Y. Adachi. Scaffold-associated regions: cis-acting determinants of chromatin structural loops and functional domains. *Curr. Opin. Genet. Dev.* **2**, 275–285 (1992).

[234] J. Dekker, K. Rippe, M. Dekker & N. Kleckner. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

[235] G. Felsenfeld & M. Groudine. Controlling the double helix. *Nature* **421**, 448–453 (2003).

[236] S. A. Sajan & R. D. Hawkins. Methods for identifying higher-order chromatin structure. *Annu. Rev. Genomics Hum. Genet.* **13**, 59–82 (2012).

[237] E. de Wit & W. de Laat. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26**, 11–24 (2012).

[238] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel & W. de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).

[239] Z. Zhao, G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti & R. Ohlsson. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006).

[240] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green & J. Dekker. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).

[241] M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Y. Chew, P. Y. H. Huang, W.-J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. S. A. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. M. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W.-K. Sung, E. T. Liu, C.-L. Wei, E. Cheung & Y. Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).

[242] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber & L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2012).

[243] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny & J. Dekker. Organization of the mitotic chromosome. *Science* **342**, 948–953 (2013).

[244] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander & E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

[245] J. Wang, X. Lan, P.-Y. Hsu, H.-K. Hsu, K. Huang, J. Parvin, T. H.-M. Huang & V. X. Jin. Genome-wide analysis uncovers high frequency, strong differential chromosomal interactions and their associated epigenetic patterns in E2-mediated gene regulation. *BMC Genomics* **14**, 70 (2013).

[246] J. Dekker. The three 'C's of chromosome conformation capture: controls, controls, controls. *Nat. Methods* **3**, 17–21 (2006).

[247] E. Yaffe & A. Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**, 1059–1065 (2011).

[248] A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul & J. Mozziconacci. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).

[249] J. C. Dohm, C. Lottaz, T. Borodina & H. Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).

[250] D. Arid, M. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum & A. Gnirke. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12** (2011).

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

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

[253] M. Schwaiger, M. B. Stadler, O. Bell, H. Kohler, E. J. Oakeley & D. Schübeler. Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. *Genes Dev.* **23**, 589–601 (2009).

[254] T. Ryba, I. Hiratani, T. Sasaki, D. Battaglia, M. Kulik, J. Zhang, S. Dalton & D. M. Gilbert. Replication timing: a fingerprint for cell identity and pluripotency. *PLoS Comput. Biol.* **7**, e1002225 (2011).

[255] F. Le Dily, D. Bae, A. Pohl, G. P. Vicent, F. Serra, D. Soronellas, G. Castellano, R. H. G. Wright, C. Ballare, G. Filion, M. A. Marti-Renom & M. Beato. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.* **28**, 2151–2162 (2014).

[256] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gulsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren & D. M. Gilbert. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).

[257] L. Liu, Y. Zhang, J. Feng, N. Zheng, J. Yin & Y. Zhang. Gesica: genome segmentation from intra-chromosomal associations. *BMC Genomics* **13**, 164 (2012).

[258] D. Filippova, R. Patro, G. Duggal & C. Kingsford. Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.* **9**, 14 (2014).

[259] C. Levy-Leduc, M. Delattre, T. Mary-Huard & S. Robin. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30**, i386–i392 (2014).

[260] J. H. Gibcus & J. Dekker. The hierarchy of the 3D genome. *Mol. Cell* **49**, 773–782 (2013).

[261] J. A. Bogan, D. A. Natale & M. L. Depamphilis. Initiation of eukaryotic DNA replication: conservative or liberal? *J. Cell. Physiol.* **184**, 139–150 (2000).

[262] M. Méchali. DNA replication origins: from sequence specificity to epigenetics. *Nat. Rev. Genet.* **2**, 640–645 (2001).

[263] A. J. McNairn & D. M. Gilbert. Epigenomic replication: linking epigenetics to DNA replication. *Bioessays* **25**, 647–656 (2003).

[264] M. I. Aladjem. Replication in context: dynamic regulation of DNA replication patterns in metazoans. *Nat. Rev. Genet.* **8**, 588–600 (2007).

[265] S. Courbet, S. Gay, N. Arnoult, G. Wronka, M. Anglana, O. Brison & M. Debatisse. Replication fork movement sets chromatin loop size and origin choice in mammalian cells. *Nature* **455**, 557–560 (2008).

[266] N. Biggs, E. Lloyd & R. Wilson. *Graph Theory, 1736-1936* (Oxford University Press, 1986).

[267] J. R. Neuman. *The World of Mathematics* (New York, Simon and Schuster, 1956).

[268] L. Akoglu & C. Faloutsos. Event detection in time series of mobile communication graphs. *Proc. of Army Science Conference* (2010).

[269] N. L. Parasanna. Applications of graph labeling in communication networks. *Orient. J. Comp. Sci. and Technol* **7(1)** (2014).

[270] P. Morgado & N. Costa. Graph-based model to transport networks analysis through GIS. In *Proceedings of European Colloquium on Quantitative and Theoretical Geography*, (2005).

[271] J. C. Johnson, S. P. Borgatti, J. J. Luczkovich & M. G. Everett. Network role analysis in the study of food webs: An application of regular role coloration. *J. Soc. Structure* **2** (2001).

[272] C. J. Stam & J. C. Reijneveld. Graph theoretical analysis of complex networks in the brain. *Nonlinear Biomed. Phys.* **1**, 3 (2007).

[273] J. A. Bondy & U. S. R. Murty. *Graph Theory* (Springer, 2008).

[274] R. Diestel. *Graph Theory* (Springer, 2010).

[275] S. Fortunato. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).

[276] M. Sales-Pardo, R. Guimera, A. Moreira & L. Amaral. Extracting the hierarchcal organization of complex systems. *PNAS* **104 (39)**, 152224–15229 (2007).

[277] G. D. Battista, P. Eades, R. Tamassia & I. G. Tollis. Algorithms for drawing graphs: an annotated bibliography. *Computational Geometry* **4**, 235–282 (1994).

[278] T. M. J. Fruchterman & E. M. Reingold. Graph drawing by force-directed placement. *Softw: Pract. Exper.* **21**, 1129–1164 (1991).

[279] S. G. Kobourov. Spring embedders and force directed graph drawing algorithms. *Handbook of Graph Drawing and Visualization* , p. 349–381 (2013).

[280] J. Sun & J. Tang. A survey of models and algorithms for social influence analysis. *Social Network Data Analytics* , 177–214 (2011).

[281] J. M. Bolland. Sorting out centrality: an analysis of the performance of four centrality models in real and simulated networks. *Social Networks* **10 (3)**, 233–253 (1988).

[282] R. Rothenberg, J. J. Potterat, D. E. Woodhouse, W. W. Darrow, S. Q. Muth & A. S. Klovdahl. Choosing a centrality measure: Epidemiologic correlates in the colorado springs study of social networks. *Soc. Net.* **17**, 273–297 (1995).

[283] K. Faust. Centrality in affiliation networks. *Soc. Netw.* **19**, 157–191 (1997).

[284] C.-Y. Lee. Correlations among centrality measures in complex networks. (2007).

[285] T. W. Valente, K. Coronges, C. Lakon & E. Costenbader. How correlated are network centrality measures? *Connect. (Tor)* **28**, 16–26 (2008).

[286] A. H. Dekker. Centrality in social networks: Theoretical and simulation approaches. In *Proceedings of SimTecT*, (2008), p. 33–38.

[287] K. Batool & M. A. Niazi. Towards a methodology for validation of centrality measures in complex networks. *PLoS One* **9**, e90283 (2014).

[288] A. Landherr, B. Friedl & J. Heidemann. A critical review of centrality measures in social networks. *Business & Information Systems Engineering* **2**, 371–385 (2010).

[289] L. S. Freeman. Centrality in social networks conceptual clarification. *Social Networks* **1**, 215 (1979).

[290] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *J. Math. Sociol.* **2**, 113 (1972).

[291] N. E. Friedkin. Theoretical foundations for centrality measures. *American J. of Sociology* **96**, 1478–1504 (1991).

[292] S. Hoory, N. Linial & A. Wigderson. Expander graphs and their applications. *Bullet. Amer. Math. Soc.* **43**, 439–561 (2006).

[293] D. Spielman. Spectral graph theory. In *Combinatorial Scientific Computing,* edited by Chapman & H. . C. Press, (2012).

[294] U. von Luxburg. A tutorial on spectral clustering. *Stat. Comput.* **17 (4)**, 395–416 (2007).

[295] D. Cvetkovic & I. Gutman. Selected topics on applications of graph spectra. *Zbornik radova* **14 (22)** (2011).

[296] A. Sandryhaila & J. M. F. Moura. Discrete signal processing on graphs. *IEEE Transactions on Signal Processing* , 1644–1656 (2012).

[297] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega & P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine* **30**, 83–98 (2013).

[298] F. R. K. Chung. *Spectral Graph Theory* (American Mathematical Society, 1997).

[299] P. Davis. *Circulant Matrices* (John Wiley and Sons, 1979).

[300] S. Mallat. *A Wavelet Tour of Signal Processing* (Academic Press, New York, 1998).

[301] R. Coifman & M. Maggioni. Diffusion wavelets. *Appl. and Comput. Harmon. Anal.* **21 (1)**, 53–94 (2006).

[302] M. Crovella & E. Kolaczyk. Graph wavelets for spatial traffic analysis. In *INFOCOM 2003.,* edited by I. Societies, volume 3. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications., (2003), p. 1848–1857.

[303] M. Jansen, G. P. Nason & B. W. Silverman. Multiscale methods for data on graphs and irregular multidimensional situations. *J. of the Roy. Stat. Soc.: Series B (Statistical Methodology)* **71 (1)**, 97–125 (2009).

[304] S. Narang & A. Ortega. Lifting based wavelet transforms on graphs. In *Proc. of APSIPA Annual Summit and Conference*. APSIPA ASC, (2009).

[305] S. Narang & A. Ortega. Perfect reconstruction two-channel wavelet filter banks for graph structured data. *IEEE Transactions on Signal Processing* **60 (6)**, 2786–2799 (2012).

[306] V. N. Ekambaram, G. Fanti, B. Ayazifar & K. Ramchandran. Critically-sampled perfect-reconstruction spline-wavelet filterbanks for graph signals. In *IEEE GlobeSip*, (2013).

[307] D. Hammond, P. Vandergheynst & R. Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis* **30**, 129–150 (2011).

[308] N. Tremblay & P. Borgnat. Multiscale community mining in networks using spectral graph wavelets. *ArXiv e-prints* , 1212.0689 (2012).

[309] N. Tremblay & P. Borgnat. Graph wavelets and community mining. *IEEE Transactions on Signal Processing* **62(20)**, 5227–5239 (2014).

[310] D. I. Shuman, C. Wiesmeyr, N. Holighaus & P. Vandergheynst. Spectrum-adapted tight graph wavelet and vertex-frequency frames. *Accepted for publication in IEEE Transactions on Signal Processing* (2015).

[311] J. Leskovec & C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, (2006), p. 631–636.

[312] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal* **23 (2)**, 298–305 (1973).

[313] M. Girvan & M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99 (12)**, 7821–7826 (2002).

[314] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (2003).

[315] M. E. J. Newman & M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).

[316] S. Redner. How popular is your paper? an empirical study of the citation distribution. *Eur.Phys.J.B* **4**, 131–134 (1998).

[317] A. J. Enright, S. Van Dongen & C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

[318] P. K. Reddy, M. Kitsuregawa, P. Sreekanth & S. S. Rao. A graph based approach to extract a neighborhood customer community for collaborative filtering. In *DNIS'02 Proceedings of the Second International Workshop on Databases in Networked Information Systems* (Springer-Verlag, London, UK, 2002), p. 188–200.

[319] S. A. Rice. The identification of blocs in small political bodies. *Am. Sociol. Rev.* **21**, 619–627 (1927).

[320] R. S. Weiss & E. Jacobson. A method for the analysis of the structure of complex organizations. *Am. Sociol. Rev.* **20**, 661–668 (1955).

[321] G. C. Homans. The human group. *Routledge* **7** (2013).

[322] L. R. Ford & D. R. Fulkerson. *Flows in networks*, edited by Princeton (Princeton University Press, 1962).

[323] B. W. Kernighan & S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal* **49 (2)**, 291–307 (1970).

[324] V. Blondel, J. Guillaume, R. Lambiotte & E. Lefebvre. Fast unfolding of communities in large networks. *J. of Stat. Mech.: Theory and Experiment* **10** (2008).

[325] A. Pothen. *Graph Partitioning Algorithms With Applications To Scientific Computing* (Kluwer, 1997).

[326] T. Hastie, R. Tibshirani & J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2009).

[327] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, (1967).

[328] A. Hlaoui & S. Wang. A direct approach to graph clustering. *Neural Networks Computational Intelligence* , 158–163 (2004).

[329] J. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *J. Cybernet.* **3**, 32–57 (1974).

[330] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms* (Kluwer Academic Publishers, Norwell, USA, 1981).

[331] M. A. M. Luca Donetti. Improved spectral algorithm for the detection of network communities. *Proceedings of the 8th Granada Seminar- Computational and Statistical Physics* , 1–2 (2005).

[332] J. Reichardt & S. Bornholdt. Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett.* **93**, 218701 (2004).

[333] I. Ispolatov, I. Mazo & A. Yuryev. Finding mesoscopic communities in sparse networks. *J. Stat. Mech.* **9**, p09014 (2006).

[334] H. Zhou. Distance, dissimilarity index, and network community structure. *Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys.* **67**, 061901 (2003).

[335] H. Zhou & R. Lipowsky. Dynamic pattern evolution on scale-free networks. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10052–10057 (2005).

[336] Y. Hu, M. Li, P. Zhang, Y. Fan & Z. Di. Community detection by signaling on complex networks. *Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys.* **78**, 016115 (2008).

[337] G. Palla, I. Derenyi, I. Farkas & T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).

[338] J. Hopcroft, O. Khan, B. Kulis & B. Selman. Tracking evolving communities in large linked networks. *Proc. Natl. Acad. Sci. USA* **101 Suppl 1**, 5249–5253 (2004).

[339] S. Asur, S. Parthasarathy & D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In *KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2007).

[340] S. Fortunato & M. Barthélemy. Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA* **104**, 36–41 (2007).

[341] J. Kumpula, J. Saramaki, K. Kaski & J. Kertesz. Limited resolution in complex network community detection with Potts model approach. *Eur. Phys. J. B.* **56 (1)**, 41–45 (2007).

[342] M. T. Schaub, J.-C. Delvenne, S. N. Yaliraki & M. Barahona. Markov dynamics as a zooming lens for multiscale community detection: non clique-like communities and the field-of-view limit. *PLoS One* **7**, e32210 (2012).

[343] R. Lambiotte. Multi-scale modularity in complex networks. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium*. IEEE, (2010), p. 546–553.

[344] J. Reichardt & S. Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **74**, 016110 (2006).

[345] A. Arenas, A. Fernandez & S. Gomez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics* **10 (5)** (2008).

[346] D. Gfeller, J.-C. Chappelier & P. De Los Rios. Finding instabilities in the community structure of complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **72**, 056135 (2005).

[347] B. Karrer, E. Levina & M. E. J. Newman. Robustness of community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **77**, 046119 (2008).

[348] D. L. Wallace. Comment. *J. of the Amer. Stat. Asso.* **78 (383)**, 569–576 (1983).

[349] E. B. Fowlkes & C. L. Mallows. A method for comparing two hierarchical clusterings. *J. of the Amer. Stat. Asso.* **78 (383)**, 553–569 (1983).

[350] M. R. Mesure et analyse d?un réseau social. Technical report, Rapport de stage de Licence 3, (2013).

[351] L. Hubert & P. Arabie. Comparing partitions. *J. of Classification* **2 (1)**, 193–218 (1985).

[352] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* **39**, 241–272 (1901).

[353] M. Meila. Comparing clustering–an information based distance. *Journal of Multivariate Analysis* **98 (5)**, 873–895 (2007).

[354] A. Lancichinetti, S. Fortunato & J. Kertesz. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* **11** (2009).

[355] P. Pons & M. Latapy. Post-processing hierarchical community structures : Quality improvements and multi-scale view. *Theoretical Computer Science* **412 (8)**, 892–900 (2011).

[356] C. Conti, B. Sacca, J. Herrick, C. Lalou, Y. Pommier & A. Bensimon. Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol. Biol. Cell* **18**, 3059–3067 (2007).

[357] J. Herrick & A. Bensimon. Global regulation of genome duplication in eukaryotes: an overview from the epifluorescence microscope. *Chromosoma* **117**, 243–260 (2008).

[358] A. Goldar, M.-C. Marsolier-Kergoat & O. Hyrien. Universal temporal profile of replication origin activation in eukaryotes. *PLoS One* **4**, e5899 (2009).

[359] L. Zaghloul, G. Drillon, R. E. Boulos, F. Argoul, C. Thermes, A. Arneodo & B. Audit. Large replication skew domains delimit GC-poor gene deserts in human. *Comput. Biol. Chem.* **53**, 153–165 (2014).

[360] G. Bernardi. The isochore organization of the human genome. *Annu. Rev. Genet.* **23**, 637–661 (1989).

[361] M. Denholtz, G. Bonora, C. Chronis, E. Splinter, W. de Laat, J. Ernst, M. Pellegrini & K. Plath. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell* **13**, 602–616 (2013).

[362] E. de Wit, B. A. M. Bouwman, Y. Zhu, P. Klous, E. Splinter, M. J. A. M. Verstegen, P. H. L. Krijger, N. Festuccia, E. P. Nora, M. Welling, E. Heard, N. Geijsen, R. A. Poot, I. Chambers & W. de Laat. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**, 227–231 (2013).

[363] O. Hyrien. Peaks cloaked in the mist: The landscape of mammalian replication origins. *J. Cell Biol.* **208**, 147–160 (2015).

[364] S. I. Takebayashi, V. Dileep, T. Ryba, J. H. Dennis & D. M. Gilbert. Chromatin-interaction compartment switch at developmentally regulated chromosomal domains reveals an unusual principle of chromatin folding. *Proc. Natl. Acad. Sci. USA* **109**, 12574–12579 (2012).

[365] G. Drillon, B. Audit, F. Argoul & A. Arneodo. Ubiquitous human 'master' origins of replication are encoded in the DNA sequence via a local enrichment in nucleosome excluding energy barriers. *J. Phys. Condens. Matter* **27**, 064102 (2015).

[366] A. Rosa & R. Everaers. Structure and dynamics of interphase chromosomes. *PLoS Comput. Biol.* **4**, e1000153 (2008).

[367] P. R. Cook & D. Marenduzzo. Entropic organization of interphase chromosomes. *J. Cell Biol.* **186**, 825–834 (2009).

[368] L. A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* **19**, 37–51 (2011).

[369] T. Cremer & C. Cremer. Chromosome Territories, Nuclear Architecture and gene Regulation in Mammalian Cells. *Nat. Rev. Genet.* **2**, 292–301 (2001).

[370] J. Dekker, M. A. Marti-Renom & L. A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390–403 (2013).

[371] P. de Gennes. *Scaling Concepts in Polymer Physics* (Cornell University Press Ithaca, 1979).

[372] A. Grosberg & A. R. Khoklov. *Statisctical Physics of Macromolecules. AIP, New York* (1994).

[373] S. Havlin & D. Ben-Arraham. Diffusion in disorderes media. *Advances in Physics* **51**, 187–292 (2002).

[374] K. Bystricky, P. Heun, L. Gehlen, J. Langowski & S. M. Gasser. Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. *Proc. Natl. Acad. Sci. USA* **101**, 16495–16500 (2004).

[375] H. Yokota, G. van den Engh, J. E. Hearst, R. K. Sachs & B. J. Trask. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. *J. Cell Biol.* **130**, 1239–1249 (1995).

[376] A. Grosberg, Y. Rabin, S. Havlin & A. Neer. Crumpled globule model of the three-dimensional structure of DNA. *Europhys. Lett.* **23**, 373–378 (1993).

[377] A. S. Belmont, Y. Hu, P. B. Sinclair, W. Wu, Q. Bian & I. Kireev. Insights into interphase large-scale chromatin structure from analysis of engineered chromosome regions. *Cold Spring Harb. Symp. Quant. Biol.* **75**, 453–460 (2010).

[378] W. A. Bickmore & B. van Steensel. Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270–1284 (2013).

[379] F. Ciabrelli & G. Cavalli. Chromatin-driven behavior of topologically associating domains. *J. Mol. Biol.* **in press**, 10.1016/j.jmb.2014.09.013 (2014).

[380] A. Vologodskii. Theoretical models of DNA topology simplification by type IIA DNA topoisomerases. *Nucleic Acids Res.* **37**, 3125–3133 (2009).

[381] A. Rosa & R. Everaers. Ring polymers in the melt state: the physics of crumpling. *Phys. Rev. Lett.* **112**, 118302 (2014).

[382] A. Khalil, J. L. Grant, L. B. Caddle, E. Atzema, K. D. Mills & A. Arneodo. Chromosome territories have a highly nonspherical morphology and nonrandom positioning. *Chromosome Res.* **15**, 899–916 (2007).

[383] T. Cremer & M. Cremer. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2**, a003889 (2010).

[384] M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo & M. Nicodemi. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. USA* **109**, 16173–16178 (2012).

[385] G. Fudenberg, G. Getz, M. Meyerson & L. A. Mirny. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.* **29**, 1109–1113 (2011).

[386] S. Holwerda & W. de Laat. Chromatin loops, gene positioning, and gene expression. *Front. Genet.* **3**, 217 (2012).

[387] O. K. Smith & M. I. Aladjem. Chromatin structure and replication origins: determinants of chromosome replication and nuclear organization. *J. Mol. Biol.* **426**, 3330–3341 (2014).

[388] C. Renard-Guillet, Y. Kanoh, K. Shirahige & H. Masai. Temporal and spatial regulation of eukaryotic DNA replication: from regulated initiation to genome-scale timing program. *Semin. Cell Dev. Biol.* **30**, 110–120 (2014).

[389] D. Jost, P. Carrivain, G. Cavalli & C. Vaillant. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res.* **42**, 9553–9561 (2014).

[390] L. Chakalova, E. Debrand, J. A. Mitchell, C. S. Osborne & P. Fraser. Replication and transcription: shaping the landscape of the genome. *Nat. Rev. Genet.* **6**, 669–677 (2005).

[391] Q. Ye, I. Callebaut, A. Pezhman, J. C. Courvalin & H. J. Worman. Domain-specific interactions of human HP1-type chromodomain proteins and inner nuclear membrane protein LBR. *J. Biol. Chem.* **272**, 14983–14989 (1997).

[392] D. Peric-Hupkes, W. Meuleman, L. Pagie, S. W. M. Bruggeman, I. Solovei, W. Brugman, S. Graf, P. Flicek, R. M. Kerkhoven, M. van Lohuizen, M. Reinders, L. Wessels & B. van Steensel. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38**, 603–613 (2010).

[393] J. M. Zullo, I. A. Demarco, R. Pique-Regi, D. J. Gaffney, C. B. Epstein, C. J. Spooner, T. R. Luperchio, B. E. Bernstein, J. K. Pritchard, K. L. Reddy & H. Singh. DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. *Cell* **149**, 1474–1487 (2012).

[394] I. Solovei, A. S. Wang, K. Thanisch, C. S. Schmidt, S. Krebs, M. Zwerger, T. V. Cohen, D. Devys, R. Foisner, L. Peichl, H. Herrmann, H. Blum, D. Engelkamp, C. L. Stewart, H. Leonhardt & B. Joffe. LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell* **152**, 584–598 (2013).

[395] W. Meuleman, D. Peric-Hupkes, J. Kind, J.-B. Beaudry, L. Pagie, M. Kellis, M. Reinders, L. Wessels & B. van Steensel. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Res.* **23**, 270–280 (2013).

[396] T. R. Luperchio, X. Wong & K. L. Reddy. Genome regulation at the peripheral zone: lamina associated domains in development and disease. *Curr. Opin. Genet. Dev.* **25**, 50–61 (2014).

[397] A. Bancaud, S. Huet, N. Daigle, J. Mozziconacci, J. Beaudouin & J. Ellenberg. Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. *EMBO J.* **28**, 3785–3798 (2009).

[398] A. Bancaud, C. Lavelle, S. Huet & J. Ellenberg. A fractal model for nuclear organization: current evidence and biological implications. *Nucleic Acids Res.* **40**, 8783–8792 (2012).

[399] S. Boyle, S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis & W. A. Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10**, 211–219 (2001).

[400] R. Berezney, D. D. Dubey & J. A. Huberman. Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma* **108**, 471–484 (2000).

[401] P. Hozak, A. B. Hassan, D. A. Jackson & P. R. Cook. Visualization of replication factories attached to nucleoskeleton. *Cell* **73**, 361–373 (1993).

[402] N. Rhind & D. M. Gilbert. DNA replication timing. *Cold Spring Harb Perspect. Biol.* **5**, a010132 (2013).

[403] E. Guillou, A. Ibarra, V. Coulon, J. Casado-Vela, D. Rico, I. Casal, E. Schwob, A. Losada & J. Mandez. Cohesin organizes chromatin loops at DNA replication factories. *Genes Dev.* **24**, 2812–2822 (2010).

[404] S. Martins, S. Eikvar, K. Furukawa & P. Collas. HA95 and LAP2 beta mediate a novel chromatin-nuclear envelope interaction implicated in initiation of DNA replication. *J. Cell. Biol.* **160**, 177–188 (2003).

[405] R. E. Boulos, A. Arneodo, P. Jensen & B. Audit. Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. *Phys. Rev. Lett.* **111**, 118102 (2013).

[406] M. Bastian, S. Heymann & M. Jacomy. Gephi: An open source software for exploring and manipulating networks. Third International AAAI Conference on Weblogs and Social Media. Available at: http://www.aaai.org/ocs/index.php/ICWSM/09/-paper/view/154, (2009).

[407] J. Kind & B. van Steensel. Stochastic genome-nuclear lamina interactions: modulating roles of lamin A and BAF. *Nucleus* **5**, 124–130 (2014).

[408] R. E. Boulos, H. Julienne, A. Baker, C.-L. Chen, N. Petryk, M. Kahli, Y. d'Aubenton-Carafa, A. Goldar, P. Jensen, O. Hyrien, C. Thermes, A. Arneodo & B. Audit. From the chromatin interaction network to the organization of the human genome into replication N/U-domains. *New J. Phys.* **16**, 115014 (2014).

[409] R.-J. Palstra, B. Tolhuis, E. Splinter, R. Nijmeijer, F. Grosveld & W. de Laat. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat. Genet.* **35**, 190–194 (2003).

[410] F. Bantignies, V. Roure, I. Comet, B. Leblanc, B. Schuettengruber, J. Bonnet, V. Tixier, A. Mas & G. Cavalli. Polycomb-dependent regulatory contacts between distant Hox loci in *Drosophila*. *Cell* **144**, 214226 (2011).

[411] S. Schoenfelder, T. Sexton, L. Chakalova, N. F. Cope, A. Horton, S. Andrews, S. Kurukuti, J. A. Mitchell, D. Umlauf, D. S. Dimitrova, C. H. Eskiw, Y. Luo, C.-L. Wei, Y. Ruan, J. J. Bieker & P. Fraser. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **42**, 53–61 (2010).

[412] Y. Shavit & P. Lio'. Combining a wavelet change point and the bayes factor for analysing chromosomal interaction data. *Mol. Biosyst.* **10**, 1576–1585 (2014).

[413] T. Sexton & G. Cavalli. The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049–1059 (2015).

[414] C. Vaillant, B. Audit & A. Arneodo. Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys. Rev. Lett.* **99**, 218103 (2007).

[415] V. Miele, C. Vaillant, Y. d'Aubenton-Carafa, C. Thermes & T. Grange. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.* **36**, 3746–3756 (2008).

[416] T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert & B. F. Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18**, 1073–1083 (2008).

[417] J. N. Burton, I. Liachko, M. J. Dunham & J. Shendure. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda)* **4**, 1339–1346 (2014).

[418] D. Hanahan & R. A. Weinberg. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).

[419] S. Ruiz, A. D. Panopoulos, A. Herreras, K.-D. Bissig, M. Lutz, W. T. Berggren, I. M. Verma & J. C. Izpisua Belmonte. A high proliferation rate is required for cell reprogramming and maintenance of human embryonic stem cell identity. *Curr. Biol.* **21**, 45–52 (2011).

[420] I. Cantone & A. G. Fisher. Epigenetic programming and reprogramming during development. *Nat. Struct. Mol. Biol.* **20**, 282–289 (2013).

[421] B. D. Pope & D. M. Gilbert. The replication domain model: regulating replicon firing in the context of large-scale chromosome architecture. *J. Mol. Biol.* **425**, 4690–4695 (2013).

[422] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay & P. Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).

[423] J. Kind, L. Pagie, H. Ortabozkoyun, S. Boyle, S. S. de Vries, H. Janssen, M. Amendola, L. D. Nolen, W. A. Bickmore & B. van Steensel. Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178–192 (2013).

[424] S. De & F. Michor. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotechnol.* **29**, 1103–1108 (2011).

[425] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. Mc Henry, R. M. Pinchback, A. H. Ligon, Y.-J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M.-S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers & M. Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).

[426] P. Goupillaud, A. Grossmann & J. Morlet. Cycle-octave and related transforms in seismic signal analysis. *Geoexploration* **23**, 85–102 (1984).

[427] A. Grossmann & J. Morlet. Decomposition of Hardy functions into square integrable wavelets of constant shape. *S.I.A.M. J. of Math. Anal.* **15**, 723–736 (1984).

[428] A. Grossmann & J. Morlet. Decomposition of functions into wavelets of constant shape and related transforms. In *Mathematics and Physics, Lectures on Recent Results,* edited by L. Streit (World Scientific, Singapore, 1985), p. 17–22.

[429] A. Arneodo, B. Audit, E.-B. Brodie of Brodie, S. Nicolay, M. Touchon, Y. d'Aubenton-Carafa, M. Huvet & C. Thermes. Fractals and wavelets: what can we learn on transcription and replication from wavelet-based multifractal analysis of DNA sequences? In *Encyclopedia of Complexity and Systems Science,* edited by R. A. Meyers, 3893–3924. Springer, New York (2009).

[430] S. Nicolay, E. B. Brodie of Brodie, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes & A. Arneodo. Bifractality of human DNA strand-asymmetry profiles results from transcription. *Phys. Rev. E* **75**, 032902 (2007).

[431] G. Samorodnisky & M. S. Taqqu. *Stable Non-Gaussian Random Processes* (Chapman and Hall, New York, 1994).

[432] J.-F. Muzy, E. Bacry & A. Arneodo. The multifractal formalism revisited with wavelets. *Int. J. Bifurc. Chaos* **4**, 245–302 (1994).