Génome-wide characterization and comparative analysis of the DNA replication program in 12 human cell lines

Hadi Kabalane


Numéro National de Thèse : 2019LYSEN063

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée par
l'École Normale Supérieure de Lyon

École Doctorale N°52
École Doctorale de Physique et Astrophysique de Lyon (PHAST)

Spécialité de doctorat : Biostatistique et Bioinformatique
Discipline : Physique

Soutenue publiquement le 18/11/2019, par :

Hadi KABALANE

Caractérisation pangénomique et analyse comparative du programme de réplication de l'ADN dans 12 lignées cellulaires humaines

---

Devant le jury composé de :

Chen, Chunlong	CR, Institut Curie, Paris	Rapporteur
Letouzé, Eric	CR, Centre de Recherche des Cordeliers, Paris	Rapporteur
Argoul, Françoise	DR, Laboratoire Ondes et Matière d'Aquitaine, Bordeaux	Examinatrice
Cayrou, Christelle	CR, Centre de Recherche en Cancérologie de Marseille	Examinatrice
Hyrien, Olivier	DR, Institut de Biologie de l'ENS Paris	Examinateur
Audit, Benjamin	DR, Laboratoire de Physique, ENS de Lyon	Directeur de thèse

Remerciements/Acknowledgments

Je profite de ces quelques lignes pour remercier tous ceux qui m'ont accompagné ces trois dernières années et sans qui ce manuscrit n'aurait jamais pu être réalisé.

Je tiens à remercier Benjamin Audit, pour m'avoir aidé et encouragé tout au long de ma thèse. Merci pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Je suis très reconnaissant a Chunlong Chen et Eric Letouzé de leur intérêt pour ce travail de thèse et d'avoir accepté de lire et d'évaluer mon manuscrit. Je souhaite remercier les examinateurs de mon jury, Olivier Hyrien, Françoise Argoul et Christelle Cayrou de leur venue et de leurs nombreuses questions pertinentes lors de ma soutenance de thèse. Grâce à eux, ma soutenance fut extrêmement enrichissante scientifiquement.

Je remercie aussi tou nos collaborateurs à l'IBENS, Paris et au LOMA; Bordeaux: Xia Wu pour les données Ok-seq, Benoit Le Tallec, Magali Hennion, Laurent Lacroix, Elizaveta Novikova, Bertrand Theulot pour toutes les discussions durant les réunions ANR. Un grand merci à Alain Arneodo dont j'ai eu la chance au cours de cette thèse de profiter de sa grande clarté et de sa rigueur scientifique sur tous les domaines de la physique à la bioinformatique qui m'ont aidé à réaliser mes analyses et à comprendre mes résultats.

Je remercie tous les membres du Laboratoire de Physique de l'ENS de Lyon représenté par son directeur Dr. Thierry Dauxois de m'avoir m'accueilli au sein du Laboratoire, et surtout les membres de l'équipe Sisyphe. Plus particulièrement aux doctorants, post doctorants et stagiaires qui sont passés au Laboratoire : Jean-Michel, Barbara, Marion, Jérémy, Yacouba, Vincent, Marcelo, Laurence, Roberto, Carlos, Eric, Salambo, Alex, Lavi.

Je voudrais exprimer ma reconnaissance envers les amis qui m'ont apporté leur soutien moral et intellectuel tout au long de ma démarche: Orsola, Soha, Sarah, Haydar, Zahraa, Riva, Batoul, Zeinab, Abed el salam, Hiba, Hussein, Ashraf et Ragheb pour les beaux moments, voyages, soirée et discussion. Merci à mes amis du Liban: George et Mona pour les beaux moment à Beyrouth, Bikfaya et mon village; Ahmad, Ali R, Ali H, Houssam, Jad, Ibrahim, Hassan, Abed pour les soirée à Beyrouth avec tous les débats intenses sur tous les sujets, Mohammad, Siraj et Hamza pour les beaux moments dans mon village. Je remercie d'une façon plus générale tous ceux qui m'ont accompagné pendant cette periode.

Et pour finir un grand merci à ma mère et mon père, qui m'ont donné l'opportunité il y'a 4 ans de suivre des études en master à Lyon, ainsi que leur soutien inconditionnel qui m'a permis de poursuivre mes études en thése. A mes frères (Hassan et Ihab) et à ma soeur Jana, merci pour les beaux moments qu'on passe à chaque fois que je retourne au Liban et pour le soutien tout au long de ma thèse.

Résumé

Le programme spatiotemporel de réplication de l'ADN est régulé au cours du développement et altéré durant la progression cancéreuse. Nous proposons une caractérisation originale de la plasticité du programme de réplication de l'ADN se basant sur le profilage de 12 lignées cellulaires humaines normales ou cancéreuses par la méthode OK-seq de purification et séquençage des fragments d'Okazaki qui permet de déterminer l'orientation de la progression des fourches de réplication (OFR) à haute résolution (10 kilo bases). L'analyse comparative des profils OFR montre que les changements réplicatifs permettent la classification des lignées cellulaires en fonction de leur tissu d'origine, la nature cancéreuse ou non de la lignée n'intervenant qu'en second ordre. Il n'apparait pas de point chaud pour l'accumulation des changements réplicatifs, ceux-ci étant largement dispersés sur tout le génome. Néanmoins, les régions riches en G+C et en gènes actifs, répliquées précocement au cours de la phase S, ont le programme de réplication le plus stable, elles présentent une forte densité de zones d'initiation de la réplication (ZI) efficaces et conservées entre lignées cellulaires. En contraste, les dernières régions répliquées, à faible densité de gènes et pauvres en G+C, présentent peu de ZI efficaces, souvent spécifiques d'un tissu ou d'une lignée. Ceci nous conduit à quantifier le degré de dissociation entre ZI et activation de la transcription. Ce travail propose un panorama original des modifications du programme de réplication au cours de la différentiation normale ou pathologique, dont un contrôle lignée cellulaire spécifique des ZI dans les déserts de gènes à réplication tardive.

Abstract

The spatiotemporal program of DNA replication is regulated during development and altered during cancer progression. We propose an original characterization of the plasticity of the DNA replication program based on the profiling of 12 normal or cancerous human cell lines by the OK-seq method of purification and sequencing of Okazaki fragments which allows to determine the orientation of the progression of replication forks (RFD) at high resolution (10 kilo bases). Comparative analysis of the RFD profiles shows that the replicative changes allow the classification of the cell lines according to their tissue of origin, the cancerous or non-cancerous nature of the cell line intervening only in second order. There is no hotspot for the accumulation of replicative changes, they are widely dispersed throughout the genome. Nevertheless, the G+C rich and active gene regions, replicated early in the S phase, have the most stable replication program, they present a high density of efficient replication initiation zones (IZ) conserved between cell lines. In contrast, the late replicated, low gene density and low G+C content regions have few efficient IZs, often specific to a tissue or lineage. This leads us to quantify the degree of dissociation between IZ and activation of transcription. This work provides an original overview of replication program changes during normal or pathological differentiation, including a cell line specific control of IZ in late-replication gene deserts.

Contents

CHAPTER 1

Introduction

Life depends on the cells ability to preserve, retrieve and translate the genetic instructions needed to build a living organism and keep it alive. This hereditary information is transmitted from a cell to its daughter cells at the time of cell division and, for one organism, from one generation to another by the reproductive cells of this organism. In each living cell, these instructions are carried by the genes, the elements containing the information which determines the characteristics of a species as well as those specific to each individual.

The biological carriers of the genetic information are deoxyribonucleic acid (DNA) molecules. Hence, DNA replication, the duplication of these molecules, is an essential mechanism for the living beings. This mechanism must be as faithful as possible, so that the genetic heritage that will be transmitted to the daughter cells is almost perfectly identical to that of the mother cell. Indeed, the short-term survival of an organism depends on the prevention of modification of its genetic content. DNA modifications such as mutations can be extremely deleterious events, associated with pathologies like cancer. However, there may be benefic mutations responsible for the long term evolution of the species. During the S phases of cell cycle where DNA replication occurs in eukaryotes, the genome is particularly vulnerable and many types of lesions can block the progression of the replication machinery. In fact, prolonged stopping of the replication machinery can lead to its collapse, double-strand breaks, and genetic instability. Stabilization and backup of blocked machinery and finally the completion of replication are essential for cell survival and genome preservation [1, 2].

In this context, the present work aim was to contribute to the better understanding of DNA replication process in human by identifying and characterizing its modifications observed between 12 human cell lines. This work is based on the recent development of a profiling protocol allowing to measure genome-wide the so-called Replication Fork Directionality (RFD)

[3]. In this introductory chapter, we will discuss how the discovery of the structure of DNA led to the modeling of the mechanism of DNA replication, and the current experimental methods to study DNA replication. Finally, we will describe the association between the organization of DNA replication and other genomic features.

1.1 Deoxyribonucleic acid structure and copying mechanism

In the 1940s, there has been a vivid devate among biologists to recognize DNA as the preincipal carrier of genetic information. This long polymer macromolecule was considered as too simple, consisting only of four types of chemically similar nucleotides. These nucleotides are composed of a phosphate group (or phosphoric acid), an aldopentose (2'-deoxy-D-ribose) and a nitrogen heterocyclic base, either one of two pyrimidines: cytosine (C) and thymine (T), or two purines: adenine (A) and guanine (G) [4]. In the early 1950s, for the first time, DNA was analyzed by X-ray diffraction, a technique that makes it possible to determine the three-dimensional atomic structure of a molecule. The first results of the X-ray diffraction showed that the DNA was very likely constituted of two co-axial molecules wound into a helix [5]. The fact that DNA is double-stranded constitutes one of the main hypothesis that led to the model of DNA structure published in 1953 by Watson and Crick [6]. In this paper, they described a structure based on specific base pairing of adenine with thymine and guanine with cytosine at the center of a double helix and holding together two single strands of DNA. The sugar-phosphate backbone of each strand of a DNA double helix has a unique chemical direction, or polarity, determined by the manner in which each sugar residue is bonded to the next, and the two strands of the double helix are anti-parallel [5–7]; that is, they go in opposite directions. Specific base complementarity between the two DNA stands implies that each strand contains all the information to reconstruct the whole double-stranded molecule. At the very end of this brief scientific blockbuster, Watson and Crick comment almost in an aside: *"It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material"*. The structure of DNA would allow for its exact duplication, an essential property required for the genetic material.

One month after their historical article in Nature, Watson and Crick published a second article [7], suggesting how DNA could be duplicated. In this article, they proposed that each strand serves as a model for the synthesis of a complementary daughter strand. For this duplication to occur, the two parental strands must separate, requiring a rupture of the hydrogen bonds between paired bases, and unfolding of the paired strands. In this model, called semi-conservative replication, each new molecule of DNA is composed of a strand derived from the original parent molecule and a newly synthesized strand (Figure 1.1). The discovery of the structure of DNA immediately suggested responses on how the information needed to define an organism can be duplicated and copied from generation to generation.

The semiconservative model indeed describes the main steps of DNA replication *in vivo*.

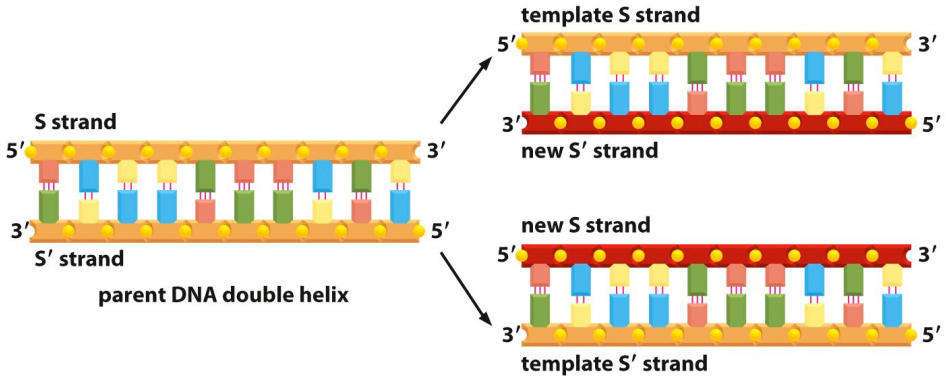

Figure 1.1: **Semiconservative model of DNA replication**. DNA replication, also known as DNA duplication, is the process in which DNA is synthesized *in vivo*. This mechanism makes it possible to obtain, from one DNA molecule, two molecules identical to the initial molecule, for distribution to the two daughter cells during cell division. Because nucleotide A will successfully pair only with T, and G with C, each strand of a DNA double helix — labeled here as the S strand and its complementary S' strand — can serve as a template to specify the sequence of nucleotides in its complementary strand. In this way, both strands of a DNA double helix can be copied precisely. In this model, the two new molecules are formed by one original (template) strand and a newly synthesized strand, this model is termed semiconservative. (Adapted from Essential Cell Biology fourth edition [8]).

However, it only captures part of the complexity involved in regulating the faithful replication of a complete genome, like the human genome made of 6.6 billion nucleotides distributed on 23 pairs of chromosomes, which requires further levels of complexity. According to the so-called "replicon" model proposed in 1963 by François Jacob, François Cuzin and Sydney Brenner [9], replication of a chromosome starts with the binding of some "initiator" protein to a specific "replicator" DNA sequence called *origin of replication*. The replicon model is the paradigm of the replication of circular, prokaryotic[1] chromosomes having a single replication origin (Figure 1.2). The initiator-replicator association makes it possible to open the DNA locally and trigger replication. The recruitment of additional factors initiates the bi-directional progression along the chromosome of two divergent *replication forks* each assuring the simultaneous polymerization of the two new DNA strands as predicted by the semiconservative replication model (Figure 1.2). This forms a "replication bubble" until the two meets at the *termination site*, opposite from the origin on the chromosome (Figure 1.2).

[1] A prokaryote is a living being whose cell structure does not contain a nucleus, and almost never any membranous organelles

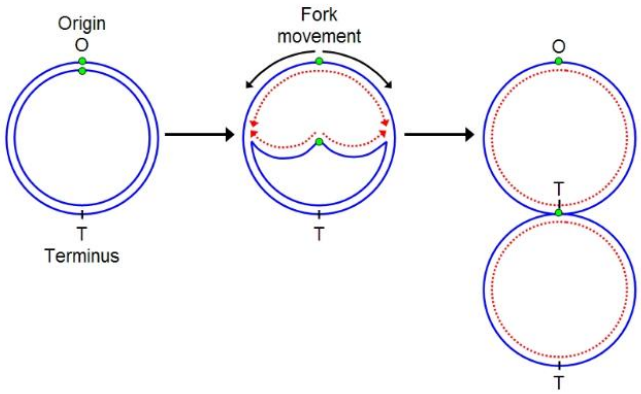

Figure 1.2: **Semiconservative model of bacterial DNA replication.** The circular double-stranded DNA chromosome of most bacteria is duplicated by bidirectional synthesis. It begins near a special site on the chromosome, called the origin (O) where the DNA opens up as hydrogen bonds between the bases in the complementary strands are enzymically broken and the strands separate, forming a replication 'bubble' between two DNA forks as the two strands separate on either side. The DNA continues unzipping at these forks, in both directions simultaneously. These so-called replication forks move around the chromosome in opposite directions as the DNA is unzipped and duplicated (hence bidirectional synthesis). DNA polymerase reads the base sequence of the unzipped single strands of parental template DNA (blue lines) and matches the sequence of bases with new bases which are polymeried to create the complementary daughter strands (red dashed lines). This continues until the replication forks reach the terminus (T) at the far-side of the chromosome, opposite the origin. (Adapted from http://cronodon.com/BioTech/Bacteria_Growth.html).

The situation in eukaryotes[2] bears some resemblance with prokaryotes, we have the same process of replication, but there are several origins of replication on each chromosomes as we see in Figure 1.3. Replication of a chromosome is initiated at a number of replication origins, diverging replication forks propagate until two converging forks collide at a *terminus of replication* resulting in replication bubble mergers until chromosome complete replication (Figure 1.3) [10, 11].

One of the specificity of the Eukaryotic cells is the cell cycle during which two phases are repeated cyclically [12, 13]. Mitosis corresponding to the division of a mother cell into two daughter cells strictly identical genetically, and the interphase which is subdivided in 3 subphases: G1, S and G2. In G1 phase, the cell performs its normal metabolism and grows to a critical size triggering S phase. In the S phase, DNA replication occurs. Finally, in the G2 phase, the cellular growth is over and the centrosomes replicate, to allow the smooth course of the mitosis. For all organisms, the DNA replication during S phase is subdivided into three stages: the ***initiation*** which, through the opening of the DNA at the level of the origins, allows the recruitment of the polymerization machinery and the start of the DNA synthesis, the ***elongation*** which corresponds to the continuation of the synthesis, and the ***termination*** which consists in the fusion of two convergent forks.

[2]The eukaryotes gather four kingdoms of the living world: animals, mushrooms, plants and protists which are characterized by the presence of a nucleus

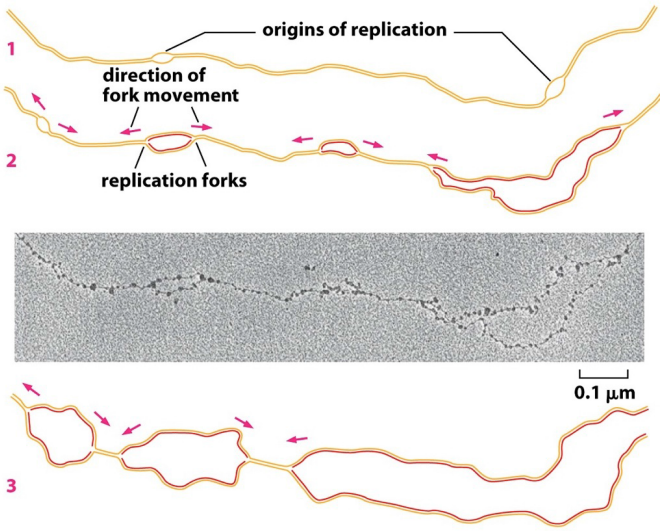

Figure 1.3: **Replication of eukaryotic chromosomes.** The electron micrograph shows DNA replicating along a portion of chromosome of an early fly embryo. The particles visible along the DNA are nucleosomes, structures made of $\sim$ 150 bp of DNA wrapped around a protein complexe. The drawings represent the same portion of chromosome as it might appear at different times during replication: (1) prior to the time of the micrograph; (2) at the time of the micrograph; (3) after the time of the micrograph and after the merger of the two rightmost replication bubbles despicted in (2). The orange lines represent the two parental DNA strands; the red lines represent the newly synthesized DNA strands. (Adapted from Essential Cell Biology fourth edition [8]).

.

1.1.1 Replication initiation

Experiments carried out using unicellular eukaryotic models, such as the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, made it possible to identify the protein actors that act at the level of the origins of replication to initiate the synthesis of DNA [14–16]. The activity of these initiation proteins is finely regulated by specific proteins, such as kinases, to ensure a unique duplication of the genome at each cell cycle and thus guarantees its stability. The goal of the initiation of replication is to set up a large number of factors, proteins necessary for the elongation of replication. It includes during late M and early G1 phases loading on the DNA of the ORC[3] (Origin Recognition Complex) protein complex and the assembly of the pre-replication[4] complexes (pre-RC) by recruitment of other factors proteins including a pair of helicases (mini chromosome mainteance protein complex (MCM) heterohexameter which consists of six gene products Mcm2-7) [17]. During S phase, teh pre-RC are activated by replication initiation factors including the cell division cycle protein cdc45 and the recruitment of additional factors such as DNA polymerases.

In yeast there are about 400 origins of replication distributed among the 16 chromosomes of this organism. In humans, it is estimated that there are more than 10 000 simultaneous

[3]ORC is a central component for eukaryotic DNA replication, and binds chromatin at replication origins.

[4]pre-RC is a protein complex that forms at the origin of replication during the initiation step of DNA replication. Formation of the pre-RC is required for the initiation of DNA replication to occur.

replication forks and between 30 000 to 50 000 origins [18, 19], because timely duplication of large linear chromosomes requires establishment of replication forks at multiple locations to complete replication in a typical 8 hours S-phase in human.

1.1.2 Elongation asymmetry

To complete the replication there is an essential step which is the propagation of the replication forks named elongation. Upon initiation of a replication origin, the two forks move away from the origin in opposite directions, unfolding the DNA double helix and replicating the DNA as it progresses (Figure 1.4). Forks move very fast in bacteria, about 1000 pairs of nucleotides per second, and about 10 times slower in eukaryotes. The slowness of fork movement in eukaryotes may be due to the difficulties of DNA replication through the more complex chromatin structure of eukaryotic chromosomes. All DNA polymerases add new nucleotides only to the 3' end of a DNA strand. As a result, a new DNA chain can only be synthesized in the 5' to 3' direction. The two template strands being antiparallel, this can easily be accommodated for the synthesis based on the DNA template strand in the 3' to 5' relative to the replication fork movement. However, what happens on the opposite DNA template strand which seems to require synthesis in the reverse orientation than the fork movement? The new DNA strand which seems to grow in the wrong direction from 3' to 5' is in fact synthesized discontinuously, in small successive separated pieces. DNA polymerase moves backward with respect to the direction of replication-fork movement so that each new DNA fragment can be polymerized in the 5' to 3' direction (Figure 1.4). The resulting small DNA fragments -called *Okazaki fragments*- are then joined to form a new continuous strand (Figure 1.4). The strand of DNA that is synthesized discontinuously is called the *lagging strand* , because the backstitching imparts a slight delay to its synthesis; the complementary strand, which is synthesized continuously, is called the *leading strand* (Figure 1.4).

This process requires fundamental actors including replicative polymerases. Polymerase δ catalyzes the extension of the continuous strand and polymerase ϵ assures the elongation of the discontinuous strand during replication [20–23]. Many other factors are involved in the progression of the replication forks to ensure the proper progression of the S phase [24–28]. For example, the profiling cell nuclear antigen (PCNA) provides a strong association with DNA contributing to the processivity of the replicative complex (replisome) [29, 30].

1.1.3 Replication termination

The termination of replication occurs when two convergent forks from two active origins meet at the end of the elongation step. DNA ligase I catalyzes the formation of the phosphodiester bond between the two neo-synthesized strands of DNA. The topological constraints generated

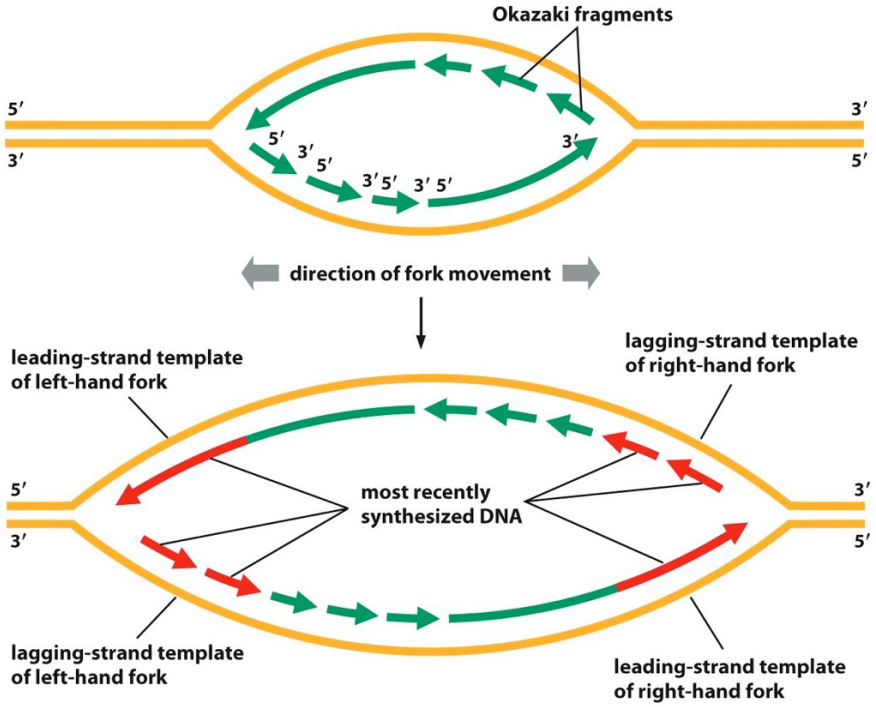

Figure 1.4: **Mechanism of DNA replication at replication forks.** The DNA molecules being replicated contain Y-shaped junctions, called replication forks. Two replication forks are formed at each origin of replication. At each fork, a replication machine moves along the DNA, separating both strands of the double helix and using each strand as a template to form a new daughter strand. The two forks move away from the origin in opposite directions, unfolding the double helix of the DNA and replicating the DNA as it goes. Replication of DNA in bacterial and eukaryotic chromosomes is therefore described as bidirectional. The upper diagram shows two replication forks moving in opposite directions; the lower diagram shows the same forks a short time later. Because both of the new strands at a replication fork are synthesized in the 5′ to 3′ direction, the lagging strand of DNA must be made initially as a series of short DNA strands, which are later joined together. To replicate the lagging strand, DNA polymerase uses a backstitching mechanism: it synthesizes short pieces of DNA (called Okazaki fragments) in the 5′ to 3′ direction and backward to fork movement along the template strand (toward the fork) before synthesizing the next fragment. (Adapted from Essential Cell Biology fourth edition [8]).

by replication are nullified by the action of DNA topoisomerase[5] I throughout replication and the daughter DNA molecules are finally deconcatenated by topoisomerase II [31].

Knowledge about the termination stage in eukaryotes is relatively limited. Unlike prokaryotes, no specific DNA sequence appears to be required for the termination stage [32]. However, at certain gene sequences there are specific sites where the replication forks are paused and the termination of the replication is induced. This is the case, for example, of regions of rDNA or of the RTS1 sequence in *S. pombe* [33] or in human the pausing replication fork near the transcription termination resulting in site-specific termination of replication [34]. Recently, the researchers identified in *S. cerevisiae* yeast 71 chromosomal regions that are terminating sites (TER) [35]. Within these regions, specialized fork gates control the termination in a polar fashion, i.e. only one of the two converging forks is stopped [36]. One of the problems

[5]These enzymes have the ability to change the topology of the DNA molecule by controlling the winding of the two strands of the molecule to allow the unfolding of the processes of the cell.

posed by the existence of such sequences is that they are more prone to "hot spots" of DNA breaks and chromosomal rearrangements [36].

1.2 Eukaryotic DNA replication program

It was early recognized thanks to autoradiography experiments that higher eukaryotes chromosome replication proceeds in *units* where DNA synthesis occur at defined time in S phase (Reviewed by Hand in 1978 [42]). Immunofluorescence staining of the synthesized DNA at the same time of S phase over two consecutive generations in mouse fibroblasts showed that replication profiles are conserved from one generation to the next, and that the same sites are replicated at the same time [43]. These experiments corroborated that each DNA sequence is replicated at a specific time of the S phase, so that the activation time of an origin of replication is not random but conditioned by the replication time of its replication unit. More recently, methods were developed to map the mean replication timing (MRT) profiles in a population of growing cells [44, 45]. These direct measurements confirmed that eukaryotic chromosomes replicate in a temporal order known as the *replication-timing program*. In mammals, they have revealed the presence of 100 kb to Mb-scale regions with similar timing, called constant timing regions (CTRs), replicating either early or late in the S-phase by coordinated activation of multiple origins. These early and late MRT plateaus were shown to be separated by rather steep timing transition regions (TTRs) of size ranging from 0.1 to 0.6Mb and initialy presumed to be replicated unidirectionally by a single fork coming from the early domain [37, 46]. It further revealed that, even though CTR locations are often conserved between cell types, MRT is cell-type-specific in mammals with up to 50% of the genome switching replication timing during development [37, 38, 47, 48]. Similarly, comparative analysis of replication profiles in 4 budding yeast species revealed species-specific replication patterns on top of an evolutionary conserved replication timing program despite chromosome rearrangements [49]. In human, MRT profiles were shown to be linked with the genome organization (Figure 1.5a). The main genomic link being that early replicating regions correspond to regions with high gene density and high GC content whereas late replicating regions correspond to regions with low gene density and low GC content [41, 50].

More recently, it was experimentally noticed that the replication rate of TTRs is not always compatible with the unidirectional progression of one replication fork [52]. In these cases, the coherent “wave” of replication from the early to the late TTR borders necessarily implies a more complex replication program. This observation was confirmed genome wide by the analysis of MRT profiles in seven human cell types including ES, somatic and HeLa cells [51, 53, 54]. This study indeed revealed that in each cell type, about half of the genome can be paved by replication U-domains where the MRT is U-shaped rather than by CTRs (Figure 1.6C). About half of the U-domains in one cell line is shared by at least another cell line (from 38.4% to 61%). The 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) reorganize

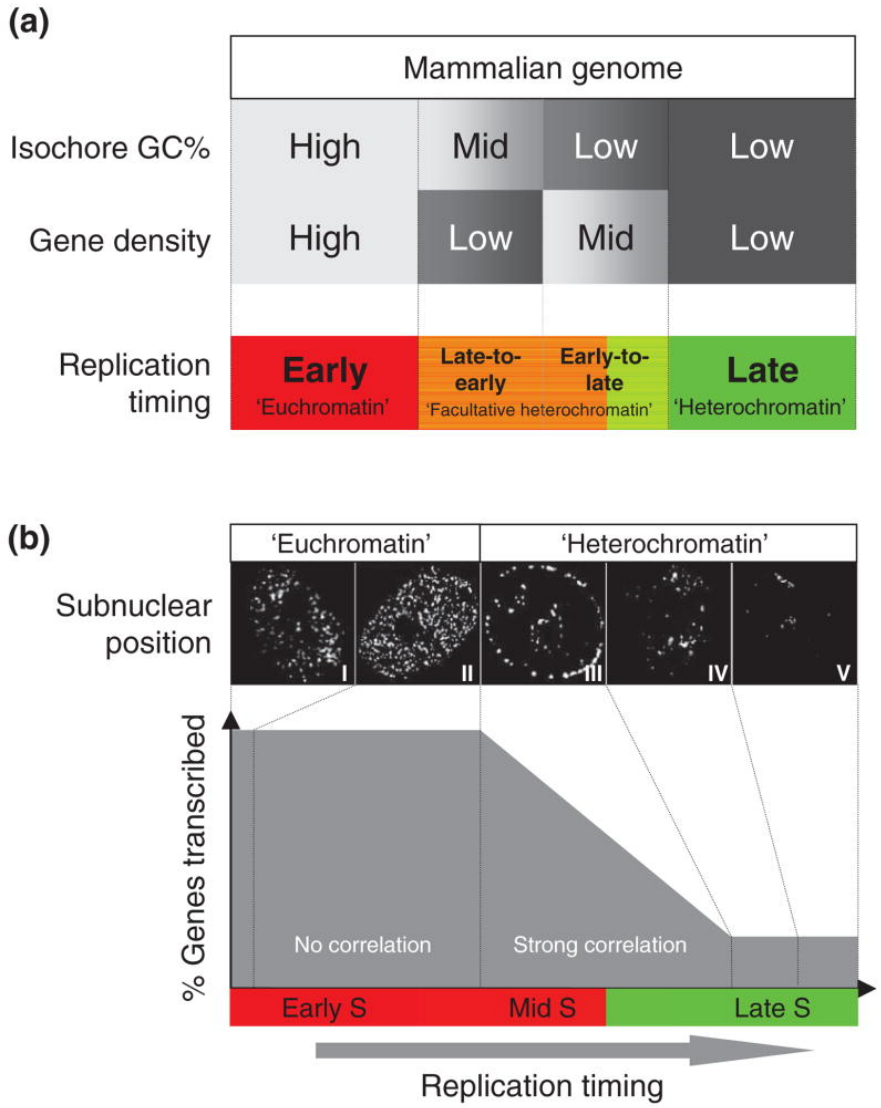

Figure 1.5: **Relationship between isochore properties and replication timing regulation, subnuclear positioning, and transcription.** (a) Isochores with unusual sequence properties are subject to replication timing regulation. The mammalian genome is partitioned into isochores with different GC content, Long Interspersed Nuclear Element (LINE) composition, and gene density, which are generally correlated. Isochores that are high in GC and gene density but low in LINE density are replicated early in S phase, whereas the alternate extremes are replicated late. Isochores with intermediate or mixed sequence features are frequently subject to replication timing regulation during differentiation (speculatively labeled "Facultative heterochromatin") [37, 38]. (b) Changes in replication timing that traverse the middle of S phase accompany changes in subnuclear position and transcriptional potential. Replication early in S phase (patterns I and II) takes place within the interior euchromatic compartment, whereas replication later in S phase (patterns III, IV, and V) takes place at the nuclear periphery (pattern III), at the nucleolar periphery (pattern III), and at internal blocks of heterochromatin (patterns IV and V) [39, 40].(Adapted from Rhind et al. [41]).

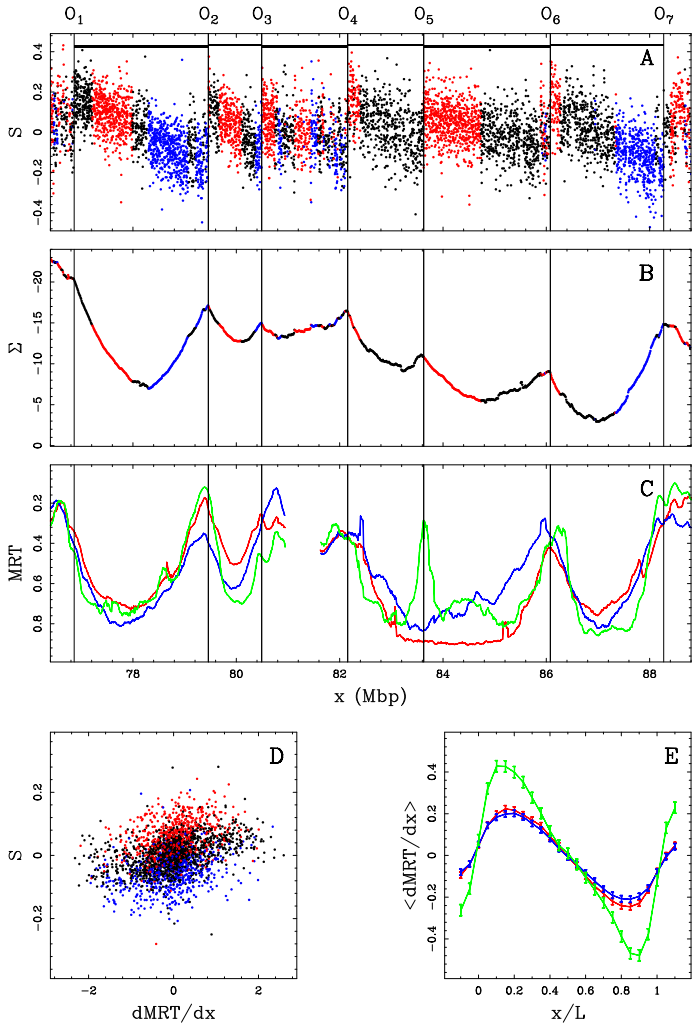

Figure 1.6: **Comparing skew $S = \frac{T-A}{T+A} + \frac{G-C}{G+C}$ and mean replication timing (MRT).**(A) S profile (Equation (1.3)) along a 11.4 Mb long fragment of human chromosome 10 that contains 6 skew N-domains (horizontal black bars) bordered by 7 putative replication origins O_1 to O_7. Each 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. (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 663 skew N-domains after rescaling their length L to unity; colors as in (C). (Adapted from Baker et al. [51]).

according to the so-called "consolidation" scenario, (ii) experience some boundary shift and (iii) emerge in a late replicating region [51] as previously observed in the mouse genome during differentiation [55]. The early initiation zones at U-domain borders were described as "master" replication origins (MaOris) [53, 56] It was proposed that replication waves likely initiate and propagate from MaOris toward the domain center via a cascade of (non-independent) origin firing, possibly by fork-stimulated initiation, resulting in the observed parabolic (U-shaped) change of MRT across each domain [51, 52, 54, 56].

If we assume a nearly deterministic replication program, where at each cell cycle nearly the

same set of replication origins is used, and where all replication origins fire at specific times, the population average replication program reflects faithfully what occurs in each cell cycle. The time of replication of each region is then a function of its distance (assuming constant fork speed) from an active origin and the time at which the origin was activated (Figure 1.8A). In this model, some origins would be activated at the beginning of the S phase, others in the middle of the S phase or at the end of the S phase, defining early, intermediate or late replication domains, respectively. Thus, it is important to distinguish between the reproducible replication timing of broad regions of the genome and the specific replication timing of individual origin for which there is in fact little evidence in higher eukaryotes. So, independently from the biochemical details of replication initiation, it is fundamental to characterize the spatial distribution of the replication origins as well as their replication timing and efficiency defined as the proportion of the cell cycle in which there are active. In this context, it was shown that budding yeast has well-defined, site-specific origins, many of which are efficient and fire in as many as 90% of S phase [57, 58]. These characteristics lead to fairly homogeneous replication kinetics [59]. So, considering the DNA replication program in budding yeast as deterministic is rather consistent.

If in *S. cerevisiae* and *S. pombe*, it was shown that replication origins correspond to specific sites called ARS (for autonomous replication sequence) and characterized by a short sequence motif and islands rich in AT [60–63], the situation is more complex in higher eukaryotes. Many methods were used to map the replication origins in human and other species [45]. Some approaches require to purify replication intermediates as in Short Nascent DNA sequencing (SNS-seq) [64, 65] and Bubble-seq [66] other approaches map the early-firing human replication origins as in Optical Replication Mapping [67] and Initiation site sequencing (Ini-seq) [68]. Even though the concordance between these methods has been the subject of discussions [68, 69], there is now increasing evidence that replication program is stochastic and that no two cell cycles use exactly the same set of replication origins and the same firing times [18, 69, 70]. In higher eukaryotes, the origins of DNA replication are activated in clusters as first observed by autoradiography [71] and later by DNA combing methods [72], where simultaneous activations would happen from an excess of potential loci [72–76]. Whatever the method used to identify the replication origin, the parameters that define the choice of origins in higher eukaryotes are beginning to emerge. These parameters depend on the sequence, the topology of the DNA, the presence or the absence of nucleosome[6] and transcription [18]. Cadoret et al., using SNS-seq, identified 283 origins of replication in HeLa cells over 1% of the genome and demonstrated that the density of active origins correlated with GC content [77]. By identifying 150 new origins of replication in this same cellular model, Karnani et al. revealed the influence of transcriptional regulation and chromatin structure on the selection of origins [78]. Moreover, the work of Karnani et al. suggested that in HeLa cells, initiation of replication takes place in AT-rich regions, near transcription initiation sites, and in open

[6]The nucleosome is the fundamental unit of chromatin, consisting of a segment of DNA wound in sequence around eight histone protein cores.

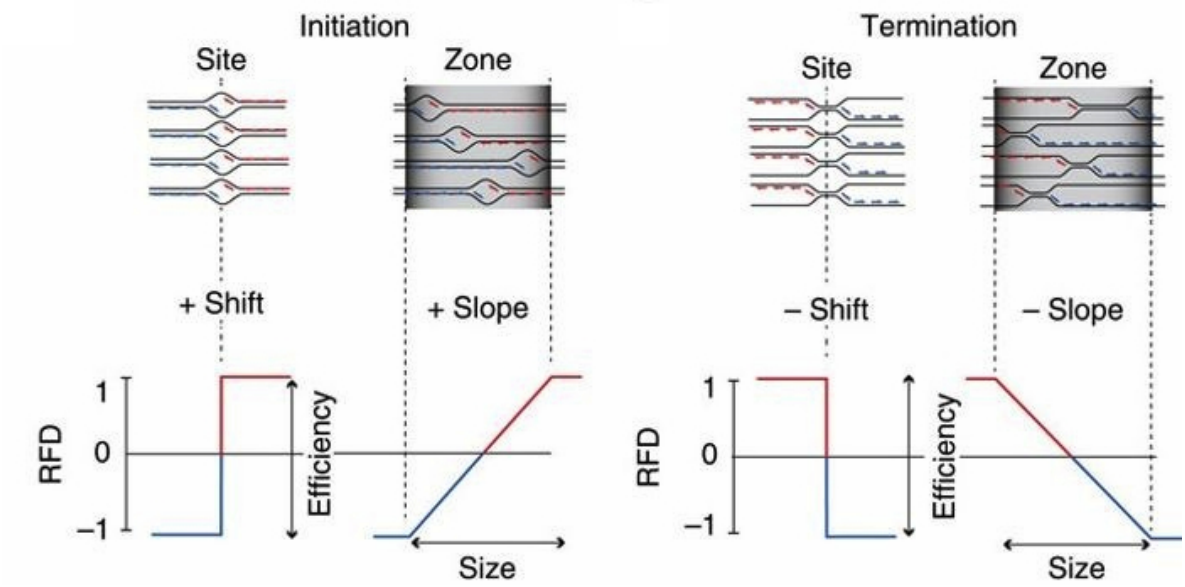

Figure 1.7: **Replication fork directionality.** (Top) Sketch representing replication bubble locations in 4 distinct S phases at a fixed origin, a dispersed initiation zone, a fixed termination site and a dispersed termination zone, from left to right respectively. The red and blue lines inside these structures indicate the Okazaki fragments synthesized during the progression of each fork. These populations of Okazaki fragments are randomly sampled from asynchronous growth cell populations in the OK-seq protocol [3]. (bottom) RFD profiles expected at the corresponding loci. (Adapted from Petryk et al. [3].)

chromatin regions [78]. Studies in *Drosophila* and mouse showed that the CpG[7] islands are associated to replication origins and that these origins are preferentially concentrated in regions containing transcriptional promoters [18, 79]. As initially reported for *Drosophila* and mouse [80] and then for human [68, 81], G-rich sequence motifs at replication origins would have the potential to form G-quadruplexes[8] contributing to origin specification. The genome wide studies also demonstrated that active origin are associated to transcription start sites (TSS) and GC rich regions [68, 81, 82]. In summary, if knowledge about replication origin properties made significant progresses, the current methods cannot precisely quantify origin efficiency and the relation between replication origin usage and specification of the replication timing program remains to be clarified.

1.3 Analysis of replication fork directionality

1.3.1 Replication Fork directionality

The stochastic nature of replication origin locations and firing times results in the orientation of the fork movement at a given locus x to depend on the considered S phase. The Replication Fork Directionality (RFD) is defined as the difference between the proportion of rightward

[7] CpG dinucleotide is a two-nucleotide DNA segment whose nucleic base sequence is CG.

[8] A G-quadruplex (G4) is a four-strand secondary structure that can be adopted by nucleic acids rich in guanine.

($p_{(+)}(x)$) and leftward ($p_{(-)}(x)$) moving forks at a given locus x:

$$RFD(x) = p_{(+)}(x) - p_{(-)}(x). \tag{1.1}$$

So, at a strong replication origin, where replication by a leftward moving fork shifts to replication by a rightward moving fork, the RFD will changes sharply from -1 to 1, resulting in a upward jumps (Figure 1.7). If the exact position of replication initiation changes from one cell cycle to the next within a so-called initiation zone (IZ), RFD will present an upward linear slope across the IZ with an amplitude that reflects the efficiency of the IZ (Figure 1.7). Correspondingly, at fixed termination, the RFD jumps sharply downward and dispersed termination sites within a termination zone (TZ) lead to a linear decrease of the RFD across the TZ (Figure 1.7). Many methods were used to compute the RFD profiles based on the replication fork asymmetry (Section 1.1.2). In the following section, we will detail two of them: the nucleotide compositional skew, which is an indirect method based on sequence composition, and the OK-seq method, which is a direct experimental method.

1.3.2 Nucleotide compositional skew

It was shown that if the replication strands are subjected to the same mutation/repair patterns, at the equilibrium, the equalities of [A]=[T] and [C]=[G] ([.] denotes the frequency of nucleotide in the DNA sequence) should be verified on each strand [83]. Inversely, when the complementary strands are subject to different mutation patterns, the nucleotide composition of strands can become asymmetrical, which is revealed by a departure from the equilibrium compositional equalities. One of the method to quantify the strand asymmetry is to calculate directly from the sequence the compositional TA skew and GC skew:

$$S_{\mathrm{TA}} = \frac{n_{\mathrm{T}} - n_{\mathrm{A}}}{n_{\mathrm{T}} + n_{\mathrm{A}}}, \qquad S_{\mathrm{GC}} = \frac{n_{\mathrm{G}} - n_{\mathrm{C}}}{n_{\mathrm{G}} + n_{\mathrm{C}}}, \tag{1.2}$$

where n_{A}, n_{C}, n_{G} and n_{T} are respectively the numbers of A, C, G and T in the analyzed sequence window.

Most bacterial genomes exhibit a GC skew, and some a TA skew [84, 85]. The skew profiles partition the circular bacterial chromosomes in two halves, one with positive skew values and the other with negative skew values separated by two sharp jumps: an upward jump at the replication origin and a downward jump at the terminus [86]. Hence, the sign of the skews correspond to the ± 1 RFD profiles of these chromosomes. This suggests that replication could be responsible for compositional strand asymmetry likely originating from replication coupled asymmetric mutation patterns. These skew profiles became an efficient way to determine or to confirm the position of bacterial replication origins and termination sites.

The relationship observed between the compositional asymmetry and the replication pro-

gram in bacteria can be generalized to eukaryotic genomes. In the human genome, as the TA and GC skews correlate [87–89], the total skew S defined as the sum of the TA and GC skews:

$$S = S_{\mathrm{TA}} + S_{\mathrm{GC}}, \tag{1.3}$$

was analyzed. It presents N shaped domains of size ranging from 0.2 to several megabases (Figure 1.6A) [90, 91]. Based on the analogy of the bacterial case the N-domains borders, skew upward jumps were proposed to be putative replication origins [88–91]. A linear decrease of the S profiles was observed between these putative replication origins (Figure 1.6A). In human the S profiles do not present sharp downward transitions but gradual downward slopes. Hence, the N-shaped skew profile cannot be trivially extend from the replicon model in bacteria. Because the inter-origin distance measured using DNA combing is about $\sim$ 50-100 kb [52], that is much smaller than the size of many detected N domains (1-3 Mb), it was suggested that some initiation and termination events must occur inside the N domains following a non-trivial random distribution which could result in the observed linear decrease of the skew. 678 N-domains bordered by 1060 putative replication origins covering $\sim$34% of the genome were identified in human [90, 91].

The GC and TA skews reflect the direction of the fork of replication because the leading and lagging strands experience different rates of nucleotide substitution [69]. Theoretical analyses [51] corroborated that the skew profiles, in both human and bacteria, result from replication fork directionality profiles by showing that under the assumption of strand-asymmetric substitution rates coupled to fork orientation, the total skew is expected to be proportional to the average fork directionality:

$$S(x) \propto RFD(x). \tag{1.4}$$

The linear decrease of S profiles along N-domains from positive (5' end) to negative (3' end) values would thus reflect a linear decrease of the RFD with a change of sign in the middle of the N-domains i.e. replication skew domains would be the signature of large-scale gradients of replication fork polarity [51].

1.3.3 Coupling between Skew and Mean Replication Timing U domains

In this section, linking replication fork polarity to MRT allows to bring another evidence of the organisation of the genome in replication domains. In Baker et al. [51], 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 replication origin. In this scenario, the replication timing profile $t(x)$ in one cell cycle is completely specified by the position x_i and the firing time $t_i = t(x_i)$ of the activated bidirectional replication origins O_i, termination time and position T_i being defined as the time and positions when the replication fork coming from O_i meets the replication fork coming from O_{i+1} (Figure 1.8A).

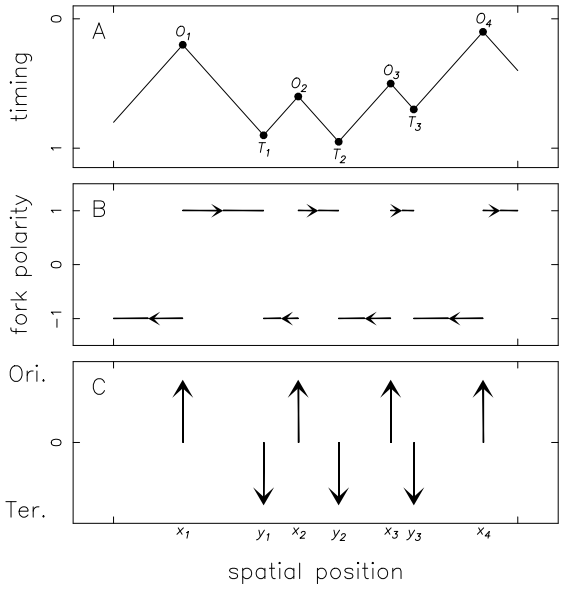

Figure 1.8: **Modeling the spatio-temporal replication program in a single cell.** (A) Replication timing $r(x)$, (B) replication fork polarity $p(x)$ and (C) spatial location of replication origins (upward arrows) and termination sites (downward arrows). $O_i = (x_i, t_i)$ corresponds to the origin 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). 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. The fundamental hypothesis is that the replication fork velocity v is constant. (Adapted from Audit et al. [54]).

The replication fork polarity profile $p(x)$ in that cell cycle is directly deduced from the position of the origins O_i and termination sites T_i (Figure 1.8B), whose spatial distributions can be represented by Dirac functions of weight $+1$ and -1, respectively (Figure 1.8C). One can deduce the fork polarity in Figure 1.8B (and the resp. origin and termination site locations in Figure 1.8C) by simply taking successive derivatives of the timing profile in Figure 1.8A. These 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 [38, 48, 92–95]. Since taking the spatial derivative commutes with statistical and spatial average, one gets:

$$\frac{d}{dx}\mathrm{MRT}(x) = \frac{1}{v}\mathrm{RFD}(x)\ , \tag{1.5}$$

$$\frac{d}{dx}\mathrm{RFD}(x) = 2(N^{Ori}(x) - N^{Ter}(x)), \tag{1.6}$$

where $\mathrm{MRT}(x)$, $\mathrm{RFD}(x)$, $N^{Ori}(x)$ and $N^{Ter}(x)$ stands for the average over many cell cylces and over some spatial resolution Δx of the replication timing, replication fork polarity and numbers of origins and termination sites per unit length.

Equations (1.4) and (1.5) reveal that the RFD provides an immediate connection between

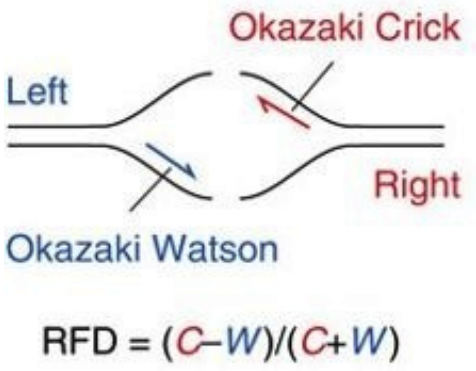

Figure 1.9: **Principle of determination of RFD based on the strand of Okazaki fragment.** Schematic leftward (blue) and rightward (red) forks with Okazaki fragments. RFD is computed as the difference between the proportions of rightward- and leftward-moving forks. (Adapted from Petryk et al 2016 [3]).

the skew S and the mean replication timing:

$$S(x) \propto \frac{d}{dx}\mathrm{MRT}(x). \tag{1.7}$$

Since the mutations responsible for the observed strand compositional asymmetries only accumulate in germline cells, this relationship was verified using, as a substitute for germline MRT, the MRT in BG02 embryonic stem cell [51]. The skew S was indeed correlated with dMRT/dx, in the BG02 embryonic stem cells (Figure 1.6D). 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 an other 6 cell lines [51]. These correlations are as important as those obtained between the dMRT/dx profiles in different cell lines [51], as well as those previously reported between the replication timing data themselves [38, 48, 94]. From equation (1.7) one also deduces that the integration of the skew S is expected to generate a profile rather similar to the MRT profile. In segments where S behaves linearly, its integral is thus expected to show a parabolic profile. The integrated S function when estimated by the cumulative skew Σ (Figure 1.6B) along N-domains, indeed displays a U-shaped (parabolic) profile likely corresponding the MRT profile in the germline. Remarkably, N-domains often correspond to genome regions where the MRT in the BG02 embryonic stem cells and other cell lines is U-shaped (Figure 1.6C). Correspondingly, averaging dMRT/dx over skew N-domains in BG02 and other cell lines resulted in N-shape profiles (Figure 1.6E) further validating equation (1.7) and that replication timing U-domains and skew N-domains both result from underlying N-shaped RFD profiles.

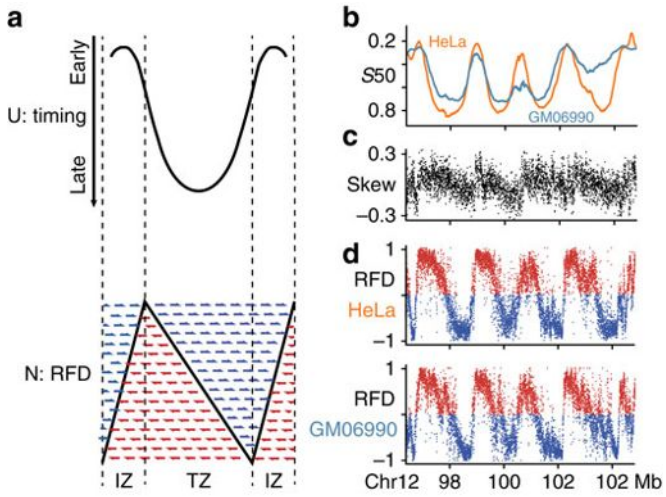

Figure 1.10: **OK-Seq corroborates the replicative organization of N/U-domains.** (a) U-domains of replication timing (top) are predicted to show an N-shaped RFD profile (bottom). The steep slopes at U-domain borders predict a high |RFD| whereas the flat slope at the central, late-replicating valley predicts a null RFD [51, 52, 56]. Blue and red arrows, expected proportions of Okazaki fragments from either strand across zones of predominant initiation (IZ) at borders and predominant termination (TZ) at center of U-domain. Dark line, expected RFD profile. The mathematical relationship dMRT/dx=RFD/v implies that RFD increases (Ascendant Segment 'AS') when the timing profile is convex ($d^2MRT/dx^2 > 0$) and decreases (Descendant segment 'DS') when it is concave ($d^2MRT/dx^2 < 0$; note that the time axis is oriented from top to bottom). (b) HeLa (orange) and GM06990 (teal blue) replication timing profiles showing four adjacent U-domains[51]; (c) nucleotide compositional skew profile showing skew N-domains [90] matching replication timing U-domains; (d) N-shaped RFD profiles determined by OK-Seq matching N-domains. (Adapted from Petryk et al. [3]).

1.3.4 Okazaki fragment sequencing (Ok-seq)

A novel method was developed to measure directly the orientation of the fork at each locus x of the human genome ([3]). This method is based on the purification and sequencing of the Okazaki fragments (the method will be detailed in Chapter 3). The Okazaki fragments (Figure 1.4) are produced by the discontinuous replication of the lagging strand in the opposite orientation to the fork movement. So by determining the strand of origin (Watson or Crick) of Okazaki fragments, we can determine the direction of the replication fork (Figure 1.9). Counting the number W (resp. C) of Okazaki fragments on the Watson strand (resp. Crick strand), we can deduce the proportion of forks replicating in the 5' $\rightarrow$ 3' ($p_{(+)} = C/(C+W)$) and the 3' $\rightarrow$ 5' direction ($p_{(-)} = W/(C+W)$) and, in turn the replication fork directionality (Eq. (1.1)):

$$RFD(x) = \frac{C - W}{C + W}. \tag{1.8}$$

The N-shaped RFD profiles observed along skew N-domains and replication timing U-domains were reproduced in the RFD profiles obtained by the Ok-seq method (Figure 1.10b,c,d) [3]. The remarkable similarity of RFD gradients derived from skew [96] or replication timing [51] or determined by Ok-seq provides strong mutual validation of these data. In addition, the

higher resolution of Ok-seq reveals the areas of initiation and termination of replication with much greater accuracy than with skew or MRT profiles.

All together, there exist a strong correlation between the RFD and MRT profiles. Under the hypothesis that the fork velocity is constant, the steep slopes at the U-domain boundaries corresponds to a high level of |RFD| whereas the flat slope in the center of the U-domains corresponds to a null RFD (Figure 1.10a). In addition, the U-domains frequently coincided with N-shaped domains detected independently from the asymmetric strand composition [51–53, 90] (Figure 1.10c), proposed to reflect an N-shaped pattern of RFD [90, 96] in germline. OK-seq fully verified the predicted RFD gradients deduced from replication timing U-domains and skew N-domains (Figure 1.10bcd). It was proposed that these linear gradients of RFD would result from cascades of replication origin firing propagating from the domain boundaries to the center [52, 53, 56].

1.4 Interplay between DNA replication program and other genomic features

The modification of the replication program between cell lines can occur at regions in any stage the replication program. For example, the change of efficiency of a replication origin could be associated with a change of firing time, of transcription level or of chromatin state at this locus. Hence it is of general interest to understand the link between replication initiation, replication timing, chromatin state and transcription.

1.4.1 Coupling between nuclear architecture and the replication program

It has been observed that within the nucleus, DNA replication is spatially organized into multiple small subnuclear domains, called replication foci [39]. The distribution pattern of replication foci reproducibly depends on S phase progression (Figure 1.5b). At the beginning of the S phase, replication foci are distributed everywhere in the nucleus. In mid of S phase, these foci tend to relocate at the nuclear membrane or at the periphery of the nucleoli. Finally, at the end of S phase, larger foci corresponding to a clustering of foci are observed at the same peripheral zones (nucleus and nucleolus). Late-stage S foci appear to be associated to heterochromatin zones (Figure 1.5b) [97]. The structure of these foci and their biological function are not fully established. It has been proposed that these foci shelter several consecutive replicons, whose synchronous initiation depends on a single initiator complex. More recently, high resolution fluorescence microscopy studies suggested that these foci correspond to a clustering of multiple replication complexes, which would be activated at the same time [98, 99]. This was also suggested by the model proposed by Berezney et al. in 2000 [39] and taken up by Cayrou et al. in 2011 [79], according to which the replicons are clustered and synchronously replicated.

1.4.2 Coupling between replication and transcription

For decades, researchers have been eager to find the connections between two fundamental genomic mechanisms, the DNA replication and the transcription of genes. Analyses in human [50] and *Drosophila* [100, 101] have shown that there is a relationship between early replication timing and active gene transcription of close-by euchromatin DNA strands. However, in *Drosophila*, transcriptional activity has to be considered over large domains (> 100 kb) rather than at the level of individual genes to understand the impact of transcription on replication timing [101]. More recent studies have confirmed this now *classic* pattern of early/late replication associated with gene expression/repression but suggested that in fact other factors such as chromatin state are more likely the causally linked factors to replication timing [48].

Using the SNS-seq, origin of replications were indentified in HeLa cells and it was found that most of them overlap with transcriptional regulatory elements [77]. Recently using the OK-seq method, Petryk et al. described 9836 (HeLa) and 5684 (GM06990) replication initiation zones corresponding to ascending segments of the RFD profiles and observed that $> 46\%$ were flanked by at least one active gene [3]. They also reported 9440 (HeLa) and 5715 (GM06990) large ($8-2600$ kb) descending segments (DS) of preferential replication termination covering 49.5% and 39.4% of the genome, individually, with 5000 (35% of the genome) shared between the two cell lines [3]. In early S phase, descending segments were associated to transcribed gene bodies and to regions of co-oriented replication and transcription. In late S phase, descending segments predominantly comprised large non-expressed DNA regions.

1.4.3 Association between DNA replication and epigenetic

The complex formed by DNA and its packaging proteins is named chromatin. In chromatin, ~ 150 bp DNA segments wrap around a bead-like structure formed by an octamer of histone proteins, to form the nucleosome. Histones are the most prevalent proteins in chromatin. There are five types of histones: H1, H2A, H2B H3 and H4, . 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 H2A and H2B dimers. The H1 protein is a linker protein (10 to 90 bp long) that completes the nucleosome. Histones carry epigenetic marks that convey a functional information by two mechanisms. First, one of the canonical histones can be replaced by a histone variant[9]. For instance, transcriptionally active regions are enriched in H3.3 and H2AZ variants 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 enviromnent. These tails can carry diverse covalent modifications that have a functional meaning. There is a specific notation to indicate modifications: H3K9ac means that a histone H3 has been acetylated on its ninth amino-acid (a lysine). Histone modifications can make chromatin looser (e.g. acetylation modifications are typically associated to active transcription) or can serve as an anchor

[9]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.

to regulatory proteins. The epigenetic term groups histones modifications described above as well as the modifications of DNA as for example cytosine methylation (Reviewed in [102–104]).

Chromatin organization and its dynamics participate in essentially all DNA-templated processes including transcription and replication. Replication origin positioning is cell line specific: even if there were consensus sequence motifs specifying replication origins, an additional regulation mechanism would be needed [105]. Experiments demonstrated that origin firing depends on the chromatin environment [106]. Therefore, mechanisms that position and control the time of firing of origins are likely associated with epigenetics marks and linked to chromatin structure [18, 40, 76, 107]. It was demonstrated that epigenetic marks and chromatin accessibility can be used to faithfully estimate the DNA replication timing profiles in a cell type specific manner [108]. From a reverse perspective, it was argued that replication timing profiles can be interpreted as an epigenetic signature of a cellular differentiation state [109]. In both views, there is strong a coupling between the DNA replication and the epigenetic marks profiles. A recent study described human chromatin as constituted of four chromatin states when analysis eleven genome wide epigenetic profiles at 100 kb resolution of MRT in 5 somatic cell lines and one embryonic stem cell (ESC) [110, 111]. These states have different MRT, namely from early to late replicating, replication proceeds though a transcriptionally active euchromatin state (C1), a repressive type of chromatin (C2) associated with polycomb complexes, a silent state (C3) not enriched in any available marks, and a gene poor HP1-associated heterochromatin state (C4). Mapping these chromatin states onto replication U-domains revealed that the associated replication fork polarity gradient corresponds to a directional path across the four chromatin states, from C1 at U-domains borders followed by C2, C3 and C4 at centers. This work provided a refined description of the replication program links to epigenetic state and transcriptional activity compared to the classical dichotomy between active replication euchromatin states (early replication in S phase) and silent and late replicating heterochromatin states. More recently, the Gilbert group [112] identified cis-regulatory elements that are collectively responsible for early replication, structural compartmentalization, local chromatin architecture, and transcription. This research corpus demonstrates there are intricate associations between DNA replication, epigenetic regulation, transcription and chromatin structure. For more details about the chromatin structure and its influence on DNA replication, we refer the reader to the review in Ref. [106].

1.5 Thesis objectives

Understanding how DNA replication program is adapted dynamically to normal or pathological differentiation cases still remain the focus of great interest. This thesis work proposes an original characterization of the plasticity of the DNA replication program during development and in association with cancer progression, based on the profiling of the high-resolution replication profile obtained for 12 cell lines by purification and sequencing of Okazaki fragments (Xia Wu, Hyrien Team, IBENS, Paris). From these experimental methods, we determine the replication fork directionality (RFD) genomic profiles at a resolution of 10 kilobase (kb), and we performed a comparative analysis of these profiles with other markers of the organization of genome activity (eg gene transcription, replication times). The manuscript is organized in six chapters. The first one is the current introduction. Chapter 2 is the description of selected cell lines and the statistical methods (Hierarchical classification, Spearman/Pearson correlation) and informatics tools (Python packages) used in the following chapters. The results are reported in the Chapters 3, 4 and 5. In the Chapter 3, we describe in details the Ok-seq biological experiment and the RFD computing methods. Then we test the robustness of the RFD profiles to classify the cell lines based on their tissue of origin by doing a comparison of the RFD profiles among the 12 cell lines. We show that our classification remains stable when considering "probes" as small as 50 Mb suggesting that replication program changes are widespread over the whole genome. We identify that the regions rich in GC are more conserved among the 12 cell lines in terms of RFD profiles. This suggests that there is a spatial susceptibility to replication program changes between cell lines. From these observations, we define the regions of the genome with an unstable replication program. In chapter 4 we extend these cell line comparisons to replication timing and transcription. On the one hand, we find the cell lines classification are similar to those obtained using the RFD profiles. On the other hand, we confirm that the large late replicating domains correspond to unstable regions in terms of replication. We also address in a direct manner the association between replicative changes and transcription changes and we recover the existence of a general coupling between replicative changes and transcription changes. Comparing the expression levels between GM06990 and Raji cells, we identify 2075 genes that change expression level and their are coupled with replicative change relatively. In chapter 5 we focus on a cellular system modeling the progression of Chronic Myeloid Leukemia (CML). First, we demonstrate that by using the RFD profiles we are able to follow gradual replicative changes occurring during CML progression. Then, performing manual annotation of the efficiency changes of replication initiation zones among the cell lines of the CML system, we observe that the targeted initiation zones are more frequently weakened than enhanced and that weakened initiation zones are associated with transcription repression. Chapter 6 presents a general conclusion and perspectives.

CHAPTER 2

Materials and Methods

2.1 Cell lines

We mainly report analyses concerning the characterization of replication and transcription in 12 cancer and non cancer human cell lines (Table 2.1). The 12 cell lines belongs to three types: Four lymphoid cell lines (GM06990, Raji, BL79, IARC385), four myeloid cell lines which form a Chronic Myeloid Leukemia 'CML' model (TF1_GFP, TF1_BCRABL_1M, TF1_BCRABL_6M, K562) and four adherent cell lines (IMR90, HeLa, TLSE19, IB118). Table 2.1 recapitulates the main cell lines properties and cancerous status. We refer the reader to [3, 113] for the details of cell sources and culture conditions.

2.1.1 Adherent cell lines

IMR90 is a normal primary fibroblast cell line, obtained from ATCC. IMR90 cells are untransformed and presenescent cells [114]. The three other adherent cell lines are cancerous. HeLa cell line is a cancer cell line established by George Gey and commonly used in cell biology and medical research. HeLa cell line comes from a metastasis sample taken from Henrietta Lacks, an African American patient with cervical cancer who died in 1951 [115]. HeLa was the first immortal cell line of human origin. Note that an immortal cell line of animal origin was created 11 years before by Wilton Earle [116]. These cells have the distinctive property of dividing indefinitely, which means that they can proliferate in laboratory culture conditions. TLSE19 and IB118 are Leiomyosarcomas (LMS) cell lines. Sarcoma refers to the malignant growth that originates from transformed cells out of mesenchymal tissues. These tumors develop in bones, muscles, ligaments, nerves, fat tissues and veins. LMS is a kind of Soft Tissue Sarcoma (STS) called threatening melanoma [117]. STSs develop over particularly short time mainly in muscles, fat, veins, or different tissues that ensure body organ structure. LMS is an

Type	Name	Description	Status
Lymphoid Derived	GM06690	GM06990, lymphoblastoid type, immortalized cell line	Normal
	Raji	Raji, a cultured line of lymphoblastoid cells derived from a Burkitt lymphoma (BL)	Cancer
	BL79	Bl79 is a BL cell line infected with EBV *in vivo*	Cancer
	IARC385	IARC385 was obtained by *in vitro* immortalization of the normal B lymphocytes of the same patient as BL79	Normal
Myeloid Derived	TF1_GFP TF1_BcrAbl_1M TF1_BcrAbl_6M	TF1 transfected with GFP alone, or transfected with a GFP_BCR_ABL1 fusion and cultured for 1 month or 6 months. These three samples are believed to represent three consecutive stages in tumourigenesis of chronic myelogenous leukemia	Cancer
	K562	K562 cells are derived from a CML patient in blast crisis (final phase)	Cancer
Connective tissues	HeLa	HeLa, cell line was derived from cervical cancer cells (epithelial origin)	Cancer
	IMR90	Human Caucasian fetal lung fibroblast	Normal
	TLSE19 IB118	TLSE19 and IB118 leiomyosarcoma cell lines were both isolated from a tumor after surgical resection. Leiomyosarcoma are sarcoma that develops in connective tissues	Cancer

Table 2.1: **Cell lines description**.First column specifies the cell lines type, the second (resp. third) column corresponds to the cell line name (resp. Description) and the fourth column corresponds to the cancer/normal status of cell lines.

uncommon sort of malignant growth [118]. The LMS do not respond well to chemotherapy or radiotherapy, and are viewed as a kind of malignancy safe. TLSE19 and IB118 were established after surgical resection from a buttock muscle tumor and a cutaneous scalp tumor, respectively.

2.1.2 Blood cell lines

All blood cells are produced in the bone marrow in a process known as hematopoiesis. All type of blood cells derive from pluripotent hematopioetic stem cells. Differentiation of these pluripotent cells results in either myeloid or lymphoid progenitor cells. Further differentiation leads to diverse cell types including myelobasts and erythrocytes in the myeloid lineage and B and T lymphocytes in the lymphoid lineage.

Lymphoid cell lines

We used two Burkitt lymphoma cell lines. The first brief description of Burkitt Lymphoma (BL) cells was established in 1897 by doctor Albert Cook [119]. In 1958, Dennis Burkitt described in details BL cell lines [120, 121]. This type of Lymphoid cancer is aggressive and widespread among the population and can affect children and adults. One reason BL is of interest here is that the overexpression of Myc gene and Epstein−Barr virus (EBV) infection distinguish BL from other Lymphomas. The overexperssion of Myc in human fibroblast induces hyper-activation of origin firing in early S phase [122, 123]. The First BL cell line is Raji. This cell line was established by R.J.V Pulvertaft in 1965 from a Burkitt lymphoma extract from an 11 years old male from Africa (Nigeria) [124]. Raji is the first continuous human cell line of hematopoietic origin. The second BL cell line is BL79 duplicated from the International Agency for Research on Cancer (IARC).

We also used two control normal B lymphoblastoid cell lines (LCL): GM06690 and IARC385. A common method to generate LCL is the transformation of peripheral B lymphocyte cells from blood by density centrifugation. While T cells are removed from the cell population, the transformation of B lymphocytes is accomplished by EBV infection. IARC385 was duplicated from IARC and was obtained from the same patient as BL79.

Myeloid cell lines

Myeloid cell lines comprise a cellular model for establishment and progression of chronic myeloid leukemia (CML), a malignant disease characterized by the Philadelphia (PH1) chromosome and the formation of the BCR_ABL1 fusion gene, whose expression is necessary and sufficient for CML formation [125]. TF1 is a BCR_ABL negative cell line established from a patient with erythroleukemia. TF1_GFP and TF1 BCR_ABL were obtained by transduction of green fluorescent protein (GFP) or a BCR_ABL_GFP fusion gene using a murine stem cell virus (MSCV)-based retroviral vector [126]. TF1_BCR_ABL cells were analyzed after culturing for 1 month (TF1_BCRABL_1M) or 6 months (TF1_BCRABL_6M) following

transduction. K562 (the last myeloid cell line) was obtained by culturing leukemic cells of pleural effusion of a patient with PH1-positive CML in blast crisis, now called K562 [127]. Hence, K562 is a late CML model.

2.2 Various Genomic data

2.2.1 GC Content

GC-content fluctuations recapitulate the non uniform organization of gene size [128], and gene density [129]. There is also a correlation between GC content and other features of the genome [130]. For example, GC-rich regions are poor (resp. rich) in LINE1[1] (resp. ALU[2]) repeat sequences. Interestingly, there is an important correlation between GC content and replication timing [50, 133], early replication being correlated with high GC content.

The GC-content was computed as :

$$GC\% = \frac{n_G + n_C}{n_A + n_T + n_G + n_C} * 100 \tag{2.1}$$

where n_A, n_T, n_G and n_C are the numbers of A, T, G and C nucleotides counted along the window of interest. We defined GC-content categories following the 5 isochores classification of the human genome [130] in light isochores L1 (GC<37%) and L2 (37%≤GC<41%) and heavy isochores H1 (41%≤GC<46 %), H2 (46%≤GC<53%) and H3 (GC≥53%). The genome coverage of L1, L2, H1, H2, H3 was 26.5%, 32%, 24.0%, 13.1%, and 4.4%, respectively, after classification based on GC content in non-overlapping 10 kb windows.

2.2.2 Comparative Genomic Hybridization data

Total genomic DNA was extracted for GM06990, Raji, IARC385 or BL79 cells to determine the copy number of each loci. We represented in Figure 2.1 for each cell line the $\log_2$ ratio of the copy number relative to a normal karyotype along of each chromosome. The red line represents the loss in copy number as we see in IARC385 cell line for the chromosomes X and 18. The blue line represents the gains in copy number as we see in chromosome 7 in IARC385 cell line. The black lines correspond to normal copy number without any gains or loss. GM06690 appears as a normal cell line where the copy number is free from amplification or deletion.

[1] The long interspersed nuclear elements are part of the repeated sequences dispersed within the DNA. Their length are usually a few thousands base pairs and their exact nature varies according to the species.

[2] An Alu sequence is a short DNA fragment belonging to the family of small interspersed nuclear elements or SINE. The Alu sequences are the most abundant transposable element of the human genome [131], distributed throughout all chromosomes, representing about 10% of the human genome [132].

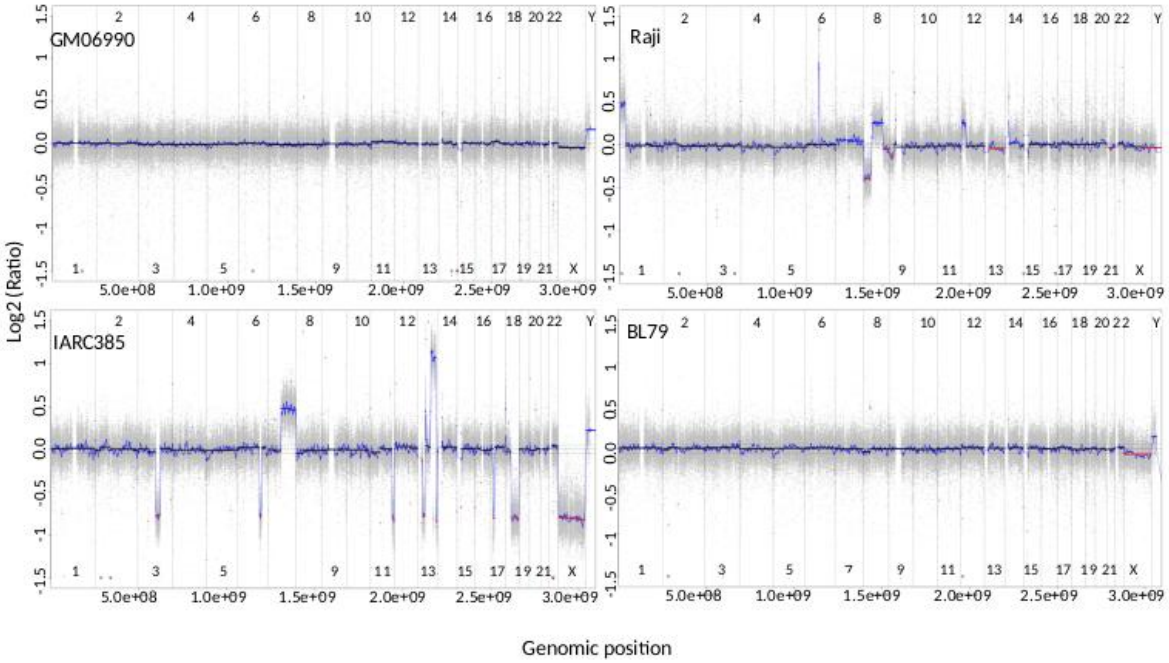

Figure 2.1: **CGH array analysis of GM06990, Raji, BL79, and IARC385.** The $\log_2$ ratio of copy number is plotted along the length of each chromosome. Red and blue horizontal segments indicate copy number losses and gains, respectively.

2.2.3 MRT computed for 7 cell lines

MRT was profiled by Repli-seq [93, 134], which consists in sequencing of newly replicated DNA of cells sorted by DNA content into consecutive compartments of S-phase. For GM06990, GM12878, K562, HeLaS3 and IMR90, alignment files of Repli-seq libraries (BAM files) for 6 S-phase fractions were obtained from the ENCODE project [48] at `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/`. The number of aligned reads ranged from 8.5 x 10^6 in K562 to 62.8 x 10^6 in IMR90 and all reads were used in downstream analyses. Repli-seq data for TF1_GFP and TF1_BCRABL_1M were obtained using a modified protocol based on the incorporation of EdU [113] instead of BrdU as in the original protocol [48, 93]. Purified nascent DNA was sequenced on illumina HiSeq platform and aligned to human genome (hg19) at the IB2C sequencing platform. Only uniquely mapping reads (186.0 x 10^6 for TF1_GFP and 143.3 x 10^6 for TF1_BCRABL_1M) were used in downstream analyses. To compute MRT profiles, tag densities in 100 kb windows were normalised and denoised as in [51, 54] using modified thresholds for the two EdU-based datasets

(TF1_GFP and TF1_BCRABL_1M). At each locus, the distribution of replication times was obtained by normalizing the denoised tag densities of each S phase fractions by their sum. MRT values were estimated as the mean of these distributions.

2.2.4 Quantification of RNA-seq for the 12 cell lines using TopHat

For Raji, GM06990, BL79, or IARC385, total RNA was extracted from 1-10 x 10^6 exponentially growing cells. Library preparation and Illumina sequencing were performed at the Ecole Normale Supérieure Genomic Platform (Paris, France). For TF1_GFP, TF1_BCRABL_1M and TF1_BCRABL_6M, three biological replicates each of total RNA prepared and provided by the genomic and microgenomic platform of Université Claude Bernard Lyon 1, France (http://profilexpert.fr). From 18.2 x 10^6 to 22.4 x 10^6 quality_controlled reads were obtained per replicate. For TLSE19 and IB118, RNA-data were provided by Frederic Chibon, INSERM, Bordeaux. RNA profiling was performed using paired-end sequencing (2 x 76 bp), yielding 137.1 x 10^6 and 104.9 x 10^6 paired-end RNA-seq reads for TLSE19 and IB118, respectively. For K562, HeLaS3 and IMR90, we used RNA-seq data from the ENCODE project. We used 2 biological replicates from which we obtained paired-end sequenced datasets. Read files in fastq format were downloaded from the European Nucleotide Archive https://www.ebi.ac.uk/ena under accession numbers SRR315336 and SRR315337 for K562 and accession numbers SRR315330 and SRR315331 for HeLaS3. For IMR90, fastq files for experiment 'EncodeCshlLongRnaSeqImr90CellPap' were downloaded from UCSC http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/; these data correspond to accession numbers SRR534301 and SRR534302.

Gene transcriptional levels for the 12 cell lines were estimated using the same computational pipeline based on the TopHat suite of softwares [135]. TopHat (version 2.1.1) and bowtie2 (version 2.2.9) were used to align RNA-seq reads to the human genome (hg19). RNA abundance were computed using Cufflinks (version 2.2.2). We fed the reference transcript annotation provided by Illumina iGenomes (ftp://igenome:G3nom3s4u$@$ussd-ftp.illumina.com/Homo_sapiens/UCSC/hg19/Homo$_$sapiens$_$UCSC$_$hg19.tar.gz) to TopHat (-G option) and cufflinks (-G option), and only considered RNA abundance estimates for genes present in the reference annotation. For each cell line we obtained, for the same set of 24,371 genes of size $>$ 300 bp, the estimated mRNA level expressed in FPKM (Fragments Per Kilobase of exon model per Million mapped fragments) and a 95% confidence interval on the FPKM ([FPKM_{lo}, FPKM_{hi}]). We considered a gene to be expressed when FPKM $\geq$ 1 and we filtered out cases where $\text{FPKM}_{hi}/\text{FPKM}_{lo} > 2$ (at most 118 genes were filtered out in TF1_BCRABL_6M). The proportion of expressed genes ranged from 43.5% in Raji to 51.3% in TF1_BCRABL_1M. Transcription level of any region R of length l_R was estimated based on the FPKM values of all the expressed genes g that overlap with the regions as:

$$\mathrm{FPKM(R)} = \sum_{\mathrm{g}} \frac{\mathrm{l}_{g \cap R}}{\mathrm{l}_R} \mathrm{FPKM(g)}, \tag{2.2}$$

where $\mathrm{l}_{g \cap R}$ is the overlap length between g and R.

2.3 Statistical Methods

2.3.1 Pearson and Spearman Correlation coefficients

Correlation coefficients are statistical measures constructed to capture the strength of the relationship between the relative fluctuations of two random variables. Their values range between -1.0 and 1.0. A correlation of $+1.0$ indicates a perfect positive correlation (in phase), while a correlation of -1.0 indicates a perfect negative correlation (in opposite phase). A correlation of 0.0 indicates no relationship between the fluctuation of the two variables.

The most common correlation coefficient is the product-moment correlation of Pearson [136]. It is simply defined as the covariance Cov(X,Y) of two variables (X,Y) normalized by their standard deviations (σ_X and σ_Y) to correct for the dependence of the covariance on the amplitude of the fluctuations, thus assuring Corr(X,Y) $\in [-1,1]$:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y}. \tag{2.3}$$

Given a sample of n joint observations (x_i, y_i) of X and Y, Corr(X,Y) is estimated as follows:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.4}$$

where $\bar{x}$ and $\bar{y}$ are the empirical means.($\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, same for $\bar{y}$).

The Spearman rank correlation coefficient is defined as the Pearson correlation coefficient of the ranks of X and Y [137]. Let r_i be the rank of x_i and s_i the rank of y_i, to compute the Spearman rank correlation coefficient we used Eq. 2.4 replacing x_i by r_i and y_i by s_i. Pearson correlation coefficient captures linear relationships between X and Y fluctuations. Spearman's is more robust as it captures any monotonic relationship. If there are no repetitive data values, a perfect Spearman rank correlation occurs with $+1$ or -1 when the two variables are a monotonic function of the other. Note that, in this thesis, the results of the correlation analyses were qualitatively unchanged if calculated using Spearman rank-correlation coefficients instead of Pearson correlations. Hence, we chose to report only results obtained using Pearson correlations.

2.3.2 Hierarchical ascending classification

Hierarchical ascending classification (HAC) is a simple iterative classification method [138]. Given a partition of a set of N objects into k classes, the core idea is to join the two "closest" classes into one class and thus obtain a new partition in k−1 classes. Starting from N classes of one object, HAC amounts to repeat this agglomeration scheme until obtaining one class of N objects. This process requires a dissimilarity measure d(o, e) between pairs of objects and an aggregation rule to derive dissimilarity between object classes. These successive groupings produce a binary classification tree that can be represented as a dendrogram, the root of which corresponds to the class grouping all the individuals. This dendrogram represents a hierarchy of partitions. We can then choose a specific partition by truncating the tree to a given level, depending on some constraints as the number of classes.

"Minimum jump" strategy or "single linkage" aggregation rule amounts to define the dissimilarity between two classes C_1 and C_2 as their "nearest neighbors" distance:

$$\Delta(C_1, C_2) = \min_{o \in C_1, e \in C_2} d(o, e) \tag{2.5}$$

It produces "chains" of classes, linked to each other like links in a chain, by pair of objects that are close to each other [139].

In opposite, "maximum jump" or "complete linkage" aggregation rule amounts to define the dissimilarity between to classes C_1 and C_2 as the dissimilarity between their most distant objects [136, 140]

$$\Delta(C_1, C_2) = \max_{o \in C_1, e \in C_2} d(o, e) \tag{2.6}$$

HAC algorithm does not take into account the nature of the dissimilarity function provided; it does not matter whether it is a true distance or another derived measure that may be more meaningful to the context under study. We can therefore choose a type of dissimilarity adapted to the context and the nature of the data. The most direct method to calculate dissimilarity between objects in a multidimensional Euclidean space consists in calculating the Euclidean distances. In this thesis we use the correlation distance to compute the dissimilarity between the transcription or the DNA replication profiles of a pair of cell lines. It is computed as 1 minus the correlation coefficient value between the pair of cell lines profiles:

$$D(CL_1, CL_2) = 1 - Corr(CL_1, CL_2) \tag{2.7}$$

2.3.3 Correlation matrices representation and comparison

For the 12 cell lines cited previously, we performed comparative analyses of their DNA replication and transcriptional programs. We illustrate the general procedure we followed using the replication fork directionality profiles discussed in Chapter 3 (Figure 2.2). First,

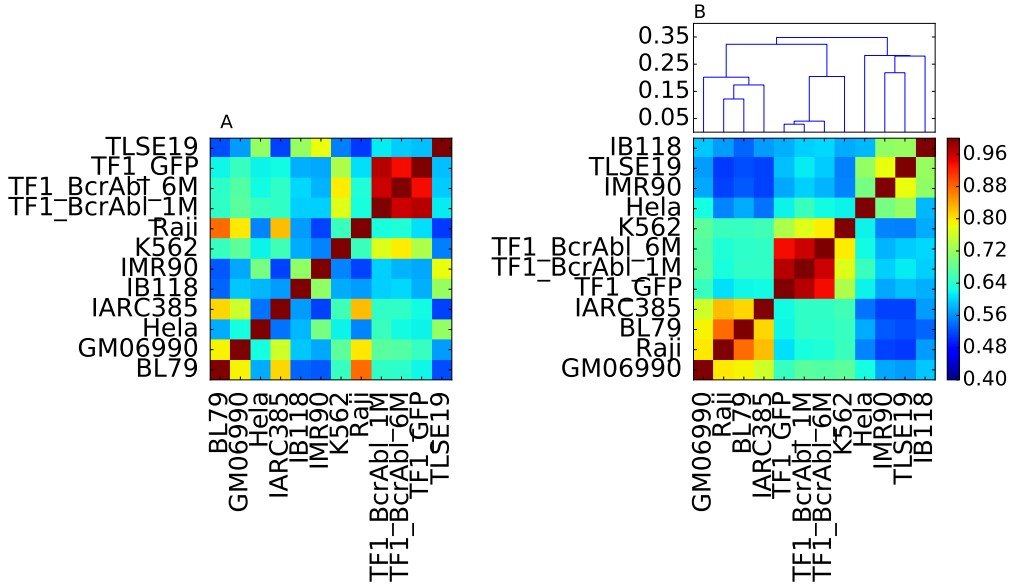

Figure 2.2: **RFD correlation matrix.** (A) RFD correlation matrix where each square correspond to a correlation coefficient value between two cell lines, cell lines are ordered by alphabetic of order. (B) (top) Dendrogram of the HAC, (bottom) same matrix as in A with a cell line order compatible with HAC.

we computed the Pearson correlation coefficients ($Corr(RFD_i, RFD_j)$) (Eq. (2.4)) between the RFD profiles of all 66 pairs of cell lines (i, j). To visualize the result, we grouped the coefficients in a symmetric matrix ($C_{RFD}(i, j) = Corr(RFD_i, RFD_j)$). As can be seen on Figure 2.2A, when arbitrarily ordering the matrix lines and columns according to cell line names alphabetical order, the correlation matrix structure cannot be easily observed. Single linkage clustering (section 2.3.2) using pairwise correlation distance (Eq. (2.7)) as the dissimilarity measure results in a hierarchical binary classification of the cell lines. Ordering the correlation matrix lines and columns according to a cell line order respecting this classification (the lines of the dendrogram representation do not cross), we obtain a correlation matrix representation where the correlation structure between cell lines is apparent (Figure 2.2B). In the example of Figure 2.2, the 3 group classification of our 12 cell lines highlighted by this dendrogram representation, nicely matches the cell type classification described in section 2.1. In the following chapters we use this representation strategy to illustrate the results of correlation analyses between genomic profiles such as RFD, MRT (C_{MRT}) and gene expression level (FPKM) ($C_{RNA-seq}$) among the 12 cell lines.

We quantify the similarity between the correlation structures of two correlation matrices in two ways. First, the classification trees can be compared, in particular we ask whether these 3 group classifications are identical or not. Second, we compute the Pearson correlation coefficient between the vectors of 66 (non-trivial) pairwise genomic profile correlation coefficients (contained in the upper triangular part excluding the diagonal of the correlation matrices). It will be referred to as the correlation coefficient between two correlation matrices.

2.3.4 Computational analyses of RFD, RNA-seq and MRT profiles

All analyses are restricted to the 22 autosomes to avoid artefacts due to the XX or XY karyotypes of the studied cell lines. RFD profile Pearson correlation C_{RFD} between a pair of cell lines was computed using non-overlapping windows with > 100 OK-seq reads in both cell lines. RNA-seq correlation C_{RNA} was computed between $\log_{10}$(FPKM) of genes expressed in both cell lines. MRT correlation C_{MRT} is computed between non-overlapping 100 kb windows with a valid MRT estimate.

Finally, when analysing GC-content, FPKM and MRT distributions in 200 kb windows depending on the change of RFD ($|\delta RFD_{C_1}^{C_2}| = |\mathrm{RFD}(C_1) - \mathrm{RFD}(C_2)|$) between two cell lines (C_1,C_2), we only consider windows with > 2000 OK-seq reads in both cell lines to guarantee that the standard deviation, using a Poisson approximation, remains < 0.023 for both RFD estimates.

2.3.5 Bioinformatic softwares

Computation and Figures are implemented in **python** (version 2.7.13) using **numpy** (version 1.11.3), **scipy** (version 0.18.3), **pandas** (version 0.23.4) and **matplotlib** (version 2.0.0) scientific computing modules. We use Jupyter notebook as an interactive platform when needed.

To compute the transcription expression level for the RNA-seq datasets we use **TopHat**, a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the high-throughput short read aligner **Bowtie**, and then analyze the mapping results to identify splice junctions between exon (Tophat 2.1.1) [135]. After alignment of the RNA-seq reads to the genome we used **cufflinks** (2.2.2) to compute the FPKM (Fragments Per Kilobase of exon model per Million mapped fragments) values.

Tophat command lines:

```
tophat-p8-G. ./genes. gtf--no-coverage-search-o. ./Tophat_ out. ./Sequence/
genomelist_ of_ Fastqfile_ per_ cell
```

-p option indicates how many threads TopHat must use to align reads. -G option provides to TopHat with a set of annotations of known gene models and / or transcripts. TopHat first extracts the transcribed sequences and uses Bowtie to align the read on this virtual transcriptome. The read that maps the transcriptome is be converted to genomic mappings and merged with the new mappings and junctions of the final output of TopHat. –no-coverage-search option disables the coverage based search for junctions, it is used to speed up computations.

The gene differential analysis is performed using the **DESeq2** [141] package and its R

version (3.5.1). We use **Featurecount** procedure to create the genes read count table across all replicates from the TopHat alignments. Differential gene expression analysis is then performed using **DESeq2** procedure on the count table.

We use **Samtools** (1.9) to read, view, create index and edit SAM and BAM format files [142]. To visualize the mapped read along the genome we use the Integrative Genome Viewer (**IGV** (2.4.8)).

CHAPTER 3

Global analysis of the DNA replication program in 12 human cell lines

3.1 Introduction

Several techniques have been used to understand the DNA replication program and describe the nature and position of replication origins in mammalian genomes [35, 59, 143–145]. The microarray hybridization and next generation sequencing based methods allows to detect and measure the abundance of replication intermediates genome wide. These technologies include SNS-seq and Bubble-seq. Using SNS-seq method, between 50000 and 250000 potential initiation sites were identified on the human genome [81, 146, 147]. In contrast, using Bubble-seq more than 100000 sites were identified in the human genome [82]. The issue using these methods is that bubble trap origins align weakly with those obtained with SNS-seq method [68, 77, 78, 82] and that different sets of origins are detcted with these two methods. Moreover, they do not provide the efficiency of the replication origin. Interestingly, Okazaki Fragments sequencing (Ok-seq) method, that maps initiation events along the whole-genome takes advantage of the deep-sequencing of strand-specific Okazaki fragments, to solve these issues. DNA replication is inherently asymmetric. About 50% of the genome is discontinuously replicated as Okazaki fragments, discovered in 1968 by Reji and Tsuneko Okazaki [148, 149]. The Okazaki fragments are short and are formed on the lagging strand during DNA replication. The Okazaki fragments, as illustrated in Figure 1.4, are in the opposite direction to the replication fork progression. Okazaki fragment synthesis on the lagging strand necessitates the repeated production of single-stranded DNA and polymerization in the opposite direction to fork progression. So sequencing the Okazaki fragments yields a direct measure of the directionality of replication fork progression location of strong firing origins and termination sites (Figure 1.7). Okazaki fragments are short, between 1000 and 2000 nucleotides long in prokaryotes and are roughly 100 to 200 nucleotides long in eukaryotes [150], and they have a short life time, estimated to few seconds. Hence, there is a low amount of Okazaki fragments in wild type cells. Hence, the biggest challenges have been to find a method to isolate the Okazaki fragments.

The Ok-seq method was first established in yeast. In this first implementation [151, 152], the isolation of Okazaki fragments was only possible thanks to the DNA ligase I repression in a degron-tagged yeast construct. For the first time, Smith and Whitehouse purified Okazaki fragment and showed that on average the Watson:Crick $\log_2$ ratio of Okazaki fragment strand changed sign when crossing Active origins in yeast [152]. In the same manner, using the deep sequencing of Okazaki fragments, the directionality of the replication fork in yeast was confirmed and modeled [151]. Very recently, Okazaki fragments sequencing (Ok-seq) was also accomplished in 2 human cells: GM06990[1] and HeLa[2] without the need for ligase inactivation [3] (for more details see section 3.2.1). This work quantified the replication fork directionality and inferred the location of replication initiation and termination zones, genome wide, at high resolution. Here we have used novel Ok-seq

[1] GM06990 is an EBV-immortalized lymphoblastoid cell line (LCL) with a near-normal karyotype
[2] HeLa is an epithelial cell line from a cervix adenocarcinoma

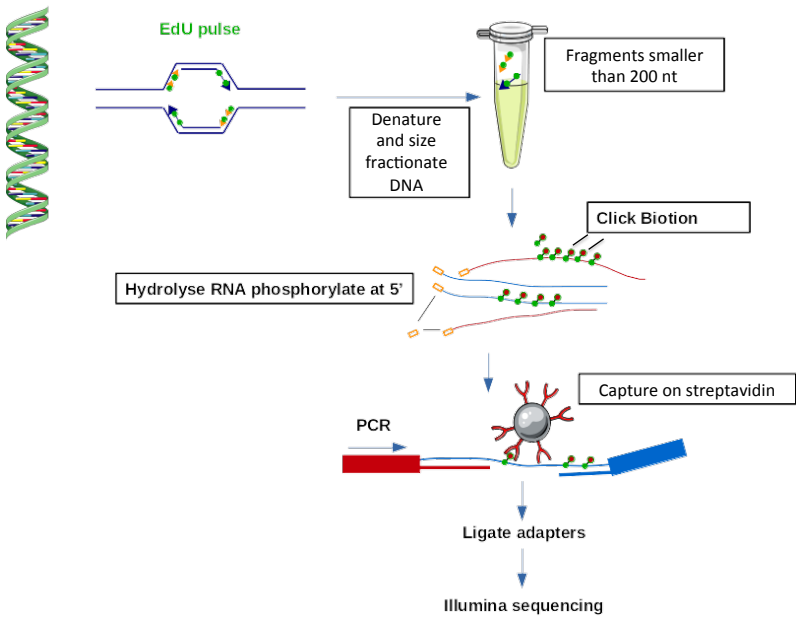

Figure 3.1: **Ok-seq protocol**. Okazaki fragment purification schema (Section 3.2.1) (see Petryk et al. [3] for more information about the method).

datasets [153] allowing us to compare the RFD profiles for a total of twelve human cell lines, including lymphoid, myeloid and adherent cell types [113]. Lymphoid cell lines, in addition to GM06990, include BL79 and Raji, two independently established Burkitt lymphoma cell lines (BLs), and IARC385, an EBV-immortalized LCL established from the same patient as BL79. Adherent cells, in addition to HeLa, include TLSE19 and IB118, two leiomyosarcoma (LMS) cell lines established from two different patients, and IMR90 human fibroblasts. Myeloid cell lines include a cellular model for establishment and progression of chronic myeloid leukemia (CML), a malignant disease characterized by the Philadelphia[3] chromosome and the formation of the BCR-ABL1 fusion gene, whose expression is necessary and sufficient for CML formation [125]. The aim of this chapter is to do a comparative analysis of the DNA replication program of these 12 human cell lines established using Ok-seq.

3.2 Profiling the DNA replication program by sequencing of Okazaki fragment

3.2.1 Purification of Okazaki fragment

The Ok-seq method in yeast requires to use conditional lethal mutants that massively increase the Okazaki fragment abundance [151, 152]. In human a new methodology was used to isolate

[3] See Chapter 5 for the definition

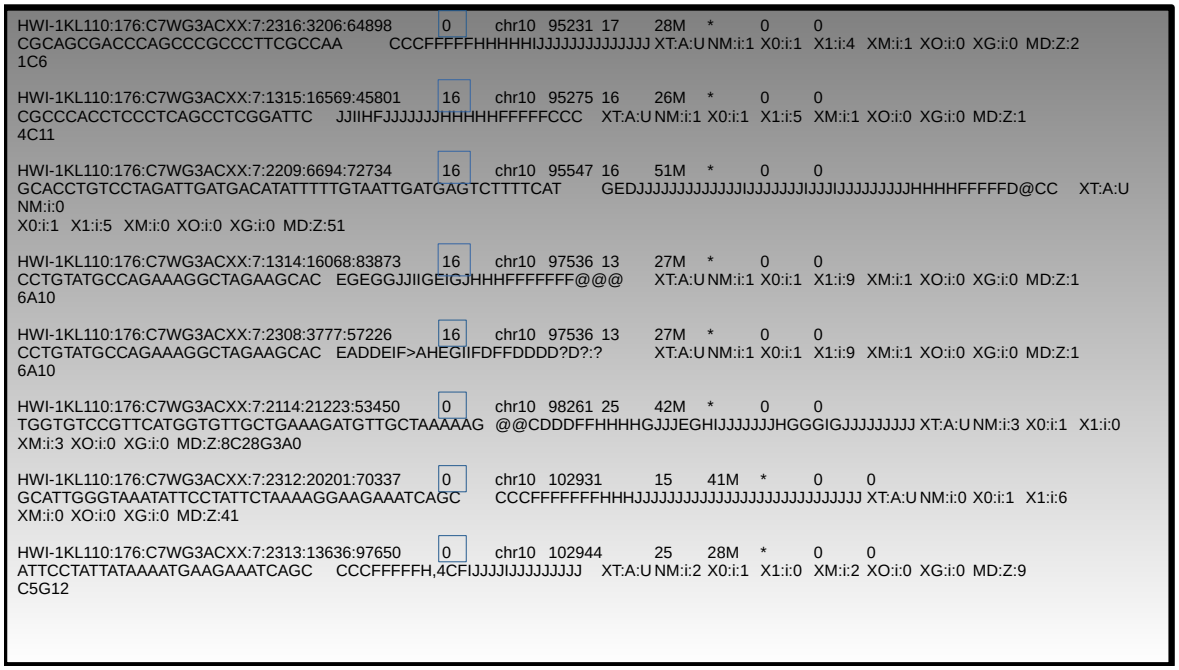

Figure 3.2: **Example of the alignment file content for one of TLSE19 biological replicates** From this file we extracted and counted the Forward strand mapping reads (flag= 0) and Reverse strand mapping reads (flag=16) to compute the RFD profiles. Flag (blue square), mapping chromosome and position are in column 2, 3 and 4 respectively. When using bwa to perform read alignment, optional field XT:A:U indicates uniquely mapping reads.

and sequence the Okazaki fragments which do not require mutant cells [3]. The new approach uses EdU as thymidine analog to label and purify nascent DNA without using genetically modified cell as in yeast. The advantage of this method is that this thymidine analog can be covalently joined to biotin or other ligands by simple click chemistry. The following steps are: denaturation, size fractionation by sucrose gradient centrifugation, collection and labeling by EdU clicked with biotin of all the small Okazaki fragments (<200 nt) (Figure 3.1). By capturing the biotin labeled DNA with streptavidin-coated magnetic beads, the Okazaki fragments are isolated. The last step is to generate Okazaki fragments libraries for sequencing and mapping the 'reads' to their strand of origin. Finally, we can use this sequenced Okazaki fragments to analyze the location and efficiency of replication origins and termination sites across the whole human genome (Section 1.3).

The technique is directly applicable to mammalian cells, which readily incorporate EdU. However, wild type yeast cells are impermeable to EdU. Therefore, the application of EdU based Ok-seq to yeast requires the use of strain modified to incorporate EdU.

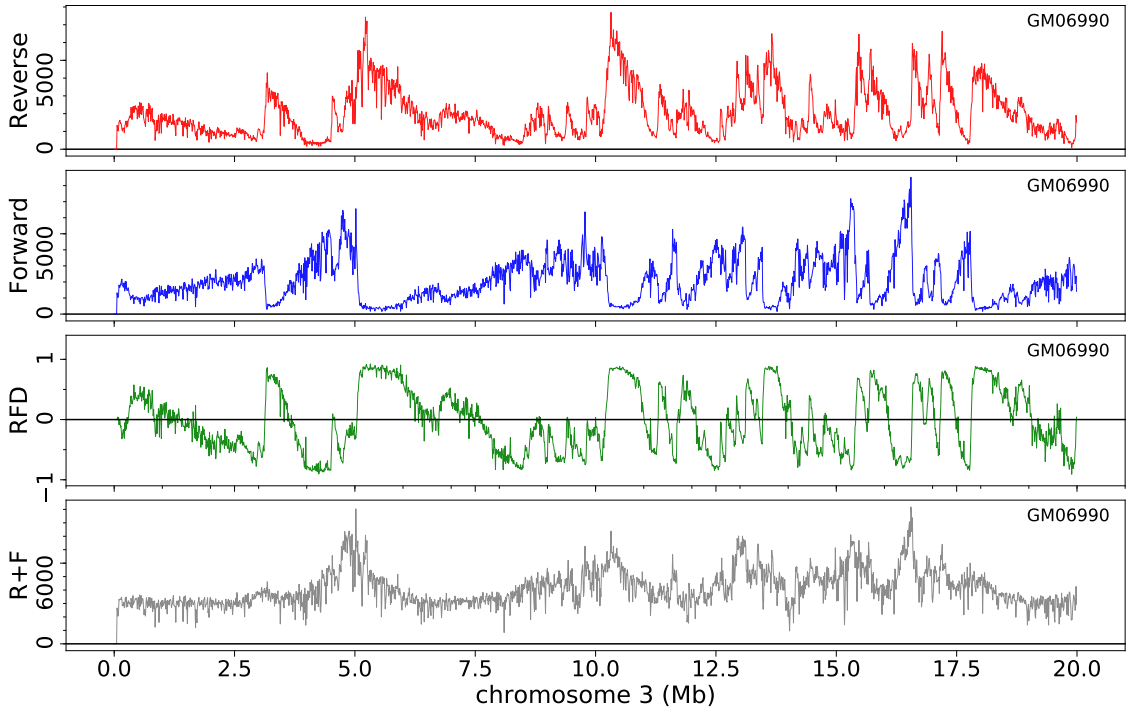

Figure 3.3: **Genome wide profiling of replication fork directionality.** From Top to bottom : number of Okazaki fragments reads mapping the reverse (R) and Forward (F) strand, corresponding RFD profile (Eq. 3.1) and total read number (R+F), computed within 10 kb non overlapping windows along 20 Mb fragment of chromosome 3 for GM06990.

3.2.2 Estimating Replication fork Directionality using OF sequencing

Sequenced reads were identified and demultiplexed using the standard Illumina software suite and adaptor sequences were removed by Cutadapt (version 1.2.1 to 1.12). Reads >10 nt were aligned to the human reference genome (hg19) using the BWA (version 0.7.4) software with default parameters. This resulted in alignment files in 'BAM' format that were downloaded from the IMAGIF sequencing platform[4] (Figure 3.2). We only considered uniquely mapped reads and we counted identical alignments (same site and strand) as one to remove PCR duplicate reads. We computed the Replication Fork Directionality (RFD) for a given window following Eq. (3.1) as:

$$RFD = (R - F)/(R + F), \tag{3.1}$$

where F represents the number of Forward strand matching reads (leftward moving forks) forks and R represents the number of Reverse strand matching reads (rightward moving forks) forks within the window; RFD takes value between -1 and 1 (Figure 3.3). In the first two top panels of Figure 3.3, we represented the number of forward reads (anti-sense forks) and

[4]Centre de Recherche de Gif-sur-yvette, CNRS

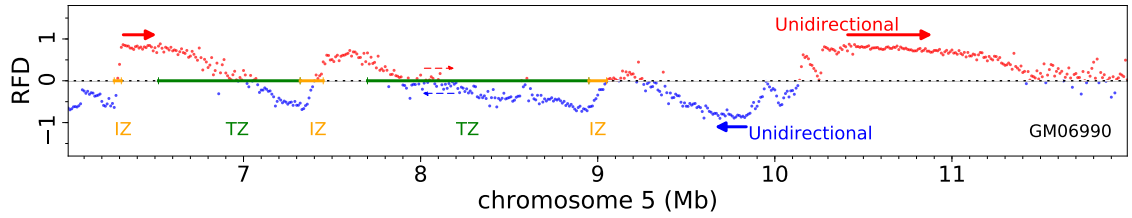

Figure 3.4: **Extraction information from Replication Fork directionality profiles.** RFD profile along a 6 Mb fragment of chromosome 5 where some zones of preferential initiation (IZ, orange), preferential termination (TZ, green) and unidirectional fork progression (thick red (resp. blue) arrows: sense (resp. antisense) fork progression) are marked; the pair of red and blue thin dashed arrows mark a region of null RFD where forks progress in equal proportion in the two directions.

reverse reads (sense forks) at 10 kb scale in a 20 Mb window. We can see that the reverse and forward reads are complementary; for example, when we have an abundance of reverse reads we have a deficiency of forward reads and vice-versa. This abundance of reads switching between forward and reverse determines the RFD profiles (Figure 3.3). Since the total amount of replication on both strands should be constant across the genome, normalization of the difference between the two strands counts by the total read count (R+F) accounts for variation in read depth due to copy number or sequence bias (Figure 3.3).

As originally described in Petryk et al. [3], RFD profiles display an alternation of quasi-linear ascending, descending and flat segments (AS, DS, FS) of varying size and slope (Figure 3.4). RFD often reached values >0.9 or <–0.9, indicating nearly complete purity of Okazaki fragments. A positive (negative) RFD value indicates that forks move predominantly rightward (leftward) in the cell population (Figure 3.4). Ascending Segment (AS) and Descending Segment (DS) therefore represent zones of predominant replication initiation (IZ) and termination (TZ), respectively (Figure 1.7). The amplitude of the RFD shift across each zone reflects its net initiation or termination efficiency. FS of high RFD are unidirectionally replicating regions. FS of null RFD, sometimes found in the middle of a TZ, are replicated equally often in both directions, presumably by random initiation and termination. A few exemplary IZs, TZs and FSs in GM06990 are illustrated for a 6 Mb segment of chromosome 5 on Figure 3.4.

3.2.3 Application to 12 human cell lines

We used Ok-seq to profile RFD genome-wide in multiple cell lines as previously described for HeLa and GM06990 [154]. As described in the previous chapter we have three types of cell lines; Four lymphoid (GM06990, Raji, BL79, IARC385), four myeloid (TF1_GFP, TF1_BCRABL_1M, TF1_BCRABL_6M, K562) and four adherent (IMR90, HeLa, TLSE19, IB118) cell lines. From 2 (IB118, GM06990) to 6 (BL79) biological replicates (BR) were sequenced per cell line (Table 3.1). GM06990_BR2 has the maximum number of filtered reads (576 millions) and TLBE19_BR1 has the minimum number of filtered reads (17 millions).

Cell line	Total number of filtered reads	Replicate	Sample name
GM06690	487286675	BR1	GM06990_BR1
	576039124	BR2	GM06990_BR2
Raji	114075570	BR1	Raji_BR1
	19116762	BR2	Raji_BR2
	31673857	BR3	Raji_BR3
	25711675	BR4	Raji_BR4
	38629936	BR5	Raji_BR5
BL79	23467045	BR1	BL79_BR1
	23146435	BR2	BL79_BR2
	25570312	BR3	BL79_BR3
	23598856	BR4	BL79_BR4
	33298270	BR5	BL79_BR5
	25745907	BR6	BL79_BR6
IARC385	51405968	BR1	IARC385_BR1
	60355939	BR2	IARC385_BR2
	47539925	BR3	IARC385_BR3
	45077665	BR4	IARC385_BR4
	48078104	BR5	IARC385_BR5
TF1_GFP	110444554	BR1	TF1_GFP_BR1
	35453132	BR2	TF1_GFP_BR2
TF1_BcrAbl_1M	212509624	BR1	TF1_BcrAbl_1M_BR1
	26698502	BR2	TF1_BcrAbl_1M_BR2
TF1_BcrAbl_6M	37171292	BR1	TF11_BcrAbl_6M_BR1
	41143513	BR2	TF1_BcrAbl_6M_BR2
K562	126730199	BR1	K562_BR1
	187257032	BR2	K562_BR2
HeLa	20015587	BR1	HeLa_BR1
	431728103	BR2	HeLa_BR2
	121305629	BR3	HeLa_BR3
IMR90	159792870	BR1	IMR90_BR1
	113705404	BR2	IMR90_BR2
TLSE19	17346782	BR1	TLSE19_BR1
	129141627	BR2	TLSE19_BR2
	40653419	BR3	TLSE19_BR3
	147103104	BR4	TLSE19_BR4
IB569	77811471	BR1	IB569_BR1
	70132616	BR2	IB569_BR2

Table 3.1: **Ok-seq sequencing statistics.** Cell lines are listed according to their order in most figures in this thesis. Number of reads in each sample are in the second column, biological replicate number and sample name in column 3 and 4.

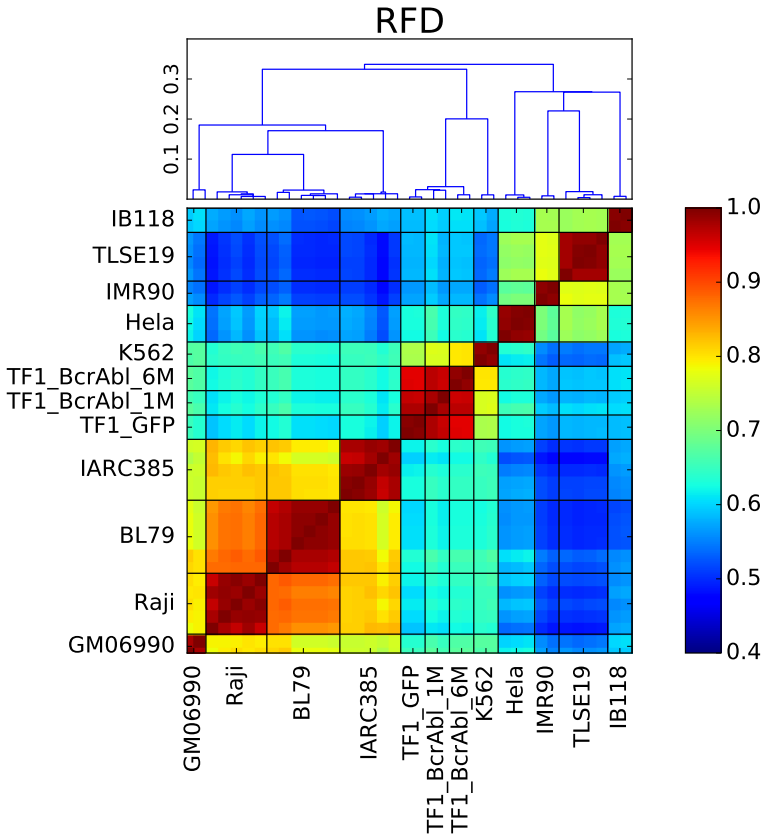

Figure 3.5: **Replicated RFD profiling are highly coherent.** Classification of RFD profiles obtained in each biological replicate experiment (Table 2.1). Classification of cell lines using RFD profiles at 50 Kb (Section 2.3.3). Pearson correlation coefficient values are color-coded from blue (0.4) to red (1) using the color bar on the rights. Each master-square in the diagonal correspond to the correlation value between biological replicates of the same cell line.

Note that filtered reads for GM06990 were obtained directly from the authors of [3]. The total number of filtered reads per cell line ranged from 78.4 x 10^6 (TF1_BCRABL_6M) to 1063.3 x 10^6 (GM06990).

RFD profiles from biological replicates of each cell lines were highly correlated, with Pearson correlation computed between 50 Kb non-overlapping windows with >100 mapped reads (R + F) ranging from 0.962 to 0.997. This is illustrated in the Figure 3.5, the correlation values between the biological replicate (BR) of the same cell line appear as red squares. The hierarchical classification (section 2.3.3) represented as a dendrogram (top of Figure 3.5), groups all BR first because of the strong similarity between them. Note that the Pearson correlation between two technical replicates of IMR90 primary cells is 0.996 and their correlations with an RFD profile of an immortalized IMR90-hTERT[5] cell line is 0.989 and 0.992 (Appendix, Figure A.1), indicating that immortalization by hTERT did not affect the replication program at 50

[5]Telomerase reverse transcriptase 'hTERT' immortalizes various normal cells in culture, thereby endowing the self-renewal properties of stem cells to non-stem cells cultures.

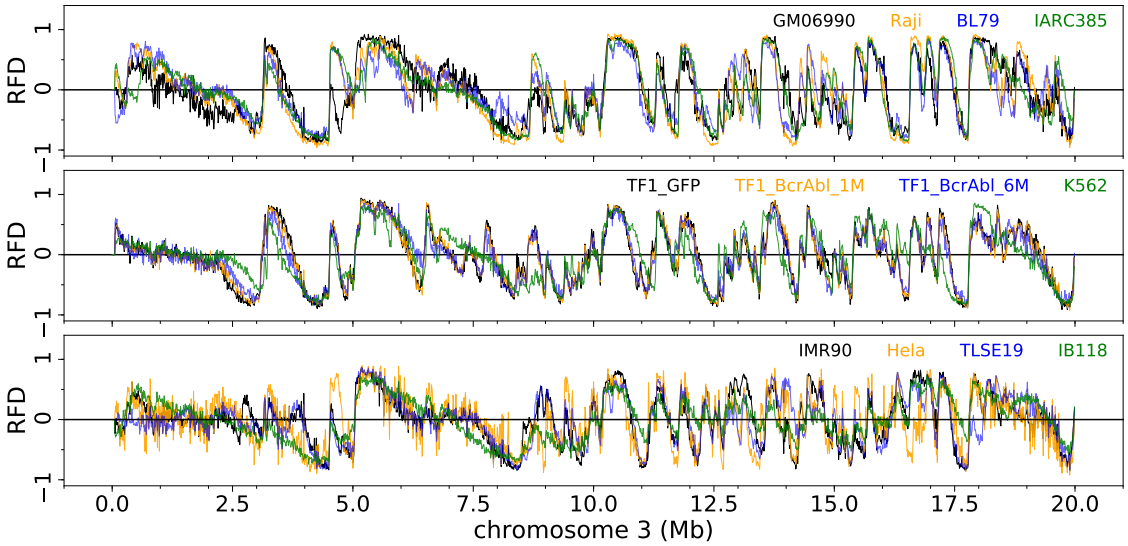

Figure 3.6: **Genome-wide profiling of replication fork directionality.** (A) Replication fork directionality (RFD) along a 20 Mb segment of chromosome 3 in 4 lymphoid (top), 4 myeloid (middle) and 4 adherent (bottom) cell lines, as indicated in the top right corners of each panel.

Kb resolution. Primary IMR90 and IMR90-htert were thus considered as the 2 BR of IMR90; IMR90_BR1 and IMR90_BR2 respectively. All BR of a given cell line have the same behavior with the other cell lines. For example, the two BR of GM06990 cell lines are more correlated to Raji and BL79 than to IARC385 with similar correlation coefficient values. Based on this observation, all BR were pooled for each cell line to produce one RFD profile per cell lines in the following analyses.

3.3 Cell lines classification based on DNA replication program

3.3.1 Cell lines classification following tissue of origin

The 12 RFD profiles of a 20 Mb segment of chromosome 3 are shown on Figure 3.6. Both shared and cell-type specific RFD patterns were observed. For example, a region replicated differently in GM06990 from other lymphoid cell lines and a region replicated similarly in all lymphoid cell lines are visible at 4.5–5.0 Mb and at 10–12 Mb, respectively, on Figure 3.6. The similarity of RFD profiles among cell lines of each group is very clear and illustrates the robustness of the RFD profiling method. However, differences in RFD profiles are informative about changes in replication program during cell differentiation. As described in section 2.3.3, to compare the 12 genome-wide RFD profiles at 10 kb, we computed the 66 pairwise correlation coefficients and collected them in a global RFD correlation matrix $C_{RFD_{10kb}}$. Hierarchical classification based on the derived correlation distance (Eq. 2.7) allowed us to order cell lines such that the correlation matrix structure be readily apparent (Figure 3.7A) (see section 2.3.3). All the differences observed among the cell lines are larger than the variation between the biological

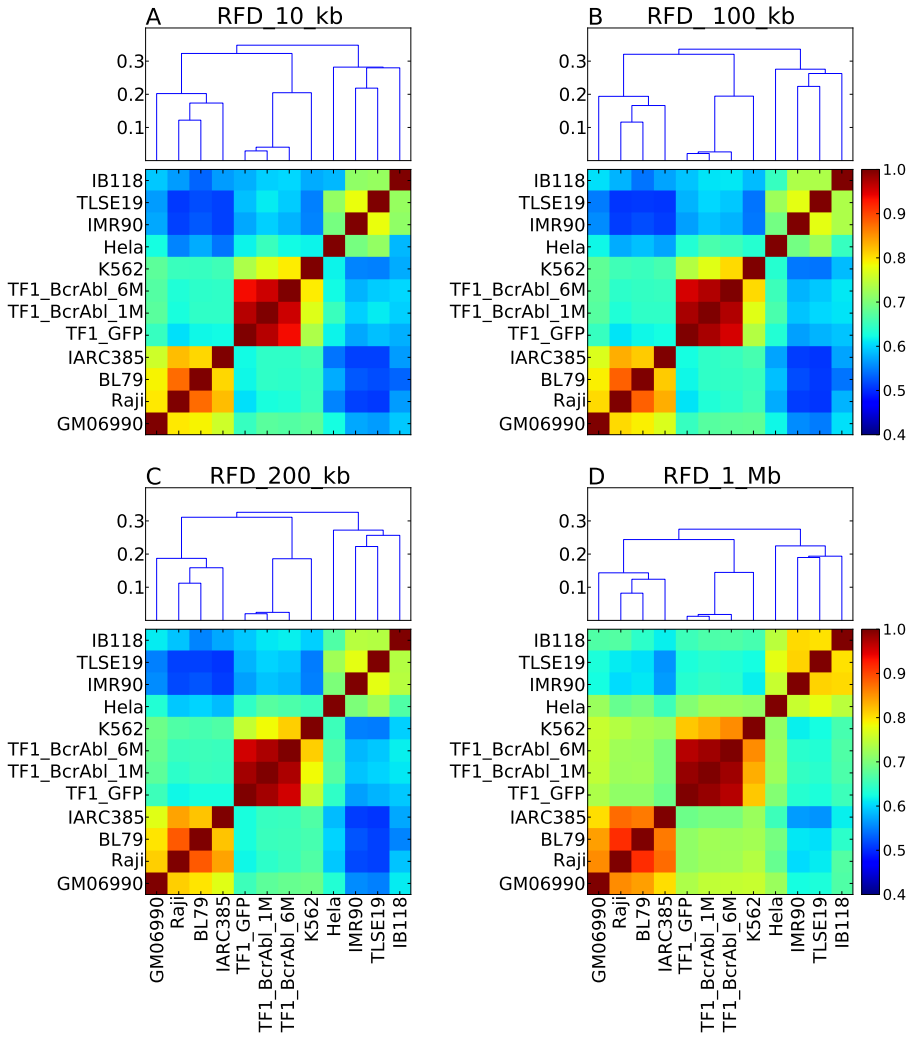

Figure 3.7: **Analysis of the scale dependence of the cell line classification based on RFD profiles.** Correlation analysis and hierarchical classification of RFD profiles (Section 2.3.3) were performed at 10kb, 100 kb, 200 kb and 1 Mb scales. The same classification of cell lines was obtained at all scales.

replicates (Figures 3.5 and 3.8). Lymphoid, myeloid and adherent cells formed 3 separate groups. Within-group correlation distances were similar by RFD so that the three groups were recovered by cutting the dendrogram at level 0.3.

Within the lymphoid group, Raji and BL79 are two BL cancer cell lines that are highly correlated between each other. Surprisingly, the two normal lymphoid cell lines GM06990 and IARC385 are more correlated to the BL cell lines (Raji and BL79 cell lines) than to each other. All types of BLs are characterized by dysregulation of the C-MYC gene, located on 8q24, by one of three possible chromosomal translocations. Karyotype analysis showed that BL79 carries a t(8;14)(q24;q32) translocation confirming its identification as a BL cell line. In contrast, a large fraction of IARC385 cells contained a t(4;11) translocation, which is not typical of BLs, but none of the three possible diagnostic translocations of BLs. In addition, genome-wide

CGH array analysis revealed few copy number variations (CNVs) in GM06990 and BL79, more in Raji and even more in IARC385 (Figure 2.1). We investigated whether Copy Number Variation (CNVs) may affect the classification of cell lines. Filtering out aneuploid regions detected by CGH array analysis in the lymphoid cell group (Figure 2.1) did not affect the classification. Thus, the RFD correlation analysis was robust to CNVs determined by CGH arrays.

Within the myeloid group, RFD profiles clustered in accordance to CML progression (Figure 3.7 A). Profile differences accumulated with BCR_ABL1 expression time in TF1, which increased resemblance to K562, a late CML. CML progression system will be specifically analyzed in chapter 5. For the adherent cell lines, including the 2 LMS cell lines (TLSE19 and IB118), the strongest resemblance was observed between TLSE19 and IMR90. TLSE19 was only slightly more correlated to IB118 than to HeLa. IB118 was only slightly more correlated to TLSE19 than to IMR90, but was more distant to HeLa. This suggests that the cancer cell lines are classified based on their tissue of origin. Thus cell type of similar morphologies, and similar differentiation pathways, tend to present similar RFD profiles.

We repeated this genome-wide analysis for three other resolutions (100 kb, 200 kb and 1 Mb, Figure 3.7 BCD). The same classification of cell lines was obtained as for 10 Kb resolution. This showed the robustness of RFD to preserve the classification in high and low resolution. In the majority of the following RFD analyses, we chose the 10 kb resolution.

3.3.2 The changes in RFD profiles are widespread among the genome

In this section, we are interested to know if the information is localized in specific regions or is distributed in the whole genome. Firstly, to investigate whether RFD differences between cell lines were concentrated in particular chromosomes, we repeated the analysis shown in Figure 3.7A for each chromosome separately (Appendix Figure B.1). The results obtained for the entire genome were recapitulated for each chromosome, with minor exceptions detailed in the legend of Figure B.1. To quantify these results, we computed the correlation between the genome wide $C_{RFD_{10kb}}$ and the correlation matrix of RFD for each chromosome $C^{chr_i}_{RFD_{10kb}}$ using the procedure defined in section 2.3.3 . The correlation between the global RFD correlation matrix and each chromosome's RFD correlation matrix was high (>0.893), albeit sensitive to chromosome size as expected (first raw of matrix and last column in the matrix of Figure 3.8). We similarly computed the pairwise correlation coefficient between individual chromosomes RFD correlation matrices (Figure 3.8). Correlations between individual chromosomes were also high (>0.79), but again sensitive to chromosome size (Figure 3.8). The minimal value of correlation is between chromosomes 16 and 22 (Figure 3.8). This result suggests that the information obtained from RFD profiles is not due to specific regions, but is distributed among all chromosomes.

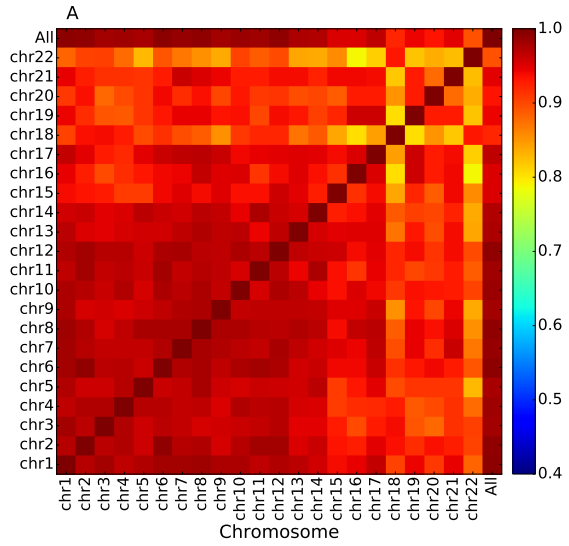

Figure 3.8: **Replication changes between cell lines are widespread through the genome.** Correlation coefficients between the global genome correlation matrix between RFD profiles shown in Figure 3.7 A and the individual chromosomes' correlation matrices shown in Figure B.1. Pearson correlation coefficient values are color-coded from blue (0.4) to red (1) using the color bar on the right.

We stratified the genome using the GC content classes described by Bernardi [155]. Why did we choose the GC content? Simply because it is common to all cell lines, and is associated with MRT and transcription, so that it is a convenient substitute for genome organization For each class of GC content we computed the RFD correlation matrix and performed the hierarchical classification illustrated by the dendrogram. The classification and the structure of the correlation matrix do not change (Figure 3.9) so we deduced that the differences of RFD profiles among the twelve cell lines are also widespread among the isochore families.

To more precisely assess how widespread RFD changes are, we generated a large number of random probes, 50 kb to 50 Mb in size, consisting each of 5–5000 randomly located 10 kb windows. For each probe, we computed its RFD correlation matrix and the correlation with the global genome correlation matrix. We then determined the statistical fluctuation of the latter correlation and the probability of observing the same classification of cell lines as with the global genome (Figure 3.10). We found that a random probe size $\geq$5 Mb was sufficient to reach a $>$90% correlation with the global genome RFD correlation matrix, and to observe with a probability $>$0.95, the same three group classifications of cell lines. These latter result shows that cell line classification is not caused by outliers data point but represent the data in general.

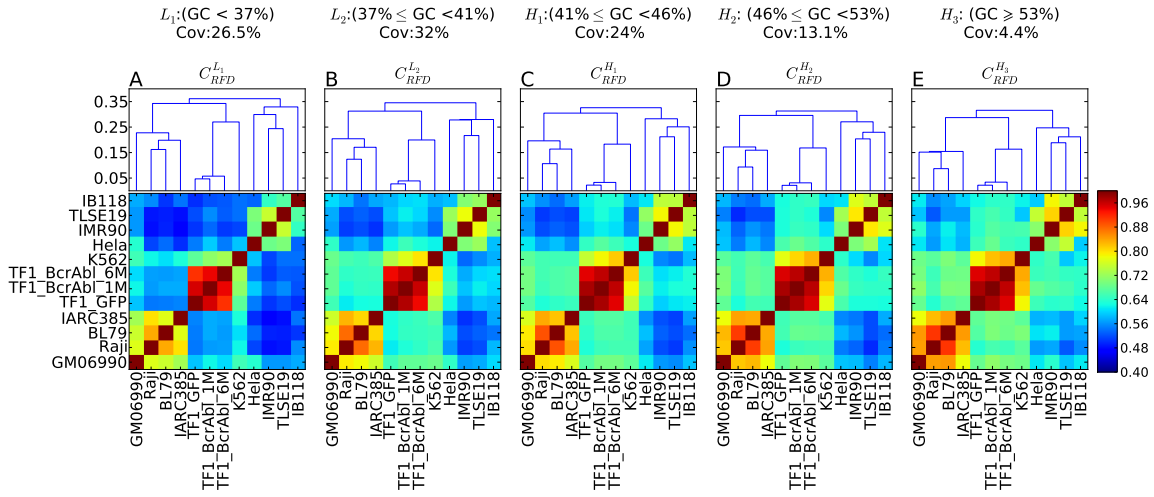

Figure 3.9: **RFD profiles are more conserved in high GC-content regions.**(A-E) Correlation matrix of RFD profiles depending on the GC content; 10 kb windows were grouped in GC-content categories following the 5 isochores classification of the human genome in light isochores L1 ($GC \leq 37$; $C^{L_1}_{RFD}$; A) and L2 ($37 \leq GC < 41$; $C^{L_2}_{RFD}$; B), and heavy isochores H1 ($41 \leq GC < 46$; $C^{H_1}_{RFD}$; C), H2 ($46 \leq GC < 53$; $C^{H_2}_{RFD}$; D) and H3 ($GC \geq 37$; $C^{H_3}_{RFD}$; E); coverage (Cov) of the sequenced human genome by each isochore family is provided in the header of each column; Pearson correlation coefficient values are colour-coded from blue (0.4) to red (1) using the colour bar on the right (Materials and Methods).

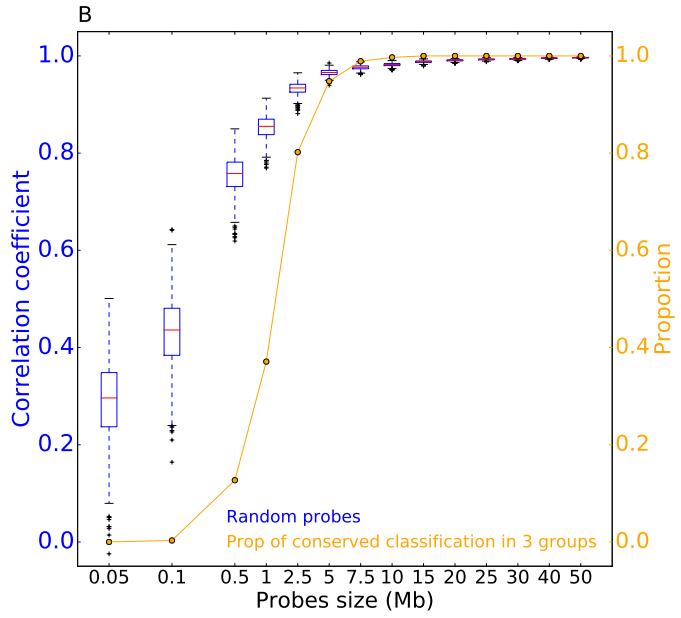

Figure 3.10: **Random 5 Mb probes are sufficient to recover the cell line classification.** (Blue, boxplot) Distribution of the correlation coefficient between the RFD correlation matrix for the complete genome (Figure 3.7A) and 1000 random probes of indicated of scale (Mb) consisting of 5 to 5,000 randomly located 10 kb windows. (Orange), probability of observing the same 3 groups classification of cell lines as in Figure 3.7A, when performing the classification using a random mode of the indicated scales.

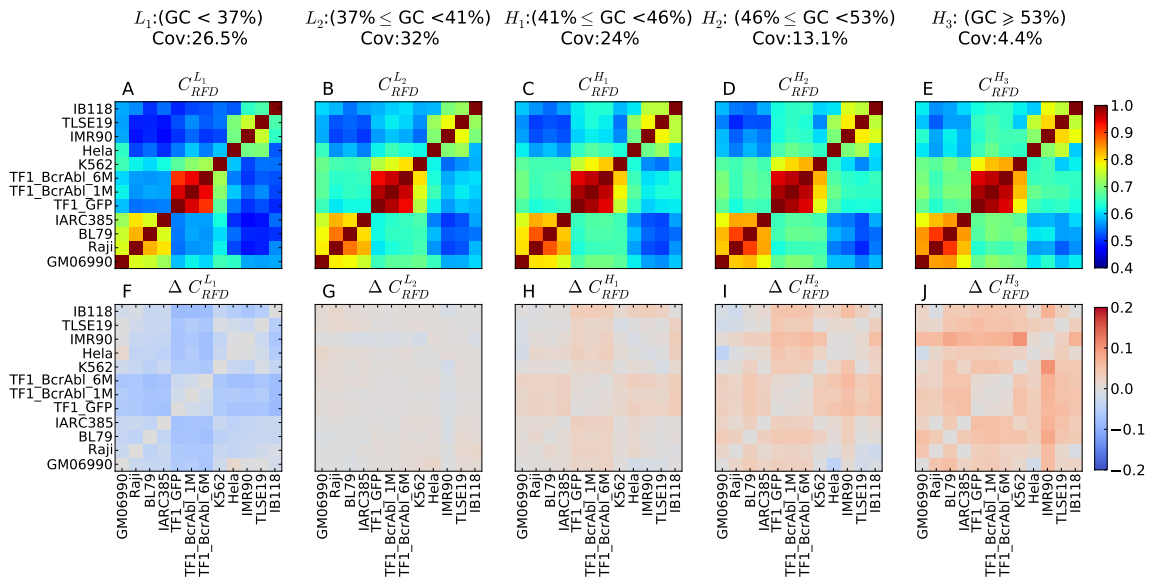

Figure 3.11: **RFD profiles are more conserved in high GC-content regions.** (A-E) Correlation matrix of RFD profiles depending on the GC content; 10 kb windows were grouped in GC-content categories following the 5 isochores classification of the human genome in light isochores L1 ($GC \leq 37$; $C^{L_1}_{RFD}$; A) and L2 ($37 \leq GC < 41$; $C^{L_2}_{RFD}$; B), and heavy isochores H1 ($41 \leq GC < 46$; $C^{H_1}_{RFD}$; C), H2 ($46 \leq GC < 53$; $C^{H_2}_{RFD}$; D) and H3 ($GC \geq 37$; $C^{H_3}_{RFD}$; E). (F-J) Matrices of correlation differences ΔC^{I}_{RFD}=C^{I}_{RFD}−C_{RFD} where I can be L1, L2, H1, H2 or H3; correlation difference values are color coded blue (resp. red) for negative (resp. positive) differences using the color bar on the right; a blue (resp. red) color indicates that RFD profiles are less (resp. more) conserved in the considered isochore than in the 22 autosomes. Matrices row and column order is the same as in Figure 3.7A.

3.4 Characterizing the heterogeneity of the DNA replication program

We have shown that 5 Mb probes are sufficient to capture the DNA replication program changes between cell lines. However, we visualized that this is not the case in 5 Mb windows (maximally localized probes). The replicative program is in fact affected by characteristics of the genome, like the GC content. This reveals an heterogeneity along the genome in the behavior of the changes of RFD profiles among cell lines. In this section, we will determine these regions, and classify the genome into replication stable and variable regions. Characterization of these regions, by linking to other genomes features, can help us to decipher the origin of this heterogeneity.

3.4.1 RFD profiles are more conserved in high GC content regions

In the previous section, we showed that the global classification of RFD profiles among the 12 cell lines is conserved when we stratified the genome using GC content levels. But, we can observe that the amplitude of the correlation coefficient in the correlation matrix increase simultaneously with the increase of GC content levels (Figure 3.11A-E). In other words, the RFD profiles among the 12 cell lines in the rich GC regions are more correlated than in poor

GC regions. Furthermore, the myeloid cell lines are more correlated to lymphoid cell lines in L1, L2, H1 regions than to adherent cell lines (Figure 3.11A-C). This suggests that even in low GC content regions we can find some high correlation between myeloid and lymphoid cell lines (Figure 3.11A). However, a better correlation between myeloid and adherent cell lines is found in high GC content regions. In contrast, even in GC-rich regions we do not have a correlation between lymphoid and adherent cell lines (Figure 3.11E). These observations can be explained by the tissue of origin of cell lines. The lymphoid and myeloid cell lines are the part of blood cells that are produced in the bone marrow, concomitantly to a specific inter-group correlation.

When the correlation coefficient differences between each GC-content class (Figure 3.11A-E) and the entire genome (Figure 3.7A) were computed (Figure 3.11F-J), most differences were negative in L1, null in L2 and increasingly positive in H1 to H3. In other words, the RFD profiles were less, equally, or more similar to each other in the L1, L2, or H1-3 fractions, respectively, than in the total genome. Therefore, RFD changes were more frequent in the GC-poor fractions of the genome. There were a few exceptions to the general trend of increasing RFD correlation in the L1<L2<H1<H2<H3 order (Figure 3.11). The correlations between HeLa and GM06990, and between IB118 and Raji, decreased rather than increased with GC content, in the order H2<H3<H1<L1<L2. In addition, the order was L1< L2<H3<H1=H2 for HeLa vs. IB118, H1=L1<H2=L2<H3 for HeLa vs. IMR90 and L1<H3<L2<H1<H2 for IARC385 vs. Raji. It was previously observed that in HeLa and GM06990, many IZs border active genes, and isolated genes expressed in only one cell type are flanked by IZs only in that cell type [3]. Given that genes are enriched in GC-rich isochores [50], such transcription-dependent changes in IZ activity are expected to decrease the RFD profile correlation more strongly in GC-rich than in GC-poor regions. The comparison of IZs in GM06990 and HeLa also revealed that in addition to these gene-bordering IZs, many cell-type specific IZs are found away from active genes, in late-replicating [3] GC-poor regions. Since GC-poor isochores form a much larger fraction of the genome than GC-rich isochores, the net density of RFD changes between HeLa and GM06990 is higher in GC-rich than in GC-poor regions, due to a predominant contribution of transcription-associated changes in GC-rich regions. Inversely, the net density of RFD changes between other cell lines is most often higher in GC-poor regions, suggesting a predominance of transcription-independent RFD changes in GC-poor regions.

Note that, low GC content level 'L1' corresponds to 26.5% and 'L2' corresponds to 32% of the genome an inversely high GC content level 'H3' corresponds to 4.4% of whole genome. To verify if within each isochores group we have a homogeneity of correlation coefficient value among the 12 cell line. We stratified the genome into 10 GC content decile groups (Appendix, Figure A.2, A.3). The aim of this step is to observe within each GC group, how the replicative changes are associated to GC content. For example, using the new stratification 'L1' and 'L2' are stratified into 5 subgroups. The global structure of correlation matrix in each subgroup is similar to the global correlation matrix of RFD among the 12

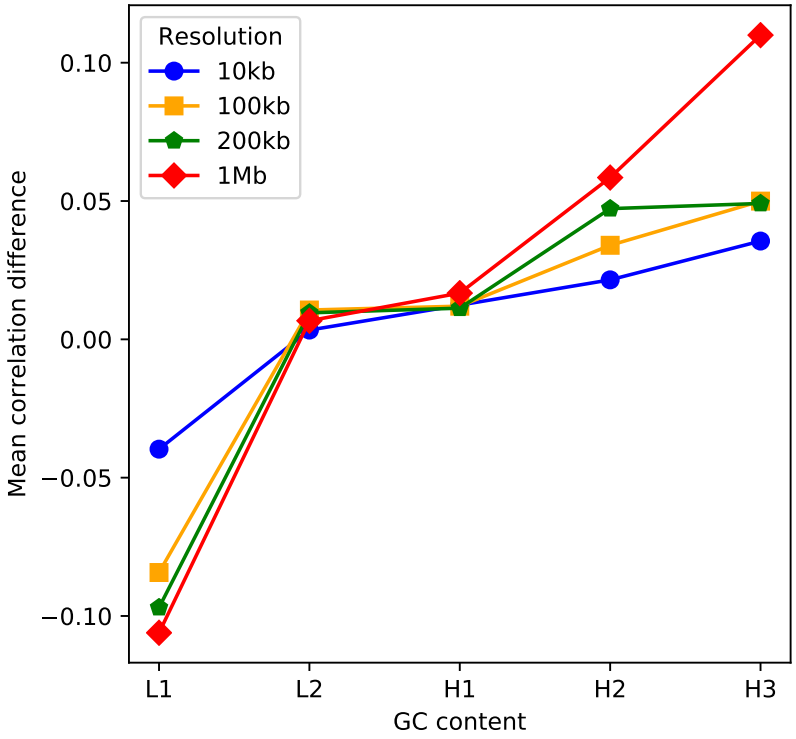

Figure 3.12: **Analyses of the scale dependence of the mean correlation difference between isochore- and global genome-based RFD correlation matrices.** The mean values of correlation differences $\Delta C^I_{RFD} = C^I_{RFD} - C_{RFD}$ (where I can be L1, L2, H1, H2 or H3) were computed for each isochore as in Figure 3.11. F-J at 10kb, 100 kb, 200 kb and 1 Mb scales. The mean correlation differences increase with GC-content at all analysis scales.

cell lines. The difference matrices of correlation of RFD profiles progressively progress from blue to light blue (Appendix, Figure A.2) and from light red to red (Appendix, Figure A.3). Thus, the new stratification showed that the GC dependence of the replicative changes is also observed within isochores classes. We obtained the same results and interpretation as with the previous stratification (Figure 3.11).

To test if the more frequent RFD changes in GC poor regions are not due to technical noise, we analyzed in Figure 3.12 the mean value of RFD difference correlation matrix between the global matrix of correlation of RFD profiles and each matrix of correlation of RFD profiles for the 5 isochores. If the RFD changes were due to noise differences, correlations differences should vanish when the scale of analysis is increased. However, the cell lines classification (Figure 3.7A) and the GC dependence of correlation differences (Figure 3.11) were conserved or even enhanced when the scale of analysis was increased from 10 kb to 100 kb, 200 kb and 1Mb (Figures 3.7 and 3.12). Therefore, these observations were not due to a higher technical noise in GC-poor regions.

We have seen in this section that RFD profile changes between cell lines do not behave the same along the genome and depend in particular on the concentration of the GC. Therefore,

can we find some regions with a high GC content that have a high level of modifications in RFD profiles among the 12 cell lines? To address this issue we have to classify the regions according to the stability of their replication program.

3.4.2 Spatial heterogeneity of DNA replication program variability in 5 Mb windows

To answer the question that we addressed at the end of the previous section, we decided to divide the genome into 5 Mb windows. The 5 Mb scale was chosen because at this scale statistical noise is small enough to not interfere with cell lines classification, based on the result of the random probes approach (Figure 3.10). We computed for each non overlapping 5 Mb window the RFD correlation matrix. We observed, for example, among all 5 Mb non overlapping windows of chromosome 1 (Appendix, Figures A.4) that the classification of the 12 cell lines using RFD profiles is not conserved in the majority of windows (totally blue or red matrices). Even though some 5 Mb regions that conserved the classification of cell lines in three groups. Moreover, we observed a strong association within the myeloid cell lines along the genome, illustrated by a red square in the majority of the RFD correlation matrices (Appendix, Figure A.4). In Figure 3.13A, we represented a 5 Mb window selected from the chromosome 1 (35-40 Mb) where the RFD profiles of 12 cell lines are similar. The correlation matrix of these RFD profiles (Figure 3.14B) illustrates that the RFD profiles among 12 cell lines are very correlated and that cell line classification according to 3 types is lost due to replication program proximity. In contrast, a disturbed RFD correlation matrix corresponding to variable RFD profiles among the 12 cell lines is observed in the region between 185 Mb and 190 Mb of chromosome 1 (Figure 3.14).

In Figure 3.15 we compare the correlation coefficient between global correlation matrix and RFD correlation matrix when using windows rather than probes (3.10) (5 kb to 50 Mb). For all scales between 5 kb and 50 Mb, we observe that the correlation of the complete genome RFD correlation matrix with the RFD correlation matrix computed in windows is smaller than with RFD correlations obtained with random probes of the same sizes. For example, for the distributions of the correlation coefficients of 5 Mb windows (Green boxplot) and 5 Mb probes (Blue boxplot) we see that the distribution of correlation coefficients of the first one vary between $\sim$0.5 and $\sim$0.75 and vary between $\sim$0.95 and $\sim$1 in the second one (Blue boxplot) (Figure 3.15). This suggests that we have more heterogeneity of RFD profiles changes among the 12 cell lines in 5 Mb window than expected scale 5 Mb but also at for all scales >50 Mb.

Identification of regions with variable DNA replication program

To identify the regions where the RFD profiles of the 12 cell lines are similar (as in Figure 3.13) or very dissimilar (as in Figure 3.14), we computed the correlation difference matrix

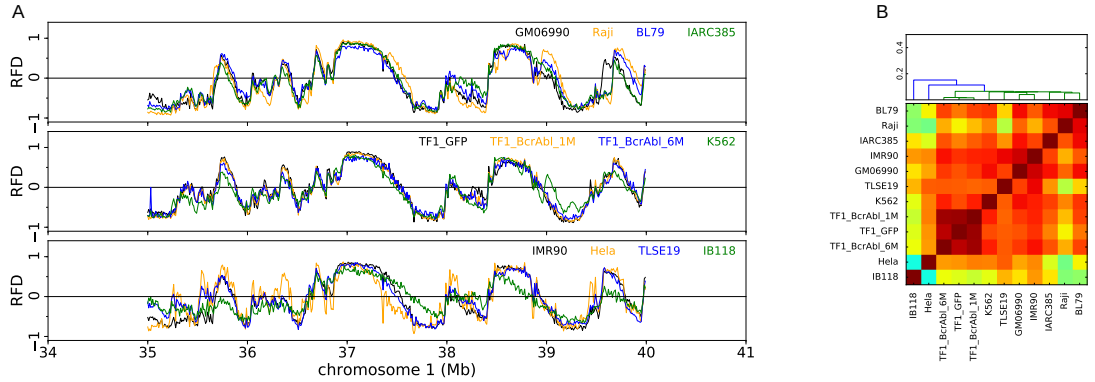

Figure 3.13: **Example of 5 Mb window with a highly stable replication program.** (A) RFD of the 12 cell lines in a 5 Mb regions. (B) Correlation matrix of the RFD profiles presented in A.

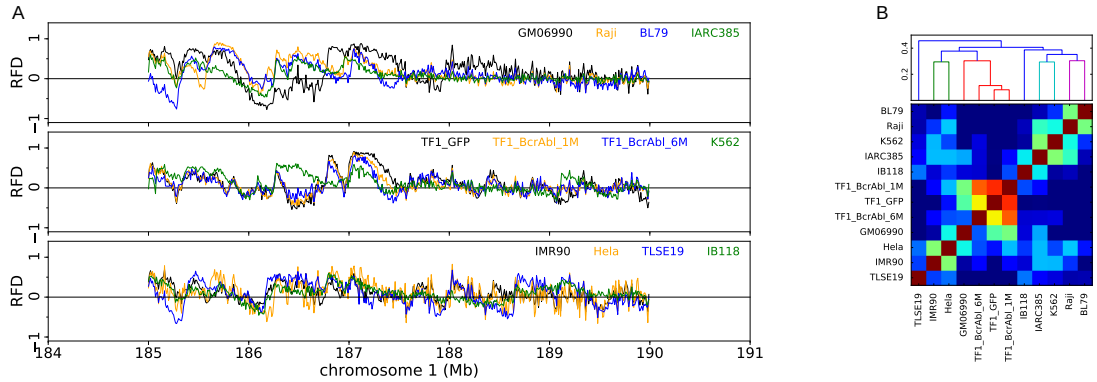

Figure 3.14: **Example of 5 Mb window with a highly variable replication program.** (A) RFD of the 12 cell lines in a 5 Mb regions. (B) Correlation matrix of the RFD profiles presented in A.

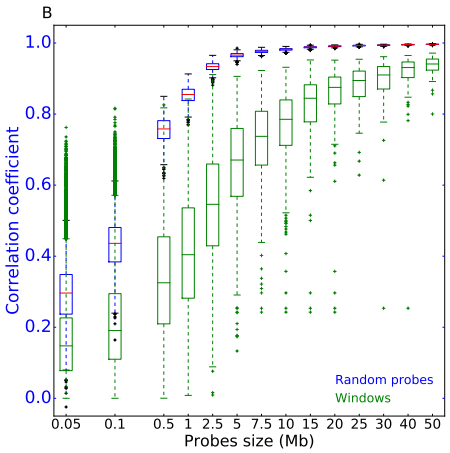

Figure 3.15: **Cell line classification is not recovered with 5 Mb windows.** (Blue) Boxplot representation of the distribution of the correlation coefficient between the RFD profile correlation matrices obtained on complete genome and for random probes of a given scale (same data as in Figure 3.10). (Green) Same analysis but for windows (made of adjacent 10 kb windows) of same scale as the random probes.

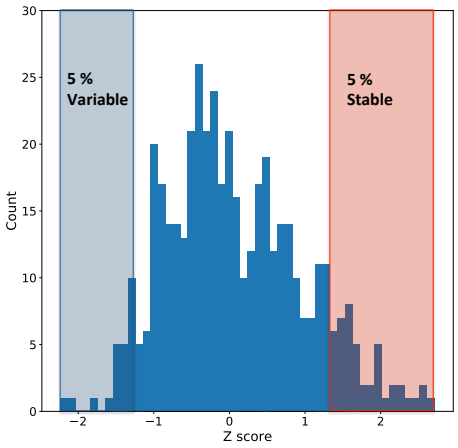

Figure 3.16: **Distribution of Z-score values.** We represented the histogram of the ratio Mean by standard deviation of the differences correlation matrix between each 5Mb window and the global one. Blue hatched the 5% unstable regions, red hatched the 5% stable regions of genome.

between each 5 Mb window W and the global matrix ($\Delta C^{W}_{RFD} = C^{W}_{RFD} - C_{RFD}$). We expected that the variable RFD profiles regions among the 12 cell lines are related to a negative mean of correlation difference matrix. In contrast, the stable RFD profiles regions among the 12 cell lines are associated to positive mean of correlation difference matrix. Thus, we computed a Z-score of each of the correlation difference matrix. This Z score was computed as the ratio between the mean and standard deviation values of RFD correlation differences:

$$Z^{w} = \frac{\frac{1}{n}\Sigma_{i=1}^{n} x_i}{\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{3.2}$$

where x_i is the set of n=66 pairwise correlation differences in window w. The histogram in Figure 3.16 showed that we have a large distribution of Z-scores, thus in many 5 Mb window regions the global cell lines classification based on RFD profiles is disturbed. To validate this method, we computed the Z-score for the GC-rich (H3, 4% of genome, Figure 3.11J) regions where the RFD profiles among the 12 cell lines were estimated as stable and we obtained 1.76, which corresponds to a stable RFD profiles among the 12 cell lines. In contrast we computed the Z-score for the GC-poor regions (L1, 26% of genome, Figure 3.11F) and we obtained -1.86, which corresponds to variable RFD regions in agreement with the previous results (section 3.4.1). Then, we selected the very similar regions by taking the highest 5% Z-score values (z>1.64). We selected the 5% lowest Z-score (z< -1.25) to identify the regions where the RFD profiles vary widely among the 12 cell lines (Figure 3.16). Then, to obtain further characterization of these regions, we compared the GC, FPKM[6] and MRT between the detected regions of stable and variable replication program (Figure 3.17). The regions with

[6]Fragments Per Kilobase of exon model per Million mapped fragments is a unit of gene expression level as estimated by RNA-seq profiling

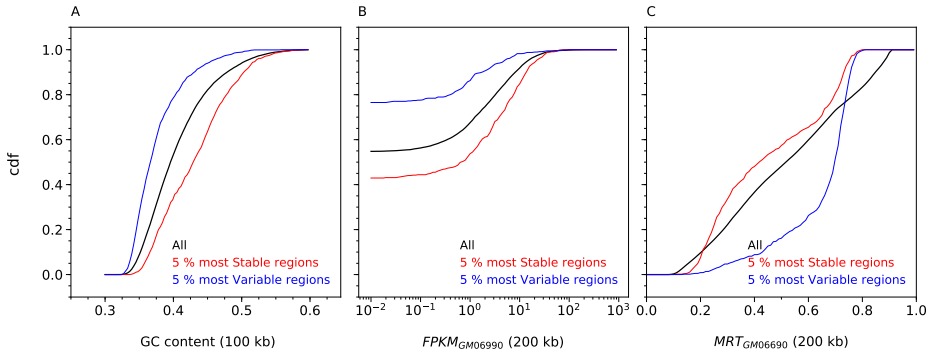

Figure 3.17: **Replication variable regions are late replicating, lowly expressed and GC poor regions.** A) Cumulative distribution function of GC content. B) Cumulative distribution function of FPKM computed in 200kb windows (Eq. 2.2). C) cumulative distribution function of MRT. Red curves corresponds to the 5% most stable regions. Blue curves corresponds to the 5% most variable regions. Black curves corresponds to the whole genome.

stable (resp. variable) RFD profiles are more associated to GC-rich (resp. AT-rich) where 50% of the regions have a GC above (resp. under) 0.42 (resp. 0.36), thus they predominantly belong to H1, H2 and H3 isochores (resp. L1). The stable (resp. variable) replication regions are highly (resp. lowly) expressed, 50% (resp. 10%) of them have a FPKM>1 (Figure 3.17). The median of the MRT in the variable regions is 0.7, this suggests that the changes in RFD profiles tend to be in late replicating regions. However, in the stable regions, it is equal to 0.4. This shows that 50% of these regions are in early replicating regions. The method used to quantify the transcription level by FPKM and Mean Replication timing will be detailed in the following chapter. In conclusion, replication program stability appears more heterogeneous than expected (Figure 3.15) in a manner correlated to genome organization (Figure 3.17).

3.4.3 Mapping of DNA replication program conservation along the human genome

Mean Correlation of RFD (MCR)

We introduced a novel method to map replication program conservation along the genome. This method is based on the RFD profiles of the 12 cell lines. First, we computed the mean of the 12 RFD profiles in each locus x at 1 kb resolution:

$$\mathrm{D(x)} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{RFD_i(x)}, \tag{3.3}$$

where x is a window of 1 kb, with RFD_i(x) is the RFD profiles of the i^{th} cell line and n=12. Then we computed for a given window (W), the correlation coefficient ($\mathrm{C_{1,i}(W)}$) between the mean RFD profile D(W) and each RFD profiles for the 12 cell lines RFD_i (W). In the final step, we calculated the mean of these correlation coefficients for the 12 cell lines:

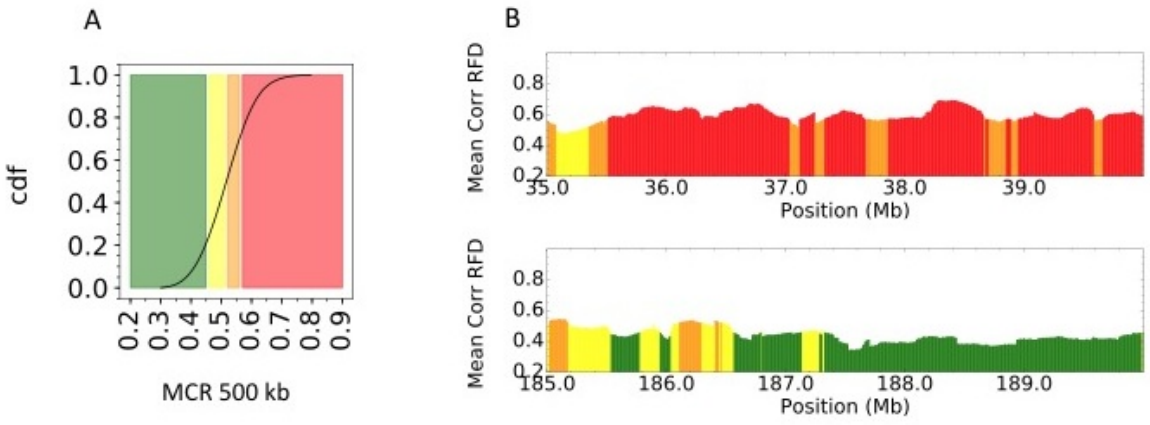

Figure 3.18: **MCR distribution.** (A) Cumulative distribution function of MCR computed in non overlapping 500 kb windows (Eq 3.4). (B) MCR profiles along the exemplary 5 Mb regions with a stable or variable replication program. (Top) Same stable regions as in Figure 3.13. (Bottom) Same variable regions as in Figure 3.14. MCR profile was computed in 500 kb windows centered every 10 kb. Color coding defined in A.

$$MCR(W) = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{C}_{1,\mathrm{i}}(\mathrm{W})). \tag{3.4}$$

At the resolution of 500 kb, MCR values are between **0.2** and **0.9** with an average equal to **0.52** as illustrated in the Figure 3.18A. We partitioned the genome in 4 classes of equal size based on the MCR values: very variable regions for MCR<0.46, moderately variable regions for MCR $\in$ [0.46, 0.52[, moderately stable regions for MCR $\in$ [0.52, 0.57[and very stable regions for MCR$\geq$0.57. To validate the MCR score, we computed the global mean difference of RFD profiles matrix among the 12 cell lines at 10 kb according to the MCR level. The difference between each mean difference RFD profiles matrix according to each MCR level and the global mean difference RFD profiles matrix among the 12 cell lines show an evolution of color from blue (low MCR) with differences larger than average to red (high MCR) with differences smaller than average (Figure 3.19). Thus, we confirmed that the regions with high MCR are the regions with less RFD differences. In contrast, regions with low MCR are regions with high RFD differences among the 12 cell lines.

To illustrate the consistence between the MCR approach and previous results obtained for 5 Mb probes (Figure 3.16), we computed the MCR profiles at 500 kb along two previously analyzed regions with a stable (Figure 3.13) and variable (Figure 3.14) replication program. As expected, the MCR score for the stable (Figure 3.18 B) regions is high and the MCR score for the variable (Figure 3.18 B) regions is low. We illustrate an overview of human chromosomes using MCR levels (Figure 3.20). When comparing the MCR profiles with the GC profiles along chromosomes [155], we observe the concordance between them, the correlation coefficient is 0.82 along the whole genome. This result suggests that the low MCR

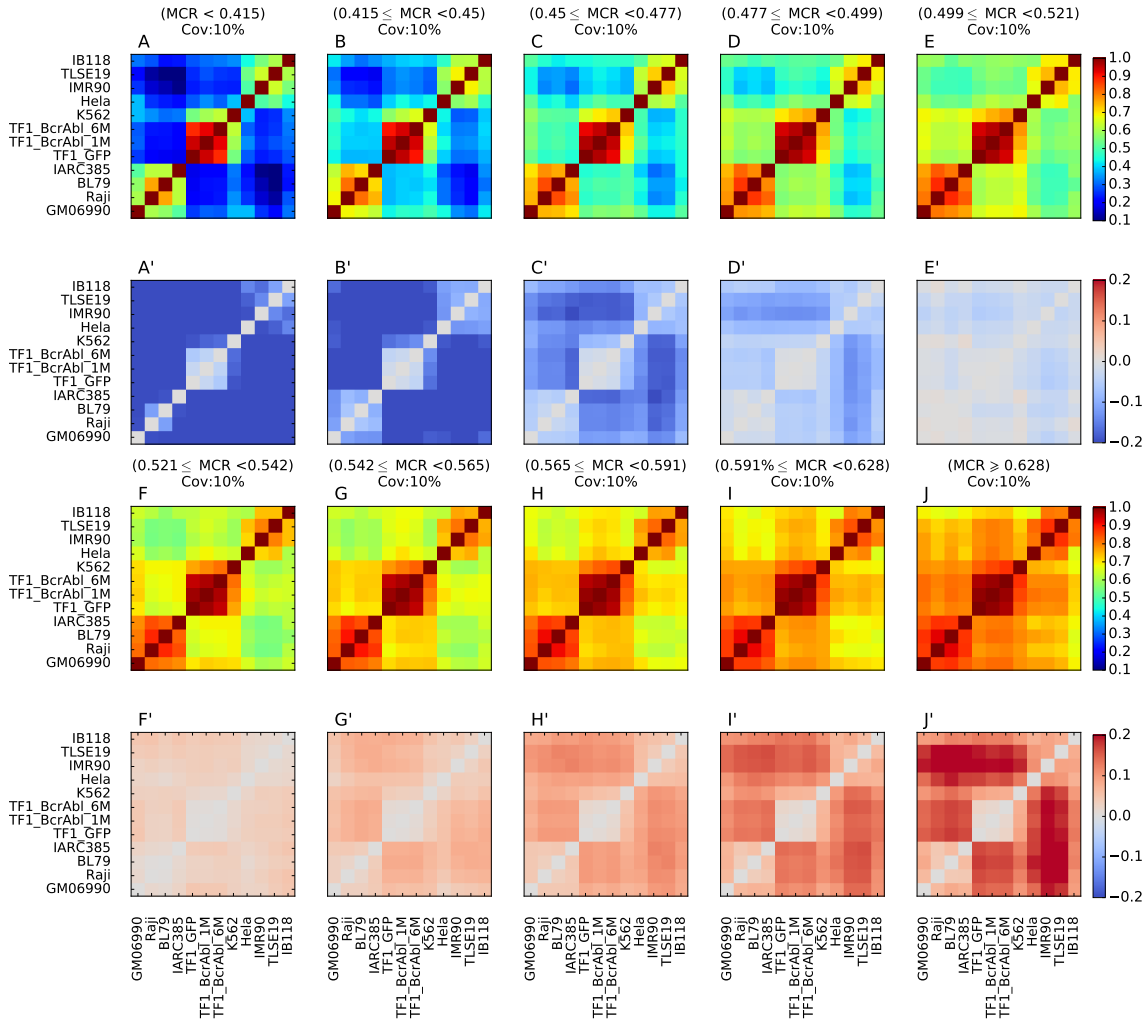

Figure 3.19: **Evolution of RFD correlation matrix with MCR score.** 10 kb windows were grouped in 10 categories of equal size based on the MCR value of the 500 kb regions centered on the windows. (A-J) Correlation matrix of RFD profiles depending on the MCR level. (A'-J') Matrices of difference of correlation matrix. Matrices row and column order is the same as in Figure 3.7A.

.

value corresponds to the low GC content regions. For example, chromosome 4 is classified as poor in GC and at the same time using the MCR score, we can see that the DNA replication program tends to be variable in this chromosome (Figure 3.20B). These results confirm the association between the low GC content and the instability of the RFD profiles among the 12 cell lines. In previous sections, we have used the GC content as a classifier of the genome. We have seen that there were more replicative changes in the low GC content regions. To test whether our MCR score ranks the regions of the genome well, we tested the level of association between the MCR and the GC-content. We computed the cdf of GC content according to the MCR level and we observed that we have a high correlation between low GC regions and low MCR regions (Figure 3.21).

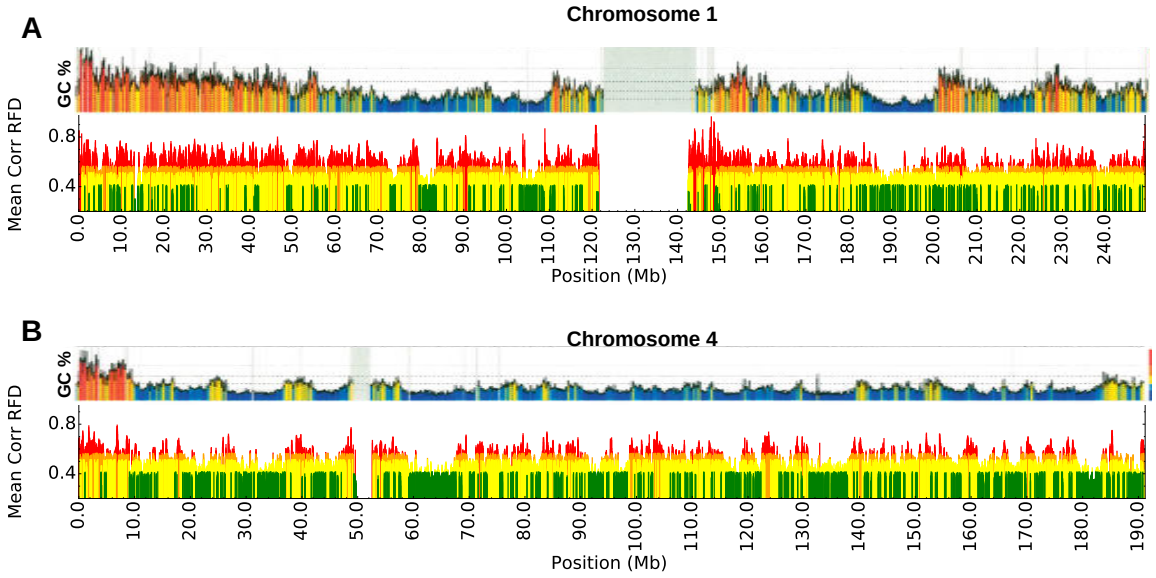

Figure 3.20: **MCR and GC are highly correlated**. We represent for the chromosome 1 (panel A) and 4 (panel B) the MCR score (bottom) and the GC content (top). We can visualize some concordance between MCR at 500 kb and GC isochore [130, 155] in chromosome 1 and 4 where we have regions rich in AT (blue) corresponding to regions of low MCR value (green).

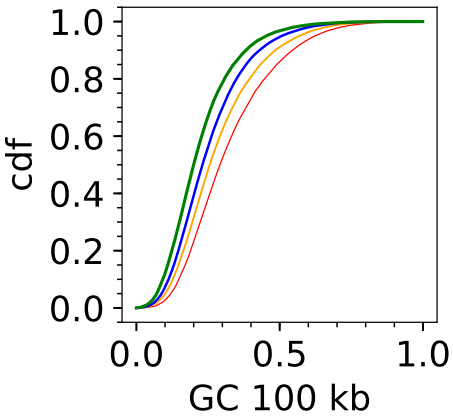

Figure 3.21: **High correlation between MCR level and GC content**.. Cdf of the GC content at 100 Kb according to the MCR level at 500 kb. From green, variable RFD profiles regions, to red highly stable RFD profiles regions.

To check if the regions with variable RFD profiles among the 12 cell lines are the same in the 3 types of cell lines, we computed the MCR score between pairs of cell lines type and per cell line type at 500 kb. For example, MCR_{ML} corresponds to MCR computed based on the 8 cell lines of myeloid and lymphoid types. MCR_M corresponds to the MCR computed

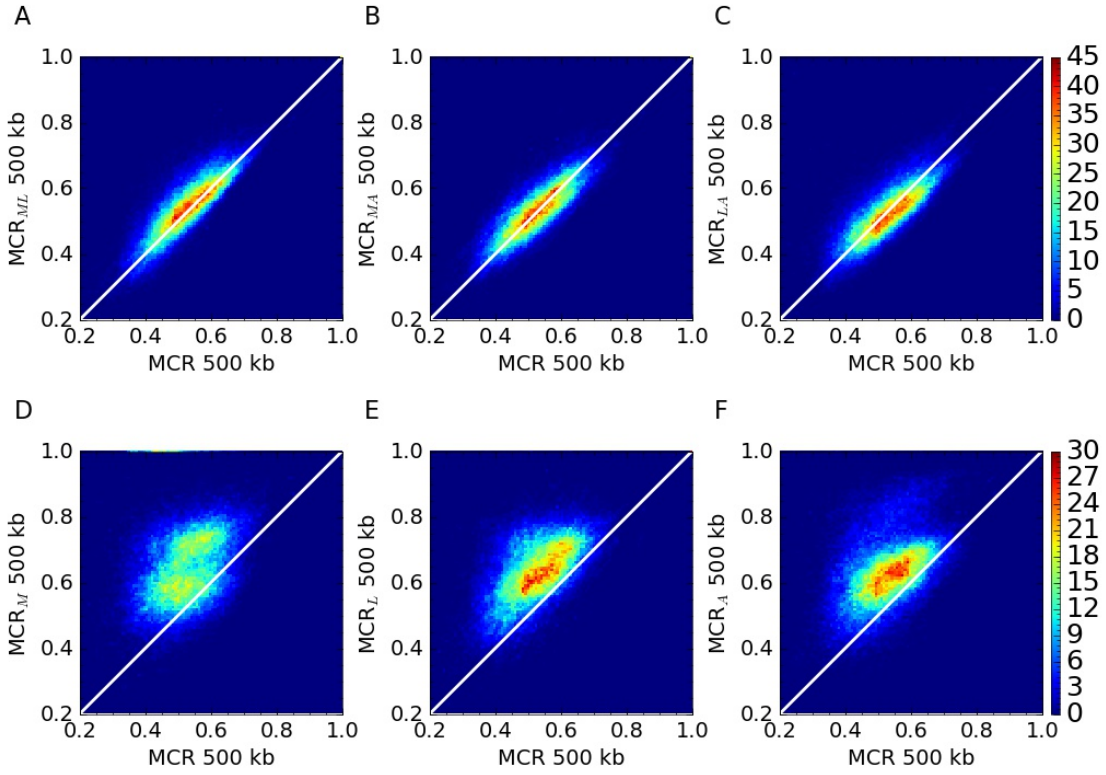

Figure 3.22: **The global MCR score are highly correlated to pairwises cell type MCR scores.**. (ABC) 2D probability distribution function (pdf) between the global MCR score and MCR score computed in pair of cell lines type MCR_{ML} (myeloid and lymphoid), MCR_{MA} (myeloid and adherent), MCR_{LA} (lymphoid and adherent). (DEF) 2D pdf between the MCR computed per cell lines type (MCR_M for myeloid cell lines, MCR_L for lymphoid cell lines, MCR_A for adhrent cell lines) and the global MCR score . Color-bar corresponding the the probability density function.

based on the 4 myeloid cell lines. Then we compared each new score to the global MCR score computed among the 12 cell lines. We observe in Figure 3.22ABC that the pairwise MCR scores for respectively MCR_{ML}, MCR_{MA}, MCR_{LA} are highly correlated with the global MCR, with respectively a correlation coefficient of 0.85, 0.84 and 0.78. This result suggests that a pairwise MCR score detects the same regions as the global MCR score. For example, we can observe in Figure 3.22A that the low (resp. high) MCR_{ML} value corresponds to low (resp. high) global MCR. Thus, the RFD profiles tend to be variable in the same regions among the 12 cell lines whatever the pair of cell groups considered. Furthermore, the comparison of the global MCR score and MCR computed per type of cell lines (MCR_M, MCR_L, MCR_A) (Figure 3.22DEF) confirmed that the variable and stable regions of RFD profiles are common among the 12 cell lines. Due to the fact that the RFD profiles of cell lines in each type are very correlated (as we shown in the global RFD correlation matrix, Figure 3.7A), the MCR_M, MCR_L and MCR_A values are above the diagonal (Figure 3.22DEF) but are highly associated to the global MCR score. For example, we observed an association between MCR_L and the global MCR score (Figure 3.22E) with a high correlation coefficient

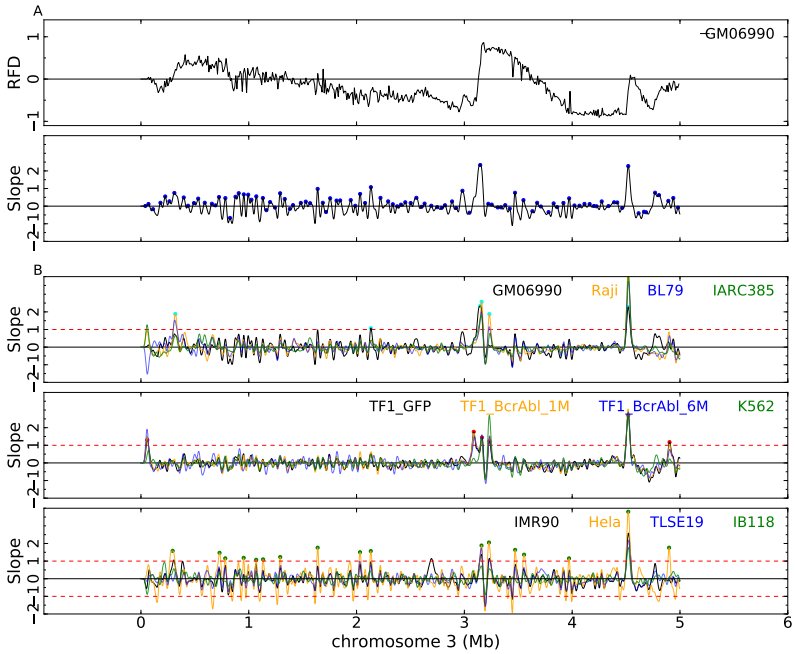

Figure 3.23: **Genome-wide profiling of slope using replication fork directionality.** (A) RFD profile (Top) and RFD slope profiles (Bottom) in GM06990 along a 5 Mb segment of chromosome 3; the blue dots correspond to the local maximum of the RFD slope. (B) RFD slope profile along the same 5 Mb segment of chromosome 3 in 4 lymphoid (top), 4 myeloid (middle) and 4 adherent (bottom) cell lines, as indicated in the top right corners of each panel. Bullets represented the local maximum of RFD slope (MS) >1.

of ∼0.6. This association was also observed between MCR_A, MCR_M and the global MCR score (Figure 3.22DF).

Finally, in this section 3.4, using the GC content as an indicator of genome organization, we showed that the RFD profiles among the 12 cell lines are more conserved in high GC content. Then we showed that along the genome we can identify many 5 Mb windows that are classified as very variable RFD profiles regions and other 5 Mb windows that are classified as very stable RFD profiles regions. To increase the resolution of detection, we computed the MCR score at 500 kb. Using MCR score we can do a very precise selection of variable regions taking into account the specificity of the region according to other features of the genome. We will adapt this score in the following analysis in order to understand the source of the heterogeneity in the DNA replication program. In the following chapter we will study the association between the DNA replication program changes and other features of the genome such as replication timing and gene expression changes.

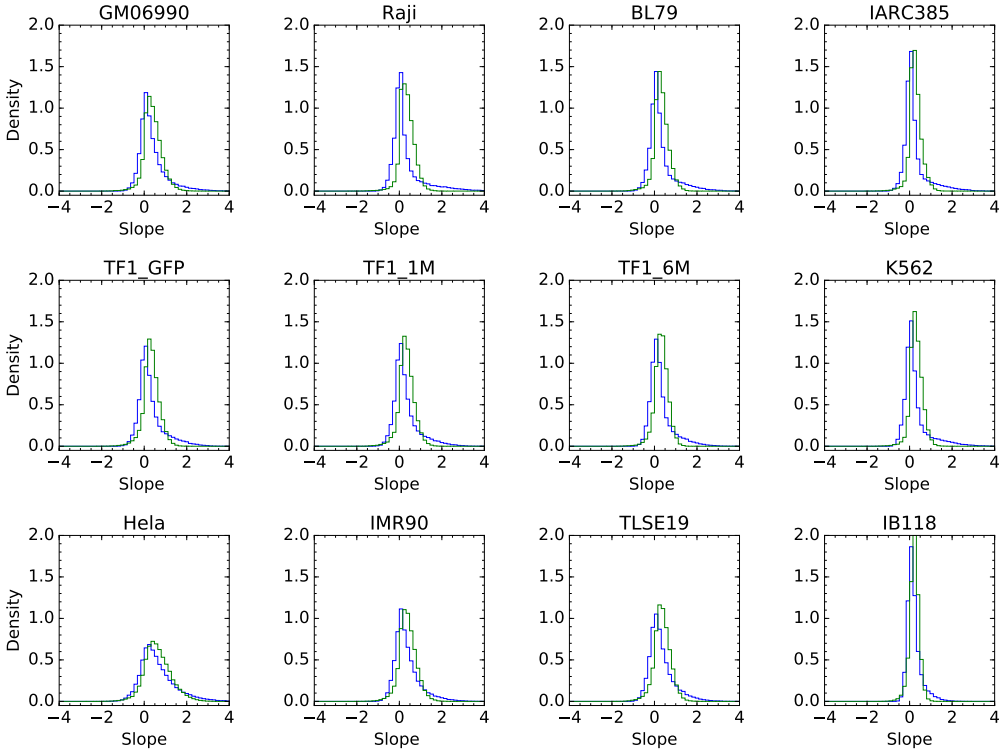

Figure 3.24: **Distribution of local extremum values of RFD slopes profiles in each cell lines.** Pdf of the slopes at the local maximum of RFD slopes profiles (blue). Pdf of the opposites of the slopes at the local minimum of RFD slopes profiles (green).

3.5 Preferential IZ and DNA replication program

3.5.1 Recovering preferential replication initiation zones

IZ are ascending segments of the RFD profiles (Figures 1.7 and 3.4, [3]) and can thus be detected as regions of maximal values of the RFD profiles slopes. A concern about experimental data is the presence of noise. Strictly speaking, the derivative of a noisy profile is not defined and, correspondingly, the naive derivative of a profile based on the numerical difference between two successive samples is ill-defined and numerically unstable. However, the derivative of a smoothed version of a noisy profile is well defined [54]. In other words, the rate of signal variation has to be estimated over a sufficiently large number of data points. Moving average of RFD profiles computed in 10 kb windows shifted by 1 kb were computed using a smooth (differentiable) window ϕ with a mid-height width of 26 kb and RFD slope was defined as the increment over 20 kb of that profile:

$$S(x) = \frac{d}{dx} RFD * \phi(x), \tag{3.5}$$

where x is the position on the genome and $*$ stand fot the convolution operator. All slopes were finally expressed in %RFD per kb. See Figure 3.23A for a RFD slope profile computed over a 5 Mb region in GM06990. We finally recorded the position and slope of the local

Cell line	MS>0.5	MS>1	MS>2
GM06990	13058	5696	1406
Raji	9333	5617	2286
BL79	9471	4989	1225
IAR385	7369	4168	1040
TF1_GFP	10348	5134	1052
TF1_BcrAbl_1M	10458	4972	1019
TF1_BcrAbl_6M	9478	4418	798
K562	8625	4726	1367
HeLa	21883	11641	3099
IMR90	13181	5438	1008
TLSE19	12211	5068	777
IB118	6414	1677	62

Table 3.2: **The number of detected origin in each cell lines.** First column represent the cell lines. second one represent the count af all detected origin per cell lines. The third (resp. forth and fifth) represent the count of origin according to the maximum slope 'MS'>0.5%RFD per kb (resp. >1%RFD per kb and > 2%RFD per kb).

maximum and local minimum of the RFD slope profiles.

As can be seen in Figure 3.23B, the RFD slopes profiles are strongly correlated among the 12 cell lines. In this Figure we see that the high RFD slopes in some regions are conserved among the 12 cell lines (as at position 4.6 MB). The RFD slope profile corresponding to the HeLa cell line is the noisiest due to the fact that the RFD profile is also noisy (Figure 3.23B). IB118 has a very low RFD amplitude that varies between ± 0.6 [153], and that is why the local maximum RFD slope profile is smaller than in the other cell lines. We selected 3 thresholds to detect the IZ in each cell lines, 0.5, 1 and 2. To choose the threshold values, we compute the distribution of the local maximum of RFD slopes (MaxSlope, corresponds to IZ) and of the distributions of the opposite of the slopes at the local minimum of RFD slopes (MinSlope, corresponds to Termination Zones (TZ)) (Figure 3.24). The RFD slopes profiles typically vary between -2 and 2 as shown in the Figure 3.24. We observe that the distributions of the MaxSlope are more asymmetric than the distributions of the MinSlope, and can reach larger values (up to 4). The two pdfs coincide at 1 in all cell lines except HeLa and IB118 due to the issues discussed above, indicating that all MaxSlope>1 likely correspond to IZ and not to noise. Thus the threshold of 1 will be mostly used in the following analysis to select the IZ location. Note that, to select IZ with high confidence we choose the threshold 2, as we see in the Figure 3.24 at this threshold, there is almost no local slope minima of the same amplitude. 0.5 was also used has a permissive threshold.

3.5.2 IZ conservation and cell lines classification

As defined in the previous section, we choose three thresholds to count in each cell line the number of IZ. For the first threshold 0.5, we detected 6414 origins in IB118 and 21883 (Table 3.2) in HeLa, and in average 9000 for the other cell lines. But, as shown in the histogram,

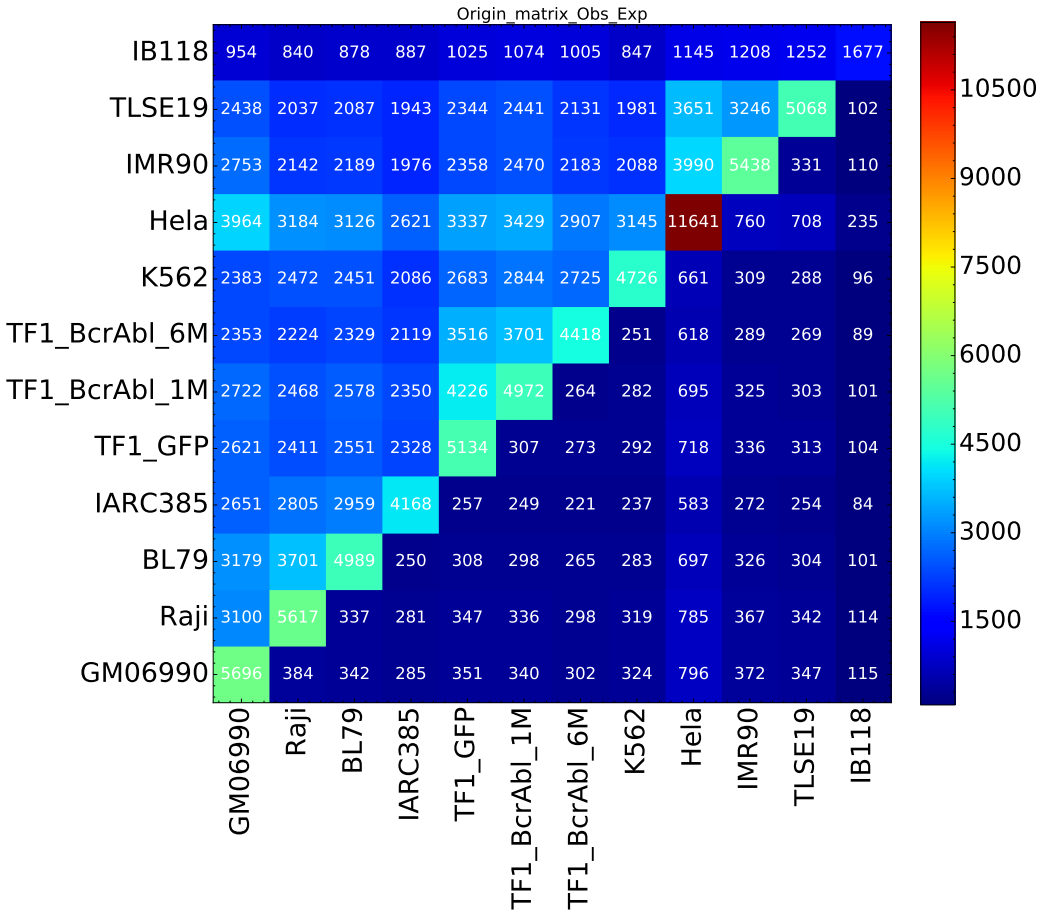

Figure 3.25: **Number of common IZ with slope>1%RFD per kb at 30 kb resolution for all pairs of cell lines.** The upper triangular part of the matrix represents the observed count of common origin with slope>1%RFD per kb between pair of cell lines in 30 kb windows. The lower triangular part of the matrix represents the expected count of common IZ computed >1%RFD per kb in each cell lines. Main diagonal is the number of detected IZ of slope>1%RFD per kb in each cell lines.

at this threshold, we would also select local slope minimum (Figure 3.24). Hence we choose the threshold 1 as we discussed previously. Using this threshold the number of detected IZ decreases in the 12 cell lines. Moreover, we obtained for GM06990 (5696) and HeLa (11641) very similar numbers of origins as previously detected using the Hidden Markov Model (HMM) method [3] (GM06990: 5684 detected origins, HeLa: 9836 detected origins). The overlapping between the two methods to select the IZ goes to 90% in GM06990 and to 84% in HeLa. So our IZ detection with this threshold is in good agreement with previous IZ detection and will be most often used in the following analyses. To be more selective and select the strongest origins for each cell lines, we can increase the threshold to 2. Now the number of detected origin decreases to 67 in IB118 and 3099 in HeLa cell line (Table 3.2).

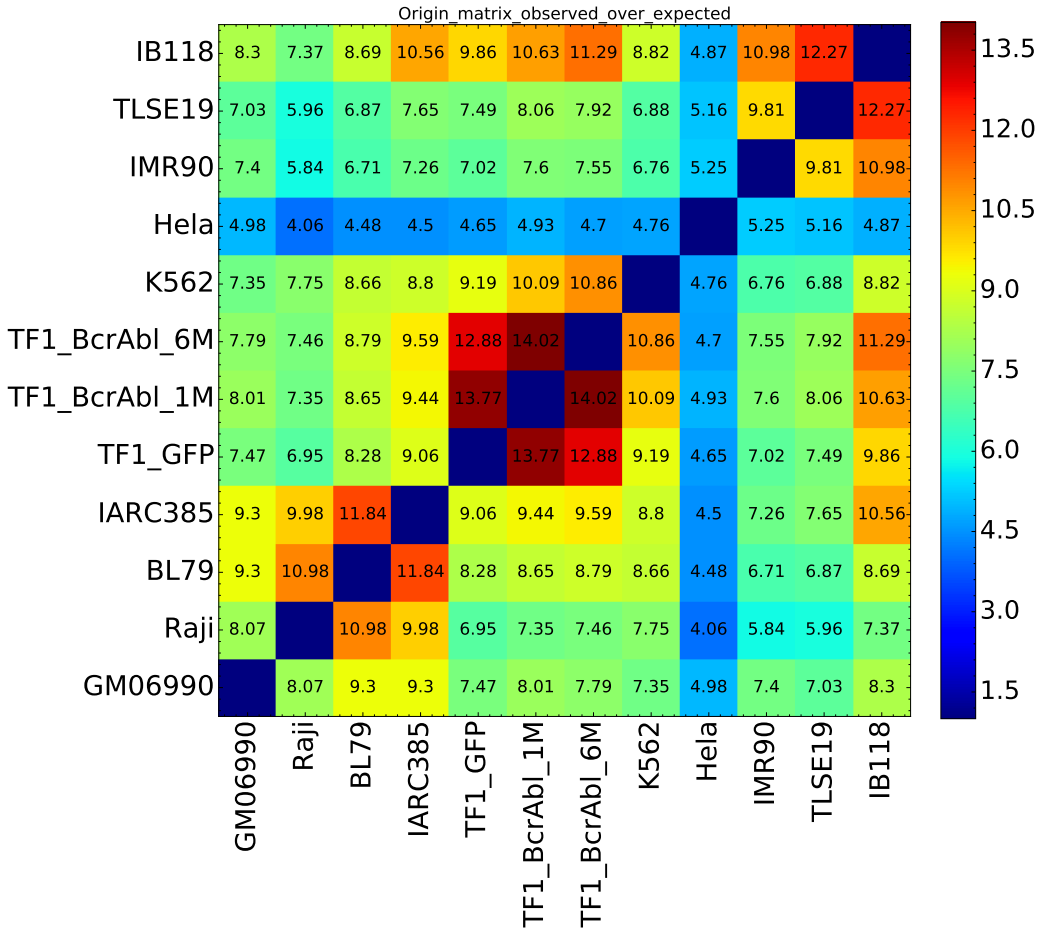

Figure 3.26: **Matrix of observed/expected numbers of common for each cell line pair for a threshold of 1%RFD per kb.** Each square of this matrix represents the normalized count of common origin between pair of cell lines. Normalization was done by dividing the observed count of common IZ (upper part of the matrix in Figure A.7) by the expected one (lower part of the matrix in Figure A.7).

IZ have a mean size of 30 kb (range 6-150 kb, [3]), so to know the number of overlapping IZ among the 12 cell lines, we computed the number of common origins per pair of cell lines at 30 kb windows using the threshold 1 and 2. Note that in this interpretation, we will exclude the results obtained in HeLa and IB118. At the threshold 1, we selected between 1943 common IZ (between TLSE19 and IARC385) and 4226 (between TF1_GFP and TF1_BcrAbl_1M) (Figure A.7). The low number of common IZ between TLSE19 and IARC385 corresponds the low RFD profiles correlation coefficient between them as shown in the Figure 3.7A. The high number of common IZ detected between TF1_GFP and TF1_BcrAbl_1M is associated with the high RFD correlation coefficient between them. In addition, we see that the number of common IZ between the TF1 cell line and the K562 also increases as expected from CML progression (CML progression system will

be specifically analyzed in Chapter 5). To validate the detected numbers we computed the expected number of common IZ between each pair of cell line at 30 kb windows as the mean for random distribution of IZ using Binomial law (Figure A.7). About 300 common IZ are expected between each two cell lines (Figure A.7) so is smaller than the observed counts. In Figure 3.26, we represented the ratio between the observed and expected common IZ between all pairs of cell lines computed by dividing the upper triangular part of the matrix (Figure A.7, observed) by the lower triangular part of the matrix (Figure A.7, expected). We systematically observed in average about >6 times more common IZ than expected. Except for HeLa and for IB118 cell line, the global matrix structure showed in the Figure 3.26 is very similar to the structure of the RFD correlation matrix showed in the previous section of this chapter (Figure 3.8 A). Thus, in each type of cell lines, the cell lines share with high concordance the same IZ. Moreover, we can see the relation between the K562 and the CML progression as in the global RFD correlation matrix (Figure 3.26). For the lymphoid cell lines, we observe a highly conserved IZs between the two BL cell lines (BL79 and Raji), furthermore IARC385 shares more IZ with BL cell lines (BL79 and Raji) than GM06990. Similar results were obtained when increasing the threshold to 2 (Figure A.8). We observed between ∼16 and ∼68 times more common IZ than expected, thus the common IZ with a higher slope would be more conserved among the 12 cell lines. These results give us a new argument to confirm that the RFD profiles are robust classifiers of cell lines based on their tissue of origin.

Finally, we increased the resolution to 10 kb to test if at high resolution there were still IZ conservations among the 12 cell lines. We repeated the same analysis as done at 30 kb windows at the threshold 1. We selected between 182 common IZ (between IMR90 and K562) and 809 common IZ (between TF1_GFP and TF1_BcrAbl_1M) (Appendix, Figure A.9). Also we observed on average about ∼3 times more than expected common IZ among the 12 cell lines (Appendix, Figure A.10). This results confirmed that even at a high resolution the IZ are very conserved among the 12 cell lines.

3.5.3 Replication stable regions are dense in efficient Initiation Zones

Finally in this chapter we study the association between the GC, the MCR score and the IZ spatial distribution. We compute the density of the IZ in 5 categories of GC content (Figure 3.27A) from poor to rich and 5 categories of MCR (Figure 3.27C) from variable to stable replication program. The density of IZ increases from GC-poor to GC-rich regions and from variable to stable RFD profiles regions. The IZ density in GC-rich regions is about ∼ 3.5 times more than in poor-GC regions (Figure 3.27A). The same result is obtained with the MCR score, the stable RFD profiles regions are denser in IZ more than the variable RFD profiles regions (Figure 3.27C). The cdf of GC content associated to the strongest IZ (Figure 3.27B, green line) are associated to high GC content values. Same results were obtained using the MCR profiles (Figure 3.27D), regions with strong amplitude of initiation zones are more stable than regions with low amplitude. This suggests that low MCR and poor GC content

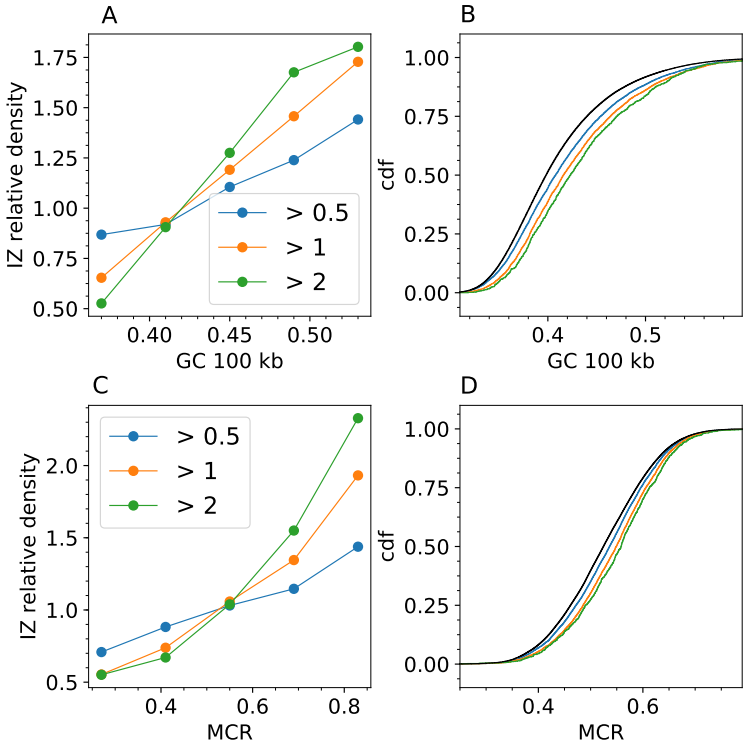

Figure 3.27: **High density of replication origins in high GC content replication stable regions.** (A) The IZ density detected using the local maximum of RFD slope (MS) profiles in 5 categories of GC at 100 kb from poor to rich ([0,0.37], [0.37,0.42], [0.42,0.47], [0.47,0.53],[0.53,1]). Blue dots all MS>0.5%RFD per kb (S1), orange dots MS>1%RFD per kb (S2) and green dots the MS>2%RFD per kb (S3). (B) Cdf of GC at 100 kb in each group of slope s(Blue line for S1, Orange line for S2, Green line for S3, Black line in the whole genome). (C) Same figure as in A but for MCR at 500 kb from variable to conserved RFD profiles. (D) Cdf of MCR at 500 kb in each group of RFD slope local maximum.

regions are corresponding to a low density of IZ with a low efficiency and that replication origins with the highest efficiency tend to localize within replication stable, high GC content regions of this IZ density.

3.6 Conclusion

As mentioned in the introduction, the replication timing (RT) has been used for several years as a good classifier of cell line. The shape of the RT profile helps to identify large domains of constant timing [37, 45, 156]. In 2008 Hiratani et al. showed that DNA replication domains were conserved between independent mESC lines [37, 157]. Ryba et al. further demonstrated the power of replication profiling to distinguish closely related cell types, and argued that temporal domains of replication are spatially compartmentalized in 2 structural and functional units of three-dimensional chromosomal architecture [38]. Therefore, classification of cell lines could be established based on this information. But the resolution of the RT profiles remained low and the domains of timing rather large (100 kb - 1 Mb). In this thesis, we

computed the RFD profiles based on the Ok-seq data, and increased the resolution to 10 kb. We have generated novel RFD profiles and have explored several approaches to compare the replication programs of twelve cancer and non-cancer cell types. To identify the changes of RFD profiles among the 12 cell lines we computed the correlation matrix at 10 kb. A global, unbiased correlation approach revealed that the RFD profiles clustered in three separate groups corresponding to lymphoid, myeloid and adherent cells. Interestingly, we observed that cancer-associated changes in replication did not blur their developmental origin signature. We did not detect any convergence of the replication programs of cancer cells from different developmental origins (e.g. LMS vs. BLs). In contrast, within lymphoid cells, we found evidence for BL-specific replication patterns. Overall, our global correlation analyses provide evidence for recurrent replication changes along specific tumor progression pathways.

GC content (isochores) classifies the genome into 5 categories (isochores: L1, L2, H1, H2, H3). Based on this partition, we computed the RFD correlation matrix of each category. We were able to reproduce a similar global classification of cell lines. Based on this, the global correlation analyses further revealed that RFD changes are widespread through the genome but more frequent in GC-poor regions. The variations of RFD profiles were distributed randomly across the genome and our global classification was conserved when we choose randomly 5 Mb probes. However, heterogeneity of correlation matrix of RFD profiles was observed in 5 Mb windows. Using a Z-score constructed on RFD correlations matrices, we isolated windows where the correlation values within and between the 3 types of cell lines is very strong suggesting that, in these regions the RFD profiles are the same among the 12 cell lines and that they thus form stable RFD regions. We also isolated windows were replication program both within and between groups appeared variable. The characterization of selected variable RFD profiles regions confirmed that those regions are associated to low GC content, low gene expression, and late replication.

The latter results encouraged us to do a global cartography of the genome to identify the zones where the RFD profiles do not change among the cell lines. We computed a new score called MCR to map the conservation level of the DNA replication program at 500 kb resolution. MCR allowed to identify the regions that are variable/stable in terms of RFD profiles among the 12 cell lines. Finally, we used the derivative of the RFD profiles to delineate IZ locations as regions of maximal RFD slope. We observe that the stable replication program regions are associated to high MCR value and correspond to the regions enriched in the most efficient IZ that are conserved among the 12 cell lines. Inversely, the variable DNA replication program regions are associated to low MCR values and correspond to region of low initiation zones density. In the following chapter we will study the association between replication program and other features of the genome as transcription, and we will use the MCR score and the IZ identification to study more specifically the association between replicative and transcription changes.

CHAPTER 4

Association between changes in replication fork polarity profiles, Mean Replication Timing and transcription

DNA serves as a template for both replication and transcription. In order to avoid collisions between the transcription machinery and the replication forks, these two processes usually take place separately in time and space [158]. The different copies of genes do not have the same regulatory sequences, so transcription can take place at some copies while others are replicated. This is the case of the genes encoding histone H4 [159]. Nevertheless, despite these regulatory systems, multiple studies showed that conflicts between transcription and replication leads to DNA damage and double-stranded breaks [160, 161].

Nevertheless cell cycle average transcription profiles revealed that genes tend to be replicated early during S phase [38]. In this chapter, we will study the global association between the replication fork directionality, the replication timing and the transcription program. We will start by comparing the previous cell line classification based on the RFD profiles showed in chapter 3 to the cell line classifications obtained when using MRT and transcription data. After that we will see how the conserved regions of RFD profiles detected among the 12 cell lines are associated to large domains of MRT. Finally, we will quantify the link between transcriptional changes and replication modifications.

4.1 Introduction

Molecularly the first correlation between the replication timing and the transcription was established in 1980 by analyzing a dozen of cell-type specific genes. Concomitantly, the first half of the S phase was shown to corresponds to the Giemsa[1]-light R bands [162] associated to GC-rich and highly transcribed regions, the second half corresponding to the Giemsa-dark G bands associated to GC-poor and low transcriptional activity regions [163–165]. In the 2000s, the emergence of microarray technologies allowing genome wide DNA replication timing analyses in many organisms as drosophila [100, 101] and human [47, 50, 166] confirmed these correlation. We confirmed these observations at the chromosomal scale. We observed a strong correlation of the chromosomes average MRT with their GC content, their average FPKM and DNase[2] sensitivity (Figure 4.1A). Note that, the correlation coefficient between MRT and DNase sensitivity (r=-0.605) was stronger than the correlation between MRT and FPKM (r=-0.56) (Figure 4.1B).

Generally, active and open euchromatin regions are replicated early in S phase, and repressed and closed heterochromatin regions are replicated late in S phase [41]. In the mammalian genome the replication timing is also associated with the subnuclear position (Figure 1.5). Using logistic regression Woodfine et al. (2003) [50] found a significant positive correlation between the gene expression probability and the timing of replication (r = 0.61). Interestingly, the correlation between MRT and transcription was not observed in budding yeast [59], suggesting that this relationship was acquired at some point during evolution [163, 168] (see review [169]). In Drosophila, it was demonstrated that most genes have a tendency to replicate in the first third of S phase with a high probability of being expressed independently of their replication time during this period. A strong relationship between the MRT and transcription was found only in 25% of the genes that replicate later in S phase [100]. The analysis of the differentiation of mouse embroyonic stem cells (mESCs) revealed that promoters of high and low CpG density, which generally have strong and weak promoter activity, respectively, showed distinct behaviors when switching to a late replication environment, where only CpG poor promoters showed a stronger tendency towards negative transcriptional regulation [37]. In cancer, replication timing changes were reported also to be associated with a concerted change in gene expression [170]. It was found that genes that replicate earlier are significantly increased in expression, whereas the genes that replicate later are significantly repressed [170]. Using Ok-seq method to compare the replication and transcription program, it was found that DNA replication initiates preferentially at the transcription start sites[3] (TSS) [3, 171]. To conclude, in human genome a global correlation between the replication and transcription program was observed but some genes did not apparently these general rules.

[1] G-banding is a technique used in cytogenetics to produce a visible karyotype by staining condensed chromosomes.

[2] DNaseI is a specific endonuclease that hydrolyses double-stranded (ds) DNA generating tri- and/or tetra-oligonucleotides having 5'-phosphate and 3'-hydroxyl termini [167].

[3] TSS is the first nucleotide of DNA that is transcribed into RNA.

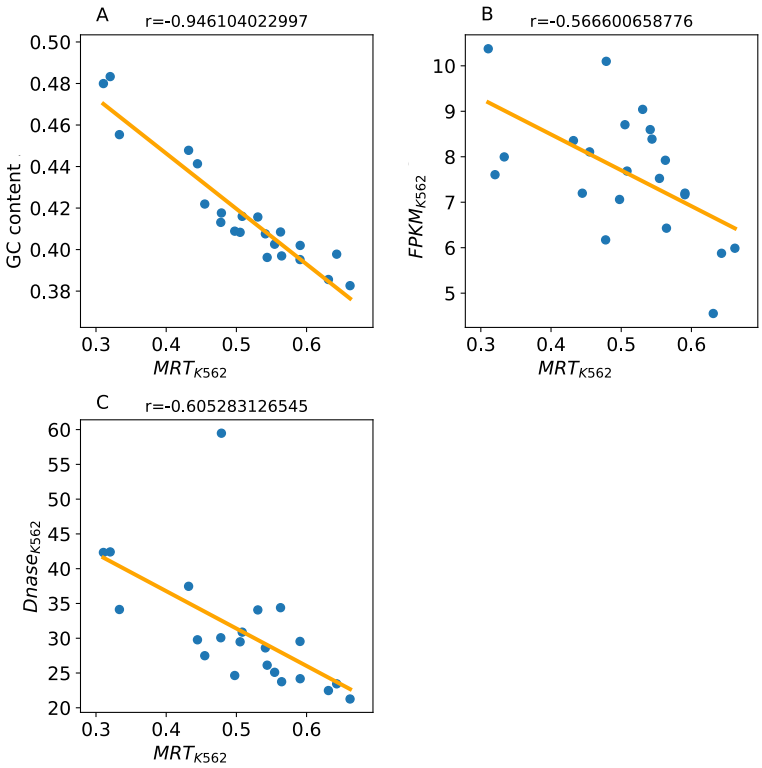

Figure 4.1: **Association of Mean Replication Timing with GC content, transcription and Dnase sensitivity in K562.** Mean of GC content (A) FPKM (B), and DNase sensitivity (C) of each chromosome as a function of the mean of MRT in each chromosome. The orange line represent the linear regression. The r value in the top of each panel represents the correlation coefficient of Pearson.

4.2 Classification of human cell lines based on the MRT and FPKM

4.2.1 Quantification of MRT profiles for 7 cell lines and RNA-seq profiles for 12 cell lines

The MRT profiles for 7 cell lines were computed as defined in the Materials and Methods (section 2.2.3). In Figure 4.2, we represent the MRT for each cell line along a 20 Mb segment of chromosome 3. We observe that the MRT of the two lymphoid cells (GM06990 and GM12878) are very similar. The same results were observed for the three myeloid and the two adherent cell lines. Between groups similarities are clearly observed.

In Figure 4.3, we represent the FPKM computed in 200 kb windows (Eq. 2.2) for each of the 12 cell lines as detailed in the Materials and Methods (section 2.2.4). We can see that FPKM profiles computed at 200 kb is similar among the 12 cell lines. But we can observe that some specific regions are particularly expressed in a specific cell line, as we see in the first 5 Mb of the chromosome 3 in TLSE19 cell (blue line). To compare the change in gene expression level among the 12 cell lines, we compute the correlation coefficients of these profiles per pairs

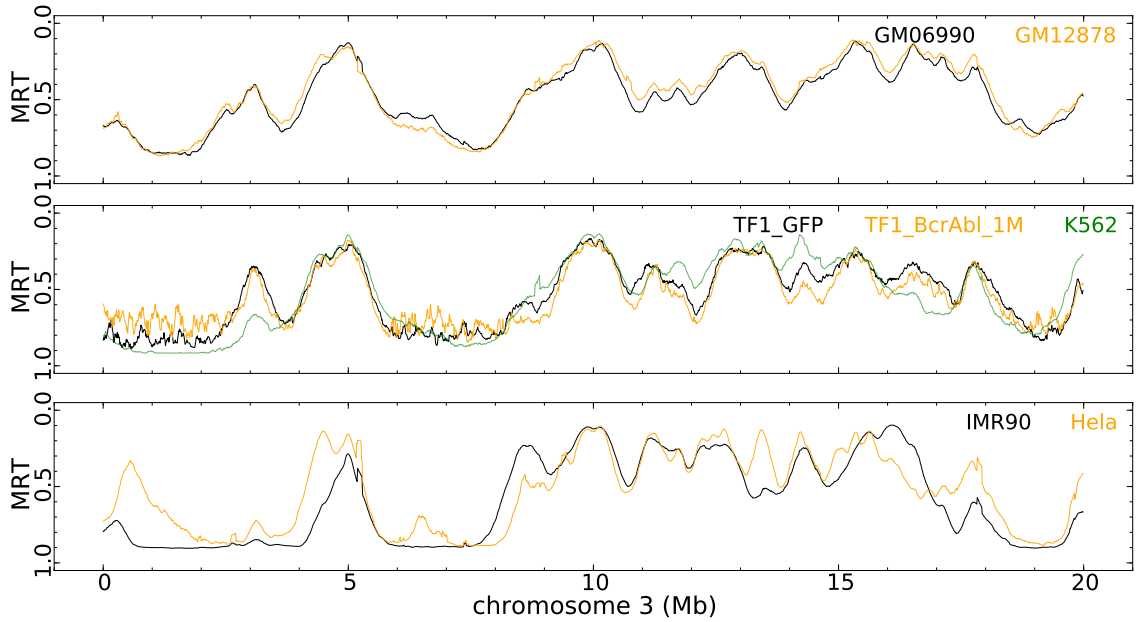

Figure 4.2: **Genome-wide profiling of replication timing.** Mean Replication Timing (MRT) along a 20 Mb segment of chromosome 3 in 2 lymphoid (top), 3 myeloid (middle) and 2 adherent (bottom) cell lines, as indicated in the top right corners of each panel.

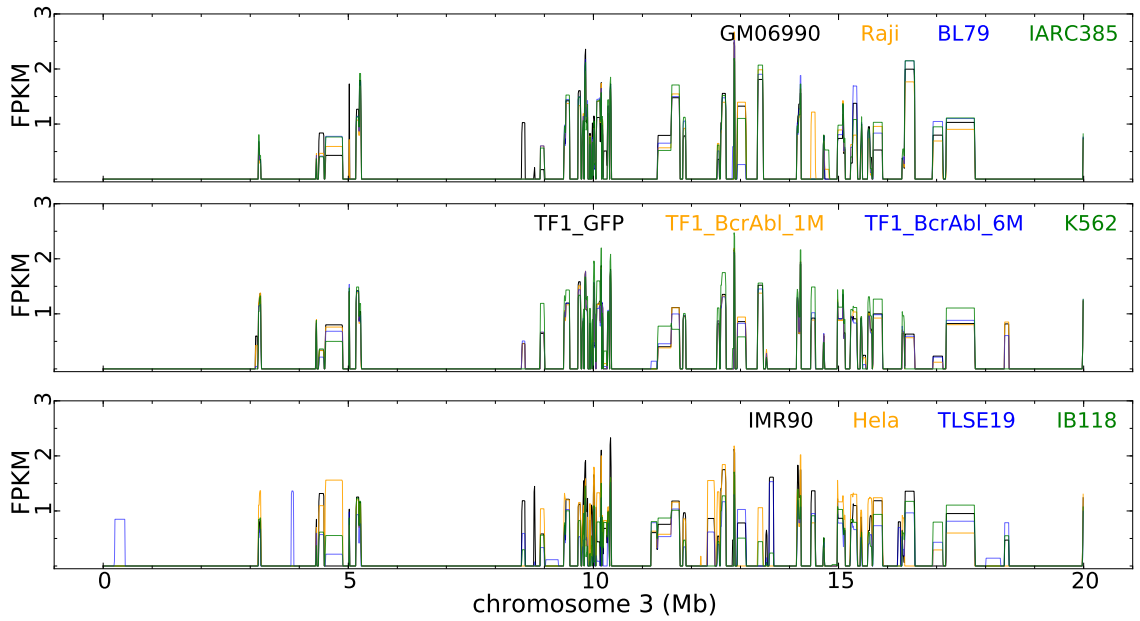

Figure 4.3: **Genome-wide profiling of gene expression.** FPKM at 200 kb (Eq. 2.2) along a 20 Mb segment of chromosome 3 in 4 lymphoid (top), 4 myeloid (middle) and 4 adherent (bottom) cell lines, as indicated in the top right corners of each panel.

of cell lines.

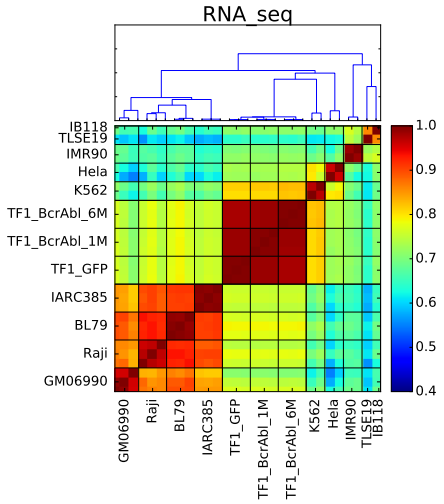

Figure 4.4: **RNA-seq biological replicate classification based on correlations between gene expression computed at 200 kb (Figure 4.3).** (Bottom) Correlation matrices between RNA-seq profiles. Pearson correlation coefficient values are color-coded from blue (0.4) to red (1) using the color bar on the right. (Top) Dendrogram representation of the hierarchical classification of RNA-seq experiments based on the corresponding correlation matrix; ordinate is the correlation distance.

4.2.2 RFD cell line classification is conserved using FPKM and MRT profiles

As described in the chapter 3, to objectively quantify differences in DNA replication program among cell lines, we computed the pairwise correlation coefficients between RFD profiles and ordered them by correlation distance using unsupervised hierarchical clustering (Figure 2.2A). In contrast to RFD profiles, RNA-seq and MRT profiles were generated by different laboratories using different methods (See sections 2.2.3 and 2.2.4). We similarly analyzed transcription (RNA-seq; Figure 4.5) and mean replication time (MRT; Figure 4.6A) data. All the observed differences between cell lines were larger than the variations between biological replicates (Pearson correlations 0.954-0.995 for RNA-seq; Figure 4.4; >0.99 for MRT). A similar classification as obtained with Ok-seq (Figure 3.7A) was observed by RNA-seq except that HeLa clustered with myeloid instead of adherent cells. We cannot exclude that HeLa clustered with myeloid cells by RNA-seq because HeLa, IMR90 and K562 RNA-seq profiles are from Encode whereas other myeloid and adherent cell profiles are generated differently. Despite the higher homogeneity of RFD methods, the correlation coefficients were generally smaller by RFD than by RNA-seq. The situation was more heterogeneous by RNA-seq where within-group correlation distances increased from lymphoid to myeloid to adherent cells and the three groups could not be recovered by cutting the dendrogram at a constant level. Within the myeloid cell group a similar, albeit weaker, progression from early to late CML was observed by RNA-seq as observed by Ok-seq (Figure 3.7A) (More details in chapter 5) (Figure 4.5).

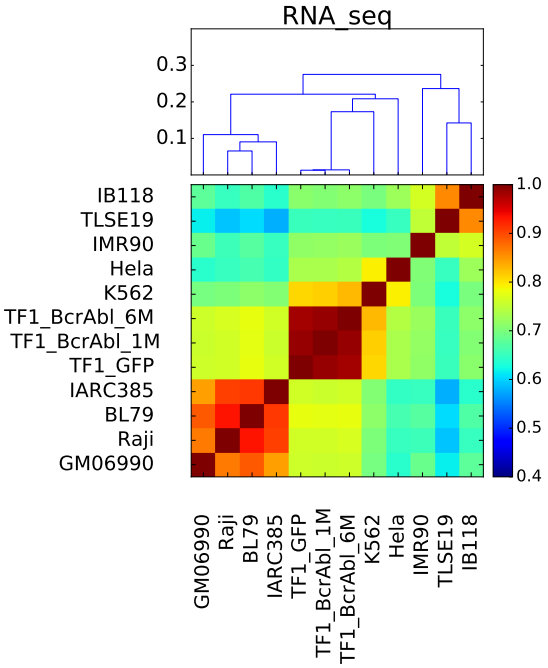

Figure 4.5: **Cell line classification based on correlations between gene expression profiles.** (Bottom) Correlation matrices between RNA-seq. C_{FPKM} Pearson correlation coefficient values are colour-coded from blue (0.4) to red (1) using the color bar on the right (Materials and Methods). (Top) Corresponding dendrogram representation of the hierarchical classification of cell lines; ordinate is the correlation distance (Eq. 2.7)

Within the lymphoid cell group, a similar classification of cell lines was obtained by RNA-seq (Figures 4.5). The two BLs (Raji, BL79) were more correlated to each other than to either LCLs (GM06990; IARC385), suggesting the existence of BL-specific replication and transcription patterns. BL79 was more correlated to IARC385 than to GM06990. This was expected since IARC385 was established from the same patient as BL79, as confirmed by HLA typing. Intriguingly, Raji and IARC385 were more correlated to each other than to GM06990. This suggested that IARC385 may share some replication and transcription signatures of BLs, and may not be as "normal" as GM06990. Indeed, the two LCLs IARC385 and GM06990 were the least correlated cell lines within this group. These results suggest either that IARC385 was established from abnormal, but non-BL blood cells from the same patient as BL79, which seems unlikely, or that IARC385 was destabilized during immortalization, which is more plausible. Both scenarios may explain the observed genomic instability associated with replication and transcription changes reminiscent of BLs. EBV transformation occurred *in vivo* for the BLs but *in vitro* for the LCLs.

Within the adherent cells, different classifications were obtained by RFD and RNA-seq (Figures 3.7, 4.5). By RNA-seq, the two LMS IB118 and TLSE19 were more correlated to each other than to IMR90 and less correlated to HeLa, which in fact clustered with myeloid cells. The cells of origin and driver mutations of LMSs are currently unclear. These results

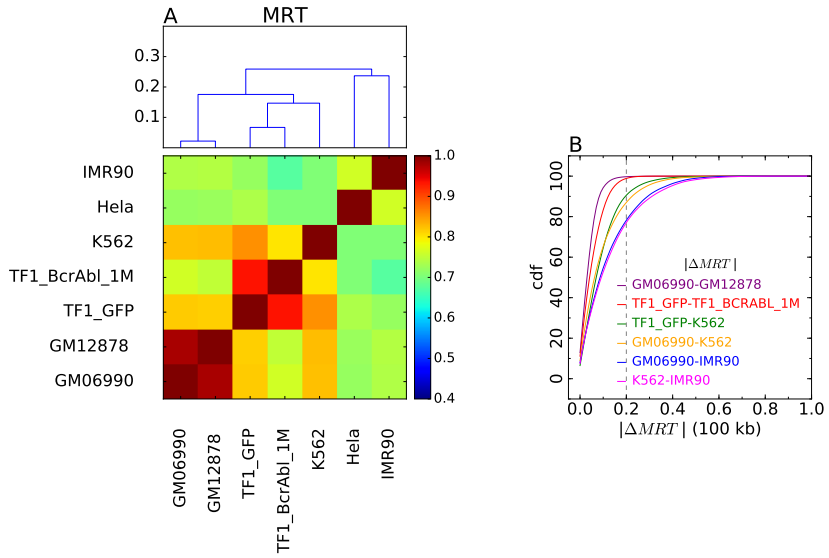

Figure 4.6: **Cell line classification based on the MRT profiles.** (A) Correlation matrices between MRT profiles ($(C_{MRT}$; A); Pearson correlation coefficient values are color-coded from blue (0.4) to red (1) using the color bar on the right (Materials and Methods). (Top) A corresponding dendrogram representation of the hierarchical classification of cell lines is shown on top of each correlation matrix; ordinate is the correlation distance. (B) Cumulative distributions of the absolute MRT changes ($|\Delta MRT|$) between cell lines. Each curve is color-coded according to the pair of cell lines indicated in the insert. The considered threshold of significance [111] ($|\Delta MRT| > 0.2$) is indicated by a vertical dotted line.

may help to distinguish different types of LMS and suggested a possible differentiation of TLSE19 and IB118, which were derived from a buttock muscle tumor and a scalp tumor, respectively. The strong correlation of the RNA-seq profiles of the two LMSs may be due to the use of different RNA-seq methodologies for LMSs and for other cell types. Alternatively, the shared expression of cancer-specific genes, in two LMSs of different cellular origins, may have resulted in strong convergence of RNA-seq but not RFD profiles. It is notable that the RNA-seq and RFD data better matched each other for IB118 than for TLSE19. This suggests that replication is more dissociated from transcription in TLSE19 than in IB118.

Lymphoid, myeloid and adherent cells formed three separate MRT clusters as in RFD (Figure 4.6A). A previous analysis reported that MRT profiles of nonleukemic cells are much less variant than those from leukemias [154]. In agreement, we observed a strong correlation of the MRT profiles from GM06990 and GM12878, two LCLs with a near-normal karyotype (Figure 4.6A). MRT profiles did not reflect the CML progression, since K562 appeared more correlated to TF1-GFP than to TF1-BCRABL-1M (Figure 4.6A). This cannot be explained by variations in MRT methodology since both TF1 derivatives were profiled by the same method. In summary, RFD profiles suggested the existence of CML-specific replication changes, not necessarily reflected in MRT profiles. To further investigate this, we compared the cumulative distributions of MRT changes between TF1-GFP, TF1-BCRABL-1M and K562 to those observed between distinct LCLs (GM06990 vs. GM12878) and distinct cell

types (K562, GM06990, IMR90) (Figure 4.6B). The proportion of MRT changes with $|\Delta MRT| > 0.2$ after one month of BCR-ABL1 expression in TF1 cells (1.13%) was slightly higher than between LCLs (0.3%), but much lower than between distinct cell types (K562 vs. IMR90: 22.9%; K562 vs. GM06990: 12.8%; IMR90 vs. GM06990: 21.8%). In contrast, the proportion of $|\Delta MRT| > 0.2$ changes between TF1-GFP and K562 (9.45%) was more similar to that previously observed between LCLs and leukemia. We concluded that RFD changes induced in our model for CML establishment and early progression are not accompanied by large-scale shifts in MRT, but that such shifts may appear during progression to late CML (K562).

In summary, globally similar results were obtained by RNA-seq, but divergences between RNA-seq and RFD classifications were also observed. These results suggested that recurrent replication changes occur in specific tumor types but that the closeness of their connection with transcription may depend on cell type.

4.2.3 Cell line classification based on RFD, FPKM and MRT profiles per chromosome

To investigate whether RNA-seq and MRT differences between cell lines were concentrated in particular regions of the genome, we repeated the analyses shown in Figures 4.5 and 4.6 for each chromosome separately (Figure B.1). The results obtained for the entire genome were recapitulated for each chromosome, with minor exceptions detailed below. We recap that lymphoid, myeloid and adherent cells formed three separate RFD clusters, except that for chromosome 19, HeLa clustered with myeloid rather than adherent cells, and for chromosome 16, HeLa was equally distant from both groups. Distances within and between groups were conserved, except that for chromosome 22, myeloid cells were closer to adherent than to lymphoid cells, and for chromosome 17, IARC385, rather than GM06990, was the most distant among lymphoid cells. As to RNA-seq data, the clustering of HeLa with myeloid rather than adherent cells was again observed for each chromosome, although for chromosomes 10 and 11 HeLa appeared equidistant from myeloid and adherent cells. In addition, IMR90 clustered with HeLa and myeloid cells rather than adherent cells for chromosomes 6 and 20, and was equidistant from both groups for chromosomes 13, 14, 18 and 21. As to MRT data, the global classification was reproduced for each chromosome except that the branching point of HeLa and K562 was somewhat unstable. The correlation coefficients were again smaller by RFD than by RNA-seq or MRT for each chromosome, but this was more pronounced for GC-poor chromosomes.

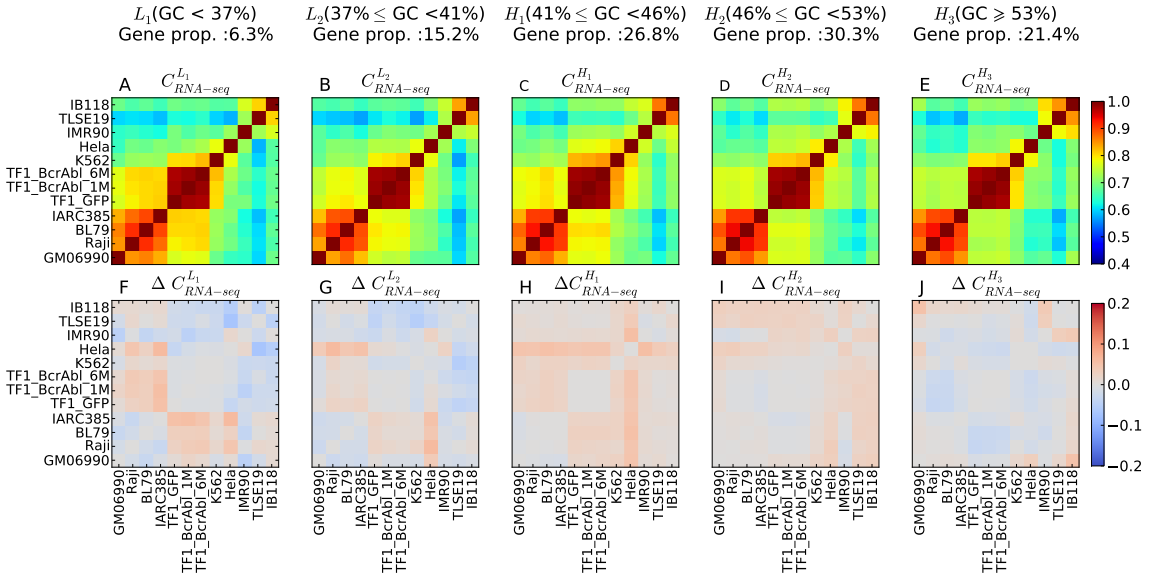

Figure 4.7: **Transcription program conservation for regions with different GC-content**. (A-E) Correlation matrix of RNA-seq profiles depending on the GC content; 10 kb windows were grouped in GC-content categories following the 5 isochores classification of the human genome in light isochores L1 ($GC \leq 37$; $C^{L_1}_{RNA-seq}$; A) and L2 ($37 \leq GC < 41$; $C^{L_2}_{RNA-seq}$; B), and heavy isochores H1 ($41 \leq GC < 46$; $C^{H_1}_{RNA-seq}$; C), H2 ($46 \leq GC < 53$; $C^{H_2}_{RNA-seq}$; D) and H3 ($GC \geq 37$; $C^{H_3}_{RNA-seq}$; E). (F-J) Matrices of correlation differences where can be L1, L2, H1, H2 or H3; correlation difference values are color coded blue (resp. red) for negative (resp. positive) differences using the color bar on the right; a blue (resp. red) color indicates that RFD profiles are less (resp. more) conserved in the considered isochore than in the 22 autosomes. Proportion of genes (Gene prop.) within each isochore family is provided in the header of each column

4.2.4 Transcription changes are not concentrated in GC poor regions

A similar correlation analysis of GC-content with RNA-seq data (Figure 4.7) was performed as previously done for the RFD data(Figure 3.11). GC content-dependent changes in correlation coefficients (Figure 4.7A-E) were less marked than for RFD (Figure 3.11). The global classification of cell lines is conserved in each GC group as we observed in the dendrogram represented in Figure B.3. Unlike RFD, correlation difference matrices of RNA-seq data showed no general tendency to follow GC content (Figure 4.7F-J). Caution is required, as RNA-seq profiles, unlike RFD profiles, were obtained by four different methodologies for (i) the lymphoid cells; (ii) the TF1 group; (iii) the LMSs; (iv) the ENCODE cell lines HeLa, K562 and IMR190. Inside the lymphoid group, correlations increased with GC-content. The only exception was the IARC385 vs. Raji comparison, with increasing RNA-seq correlations in the order L1<L2<H3<H1<H2, which in fact closely followed the atypical RFD correlation order L1<H3<L2<H1<H2 noted in Chapter 3 for this pair of cell lines. Thus, in the lymphoid group, the GC-dependence of RFD and RNA-seq correlations closely matched each other. This was also the case for the TF1 group and the LMS group. A more complex situation was observed for the ENCODE cell lines. The order for adherent cells HeLa vs. IMR90 was H1=L1<H2=L2<H3 by RFD but H3<L1<L2<H2<H1 by RNA-seq, thus deviating from GC-content in a different manner for RNA-seq and RFD. Within the ENCODE group, the myeloid vs. adherent comparisons re-

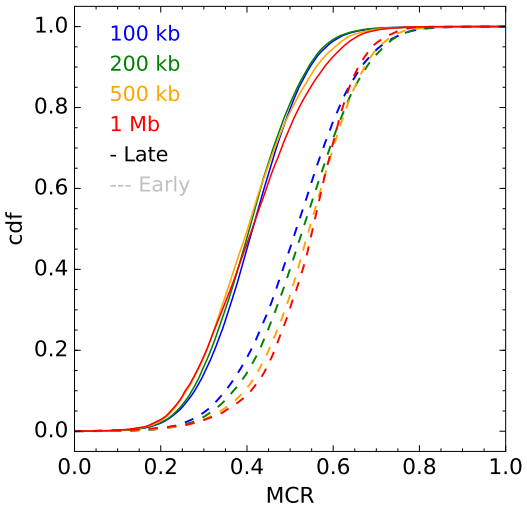

Figure 4.8: **MCR in late and early timing regions.** The lines represent the cdf of MCR in consistently late replicating regions ($\widehat{MRT}$ >0.7), the dashed lines represent the cdf of MCR in consistently early replicating regions ($\widehat{MRT}$ <0.3). Each color correspond to a scale of computation of MCR , as indicated.

vealed a more subtle deviation of replication and transcription changes. For K562 vs. HeLa, the order was L1<L2<H1<H2=H3 by RFD but L2<L1<H3<H2<H1 by RNA-seq. For K562 vs. IMR90, the order was L1<L2<H1<H2<H3 by RFD but L1<L2< H3<H2=H1 by RNA-seq. Other inter-group comparisons also showed a different ordering of RNA-seq and RFD correlations, but in these cases we cannot exclude an effect of the different RNA-seq methodologies. To summarize, contrarily to RFD changes the RNA-seq profiles did not reveal a consistent tendency for transcription changes to increase or to decrease with GC content. This suggests that replication changes were at least partly dissociated from transcription changes, to an extent that depended on the cellular context.

4.3 Association between Replication Fork Polarity conservation and Mean Replication Timing

It was suggested that 50% of the genome contains large late and early domains of replication timing. The other half of MRT profiles is made of U-shaped domains [53]. A very high association was observed between the nucleotide compositional Skew N-domains and the replication timing U domains. About half of detected U-domains are shared between different cell lines [111] and this is a hallmark of the robustness of MRT profiles. In this thesis, MRT profiles shown in Figure 4.2 help us to study the link between regions of stable and variable RFD profiles among the 12 cell lines and the constant timing regions.

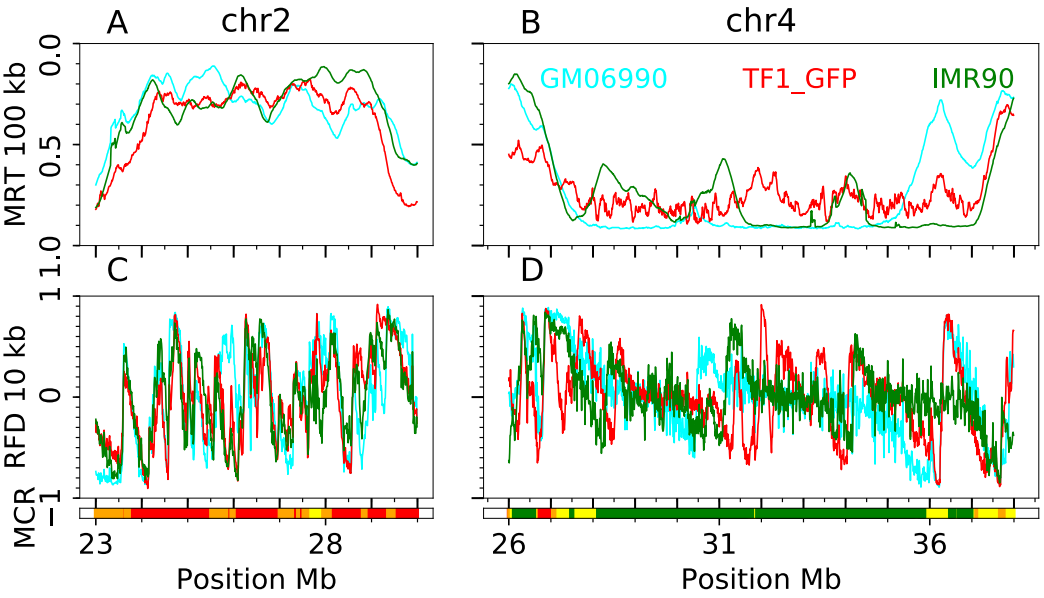

Figure 4.9: **Identification of large (>1.5 Mb) stable DNA replication timing program domains.** We represent the RFD profiles at 10 kb (C, resp. D), MCR at 500 kb and Mean Replication Timing (MRT) at 100 kb (A, resp. B) of 7 (resp. 12) Mb of early (resp. late) replicating regions of chromosome 2 (resp. chromosome 4). MCR code is defined in Figure 3.18. Other examples of large early and late replication domains are provided in Figure B.4.

4.3.1 Large domains of MCR are correlated to Constant Timing Regions (CTR)

We computed the mean MRT ($\widehat{MRT}$) profiles, which is the average of MRT profiles at 100 kb of 5 cell lines. An $\widehat{MRT}$ close to 0 corresponds to consistently early replicating regions among the 5 cell lines. In opposite an $\widehat{MRT}$ close to 1 corresponds to consistently late replicating regions among the 5 cell lines. To study the association between the MRT and the RFD profiles we associated the MCR profiles at 500 kb to the consistently late replicating regions where $\widehat{MRT}$ >0.7 and to consistently early replicating regions where $\widehat{MRT}$ <0.3. The result is illustrated in Figure 4.8, where the dashed lines correspond to consistently early replicating regions and the continuous lines correspond to consistently late replication regions. We observed that the early regions are associated to the stable RFD profiles (high MCR values). In contrast, the late conserved replicating regions are associated to the variable RFD profiles (low MCR value). Thus, in the regions where the replication timing profiles are consistently late, the RFD profiles are variable among 12 cell lines. The same result was obtained if we changed the resolution of MCR to 100 kb, 200 kb and 1 Mb. To confirm this result, we illustrate in the

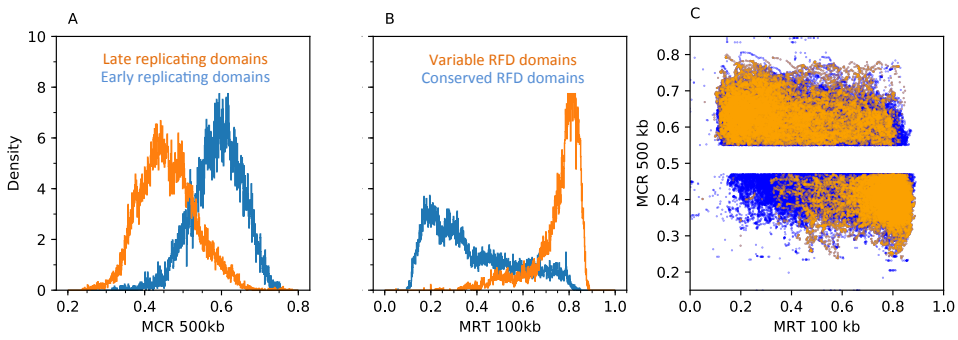

Figure 4.10: **Association between large domain (>1.5 Mb) of constant MRT and MCR.** (A) Histogram of MCR profiles at 500 kb associated with late timing regions in Orange and early timing regions in Blue. (B) Histogram of MRT profiles of TF1_GFP at 100 kb associated with variable RFD regions (Orange) and stable RFD regions (Blue). (C) Scatter plot between the MCR and MRT; orange dot correspond to the large domain that overlap in the two profiles.

Figure 4.9A (resp. B) large domains (>1.5 Mb) of consistently early (resp. late) MRT. The large early replicating domain (>1.5 Mb) is associated to a high MCR values (Figure 4.9A) and the large late replicating domains is associated to a low MCR values (Figure 4.9B). Thus we observed an association between the large replication timing domains and large domains (>1.5 Mb) of constant MCR. To investigate if this finding is a global result, we associated $\widehat{MRT}$ at 100 kb to all high MCR (MCR>0.55) and low MCR (MCR<0.47) profiles at 500 kb (Figure 4.10C, blue). Then we associated the $\widehat{MRT}$ to large domains of high and low MCR (>1.5 Mb) (Figure 4.10C, orange). We observed that the large domains of low MCR are concentrated in late replicating regions ($\widehat{MRT}$>0.6, orange dot). But the $\widehat{MRT}$ in large domains of high MCR are more dispersive, so we can have some regions which are stable in terms of RFD profiles but are replicating late. Thus, the large domains of variable RFD profiles are majoratively associated to late replicating regions. In Figure 4.10A, we selected all large domains (>1.5 Mb) of late and early replication and we associated them to MCR values at 500 kb. We observed that the late replicating domains are associated to low MCR values (orange distribution), in opposite the early replicating domains are associated to high MCR values (blue distribution). The reciprocal analysis is shown in the Figure 4.10B, we associated the $\widehat{MRT}$ to the large domains (>1.5 Mb) of variable (MCR<0.47) and stable (MCR>0.55) RFD profiles. We confirmed the very strong association between the variable RFD profiles domains and the high $\widehat{MRT}$ values (orange distribution). The stable RFD profiles domains are mostly associated to early replicating regions. However, the latter association is weaker than the former. These results suggested that globally we have a strong association between large domains (> 1.5 Mb) of late (resp. early) replication timing and large domains (> 1.5 Mb) of variable (resp. variable) RFD profiles. Finally, in Figure 4.9CD, we represented the RFD profiles at 10 kb for three cell lines (GM06990, TF1_GFP and IMR90) that represent each type of cell line group, in respectively early and late constant timing regions. The large

early CTR (> 1.5 Mb) corresponds to regions with high efficiency of conserved initiation zones between the three cell lines (Figure 4.9C). This suggests that constant regions of RFD profiles are rich in replication IZ and associated to conserved early MRT profiles. In contrast, in the large late CTR, the RFD profiles are very variable among the 3 cell lines, with only few IZ that appear to be cell line specific as the case in TF1_GFP at the position 32 Mb (Figure 4.9D). In the following, section we will compute the density of IZ associated to late and early replicating regions.

4.3.2 High Density of conserved initiation zones in early replicating regions

We use the local maximum of RFD slope profiles as computed in the previous chapter to confirm and better understand the results obtained with the MCR profiles. A large domain of consistently early replication timing is presented in Figure 4.11. In this region of low $\widehat{MRT}$ (>0.33) the MRT profiles for 5 cell lines all present an early CTR, as expected (Figure 4.11). We observed that the high MCR level is correlated to early replication but some surrounding regions are also associated to late replication. This confirmed the previous results showed in Figure 4.10AC. We further analyzed the RFD slopes profiles among the 12 cell lines in this region (Figure 4.11C). We observed that this region is associated to a high density of conserved IZ (corresponding to slope maxima >1) among the 12 cell lines. In contrast, in the region between 270 and 274 Mb, which corresponds to a late replicating domains, we do not observed any IZ among the 12 cell lines. This is also true in large domains of late replication presented in Figure 4.12. Here we chose a region where we have two consecutive large late domains. We observe a high concordance between the $\widehat{MRT}$ and the MCR scores, large domains of low MCR value are highly linked to large late replicating domains. These is compatible with the result shown in Figure 4.10BC. This two large domains are associated to a very low density of IZ among the 12 cell lines. Thus we concluded that the early replicating domains are likely linked to a high density of conserved IZ among the 12 cell lines and the late replicating domains are associated to low density of cell line specific IZ.

To further study the conservation of IZ among the cell lines, we look for some correlations between the replication IZ, the MRT and MCR profiles in a pair of lymphoid cell lines (GM06990 and Raji). We selected all the RFD slope maxima (SM) > 1%RFD per kb in GM06990. Then we estimated the RFD slope change at these loci in Raji. In Figure 4.13, we observed that the large RFD SM, which correspond to efficient IZ in GM06990 (SM>2%RFD per kb), keep a strong slope in Raji cell line. SM locations in GM06990 corresponding to a weaker slope in Raji are associated to the late replicating regions (Figure 4.13A, orange dot) and variable RFD profiles (Figure 4.13B, orange dot). Note that in this case, the slope in Raji dropped below the detection threshold for IZ (1%RFD per kb). These results suggest that the RFD profiles are conserved when the amplitude of the IZ is very strong. The middle amplitude values of IZ have more chances to be modified, (weakened or silenced IZ) when we compare them between two cell lines. In Figure B.8 using the same procedure, we present all

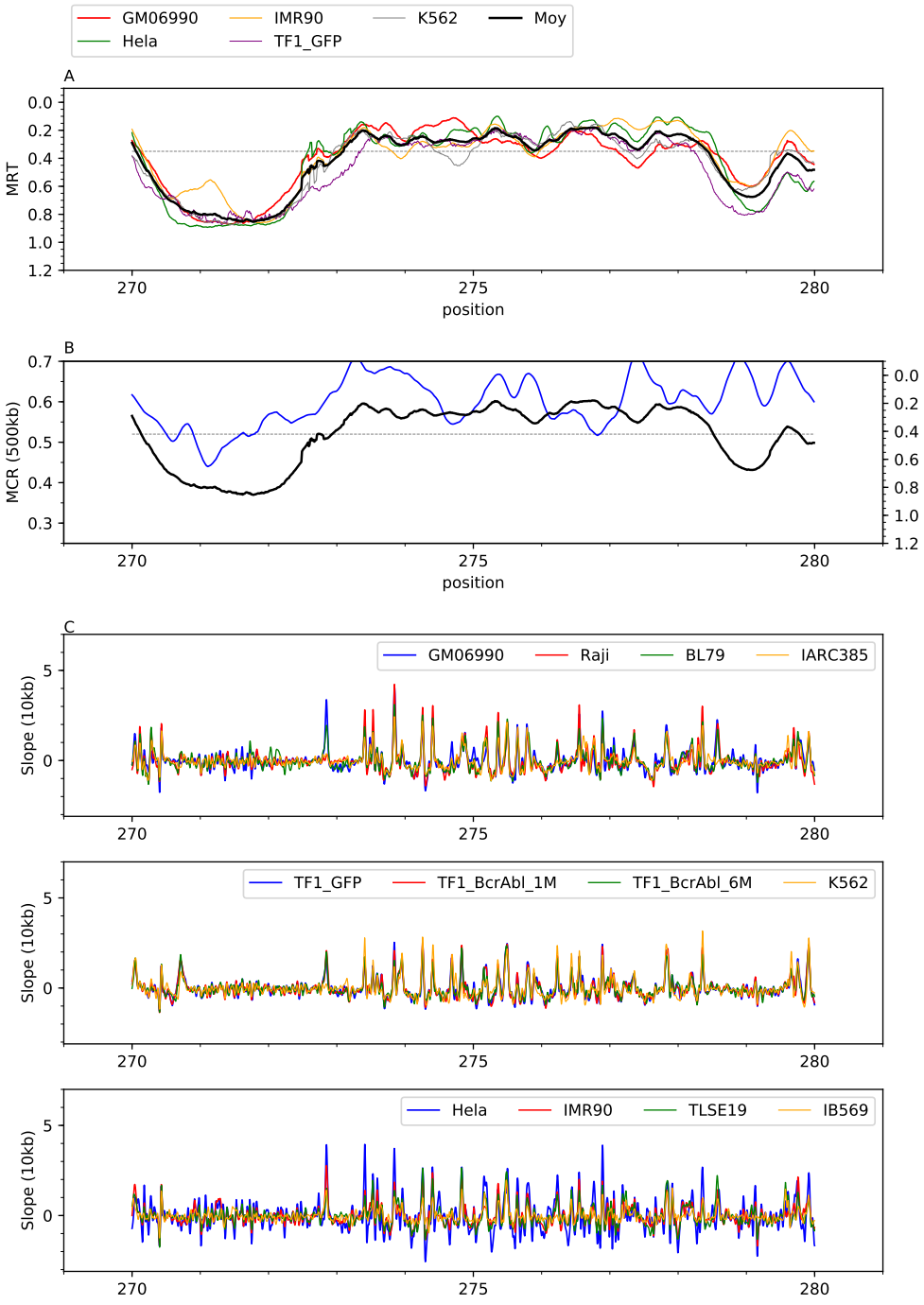

Figure 4.11: **MRT, MCR and RFD slope profiles in 10 Mb early stable replicating regions.** (A) MRT profiles of 5 cell lines at 100 kb in 10 Mb early replicating regions. (B) MCR profiles at 500 kb in stable RFD profiles. (C) RFD slope profiles for the 12 cell lines.

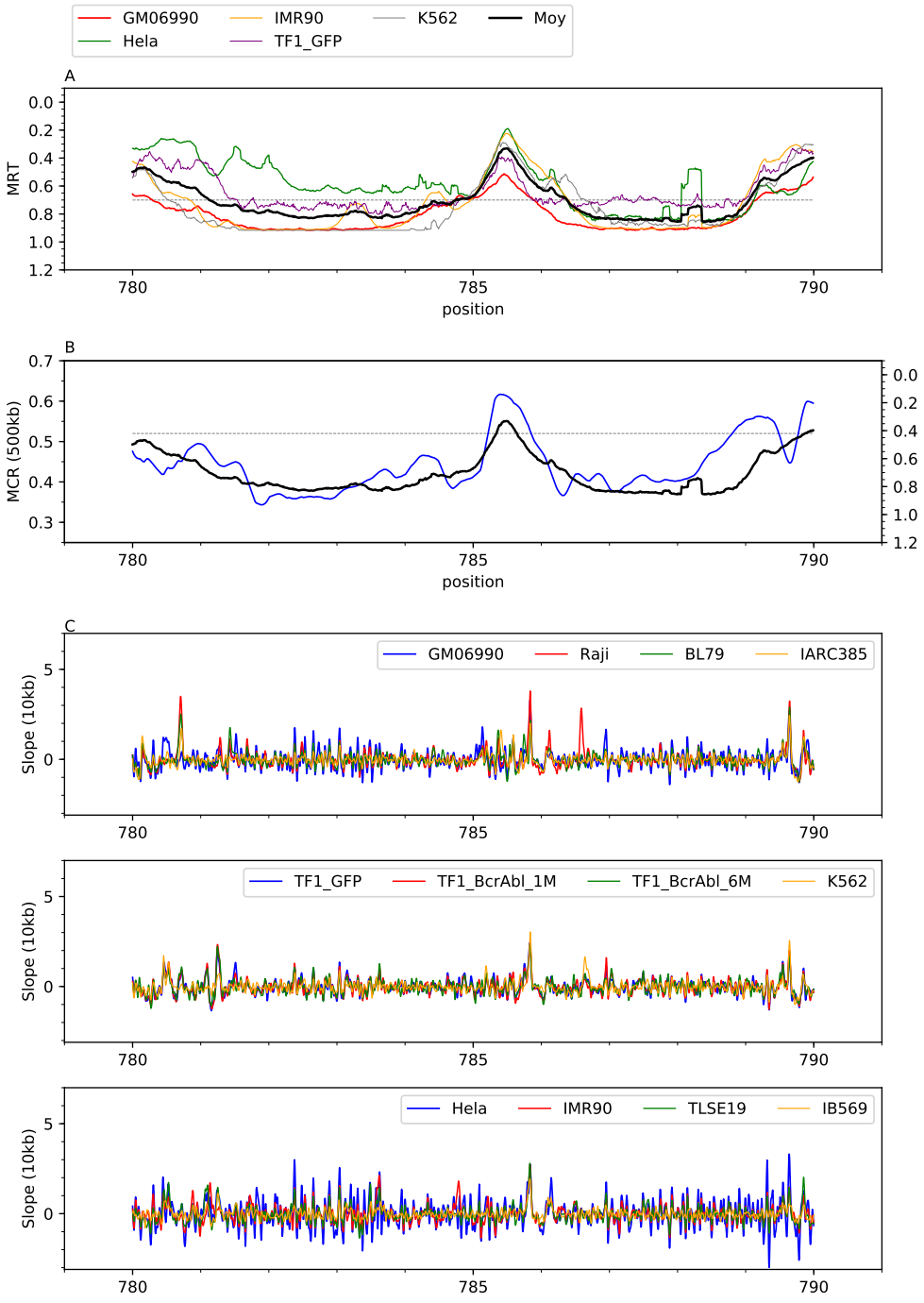

Figure 4.12: **MRT, MCR and RFD slope profiles in 10 Mb late variable replicating regions.** (A) MRT profiles of 5 cell lines at 100 kb in 10 Mb late replicating regions. (B) MCR profiles at 500 kb in variable RFD profiles. (C) Slope profiles for the 12 cell lines.

the pairwise comparison of the RFD slope distribution associated to late and early replicating regions between GM06990 and all other cell lines. We observed in the first histogram (GM06990-G06990) that the late replicating regions (red line) are linked to weaker IZ. The comparison between GM06990 and Raji reveals that the global distribution of RFD slopes in Raji associated to strong SM in GM06990 represent two peaks corresponding respectively to late replicating regions with weakened slopes and early replicating regions with conserved slopes. This result is a confirmation of the results showed in Figure 4.13. In the appendix of the chapter 4, we present all the histogram comparison between SM and slopes per pair of cell line using the same procedure (Appendix Figure B.8-B.19). Similar results are observed for all pairs.

As a confirmation, we computed the relative density of the SM in 5 categories of MRT in TF1_GFP (Figure 4.14A) as done previously for GC and MCR (Figure 3.27C). The density decreases from early to late replicating regions. The IZ density in early replication regions is about ~ 6 times larger than in late replication regions at the threshold $MS > 1\%$RFD per kb. This was also observed for the two other thresholds. The cdf of MRT_{TF1_GFP} shows the difference in the distribution of MRT_{TF1_GFP} profiles depending on the SM selection threshold and the distribution in whole genome (Figure 4.14B). The strong IZs are associated to early replication regions. This confirms that low MCR and late replicating regions are corresponding to the low amplitude of initiation zones (IZ).

To conclude we described two categories of regions with a conserved replication program. The first one corresponds to the regions that replicate early, dense in conserved IZ have a stable RFD profiles among the 12 cell lines and covers about $\sim 8\%$ of the genome. The second one corresponds to regions that replicate late with low density of cell line specific IZ and variable RFD profiles and that covers about $\sim 8\%$ of the genome. Hence, it is important to study why the replication IZ efficiency is changing among cell lines, and to correlate this change to other features of the genome such as gene expression changes.

4.4 Quantification of the links between the replication program modifications and transcriptional changes

In all recent genome-wide studies, where the replication origins were identified, a link between transcription and replication origins position was detected. It appears that current studies revealed some origins of replication lying within R-loops[4]. These tri-catenary structures correspond to the hybridization of an RNA on one of the two strands of DNA [172, 173])). These results raised the hypothesis that the origins of replication activated at the beginning of S phase could be localized nearby of R-loops. These results echo an article published in

[4]R-loop is a three-stranded nucleic acid structure, composed of a DNA, RNA hybrid and the non-template single-stranded DNA

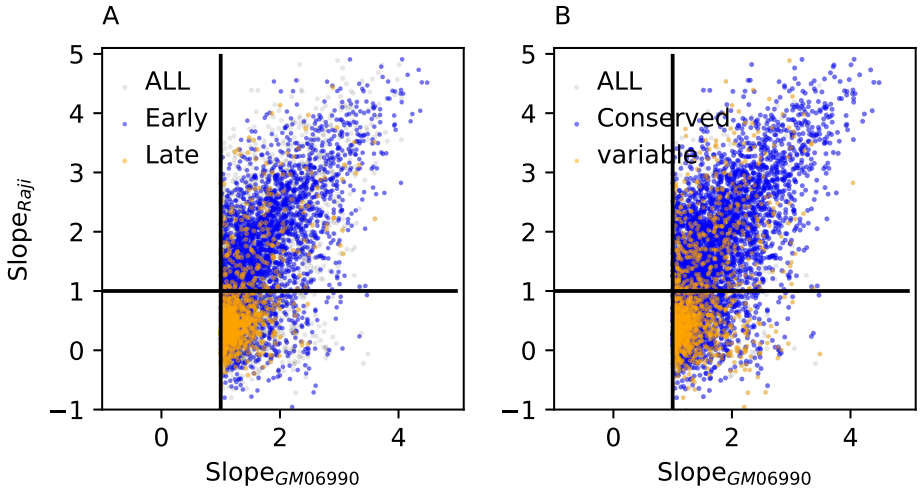

Figure 4.13: **Only strong IZ in GM06990 are conserved in Raji.** (A) (resp. B) Scatter plot of the RFD slopes in Raji associated to the local maximum of RFD Slope (SM) in GM06990 >1. Blue dots represent the early replicating regions according to the MRT of GM06990 at 100 kb (resp. stable replicating regions according to MCR at 500 kb). Orange dots represent the late replicating regions according to the MRT of GM06990 at 100 kb (resp. variable replicating regions according to MCR at 500 kb). Grey dots represent the rest of the genome.

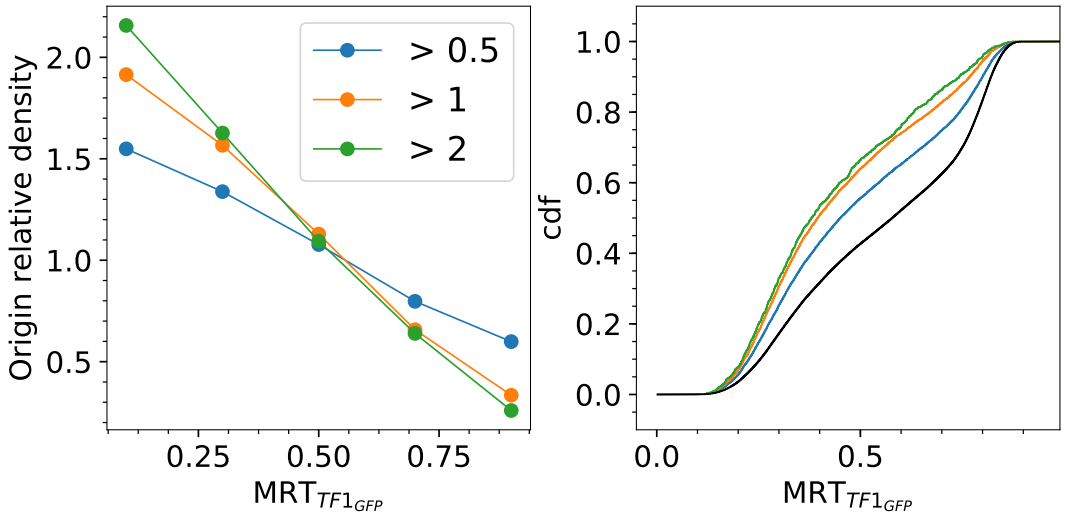

Figure 4.14: **High density of origin in early replication program regions.** (left) Replicative density of IZ detected using the slope of RFD profiles in 5 categories of MRT$_{TF1_GFP}$ at 100 kb from early to late ([0, 0.2], [0.2, 0.4], [0.4, 0.6], [0.6, 0.8], [0.8, 1]). Blue dots for all slopes>0.5%RFD per kb (S1), orange dots slopes>1%RFD per kb (S2) and green dots for slopes>2%RFD per kb (S3). (right) Cdf of MRT$_{TF1_GFP}$ in each group of slope (Blue line for S1, Orange line for S2, Green line for S3, Black line for MRT$_{TF1_GFP}$ at 100 kb in the whole genome).

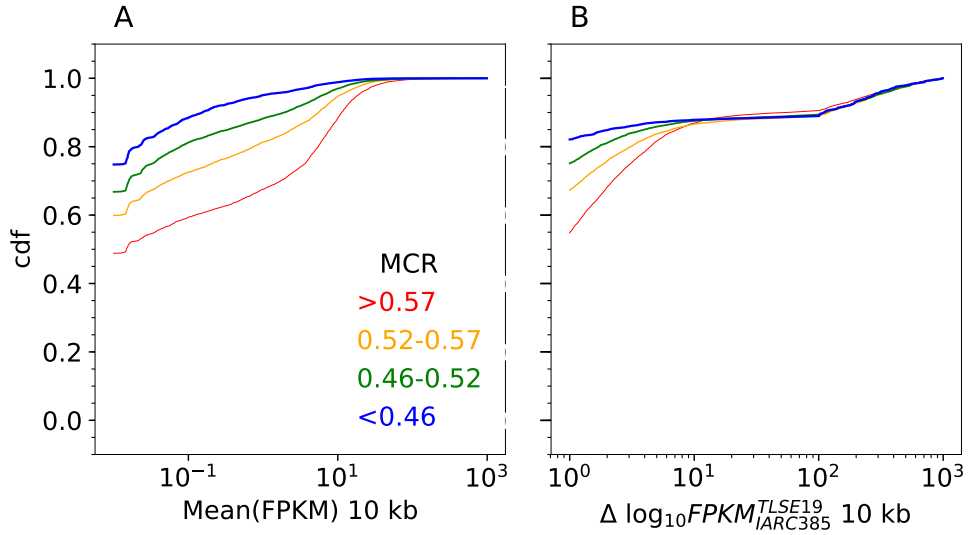

Figure 4.15: **Transcription associated to replication change.** (A) We represent the cumulative distribution function of the average FPKM value on the 12 cell lines according to their MCR level. (B) We represent the cdf of the relative changes between the pair of cell line with the lowest correlation in the RFD profiles (which is IARC385 and TLSE19, Figure 3.7A).

2015 that raised the question of the presence of replication origins at the level of R-loops [174]. Moreover, the primary results obtained in two studies conducted in human cells [77] and on mouse [175] showed that gene rich regions are also rich in replication origins and that a large part of the replication origins are at the level of transcription promoters. In this case, replication initiation sites could coincide with Transcription Starting Sites (TSS). Cayrou et al. [79], showed that the strong representation of TSS in mouse genome is in fact a consequence of their association with CpG islands. In Drosophila, among the activated origins located within GC rich regions, the TSS were concomitant with gene transcription, transcriptional regulation by the mythylation of lysine 36 of histone H3 (H3K36me1) (reviewed in [176]), and RNA polymerase II binding [177].

Here, we used the MCR scores to delineate regions that are variable in terms of DNA replication program among the 12 cell lines. In Figure 4.15A, we represent the cumulative distribution functions of the average of FPKM values on the 12 cell lines according to their MCR levels. We found that regions with low MCR are associated to regions which are almost not expressed. About half of the regions with high MCR level (>0.57) are associated to expressed regions (Figure 4.15A). Moreover, the expression differences between IARC385 and TLSE19, where we observed the most RFD changes, tend to confirm that the RFD changes are associated to the largest relative FPKM changes (Figure 4.15B). However, we observed that the highly expressed genes preferentially lay within stable RFD profiles regions (Figure 4.15A).

4.4.1 Coupling between relative transcriptional changes and replication changes

In order to analyze mean expression changes between 2 cell lines, we can either consider the relative transcriptional differences (4.2) or the absolute transcriptional differences (4.1) between pairs of cell lines.

$$\Delta_{absolute}(FPKM_{cell_1}^{cell_2}) =< |FPKM_{cell_1} - FPKM_{cell_2}| >, \tag{4.1}$$

$$\Delta_{relative}(FPKM_{cell_1}^{cell_2}) =< |\log_{10}\frac{FPKM_{cell_1}}{FPKM_{cell_2}}| >, \tag{4.2}$$

where $< . >$ stands for the average over all windows. We compile all pairwise average absolute expression differences (Eq. (4.1)) in a global absolute expression difference matrix (Figure 4.16). Then we computed the same matrix but for each MCR levels from RFD variable regions (MCR<0.42, wich corresponds to 8% of all genes) to RFD conserved regions (MCR>0.57, which corresponds to 37.5% of all genes) (Figure 4.17A-D). Finally, we compute the differences between the average absolute difference matrix in each MCR level and the global absolute difference matrix (Figure 4.17E-H). The blue matrix in Figure 4.17E reveals that we have less absolute transcriptional changes in the variable RFD regions except when associating GM06990 and the other lymphoid cell lines. In contrast, the red matrix in Figure 4.17H illustrates that we have more absolute transcriptional changes in stable RFD regions. This suggests that we have a coupling between the absolute transcriptional changes and replication changes. Complementary, we computed the global relative expression difference (Eq. (4.2)) matrix among the 12 cell lines then we follow the same procedure as done previously with the absolute change (Figure 4.18). The blue matrix in Figure 4.19H reveals that we have less relative transcriptional changes in the stable RFD regions. In contrast the red matrix in Figure 4.19E illustrates that we have more relative transcriptional changes in variable RFD profiles. Thus, we observe a coupling between relative transcriptional changes and replication changes but in the opposite direction to the coupling observed with the absolute differences. Note that the results obtained using the absolute FPKM differences are likely due to the level of gene expression, since the stable RFD profile regions are more highly expressed than the variable RFD profile regions and absolute expression changes are correlated to the average expression level. To summarize the results, we computed the average of the difference matrices (obtained over the 66 independent terms) for both absolute and relative expression changes (Appendix, Figure B.5). This Figure B.5 confirms that we have a global association between the replication program changes and the relative transcription program changes but in opposite directions.

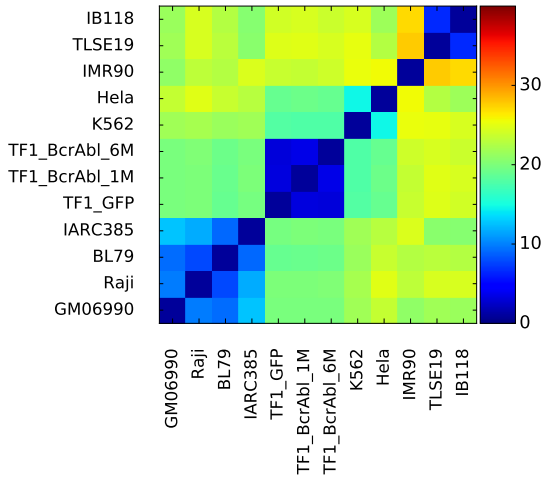

Figure 4.16: **Average absolute difference between gene expression profiles.** Matrice of average FPKM differences per cell line pair (Eq. 4.1)). Color bar represents the absolute difference values.

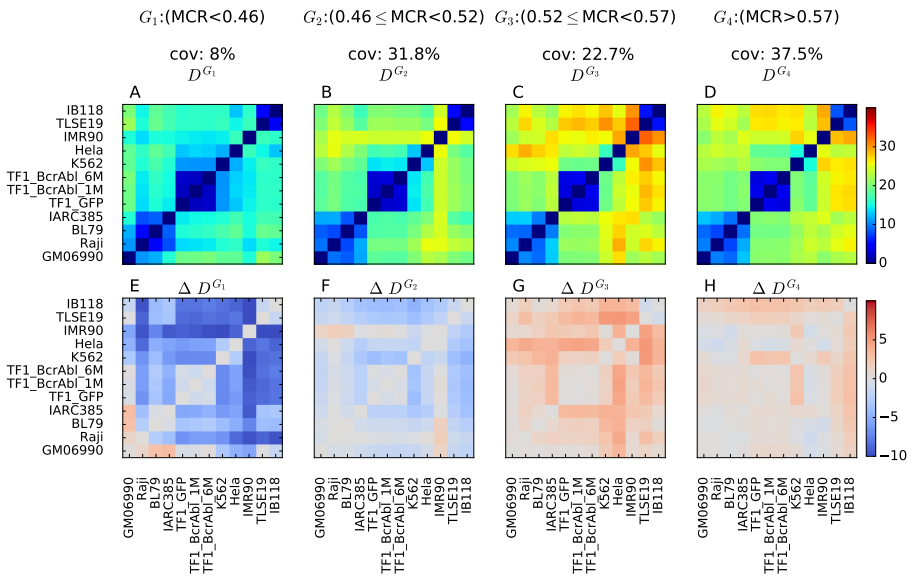

Figure 4.17: **Average absolute difference matrices of FPKM profiles depending on the MCR level.** 10 kb windows were grouped in MCR categories following the 4 groups classification of the human genome in extremely stable RFD profiles regions with MCR≥0.57 (G1, A) , the moderately stable RFD profiles regions with 0.52≤MCR<0.57 (G2, B), the moderately variable RFD profiles regions with 0.42≤MCR<0.52 (G3, C) and the extremely variable RFD profiles regions with MCR<0.42 (G4,D); (F-H) Matrices of differences between average absolute FPKM difference matrices in each MCR levels and the genome-wide matrix (Figure 4.16)

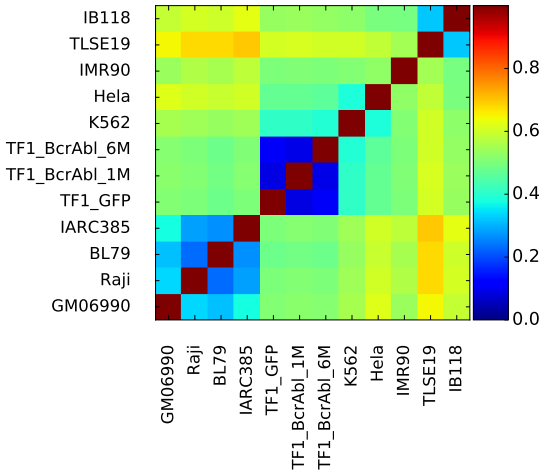

Figure 4.18: **Average relative differences between gene expression profiles.** Same as 4.16 but for relative differences.

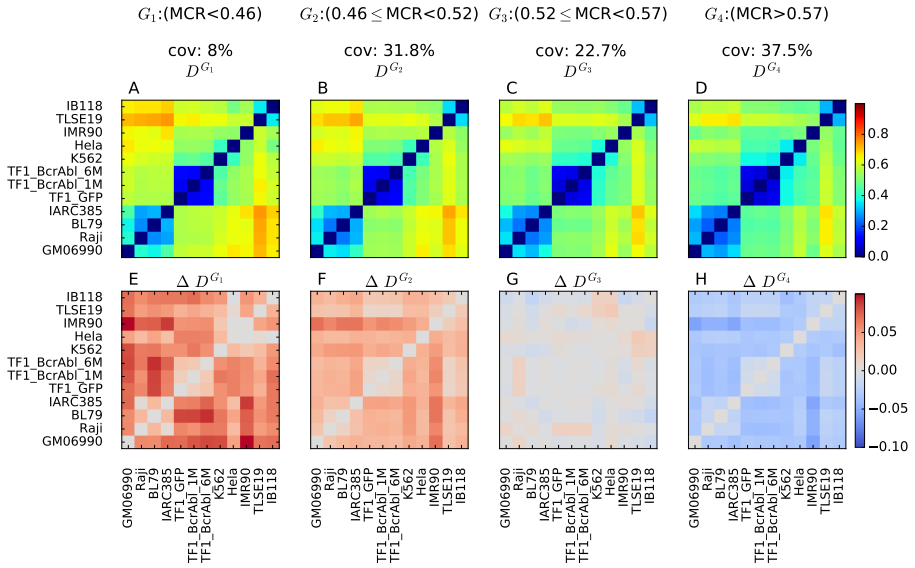

Figure 4.19: **Average relative difference matrices of FPKM profiles depending on the MCR level.** Same as 4.17 but for relative differences.

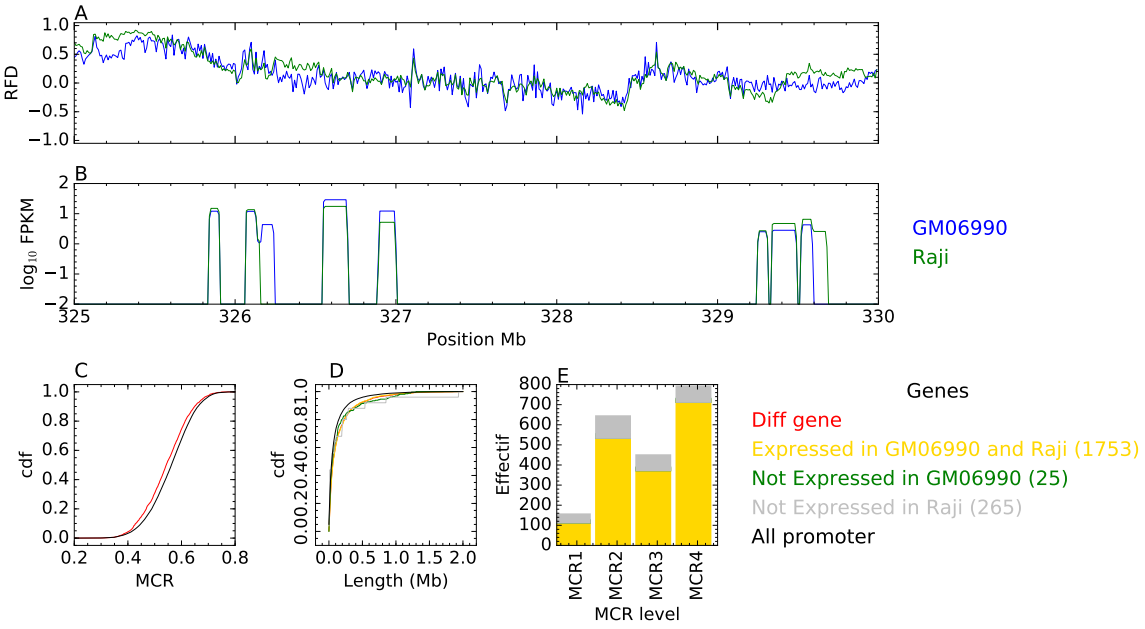

Figure 4.20: **Association between gene expression and RFD profiles in GM06990 and Raji.** (A) RFD profiles and (B) FPKM profiles of GM06990 (blue) and Raji (green) computed in 10 kb windows along a 5 Mb region of chromosome 2. (C) Cdf of the MCR values at the differentially expressed gene promoters (red) and at all promoters (black). (D) Cdf of the length of gene expressed in both GM06990 and Raji (orange), all genes (black), differentially expressed genes (red), expressed in Raji only (green) and expressed in GM06990 only (silver). (E) Counts of MCR level at the promoters of gene expressed in GM06990 and Raji (orange), all promoters (black), differentially expressed genes (red), genes expressed in Raji only (green) and genes expressed in GM06990 only (silver).

4.4.2 No systematic coupling between initiation zone efficiency changes and FPKM changes

To complete the analysis, we increased the resolution of the analysis by comparing the gene expression level in pairs of cell lines. We selected two lymphoid cell lines: GM06990 and Raji. Using Feature count package (section 2.3.5), we counted mapped RNA-seq reads for each genes in each cell lines as the input to DESeq2 analysis. Then, we run the DESeq2 to detect the significant gene expression changes between the two cell lines. The advantage of applying the analysis using the RNA-seq data from the 12 cell lines instead of the data from the 2 considered cell lines is that the pairwise comparison in this case takes into account the variability of genes in all cell lines. The result of DESeq2 showed that cell lines of the same type are classified together based on the hierarchical classification (Appendix, Figure B.6) and it is very compatible with RFD classification results obtained in the Chapter 3. Moreover, the results of the principal component analysis (Appendix, Figure B.7) showed that the HeLa cell line is close to the adherent cell line group and myeloid cell line group. This is very compatible with the RNA-seq correlation matrix (Figure 4.5) where HeLa was classified with myeloid cell line instead of Adherent cell lines. Thus, we have a confirmation of the previous classification results based on RFD and RNA-seq.

We only considered the significantly differentially expressed genes having a $\log_2$ fold change of gene expression >2. We chose this high threshold value to be specific in the selection of differentially expressed genes. Under this assumption, we selected 2045 significant gene expression changes between Raji and GM06990. First, we observed that we selected the longest length genes (Figure 4.20D). Moreover, we observed more genes specifically expressed in GM06990 than in Raji (Figure 4.20E). More importantly, we observed that differentially expressed genes have a lower MCR at their promoters compared to the complete gene set (Figure 4.20C). Therefore, the significant gene expression changes are associated with late replicating regions based on the association between the large domains of late replication timing values and low MCR domains (section 4.3.1). This confirmed our observation that low MCR regions are associated to the largest relative changes of expression (section 4.4.1).

Link between gene expression change and IZ at promoters

We further asked whether the local replication program changes are coupled with local gene expression changes? We thus correlated the IZ efficiency changes to changes of the RFD slopes to the output of DESeq2. For each differentially expressed genes, we compared the RFD slope values at the promoter in GM06990 (Normal) and Raji (Cancer), 2 cell lines of the same type. The histogram in Figure 4.21A shows that the RFD slope changes follow on average the gene expression changes as expected. The blue line shows that we have a larger RFD slope in GM06990 than in Raji when the $\text{FPKM}_{GM06990}$ is greater than the FPKM_{Raji}. This result confirms the previous finding in [3, 171], where the replication initiation efficiency increased near the active TSS increasing co-orientation of replication and transcription at gene 5' ends. However, covariation between FPKM and RFD slope changes at differentially expressed gene promoter was not systematic as negative Δ $\text{slope}^{Raji}_{GM06990}$ were observed for $\text{FPKM}_{GM06990} > \text{FPKM}_{Raji}$ and positive genes for $\text{FPKM}_{GM06690} < \text{FPKM}_{Raji}$. We also chose to compare 2 cell lines with very close RFD profiles, TF_GFP and TF1_BcrAbl_1M (Figure 4.21B). We do not observe a clear co-variation between replication and transcription, likely because RFD changes are too few. Finally, we decided to manually annotate the association between differentially expressed genes and RFD profiles changes in GM06990 and Raji. We did not put any threshold on expression changes for this analysis, so we used 6060 instead of 2045 differentially expressed genes. Only the first 4 chromosome were annotated. We visually analyzed the RFD profiles in GM06990 and Raji $\pm$100 kb of each. Then we defined 4 types of observed changes. Type I when we observed a specific IZ for GM06990 within the $\pm$100 kb window (Figure 4.22A). Type II when we observed to a specific IZ for Raji (Figure 4.22B). Type III when we observed for the 2 cell lines an IZ that may change position and efficiency or have the same position and efficiency (Figure 4.22C). And the case that does not belongs to any of these 3 types are called 'Other' type. We find that 38% (653/1722) of the significant expression genes changes are associated to type III (Table 4.1) where an IZ is observed in the two cell lines, and 25% (427/1722) of the significant expression genes changes are associated to type I and II, where an IZ is observed in the cell line of greater promoter activity only.

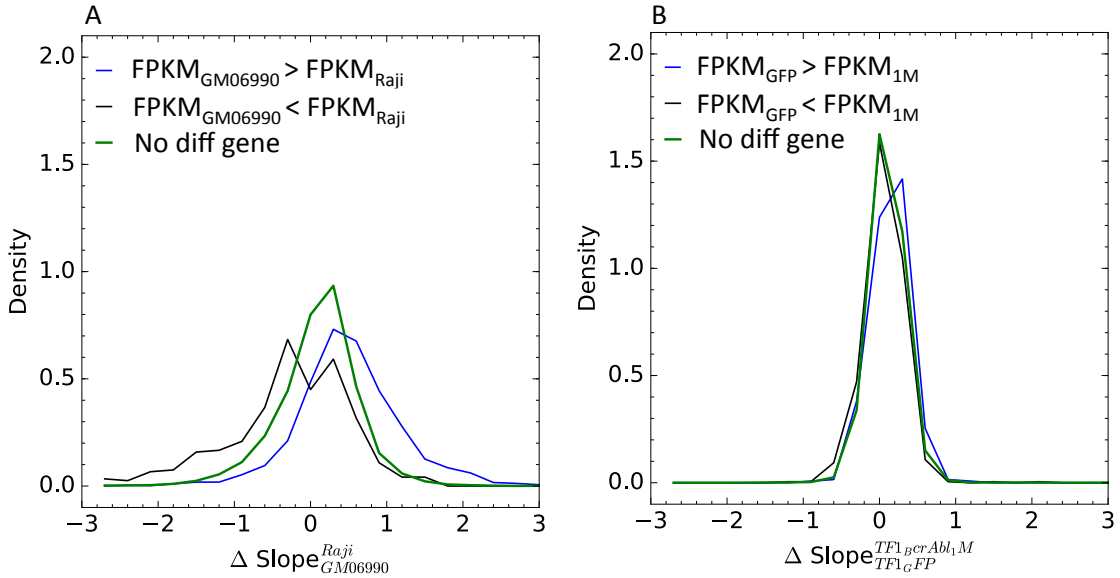

Figure 4.21: **Association between changes in replication initiation efficiency and in gene expression.** (A) The slope differences between GM06990 and Raji ($\Delta\text{slope}^{Raji}_{GM06990} = slope_{GM06990}\text{-slope}_{Raji}$) at the promoter of significantly expressed genes with $\text{FPKM}_{GM06990} > \text{FPKM}_{Raji}$ (green), with $\text{FPKM}_{GM06990} < \text{FPKM}_{Raji}$ (black), with no significant gene expression change in green. (B) Same analysis between TF1_GFP and TF1_BcrAbl_1M.

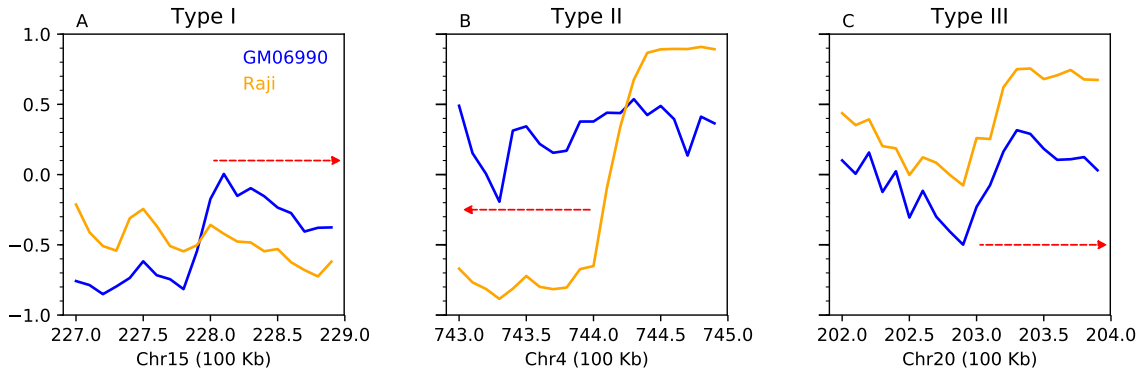

Figure 4.22: **Type of replication changes.** We illustrated the 3 types of RFD profiles differences between GM06990 (blue) and Raji (Orange). RFD profiles were computed at 10 kb. The red arrows correspond to the orientation of the genes. From left to right : (Type I) corresponds to GM06990 specific IZ, (Type II) corresponds to Raji specific IZ and (Type III) corresponds to the common IZ detected in the two cell lines that may change efficiency or shift position

Thus in summary, on the first hand, we detected a coupling between transcriptional changes at some loci but initiation zones changes and on the second hand, no IZ efficiency changes at other loci of differentially expression.

Chromosome	Strand	Type I	Type II	Type III	Other
Chr1	+	42	37	130	131
	−	44	55	144	131
	Total	86	87	274	262
Chr2	+	37	24	77	68
	−	28	23	76	72
	Total	65	47	153	140
Chr3	+	16	20	66	75
	−	13	19	73	72
	Total	29	39	139	147
Chr4	+	19	16	41	46
	−	24	15	46	47
	Total	43	31	87	93

Table 4.1: **Manual annotation of initiation zones efficiency changes at promoter regions of differentially expressed genes between GM06990 and Raji in for chromosomes 1, 2, 3 and 4.** First column represents the chromosome. Second one represents genes orientation. The third represents the count of GM06990 specific IZ (**Type I**), the fourth represent the count of Raji specific IZ (**Type II**), the fifth column for the common IZ detected in the two cell lines that may change efficiency or shift position (**Type III**), finally the last column represents the remaining cases (**Other**).

Link between IZ efficiency changes and transcription changes

The reciprocal analysis considering the regions of changing RFD rather than those with changing FPKM promoter confirmed that we have a decoupling between replication and transcription changes in some regions of the genome (Figure 4.23). We selected the location where the initiation efficiency of GM06990 are higher than in Raji. We first computed the slope difference profile as slope differences : $\Delta\text{slope}^{Raji}_{GM06990}$ =slope$_{GM06990}$-slope$_{Raji}$ in 10 kb non overlapping windows. We then selected the position of the largest $\Delta\text{slope}^{Raji}_{GM06990}$>1.5%RFD per kb. Then to each selected position we associated the FPKM value of the nearest transcribed gene (FPKM>1) for both cell lines. Finally, we computed the expression differences ($\Delta FPKM^{Raji}_{GM06990} = FPKM_{GM06990} - FPKM_{Raji}$) and the relative expression changes ($\log_{10}\frac{FPKM_{GM06990}}{FPKM_{Raji}}$). We observed that the distribution of both expression differences and relative expression changes between GM06990 and Raji is shifted to the right of the corresponding distributions over the complete gene set. We thus observed a global increase of gene expression at region of enhanced replication initiation activity, as expected. In the same manner, we inversed the analysis by taking into account that the regions where the RFD slope in Raji is higher than in GM06990. We observed that the distributions of FPKM differences and relative expression changes between GM06990 and Raji is shifted to the left of the global distribution, corroborating our first observation. We repeated the same analysis for the two lymphoid cell lines (TF1_GFP and TF1_BcrAbl_1M) (Figure 4.23CD), but did not observe expression modifications at the selected regions of high replication initiation efficiency change. This is likely due to the small expression modifications between these two related cell lines.

In order to confirm these results, we then focused on replication IZ identified as local

maxima of the RFD slope profiles. Analysis of 452 IZ strongly efficient in lymphoblastoid cell line GM06990 (>2%RFD per kb) and significantly less efficient in Burkitt's lymphoma cell line Raji (efficiency loss > 0.5%RFD per kb) in relation to gene expression data in these cell lines allowed us to further quantify gene expression changes in regions of IZ efficiency weakening. 119 IZ (26.3%) are not associated to transcription in either cell line (closest expressed protein coding gene at least 50kb away from the IZ). For the 333 remaining IZ, the closest expressed gene changes expression between GM06990 and Raji in 162 cases ($|FPKM_{GM06990}\text{-}FPKM_{Raji}|>2$), correspond to 116 gene repressions ($FPKM_{GM06990} > FPKM_{Raji}$) and 46 gene activations. In other words, if 71.6% (116/162) of gene expression changes associated to IZ efficiency losses are gene repression as expected for a co-regulation of replication and transcription, 74.3% ((452-116)/452) of the IZ efficiency losses are not associated to transcriptional repression. Reproducing the analysis between closely cell lines TF1_GFP and TF1_BcrAbl_1M provided similar results with 86.0% (37/43) of IZ efficiency losses not being associated to transcriptional repression, in particular because in this case gene expression changes are very few.

These results confirm previously reported correlations between transcription and replication changes but underline that it is not systematic and that the source of the link between IZ efficiency variations and transcriptional changes during cell differentiation remains to be fully understood.

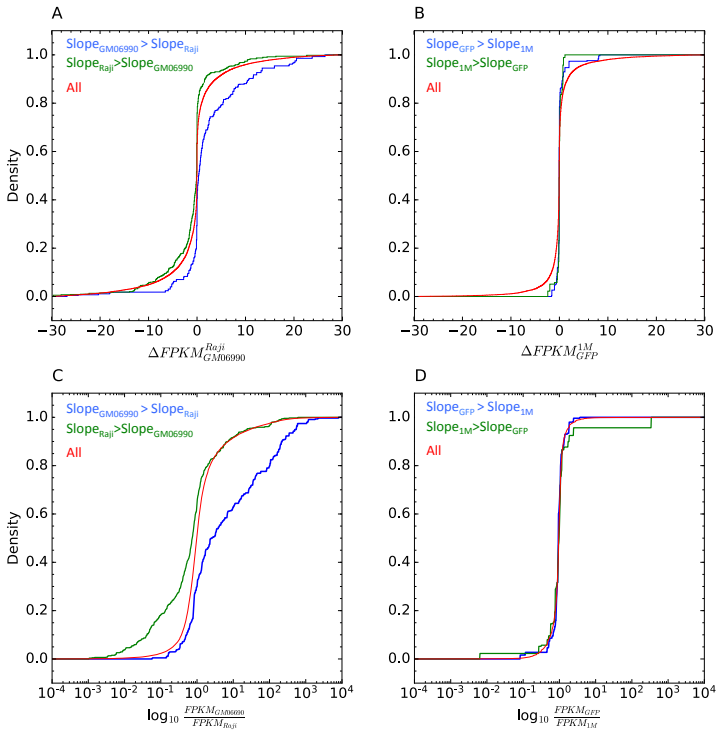

Figure 4.23: **Association between gene expression change according to largest** $\Delta slope$ We represent the absolute (resp. relative) $\Delta FPKM$ according to the largest $\Delta Slope$ in two comparison (A, resp.C) between GM06990 and Raji, (B, resp. D) between TF1_GFP and TF1_BcrAbl_1M.

4.5 Conclusion

It is widely accepted that the replication of the human genome proceeds in a time-ordered manner, correlated with parameters as gene activity, chromatin structure, and nuclear location. However, the evidence supporting this view depends on the scale of analysis. For example, chromosome replication bands have shown a correlation between R bands[5] and gene density. Fine scale, genome-wide analysis revealed that these correlation patterns were not as strong. The early and late replication domains detected by Repli-seq method correlate very well with the subnuclear A (euchromatic, active) and B (heterochromatic, inactive) compartments defined by Hi-C [38, 94]. This observation provided a better understanding of the complex association between replication timing and transcription. Moreover, other analyses using the replication timing of individual genes revealed a strong correlation between transcription and early replication [178–182]. Expressed genes replicate at early stages of S phase, whereas not transcribed genes replicate at late stages of S phase [183–185]. However, it was also found that same active genes are additionally found in the late replicating regions and that in the early replicating regions there are dormant genes [37, 46]. This result was also observed recently by Gilbert team [186], as they found that the average changes in transcription are coordinated with replication timing changes but can be anti-correlated for individual gene. Experiments in chicken cells illustrated this complex connection between transcription and replication timing [187]. It was demonstrated that origins flanked by the transcription factor binding site USF1[6] (for "Upstream Stimulatory Factor") contains HS4. The HS4 insulator has the capacity to impose a shift to earlier replication. This adjustment in replication timing does not require that transcriptional action be related with it despite the fact that it is enhanced when a gene is transcribed close-by.

In a first part of this chapter, we extended the global unbiased correlation approach to demonstrate that MRT profiles clustered in three separate groups corresponding to lymphoid, myeloid and adherent cells. Similar results were obtained by RNA-seq, except that HeLa clustered with myeloid instead of adherent cells. Therefore, cancer-associated changes in replication do not blur their developmental origin signature, although changes in gene expression may sometimes do so. It is notable that the two LMSs are more correlated to each other by RNA-seq than by RFD, which suggested that their cell of origin may be different and that the selection for a tumor phenotype may have resulted in a stronger convergence of their transcription than their replication program. We did not detect any evidence for convergence of the replication or transcription programs of cancer cells from different developmental origins (e.g. LMS vs. BLs). In contrast, within lymphoid cells, we find evidence for BL-specific replication and transcription patterns. Overall, our global correlation analyses provide evidence for the existence of recurrent replication and transcription changes along specific tumor progression pathways. Interestingly, RFD changes induced by BCR-ABL1 expression

[5] R bands are guanine-cytosine rich, and adenine-thymine rich regions are more easily denatured by heat.
[6] This protein is able to induce transcription through a pyrimidine-rich promoter element.

in early CML were not associated with large-scale MRT switches comparable to those observed between early and late CML or previously reported between leukemias and control LCLs [154]. The global correlation analyses further revealed that RFD changes between cell lines are widespread through the genome but more frequent in GC-poor regions. In contrast, RNA-seq changes do not vary uniformly with GC content. These results strengthen the notion that replication changes are dissociated from transcription changes, to an extent that specifically depends on the compared cell types.

In a second part, we aimed to understand and characterize the stable and variable regions of RFD profiles among the 12 cell lines detected using the MCR profiles. We found that the largest late replicating regions are associated to large domains of low MCR profiles (i.e. variable regions of RFD). In contrast the largest early replicating regions are associated to large domains of high MCR profiles (i.e. stable regions of RFD). Then, we determined replication initiation zones locations based on the RFD slope profiles. Early replication and RFD stability was associated with a high density of the most efficient IZ that tend to be conserved between cell lines. In contrast, regions of late replication associated to variable RFD profiles were characterized by a low density of IZ with comparatively smaller efficiency that tend to be cell line specific. Furthermore, we quantified directly the association between the replicative changes and transcription changes. We found that there is a coupling between RFD change and relative transcriptional change on average. These conclusions confirmed the results in the literature [3, 171]. We validated this analysis by doing pairwise comparison of gene expression and RFD slope profiles. We observed that the significant transcriptional changes are not systematically associated to a large RFD changes. In the same manner, we found that the initiation zones efficiency changes are not systematically associated to large transcription modifications. We concluded that there is no causal association between transcriptional changes and initiation zones efficiency changes.

CHAPTER 5

Application in Chronic Myeloid Leukemia System

To characterize cancer cells, many tools have been developed. These include the determination of phenotypic properties, using cell culture [188]. Other studies focused on the specific differences in genes expression patterns between the normal and cancer cell lines [189]. Ross et al. [190] explored the relationships between patterns of gene expression in breast tumor derived cell lines and those from clinical tumor specimens. It was also shown that these cell lines and tumor samples have distinct gene expression patterns [190]. The unsupervised classification of this cell lines led to the separation of the breast tumor derived cell lines from those from clinical tumor specimens. Along the same line, sets of genes responsible for the differences between solid tumors and cell lines in their response to cancer have been identified by Serial Analysis of Gene Expression [191]. Recently, a comparative analysis of the replication timing profiles in six human cancer cell lines was performed and defined common replication timing regions between them [192].

In order to study how the DNA replication program differs between cancerous and non-cancerous cell lines, we will compared the RFD profiles of 4 myeloid cell lines that form a system of evolution of a Chronic Myeloid Leukemia cancer tumor with time.

5.1 Introduction

The research is underway to determine the metastatic behavior of cancer cells *in vivo*. Cancer is linked to the uncontrolled proliferation[1] of cells resulting in disruption of tissue homeostasis [2]. Cancer is first and foremost a DNA disease and the environment is associated with this process. The development of a tumor occurs in successive stages, a single alteration of the DNA is not enough. The genome of cancer cells becomes more distinct from that of normal cells as cancer progresses [193]. Unlike normal cells, cancer cells are characterized by genetic instability, which is related to a deficiency of the genome repair systems. This instability allows the accumulation of DNA alterations [194] that can be from genetic or epigenetic source. These genetic abnormalities may be due to the intervention of exogenous or endogenous factors. Alterations include alterations during replication (accidental) and alterations after replication (physical, chemical, biological, etc.) [4]. These genetic instabilities cause so called DNA Replication Stress (RS). RS is a term that extensively characterizes obstacles in DNA replication and by large incorporates slowing down and breakdown of DNA replication forks [195, 196]. The meaning of replication stress is constantly developing and hard to define accurately. RS emerges from a wide range of sources and has various ramifications in the cell [195]. It can be a re-replication or a silencing of the replication origins [195]. In addition, there are other extrinsic factors that damage DNA. Intracellular problems also cause RS such as sites that are difficult to replicate, chromatin accessibility, or collisions between the replication and transcription machines [195, 197, 198].

Oncogene expression can induce RS and trigger DNA damage from the earliest tumorigenesis stages [199–203]. In precancerous lesions, RS induces a DNA damage response (DDR) that can trigger senescence or apoptosis. Tumorigenesis proceeds when the DDR is downregulated (e.g. by p53 mutation), favoring cell proliferation with genome instability [199–203]. Oncogenes have been proposed to trigger RS by multiple mechanisms: reduced or increased origin firing, exhaustion of limiting nucleotides or replication factors, increased transcription and replication-transcription conflicts. For example, in *Xenopus* egg extracts, in which no transcription takes place, addition of recombinant Myc increases origin firing, fork stalling, and DNA breakage in a manner dependent on Cdc45, a limiting origin firing factor, and these effects are recapitulated by addition of recombinant Cdc45 alone [122, 204]. In contrast, overexpression of HRASv12 in cultured cells stimulates RNA synthesis and RS in a manner dependent on the TATA-binding protein (TBP), a general transcription factor, and these effects are recapitulated by overexpression of TBP alone; increased origin firing seems to be a consequence rather than a cause of RS in this case [205]. Recently, a novel nascent DNA mapping assay was used to show that overexpression of Cyclin E1 or MYC, which shortens G1 phase, induces novel intragenic origins, normally erased by transcription during G1, that

[1] Cell proliferation is the process that leads to an increase in the number of cells. It is determined by balancing scissions and cell loss by cell death or differentiation. Increase cell proliferation in tumors.

[2] Process involved in maintaining a fixed internal state in specific tissues of the body, including control of cell proliferation and cell death and control of metabolic function.

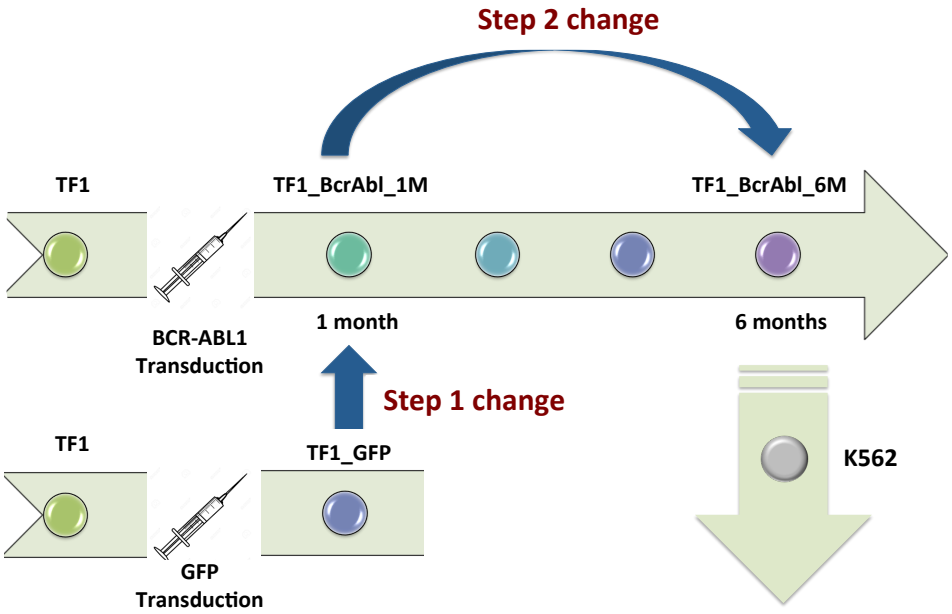

Figure 5.1: **Early CML progression model.** The CML, progression model includes the TF1_GFP cell line, the TF1_BCRABL_1M cultured 1 month after the transduction of the BCR-Abl1 gene and TF1_BCRABL_6M cultured 6 months after the transduction of the BCR-Abl1 gene. Step 1 (resp. 2) changes correspond to the changes of RFD profiles between TF1_GFP and TF1_BcrAbl_1M (resp. TF1_BcrAbl_1M and TF1_BcrAbl_6M)

are particularly prone to fork collapse due to conflict with transcription [206]. However, this study only interrogated the earliest-replicating, gene-rich part of the genome, and ectopic origins were only induced in cells with the shortest G1 phase. It remains unclear if oncogene expression can more globally disrupt the spatiotemporal program of DNA replication. Many techniques exist to clarify the RS in direct measurement of firing replication origin and replication fork progression, for example, DNA fiber stretching techniques. The latter allows to visualize the replication program, but its values are probably limited to the genomic position of the observed fiber. Recent techniques as nanopore sequencing [207, 208] will soon allow the direct detection of stalled forks on genomic positions. However, in this chapter we will use the RFD profiles to compare the efficiency of the initiation zones [3] in a cell line system of cancer proportion.

5.2 Chronic Myeloid Leukemia System

To study the association between cancer and replication program, we choose to compare the IZ efficiency change in a Chronic Myeloid Leukeamia (CML) [127, 209] progression system. This cancer is characterized by the presence of the Philadelphia chromosome that contains a fusion gene called BCR-ABL1 resulting from the combination of the BCR and ABL genes as shown in the Figure C.1 [127, 209]. The BCR gene is located on the chromosome 22 and the

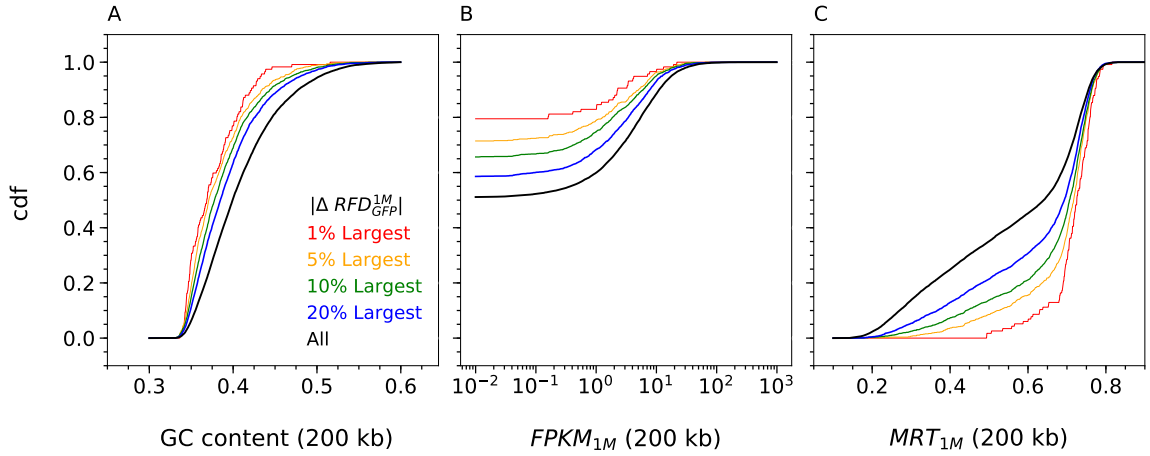

Figure 5.2: **The largest RFD changes induced by 1 month of BCR_ABL1 expression are observed in GC-poor, lowly-expressed and late replicating regions.** Cumulative distribution functions (cdf) of GC content (A), transcription in TF1_BrcAbl_1M $FPKM_{1M}$ (B) and MRT in TF1_BrcAbl_1M (C) computed in non-overlapping 200 kb windows of the 22 autosomes. Cdfs were determined for all windows (all, black) or selected windows with the 20% (blue), 10% (green), 5% (yellow) and 1% (red) largest RFD changes between TF1_GFP and TF1_BrcAbl_1M.

ABL gene is located on chromosome 9. A translocation between the ABL of the chromosome 9 that contains the ABL gene and a part of chromosome 22 leads to the fusion of the ABL gene with the BCR gene to form the BCR-ABL1 gene (Appendix, Figure C.1). An early Chronic Myeloid Leukemia (CML) System was constructed based on the BCR-ABL1 negative TF1 cell line (Figure 5.1, section 2.1). BCR-ABL1 gene was transduced with GFP (TF1_GFP) or a BCR_ABL fusion (TF1_BcrAbl). The latter were analyzed after culturing for 1 month (TF1_BcrAbl_1M) or 6 months (TF1_BcrAbl_6M) following transduction. These constructions were also compared with K562 a late stage of CML model (section 2.1)). We will analyze the initiation zones efficiency changes between TF1_GFP and TF1_BcrAbl_1M as step 1 and between TF1_BcrAbl_1M and TF1_BcrAbl_6M as step 2 (Figure 5.1).

5.3 TF1 cell lines clustered in accordance to CML progression in RFD and RNA_seq based classifications.

We computed the differences in RFD profiles in non overlapping 200 kb windows between the TF1_GFP and TF1_BcrAbl_1M, and selected respectively the 1, 5, 10 and 20% largest RFD changes. We associated to these windows the GC content, MRT in TF1_BcrAbl_1M and gene expression level of TF1_BcrAbl_1M. We showed that the largest changes in the replication fork directionality are associated with poor GC content, late replication timing and low expression genes (Figure 5.3). Same results were obtained for the largest RFD changes between TF1_BcrAbl_1M and TF1_BcrAbl_6M. These observation are consistent with the results reported in chapters 3 and 4.

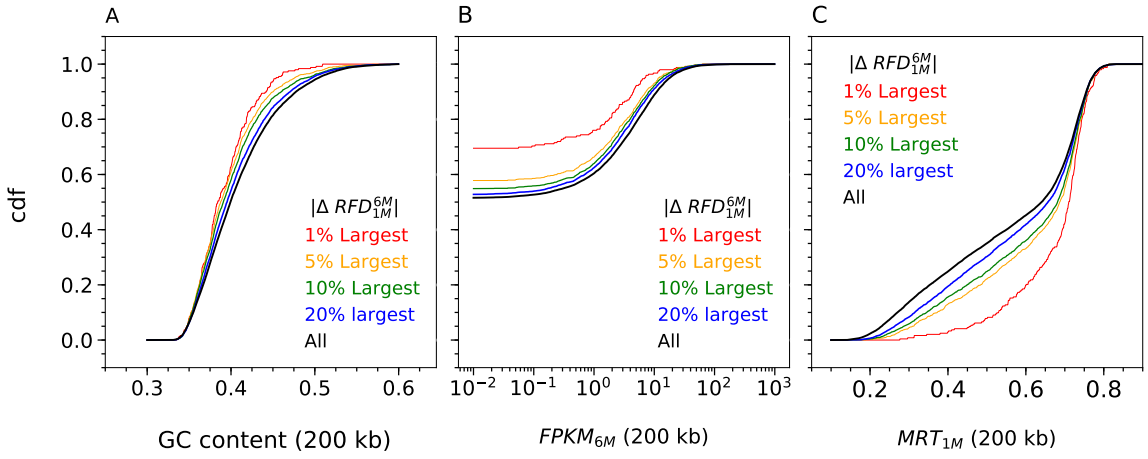

Figure 5.3: **The largest RFD changes between 1 month and 6 months of BCR_ABL1 expression are observed in GC-poor, lowly-expressed and late replicating regions.** Cumulative distribution functions (cdf) of GC content (A), transcription in TF1_BrcAbl_1M (B) and MRT in TF1_BrcAbl_1M (C) computed in non-overlapping 200 kb windows of the 22 autosomes. Cdfs were determined for all windows (all, black) or selected windows with the 20% (blue), 10% (green), 5% (yellow) and 1% (red) largest RFD changes between TF1_BrcAbl_1M and TF1_BrcAbl_6M.

As discussed in chapter 3, RFD profiles allow to classify the cell lines in accordance to their tissue of origin. Here, we zoom in the global RFD correlation matrix to clearly see the relation between the TF1 cell lines and K562 (Figure 5.4A). The correlation values between K562 and the TF1 cell lines increase in accordance with early CML progression. This result confirms the robustness of the RFD as a good identificator of the replicative change in this cancer progression system. It could be related to specific origin activation or silencing. We reached similar conclusions the similar using RNA-seq (Figure 5.4B). In contrast, the relationship between K562 and TF1 cell lines is inverted when using MRT. Note that RFD profiles are more discriminant than the RNA-seq, since the differences between the correlation values of TF1_GFP and TF1_BCRABL_6M with K562 are about 0.2 for RFD profiles and < 0.05 for RNA-seq profiles (Figure 5.4). However, in the two cases the evolution of tumor cancer with time is clearly illustrated.

The interplay between the replication origin activity and the genome instability in yeast uncovered a role for the organization of the DNA replication in delimiting the genetic instability [210]. Moreover a thousand of mutations were identified and found to be heterogeneous within and between tumors [211]. They could be generated by replication stress associated instability [212]. Hence, we hypothesized that the detected differences in RFD correlation values are related to the efficiency changes of the replication Initiation Zones. We thus compared the IZ among the three TF1 cell lines as a possible signature of RS.

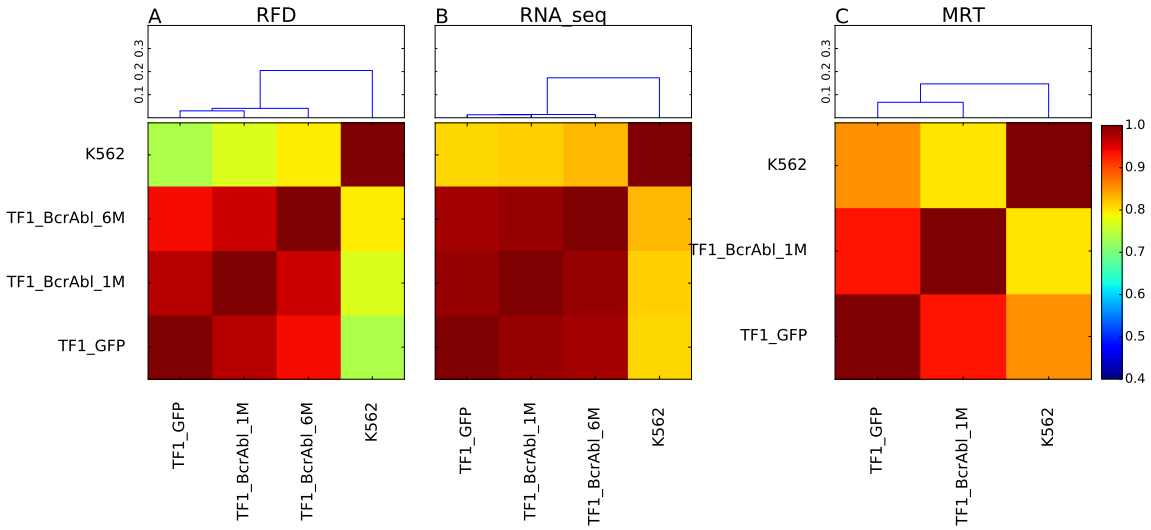

Figure 5.4: **The cell lines clustered in accordance to CML progression.**(A-C) Correlation matrices between RFD profiles in accordance to CML progression (A), RNA-seq (B) and MRT profiles (C); Pearson correlation coefficient values are colour-coded from blue (0.4) to red (1) using the colour bar on the right. (Top) A corresponding dendrogram representation of the hierarchical classification of cell lines is shown on top of each correlation matrix; ordinate is the correlation distance.

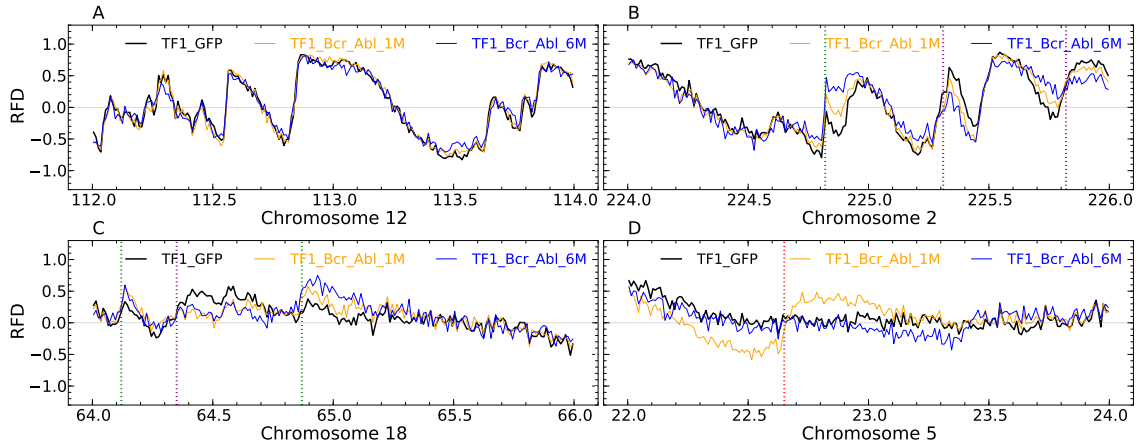

Figure 5.5: **Manual annotation of RFD profile changes during the first two steps of the CML progression model.**(A-D) RFD profiles computed in 10 kb non-overlapping windows for TF1_GFP (black), TF1_BrcAbl_1M (yellow) and TF1_BrcAbl_6M (blue) are visualized in 2 Mb regions along the 22 human autosomes. (A) Region along chromosome 12 where RFD profiles in the 3 cell lines do not present any significant difference. (B) Region along chromosome 2 illustrating IZs whose efficiency is repeatedly enhanced (green) or weakened (purple) in Steps 1 and 2 of the model CML progression. (C) Region along chromosome 18 having two IZ efficiency changes in Step 1 (enhanced, leftmost green; weakened, purple) that are confirmed in Step 2. The rightmost green line marks an IZ which is repeatedly enhanced at Steps 1 and 2 (as in B). (D) Region along chromosome 5 illustrating an IZ activated during Step 1 (red line) and silenced during Step 2.

5.4 Manual annotation of initiation zones efficiency changes

RFD profiles changes in the first two steps of the CML initiation and progression model (Step 1: TF1_GFP − > TF1_BCRABL_1M (Figure 5.1); Step 2: TF1_BCRABL_1M − > TF1_BCR_ABL_6M (Figure 5.1)) were annotated by manually scanning the profiles in 2 Mb windows. An IZ present in the initial state was annotated as Silenced if no IZ was present in the following state. Weakened when IZ efficiency decreased between the two consecutive states (Figure 5.5C). Enhanced when IZ efficiency increased between the two consecutive states (Figure 5.5B). IZs inactive in the initial state but active in the final state were annotated New (Figure 5.5D). In contrast, we do not observed any change in MRT profiles (Appendix V Figure C.2A) or transcription profile (Appendix V Figure C.2B). The regions where we have a change are regions desert of active genes and late replicating. Moreover, each locus annotated for a Step 1 or Step 2 change (total 1027) was also annotated for its status in Step 3 (TF1_BCRABL_6M − > K562). Step 3-specific changes were too numerous to be manually annotated. The 1027 manually annotated loci included 253 IZs which changed efficiency at Step 1 but not Step 2, 551 IZs efficiency changes during Step 2 but not Step 1, and 223 IZs efficiency changes in both Step 1 and Step 2. In total, this database encompasses 476 and 774 and 716 efficiency changes in Steps 1 and 2 and 3, respectively (Table 5.1).

5.5 BCRABL1 expression continuously induces replication changes in gene desert

5.5.1 The targeted initiation zones were more frequently weakened than enhanced

We focused further our analysis on the early CML progression model, where changes in RFD can be directly attributed to changes in expression of the BCR_ABL1 oncogene in TF1 cells. Consistent with their high global correlation coefficients (>0.95), we observed a striking identity of the RFD profiles of the three cell lines over most of the genome, as exemplified in Figure 5.5A. This facilitated the detection and manual annotation of IZ efficiency changes, scored as new, enhanced, weakened or silenced IZs, at each step of CML progression (section 5.4). A few examples of annotated 2Mb segments are shown on Figure 5.5. RNA-seq and MRT profiles of the same regions are shown on Figure C.2. Along the genome, 476 changes during Step 1 and 774 changes during Step 2 were observed (Table 5.1). The distributions of IZ efficiency changes were strikingly similar at Step 1 and Step 2 (Figure 5.6A). Weakened IZs were by far the most frequent at each step (55% and 66%, respectively) and enhanced IZs the second most frequent. We then analyzed the evolution of the 476 Step 1 changes during Step 2 (Figure 5.6B). The behavior of the Step 1 changes during Step 2 significantly depended on the type of change at Step 1 ($P < 0.001$ using a χ^2 test of independence). Step 1 changes, whatever their direction, were most frequently

		TF1_BrcAbl_1M			
		Active IZ		Inactive IZ	Total
TF1_GFP		*Enhanced*	*Weakened*	*Silenced*	
	Active IZ	127	260	39	426
		New			
	Inactive IZ	50		–	50
	Total	437		39	476

		TF1_BrcAbl_6M			
		Active IZ		Inactive IZ	Total
TF1_BrcAbl_1M		*Enhanced*	*Weakened*	*Silenced*	
	Active IZ	123	514	86	723
		New			
	Inactive IZ	51		–	51
	Total	688		86	774

		K562			
		Active IZ		Inactive IZ	Total
TF1_BrcAbl_6M		*Enhanced*	*Weakened*	*Silenced*	
	Active IZ	157	230	315	702
		New			
	Inactive IZ	14		–	14
	Total	401		315	716

Table 5.1: **Database of replication initiation zone change of efficiency.** Summary of the manual annotation of RFD profiles for changes in IZ efficiency between TF1_GFP and TF1_BrcAbl_1M (top), TF1_BrcAbl_1M and TF1_BrcAbl_6M (middle) and, TF1_BrcAbl_6M and K562 cell lines (bottom). Note that for the latter case, only the 1027 loci presenting an IZ efficiency change in at least one of the two first comparisons were analysed. IZ change types (*New*, *Enhanced*, *Weakened* and *Silenced*) were organised to highlight the active or inactive status of IZ in each cell line.

(253/476) confirmed during Step 2. New or enhanced IZs at Step 1 that changed again at Step 2 (n=88) were most frequently enhanced (n=50). In contrast, weakened IZs at Step 1 that changed again at Step 2 (n=131) were most frequently further weakened (n=89) or silenced (n=33). Therefore, there was a significant tendency for IZ efficiency changes during Step 2 to follow the same direction as observed during Step 1. These results demon-

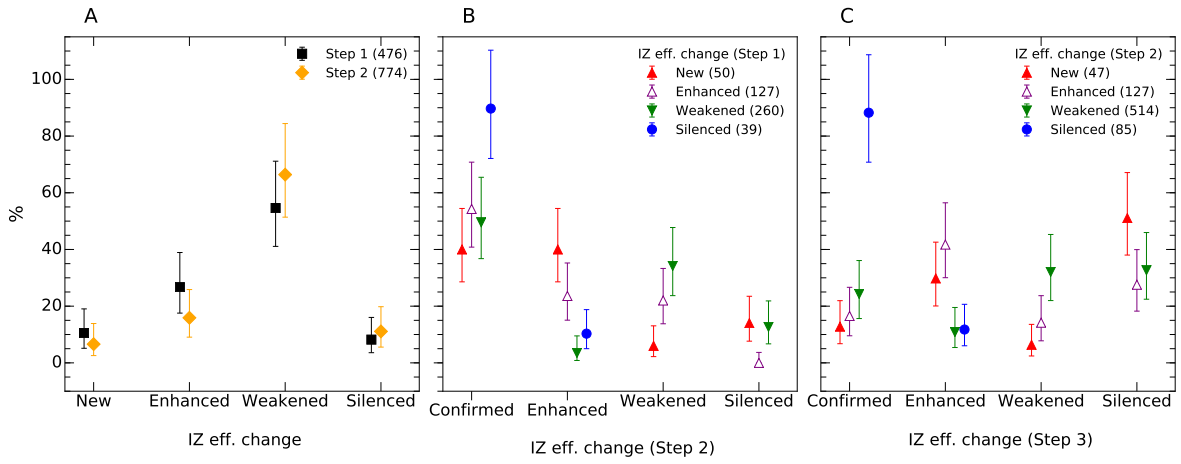

Figure 5.6: **Persistence of replication initiation zones efficiency changes in the CML progression model.**(A) Distribution of IZ efficiency changes (New, Enhanced, Weakened and Silenced) in Step 1 (black; n= 476) and Step 2 (yellow; n=774) of CML progression). (B) Evolution of Step 1 changes during Step 2, reported separately for each type of change. (C) Evolution of Step 2 changes in K562, reported as in (B). Error bars represent a 95% confidence interval assuming data counts follow a Poisson distribution.

strate that BCR_ABL1 expression gradually alters the efficiency of specific IZs in TF1 cells during long-term growth, resulting in a progressive destabilization of their replication program.

We then analyzed how IZs that changed during Step 2 behave in K562 cells, a model for advanced CML (Figure 5.6C). The behavior of Step 2 changes in K562 significantly depended of the type of change at Step 2 ($P < 0.001$, χ^2 test). Among the 85 IZs that had been silenced at Step 2, 75 (88%) remained silent in K562. Among the 514 IZs that had been weakened, a majority (65%) were further weakened (n=165) or silenced (n=168). In contrast, a majority (58%) of the 127 IZs that had been enhanced were further enhanced (n=53) or confirmed (n=21). Nevertheless, among the 47 new IZs appeared at Step 2, only a small half (43%) were confirmed (n=6) or enhanced (n=14) whereas the majority (57%) was silenced (n=24) or weakened (n=3) in K562. In summary, there was a significant overall tendency for the activity changes observed in K562 at these 773 IZs to occur in the same direction as at Step 2, reminiscent of the tendency of Step 2 changes to occur in the same direction as Step 1. However, a stronger tendency to silencing was observed in K562 than at Step 2 (Table 5.1).

Complementarily, we decided to classify the detected IZ efficiency changes according to their Mean Correlation Replication score (MCR) (section 3.4.3). We computed the cumulative distribution function of MCR at 500 kb for each innovation in each step. We observed that the weakened IZ efficiency changes are more associated to high MCR value in step 2 ∼50% of weakened IZ efficiency changes between TF1_BCR_ABL_1M and TF1_BCR_ABL_6M are higher than 0.57 (Figure 5.7 right). It corresponds to regions where RFD profiles are very conserved. In contrast, 50% of weakened IZ in step 1 had a MCR under ∼ 0.48 (Figure 5.7 left). This suggests that the weakened IZ changes in Step 1 are more associated to variable

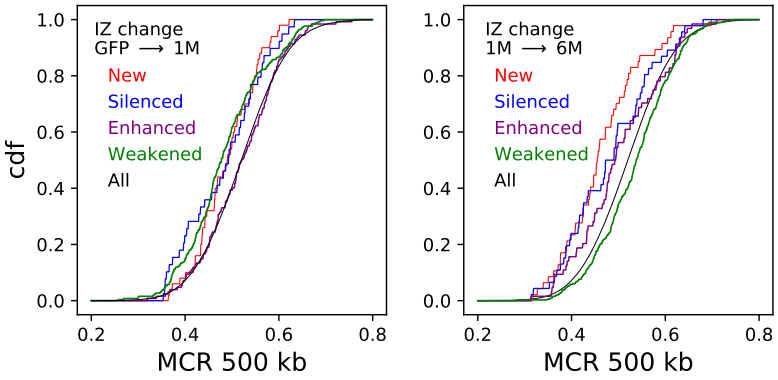

Figure 5.7: **Weakened initiation zones efficiency changes in step 2 are more associated to conserved RFD profiles.** Cumulative distribution function of MCR at 500 Kb in accordance to the IZ changes in step 1 (left) and step 2 (right). Red lines for New firing origin. Blue lines for silenced origin. Purple for Enhanced IZ, and Green lines for Weakened IZ.

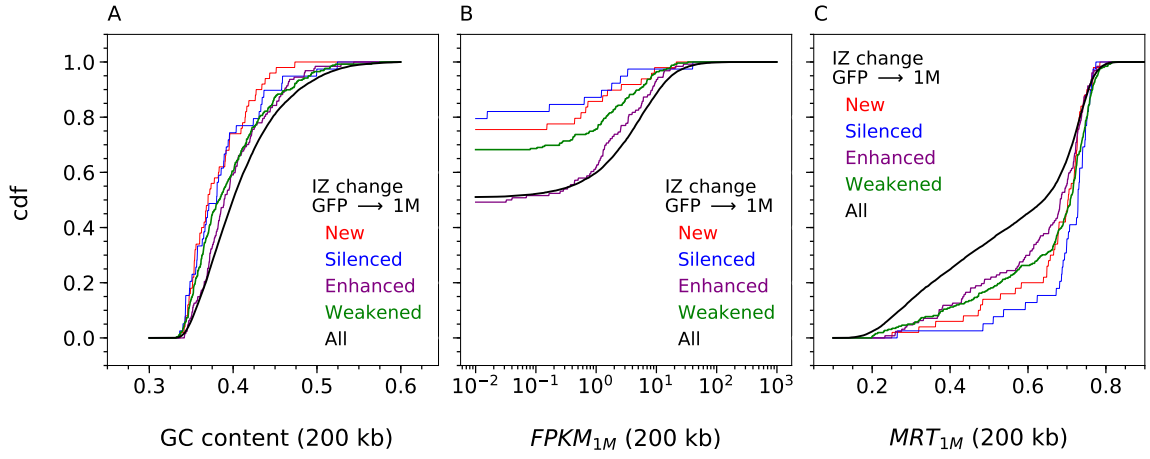

Figure 5.8: **Initiation zones efficiency changes in response to 1 month of BCR_ABL1 expression are observed in GC-poor, lowly-transcribed and late replicating regions.** Cumulative distribution functions (cdf) of GC content (A), transcription in TF1_BCRABL_1M (B) and MRT in TF1_BCRABL_1M (C) computed in non-overlapping 200 kb windows of the 22 autosomes. Cdfs were determined for all windows (all, black) or limited to windows with Silenced (blue), Weakened (green), Enhanced (violet) and New (red) IZ. Similar results are obtained using transcription and MRT data of TF1_GFP (Figure C.3).

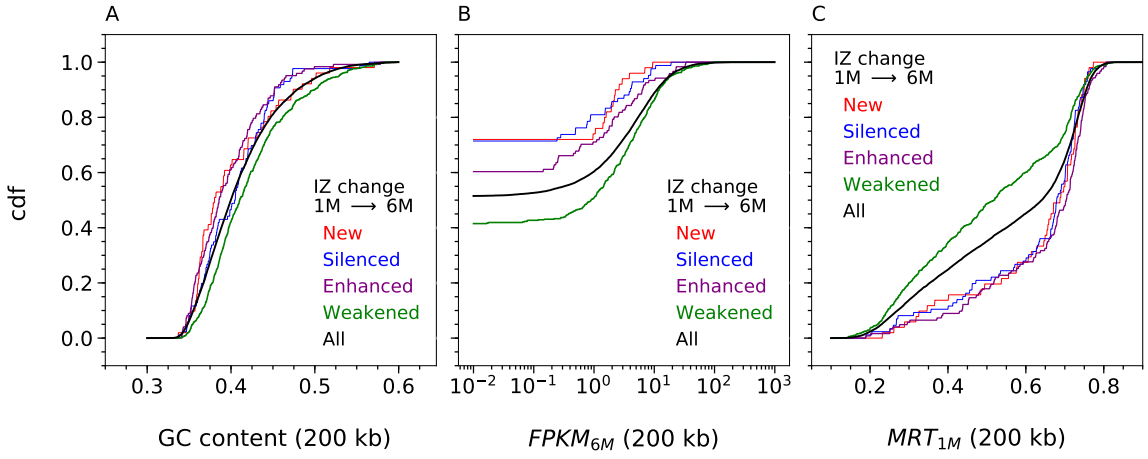

Figure 5.9: **Initiation zones efficiency changes between 1 month and 6 months of BCR_ABL1 expression are observed in GC-poor, lowly-transcribed and late-replicating regions, except for weakened IZs which show the opposite tendency.** Cumulative distribution functions (cdf) of GC content (A), transcription in TF1_BCRABL_6M (FPKM_6M) (B) and MRT in TF1_BCRABL_1M (C) computed in non-overlapping 200 kb windows of the 22 autosomes. Cdfs were determined for all windows (all, black) or limited to windows with Silenced (blue), Weakened (green), Enhanced (violet) and New (red) IZ between 1 month and 6 months BCR_ABL1 expression.

RFD profiles than the IZ efficiency changes that occurs during the step 2. In contrast, the enhanced IZ efficiency changes are more conserved in step 1 (Figure 5.7 left) than step 2 (Figure 5.7 right). Moreover, we observed that the large changes in IZ efficiency (New and silenced) in the two steps are always associated with variable RFD profiles in accordance with the observation cell line specific IZ were characteristic of low MCR domains (Chapter 4). These results suggest that the Enhanced and Weakened IZ efficiency changes might be related to different replication modifying processes than New Silenced IZ events.

In summary, BCR_ABL1 expression during early CML progression changed replication predominantly in GC-poor, lowly expressed and late replicating regions. The targeted IZs were more frequently weakened than enhanced. Targeted IZs in early CML tended to further change activity in the same direction at later tumour progression stages. Therefore, BCR_ABL1 had a long-lasting action on IZs but the direction of the change depended on the targeted region. These results suggest a potential mechanism for generating RS and genome instability independently of transcription by perturbed replication of GC-poor, late-replicating gene deserts.

5.5.2 Weakened initiation zones between 1 month and 6 months of BCR-ABL1 expression are associated with transcription repression.

Many studies suggest a strong correlation between replication timing and transcription in drosophila [100] and mammals [50, 92]. They found a strong correlation between DNA replication at the beginning of the S phase and transcriptional activity. Replicating changes are associated with transcription changes for weak promoters more than strong promoters [37].

In the Figure 5.8 we associated for the step 1 changes categories (New, silenced, Enhanced and weakened) the GC content, MRT and genes expression level represented by the FPKM of TF1_BcrAbl_1M at 200 kb. We confirmed that Step 1 changes preferentially occurred in GC-poor, lowly expressed and late replicating regions (Figure 5.8). Interestingly, this tendency was more pronounced for new or silenced IZs than for weakened or enhanced IZs, in other words for more extreme changes. Similar results were obtained whether RNA-seq and MRT data from TF1_BCRABL_1M (Figure 5.8) or from TF1_GFP (Appendix, Figure C.3) were used. A more complex situation was observed for Step 2 changes (Figure 5.9). New, enhanced and silenced IZs again concentrated in GC-poor, lowly expressed and late replicating regions, but weakened IZs now were more often found in GC-rich, highly expressed, early replicating DNA, coherently with their preferential location in high MCR regions at this step (Figure 5.7). To address whether weakened IZs at Step 2 were associated with nearby gene transcription changes, we plotted the RNA expression (by 200 kb windows) ratio in TF1_BCRABL_6M over TF1_BCRABL_1M, as a function of their mean expression level (Figure 5.10A, MA plot) and computed the cumulative distribution of the expression changes (Figure 5.10B), for the total genome or for windows containing at least one weakened IZ at Step 2. The results indicate that IZ weakening events at Step 2 were significantly associated with nearby transcription repression. We observed that correlation coefficients between cell lines were generally smaller by RFD (Figure 3.7A) than by RNA-seq (Figure 4.5). We investigated in the CML system whether, when focusing on early replicating regions, changes in RNA predicted changes in RFD and IZs. The distribution of RNA expression changes in early replicating windows with the largest RFD changes at Step 2, was shifted towards RNA repression when compared to all early windows (Figure 5.11AB), consistent with the association of weakened IZs with transcription repression (Figure 5.10). However, we found no dependence of the distribution of Step 2 RFD changes on RNA expression changes in early replicating regions (Figure 5.11C). Therefore, although the largest RFD changes in early replicating regions were associated with transcription repression, changes in RNA expression did not reciprocally predict RFD changes.

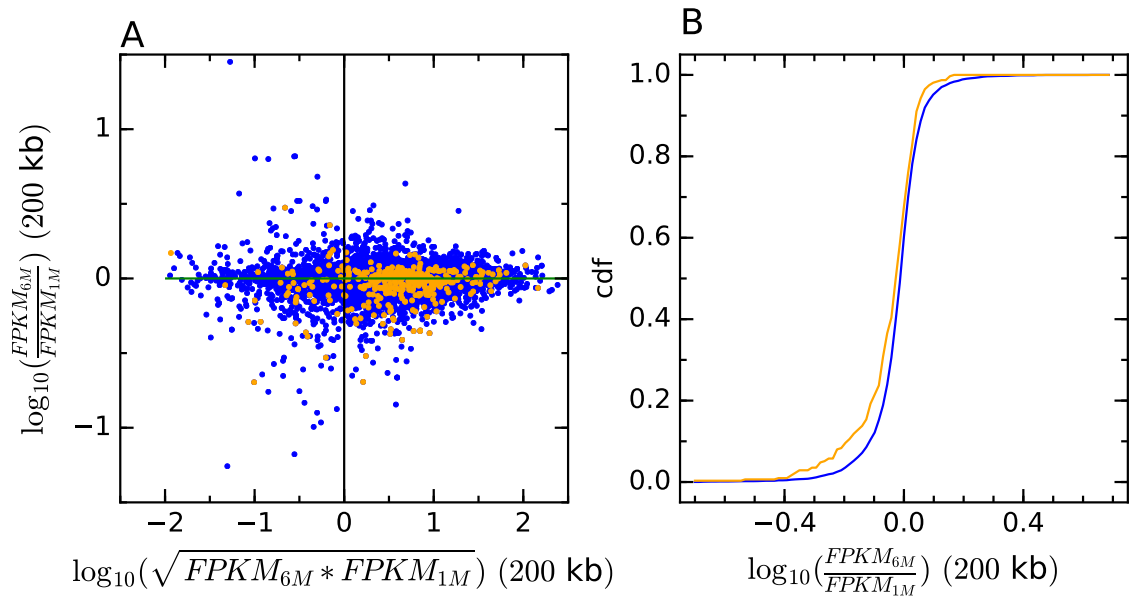

Figure 5.10: **Weakened initiation zones between 1 month and 6 months of BCR_ABL1 expression are associated with transcription repression.**(A) $\log_{10}$-ratio of FPKM values in TF1_BCRABL_6M over TF1_BCRABL_1M as a function of their $\log_{10}$ (geometric) mean. FPKM values were computed in non-overlapping 200 kb windows. Only windows expressed in both cell lines were considered (FPKM > 0.01). Windows containing at least one weakened IZ at Step 2 are in orange, the rest is in blue. (B) Cdf of the FPKM $\log_{10}$-ratios for windows with FPKM (geometric) mean > 0. Blue and orange, as in (A).

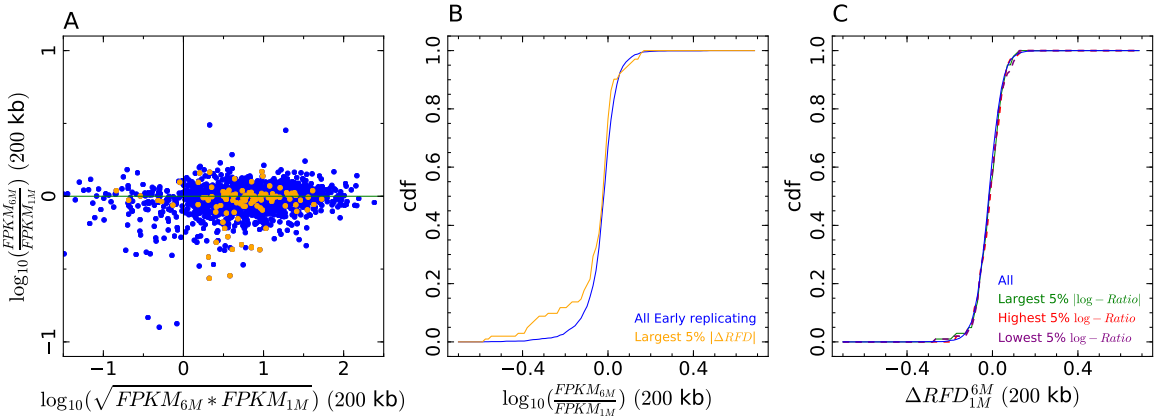

Figure 5.11: **The largest RFD changes in early replicating regions between 1 month and 6 months of BCR_ABL1 expression are associated with transcription repression, but transcription changes do not predict RFD changes.**(A,B), Same as Figure 5.10AB but for all early replicating regions (MRT < 0.33). The 5% with the largest RFD changes are in orange, the rest is in blue. (C) Cdf of ΔRFD_{6M}^{1M} between TF1_BCRABL_6M and TF1_BCRABL_1M for windows with FPKM (geometric) mean > 0. Results in green, red and purple correspond to the 5% of the regions plotted in blue with the largest absolute FPKM ratio,the largest FPKM ratio and the lowest FPKM ratio, respectively.

5.6 Conclusion

IZ efficiency changes initiated by BCR_ABL1 expression are accentuated during prolonged BCR_ABL1 expression in TF1 and furthermore in the late CML cell line K562. BCR_ABL1 therefore has a long-lasting action on IZ efficiency, in a direction that depends on each IZ but is more often repressive. Changes in IZ efficiency induced by BCR_ABL1 mostly occur in GC-poor, lowly expressed, late-replicating DNA regions and are predominantly repressive. These results suggest that BCR_ABL1 expression may generate RS independently of transcription by disturbing the replication of GC-poor, late-replicating gene deserts. This is in apparent contrast with other oncogenes such as MYC or Cyclin E which, by shortening G1 phase, induce early firing of intragenic origins normally erased by transcription during G1 phase [206]. However, late-replicating, gene-poor regions of the genome were not interrogated in the latter study.

The BCR_ABL1 kinase affects a wide range of intracellular signaling pathways [213] which might directly or indirectly modulate origin firing. The origin licensing factor Cdc6 is upregulated in a BCR_ABL1-dependent manner in primary CML and K562 cells [214], which may explain the increased activity of some origins. The DNA damage response (DDR) pathway is activated in chronic CML [215], and DDR activation is known to repress late-firing origins. Even though IZ changes were predominantly repressive, both repression and activation events were observed, and the direction of most changes was conserved during CML progression, suggesting that different classes of late IZs have opposite responses to BCR_ABL1. It is possible that the down regulation of some late IZs by BCR_ABL1 indirectly stimulates other late IZs due to increased availability of limiting origin firing factors. It has previously been argued that replication of the human genome involves a superposition of efficient initiation at "master" IZs detected in RFD profiles followed by more random, cryptic initiation between them [3, 51, 52, 56]. While we only detect silencing of master IZs, it is possible that BCR_ABL1 has a much broader effect on dispersed initiation.

CHAPTER 6

Conclusion and perspectives

We have analyzed novel RFD, gene expression and MRT datasets and have explored several approaches to compare the replication and transcription programs of twelve cancer and non-cancer cell types. A global, unbiased correlation approach revealed that the RFD (Figure 3.7) and MRT (Figure 4.6) profiles are clustered in three separate groups corresponding to lymphoid, myeloid and adherent cells. Similar results were obtained by RNA-seq (Figure 4.5), except that HeLa clustered with myeloid instead of adherent cells. Therefore, cancer-associated changes in replication do not blur their developmental origin signature, although changes in gene expression may sometimes do so. It is notable that the two LMSs are more correlated to each other by RNA-seq than by RFD, which suggests that their cell of origin may be different and that the selection for a tumour phenotype may have resulted in a stronger convergence of their transcription than their replication program. We did not detect any evidence for convergence of the replication or transcription programs of cancer cells from different developmental origins (e.g. LMS vs. BLs). In contrast, within lymphoid cells, we found evidence for BL-specific replication and transcription patterns. Furthermore, within myeloid cells, we found that expression of the BCR_ABL1 oncogene in TF1 cells, which models the establishment and early progression of CML, altered the RFD and transcription profiles of TF1 cells in a manner that increased their resemblance to K562, a late CML cell line (Figure 5.4). Overall, our global correlation analyses provide evidence for the existence of recurrent replication and transcription changes along specific tumour progression pathways. Interestingly, the RFD changes induced by BCR_ABL1 expression in early CML were not associated with large-scale MRT switches comparable to those observed between early and late CML or previously reported between leukemias and control LCLs [215]. The global correlation analyses further revealed that RFD changes between cell lines are widespread through the genome but more frequent in GC-poor regions (Figure 3.11). In contrast, RNA-seq changes do not vary uniformly with GC content (Figure B.3). These

results strengthen the hypothesis that replication changes are dissociated from transcription changes, to an extent that specifically depends on the compared cell types. More detailed investigations of the CML model reinforced and refined these conclusions. The largest RFD changes induced by 1 month of BCR_ABL1 expression in TF1 cells are concentrated in GC-poor, lowly-expressed and late replicating regions (Figure 5.2). A similar, albeit less pronounced, tendency is observed after 6 months (Figure 5.3). For example, 98% and 92% of the 1% largest RFD changes after 1 month and 6 months, respectively, of BCR_ABL1 expression, occur in the latest half of S phase (MRT>0.5). Visual examination of the TF1 RFD profiles after 0, 1 or 6 months of BCR_ABL1 expression shows that the three profiles are strikingly identical over most of the genome, consistent with their high global correlation coefficients (>0.95). This striking conservation has allowed us to manually detect and annotate BCR_ABL1 induced changes of IZ efficiency at 1027 loci and to follow their fate during early and late CML progression (Figure 5.5). The targeted IZs are more often downregulated (∼2/3) than upregulated (∼1/3), and these changes are more often enhanced than reverted over months of BCR_ABL1 expression in TF1 and in the late CML K562 (Figure 5.8). Other RFD changes are restricted to the immediate neighbourhood of affected IZs, consistent with a changed distribution of termination events due to IZ efficiency changes. Large MRT shifts are not observed. Similarly to global RFD changes, IZ efficiency changes are enriched in GC-poor, lowly-expressed and late-replicating regions and this tendency is less pronounced after 6 months than after 1 month of BCR_ABL1 expression (Figure 5.9). More specifically, we observed that IZ weakening events become more frequent in the GC-rich, early replicating, highly expressed portion of the genome after 6 months of BCR_ABL1 expression, and such changes are associated with transcription repression. In summary, BCR_ABL1 has a long-lasting action on IZ efficiency, in a direction that depends on each IZ but is more often repressive, with an initially strong preference for late-replicating gene deserts and a progressive shift to other genome compartments during prolonged expression. Previous study of HeLa and GM06990 RFD profiles revealed the existence of three types of IZs [125]. Type 1 and type 2 IZs are circumscribed on one or both sides by active genes, and fire early in S phase. Type 3 IZs, on the other hand, are not associated with active genes and fire predominantly late in S phase. The mechanisms that delimit type 3 IZs remain unclear, but the findings reported here suggest that BCR_ABL1 specifically affects non-transcriptional mechanisms that set the boundaries and/or regulate the activity of type 3 IZs, although BCR_ABL1 can also weaken type 1/2 IZs in association with transcription repression after 6 months of expression.

We observed an heterogeneity of RFD profile fluctuations along the genome in accordance to GC content level: RFD changes between cell lines accumulate in low GC content regions. We computed a new score called MCR to identify the zones where the RFD profiles are variable or stable among the 12 cell lines. MCR very precisely captured RFD profile correlation intensity fluctuations along the genome (Figure 3.19). Differential susceptibility to replication program changes appeared to be similar in the 3 cell line groups when computing

the MCR scores restricted to each group and each pair of groups (Figure 3.22). We observed that regions late (resp. early) replication in most cell lines correlated with variable (resp. stable) RFD profiles (Figure 4.8). In particular, the large (> 1.5 Mb) domains of variable replication program almost always corresponded to regions systematically replicating in late S phase (Figure 4.10). We used the derivative of the RFD profiles to delineate IZ locations as regions of maximal RFD slope. We observed a highly significant conservation of IZ between cell lines (ratio of observed over expected conservation $\gtrsim 6$ in all pairwise comparisons) allowing to recover the main features of the cell line classification (Figure 3.26). We found a higher density of IZ in early replicating and very conserved RFD profiles regions than in late and variable RFD profile regions (Figures 3.27 and 4.14). For example, the density fold change in TF1_GFP between early and late replicating regions was larger than 5. Moreover, IZ are more efficient and more conserved between cell lines in the latter than the former Figure 4.13. This suggested a cell line specific control of IZ in late replicating domains during normal and pathological differentiation. Late replication are associated to low gene density suggesting a decoupling between replication program changes and gene expression changes, at least in these regions.

An increase in DNA damage and genomic rearrangements is caused by deleterious replication-transcriptional conflicts in eukaryotes [216, 217]. Furthermore, during stem cell differentiation the replication timing changes are correlated with changes in gene activity and subnuclear position [37, 55, 156, 218–220]. To understand the reported covariation between the replication and transcription, we included gene transcription data in the analysis. Using the MCR scores, we grouped the transcribed genes in 4 classes associated respectively to extremely variable RFD profiles, moderately variable RFD profiles, moderately conserved RFD profiles and extremely conserved RFD profiles and we found that relative transcriptional changes are higher in the variable RFD profile regions than in the conserved one (Figure B.5). Previous works have demonstrated a covariation between the relative transcription changes and the replication program changes from a gene promoter perspective. Transcriptional changes were first identified and, RFD profiles [3, 171] or MRT profiles [186] changes were then questioned at these loci. Similarly, we selected differentially expressed genes between GM06990 and Raji and analyzed the change of replication initiation potential at their promoters by comparing their RFD slopes. Higher initiation potential was on average associated to higher expression level but not systematically (Figure 4.21). We confirmed this analysis by manually annotating the presence/absence of an IZ ± 100 kb of the promoter of all differentially expressed genes between GM06990 and Raji (Table 4.1). We found that, for 38% of differentially expressed genes, an IZ was present in proximity to their promoter in the two cell lines and that, for 25% of them, an IZ was observed in the cell line of greater promoter activity only, as expected for a direct link between promoter activity and replication initiation. These results suggested that the presence of a replicative change at promoters of changing activity is in fact limited. This confirmed previous studies where it was shown that the coupling between the replication and transcription changes is not systematic for all genes.

Originally, we studied the reciprocal association at replication IZ (IZ point of view). First, we observed that the regions with a large initiation potential increase between the GM06990 and Raji are associated to increased transcription (Figure 4.23). Focusing specifically on detected IZ, we selected 452 highly efficient GM06990 IZs presenting a significant weakening in Raji. We observed that not all the IZ efficiency changes are associated to a large FPKM change, as (i) 119 IZ (26.3%) are not associated to transcription in either cell lines (closest expressed protein coding gene at least 50kb away from the IZ), (ii) 46 (10.2%) were associated to gene activation in Raji and (iii) 171 (37.8%) were associated to not differential expressed genes. Thus, we detected a global correlation between transcriptional changes and replication program changes as in the other studies [3, 171, 186] and we quantified a significant extent of decoupling when considering individual genes and individual IZ, underlying that other factors remain to be identified.

OK-seq method has will undoubtedly be generalised to other systems. In the hypothesis that Okazaki fragment (OF) has a constant life time all along the genome, it is expected that Ok-seq coverage is homogeneous (up to mappability or other biases). We observed that it is in fact heterogeneous, with coverage peaks sometimes corresponding to high efficiency IZ (Figure 3.3 around position 5 Mb). A possible explanation of this observation is that OF biochemistry could depend on genomic or epigenomic context. This prompt our experimental collaborators at IBENS, Paris to extend the OK-seq protocol to paired-end sequencing. Indeed, these new Ok-seq experiments should in principle enable us to reconstruct the OF size statistics at every locus as the distribution of paired-end reads outer length (Figure D.1). Paired-end Ok-seq experiments performed on previously profiled cell lines (GM06990, Raji, BL79 and IARC385) already demonstrated that this now approach is highly compatible with the original protocol (Figure D.5) and that at least for 3 cell lines (Raji, BL79, IARC385) the observed genome-wide estimated OF length distribution is compatible with the expected OF full size of $\sim$ 150 bp (Figures D.2, D.3 and D.4). This new data demonstrate the potential of Ok-seq for genomic analyses of the dynamics of OF processing. OK-seq experiment can also be designed to ask specific question about the implication of oncogenes on RS. The Myc expression in hematopoietic cells can induce tumorigenesis. Myc over-expression is believed to be one of the primary events in the malignant transformation of Burkitt's lymphoma. Overexpression of Myc increases cell size and energy production, induces proliferation and genomic instability, and favors apoptosis. Moreover, Myc over expression can cause RS from the earliest tumorigenesis. New data constructed to question the role of C-myc overexpression on RS are new available (Hyrien team, IBENS, Paris). They are based on a B lymphocyte cell line (P493) where EBNA and Myc where transfected under the control of inducible promoters (Figure D.6). EBNA is required to assure cell proliferation in the absence of Myc expression. The P493 cell line was obtained from a B-cell line by allowing expression of one or both of the EBNA and Myc (Table D.1) and each culture was profiled by Ok-seq. The RFD profiles of P493 cultures with the expression of EBNA and Myc or with EBNA expression alone are very correlated to each other but are less correlated to the RFD profiles with Myc expression

alone. Hence, EBNA expression appears to be a stronger factor influencing RFD profiles than Myc (Figure D.7). This shows that OK-seq profiling has the potential to inform us on the effect of Myc on the replication program, even though in the current system there will be a challenge to factor out the effect of EBNA expression.

To conclude, in Chapter 4 we have identified 2 classes of large genomic zones with a robust replication timing across cell lines that do not behave in the same way. The first one corresponds to the large domains of early replication associated with a conserved RFD profile with a high density of replication initiation zones. The second class corresponds to robustly late replication domains with a variable RFD profile between cell line and a low density of initiation zones. Along the latter class, we visualized late replicating cell line specific initiation zones. Experiments demonstrated that the activity of replication origins depends on the chromatin environment [106]. Late replicating regions are associated with a low density of transcribed genes and there is a strong correlation between chromatin modifications associated with active transcription and early replication. So, we can hypothesize that these late initiation zones are associated with replication specific epigenetics marks, allowing the identification of the specific epigenetics mark related to replication initiation independently of transcription.

Recently high-resolution Hi-C interaction maps [221] have revealed that the human genome is organized into distinct units, the so-called *Topologically Associating Domains* (TADs), where genomic interactions are strong within a domain and depleted between domains [222]. TADs were suggested to be a stable property of the human genome as they appeared to be conserved between different cell lines [222]. The comparative analysis of replication timing data and Hi-C correlation matrix in human revealed that early and late replicating loci occur in separated compartments of open and closed chromatin, respectively [38, 221]. Moreover, the boundaries of the cell type specific replication timing domains often coincide with the boundaries of insulated compartments of chromatin interaction [223] and recent study showed that the TAD borders are enriched in initiations zones [3]. These results question whether the cell line specific initiation zones observed in the late replicating regions are associated with cell line specific TAD borders.

APPENDIX A

Supplementary figures for Chapter III

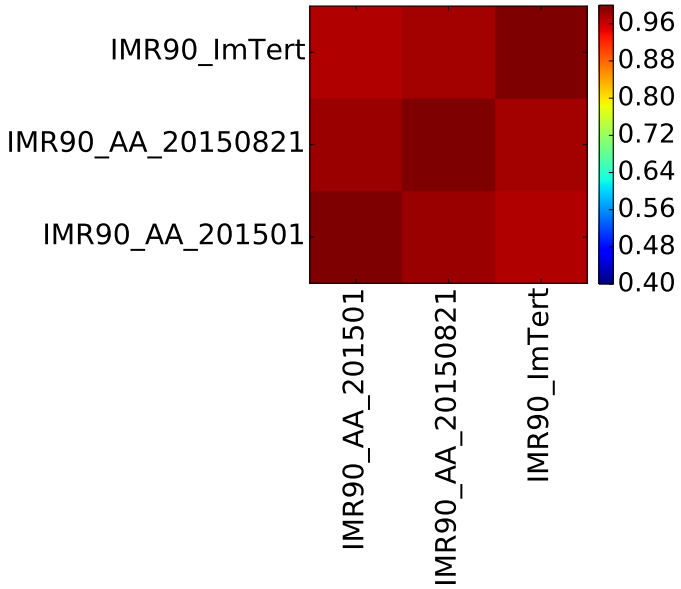

Figure A.1: **RFD profiles correlation matrix between 2 technical replicates of IMR90 and 1 replicate of hTERT immoratilized IMR90 cell lines at 10 kb.**

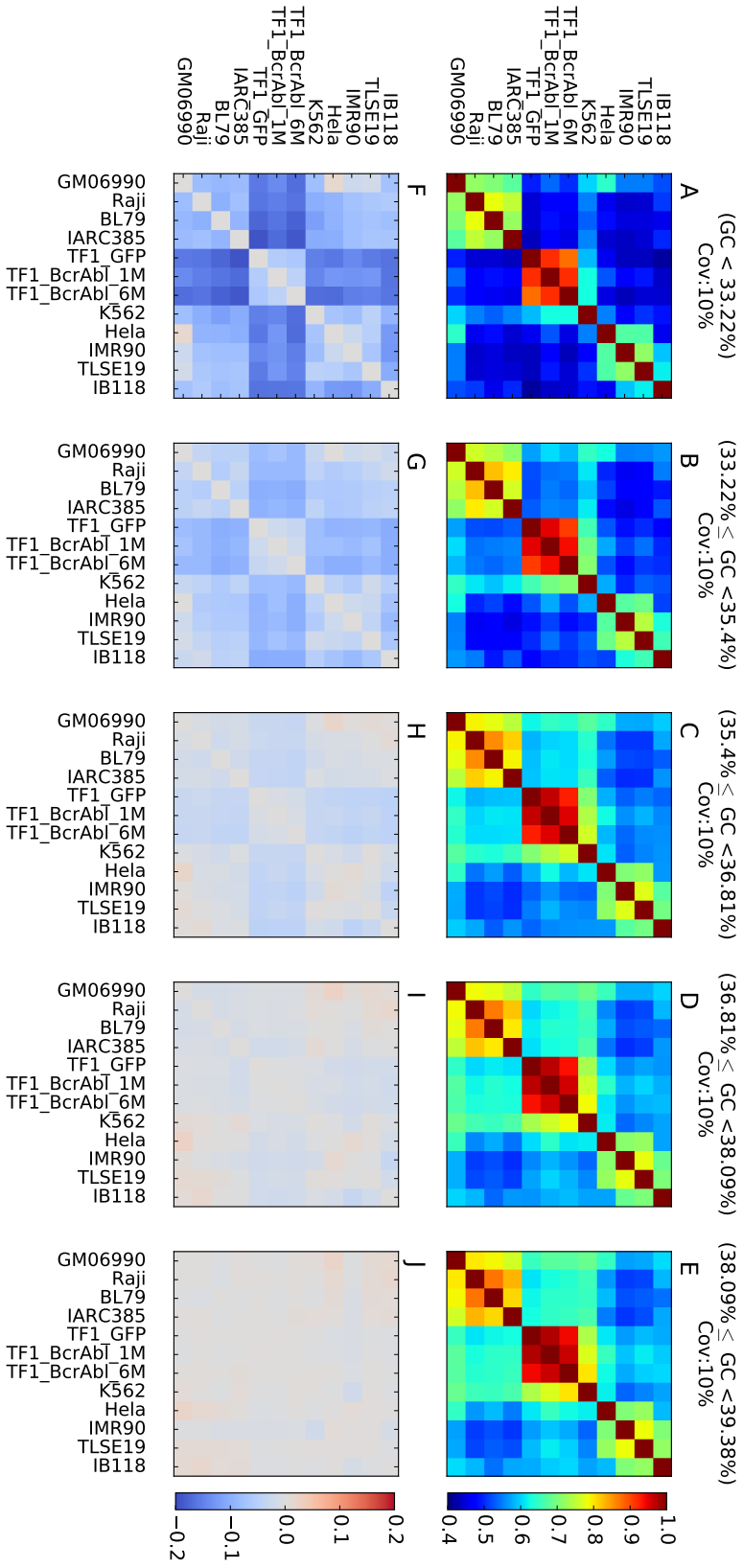

Figure A.2: **Same as Figure 3.11. When 10 kb windows were grouped within GC content decile.**

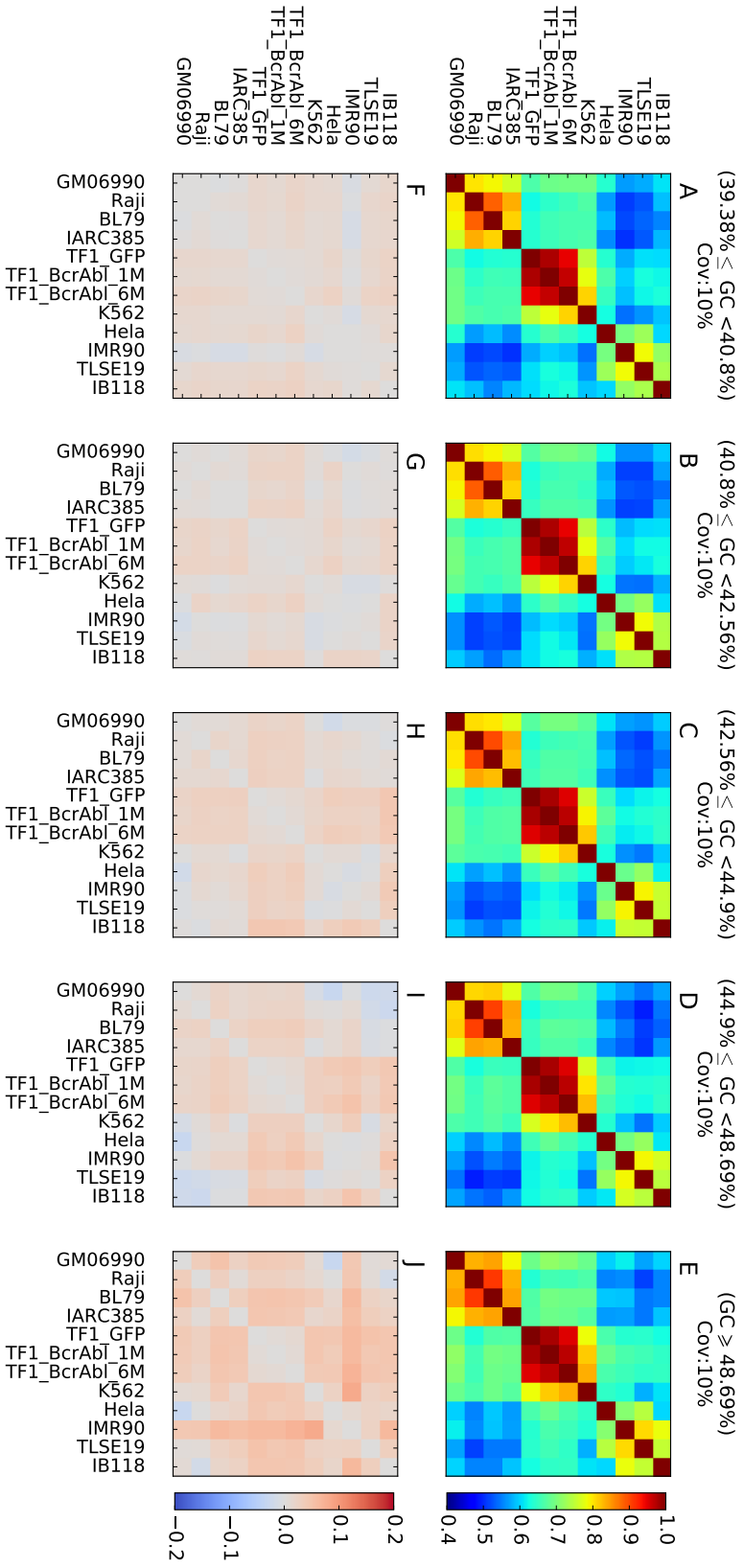

Figure A.3: **Same as Figure 3.11. When 10 kb windows were grouped within GC content decile. Last 5 deciles are presented.**

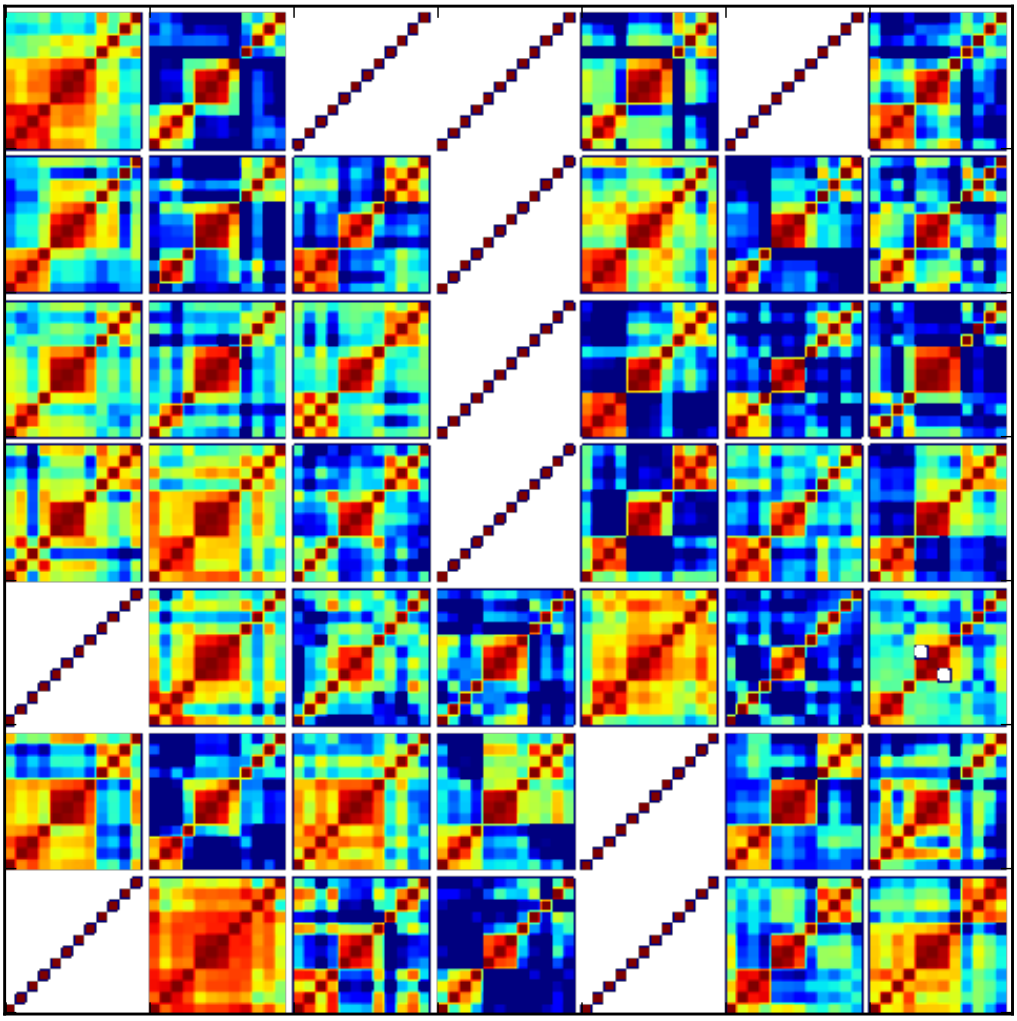

Figure A.4: **Detection of Stable and Variable regions.** RFD correlation matrices of the 5Mb non overlapping window of chromosome 1. Each square represent a 5 Mb RFD correlation matrix among the 12 cell lines at 10 kb. Pearson correlation coefficient values are color-coded from blue (0.4) to red (1) as in Figure 3.7. Order is upward then rightward from bottom left corner.

Figure A.5: **Overview of human chromosomes and their MCR levels**. We map the chromosomes between 1 and 11 using the MCR at 100 kb. Colors code correspond to the 4 MCR levelas in Figure 3.18.

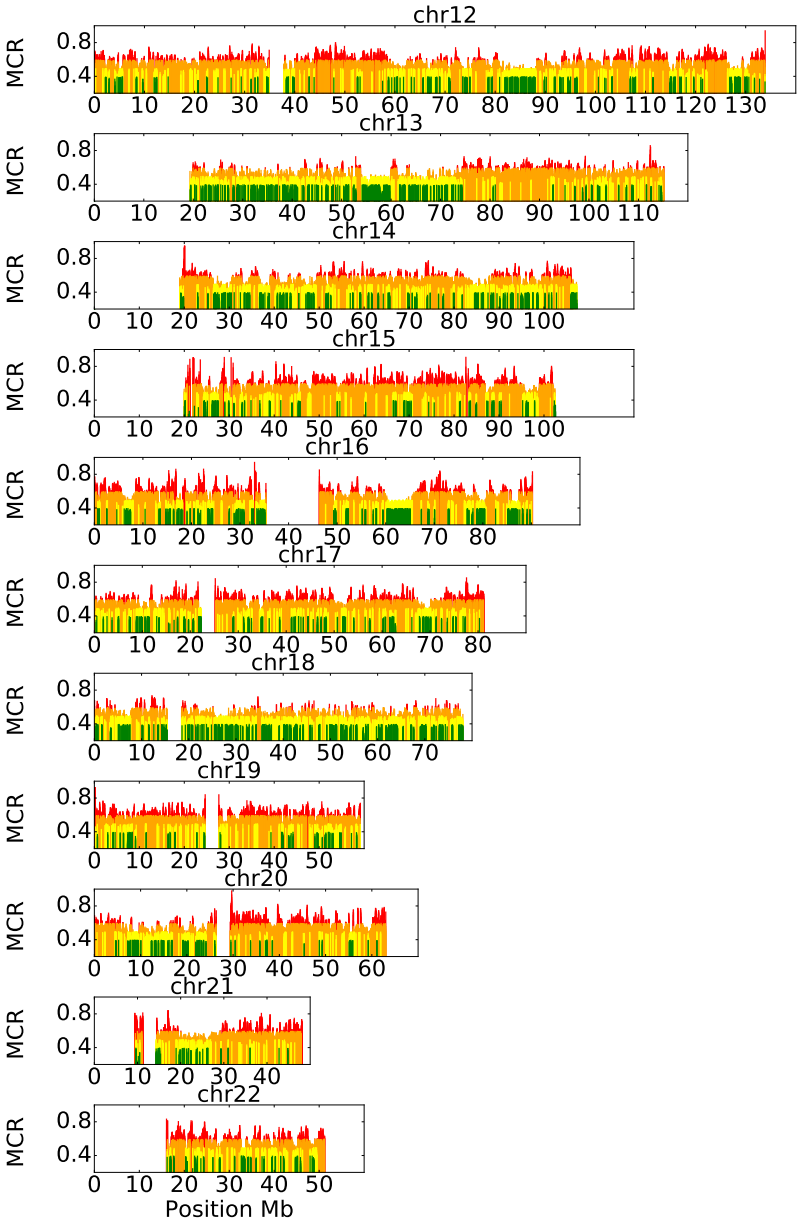

Figure A.6: **Overview of human chromosomes and their MCR levels**. We map the chromosomes between 12 and 22 using the MCR at 100 kb. Colors code correspond to the 4 MCR levelas in Figure 3.18.

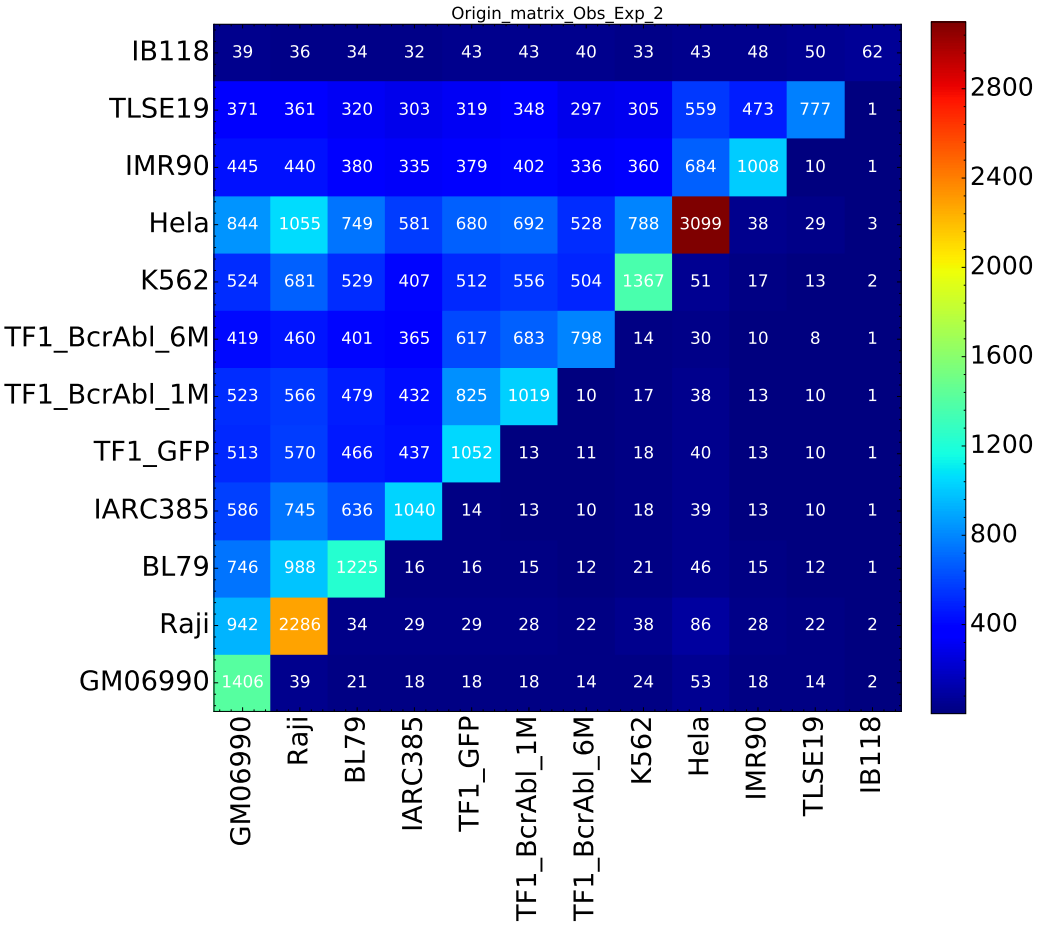

Figure A.7: **Number of common IZ with slope>2%RFD per kb at 30 kb resolution for all pairs of cell lines.** The upper triangular part of the matrix represents the observed count of common origin with slope>2%RFD per kb between pair of cell lines in 30 kb windows. The lower triangular part of the matrix represents the expected count of common IZ computed >2%RFD per kb in each cell lines. Main diagonal is the number of detected IZ of slope>2%RFD per kb in each cell lines.

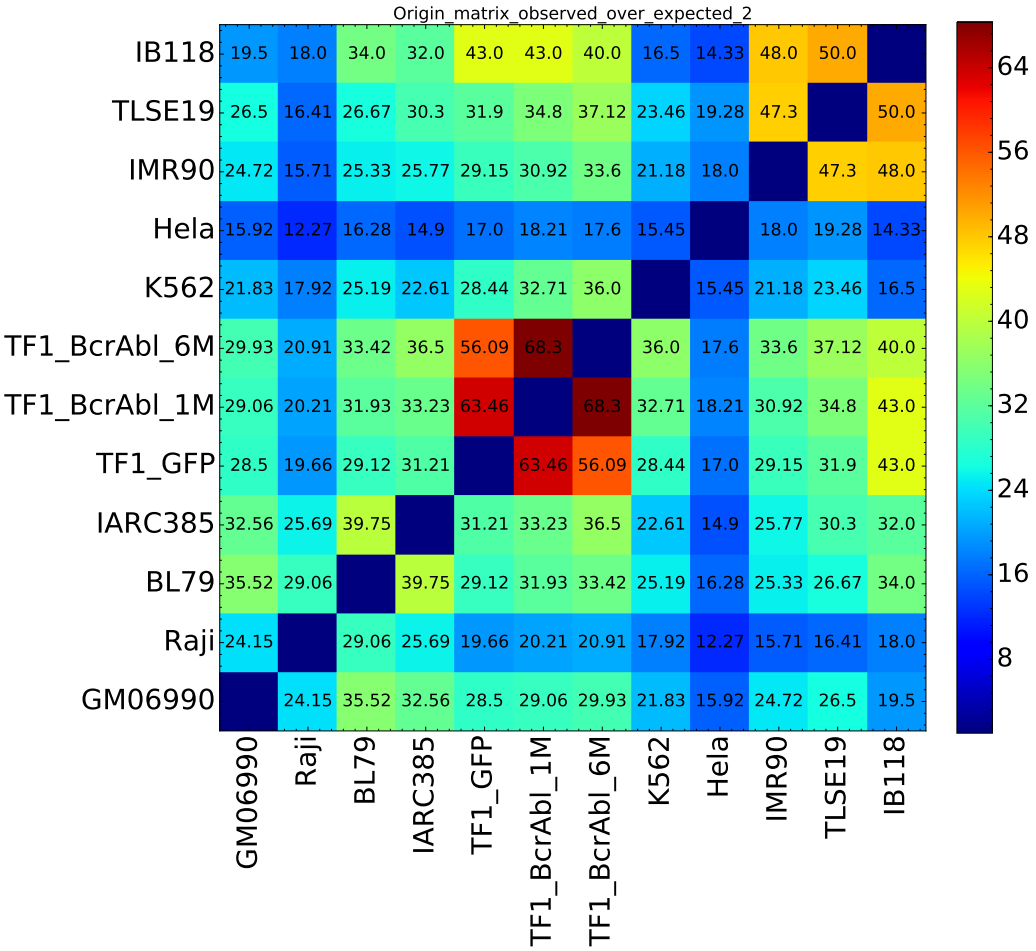

Figure A.8: **Matrix of observed/expected numbers of common for each cell line pair for a threshold of 2%RFD per kb**. Each square of this matrix represents the normalized count of common origin between pair of cell lines. Normalization was done by dividing the observed count of common IZ (upper part of the matrix in Figure A.7) by the expected one (lower part of the matrix in Figure A.7).

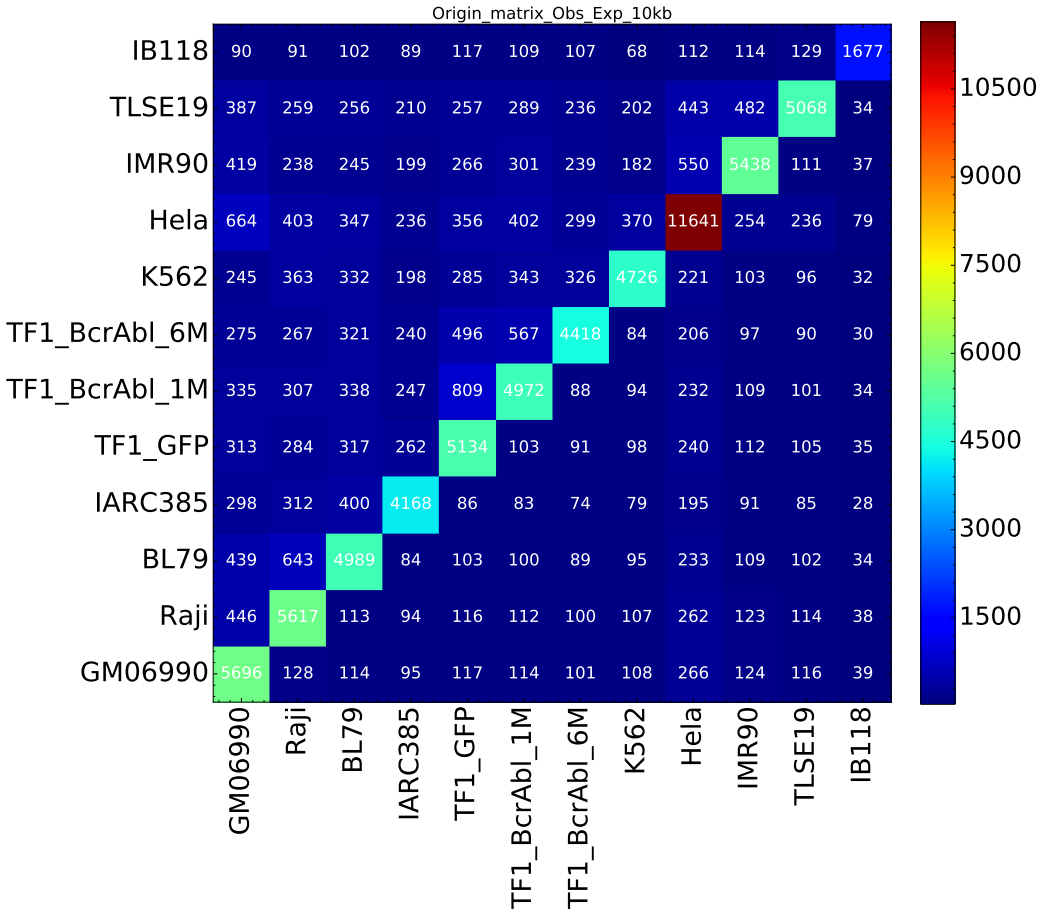

Figure A.9: **Number of common IZ with slope>1%RFD per kb at 10 kb resolution for all pairs of cell lines.** The upper triangular part of the matrix represents the observed count of common origin with slope>1%RFD per kb between pair of cell lines in 10 kb windows. The lower triangular part of the matrix represents the expected count of common IZ computed >1%RFD per kb in each cell lines. Main diagonal is the number of detected IZ of slope>1%RFD per kb in each cell lines.

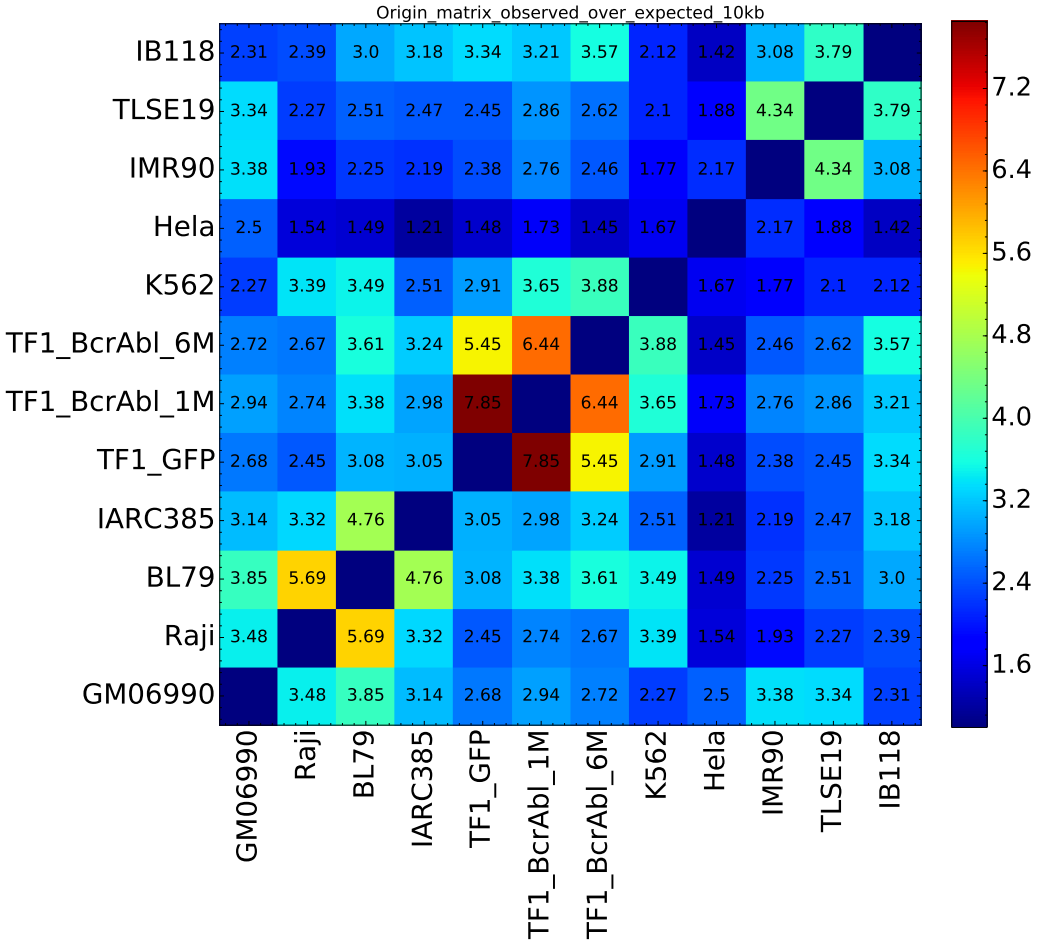

Figure A.10: **Matrix of observed/expected numbers of common for each cell line pair for a threshold of 1%RFD per kb.** Each square of this matrix represents the normalized count of common origin between pair of cell lines. Normalization was done by dividing the observed count of common IZ (upper part of the matrix in Figure A.9) by the expected one (lower part of the matrix in Figure A.9).

APPENDIX B

Supplementary figures for Chapter IV

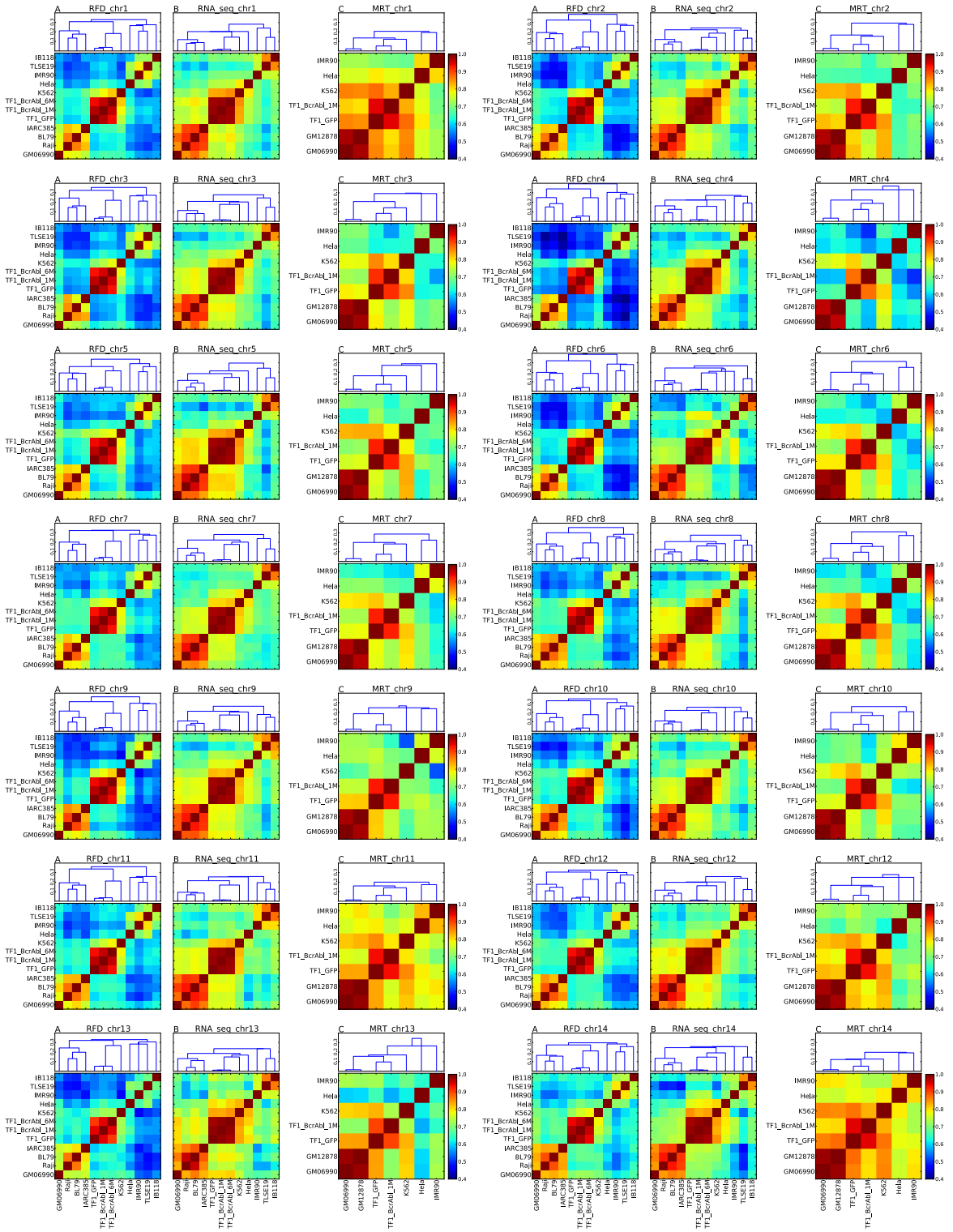

Figure B.1: **Chromosomes 1 to 14 are shown, see Figure B.2.**

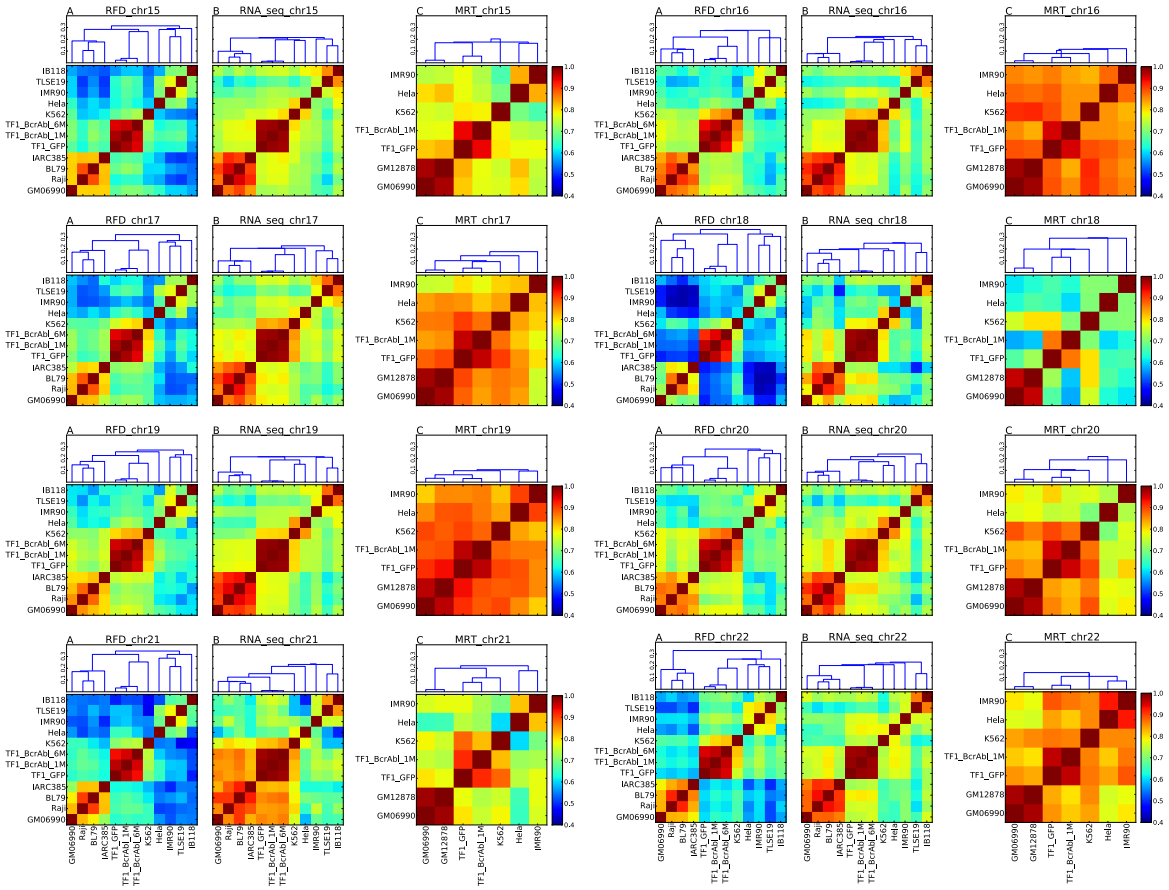

Figure B.2: **Cell line classification based on correlations between replication and gene expression profiles for each chromosome.** Correlation matrices between RFD profiles (C_{RFD} ; A), RNA-seq (C_{FPKM} ; B) and MRT profiles (C_{MRT}; C); Pearson correlation coefficient values are colour-coded from blue (0.4) to red (1) using the colour bar on the right. A corresponding dendrogram representation of the hierarchical classification of cell lines is shown on top of each correlation matrix; ordinate is the correlation distance. Chromosomes 15 to 22 are shown.

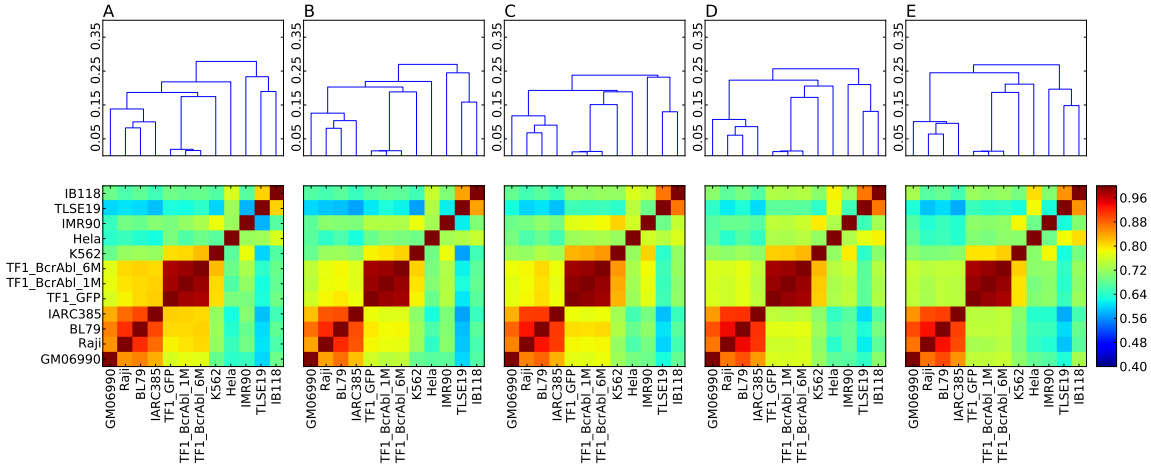

Figure B.3: **Transcription changes are not concentrated in GC poor regions.**(A-E) Correlation matrix of RNA-seq profiles depending on the GC content; 10 kb windows were grouped in GC-content categories following the 5 isochores classification of the human genome in light isochores L1 ($GC \leq 37$; ; A) and L2 ($37 \leq GC < 41$; B), and heavy isochores H1 ($41 \leq GC < 46$; C), H2 ($46 \leq GC < 53$; D) and H3 ($GC \geq 37$; $C_{RFD}^{H_3}$; E); Pearson correlation coefficient values are colour-coded from blue (0.4) to red (1) using the colour bar on the right.

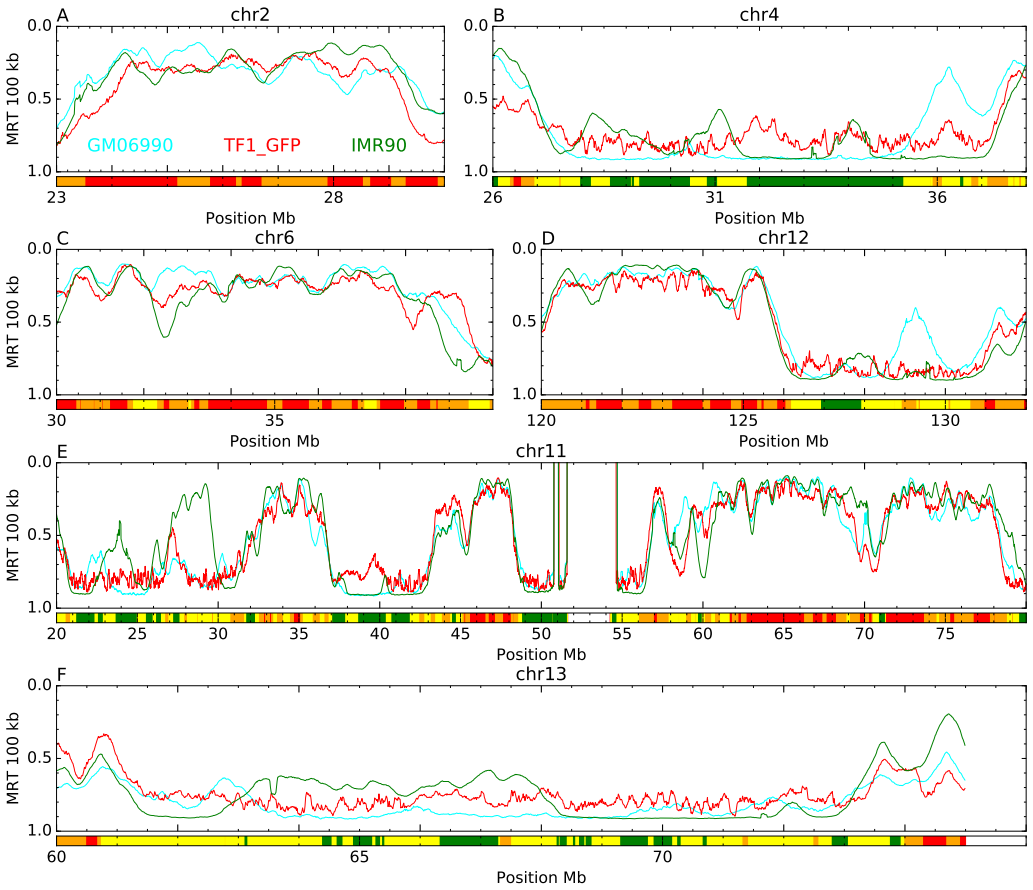

Figure B.4: **Association between MRT and MCR at 500 kb** Large early (A,C), late (B,F) and both (D,E) CTR are associated with their MCR at 500 kb. The GM06990 (Cyan) correspond to Lymphoid cell lines, the TF1_GFP (red) correspond to Myeloid cell lines and IMR90 (green) correspond to connective tissue. The color code of the MCR: green (MCR<0.42, 25 % of genome), yellow ($0.42 \leq MCR < 0.52$, 25% of genome) for variable RFD profiles; orange ($0.52 \leq MCR < 0.57$, 25% of genome),red (MCR≥0.57, 25 % of genome) for conserved RFD profiles.

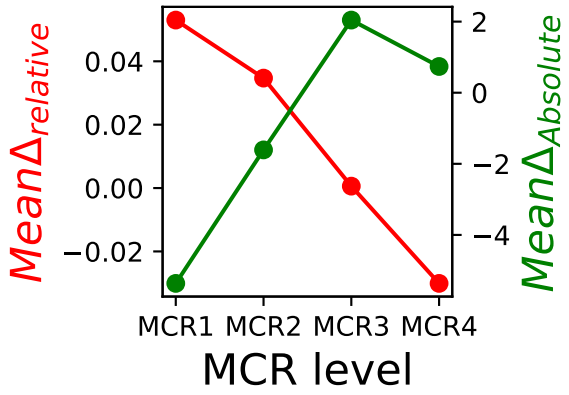

Figure B.5: **Coupling between Replication and relative transcription change**. Average of the difference matrices for the both (green) absolute (Figure 4.17)) and (red) relative (Figure 4.19) expression changes.

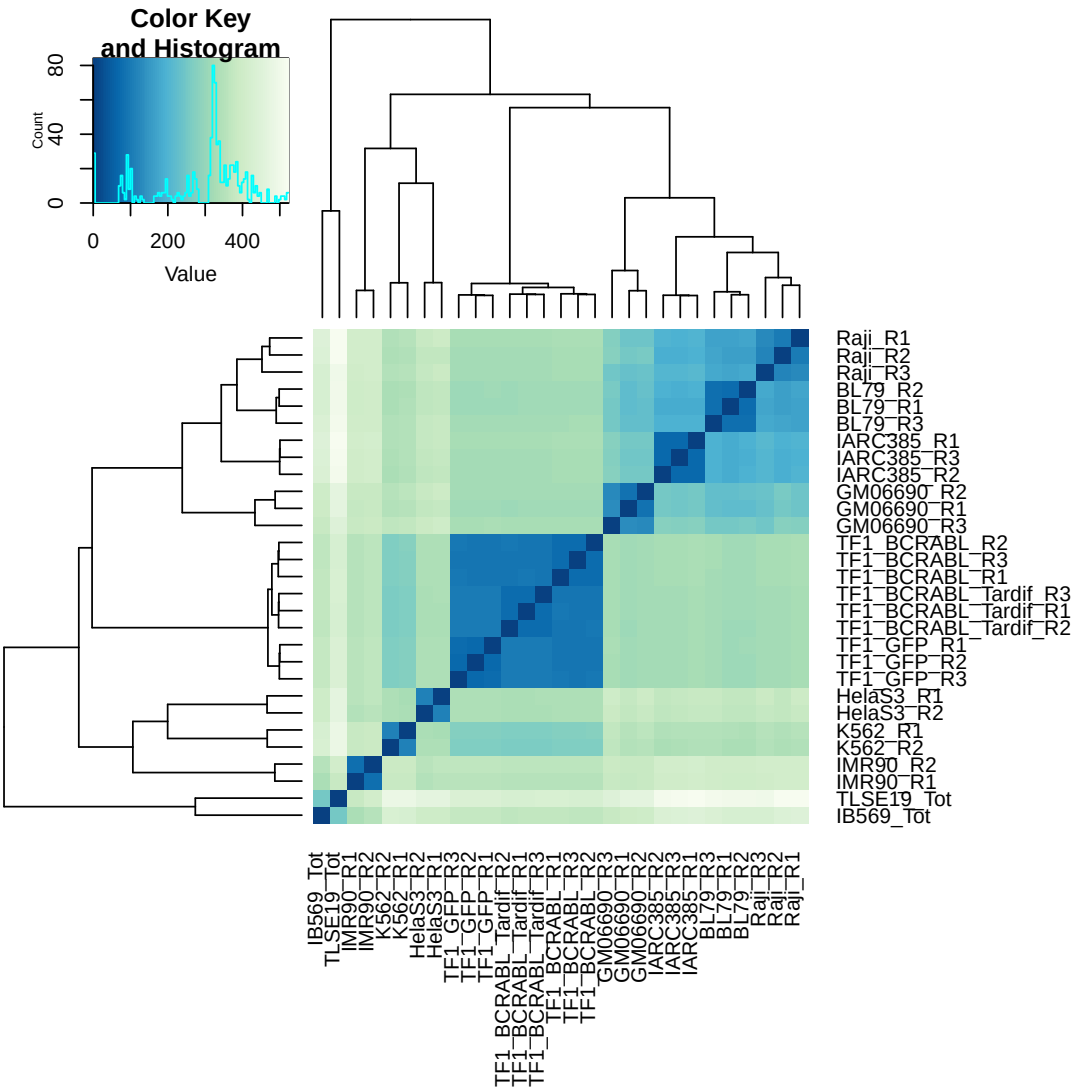

Figure B.6: **Classification of RNA-seq samples based on the numbers of the gene expression change as computed by DESeq2.**

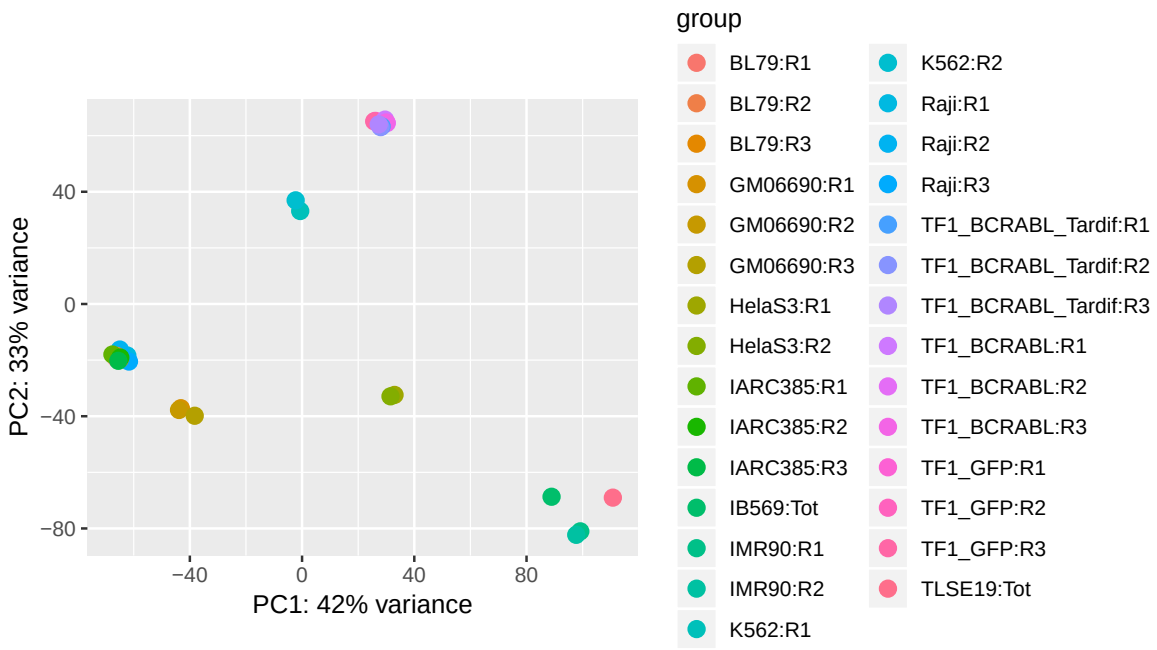

Figure B.7: **Principal component of the gene expression profiles for all RNA-seq sqmples as computed by DESeq2.**

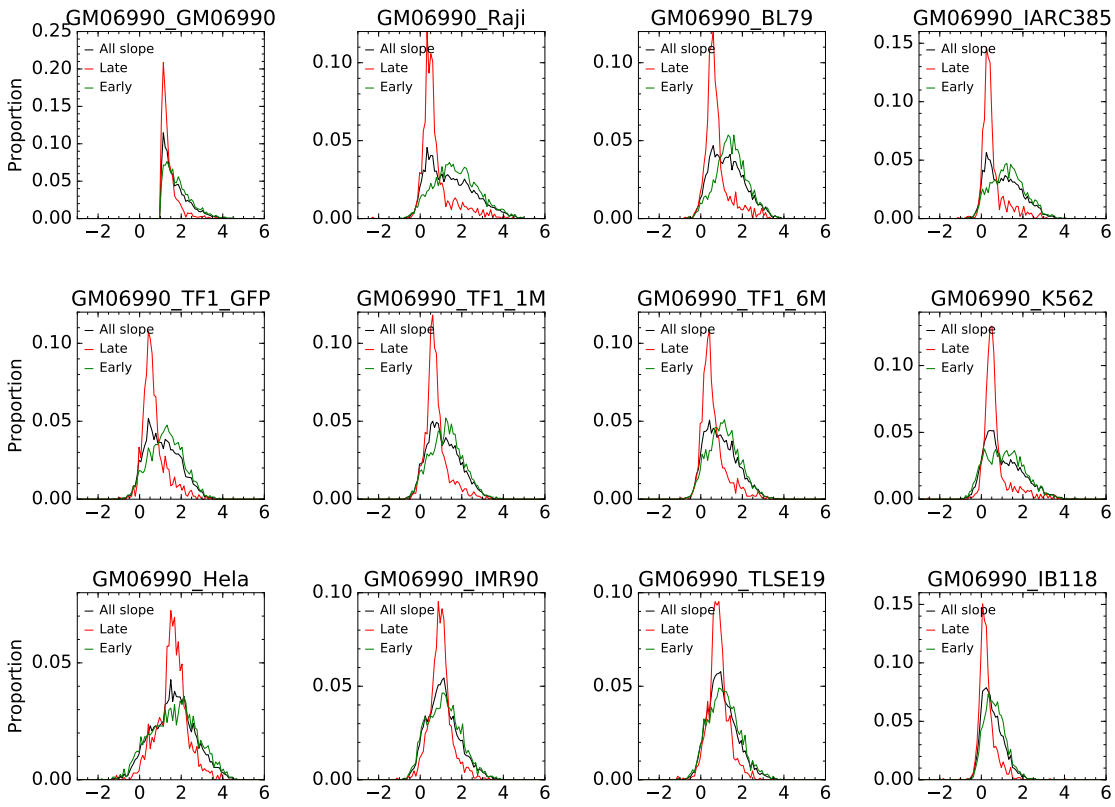

Figure B.8: **Pdf of RFD slopes in each cell line at the location of IZ selected in GM06990.** Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in GM06990. (Black) complete genome. (Red) loci where $\text{MRT}_{GM06990}$ >0.7. (Green) loci where $\text{MRT}_{GM06990}$ <0.3.

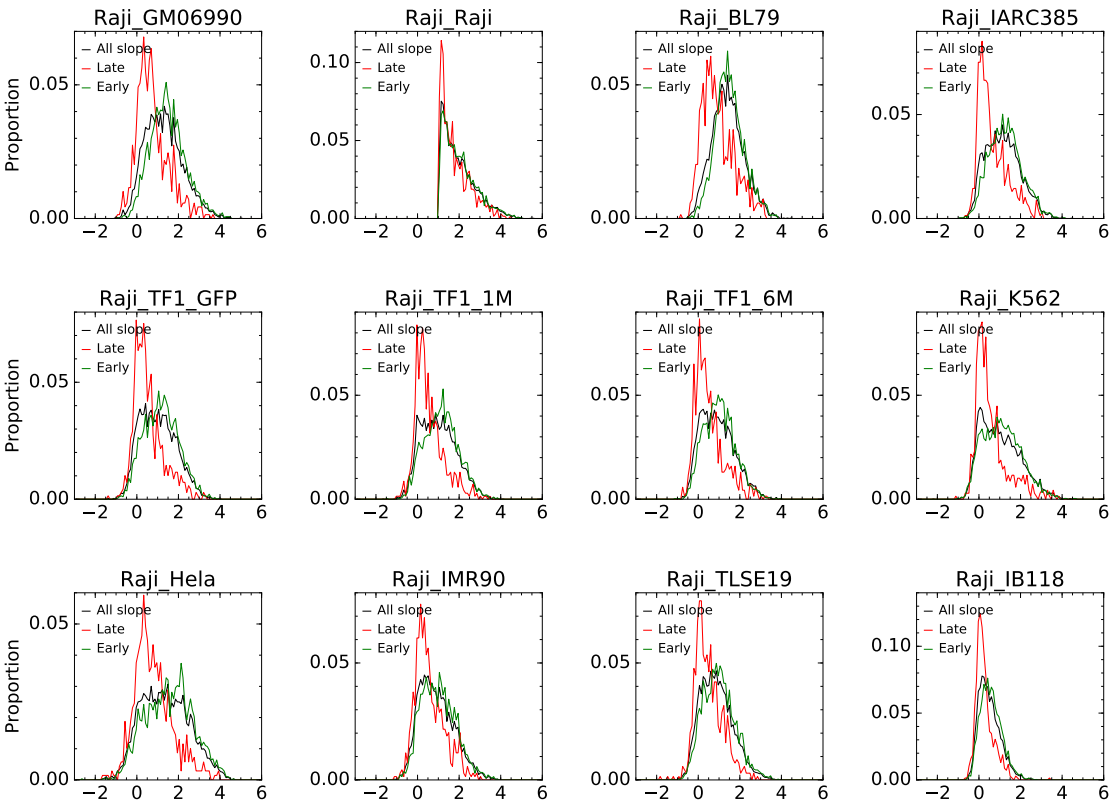

Figure B.9: **Pdf of RFD slopes in each cell line at the location of IZ selected in Raji.** Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in Raji. (Black) complete genome. (Red) loci where $MRT_{GM06990}$ >0.7. (Green) loci where $MRT_{GM06990}$ <0.3.

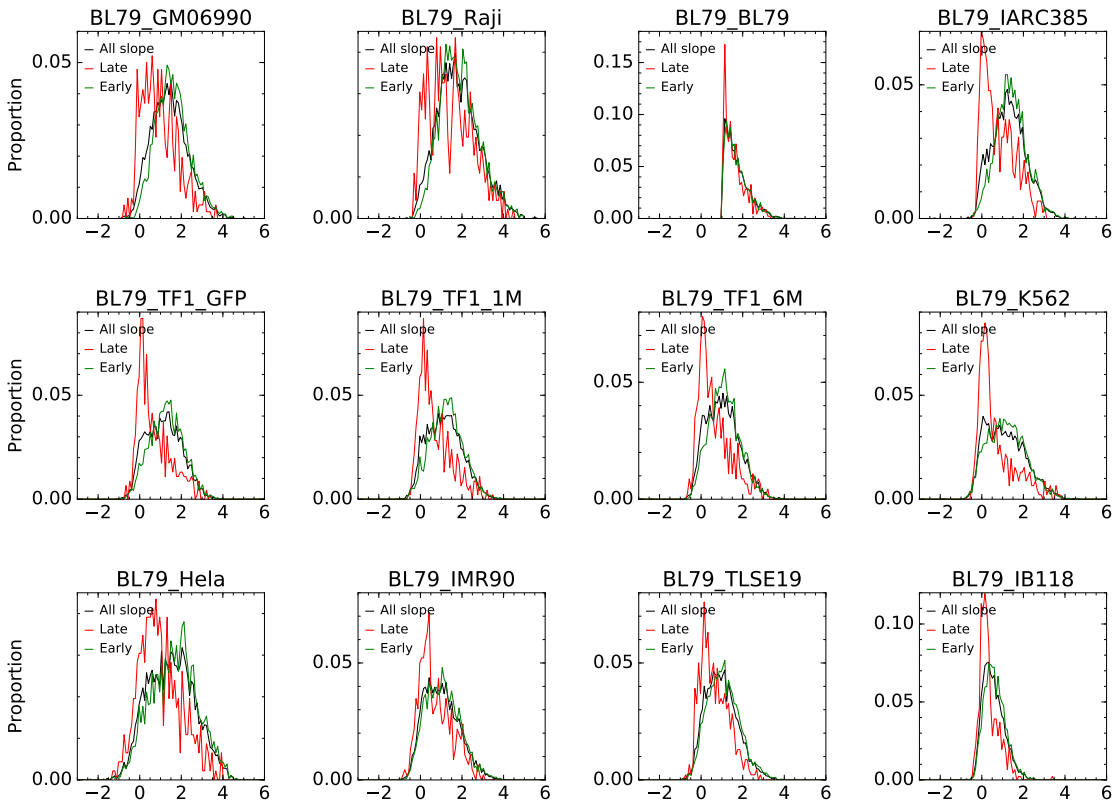

Figure B.10: **Pdf of RFD slopes in each cell line at the location of IZ selected in BL79**. Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in BL79. (Black) complete genome. (Red) loci where $\text{MRT}_{GM06990}$ >0.7. (Green) loci where $\text{MRT}_{GM06990}$ <0.3.

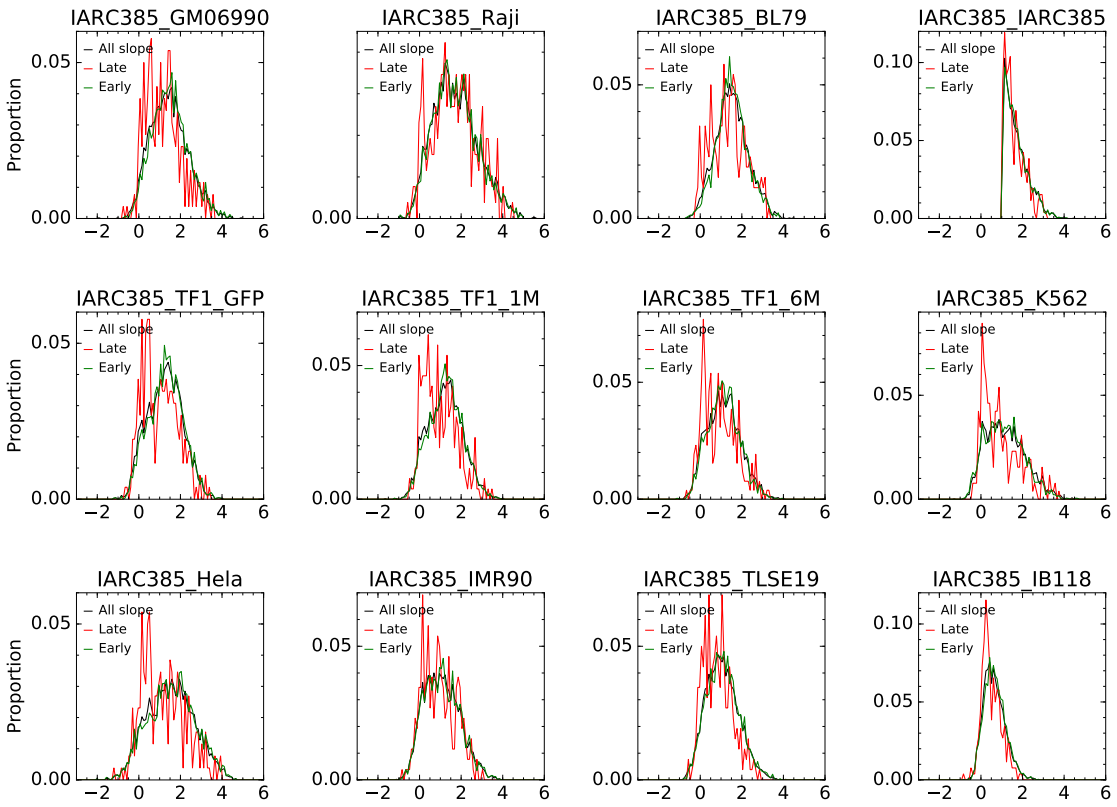

Figure B.11: **Pdf of RFD slopes in each cell line at the location of IZ selected in IARC385**. Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in IARC385. (Black) complete genome. (Red) loci where $MRT_{GM06990}$ >0.7. (Green) loci where $MRT_{GM06990}$ <0.3.

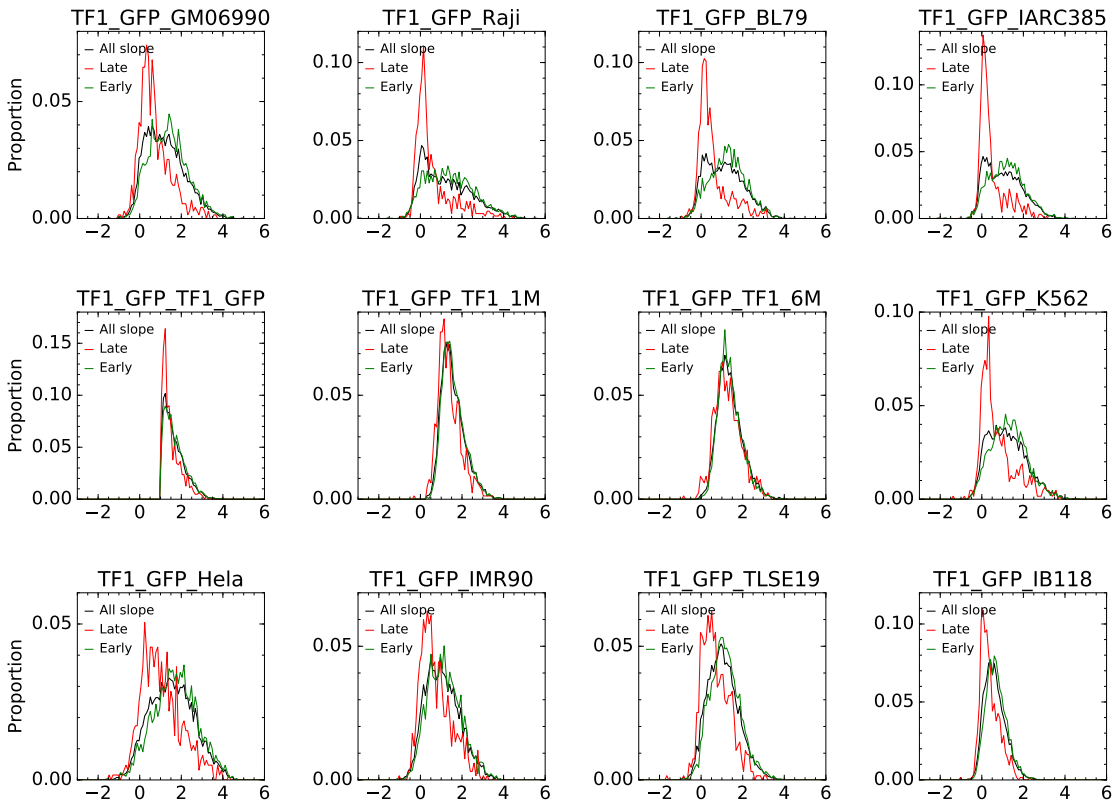

Figure B.12: **Pdf of RFD slopes in each cell line at the location of IZ selected in TF1_GFP.** Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in TF1_GFP. (Black) complete genome. (Red) loci where MRT_{TF1_GFP} >0.7. (Green) loci where MRT_{TF1_GFP} <0.3.

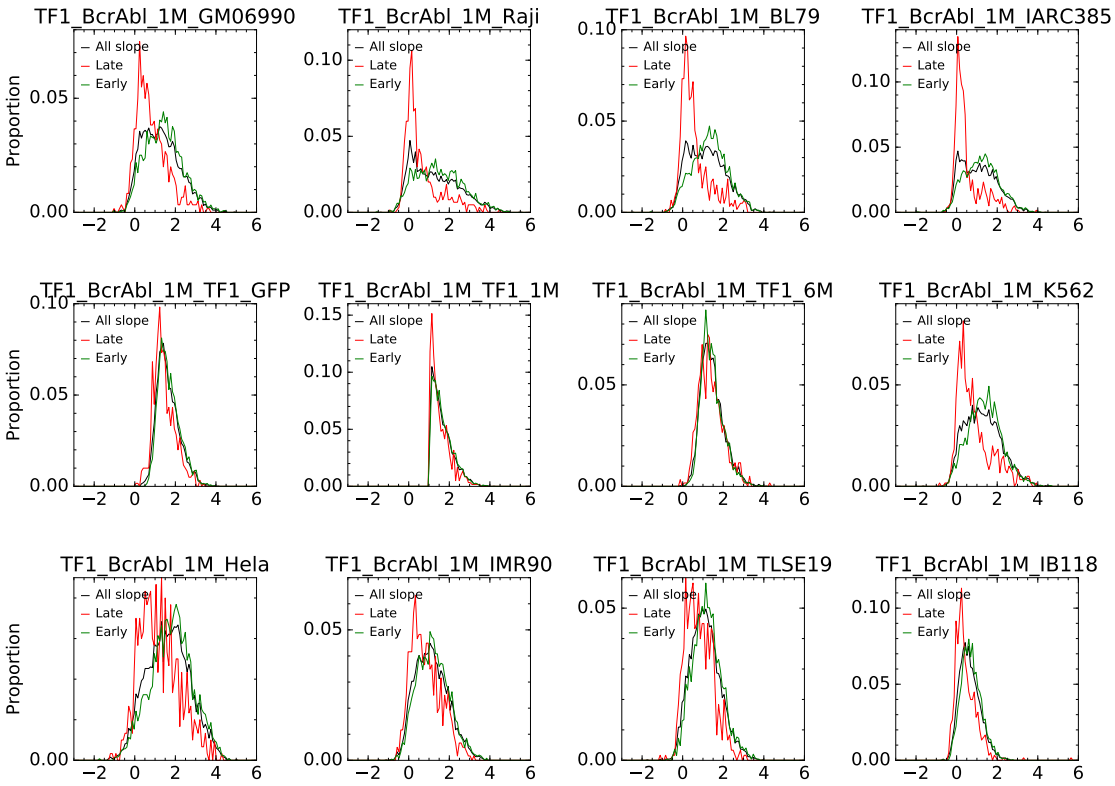

Figure B.13: **Pdf of RFD slopes in each cell line at the location of IZ selected in TF1_BcrAbl_1M.** Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in TF1_BcrAbl_1M. (Black) complete genome. (Red) loci where MRT_{TF1_GFP} >0.7. (Green) loci where MRT_{TF1_GFP} <0.3.

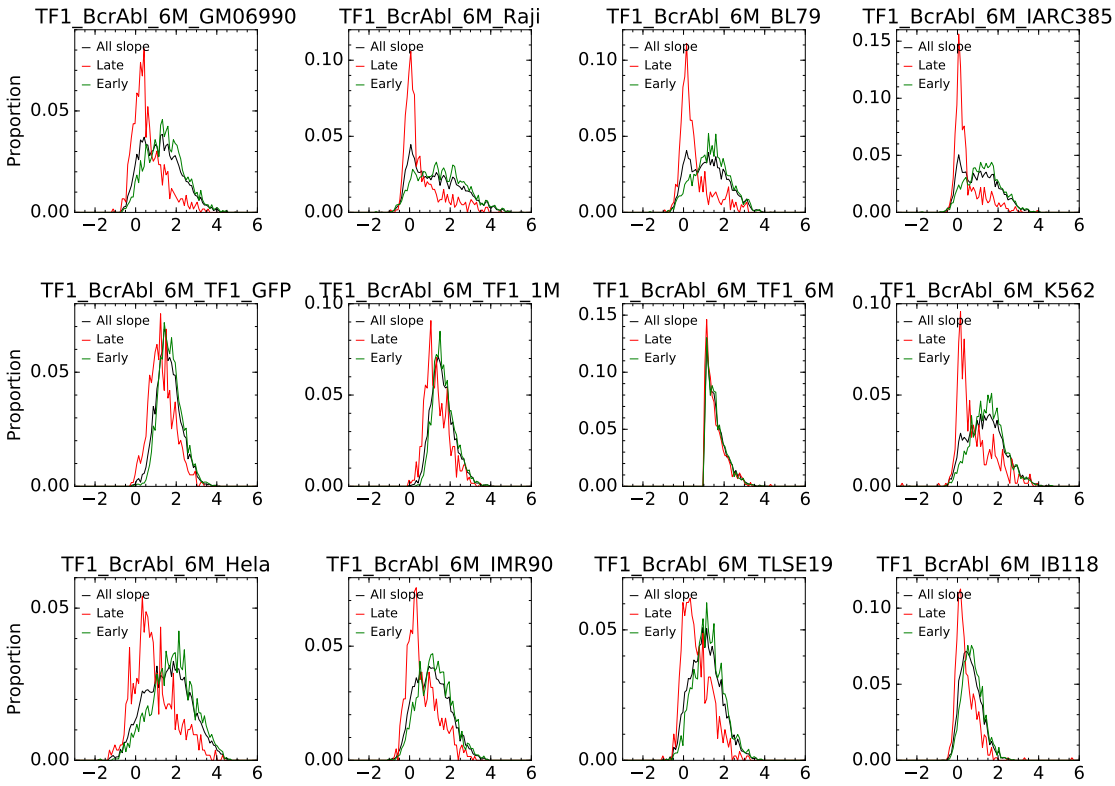

Figure B.14: **Pdf of RFD slopes in each cell line at the location of IZ selected in TF1_BcrAbl_6M.** Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in TF1_BcrAbl_6M. (Black) complete genome. (Red) loci where MRT_{TF1_GFP} >0.7. (Green) loci where MRT_{TF1_GFP} <0.3.

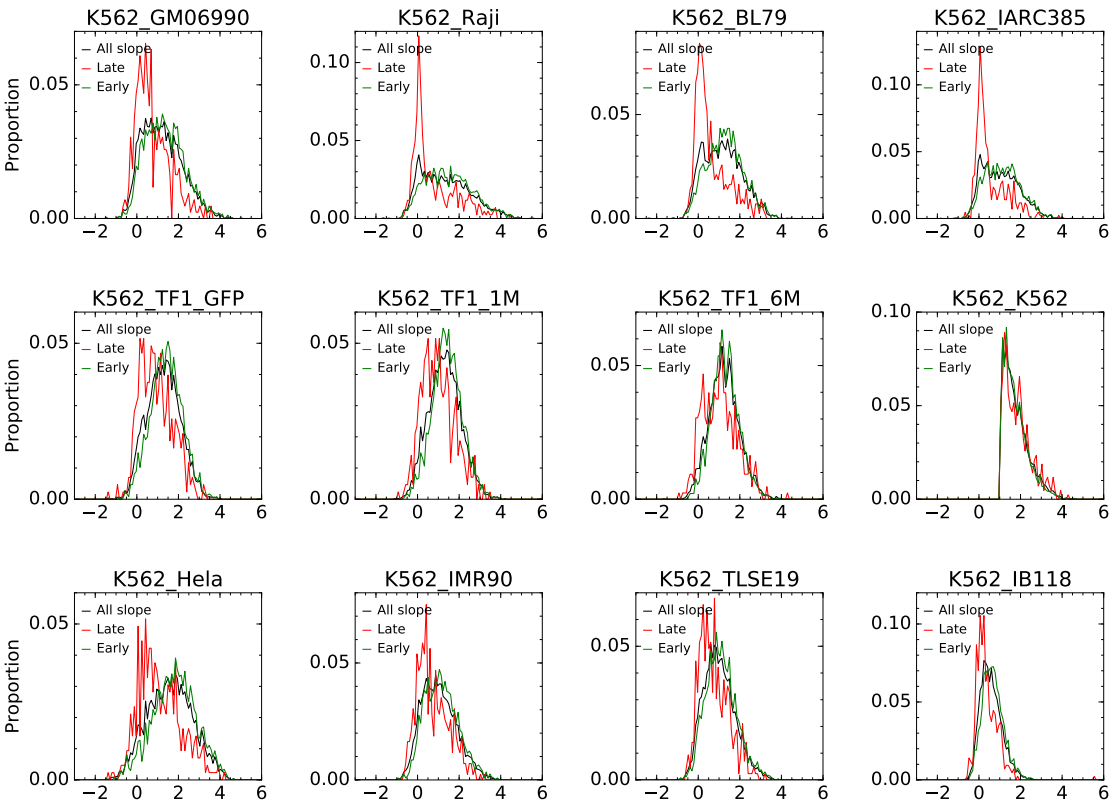

Figure B.15: **Pdf of RFD slopes in each cell line at the location of IZ selected in K562**. Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in K562. (Black) complete genome. (Red) loci where MRT_{TF1_GFP} >0.7. (Green) loci where MRT_{TF1_GFP} <0.3.

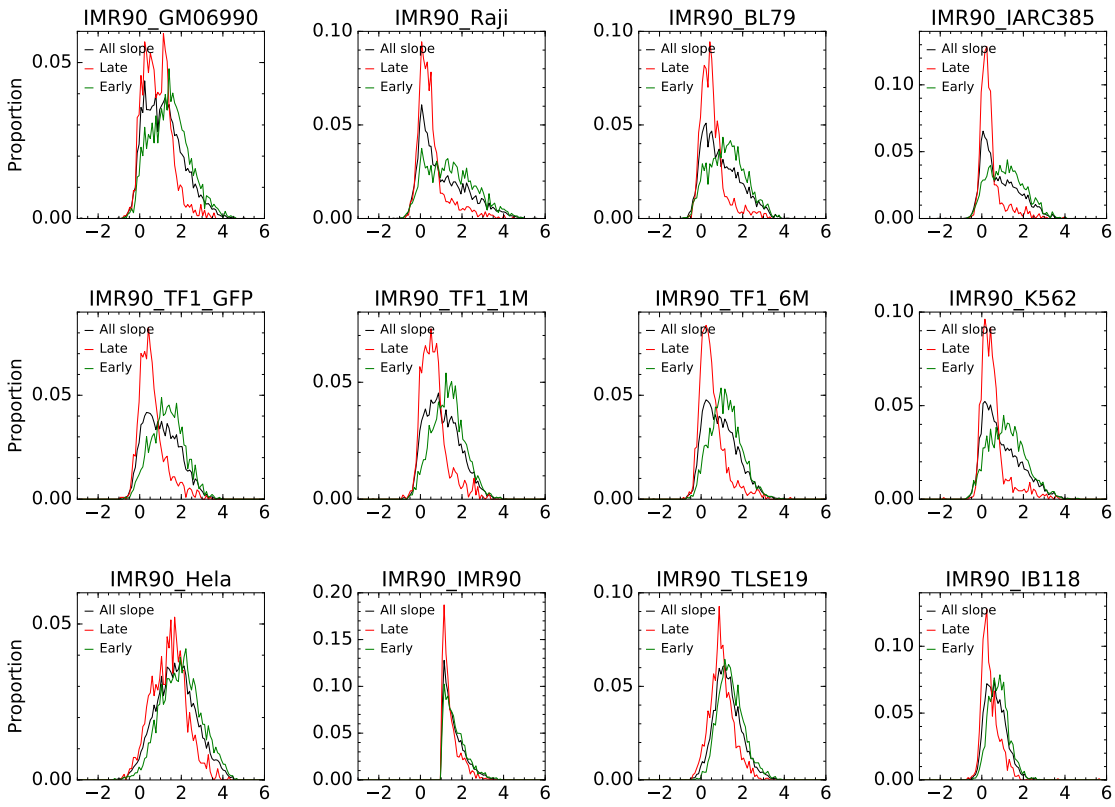

Figure B.16: **Pdf of RFD slopes in each cell line at the location of IZ selected in IMR90**. Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in IMR90. (Black) complete genome. (Red) loci where MRT_{IMR90} >0.7. (Green) loci where MRT_{IMR90} <0.3.

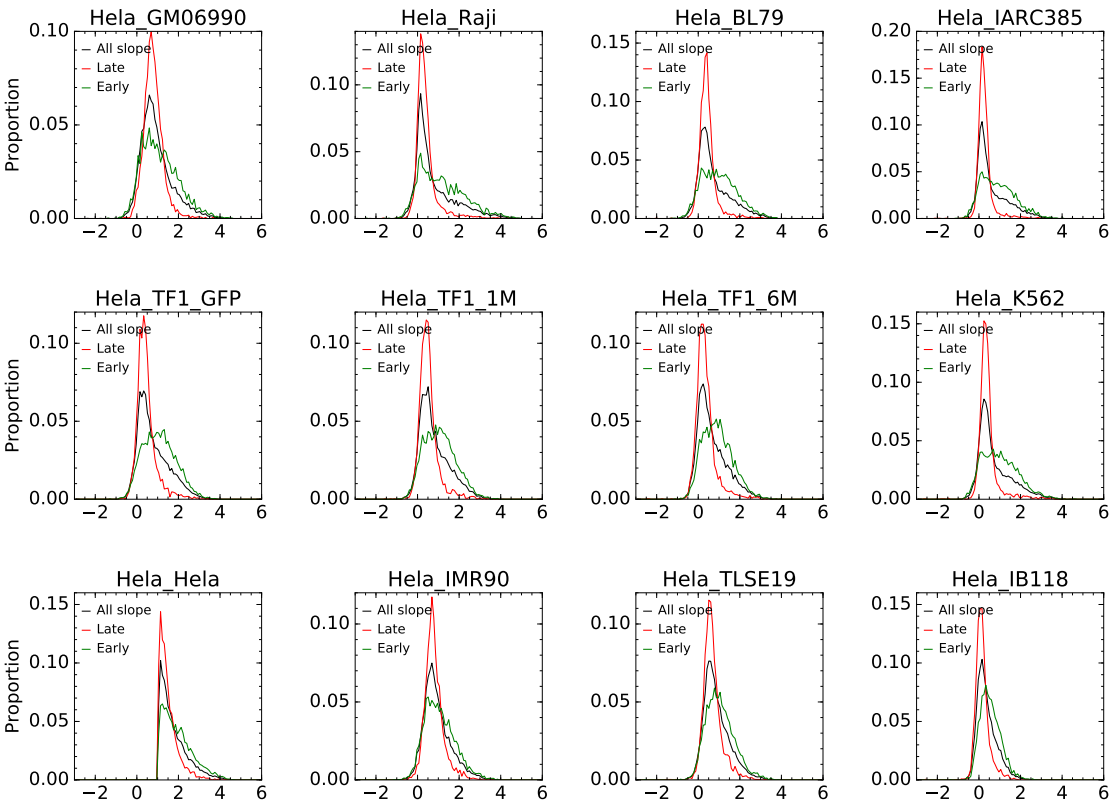

Figure B.17: **Pdf of RFD slopes in each cell line at the location of IZ selected in HeLa.** Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in HeLa. (Black) complete genome. (Red) loci where MRT_{IMR90} >0.7. (Green) loci where MRT_{IMR90} <0.3.

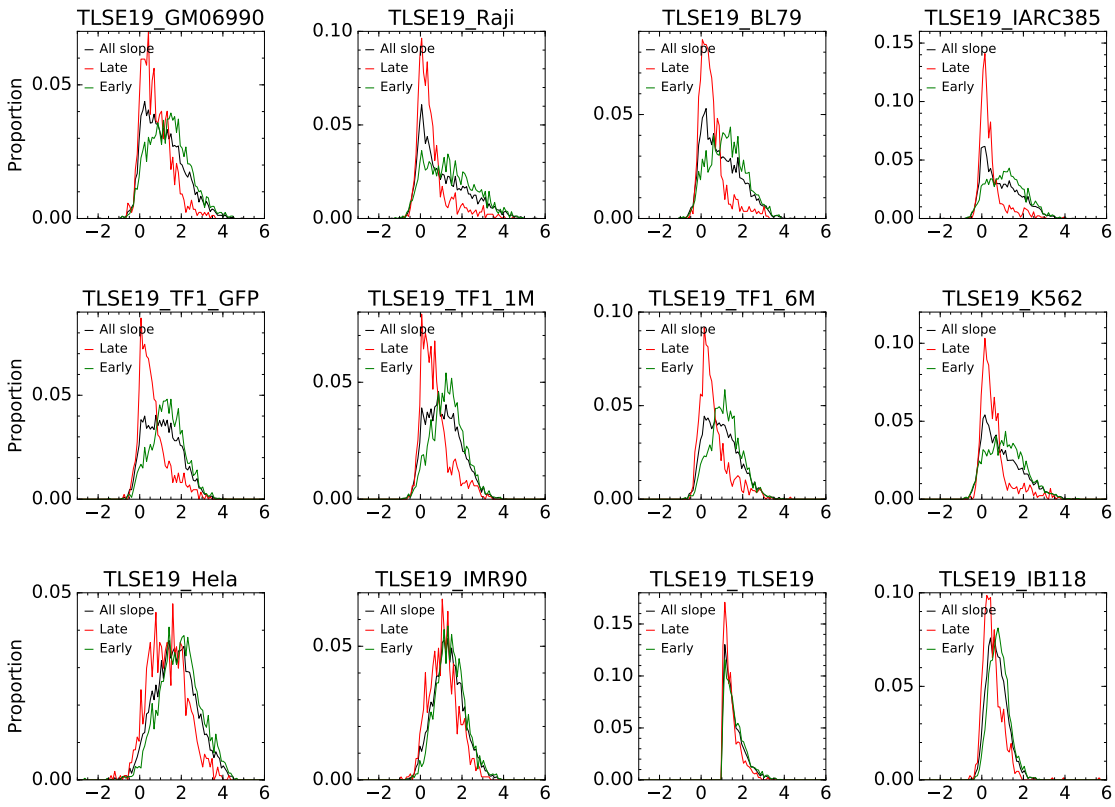

Figure B.18: **Pdf of RFD slopes in each cell line at the location of IZ selected in TLSE19**. Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in TLSE19. (Black) complete genome. (Red) loci where MRT_{IMR90} >0.7. (Green) loci where MRT_{IMR90} <0.3.

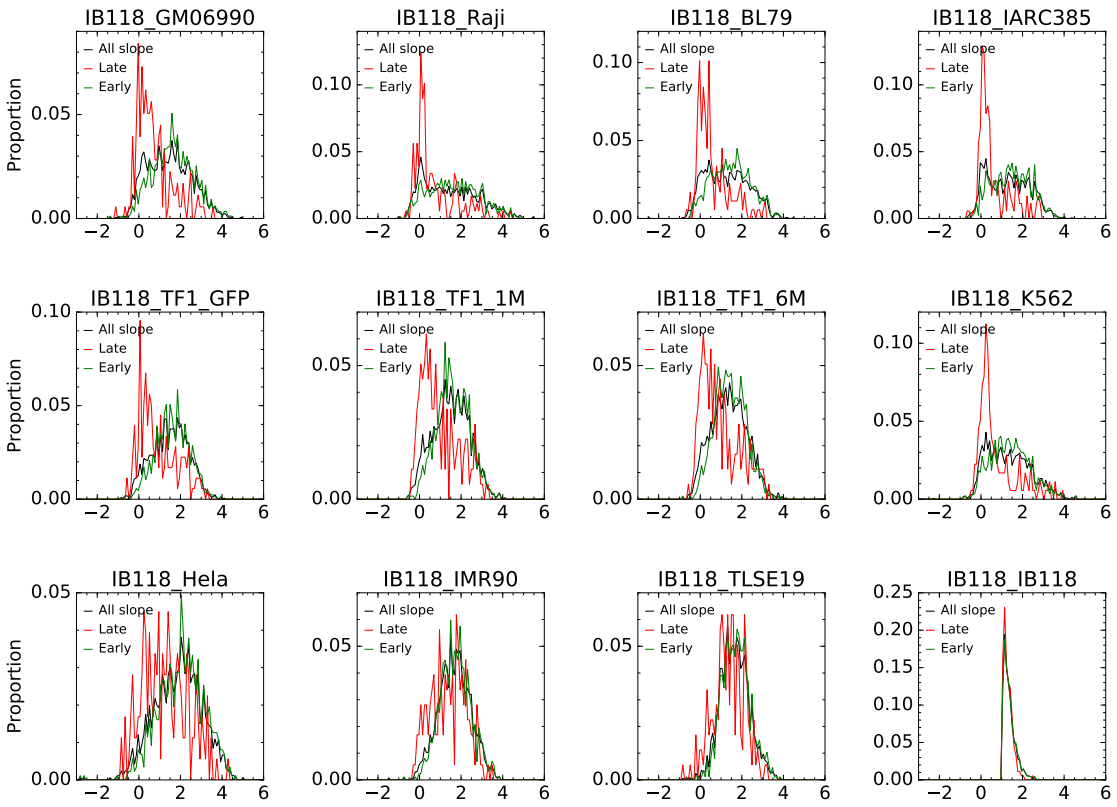

Figure B.19: **Pdf of RFD slopes in each cell line at the location of IZ selected in IB118.** Pdf of the RFD slopes in the indicated cell lines at the location of slope maxima with MS>1%RFD per kb detected in IB118. (Black) complete genome. (Red) loci where MRT_{IMR90} >0.7. (Green) loci where MRT_{IMR90} <0.3.

APPENDIX C

Supplementary figures for Chapter V

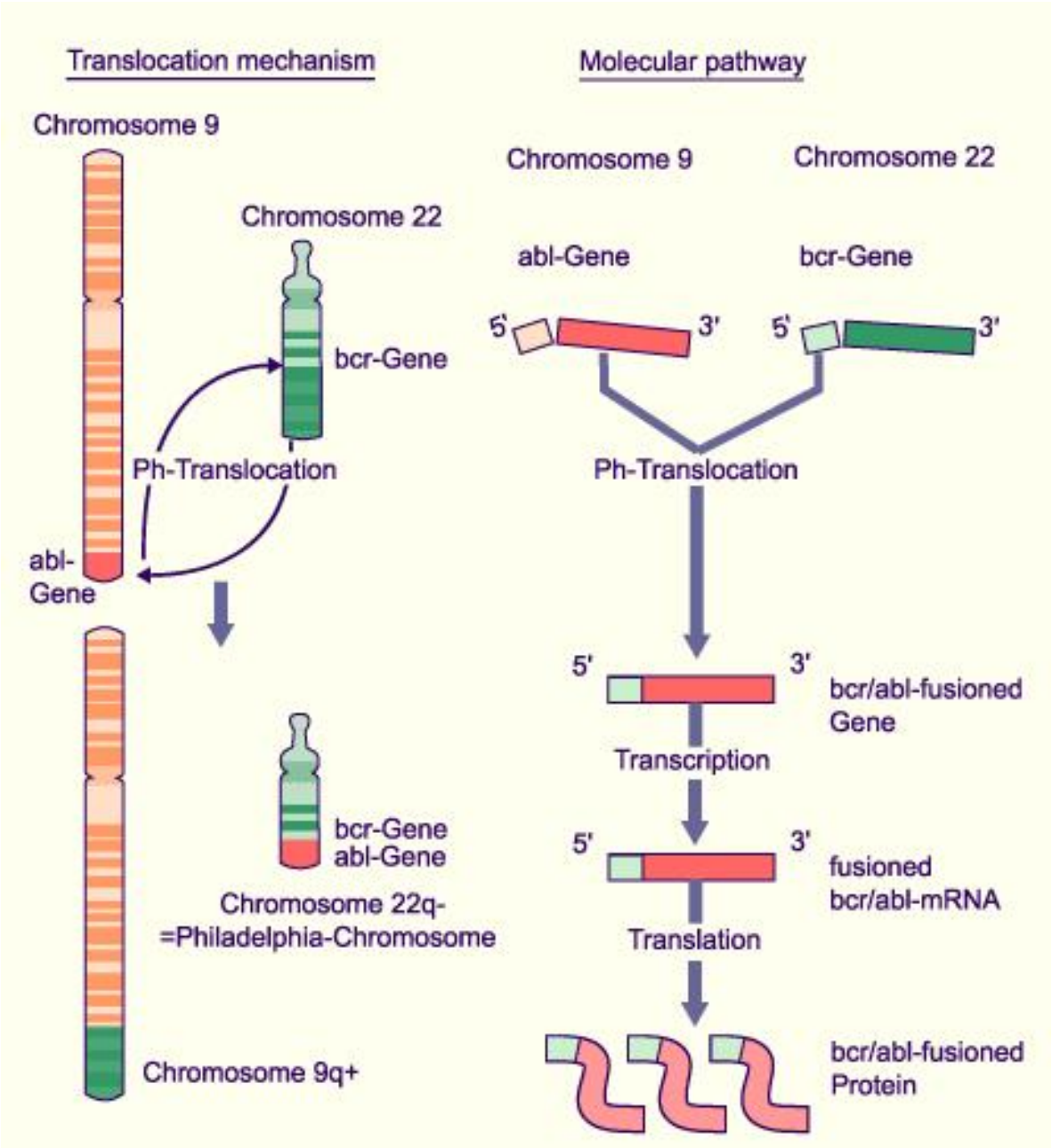

Figure C.1: **Molecular mechanism of the CML**. A break occurs in the abl gene on chromosome 9 and another break in the bcr gene on chromosome 22. The abl gene is a proto-oncogene. The translocation causes the fusion of these two genes bcr/abl. By its fusion with another gene, the proto-oncogene abl is then transformed into an oncogene. Therefore a new mRNA is produced: bcr/abl-mRNA. A bcr/abl fusion protein is thus synthesized. It induces a tyrosine kinase activity above normal which stimulates an over proliferation of precursor bone marrow cells. (Adapted from http://www.embryology.ch/francais/kchromaber/popupchromaber/02abweichende/mfphilainterak/01.html)

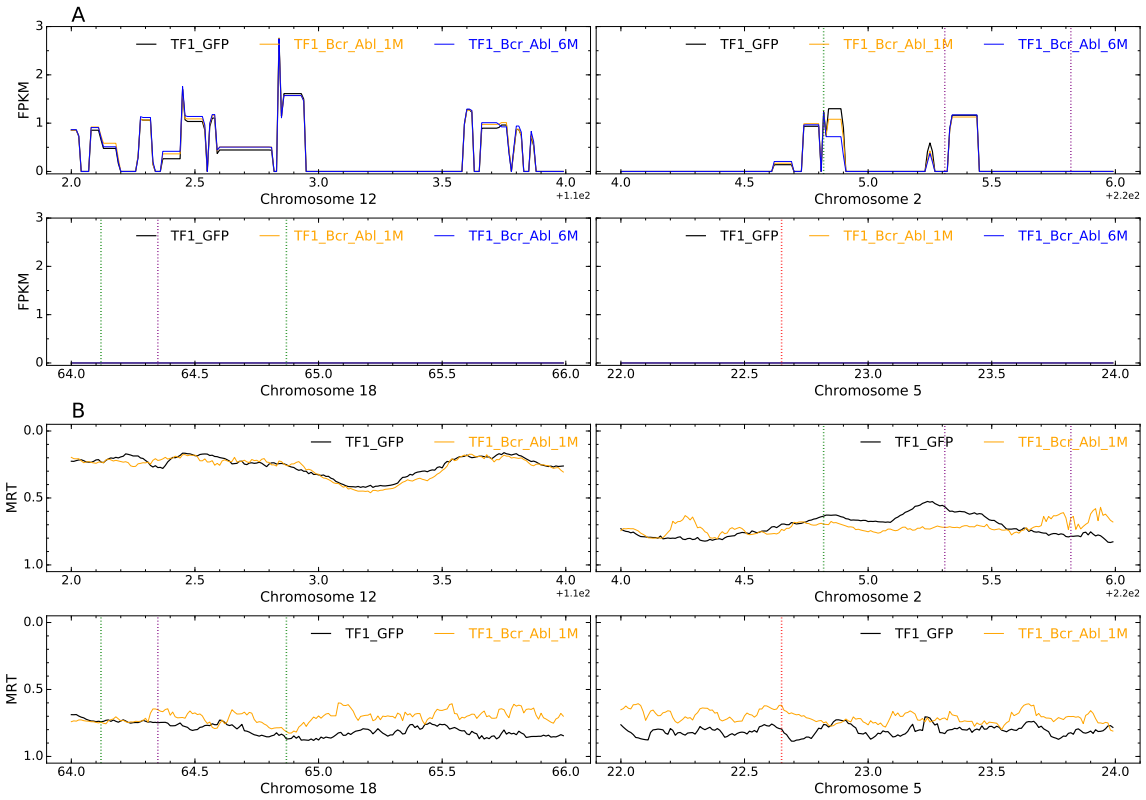

Figure C.2: **RNA-seq (A) and MRT (B) profiles of the same 2Mb regions and the same cell lines as in Figure 5.5.**

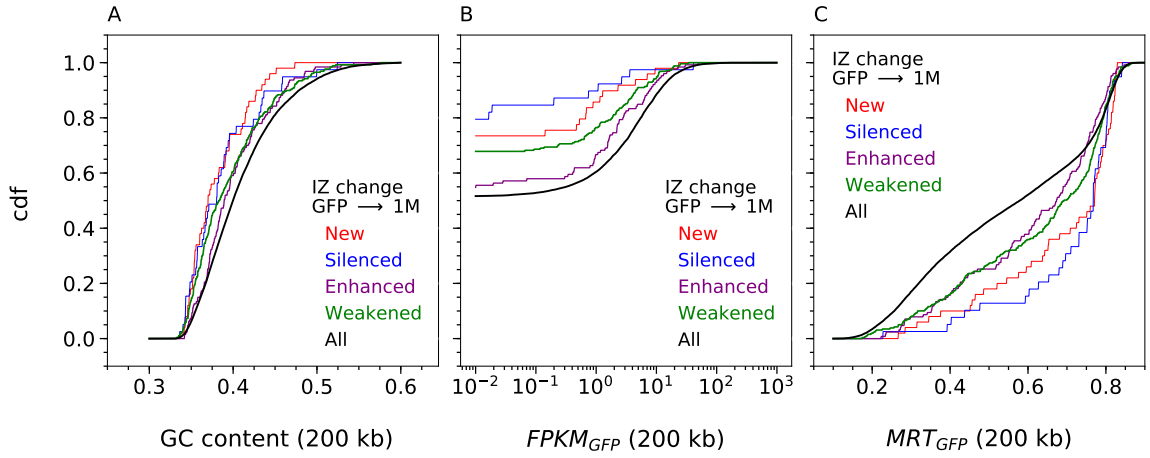

Figure C.3: **Initiation zones efficiency changes in response to 1 month of BCR_ABL1 expression are observed in GC-poor, lowly-transcribed and late replicating regions.** Cumulative distribution functions (cdf) of GC content (A), transcription in TF1_GFP (B) and MRT in TF1_GFP (C) computed in non-overlapping 200 kb windows of the 22 autosomes. Cdfs were determined for all windows (all, black) or limited to windows with Silenced (blue), Weakened (green), Enhanced (violet) and New (red) IZ. Similar results are obtained using transcription and MRT data of TF1-GFP.

APPENDIX D

Supplementary figures for Chapter VI

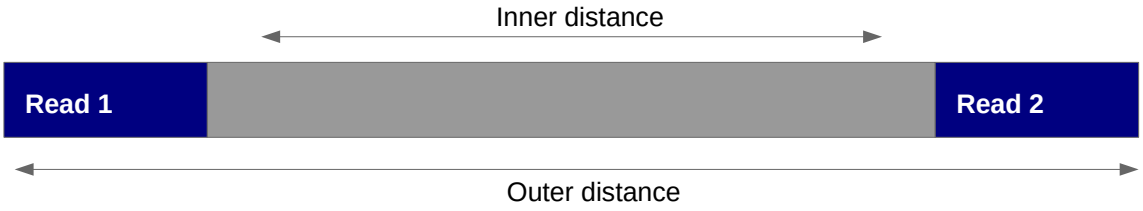

Figure D.1: **Example of paired-end sequencing.**

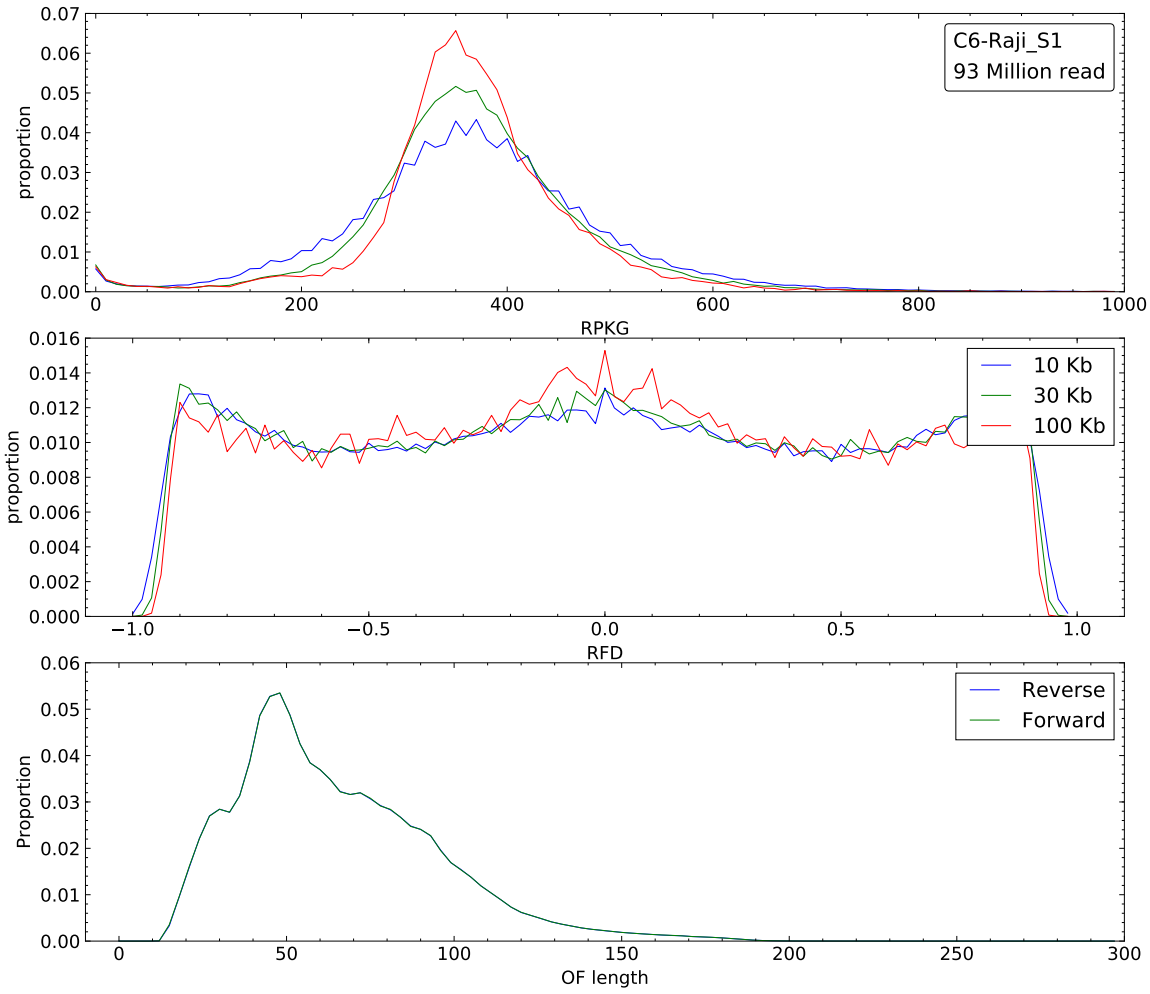

Figure D.2: **Distribution of number of reads, RFD amplitude and Okazaki fragment length in Raji cell line.** Pdf of paired-end read count (RPKG, top), RFD (middle) and Okazaki fragment length estimated as paired-end reads outer distance (Figure D.1). RPKG and RFD were computed in 10 kb (blue), 30 kb (green) and 100 kb (red) non-overlapping windows.

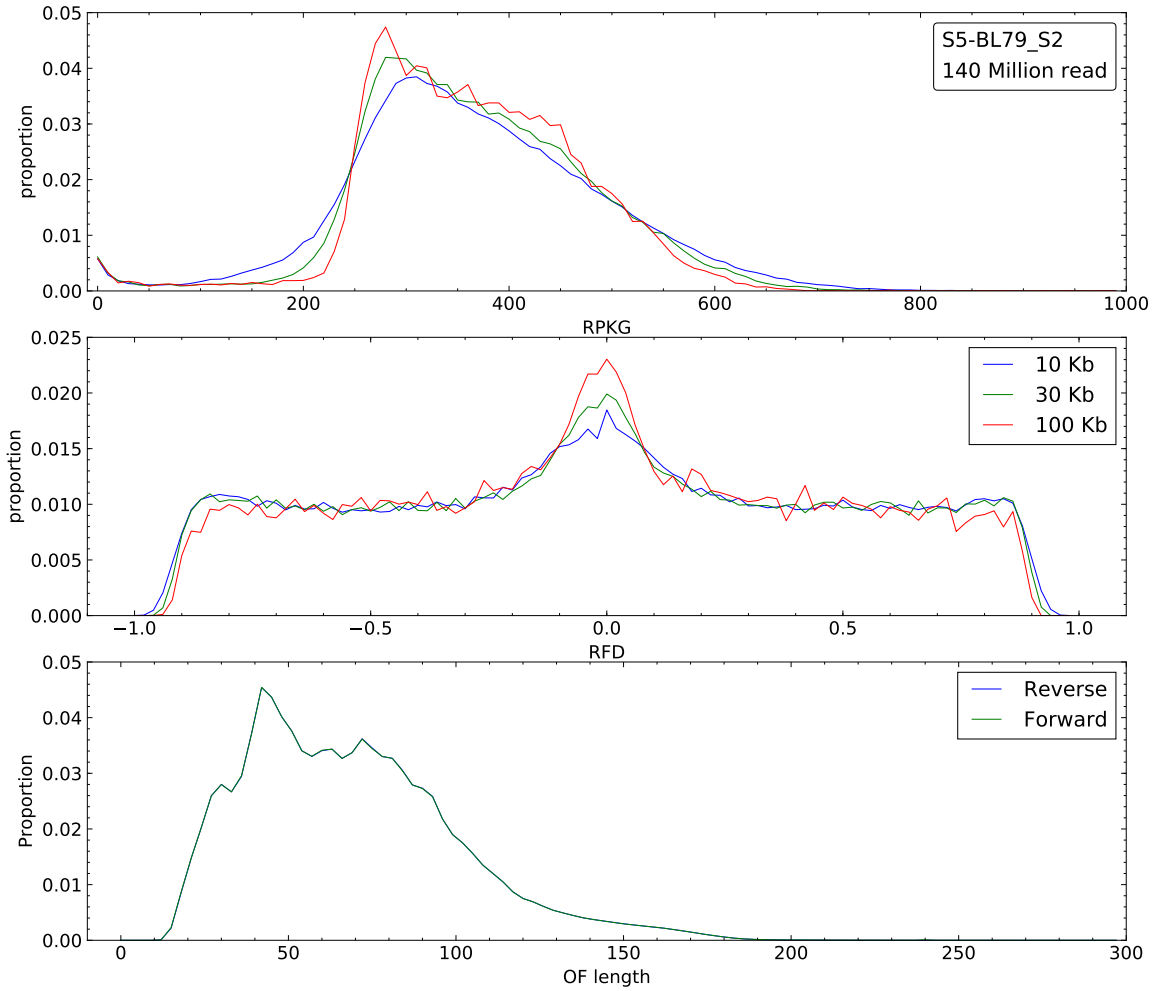

Figure D.3: **Distribution of number of reads, RFD amplitude and Okazaki fragment length in BL79 cell line.** Pdf of paired-end read count (RPKG, top), RFD (middle) and Okazaki fragment length estimated as paired-end reads outer distance (Figure D.1). RPKG and RFD were computed in 10 kb (blue), 30 kb (green) and 100 kb (red) non-overlapping windows.

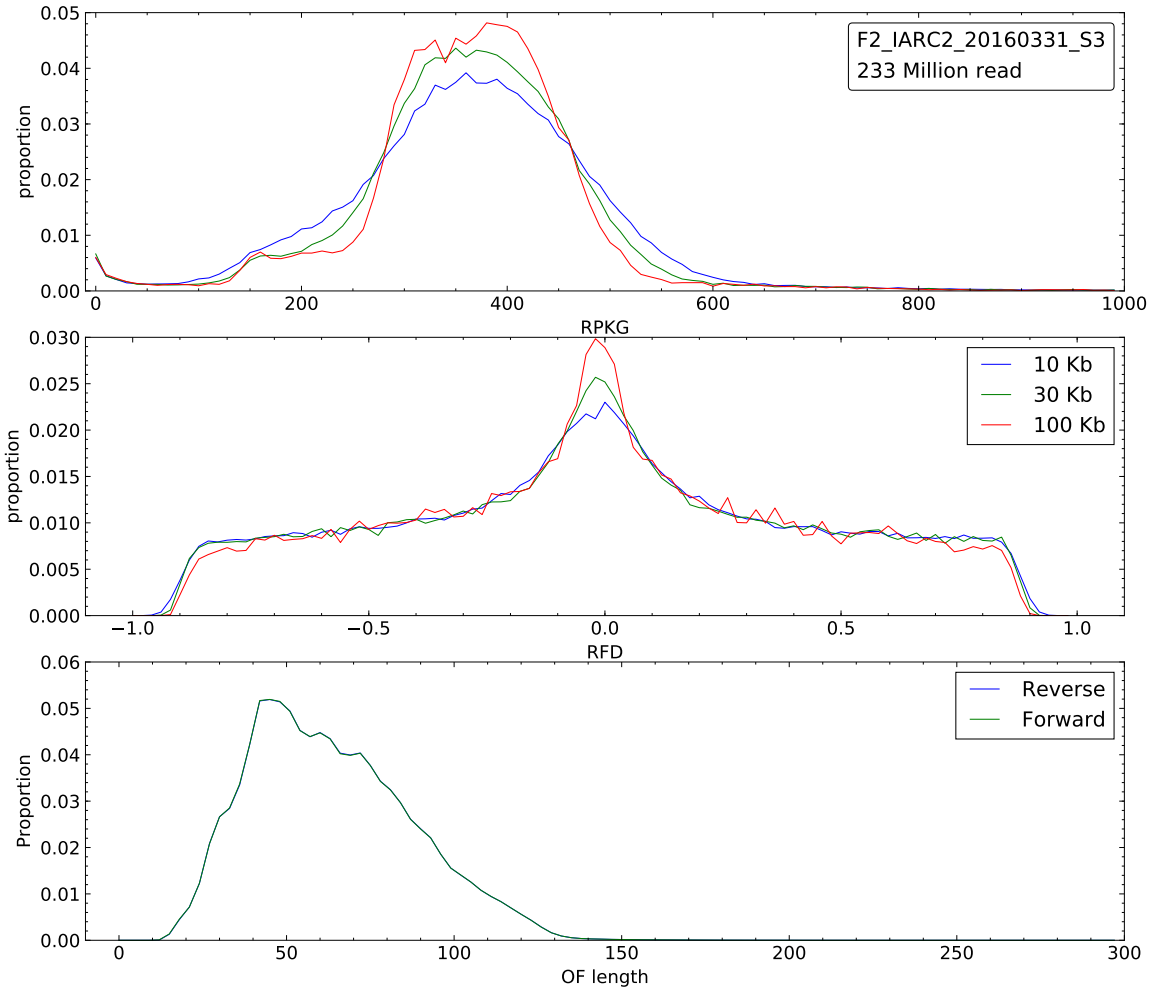

Figure D.4: **Distribution of number of reads, RFD amplitude and Okazaki fragment length in IARC2 cell line.** Pdf of paired-end read count (RPKG, top), RFD (middle) and Okazaki fragment length estimated as paired-end reads outer distance (Figure D.1). RPKG and RFD were computed in 10 kb (blue), 30 kb (green) and 100 kb (red) non-overlapping windows.

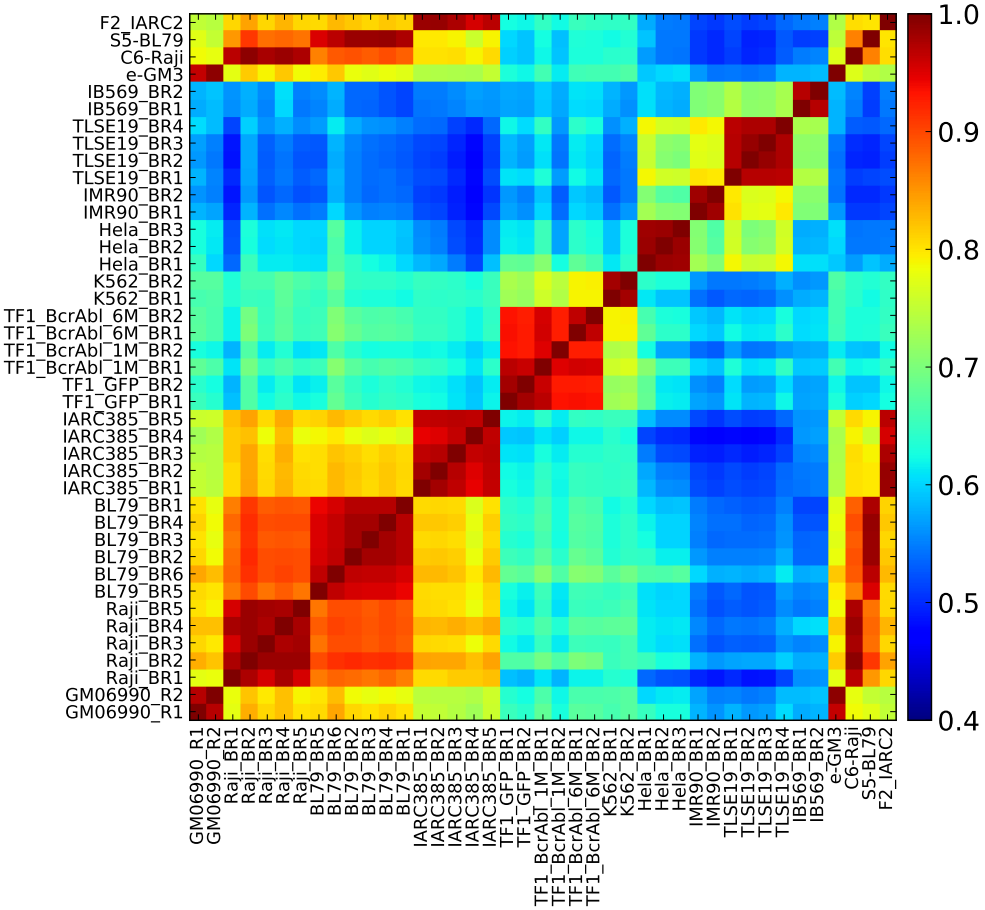

Figure D.5: **Correlation matrix of RFD profiles at 10 kb with paired-end data.** Same as the Figure 3.5 but we added 4 lymphoid paired-end cell lines. e-GM3 is a GM06990 dataset, C6_Raji is a Raji dataset, S5_BL79 is a BL79 dataset and F2_IARC2 is a IARC385 dataset.

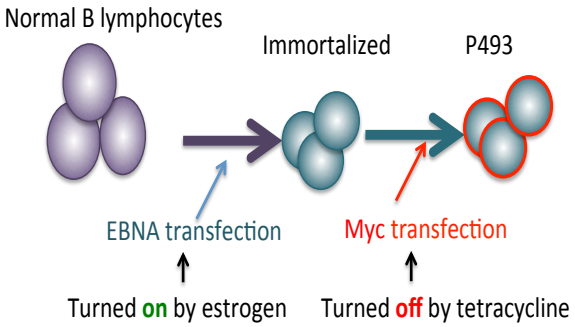

Figure D.6: **Procedure to obtain P493 cell lines.** A normal B lymphocyte was first transfected with EBNA under the control of a promoter that is turned on in the presence of estrogen. It was further transfected by Myc under the control of a promoter that is turned off in presence of tetracycline. Cell line was grown in 3 conditions, with only EBNA expression for long term, with only Myc over-expression for long term or with both Myc over-expression and EBNA expression for long term.

Short name	Expression of genes	Culture conditions
P493_mEl	With only EBNA expression for long term ($\geq$ 30 days)	+estrogen +tetracycline for $\geq$ 30 days
P493_MEl	With both Myc over-expression and EBNA expression for long term ($\geq$ 30 days)	+estrogen −tetracycline for $\geq$ 30 days
P493_Mel	With only Myc over-expression for long term ($\geq$ 30 days)	−estrogen −tetracycline for $\geq$ 30 days
P493_mEl_MEs	With only EBNA expression for long term ($\geq$ 30 days), then with both Myc over-expression and EBNA expression for short term (1 day)	+estrogen +tetracycline for $\geq$ 30 days, then +estrogen −tetracycline for 1 day
P493_mEl_Mes	With only EBNA expression for long term ($\geq$ 30 days), then with only Myc over-expression for short term (1 day)	+estrogen +tetracycline for $\geq$ 30 days, then −estrogen −tetracycline for 1 day
P493_Mel_mEs	With only Myc over-expression for long term ($\geq$ 30 days), then with only EBNA expression for short term (1 day)	−estrogen −tetracycline for $\geq$ 30 days, then +estrogen +tetracycline for 1 day
P493_MEl_mEs	With both Myc over-expression and EBNA expression for long term ($\geq$ 30 days), then with only EBNA expression for short term (1 day)	+estrogen −tetracycline for $\geq$ 30 days, then +estrogen +tetracycline for 1 day

Table D.1: **Description of the P493 cell line datasets.** First column represent the short name of cell lines. second column represent the state of expressed gene, third column represent the procedure of culture. "M": Myc was over-expression; "m": Myc was not over-expressed; "E": EBNA was expressed; "e": EBNA was not expressed; "l": long-term ($\geq$ 1 month) culture; "s": short-term (1 day) culture.

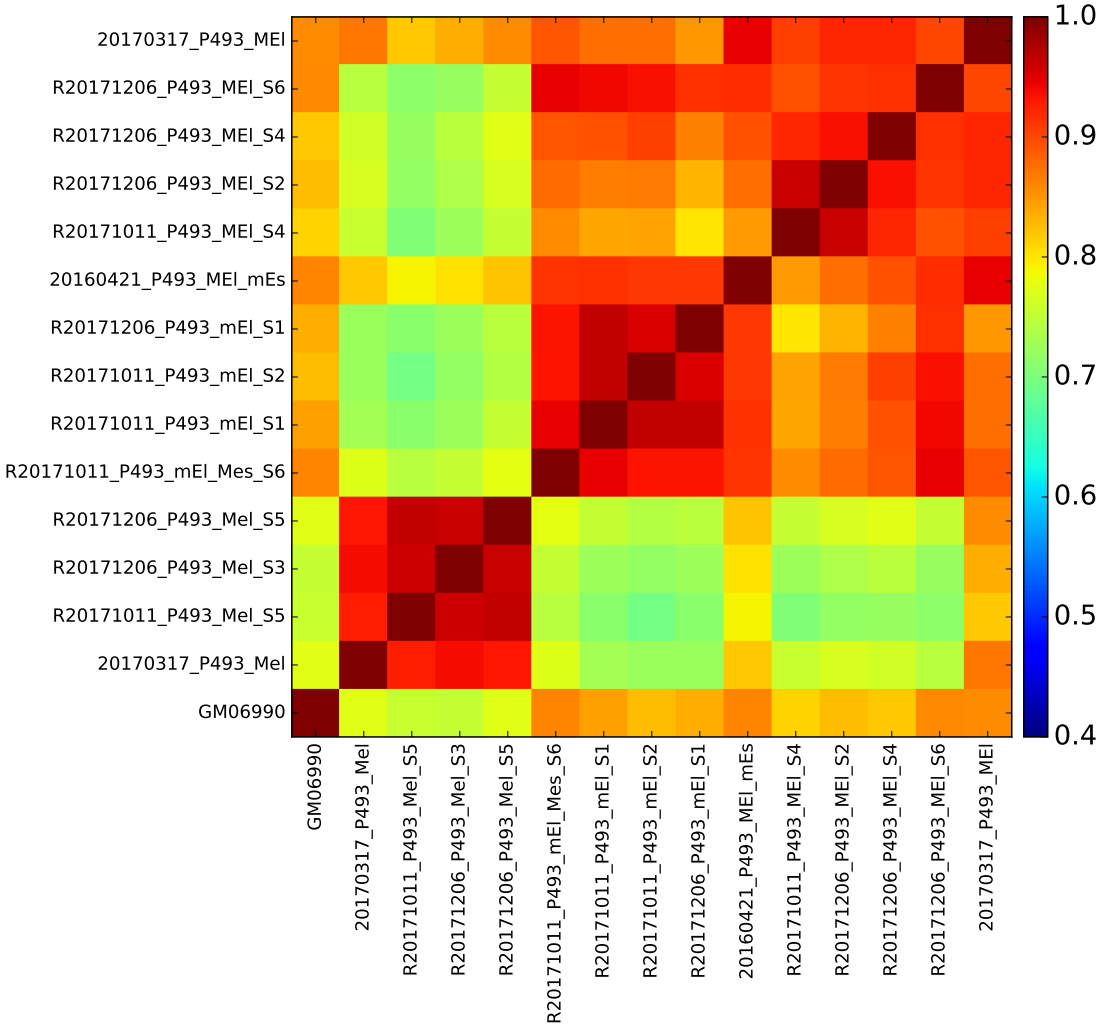

Figure D.7: **RFD correlation matrix between GM06990 and all P493 cell lines.** RFD correlation matrix of cell lines at 10 Kb. Pearson correlation coefficient values are color-coded from blue (0.4) to red (1) using the color bar on the rights.

List of Figures

List of Tables

Bibliography

[1] K. A. Cimprich and D. Cortez, "ATR: an essential regulator of genome integrity," *Nature reviews Molecular cell biology*, vol. 9, no. 8, p. 616, 2008. 11

[2] R. D. Paulsen and K. A. Cimprich, "The ATR pathway: fine-tuning the fork," *DNA repair*, vol. 6, no. 7, pp. 953–966, 2007. 11

[3] N. Petryk, M. Kahli, Y. d'Aubenton Carafa, Y. Jaszczyszyn, Y. Shen, M. Silvain, C. Thermes, C.-L. Chen, and O. Hyrien, "Replication landscape of the human genome," *Nature communications*, vol. 7, p. 10208, 2016. 12, 22, 26, 27, 29, 33, 46, 47, 48, 50, 52, 59, 70, 72, 73, 78, 99, 105, 109, 120, 123, 124, 125

[4] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, *Molecular Biology of the Cell*. Garland, 4th ed., 2002. 12, 108

[5] R. E. Franklin and R. G. Gosling, "Molecular configuration in sodium thymonucleate," *Nature*, vol. 171, no. 4356, p. 740, 1953. 12

[6] J. D. Watson, F. H. Crick, *et al.*, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953. 12

[7] J. D. Watson and F. H. Crick, "Genetical implications of the structure of deoxyribonucleic acid," *Nature*, vol. 171, no. 4361, pp. 964–967, 1953. 12

[8] B. Alberts, D. Bray, K. Hopkin, A. D. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential cell biology*. Garland Science, 2015. 13, 15, 17

[9] F. Jacob, S. Brenner, and F. Cuzin, "On the regulation of DNA replication in bacteria," in *Cold Spring Harbor symposia on quantitative biology*, vol. 28, pp. 329–348, Cold Spring Harbor Laboratory Press, 1963. 13

[10] S. P. Bell and A. Dutta, "DNA replication in eukaryotic cells," *Annual review of biochemistry*, vol. 71, no. 1, pp. 333–374, 2002. 14

[11] M. L. DePamphilis, J. J. Blow, S. Ghosh, T. Saha, K. Noguchi, and A. Vassilev, "Regulating the licensing of DNA replication origins in metazoa," *Current opinion in cell biology*, vol. 18, no. 3, pp. 231–239, 2006. 14

[12] M. DePamphilis and S. Bell, "Genome duplication (concepts, mechanisms, evolution, and disease) garland science," 2011. 14

[13] J. F. Diffley and K. Labib, "The chromosome replication cycle," *Journal of cell science*, vol. 115, no. 5, pp. 869–872, 2002. 14

[14] H. Araki, "Initiation of chromosomal DNA replication in eukaryotic cells; contribution of yeast genetics to the elucidation," *Genes & genetic systems*, vol. 86, no. 3, pp. 141–149, 2011. 15

[15] J. J. Wyrick, J. G. Aparicio, T. Chen, J. D. Barnett, E. G. Jennings, R. A. Young, S. P. Bell, and O. M. Aparicio, "Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins," *Science*, vol. 294, no. 5550, pp. 2357–2360, 2001.

[16] Y. Marahrens and B. Stillman, "A yeast chromosomal origin of DNA replication defined by multiple functional elements," *Science*, vol. 255, no. 5046, pp. 817–823, 1992. 15

[17] M. W. Parker, M. R. Botchan, and J. M. Berger, "Mechanisms and regulation of DNA replication initiation in eukaryotes," *Critical reviews in biochemistry and molecular biology*, vol. 52, no. 2, pp. 107–144, 2017. 15

[18] M. Méchali, "Eukaryotic DNA replication origins: many choices for appropriate answers," *Nature reviews Molecular cell biology*, vol. 11, no. 10, p. 728, 2010. 16, 21, 22, 30

[19] C. Cvetic and J. C. Walter, "Eukaryotic origins of DNA replication: could you please be more specific?," in *Seminars in cell & developmental biology*, vol. 16, pp. 343–353, Elsevier, 2005. 16

[20] E. Johansson and N. Dixon, "Replicative DNA polymerases," *Cold Spring Harbor perspectives in biology*, vol. 5, no. 6, p. a012799, 2013. 16

[21] R. E. Johnson, R. Klassen, L. Prakash, and S. Prakash, "A major role of DNA polymerase δ in replication of both the leading and lagging DNA strands," *Molecular cell*, vol. 59, no. 2, pp. 163–175, 2015.

[22] E. E. Henninger and Z. F. Pursell, "DNA polymerase ε and its roles in genome stability," *IUBMB life*, vol. 66, no. 5, pp. 339–351, 2014.

[23] S. A. N. McElhinny, D. A. Gordenin, C. M. Stith, P. M. Burgers, and T. A. Kunkel, "Division of labor at the eukaryotic replication fork," *Molecular cell*, vol. 30, no. 2, pp. 137–144, 2008. 16

[24] V. Aria and J. T. Yeeles, "Mechanism of bidirectional leading-strand synthesis establishment at eukaryotic DNA replication origins," *Molecular cell*, vol. 73, no. 2, pp. 199–211, 2019. 16

[25] Y.-H. Kang, W. C. Galal, A. Farina, I. Tappin, and J. Hurwitz, "Properties of the human Cdc45/Mcm2-7/GINS helicase complex and its action with DNA polymerase ε in rolling circle DNA synthesis," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 6042–6047, 2012.

[26] D. Maiorano, O. Cuvier, E. Danis, and M. Méchali, "MCM8 is an MCM2-7-related protein that functions as a DNA helicase during replication elongation and not initiation," *Cell*, vol. 120, no. 3, pp. 315–328, 2005.

[27] D. Maiorano, M. Lutzmann, and M. Méchali, "MCM proteins and DNA replication," *Current opinion in cell biology*, vol. 18, no. 2, pp. 130–136, 2006.

[28] M. E. Douglas, F. A. Ali, A. Costa, and J. F. Diffley, "The mechanism of eukaryotic CMG helicase activation," *Nature*, vol. 555, no. 7695, p. 265, 2018. 16

[29] T. Mondol, J. L. Stodola, R. Galletto, and P. M. Burgers, "PCNA accelerates the nucleotide incorporation rate by DNA polymerase δ," *Nucleic acids research*, 2019. 16

[30] G. Maga and U. Hübscher, "Proliferating cell nuclear antigen (PCNA): a dancer with many partners," *Journal of cell science*, vol. 116, no. 15, pp. 3051–3060, 2003. 16

[31] R. M. Jones and E. Petermann, "Replication fork dynamics and the DNA damage response," *Biochemical Journal*, vol. 443, no. 1, pp. 13–26, 2012. 17

[32] D. Santamaria, E. Viguera, M. L. Martínez-Robles, O. Hyrien, P. Hernandez, D. B. Krimer, and J. B. Schvartzman, "Bi-directional replication and random termination," *Nucleic acids research*, vol. 28, no. 10, pp. 2099–2107, 2000. 17

[33] T. Eydmann, E. Sommariva, T. Inagawa, S. Mian, A. Klar, and J. Z. Dalgaard, "Rtf1-mediated eukaryotic site-specific replication termination," *Genetics*, 2008. 17

[34] R. D. Little, T. H. Platt, and C. L. Schildkraut, "Initiation and termination of DNA replication in human rRNA genes.," *Molecular and Cellular Biology*, vol. 13, no. 10, pp. 6600–6613, 1993. 17

[35] D. Fachinetti, R. Bermejo, A. Cocito, S. Minardi, Y. Katou, Y. Kanoh, K. Shirahige, A. Azvolinsky, V. A. Zakian, and M. Foiani, "Replication termination at eukaryotic chromosomes is mediated by top2 and occurs at genomic loci containing pausing elements," *Molecular cell*, vol. 39, no. 4, pp. 595–605, 2010. 17, 46

[36] S. Lambert and A. M. Carr, "Checkpoint responses to replication fork barriers," *Biochimie*, vol. 87, no. 7, pp. 591–602, 2005. 17, 18

[37] I. Hiratani, T. Ryba, M. Itoh, T. Yokochi, M. Schwaiger, C.-W. Chang, Y. Lyou, T. M. Townes, D. Schübeler, and D. M. Gilbert, "Global reorganization of replication domains during embryonic stem cell differentiation," *PLoS biology*, vol. 6, no. 10, p. e245, 2008. 18, 19, 75, 78, 104, 117, 123

[38] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert, "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types," *Genome research*, vol. 20, no. 6, pp. 761–770, 2010. 18, 19, 25, 26, 75, 77, 104, 125

[39] R. Berezney, D. D. Dubey, and J. A. Huberman, "Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci," *Chromosoma*, vol. 108, no. 8, pp. 471–484, 2000. 19, 28

[40] A. J. McNairn and D. M. Gilbert, "Epigenomic replication: linking epigenetics to DNA replication," *Bioessays*, vol. 25, no. 7, pp. 647–656, 2003. 19, 30

[41] N. Rhind and D. M. Gilbert, "DNA replication timing," *Cold Spring Harbor perspectives in biology*, vol. 5, no. 8, p. a010132, 2013. 18, 19, 78

[42] R. Hand, "Eucaryotic DNA: organization of the genome for replication," *Cell*, vol. 15, no. 2, pp. 317–325, 1978. 18

[43] H. Ma, J. Samarabandu, R. S. Devdhar, R. Acharya, P.-c. Cheng, C. Meng, and R. Berezney, "Spatial and temporal dynamics of DNA replication sites in mammalian cells," *The Journal of cell biology*, vol. 143, no. 6, pp. 1415–1425, 1998. 18

[44] S. Farkash-Amar and I. Simon, "Genome-wide analysis of the replication program in mammals," *Chromosome research*, vol. 18, no. 1, pp. 115–125, 2010. 18

[45] D. M. Gilbert, "Evaluating genome-scale approaches to eukaryotic DNA replication," *Nature reviews Genetics*, vol. 11, no. 10, p. 673, 2010. 18, 21, 75

[46] S. Farkash-Amar, D. Lipson, A. Polten, A. Goren, C. Helmstetter, Z. Yakhini, and I. Simon, "Global organization of replication time zones of the mouse genome," *Genome research*, pp. gr–079566, 2008. 18, 104

[47] E. J. White, O. Emanuelsson, D. Scalzo, T. Royce, S. Kosak, E. J. Oakeley, S. Weissman, M. Gerstein, M. Groudine, M. Snyder, *et al.*, "DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states," *Proceedings of the National Academy of Sciences*, vol. 101, no. 51, pp. 17771–17776, 2004. 18, 78

[48] R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, and J. A. Stamatoyannopoulos, "Sequencing newly replicated DNA reveals widespread plasticity in human replication timing," *Proceedings of the National Academy of Sciences*, vol. 107, no. 1, pp. 139–144, 2010. 18, 25, 26, 29, 37

[49] C. A. Müller and C. A. Nieduszynski, "Conservation of replication timing reveals global and local regulation of replication origin activity," *Genome research*, vol. 22, no. 10, pp. 1953–1962, 2012. 18

[50] K. Woodfine, H. Fiegler, D. M. Beare, J. E. Collins, O. T. McCann, B. D. Young, S. Debernardi, R. Mott, I. Dunham, and N. P. Carter, "Replication timing of the human genome," *Human molecular genetics*, vol. 13, no. 2, pp. 191–202, 2003. 18, 29, 36, 59, 78, 117

[51] A. Baker, B. Audit, C.-L. Chen, B. Moindrot, A. Leleu, G. Guilbaud, A. Rappailles, C. Vaillant, A. Goldar, F. Mongelard, *et al.*, "Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines," *PLoS computational biology*, vol. 8, no. 4, p. e1002443, 2012. 18, 20, 24, 26, 27, 28, 37, 120

[52] G. Guilbaud, A. Rappailles, A. Baker, C.-L. Chen, A. Arneodo, A. Goldar, Y. d'Aubenton Carafa, C. Thermes, B. Audit, and O. Hyrien, "Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome," *PLoS computational biology*, vol. 7, no. 12, p. e1002322, 2011. 18, 20, 24, 27, 28, 120

[53] B. Audit, L. Zaghloul, A. Baker, A. Arneodo, C.-L. Chen, Y. d'Aubenton Carafa, and C. Thermes, "Megabase replication domains along the human genome: relation to chromatin structure and genome organisation," in *Epigenetics: Development and Disease*, pp. 57–80, Springer, 2013. 18, 20, 28, 86

[54] B. Audit, A. Baker, C.-L. Chen, A. Rappailles, G. Guilbaud, H. Julienne, A. Goldar, Y. D'Aubenton-Carafa, O. Hyrien, C. Thermes, *et al.*, "Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm," *Nature protocols*, vol. 8, no. 1, p. 98, 2013. 18, 20, 25, 37, 70

[55] I. Hiratani, T. Ryba, M. Itoh, J. Rathjen, M. Kulik, B. Papp, E. Fussner, D. P. Bazett-Jones, K. Plath, S. Dalton, *et al.*, "Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis," *Genome research*, vol. 20, no. 2, pp. 155–169, 2010. 20, 123

[56] O. Hyrien, A. Rappailles, G. Guilbaud, A. Baker, C.-L. Chen, A. Goldar, N. Petryk, M. Kahli, E. Ma, Y. d'Aubenton Carafa, *et al.*, "From simple bacterial and archaeal replicons to replication N/U-domains," *Journal of molecular biology*, vol. 425, no. 23, pp. 4673–4689, 2013. 20, 27, 28, 120

[57] W. L. Fangman and B. J. Brewer, "Activation of replication origins within yeast chromosomes," *Annual review of cell biology*, vol. 7, no. 1, pp. 375–402, 1991. 21

[58] C. Newlon, L. Lipchitz, I. Collins, A. Deshpande, R. Devenish, R. Green, H. Klein, T. Palzkill, R. Ren, and S. Synn, "Analysis of a circular derivative of saccharomyces

cerevisiae chromosome III: a physical map and identification and location of ars elements.," *Genetics*, vol. 129, no. 2, pp. 343–357, 1991. 21

[59] M. Raghuraman, E. A. Winzeler, D. Collingwood, S. Hunt, L. Wodicka, A. Conway, D. J. Lockhart, R. W. Davis, B. J. Brewer, and W. L. Fangman, "Replication dynamics of the yeast genome," *science*, vol. 294, no. 5540, pp. 115–121, 2001. 21, 46, 78

[60] N. M. Berbenetz, C. Nislow, and G. W. Brown, "Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure," *PLoS genetics*, vol. 6, no. 9, p. e1001092, 2010. 21

[61] C. Heichinger, C. J. Penkett, J. Bähler, and P. Nurse, "Genome-wide characterization of fission yeast DNA replication origins," *The EMBO journal*, vol. 25, no. 21, pp. 5171–5179, 2006.

[62] C. A. Nieduszynski, Y. Knox, and A. D. Donaldson, "Genome-wide identification of replication origins in yeast by comparative genomics," *Genes & development*, vol. 20, no. 14, pp. 1874–1879, 2006.

[63] J. Xu, Y. Yanagisawa, A. M. Tsankov, C. Hart, K. Aoki, N. Kommajosyula, K. E. Steinmann, J. Bochicchio, C. Russ, A. Regev, *et al.*, "Genome-wide identification and characterization of replication origins by deep sequencing," *Genome biology*, vol. 13, no. 4, p. R27, 2012. 21

[64] L. T. Vassilev, W. C. Burhans, and M. L. DePamphilis, "Mapping an origin of DNA replication at a single-copy locus in exponentially proliferating mammalian cells.," *Molecular and cellular biology*, vol. 10, no. 9, pp. 4685–4689, 1990. 21

[65] W. C. Burhans, L. T. Vassilev, M. S. Caddle, N. H. Heintz, and M. L. DePamphills, "Identification of an origin of bidirectional DNA replication in mammalian chromosomes," *Cell*, vol. 62, no. 5, pp. 955–965, 1990. 21

[66] L. D. Mesner, V. Valsakumar, N. Karnani, A. Dutta, J. L. Hamlin, and S. Bekiranov, "Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription," *Genome research*, vol. 21, no. 3, pp. 377–389, 2011. 21

[67] K. Klein, W. Wang, T. Borrman, S. Chan, D. Zhang, Z. Weng, A. Hastie, C. Chen, D. M. Gilbert, and N. Rhind, "Genome-wide identification of early-firing human replication origins by optical replication mapping," *bioRxiv*, p. 214841, 2017. 21

[68] A. R. Langley, S. Gräf, J. C. Smith, and T. Krude, "Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq)," *Nucleic acids research*, vol. 44, no. 21, pp. 10230–10247, 2016. 21, 22, 46

[69] O. Hyrien, "Peaks cloaked in the mist: the landscape of mammalian replication origins," *J Cell Biol*, vol. 208, no. 2, pp. 147–160, 2015. 21, 24

[70] N. Rhind, "DNA replication timing: random thoughts about origin firing," *Nature cell biology*, vol. 8, no. 12, p. 1313, 2006. 21

[71] J. A. Huberman and A. D. Riggs, "On the mechanism of DNA replication in mammalian chromosomes," *Journal of molecular biology*, vol. 32, no. 2, pp. 327–341, 1968. 21

[72] R. Lebofsky, R. Heilig, M. Sonnleitner, J. Weissenbach, and A. Bensimon, "DNA replication origin interference increases the spacing between initiation events in human cells," *Molecular biology of the cell*, vol. 17, no. 12, pp. 5337–5345, 2006. 21

[73] P. Norio, S. Kosiyatrakul, Q. Yang, Z. Guan, N. M. Brown, S. Thomas, R. Riblet, and C. L. Schildkraut, "Progressive activation of DNA replication initiation in large domains of the immunoglobulin heavy chain locus during b cell development," *Molecular cell*, vol. 20, no. 4, pp. 575–587, 2005.

[74] M. Anglana, F. Apiou, A. Bensimon, and M. Debatisse, "Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing," *Cell*, vol. 114, no. 3, pp. 385–394, 2003.

[75] P. Pasero, A. Bensimon, and E. Schwob, "Single-molecule analysis reveals clustering and epigenetic regulation of replication origins at the yeast rDNA locus," *Genes & development*, vol. 16, no. 19, pp. 2479–2484, 2002.

[76] S. Courbet, S. Gay, N. Arnoult, G. Wronka, M. Anglana, O. Brison, and M. Debatisse, "Replication fork movement sets chromatin loop size and origin choice in mammalian cells," *Nature*, vol. 455, no. 7212, p. 557, 2008. 21, 30

[77] J.-C. Cadoret, F. Meisch, V. Hassan-Zadeh, I. Luyten, C. Guillet, L. Duret, H. Quesneville, and M.-N. Prioleau, "Genome-wide studies highlight indirect links between human replication origins and gene regulation," *Proceedings of the National Academy of Sciences*, 2008. 21, 29, 46, 94

[78] N. Karnani, C. M. Taylor, A. Malhotra, and A. Dutta, "Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection," *Molecular biology of the cell*, vol. 21, no. 3, pp. 393–404, 2010. 21, 22, 46

[79] C. Cayrou, P. Coulombe, A. Vigneron, S. Stanojcik, O. Ganier, E. Rivals, A. Puy, S. Laurent-Chabalier, R. Desprat, M. Mechali, *et al.*, "Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features," *Genome research*, pp. gr–121830, 2011. 22, 28, 94

[80] C. Cayrou, P. Coulombe, A. Puy, S. Rialle, N. Kaplan, E. Segal, and M. Méchali, "New insights into replication origin characteristics in metazoans," *Cell Cycle*, vol. 11, no. 4, pp. 658–667, 2012. 22

[81] E. Besnard, A. Babled, L. Lapasset, O. Milhavet, H. Parrinello, C. Dantec, J.-M. Marin, and J.-M. Lemaitre, "Unraveling cell type–specific and reprogrammable human replication origin signatures associated with g-quadruplex consensus motifs," *Nature Structural and Molecular Biology*, vol. 19, no. 8, p. 837, 2012. 22, 46

[82] L. D. Mesner, V. Valsakumar, M. Cieślik, R. Pickin, J. L. Hamlin, and S. Bekiranov, "Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early-and late-firing origins," *Genome research*, pp. gr–155218, 2013. 22, 46

[83] J. Lobry, "Properties of a general model of DNA evolution under no-strand-bias conditions," *Journal of molecular evolution*, vol. 40, no. 3, pp. 326–330, 1995. 23

[84] E. P. Rocha, A. Danchin, and A. Viari, "Universal replication biases in bacteria," *Molecular microbiology*, vol. 32, no. 1, pp. 11–16, 1999. 23

[85] E. P. Rocha, "The replication-related organization of bacterial genomes," *Microbiology*, vol. 150, no. 6, pp. 1609–1627, 2004. 23

[86] J. Lobry, "Asymmetric substitution patterns in the two DNA strands of bacteria.," *Molecular biology and evolution*, vol. 13, no. 5, pp. 660–665, 1996. 23

[87] M. Touchon, S. Nicolay, A. Arnéodo, Y. d'Aubenton Carafa, and C. Thermes, "Transcription-coupled TA and GC strand asymmetries in the human genome," *FEBS letters*, vol. 555, no. 3, pp. 579–582, 2003. 24

[88] M. Touchon, S. Nicolay, and B. Audit, "Brodie of brodie eb, d'aubenton-carafa y, arneodo a, thermes c: Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins," *Proc Natl Acad Sci USA*, vol. 102, no. 28, pp. 9836–9841, 2005. 24

[89] M. Touchon, S. Nicolay, B. Audit, Y. d'Aubenton Carafa, A. Arneodo, C. Thermes, *et al.*, "Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins," *Proceedings of the National Academy of Sciences*, vol. 102, no. 28, pp. 9836–9841, 2005. 24

[90] M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton Carafa, A. Arneodo, and C. Thermes, "Human gene organization driven by the coordination of replication and transcription," *Genome research*, vol. 17, no. 9, pp. 000–000, 2007. 24, 27, 28

[91] A. Baker, S. Nicolay, L. Zaghloul, Y. d'Aubenton Carafa, C. Thermes, B. Audit, and A. Arneodo, "Wavelet-based method to disentangle transcription-and replication-associated strand asymmetries in mammalian genomes," *Applied and Computational Harmonic Analysis*, vol. 28, no. 2, pp. 150–170, 2010. 24

[92] K. Woodfine, D. M. Beare, K. Ichimura, S. Debernardi, A. J. Mungall, H. Fiegler, V. P. Collins, N. P. Carter, and I. Dunham, "Replication timing of human chromosome 6," *Cell Cycle*, vol. 4, no. 1, pp. 172–176, 2005. 25, 117

[93] C.-L. Chen, A. Rappailles, L. Duquenne, M. Huvet, G. Guilbaud, L. Farinelli, B. Audit, Y. d'Aubenton Carafa, A. Arneodo, O. Hyrien, *et al.*, "Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes," *Genome research*, pp. gr–098947, 2010. 37

[94] E. Yaffe, S. Farkash-Amar, A. Polten, Z. Yakhini, A. Tanay, and I. Simon, "Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture," *PLoS genetics*, vol. 6, no. 7, p. e1001011, 2010. 26, 104

[95] R. Desprat, D. Thierry-Mieg, N. Lailler, J. Lajugie, C. Schildkraut, J. Thierry-Mieg, and E. E. Bouhassira, "Predictable dynamic program of timing of DNA replication in human cells," *Genome research*, vol. 19, no. 12, pp. 2288–2299, 2009. 25

[96] C.-L. Chen, L. Duquenne, B. Audit, G. Guilbaud, A. Rappailles, A. Baker, M. Huvet, Y. d'Aubenton Carafa, O. Hyrien, A. Arneodo, *et al.*, "Replication-associated mutational asymmetry in the human genome," *Molecular biology and evolution*, vol. 28, no. 8, pp. 2327–2337, 2011. 27, 28

[97] H. Nakayasu and R. Berezney, "Mapping replicational sites in the eucaryotic cell nucleus.," *The Journal of Cell Biology*, vol. 108, no. 1, pp. 1–11, 1989. 28

[98] D. Baddeley, V. Chagin, L. Schermelleh, S. Martin, A. Pombo, P. Carlton, A. Gahl, P. Domaing, U. Birk, H. Leonhardt, *et al.*, "Measurement of replication structures at the nanometer scale using super-resolution light microscopy," *Nucleic acids research*, vol. 38, no. 2, pp. e8–e8, 2009. 28

[99] Z. Cseresnyes, U. Schwarz, and C. M. Green, "Analysis of replication factories in human cells by super-resolution light microscopy," *BMC cell biology*, vol. 10, no. 1, p. 88, 2009. 28

[100] D. Schübeler, D. Scalzo, C. Kooperberg, B. van Steensel, J. Delrow, and M. Groudine, "Genome-wide DNA replication profile for drosophila melanogaster: a link between transcription and replication timing," *Nature genetics*, vol. 32, no. 3, p. 438, 2002. 29, 78, 117

[101] D. M. MacAlpine, H. K. Rodríguez, and S. P. Bell, "Coordination of replication and transcription along a drosophila chromosome," *Genes & development*, vol. 18, no. 24, pp. 3094–3105, 2004. 29, 78

[102] B. E. Bernstein, A. Meissner, and E. S. Lander, "The mammalian epigenome," *Cell*, vol. 128, no. 4, pp. 669–681, 2007. 30

[103] A. D. Goldberg, C. D. Allis, and E. Bernstein, "Epigenetics: a landscape takes shape," *Cell*, vol. 128, no. 4, pp. 635–638, 2007.

[104] P. Winata, M. William, V. Keena, K. Takahashi, and Y. Y. Cheng, "DNA methylation in mammalian cells," in *Gene Expression and Regulation in Mammalian Cells-Transcription Toward the Establishment of Novel Therapeutics*, IntechOpen, 2018. 30

[105] D. M. Gilbert, "In search of the holy replicator," *Nature Reviews Molecular Cell Biology*, vol. 5, no. 10, p. 848, 2004. 30

[106] D. M. MacAlpine and G. Almouzni, "Chromatin and DNA replication," *Cold Spring Harbor perspectives in biology*, vol. 5, no. 8, p. a010207, 2013. 30, 125

[107] O. K. Smith and M. I. Aladjem, "Chromatin structure and replication origins: determinants of chromosome replication and nuclear organization," *Journal of molecular biology*, vol. 426, no. 20, pp. 3330–3341, 2014. 30

[108] Y. Gindin, M. S. Valenzuela, M. I. Aladjem, P. S. Meltzer, and S. Bilke, "A chromatin structure-based model accurately predicts DNA replication timing in human cells," *Molecular systems biology*, vol. 10, no. 3, 2014. 30

[109] I. Hiratani and D. M. Gilbert, "Replication timing as an epigenetic mark," *Epigenetics*, vol. 4, no. 2, pp. 93–97, 2009. 30

[110] H. Julienne, A. Zoufir, B. Audit, and A. Arneodo, "Human genome replication proceeds through four chromatin states," *PLoS computational biology*, vol. 9, no. 10, p. e1003233, 2013. 30

[111] H. Julienne, B. Audit, and A. Arneodo, "Embryonic stem cell specific "master" replication origins at the heart of the loss of pluripotency," *PLoS computational biology*, vol. 11, no. 2, p. e1003969, 2015. 30, 83, 86

[112] J. Sima, A. Chakraborty, V. Dileep, M. Michalski, K. N. Klein, N. P. Holcomb, J. L. Turner, M. T. Paulsen, J. C. Rivera-Mulia, C. Trevilla-Garcia, *et al.*, "Identifying cis elements for spatiotemporal control of mammalian DNA replication," *Cell*, vol. 176, no. 4, pp. 816–830, 2019. 30

[113] X. Wu, H. Kabalane, M. Kahli, N. Petryk, B. Laperrousaz, Y. Jaszczyszyn, G. Drillon, F.-E. Nicolini, G. Perot, A. Robert, *et al.*, "Developmental and cancer-associated plasticity of DNA replication preferentially targets gc-poor, lowly expressed and late-replicating regions," *Nucleic acids research*, vol. 46, no. 19, pp. 10157–10172, 2018. 33, 37, 47

[114] J. Yu, M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J. Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, *et al.*, "Induced pluripotent stem cell lines derived from human somatic cells," *science*, vol. 318, no. 5858, pp. 1917–1920, 2007. 33

[115] B. P. Lucey, W. A. Nelson-Rees, and G. M. Hutchins, "Henrietta lacks, HeLa cells, and cell culture contamination," *Archives of pathology & laboratory medicine*, vol. 133, no. 9, pp. 1463–1467, 2009. 33

[116] H. A. McGehee, "Johns hopkins–the birthplace of tissue culture: the story of Ross G. Harrison, Warren H. Lewis and George O. Gey.," *The Johns Hopkins Medical Journal*, vol. 136, no. 3, p. 142, 1975. 33

[117] A. Ferrari, I. Sultan, T. T. Huang, C. Rodriguez-Galindo, A. Shehadeh, C. Meazza, K. K. Ness, M. Casanova, and S. L. Spunt, "Soft tissue sarcoma across the age spectrum: A population-based study from the surveillance epidemiology and end results database," *Pediatric blood & cancer*, vol. 57, no. 6, pp. 943–949, 2011. 33

[118] E. Letessier, A. Hamy, J. Bailly, J. Paineau, and J. Visset, "Leiomyosarcomas of the rectum. amputation of the rectum or local resection?," in *Annales de chirurgie*, vol. 46, pp. 442–444, 1992. 35

[119] M. Hutt, "Historical introduction, burkitt's lymphoma, nasopharyngeal carcinoma and kaposi's sarcoma," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 75, no. 6, pp. 761–765, 1981. 35

[120] D. Burkitt, "Burkitt's lymphoma outside the known endemic areas of africa and new guinea," *International journal of cancer*, vol. 2, no. 6, pp. 562–565, 1967. 35

[121] I. T. Magrath, "African burkitt's lymphoma. history, biology, clinical features, and treatment.," *The American journal of pediatric hematology/oncology*, vol. 13, no. 2, pp. 222–246, 1991. 35

[122] D. Dominguez-Sola, C. Y. Ying, C. Grandori, L. Ruggiero, B. Chen, M. Li, D. A. Galloway, W. Gu, J. Gautier, and R. Dalla-Favera, "Non-transcriptional control of DNA replication by c-myc," *Nature*, vol. 448, no. 7152, p. 445, 2007. 35, 108

[123] D. Dominguez-Sola and J. Gautier, "Myc and the control of DNA replication," *Cold Spring Harbor perspectives in medicine*, vol. 4, no. 6, p. a014423, 2014. 35

[124] R. Pulvertaft, "A study of malignant tumours in nigeria by short-term tissue culture," *Journal of Clinical Pathology*, vol. 18, no. 3, p. 261, 1965. 35

[125] G. Q. Daley, R. A. Van Etten, and D. Baltimore, "Induction of chronic myelogenous leukemia in mice by the p210bcr/abl gene of the philadelphia chromosome," *Science*, vol. 247, no. 4944, pp. 824–830, 1990. 35, 47, 122

[126] B. Laperrousaz, S. Jeanpierre, K. Sagorny, T. Voeltzel, S. Ramas, B. Kaniewski, M. Ffrench, S. Salesse, F. E. Nicolini, and V. Maguer-Satta, "Primitive CML cell expansion relies on abnormal levels of BMPs provided by the niche and bmprib overexpression," *Blood*, pp. blood–2013, 2013. 35

[127] C. B. Lozzio and B. B. Lozzio, "Human chronic myelogenous leukemia cell-line with positive philadelphia chromosome," *Blood*, vol. 45, no. 3, pp. 321–334, 1975. 36, 109

[128] L. Duret, D. Mouchiroud, and C. Gautier, "Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores," *Journal of Molecular Evolution*, vol. 40, no. 3, pp. 308–317, 1995. 36

[129] M. Costantini, R. Cammarano, and G. Bernardi, "The evolution of isochore patterns in vertebrate genomes," *BMC genomics*, vol. 10, no. 1, p. 146, 2009. 36

[130] G. Bernardi, "Misunderstandings about isochores. part 1," *Gene*, vol. 276, no. 1-2, pp. 3–13, 2001. 36, 67

[131] I. H. G. S. Consortium *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, p. 860, 2001. 36

[132] L. Zhang, J.-G. Chen, and Q. Zhao, "Regulatory roles of Alu transcript on gene expression," *Experimental cell research*, vol. 338, no. 1, pp. 113–118, 2015. 36

[133] M. Costantini and G. Bernardi, "Replication timing, chromosomal bands, and isochores," *Proceedings of the National Academy of Sciences*, vol. 105, no. 9, pp. 3433–3437, 2008. 36

[134] R. E. Thurman, N. Day, W. S. Noble, and J. A. Stamatoyannopoulos, "Identification of higher-order functional domains in the human ENCODE regions," *Genome research*, vol. 17, no. 6, pp. 917–927, 2007. 37

[135] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with tophat and cufflinks," *Nature protocols*, vol. 7, no. 3, p. 562, 2012. 38, 42

[136] G. Saporta, *Probabilités, analyse des données et statistique*. Editions Technip, 2006. 39, 40

[137] A. J. Izenman, "Modern multivariate statistical techniques," *Regression, classification and manifold learning*, 2008. 39

[138] M. Bruynooghe, "Classification ascendante hiérarchique des grands ensembles de données: un algorithme rapide fondé sur la construction des voisinages réductibles," *Les cahiers de l'analyse de données*, vol. 3, pp. 7–33, 1978. 40

[139] R. Sibson, "SLINK: an optimally efficient algorithm for the single-link cluster method," *The computer journal*, vol. 16, no. 1, pp. 30–34, 1973. 40

[140] D. Defays, "An efficient algorithm for a complete link method," *The Computer Journal*, vol. 20, no. 4, pp. 364–366, 1977. 40

[141] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome biology*, vol. 15, no. 12, p. 550, 2014. 42

[142] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009. 43

[143] N. Yabuki, H. Terashima, and K. Kitada, "Mapping of early firing origins on a replication profile of budding yeast," *Genes to Cells*, vol. 7, no. 8, pp. 781–789, 2002. 46

[144] W. Feng, D. Collingwood, M. E. Boeck, L. A. Fox, G. M. Alvino, W. L. Fangman, M. K. Raghuraman, and B. J. Brewer, "Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication," *Nature cell biology*, vol. 8, no. 2, p. 148, 2006.

[145] M. D. Sekedat, D. Fenyö, R. S. Rogers, A. J. Tackett, J. D. Aitchison, and B. T. Chait, "GINS motion reveals replication fork progression is remarkably uniform throughout the yeast genome," *Molecular systems biology*, vol. 6, no. 1, 2010. 46

[146] C. Cayrou, B. Ballester, I. Peiffer, R. Fenouil, P. Coulombe, J.-C. Andrau, J. van Helden, and M. Méchali, "The chromatin environment shapes DNA replication origin organization and defines origin classes," *Genome research*, 2015. 46

[147] R. Mukhopadhyay, J. Lajugie, N. Fourel, A. Selzer, M. Schizas, B. Bartholdy, J. Mar, C. M. Lin, M. M. Martin, M. Ryan, *et al.*, "Allele-specific genome-wide profiling in human primary erythroblasts reveal replication program organization," *PLoS genetics*, vol. 10, no. 5, p. e1004319, 2014. 46

[148] K. Sugimoto, T. Okazaki, and R. Okazaki, "Mechanism of DNA chain growth, ii. accumulation of newly synthesized short chains in E. coli infected with ligase-defective T4 phages," *Proceedings of the National Academy of Sciences*, vol. 60, no. 4, pp. 1356–1362, 1968. 46

[149] R. Okazaki, T. Okazaki, K. Sakabe, K. Sugimoto, and A. Sugino, "Mechanism of DNA chain growth. i. possible discontinuity and unusual secondary structure of newly synthesized chains," *Proceedings of the National Academy of Sciences*, vol. 59, no. 2, pp. 598–605, 1968. 46

[150] B. Alberts, D. Bray, J. Lewis, M. Minkowski, and G. Rolland, "Biologie moléculaire de la cellule," 1990. 46

[151] S. R. McGuffee, D. J. Smith, and I. Whitehouse, "Quantitative, genome-wide analysis of eukaryotic replication initiation and termination," *Molecular cell*, vol. 50, no. 1, pp. 123–135, 2013. 46, 47

[152] D. J. Smith and I. Whitehouse, "Intrinsic coupling of lagging-strand synthesis to chromatin assembly," *Nature*, vol. 483, no. 7390, p. 434, 2012. 46, 47

[153] X. Wu, *Determination of DNA replication program changes between cancer and normal cells by sequencing of Okazaki fragments*. PhD thesis, Paris Sciences et Lettres, 2016. 47, 71

[154] T. Ryba, D. Battaglia, B. H. Chang, J. W. Shirley, Q. Buckley, B. D. Pope, M. Devidas, B. J. Druker, and D. M. Gilbert, "Abnormal developmental control of replication timing domains in pediatric acute lymphoblastic leukemia," *Genome research*, pp. gr–138511, 2012. 50, 83, 105

[155] G. Bernardi, "Isochores and the evolutionary genomics of vertebrates," *Gene*, vol. 241, no. 1, pp. 3–17, 2000. 56, 65, 67

[156] I. Hiratani, A. Leskovar, and D. M. Gilbert, "Differentiation-induced replication-timing changes are restricted to AT-rich/long interspersed nuclear element (line)-rich isochores," *Proceedings of the National Academy of Sciences*, vol. 101, no. 48, pp. 16861–16866, 2004. 75, 123

[157] S. Takahashi, H. Miura, T. Shibata, K. Nagao, K. Okumura, M. Ogata, C. Obuse, S.-i. Takebayashi, and I. Hiratani, "Genome-wide stability of the DNA replication program in single mammalian cells," *Nature genetics*, p. 1, 2019. 75

[158] A. Aguilera and T. García-Muse, "Causes of genome instability," *Annual review of genetics*, vol. 47, pp. 1–32, 2013. 77

[159] W. F. Holmes, C. D. Braastad, P. Mitra, C. Hampe, D. Doenecke, W. Albig, J. L. Stein, A. J. Van Wijnen, and G. S. Stein, "Coordinate control and selective expression of the full complement of replication-dependent histone H4 genes in normal and cancer cells," *Journal of Biological Chemistry*, vol. 280, no. 45, pp. 37400–37407, 2005. 77

[160] N. Kim and S. Jinks-Robertson, "Transcription as a source of genome instability," *Nature Reviews Genetics*, vol. 13, no. 3, p. 204, 2012. 77

[161] S. Tuduri, L. Crabbé, C. Conti, H. Tourrière, H. Holtgreve-Grez, A. Jauch, V. Pantesco, J. De Vos, A. Thomas, C. Theillet, *et al.*, "Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription," *Nature cell biology*, vol. 11, no. 11, p. 1315, 2009. 77

[162] G. Holmquist, M. Gray, T. Porter, and J. Jordan, "Characterization of giemsa dark-and light-band DNA," *Cell*, vol. 31, no. 1, pp. 121–129, 1982. 78

[163] D. M. Gilbert, "Replication timing and transcriptional control: beyond cause and effect," *Current opinion in cell biology*, vol. 14, no. 3, pp. 377–383, 2002. 78

[164] D. E. Comings, "Mechanisms of chromosome banding and implications for chromosome structure," *Annual review of genetics*, vol. 12, no. 1, pp. 25–46, 1978.

[165] T. Ryba, D. Battaglia, B. D. Pope, I. Hiratani, and D. M. Gilbert, "Genome-scale analysis of replication timing: from bench to bioinformatics," *Nature protocols*, vol. 6, no. 6, p. 870, 2011. 78

[166] Y. Jeon, S. Bekiranov, N. Karnani, P. Kapranov, S. Ghosh, D. MacAlpine, C. Lee, D. S. Hwang, T. R. Gingeras, and A. Dutta, "Temporal profile of replication of human chromosomes," *Proceedings of the National Academy of Sciences*, vol. 102, no. 18, pp. 6419–6424, 2005. 78

[167] M. Laskowski Sr, "12 deoxyribonuclease i," in *The enzymes*, vol. 4, pp. 289–311, Elsevier, 1971. 78

[168] A. D. Donaldson, "Shaping time: chromatin structure and the DNA replication programme," *Trends in genetics*, vol. 21, no. 8, pp. 444–449, 2005. 78

[169] I. Hiratani and S. Takahashi, "DNA replication timing enters the single-cell era," *Genes*, vol. 10, no. 3, p. 221, 2019. 78

[170] Q. Du, S. A. Bert, N. J. Armstrong, C. E. Caldon, J. Z. Song, S. S. Nair, C. M. Gould, P.-L. Luu, T. Peters, A. Khoury, *et al.*, "Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer," *Nature communications*, vol. 10, no. 1, p. 416, 2019. 78

[171] Y.-H. Chen, S. Keegan, M. Kahli, P. Tonzi, D. Fenyö, T. T. Huang, and D. J. Smith, "Transcription shapes DNA replication initiation and termination in human cells," *Nature structural & molecular biology*, vol. 26, no. 1, p. 67, 2019. 78, 99, 105, 123, 124

[172] F. Chedin, "R-loop structures are novel, conserved, functional elements in mammalian genomes," 2015. 92

[173] L. A. Sanz, S. R. Hartono, Y. W. Lim, S. Steyaert, A. Rajpurkar, P. A. Ginno, X. Xu, and F. Chédin, "Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals," *Molecular cell*, vol. 63, no. 1, pp. 167–178, 2016. 92

[174] R. Lombraña, R. Almeida, A. Álvarez, and M. Gómez, "R-loops and initiation of DNA replication in human cells: a missing link?," *Frontiers in genetics*, vol. 6, p. 158, 2015. 94

[175] J. Sequeira-Mendes, R. Díaz-Uriarte, A. Apedaile, D. Huntley, N. Brockdorff, and M. Gómez, "Transcription initiation activity sets replication origin efficiency in mammalian cells," *PLoS genetics*, vol. 5, no. 4, p. e1000446, 2009. 94

[176] E. J. Wagner and P. B. Carpenter, "Understanding the language of lys36 methylation at histone h3," *Nature reviews Molecular cell biology*, vol. 13, no. 2, p. 115, 2012. 94

[177] F. Comoglio, T. Schlumpf, V. Schmid, R. Rohs, C. Beisel, and R. Paro, "High-resolution profiling of drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins," *Cell reports*, vol. 11, no. 5, pp. 821–834, 2015. 94

[178] J. Braunstein, D. ScHuLZE, T. DelGiudice, A. Furst, and C. Schildkraut, "The temporal order of replication of murine immunoglobulin heavy chain constant region sequences corresponds to their linear order in the genome," *Nucleic acids research*, vol. 10, no. 21, pp. 6887–6902, 1982. 104

[179] D. M. Gilbert, "Temporal order of replication of xenopus laevis 5s ribosomal rna genes in somatic cells," *Proceedings of the National Academy of Sciences*, vol. 83, no. 9, pp. 2924–2928, 1986.

[180] M. A. Goldman, G. P. Holmquist, M. C. Gray, L. A. Caston, and A. Nag, "Replication timing of genes and middle repetitive sequences," *Science*, vol. 224, no. 4650, pp. 686–692, 1984.

[181] R. S. Hansen, T. K. Canfield, M. M. Lamb, S. M. Gartler, and C. D. Laird, "Association of fragile x syndrome with delayed replication of the fmr1 gene," *Cell*, vol. 73, no. 7, pp. 1403–1409, 1993.

[182] M. Schmidt and B. R. Migeon, "Asynchronous replication of homologous loci on human active and inactive x chromosomes.," *Proceedings of the National Academy of Sciences*, vol. 87, no. 10, pp. 3685–3689, 1990. 104

[183] G. P. Holmquist, "Role of replication time in the control of tissue-specific gene expression.," *American journal of human genetics*, vol. 40, no. 2, p. 151, 1987. 104

[184] S. Selig, K. Okumura, D. Ward, and H. Cedar, "Delineation of DNA replication time zones by fluorescence in situ hybridization.," *The EMBO journal*, vol. 11, no. 3, pp. 1217–1225, 1992.

[185] D. M. MacAlpine and S. P. Bell, "A genomic view of eukaryotic DNA replication," *Chromosome Research*, vol. 13, no. 3, pp. 309–326, 2005. 104

[186] V. Dileep, K. A. Wilson, C. Marchal, X. Lyu, P. A. Zhao, B. Li, A. Poulet, D. A. Bartlett, J. C. Rivera-Mulia, Z. S. Qin, *et al.*, "Rapid irreversible transcriptional reprogramming in human stem cells accompanied by discordance between replication timing and chromatin compartment," *Stem cell reports*, 2019. 104, 123, 124

[187] V. Hassan-Zadeh, S. Chilaka, J.-C. Cadoret, M. K.-W. Ma, N. Boggetto, A. G. West, and M.-N. Prioleau, "Usf binding sequences from the hs4 insulator element impose early replication timing on a vertebrate replicator," *PLosS biology*, vol. 10, no. 3, p. e1001277, 2012. 104

[188] J. B. Kim, R. Stein, and M. J. O'hare, "Three-dimensional in vitro tissue culture models of breast cancer—a review," *Breast cancer research and treatment*, vol. 85, no. 3, pp. 281–291, 2004. 107

[189] A. Ertel, A. Verghese, S. W. Byers, M. Ochs, and A. Tozeren, "Pathway-specific differences between tumor cell lines and normal and tumor tissue cells," *Molecular cancer*, vol. 5, no. 1, p. 1, 2006. 107

[190] D. T. Ross and C. M. Perou, "A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines," *Disease markers*, vol. 17, no. 2, pp. 99–109, 2001. 107

[191] W. D. Stein, T. Litman, T. Fojo, and S. E. Bates, "A serial analysis of gene expression (SAGE) database analysis of chemosensitivity: comparing solid tumors with cell lines and comparing solid tumors from different tissue origins," *Cancer research*, vol. 64, no. 8, pp. 2805–2816, 2004. 107

[192] D. Hadjadj, T. Denecker, C. Maric, F. Fauchereau, G. Baldacci, and J.-C. Cadoret, "Characterization of the replication timing program of 6 human model cell lines," *Genomics data*, vol. 9, pp. 113–117, 2016. 107

[193] D. Zink, A. H. Fischer, and J. A. Nickerson, "Nuclear structure in cancer cells," *Nature reviews cancer*, vol. 4, no. 9, p. 677, 2004. 108

[194] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *science*, vol. 339, no. 6127, pp. 1546–1558, 2013. 108

[195] M. K. Zeman and K. A. Cimprich, "Causes and consequences of replication stress," *Nature cell biology*, vol. 16, no. 1, p. 2, 2014. 108

[196] M. Berti and A. Vindigni, "Replication stress: getting back on track," *Nature structural & molecular biology*, vol. 23, no. 2, p. 103, 2016. 108

[197] A. Mazouzi, G. Velimezi, and J. I. Loizou, "DNA replication stress: causes, resolution and disease," *Experimental cell research*, vol. 329, no. 1, pp. 85–93, 2014. 108

[198] A. Tubbs and A. Nussenzweig, "Endogenous DNA damage as a source of genomic instability in cancer," *Cell*, vol. 168, no. 4, pp. 644–656, 2017. 108

[199] J. Bartkova, Z. Hořejší, K. Koed, A. Krämer, F. Tort, K. Zieger, P. Guldberg, M. Sehested, J. M. Nesland, C. Lukas, *et al.*, "DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis," *Nature*, vol. 434, no. 7035, p. 864, 2005. 108

[200] J. Bartkova, N. Rezaei, M. Liontos, P. Karakaidos, D. Kletsas, N. Issaeva, L.-V. F. Vassiliou, E. Kolettas, K. Niforou, V. C. Zoumpourlis, *et al.*, "Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints," *Nature*, vol. 444, no. 7119, p. 633, 2006.

[201] V. G. Gorgoulis, L.-V. F. Vassiliou, P. Karakaidos, P. Zacharatos, A. Kotsinas, T. Liloglou, M. Venere, R. A. DiTullio Jr, N. G. Kastrinakis, B. Levy, *et al.*, "Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions," *Nature*, vol. 434, no. 7035, p. 907, 2005.

[202] T. D. Halazonetis, V. G. Gorgoulis, and J. Bartek, "An oncogene-induced DNA damage model for cancer development," *science*, vol. 319, no. 5868, pp. 1352–1355, 2008.

[203] H. Gaillard, T. Garcia-Muse, and A. Aguilera, "Replication stress and cancer," *Nature Reviews Cancer*, vol. 15, no. 5, p. 276, 2015. 108

[204] S. V. Srinivasan, D. Dominguez-Sola, L. C. Wang, O. Hyrien, and J. Gautier, "Cdc45 is a critical effector of myc-dependent DNA replication stress," *Cell reports*, vol. 3, no. 5, pp. 1629–1639, 2013. 108

[205] P. Kotsantis, L. M. Silva, S. Irmscher, R. M. Jones, L. Folkes, N. Gromak, and E. Petermann, "Increased global transcription activity as a mechanism of replication stress in cancer," *Nature communications*, vol. 7, p. 13087, 2016. 108

[206] M. Macheret and T. D. Halazonetis, "Intragenic origins due to short g1 phases underlie oncogene-induced DNA replication stress," *Nature*, vol. 555, no. 7694, p. 112, 2018. 109, 120

[207] M. Hennion, J.-M. Arbona, C. Cruaud, F. Proux, B. Le Tallec, E. Novikova, S. Engelen, A. Lemainque, B. Audit, and O. Hyrien, "Mapping DNA replication with nanopore sequencing," *bioRxiv*, p. 426858, 2018. 109

[208] C. A. Mueller, M. A. Boemo, P. Spingardi, B. Kessler, S. Kriaucionis, J. T. Simpson, and C. A. Nieduszynski, "Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads," *BioRxiv*, p. 442814, 2018. 109

[209] R. Kurzrock, J. U. Gutterman, and M. Talpaz, "The molecular genetics of philadelphia chromosome–positive leukemias," *New England Journal of Medicine*, vol. 319, no. 15, pp. 990–998, 1988. 109

[210] B. Gómez-Escoda and P.-Y. J. Wu, "The organization of genome duplication is a critical determinant of the landscape of genome maintenance," *Genome research*, vol. 28, no. 8, pp. 1179–1192, 2018. 111

[211] N. McGranahan and C. Swanton, "Clonal heterogeneity and tumor evolution: past, present, and the future," *Cell*, vol. 168, no. 4, pp. 613–628, 2017. 111

[212] S. Negrini, V. G. Gorgoulis, and T. D. Halazonetis, "Genomic instability—an evolving hallmark of cancer," *Nature reviews Molecular cell biology*, vol. 11, no. 3, p. 220, 2010. 111

[213] J. D. Rowley, M. M. Le Beau, and T. H. Rabbitts, *Chromosomal translocations and genome rearrangements in cancer*. Springer, 2015. 120

[214] J.-H. Zhang, Y.-L. He, R. Zhu, W. Du, and J.-H. Xiao, "Deregulated expression of Cdc6 as BCR/ABL-dependent survival factor in chronic myeloid leukemia cells," *Tumor Biology*, vol. 39, no. 6, p. 1010428317713394, 2017. 120

[215] M. Takagi, M. Sato, J. Piao, S. Miyamoto, T. Isoda, M. Kitagawa, H. Honda, and S. Mizutani, "Atm-dependent DNA damage-response pathway as a determinant in chronic myelogenous leukemia," *DNA repair*, vol. 12, no. 7, pp. 500–507, 2013. 120, 121

[216] S. Hamperl, M. J. Bocek, J. C. Saldivar, T. Swigut, and K. A. Cimprich, "Transcription-replication conflict orientation modulates R-loop levels and activates distinct DNA damage responses," *Cell*, vol. 170, no. 4, pp. 774–786, 2017. 123

[217] P. L. T. Tran, T. J. Pohl, C.-F. Chen, A. Chan, S. Pott, and V. A. Zakian, "PIF1 family DNA helicases suppress R-loop mediated genome instability at tRNA genes," *Nature communications*, vol. 8, p. 15025, 2017. 123

[218] J. C. Rivera-Mulia, Q. Buckley, T. Sasaki, J. Zimmerman, R. A. Didier, K. Nazor, J. F. Loring, Z. Lian, S. Weissman, A. J. Robins, *et al.*, "Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells," *Genome research*, vol. 25, no. 8, pp. 1091–1103, 2015. 123

[219] R. R. Williams, V. Azuara, P. Perry, S. Sauer, M. Dvorkina, H. Jørgensen, J. Roix, P. McQueen, T. Misteli, M. Merkenschlager, *et al.*, "Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus," *Journal of cell science*, vol. 119, no. 1, pp. 132–140, 2006.

[220] P. Perry, S. Sauer, N. Billon, W. D. Richardson, M. Spivakov, G. Warnes, F. J. Livesey, M. Merkenschlager, A. G. Fisher, and V. Azuara, "A dynamic switch in the replication timing of key regulator genes in embryonic stem cells upon neural induction," *Cell Cycle*, vol. 3, no. 12, pp. 1619–1624, 2004. 123

[221] 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, *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *science*, vol. 326, no. 5950, pp. 289–293, 2009. 125

[222] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, *et al.*, "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions," *Nature*, vol. 453, no. 7197, p. 948, 2008. 125

[223] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, *et al.*, "Topologically associating domains are stable units of replication-timing regulation," *Nature*, vol. 515, no. 7527, p. 402, 2014. 125