Cell type-specific dynamics of DNA replication timing in primary cancer whole-genome sequences

Inaugural Dissertation

zur

Erlangung des Doktorgrades
Dr. nat. med.

der Medizinischen Fakultät
und
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Tsun-Po Yang

aus Taipei, Taiwan

Copy-Star, Köln

2022

Betreuer:	Prof. Dr. Martin Peifer
Referenten:	Prof. Dr. Matthias Fischer
	Prof. Dr. Björn Schumacher

Datum der mündlichen Prüfung: Mittwoch, 9. Februar 2022

Summary

Although the whole-genome tumour DNA from human primary cancers have been routinely sequenced in recent large-scale cancer sequencing studies to identify somatic mutations, little is known about their roles in DNA replication timing. However, mechanisms underlying this timing programme are also implicated in transcriptional activities and the mutational landscape of the cancer genome in a cell type-specific manner. Understanding where exactly, and how differently, DNA replication is initiated and terminated across different cancer genomes is of fundamental importance and will help to understand the cellular plasticity that give rise to cancer and help cancer cells survive during cancer cell proliferation.

In this thesis, I propose to fully explore the entire primary cancer whole-genome sequences, and hypothesise that they may provide a snapshot in time, in space, and in specific cell type of the tumour replication timing programme.

In Chapter 2, I measure the proportion of S phase cells present in a primary tumour using whole-genome sequencing (WGS) data, and use it to separate tumour samples based on their cell cycle status, referred to as *in silico* sample sorting method. Upon *in silico* sorting of primary tumour samples, in Chapter 3, I adapt the S to G1 read depth ratio approach, and apply it to directly profile the tumour replication timing (RT) from 256 cancer whole genomes of three tumour types. Finally, I demonstrate that the temporal dynamics of tumour replication timing is preserved in closely related normal tissues, as well as in lineage-specific cancer cell lines, suggesting the cellular plasticity of the timing programme captured by my direct profiling approach.

Furthermore, in Chapter 4, I introduce a novel resampling-based replication fork directionality (RFD) methodology to model the stochastic but symmetrical nature of bi-directional replication, and use it to simultaneously fine map the replication origins and termini at 1 kb (kilobase) resolution using the same primary cancer WGS data. Unexpectedly, I find that the genome-wide distribution of termination events is tightly coordinated with the initiation activities in both the normal and cancer genomes, which has not been previously reported in humans using directional sequencing of Okazaki fragments (OK-seq) *in vitro*. However, the distribution of my reconstructed RFD

domains suggests that replication termini are determined by, and located between two activating origin firings in the human genome, which is consistent with the consensus notion widely reported in yeast.

Nevertheless, I find that the spatial landscapes of my reconstructed RFD domains are also preserved in closely related normal tissues and lineage-specific cancer cell lines, in line with the cellular plasticity of tumour RT shown in Chapter 3. Furthermore, in Chapter 5, I demonstrate that my reconstructed RFD domains are significantly coupled with the transcriptional activities across three tumour types, thus providing strong support for a *bona fide* mapping on fork initiation, progression, and termination by my novel resampling-based RFD methodology.

Altogether, my novel *in silico* framework allows one to assess the tumour RT and RFD domains directly using primary cancer whole-genome sequences without the need for *in vitro* sorting procedures, and therefore opens up opportunities for the routinely performed WGS data from the broader cancer genomics community. The cell type specificity of the tumour timing programme recapitulated by my novel *in silico* framework also adds a new spatiotemporal perspective to the three-dimensional cancer genome, thus could provide new insights into the identification of potential cancer targets that is topologically preserved in specific cancer types.

Zusammenfassung

Große Tumor DNA Sequenzierungsstudien zielen auf die umfassende Beschreibung und Funktion von somatischen Mutationen in Primärtumoren ab, dennoch ist wenig über den Einfluss von „DNA Replication Timing“ hierauf bekannt. Mechanismen, die dem Timing Programm zu Grunde liegen, stehen in Verbindung mit den Zelltyp spezifischen transkriptionellen Aktivitäten und dem Mutationsspektrum in Krebsgenomen. Ein besseres Verständnis wo und wie verschiedenartig DNA Replikation in Krebsgenomen initiiert oder terminiert ist würde dazu beitragen die zellulare Plastizität, welche den Krebszellen eine weitere Proliferation ermöglicht, zu verstehen.

In der vorliegenden Arbeit, werde ich Genomsequenzierungen (WGS) von Primärtumoren untersuchen mit dem Ziel einen zeitlichen, räumlichen und Zelltyp bzw. Krebstyp spezifischen Einblick in das Replication Timing Programmes zu bekommen.

Hierzu werde ich in Kapitel 2 ein Maß für den Anteil der Zellen, die sich in der S-Phase befinden, aus WGS Daten ableiten um somit einzelne Tumorproben nach ihrem Zellzyklus zu klassifizieren. Diese Methode bezeichne ich als „*In Silico* Sorting“. Um die Tumor Replication Timing (RT) direkt aus WGS Daten zu bestimmen, werde ich in Kapitel 3 meine *In Silico* Sorting Methode benutzen (in Anlehnung an das S- zu G1-Zellzyklusphasen Verhältnis bei FACS sortierten WGS Daten von Zelllinien) und auf 256 WGS Datensätzen von drei Tumorentitäten anwenden. Hiermit werde ich zeigen, dass die zeitliche Dynamik des Tumor Replication Timing Programmes in dem jeweilig korrespondierenden Normalgewebe und in Zelllinien des gleichen Tumortyps erhalten bleibt, was suggeriert, dass die zelluläre Plastizität des Timing Programmes durch meinen Ansatz abgebildet wird.

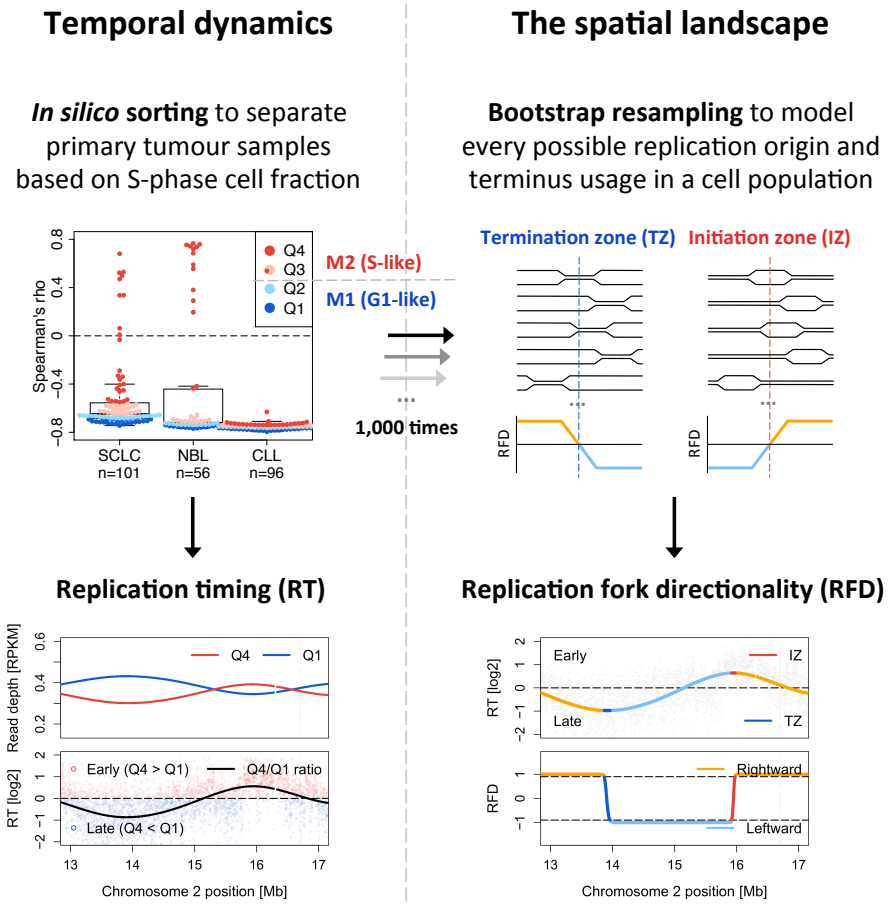
Ferner werde ich in Kapitel 4 eine neue Methode zur „Replication Fork Directionality” (RFD) vorstellen, mit der sich die stochastische aber symmetrische Bidirektionale DNA Replikation modellieren lässt um somit Replikationsursprünge und deren Terminierung in einer 1 kbp (Kilobasenpaare) Auflösung bestimmen zu können. Dies ergab, dass die Genomweite Verteilung Replikationsursprünge und der Termini sowohl im Normal- wie auch im Tumorgewebe stark miteinander verknüpft sind. Eine alternative Methode, die RFD mittels einer gerichteten Sequenzierung von Okazaki Fragmenten *in vitro*

bestimmt kommt allerdings zu einem anderen Ergebnis. In Gegensatz liegt meine Bestimmung der RFD nahe, dass sich Replikationsursprünge und Terminationen in geordneter weise abwechseln und ist somit im Einklang mit dem, was in Hefegenomen gefunden wurde.

Ähnlich zu den RT Profilen aus Kapitel 3, konnte ich zeigen, dass räumliche RFD Profile von Tumoren denen derer korrespondierenden Normalgeweben oder Zelllinien des gleichen Tumortyps am ähnlichsten sind. Darüber hinaus konnte ich in Kapitel 5 zeigen, dass meine rekonstruierten RFD Profile signifikant mit transkriptionellen Aktivitäten in den drei Tumorentitäten korreliert ist. Diese Ergebnisse weisen stark auf die Validität meiner RFD Methode zur Bestimmung der Initiation, Progression und Terminierung der Replikationsgabel.

Zusammenfassend erlaubt die hier vorgestellte *in silico* Methodik Replication Timing und Replication Fork Directionality direkt aus WGS von Primärtumoren zu bestimmen und steht folglich einer breiteren Gruppe von Krebsgenomforschern zur Verfügung. Da meine *in silico* Methoden zur Bestimmung des Replication Timing Programmes Zelltyp spezifisch sind, könnte diese neue Erkenntnisse über die Identifikation von potentiell neuen Ansatzpunkten für eine zielgerichtete Therapie in topologisch konservierten Regionen liefern.

Graphical summary



Contents

Chapter 1

Introduction

DNA sequences that underwent somatic alterations across primary cancer samples have been comprehensively studied in recent large-scale sequencing projects over the past decades (Yates and Campbell 2012; Campbell et al. 2020). However, the vast majority of accurately replicated sequences of the same primary cancer genomes have not been widely explored, and little is known about their roles in cancers. Indeed, cancers are the consequences as well as the survivors of genomic aberrations through an extremely robust DNA replication timing programme during cancer cell proliferation (Marchal, Sima, and Gilbert 2019). Moreover, the mechanisms underlying this timing programme are also implicated in transcriptional activity (Lawrence et al. 2013; Pourkarimi et al. 2016; Chen et al. 2019), developmental regulation (Ryba et al. 2010; Pope et al. 2014), and the mutational landscape of the cancer genome (Lawrence et al. 2013; Polak et al. 2015; Haradhvala et al. 2016; Du et al. 2019; Li et al. 2019). Therefore, in this thesis I propose to analyse cancer genome in a complementary approach by looking at 'the other side of the same coin' through the analysis of accurately replicated sequences of the cancer genomes. A thorough understanding of the replication timing programme in cancers is essentially important to fully understand the cellular mechanisms that give rise to cancer and help cancer cells survive during the course of cancer development.

1.1 The human cancer genome

1.1.1 Somatic alterations in cancer

Replication of the entire human genome is required during the synthesis phase of each round of cell cycle. Failure to accurately replicate or repair DNA sequences leads to damage and errors in the genome over time (Tubbs and Nussenzweig 2017). One of the current theories of cancer development suggests that accumulation of such somatic alterations in the DNA may provide selective advantage to the cancerous cells throughout their evolutionary lifetime (Jolly and Van Loo 2018). With recent technological advances in sequencing, several types of somatic alterations have been well identified across primary cancer specimens, including point mutations, copy number alterations, and genome rearrangements. Therefore, to date, the identification of cancer driver genes has been traditionally focused on DNA sequences that underwent somatic alterations across primary cancer samples, such as comprehensive studies from the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) communities.

1.1.2 Survivorship bias

However, only around 4-5 of the coding and non-coding somatic alterations found in tumour cells are under positive selection and drive cancer development (Martincorena et al. 2017; Campbell et al. 2020). In contrast, the majority of the alterations are known to be passengers and have no or neural selection effects to the cancer cells (Dietlein et al. 2018). Therefore, it is undeniable that cancer develops also as a survivor of genomic aberrations under neutral selection (Nik-Zainal and Hall 2019). Furthermore, recent cancer dependency screenings have begun to address a fundamental bias in current cancer driver gene approaches; that is, there are also genes that are not mutated but are nonetheless essential for proliferation and survival of cancer cells (Hahn et al. 2021; Malone et al. 2021).

Indeed, if we assume that mutations on essential genes required for cell proliferation and survival are deleterious, then it is reasonable to presume that these mutations resulted in cell death and did not become a data point when we sequence the tumour samples. This is known as survivorship bias (Mangel and Samaniego 1984; Pyatnitskiy

et al. 2015). Toward this end, essential genes are expected to contain fewer mutations, whereas non-essential genes are expected to contain higher mutational burdens. It is therefore important to study DNA sequences that underwent somatic alterations, but also the sequences that accurately replicated in the same cancer genome.

1.1.3 DNA replication timing shapes the mutational landscape of the cancer genome

In recent large-scale cancer sequencing studies, the distribution and patterns of somatic mutations have been identified to vary among different parts of chromosomes and different cancer types (Lawrence et al. 2013; Haradhvala et al. 2016). Variations in DNA replication timing, chromatin compartments, and gene expression levels have all been proposed as major determinants underlying the mutational processes that constantly change the somatic genome (Polak et al. 2015; Haradhvala et al. 2016). Collectively, chromatin accessibility and replication timing account for up to 86% of the variation in mutation densities along cancer genomes (Polak et al. 2015). Furthermore, replication timing and epigenome remodelling have also been associated with the nature of chromosomal rearrangements in cancers (Du et al. 2019). Therefore, understanding where exactly, and how differently, DNA replication is initiated and terminated across different cell types is an important prerequisite for studying the mutational strand asymmetries (Haradhvala et al. 2016), for modelling the cancer evolution (Nik-Zainal et al. 2012), and for identifying cancer target genes (Lawrence et al. 2013).

1.2 DNA replication

1.2.1 The replication timing programme

During the synthesis phase of the cell cycle, human chromosomes are not replicated linearly from telomere to telomere (Miga et al. 2020). Instead, DNA replication stochastically initiates and terminates from different parts of the chromosome at different times (Hawkins et al. 2013; Petryk et al. 2016; Marchal, Sima, and Gilbert

2019). This process, despite its stochasticity, is strictly orchestrated in a temporal and spatial order known as the DNA replication-timing programme in higher eukaryotes (Koren et al. 2014; Pope et al. 2014; Fragkos et al. 2015). The temporal order of DNA replication, from early to late, is highly associated with the spatial landscape of chromatin compartments, from open to close, respectively (Pope et al. 2014; Petryk et al. 2016; Du et al. 2019). Therefore, the timing programme is known to be topologically preserved in distinct cell types in humans (Pope et al. 2014). This also indicates that the replication domains, including origins and termini, are among the most cell type-specific genomic properties in the human genome (Rhind and Gilbert 2013; Haradhvala et al. 2016).

1.2.2 Profiling of replication timing (RT) in the cancer genome

By comparing the difference between replicated and un-replicated DNA, whole-genome sequencing (WGS) of DNA has been developed to profile replication timing (RT). Currently, WGS of DNA from flow-sorted cancer cell lines can provide a snapshot in time and space of the timing programme from a specific tumour type (Woodfine et al. 2004; Hansen et al. 2010; Takahashi et al. 2019). However, not all cell or cancer types can be established as proliferative and immortalised cell lines (Masters 2000; Sasaki et al. 2017), and appropriate cell lines are limited for the cells of origin of distinct tumours (Sutherland et al. 2011; Joshy George et al. 2016). Therefore, to date, a single reference timing profile derived from flow-sorted lymphoblastoid cell lines has been widely used in the cancer genomics community to study replication timing in multiple types of cancers (Koren et al. 2014; Polak et al. 2015; Haradhvala et al. 2016; Li et al. 2019). This major one-size-fits-all limitation has hampered the possibility to investigate the most cell type-specific genomic properties, i.e. replication initiation and termination domains, in the human cancer genome (Ryba et al. 2010; Haradhvala et al. 2016).

Previously, an alternative method has shown that when a cell population contains appropriate high proportion of S phase cells, its sequencing read depth patterns would become highly correlated with the reference timing profile (Koren et al. 2014; Marchal et al. 2018). Despite these megabase-scale read depth patterns are also reportedly to be individual- or sequence-specific (Ryba et al. 2012; Koren et al. 2014; Sasaki et al.

2017), the authors further proposed to directly use them as *de facto* replication timing profiles. Nevertheless, methods for direct reconstruction of cell type-specific replication timing from primary bulk tumours without the need for *in vitro* flow sorting procedures have not been previously reported.

1.2.3 Replication initiation and termination

Despite intensive studies, the usage and timing of replication origins and termini are among the least understood genomic properties of the human genome (Petryk et al. 2016; Chen et al. 2019). This is because the heterogeneity of the timing programme is two-fold; that is, not only the usage of replication origins are different between cell types, even cells of the same type use different origins to initiate replication in each cell cycle (Hawkins et al. 2013; Audit et al. 2013; Bartholdy et al. 2015; Takahashi et al. 2019). Therefore, in principle, rather than aiming to identify the individual initiation and termination events at single-cell or single-molecular level (Chen et al. 2019; Takahashi et al. 2019), current replication profiling and domain mapping methods focus on measurements of the central tendency, i.e. average replication timing, in a cell population of the same type (Woodfine et al. 2004; Audit et al. 2013; Marchal et al. 2018).

Although timing profile itself can already be used to infer the position of replication origins and termini along its timing transition slopes using algorithms (Audit et al. 2013; Zhao, Sasaki, and Gilbert 2020), these methods rely only on one fixed snapshot of timing profile and therefore could not fully reflect the stochastic nature of the timing programme.

1.2.4 Mapping of replication fork directionality (RFD) domains

Upon stochastic origin firings, replication forks progress bi-directionally away from the initiation sites during the synthesis phase of the cell cycle (Ticau et al. 2015; Aria and Yeeles 2019). Therefore, the nature of DNA replication is stochastic but symmetrical. Most recently, directional sequencing of Okazaki fragments (OK-seq) has been

developed to map the replication fork directionality (RFD) domains along the chromosomes (Smith and Whitehouse 2012; McGuffee, Smith, and Whitehouse 2013; Petryk et al. 2016), by comparing the difference between the proportions of rightward- and leftward-moving forks.

However, RFD domains mapped by the OK-seq method do not always overlap with the RT profiled by the WGS data (Pope et al. 2014; Petryk et al. 2016; Tubbs et al. 2018). A plausible explanation for the discrepancy between the two *in vitro* methods is that they were independently measuring different replication events from different cell populations at different times, in line with the heterogeneity of the timing programme. To date, methods for simultaneous reconstruction of RT profiles and RFD domains from the same cell population have not been previously reported.

More broadly, it is worth noting that OK-seq only employs nascent Okazaki sequences from the discontinuously replicated lagging strands (Smith and Whitehouse 2012). Therefore, methods for reconstruction of RFD domains using sequences from both the leading and lagging strands have also not been previously reported.

1.2.5 Transcription shapes the landscape of DNA replication

During the synthesis phase of the cell cycle, DNA replication machineries must interfere, overcome, or compete with a range of genomic properties on the same DNA template; for examples, transcription machineries, chromatin features, and G-quadruplex (G4) structures (Besnard et al. 2012; Haradhvala et al. 2016; Hamperl and Cimprich 2016; Du et al. 2019). Despite intensive investigation, it remains poorly understood where exactly, and how differently, DNA replication starts and ends in the human genome. However, it is well known that early-replicating regions are enriched with highly expressed genes (Pope et al. 2014; Polak et al. 2015; Haradhvala et al. 2016; Marchal, Sima, and Gilbert 2019). Most recently, OK-seq mappings have further revealed that origin firing preferentially initiates at the transcription start site (TSS) of highly transcribed genes (Petryk et al. 2016; Chen et al. 2019).

1.3 Research objectives

In this thesis, I propose to fully explore the entire WGS data from primary bulk tumours, and hypothesise that they may provide a snapshot in time, in space, and in specific cell type of the tumour replication timing programme. In summary, the aims of this study are to (i) measure the proportion of S phase cells within a primary tumour sample using WGS data, referred to as *in silico* sorting; (ii) profile the temporal dynamics of tumour replication timing directly from primary tumour samples rather that from cancer cell lines; and (iii) map the spatial landscape of replication origins and termini simultaneously from the same primary cancer WGS data using a novel resampling-based RFD methodology in a number of steps (as shown in Graphical Summary).

Chapter 2

Using DNA sequencing data to infer the proportion of S phase cells within a primary tumour sample

2.1 Introduction

The separation of proliferating cells from resting cells is the first and foremost procedure when profiling DNA replication timing. But not all cell or cancer types can be established as cell lines (Masters 2000; Sasaki et al. 2017), followed by *in vitro* flow sorting procedures. This has proven to be even challenging for the low proliferating hematological cancers, hence patient-derived xenografts of human leukemia are subsequently established as an experimental alternative (Sasaki et al. 2017).

Previously, a study has shown that when a cell population contains appropriate high proportion of S phase cells, its sequencing read depth patterns would become highly correlated with the reference timing profile, and can be even directly used as *de facto* timing profiles (Koren et al. 2014; Marchal et al. 2018).

Here, in this chapter, instead of directly using read depth patterns from the highly proliferating primary tumour samples as *de facto* timing profiles, I propose that ordering the extent of individual tumours' read depth correlations (i.e. similarities) with the reference timing profile, a novel separation method could then be introduced, which I refer to as *in silico* sorting. Subsequently, by comparing the difference between

proliferating and resting tumour samples hitherto separated by this *in silico* sorting procedure, I envision that the average replication timing of a distinct cancer or cell type could also be directly reconstructed, in line with the existing *in vitro* profiling methods. By doing so, I could also average out sequence- or patient-specific heterogeneities in a cell population.

2.2 Material and methods

2.2.1 Datasets

Whole-genome sequencing data of six unsynchronised, flow-sorted lymphoblastoid cell lines (LCLs; with one experiment repetition) (Koren et al. 2012) were reconstructed as the reference replication timing profile in this study. Cancer whole genomes from three cancer types, including 101 small cell lung cancer (SCLC) (George et al. 2015), 56 neuroblastoma (NBL) (Peifer et al. 2015), and 99 chronic lymphocytic leukemia (CLL) (Puente et al. 2011) were reconstructed as the tumour replication timing profiles. Sequencing data of 92 (out of 101) adjacent normal lung tissues from the same SCLC patients was reconstructed as normal replication timing (SCLC-NL) (George et al. 2015). Finally, the validation sequencing data of 8 unsynchronised, neuroblastoma cell lines (NBL-CLs) was also included in this chapter to validate my *in silico* sorting method.

Table 2.1: Whole-genome sequencing data

Published data	Source	Identifier
Lymphoblastoid cell lines (LCL)	Koren et al. 2012	SRA052697
Small cell lung cancer (SCLC)	George et al. 2015	EGAS00001000925
Neuroblastomas (NBL)	Peifer et al. 2015	EGAS00001001308
Chronic lymphocytic leukemia (CLL)	Puente et al. 2011	EGAS00000000092
Neuroblastomas cell line (NBL-CL)	Rosswog et al. 2021	ENA: PRJEB45367

2.2.2 Quantification and normalisation of whole-genome sequencing

All paired reads were aligned to GRCh37 human reference genome (hs37d5) using BWA-MEM (version 0.7.15-r1140) with the parameters -T 0 -M -Y by our in-house analysis pipeline as previously described (Peifer et al. 2012). For LCL reference genome, single-end reads were aligned by GATK (version 4.0.0) following its Best Practices pre-processing workflow. To ensure cross-comparability, the human reference genome was uniformly partitioned into ~2.6 million non-overlapping 1 kb windows (Hawkins et al. 2013; Sasaki et al. 2017). Sequencing read depth/counts per 1 kb window were calculated by our Sclust copy number analysis tool (Cun et al. 2018), followed by local GC content correction for each window. For cancer genomes, read counts were further corrected for copy number states estimated by Sclust to avoid bias from somatic alteration events, i.e. amplifications and deletions. Partitioned windows with zero reads were precluded from downstream analyses. Reads per kilobase, per million (RPKM) normalisation was performed in each library (Sasaki et al. 2017; Tubbs et al. 2018).

Median normalised read counts (median RPKM) were then calculated across different samples or cells in the respective subgroups or cell populations (Woodfine et al. 2004). By doing so, one can also averaged out the putative sequence-specific replication timing between individuals (Ryba et al. 2012; Koren et al. 2014; Sasaki et al. 2017), copy number variations between normal samples (Marchal et al. 2018), and potential asynchronous replication of chromosomes between tumour samples (Marchal, Sima, and Gilbert 2019).

2.2.3 Reconstruction of S/G1 reference replication timing profile

The LCL reference replication timing profile was determined from the S/G1 (S to G1) read depth ratio, by comparing the difference between proliferating and resting reads as broadly described (Woodfine et al. 2004; Ryba et al. 2011; Koren et al. 2012; Tubbs et al. 2018). S/G1 ratios were log2 transformed, and scaled to genome-wide mean of 0 and SD of 1 (Koren et al. 2012) to ensure cross-comparability. Early and late replication timing regions were defined by log2 ratio greater than or less than 0. Smoothed read

depth and timing profiles were performed on total 1 kb windows using smooth.spline package with the default parameters in R (https://www.r-project.org).

2.2.4 Visualisation of replication timing profiles

Chromosomal profiles were generated by purpose-written R scripts using chromosome and cytoband information (chromInfo.txt.gz and cytoBand.txt.gz) downloaded from UCSC Genome Browser (GRCh37/hg19). Published consensus replication timing profile was downloaded from author's website (http://mccarrolllab.org/resources) (Koren et al. 2012), which was calculated by pooling total read counts from all LCL cells at ~2 kb variable-size, equal-coverage windows. Genomic coordinates were converted to hg19 using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Given the resolution differences, this consensus timing profile was merely used for visual comparison in Figure 2.1A (bottom panel, green line). As for the reference timing profile reconstructed by my approach at 1 kb resolution can be seen in Figure 3.1C, 3.3A and 3.4A,B (green smoothed lines).

2.2.5 Statistical analysis

Spearman's rank correlation coefficient was performed in read depth correlation analysis, and a Spearman's rho was given.

2.3 Results

2.3.1 Opposing sequencing read depth pattern between S and G1 phase cells

The separation of proliferating cells from resting cells is the first and foremost procedure when profiling DNA replication timing, therefore I set out to first study DNA content differences between cell cycle time points. Whole-genome sequencing of 6 unsynchronised, flow-sorted lymphoblastoid cell lines (LCLs) were used as my

reference for S and G1 phase reads (Koren et al. 2012). Then, sequencing reads were uniformly partitioned at 1 kb resolution (Hawkins et al. 2013; Sasaki et al. 2017), followed by local GC content correlation and reads per kilobase million (RPKM) normalisation. To measure the central tendency in a cell population (Woodfine et al. 2004), I calculated median read counts across different experiments per kb window in the respective S and G1 cell populations (Figure 2.1A, top panel). Finally, I reconstructed reference timing profiles using the log2 S/G1 read depth ratio as widely described (Woodfine et al. 2004; Ryba et al. 2011; Koren et al. 2012) (Figure 2.1A, bottom panel).

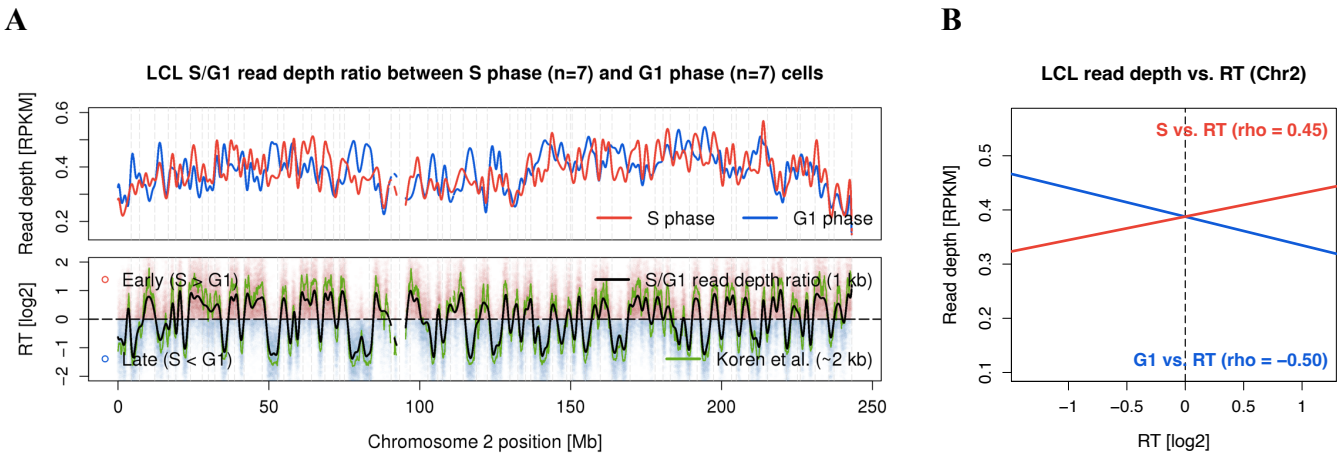


Figure 2.1: (A) LCL reference replication timing (RT) profile for human chromosome 2. (Top panel) Normalised S phase (red smoothed line) and G1 phase (blue) reads at 1 kb resolution across 7 LCLs (median RPKM). (Bottom panel) Reference RT profile (black smoothed line) inferred from log2 S/G1 read depth ratio (genome-wide mean of 0 and SD of 1). Early- and late-replicating regions are defined by a log2 ratio greater than (red dots) or less than (blue) 0. (B) Simplified scatterplot showing S phase reads positively correlated with the reference RT (red linear regression line), whereas G1 phase reads negatively correlated (blue) in chromosome 2. Spearman's rank correlations are given (see unsimplified scatterplots in Figure A.1 and A.2).

Patterns of S phase reads are known to correlate with the timing profiles (Koren et al. 2014; Marchal et al. 2018), and have been extensively studied when profiling DNA replication timing (Ryba et al. 2012; Hawkins et al. 2013; Koren et al. 2014). In contrast, G1 phase reads are less studied, and only served as controls in each uniformly-partitioned window (Hawkins et al. 2013; Sasaki et al. 2017), or fixed to a constant

coverage in each variable-sized window (Koren et al. 2012; Siefert et al. 2017). Interestingly, I observed that G1 phase reads appeared not to be a flat line, and still exhibited peak and valley patterns along the chromosome arms even after GC correction (Figure 2.1A, blue line in top panel). This observation prompted me to speculate that there might be a background signal from an unknown source, and suggested that it could have underpinning impacts on both S and G1 phase reads. Nevertheless, this unknown background signal would be eventually cancelled out in the canonical S/G1 ratio approach.

To determine whether patterns of G1 phase reads are also associated with the timing profiles, I performed read depth correlation analyses (Koren et al. 2014). Using chromosome 2 as an example, I first observed that S phase reads had a positive correlation with the timing profile (Figure 2.1B, red line); whereas, in contrast, G1 phase reads had a negative correlation with the timing profile (Figure 2.1B, blue). Intriguingly, I found that this opposite read depth correlation trend is ubiquitous across 22 autosomes in the LCL reference genome (Figure 2.2). Moreover, I also observed that G1 phase reads correlated more strongly but negatively with the timing profile (Figure 2.3A), further suggesting an indispensible role of G1 phase reads when profiling DNA replication. Together, I hypothesised that such distinct, opposing correlation trend could allow the discrimination of any given cell population into S or G1 phase-like cell cycle status.

2.3.2 *In silico* sorting of primary tumour samples using cancer whole-genome sequences

To test this hypothesis, I next examined whether patterns of the primary cancer whole genomes are also associated with the timing profiles. Whole-genome sequencing of primary tumour samples across three different cancer types, including 101 small cell lung cancers (SCLC) (George et al. 2015), 56 neuroblastomas (NBL) (Peifer et al. 2015), and 96 chronic lymphocytic leukemia (CLL) (Puente et al. 2011) were used to initiate this idea, which I refer to as *in silico* sorting hereafter. To ensure cross-

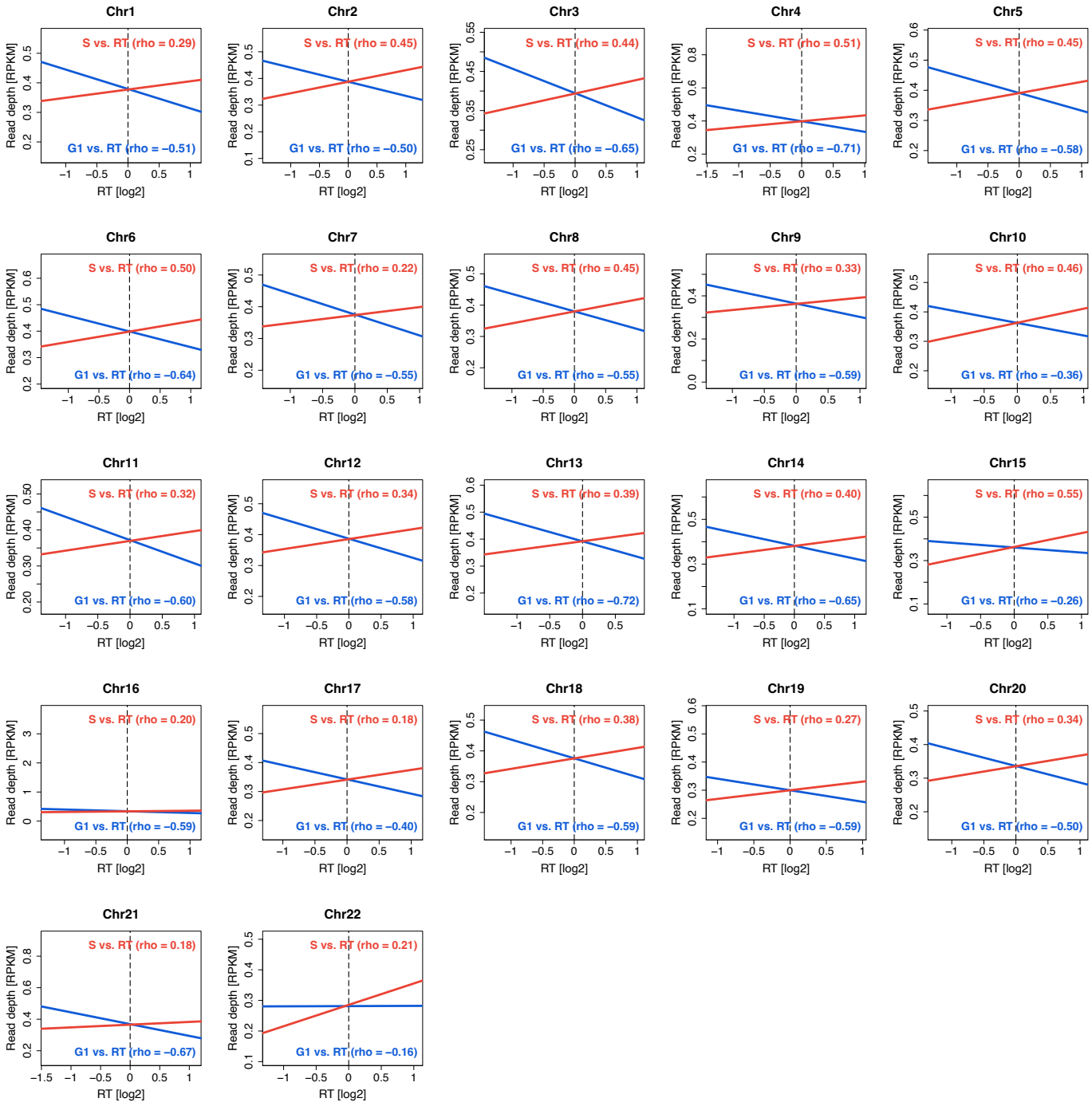


Figure 2.2: Comparison of read depth correlations with the reference timing profile across 22 human autosomes. Overlays of two simplified scatterplots (from Figure A.1 and A.2) demonstrating a distinct, opposing correlation trend between S phase and G1 phase read depths, when compared them with the RT profile (the log2 S/G1 ratio; rho = Spearman's rank correlation coefficient; solid regression line = Linear regression)

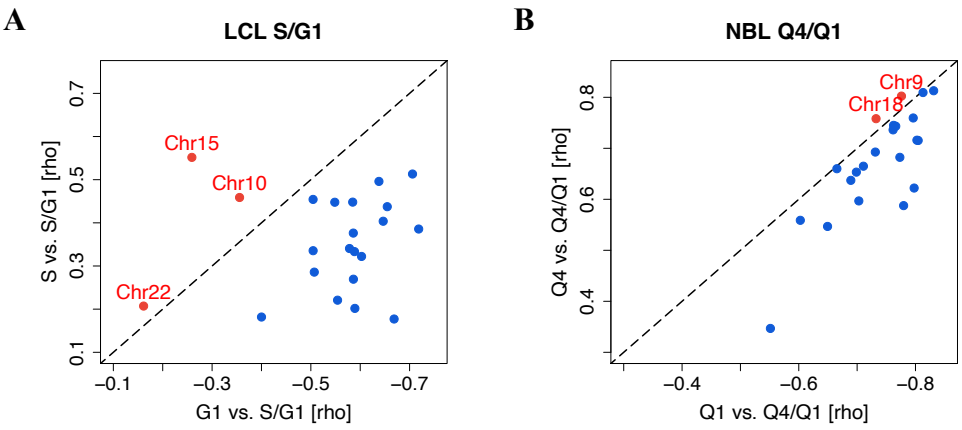


Figure 2.3: G1 phase reads correlated more strongly but negatively with the timing profiles. (A) A stronger but negative correlation trend was observed between G1 reads and the reference RT (S/G1 ratio) across most autosomes (excepted for chr10, chr15 and chr22) in the reference LCL genome, thus challenging the significance of S phase reads as widely accepted. (B) A similar correlation trend was observed between NBL Q1 (i.e. the most G1 phase-like) reads and the NBL tumour timing profiles (Q4/Q1 ratio) across most autosomes (expected for chr9, chr18) in the NBL cancer genome.

comparability, cancer whole genomes were also uniformly partitioned at 1 kb resolution, followed by local GC content correction and RPKM normalisation. Additionally, to avoid bias from potential somatic alteration events in the cancer genomes, i.e. amplifications and/or deletions, sequencing read counts were further corrected for copy number states in the respective tumour samples. Consistent with the opposing correlation trend observed earlier in the LCL reference genome (Figure 2.4A), I found that patterns of the tumour reads were also proportionally correlated with the reference timing profiles across 22 autosomes in the respective cancer types (Figure 2.4B).

Subsequently, I conducted *in silico* sorting prediction (y-axis in Figure 2.4C) by ordering the degree of individual tumours' overall read depth correlation (i.e. similarities) with the reference timing profiles in the respective cancer types. I found that SCLC tumours exhibited the highest median overall read depth correlation among

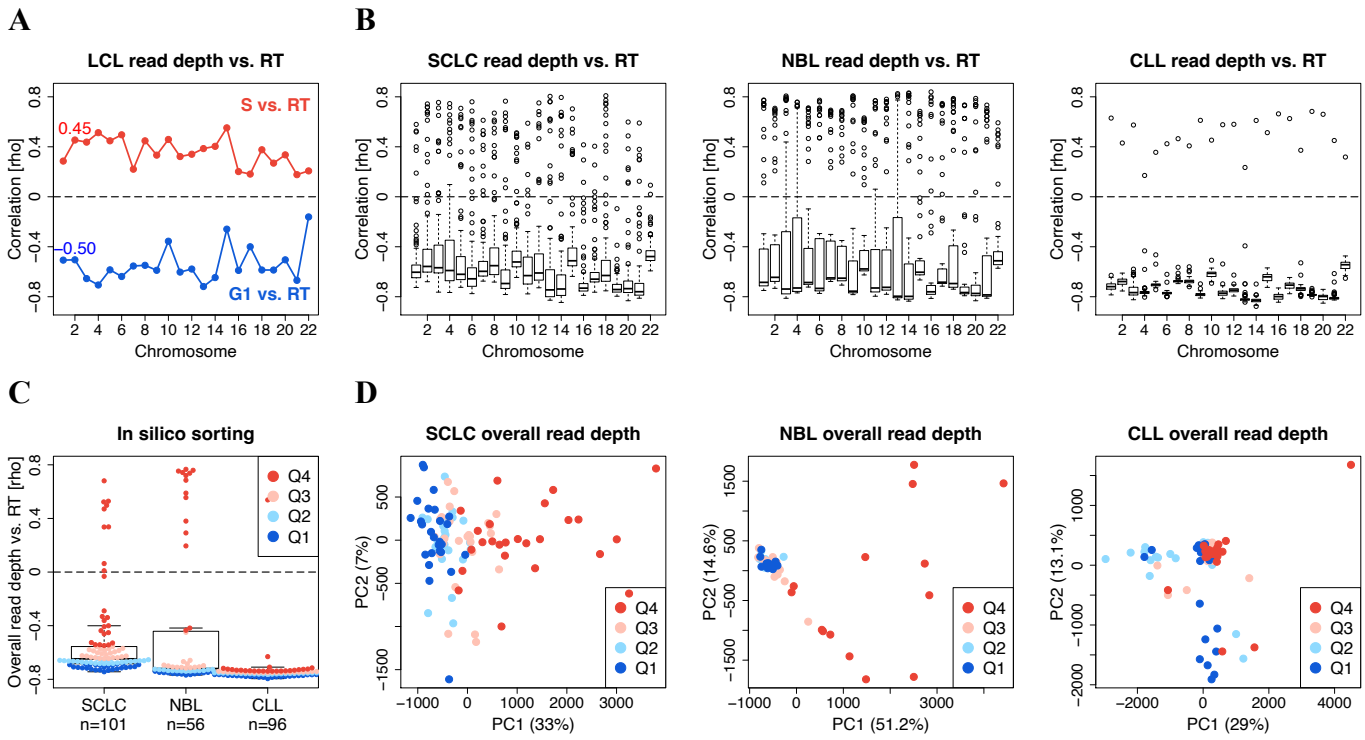


Figure 2.4: *In silico* sorting of primary tumours using cancer whole-genome sequences (A) Opposing read depth correlation trends with the reference RT profile between S phase (red) and G1 phase (blue) reads across 22 human autosomes (full scatterplots in Figure 2.4). (B) Proportional read depth correlation trends with the reference RT profile among primary cancer genomes across 22 autosomes in three cancer types. (C) Boxplot showing our *in silico* sorting predictions (y-axis) for 256 primary tumour samples. Tumour samples are then grouped into equal-sized four quartiles in the respective cancer types. (D) Unsupervised principal component analysis (PCA) of tumour samples' overall read depth patterns, showing that the direction of PC1 (x-axis) is generally in parallel to the distribution of the quartile subgroups (as in A).

the three cancer types, indicating a higher proportion of S phase cells in their cancer cell populations. A possible explanation is that due to the bi-allelic *RB1* loss in nearly all of the SCLC tumours (George et al. 2015), defects of the cell cycle arrest is a hallmark of SCLC (Burkhart and Sage 2008), resulting cell proliferation in an unchecked manner (Hanahan and Weinberg 2011). On the other hand, all but one CLL cells appeared to be negatively correlated with the reference timing profiles (Figure 2.4C), indicating a lower proportion of S phase cells. This can also be explained by the low proliferative activity widely reported in human hematological cancers (Sasaki et al. 2017). Finally, tumour samples were grouped into equal-sized four quartiles in the respective cancer

types based on their overall correlations (coloured legends in Figure 2.4C). Together, I conclude that primary tumour samples can be computationally discriminated using cancer whole-genome sequences, and their overall read depth patterns can be used to infer the proportion of S phase cells in a cell population.

To determine whether the cell cycle distribution predicted by my *in silico* sorting was directly associated with their sequencing reads, I performed unsupervised principle component analysis (PCA) on tumours' overall read depth patterns in the respective cancer types. The PCA analyses showed that the first principle component (PC1) was consistent with the quartile distribution (coloured legends in Figure 2.4D), which were predicted independently as shown earlier (y-axis in Figure 2.4C). Moreover, PC1 not only reflected this visible trend of quartile distribution, it also explained up to 33%, 51.2% and 29% of the variance in SCLC, NBL and CLL cancer genomes, respectively (Figure 2.4D). This suggests that tumours' overall read depth patterns play a dominant role in the quartile distribution predicted by our *in silico* sorting.

2.3.3 Flow cytometry reflects *in silico* sorting prediction in 8 neuroblastoma cell lines

To experimentally validate the cell cycle distribution predicted by our *in silico* method, I performed *in silico* sorting in a separate cohort of 8 neuroblastoma cell lines (NBL-CLs), followed by flow cytometry analysis *in vitro.* Whole-genome sequencing of these 8 unsynchronised NBL-CLs were used to conduct *in silico* sorting prediction by ordering the degree of cancer cell lines' overall read depth correlations with the reference timing profiles (y-axis in Figure 2.5A), in line with the premises approach (Figure 2.4C). I reasoned that the forth quartile (Q4) GLBGA and NGP cells would have the highest proportion of S phase cells, and the first quartile (Q1) GIMEN and LAN6 cells the highest proportion of G1 phase cells. Subsequently, *in vitro* cell cycle distribution was measured by flow cytometry (Figure 2.5D), and the proportion of DNA contents in each cell cycle phase was calculated using Dean-Jett-Fox algorithm (Figure 2.5C).

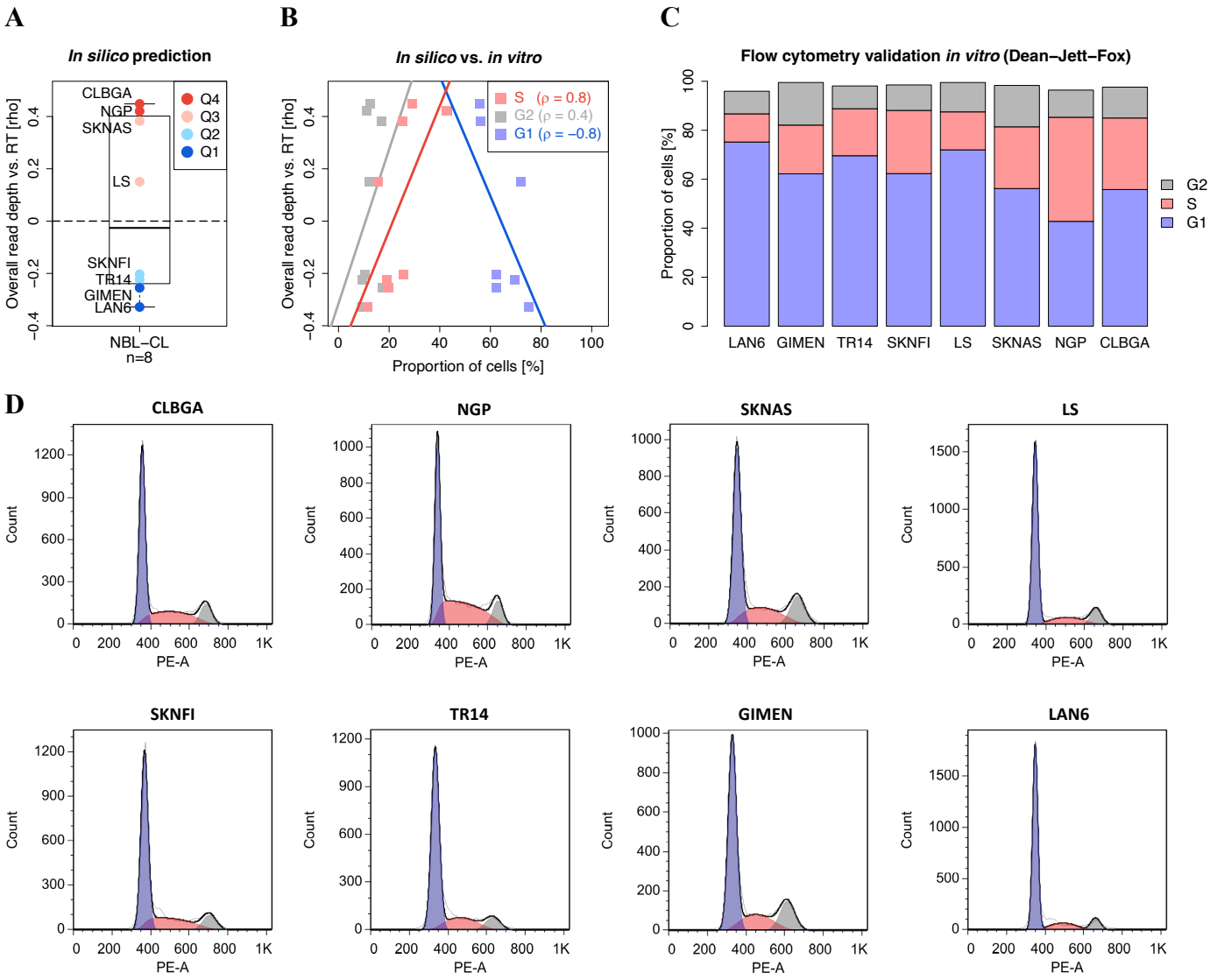


Figure 2.5: Flow cytometry validation reflects *in silico* sorting prediction in 8 neuroblastoma cell lines (NBL-CLs). (A) Boxplot showing the proportion of S phase cells predicted by our *in silico* sorting in 8 NBL-CLs. (B) Scatterplot showing correlations between our *in silico* predictions (y-axis; as in A) and *in vitro* flow cytometry validations (x-axis; as in C) in the respective cell cycle phases. (C) Barchart showing the proportion of cells in each cell cycle phase calculated by the Dean-Jett-Fox algorithm following flow cytometry analysis (as in D). (D) Histograms showing the number of cells (y-axis) with distinct DNA contents (2n, 2-4n and 4n; x-axis) in the respective cell cycle phases (G0/G1, S and G2/M) measured by flow cytometry in 8 NBL-CLs.

Finally, the *in silico* prediction was compared with the *in vitro* validation (Figure 2.5B). I found that my proposed *in silico* predictions had a strong positive correlation with the sorted S phase cells (red line; Spearman's rho = 0.8), a strong negative correlation with the sorted G0/G1 phase cells (blue; rho = -0.8), and no correlation with the sorted G2/M phase cells (gray; rho = 0.4). Therefore, I conclude that the overall read depth patterns of a cancer whole genome is strongly correlated with the cell cycle distribution, and can be directly used to infer the proportion of S phase cells in a cell population.

2.4 Discussion

In this chapter, I presented a novel *in silico* sorting method to measure the proportion of S phase cells using whole-genome sequencing data, and verified it with flow cytometry *in vitro*. Through the analysis of accurately replicated DNA sequences of the primary tumour samples, my novel approach highlights that their overall sequencing read depth patterns can be directly used to perform *in silico* prediction without the need for *in vitro* sorting procedures. Therefore, after the separation of proliferating and resting primary tumour samples, and pooling them into a virtual population of the same type, I hypothesise that the average replication timing profile from a distinct cancer type could then be directly reconstructed as described in detail in Chapter 3.

Chapter 3

Temporal dynamics of tumour replication timing in primary cancer whole-genome sequences

3.1 Introduction

Recent studies have begun to address the cellular heterogeneity of the replication timing programme in humans (Hansen et al. 2010; Pope et al. 2014; Klein et al. 2019). Currently, whole-genome sequencing data from flow-sorted cancer cell lines can provide a snapshot in time and space of the replication programme that is topologically preserved in distinct cellular types (Koren et al. 2012; Fragkos et al. 2015). However, not all cell or cancer types can be established as cell lines, followed by cell sorting *in vitro* (Masters 2000; Sasaki et al. 2017). Therefore, to date, a single reference timing profile derived from sorted lymphoblastoid cell lines has been widely used in the cancer genomics community to study replication timing in multiple types of cancers (Koren et al. 2014; Polak et al. 2015; Haradhvala et al. 2016; Li et al. 2019). This major one-size-fits-all limitation has hampered the possibility to investigate the most cell type-specific genomic properties, i.e. replication initiation and termination domains, in the human cancer genome (Ryba et al. 2010; Haradhvala et al. 2016).

Previously, an alternative profiling method has shown that when a cell line population contains an appropriate high proportion of cells in S phase, its sequencing read depth

patterns would become highly correlated with the reference timing profile (Koren et al. 2014; Marchal et al. 2018). Despite these megabase-scale read depth patterns are also reportedly to be individual- or sequence-specific (Ryba et al. 2012; Koren et al. 2014; Sasaki et al. 2017), the authors further proposed to directly use them as *de facto* replication timing profiles. Nevertheless, methods to reconstruct cell type-specific replication timing directly from primary cancer genomes without the need for *in vitro* sorting procedures have not been previously reported.

Upon *in silico* sorting of primary tumour samples into S and G1 phase-like cell cycle status as described above (Figure 2.4C), I hypothesised that tumour replication timing could then be directly profiled in a cancer- or cell type-specific manner. To this end, cell cycle enrichment (i.e. cell synchronization) is not required in the isolation process of current established methods when profiling average replication timing in a given cell population (Woodfine et al. 2004; Audit et al. 2013; Marchal et al. 2018). Therefore, I reasoned that I can directly inferred replication timing by using the canonical S/G1 read depth ratio approach as widely described (Woodfine et al. 2004; Ryba et al. 2011; Koren et al. 2012). Notably, I have reported above that it is the G1 phase reads that correlated more strongly but negatively with the reference timing profile, and neither S nor G1 phase reads match the full extent of the timing profile (Figure 2.3A; maximum Spearman's |rho| = 0.72), suggesting an unknown background signal. Therefore, instead of directly using S phase-like reads as *de facto* timing profiles as previously described in an alternative method (Koren et al. 2014; Marchal et al. 2018), I reasoned that the canonical S/G1 ratio approach is nevertheless required for cancelling out this unknown background in both S and G1 phase reads when profiling DNA replication.

3.2 Material and methods

3.2.1 Signal-to-noise ratio

The strength of replication timing (RT) signal was determined by standard deviation of the timing profile (e.g. black smoothed spline in Figure 3.1A; bottom panel). The strength of RT noise was determined by standard deviation of the difference between

original read depth ratios (e.g. un-smoothed red and blue dots in Figure 3.1A; bottom panel) and the timing profile (i.e. smoothed curve).

3.2.2 Replication timing skew (RTS)

The RTS values were computed as the difference between the proportions of early-replicating (E) and late-replicating (L) 1 kb windows in each chromosome as:

$$RTS = (E - L) / (E + L).$$

3.3 Results

3.3.1 Direct profiling of tumour replication timing using human cancer whole-genome sequences

First, I used the most variable (PC1 = 51.2% in Figure 2.4D) 28 out of 56 NBL tumours to perform my direct replication profiling approach, i.e. between the most S phase-like Q4 and the most G1 phase-like Q1 tumours predicted earlier by my novel *in silico* sorting approach (as in Figure 2.4C). To average out potential asynchronous replication of chromosomes (Marchal, Sima, and Gilbert 2019) and, as importantly, putative sequence-specific replication timing between individuals (Ryba et al. 2012; Koren et al. 2014; Sasaki et al. 2017), I measured median read counts across tumour samples in the respective subgroups (Woodfine et al. 2004). Intriguingly, I found an inverted, near mirroring read depth pattern between NBL Q4 and Q1 tumour reads (Figure 3.1A, top panel), suggesting an opposing read depth correlation trend between S and G1 phase reads as shown earlier in the LCL reference genome (Figure 2.1B). By comparing Q4 and Q1 tumour reads with the LCL reference timing profiles, I also reproduced this opposing trend across autosomes in the NBL cancer genome (Figure 3.1B, red and blue lines). Finally, NBL tumour timing profiles were reconstructed using a log2 Q4/Q1 read

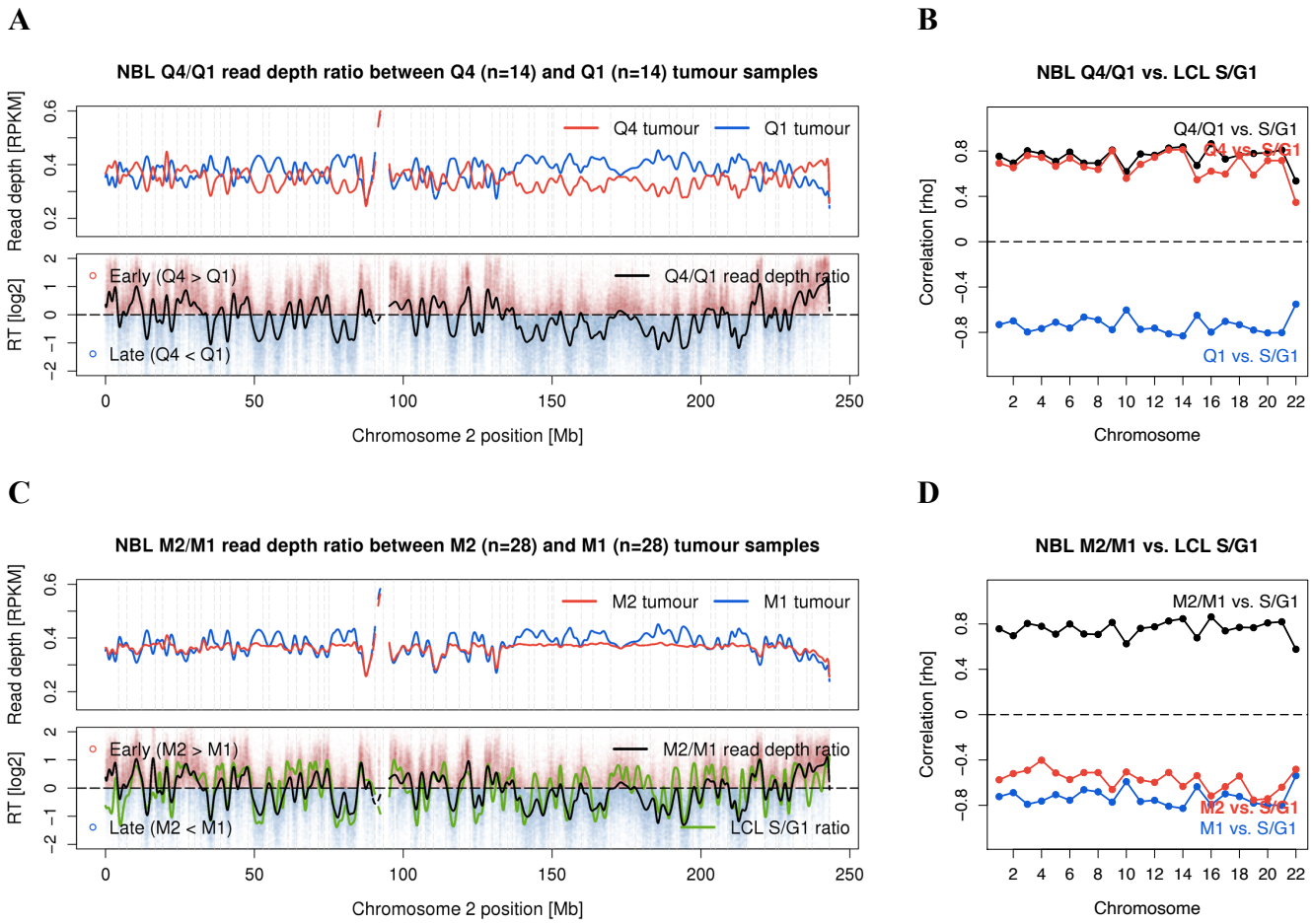


Figure 3.1: Direct profiling of tumour replication timing from 56 neuroblastoma (NBL) primary cancer whole genomes. (A) NBL tumour timing profile for chromosome 2 inferred from Q4 and Q1 tumours (as in Figure 1E). (Top panel) Normalised Q4 (red smoothed line) and Q1 (blue) tumour reads at 1 kb resolution across 28 NBLs (median RPKM). (Bottom panel) Tumour timing profile (black smoothed line) inferred from the log2 Q4/Q1 read depth ratio (genome-wide mean of 0 and SD of 1). (B) Correlations between tumour reads and LCL reference profiles across 22 autosomes. The composite black line showing NBL Q4/Q1 timing profiles were highly correlated with the LCL reference profiles (median rho = 0.77). (C) NBL tumour timing profile for chromosome 2 inferred from M2 (the second median; Q4+Q3) and M1 tumours (the first median; Q2+Q1). (Top panel) Normalised M2 (red) and M1 (blue) tumour reads across 56 NBLs. (Bottom panel) Tumour timing profile (black) inferred from the log2 M2/M1 read depth ratio. (D) Correlations between tumour reads and LCL reference profiles across 22 autosomes. M2 reads flip to be negatively correlated with the LCL reference profile, when compared to the Q4 reads (as in B).

depth ratio at 1 kb resolution across 28 NBL tumours (Figure 3.1A, bottom panel), which had a strong correlation with the LCL reference timing profiles across autosomes (median rho = 0.77; Figure 3.1B, black line). Consistently, I also observed that it is the

NBL Q1 (i.e. the most G1-like) tumour reads correlated more strongly but negatively with the NBL Q4/Q1 tumour timing profiles (Figure 2.3B). Moreover, neither Q4 nor Q1 tumour reads matched the full extent of the NBL Q4/Q1 tumour timing profiles (Figure 3.1B and 2.3B), in line with earlier observation in the LCL reference genome (Figure 2.3A).

Second, to confirm the reproducibility of my direct approach, I used the total number of 56 NBL tumours by dividing them into two equal-sized subgroups, i.e. between the S phase-like M2 (second media) and the G1 phase-like M1 (first median) tumours. Curiously, M2 tumour reads appeared to be a near flat line with very few peak and valley patterns along the chromosome arms (Figure 3.1C, red line in top panel). I reasoned that the visible S phase reads from Q4 tumours were consequently balanced out by the inverted G1 phase reads from Q3 tumours, when the two quartile subgroups were merged into M2 tumours. This could also explain why M2 reads flipped to be negatively correlated with the reference timing profiles across the autosomes (Figure 3.1D, red line), when compared to the Q4 reads (Figure 3.1B, red line). Accordingly, NBL tumour timing profiles were reconstructed using a log2 M2/M1 read depth ratio at 1 kb resolution across 56 NBL tumours (Figure 3.1C, bottom panel), which consistently had a strong correlation with the LCL reference timing profiles across autosomes (median rho = 0.77; Figure 3.1D, black line). Nevertheless, comparisons between the two hitherto NBL tumour timing profiles revealed that M2/M1 timing profiles were almost indistinguishable from their Q4/Q1 counterparts across autosomes (Figure 3.2A, black line in NBL; Spearman's rhos ranging from 0.98 to 1), suggesting the robustness of the canonical S/G1 ratio approach.

Third, to further confirm the reproducibility of my direct replication profiling approach, I repeated the above two steps to reconstruct SCLC and CLL tumour timing profiles from the rest of tumour samples (Figure A.3 and A.4). Consistently, I found that M2/M1 timing profiles were ubiquitously, indistinguishable from their Q4/Q1 counterparts across autosomes in the respective cancer types (Figure 3.2A, black lines; Spearman's rhos ranging from 0.96 to 1).

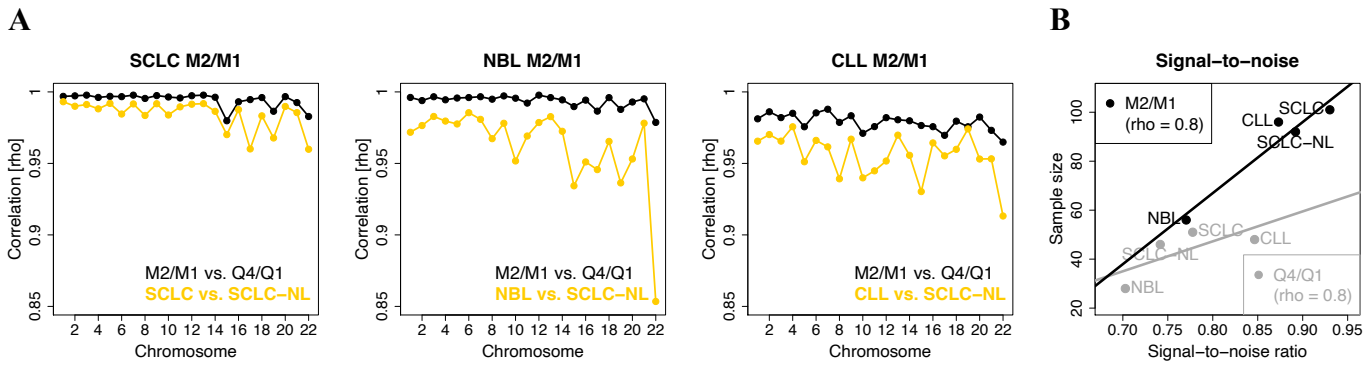


Figure 3.2: (A) Comparisons between timing profiles showing (Black lines) M2/M1 and Q4/Q1 tumour timing profiles were almost indistinguishable (median rho = 1, 1, and 0.98) across autosomes in the respective cancer types. (Yellow lines) SCLC-NL normal timing profiles were highly correlated with their SCLC tumour counterparts (median rho = 0.99), compared with NBL (median rho = 0.97) and CLL (median rho = 0.96) across autosomes. (B) Signal-to-noise ratio analyses (see Methods) among M2/M1 (black dots) and Q4/Q1 (gray) tumour timing profiles across the three cancer types (SCLC, NBL and CLL) and one matched normal type (SCLC-NL), showing the timing profiles were significantly positively correlated with the number of samples (Spearman's rho = 0.8).

To determine the resolution between the two hitherto timing profiles, I performed signal-to-noise ratio (SNR) analyses. I identified that the resolution of timing profiles (i.e. SNR; x-axis in Figure 3.2B) had a strong positive correlation (black and gray lines; rhos = 0.8) with the size of their samples (y-axis). I also found that M2/M1 timing profiles (black dots in Figure 3.2B) in general had slightly higher resolution than their Q4/Q1 counterparts (gray), despite they are almost indistinguishable. Therefore, to simplify my approach and to maximise the statistical power, I hereafter only reported M2/M1 timing profiles in the following analyses. Altogether, I conclude that tumour replication timing can be directly profiled from the primary tumour sample using cancer whole genomes, and my novel direct profiling approach is robust and reproducible across different cancer types derived from independent studies.

3.3.2 Tumour replication timing is preserved in closely related normal tissues and lineage-specific cancer cell lines

To determine whether tumour timing profiles were cell type-specific, I further reconstructed normal timing profiles from 92 adjacent non-neoplastic lung tissues

derived from the same SCLC patients (SCLC-NL) (George et al. 2015) (Figure A.6). Intriguingly, I identified that SCLC-NL normal timing profiles correlated more strongly with their SCLC tumour counterparts across autosomes (median rho = 0.99) than NBL (median rho = 0.97) and CLL (median rho = 0.96) tumour timing profiles (Figure 3.2A, three yellow lines). This indicates that the cellular plasticity of the timing programme is preserved and shared between SCLC tumours and their adjacent normal lung tissues (Pope et al. 2014; Petryk et al. 2016; Du et al. 2019).

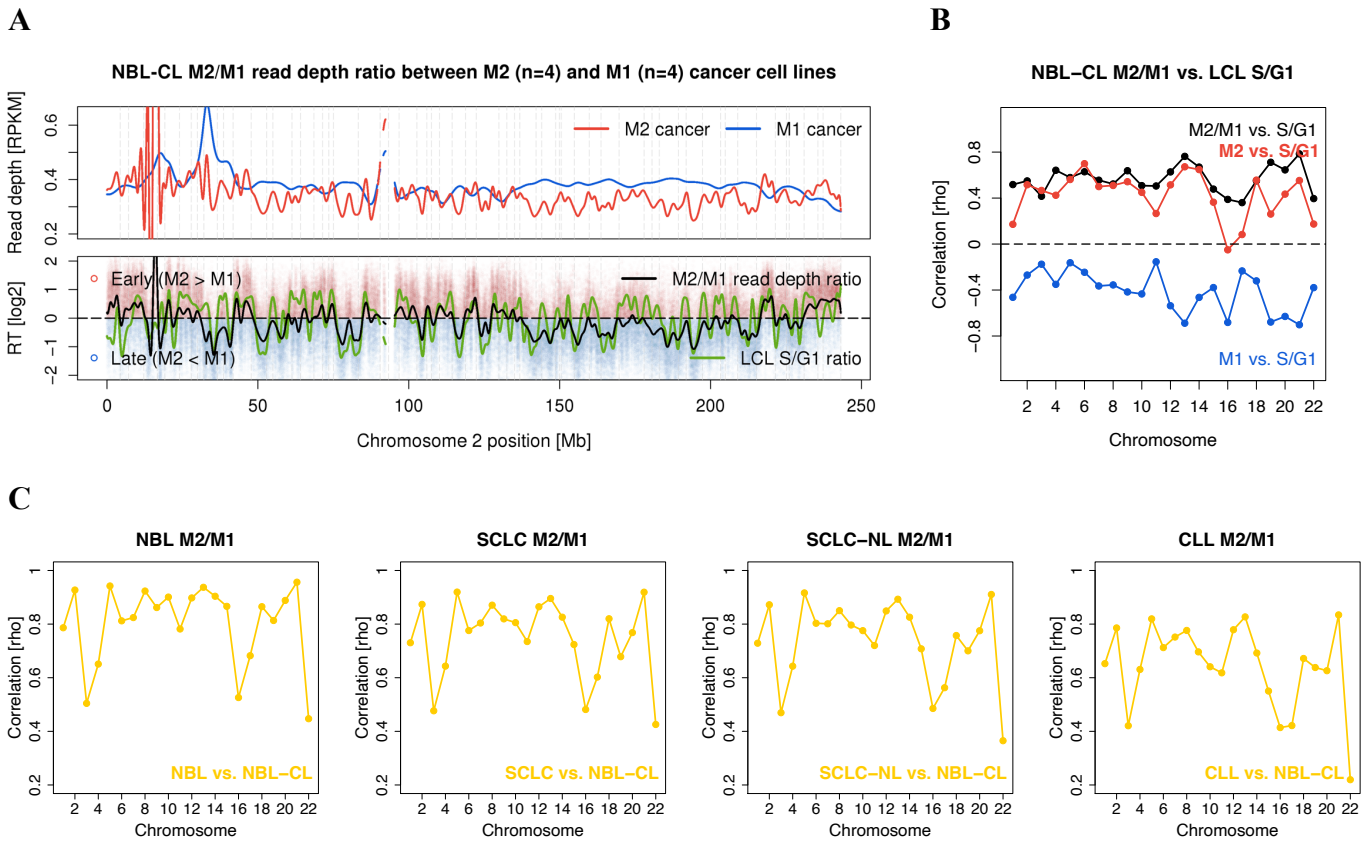


Figure 3.3: Direct profiling of tumour replication timing from 8 NBL-CL cancer cell lines using cancer whole genomes. (A and B) (A) Tumour timing profile for chromosome 2 as inferred between NBL-CL M2 and M1 cell lines (see *in silico* sorting predictions in Figure 2.5A). (B) Intriguingly, NBL-CL timing profiles had the lowest correlations with the LCL reference timing profiles (black line; median rho = 0.55), despite they were both derived from the same cell culture environment and, most importantly, with similar low number of samples (n = 8 and 7, respectively). (C) Comparisons between timing profiles showing that NBL-CL tumour timing profiles correlated more strongly with their NBL tumour counterparts across autosomes (median rho = 0.86), compared with SCLC (median rho = 0.79) and CLL (median rho = 0.66) tumour timing profiles.

To confirm the reproducibility, I further reconstructed tumour timing profiles from the validation cohort of 8 neuroblastoma cell lines (NBL-CLs) (Figure 3.3A). Consistently, I identified that NBL-CL tumour timing profiles correlated more strongly with their NBL tumour counterparts across autosomes (median rho = 0.86) than SCLC (median rho = 0.79) and CLL (median rho = 0.66) tumour timing profiles (Figure 3.3C, yellow lines). Intriguingly, NBL-CL timing profiles had the lowest correlations with the LCL reference timing profiles (Figure 3.3B, black line; median rho = 0.55), despite they were both derived from the similar cell culture environment and, most importantly, with similar low number of samples (n = 8 and 7, respectively). Together, I demonstrate that the topology of reconstructed tumour timing profiles is preserved and shared between SCLC and its matched normal lung tissues, as well as between NBL and its cancer cell lines that originate from the same cell type of origin.

3.3.3 Early- and late-replicating compartments are intrinsic and conserved across different cell types in the human genome

To determine whether the spatial dynamics of replication timing were influenced by the times when chromosomes enter the first half of the S phase (Woodfine et al. 2004; Marchal, Sima, and Gilbert 2019), I then investigated the imbalance between early- and late-replicating regions in the respective chromosomes. Intriguingly, I observed polarised distribution across human chromosomes. For example, chromosome 17 was dominated by the plateau-like, early-replicating compartments (Figure 3.4A, log2 ratios > 0 in bottom panel); whereas, in contrast, chromosome 13 was dominated by the basin-like, late-replicating compartments (Figure 3.4B, log2 ratios < 0) in both SCLC cancer and LCL reference genomes (black and green lines in Figure 3.4A and 3.4B). To quantify this imbalance, replication timing skew (RTS) values were calculated as the difference between the proportions of early- (E) and late-replicating (L) 1 kb windows in the respective chromosomes (Figure 3.4C). Strikingly, skewed values derived from the SCLC cancer genome were significantly correlated with those from the LCL reference genome (Figure 3.4D; Spearman's rho = 0.91, P = 4.48E-09), despite they were originated from different cell types. In fact, I identified this universal early-to-late

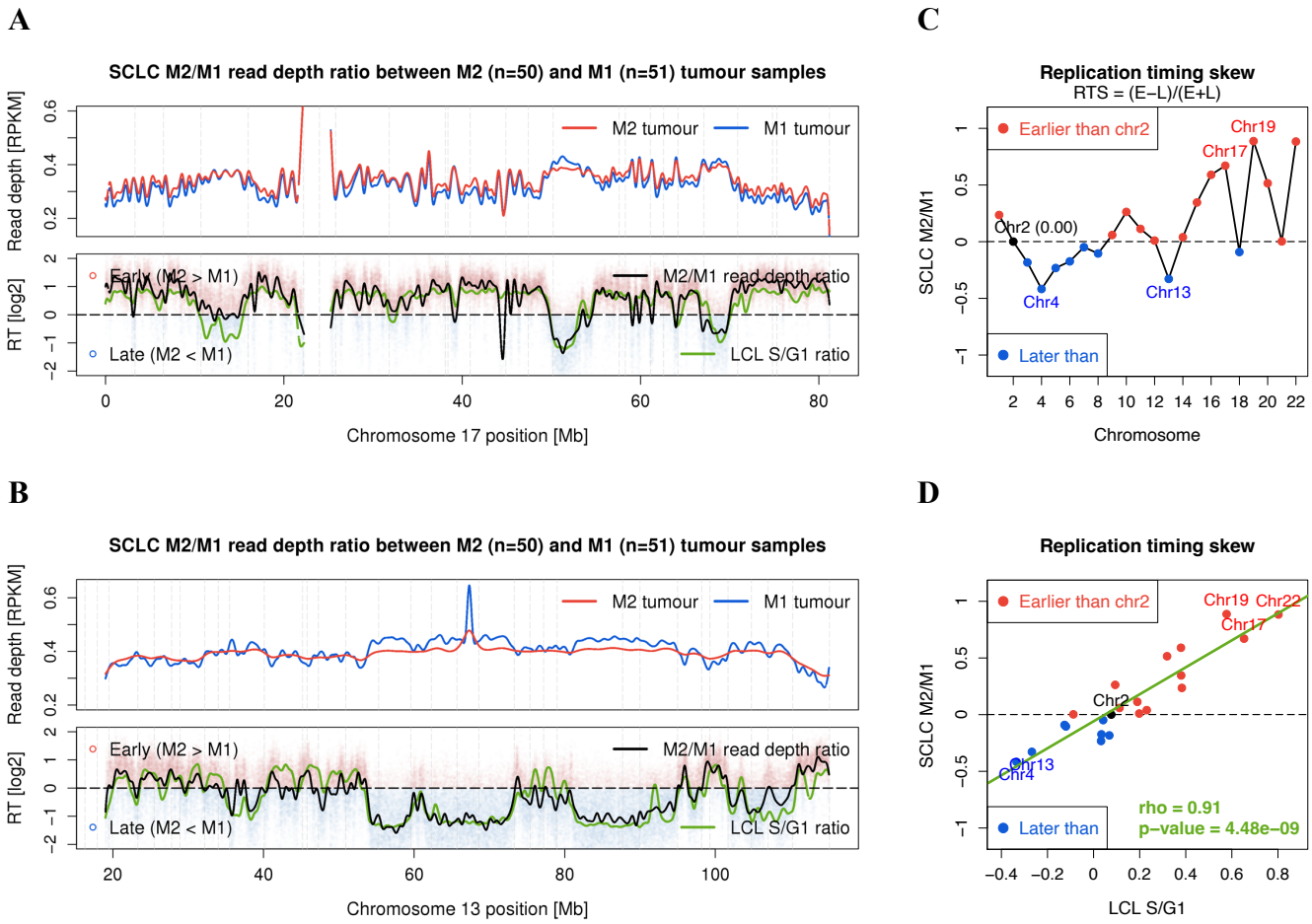


Figure 3.4: Replication timing is intrinsically conserved across different cell types in the human genome. (A and B) (A) Chromosome 17 is dominated by the plateau-like, early-replicating regions (log2 ratios > 0 in bottom panel), whereas (B) chromosome 13 is dominated by the basin-like, late-replicating regions (log2 ratios < 0) in both SCLC tumour and LCL reference timing profiles (black and green lines). (C) Replication timing skew (RTS) in the indicated chromosomes shows that chromosomes 17, 19 and 22 are among the earliest to enter the first-half of S phase in the human genome (red dots), and chromosomes 4 and 13 the latest (blue dots). (D) RTS values derived from the SCLC tumour timing profiles correlated significantly with those from the LCL reference timing profiles (Spearman's rho = 0.91, *P* = 4.48E-09; green regression line for visualization only). This universal early-and-late division is ubiquitous across different normal and tumour timing profiles (see Figure A.8).

were originated from different cell types. In fact, I identified this universal early-to-late skew across different normal and cancer genomes, including SCLC-NL, NBL and CLL (Figure A.7; rhos = 0.91, 0.91 and 0.89), indicating a strictly conserved replication-timing programme in the human genome. Taken together, I can confirm that on average chromosomes 17, 19 and 22 are among the earliest to replicate in the first half of the S

phase, and chromosomes 4 and 13 the latest across different cell types in the human genome (Woodfine et al. 2004) (Figure 3.4C and 3.4D).

However, I observed that patterns of focal peaks and valleys, which represent the exact locations of replication origins and termini, appeared to be equally distributed in the megabase-sized, early-replicating compartments (Figure 3.4A), as well as in the megabase-sized, late-replicating compartments (Figure 3.4B). I reasoned that this is because the early-and-late replication division is a comparative measurement on the genome-wide level, and cannot be used to determine where exactly and how differently DNA replication is initiated and terminated along the chromosome arms. This is also consistent with the notion that replication origins and termini are among the most cell type-specific genomic properties in the human genome (Rhind and Gilbert 2013; Haradhvala et al. 2016), despite that the replication-timing programme is intrinsic and conserved across different type of cells as I observed above (Figure 3.4D). Therefore, to fully understand the dynamics of genome replication, I proposed to further fine map the replication-domain landscape as described in detail in Chapter 4.

3.4 Discussion

Combining Chapter 2 and 3, I presented a novel *in silico* framework to assess the temporal dynamics of the replication timing programme directly from primary tumour samples rather than from cancer cell lines. Through the analysis of accurately replicated DNA sequences of the primary cancer genomes, I showed that the cellular plasticity of tumour replication timing is topologically preserved in closely related normal tissues, as well as in lineage-specific cancer cell lines.

My complementary approaches, i.e. looking at 'the other side of the same coin', highlight that the majority of primary cancer genomes are accurately replicated during cancer cell proliferation thus provides a snapshot in time of the replication programme. Based on my *in silico* sorting predictions between primary tumours, their overall sequencing read depth patterns can also be directly used to infer the proportion of S phase cells within a cell population. However, my results suggest that it is the G1 phase

reads in a cell population that correlate more strongly, but negatively with the replication timing profiles (Figure 2.3A), thus challenging the direct use of S phase reads as *de facto* timing profiles (Koren et al. 2014; Marchal et al. 2018), and further emphasising the indispensible role of G1 phase reads when profiling DNA replication.

Upon adapting the canonical S/G1 read depth ratio method, my direct replication profiling approach allows one to average out the individual- or sequence-specific replication timing in a cell population (Ryba et al. 2012; Koren et al. 2014; Sasaki et al. 2017), as well as an unknown background signal, both of which have not been properly addressed in the alternative profiling method, as noted earlier. Together, the replication timing programme recapitulated by my novel *in silico* framework is not only intrinsically conserved across different normal and cancer genomes (Figure 3.4D), but is also topologically preserved in distinct cancer types, e.g. small cell lung cancers and neuroblastomas (Figure 3.2A and 3.3C).

My findings also highlight that why we must be cautious to use a single one-size-fits-all reference timing profile in studying multiple cancer types (Koren et al. 2014; Polak et al. 2015; Li et al. 2019). However, my direct profiling approach allows one to assess the cell type-specific tumour replication timing programme directly from primary cancer genomes without the need for *in vitro* sorting procedures, and therefore opens up opportunities for the increasing numbers of whole-genome sequencing data published by the broader cancer genomics community.

Chapter 4

The spatial landscapes of replication initiation and termination in primary cancer whole genomes

4.1 Introduction

Another major challenge in our understanding of human DNA replication is the heterogeneity of the timing programme is two-fold; that is, not only the replication origins are different between cell types, but even cells of the same type use different origins to initiate replication in each cell cycle (Hawkins et al. 2013; Takahashi et al. 2019; Bartholdy et al. 2015). Upon stochastic origin firings, replication forks progress bi-directionally away from the initiation sites during the synthesis phase of the cell cycle (Ticau et al. 2015; Aria and Yeeles 2019). Therefore, the nature of DNA replication is stochastic but symmetrical. Most recently, by comparing the difference between the proportions of rightward- and leftward-moving forks, directional sequencing of Okazaki fragments (OK-seq) has been developed to map the replication fork directionality (RFD) domains along the chromosomes (Smith and Whitehouse 2012; McGuffee, Smith, and Whitehouse 2013; Petryk et al. 2016). However, RFD domains mapped by the OK-seq method do not always overlap with the RT profiled by the WGS data (Pope et al. 2014; Petryk et al. 2016; Tubbs et al. 2018). A plausible explanation for the discrepancy between these two *in vitro* methods is that they were

independently measuring on different replication events from different cell populations at different times. This is in line with the aforementioned replication heterogeneities. To date, methods for simultaneous reconstruction of replication profiling and RFD mappings from the same cell population at the same time have not been previously reported.

More broadly, although timing profile itself can already be used to infer RFD domains (Audit et al. 2013; Zhao, Sasaki, and Gilbert 2020), these methods rely solely on one fixed timing profile and therefore could not fully reflect the stochastic nature of the timing programme. Finally, it is worth noting that OK-seq mappings only employed nascent sequences from the lagging strands (Smith and Whitehouse 2012); whereas replication profiling methods comprised complete sequences from both strands. Therefore, methods for reconstruction of RFD domains using sequences from both the leading and lagging strands have also not been previously reported.

Here, given that each timing profile represents a possible combination of replication origins and termini in a cell population, I envision that I could mathematically model every possible origin and terminus usage in the population using bootstrap resampling, followed by reconstruction of resampled timing profiles multiple times. Subsequently, by adapting the principle of OK-seq and applying it to the resampled profiles, a novel bootstrap-based RFD mapping methodology could then be introduced.

4.2 Material and methods

4.2.1 Defining the direction of replication fork movement

Replication timing profile can be used to determine the direction of replication fork movement (Audit et al. 2013; Hawkins et al. 2013). Due to the bi-directional replication (Ticau et al. 2015; Aria and Yeeles 2019), rightward- and leftward-moving forks appear to be descending and ascending slopes along the timing profile, respectively (Figure 4.1A, top). It is worth noting that the definition of rightward and leftward forks here is

based on the leading strand forks progressing bi-directionally away from the initiation sites.

4.2.2 Bootstrap-based replication fork directionality (RFD)

Non-parametric bootstrap resampling was performed with replacement 1,000 times for each genome, followed by reconstruction of 1,000 resampled replication timing profiles (Figure 4.1D). For each 1 kb window, the direction of replication fork movement for each resampled timing profile were calculated as described above by measuring the slopes along the timing profile. Then, in line with the principle of directional sequencing of Okazaki fragments (OK-seq) (Petryk et al. 2016), my proposed bootstrap-based RFD values (Figure 4.1B) were computed as the difference between the proportions of resampled rightward- (R) and leftward-moving (L) forks for each 1 kb window as:

$$RFD = (R - L) / (R + L).$$

4.2.3 Mapping timing transition region (TTR) using RFD values

TTR domains are elongated by uni-directional forks (Figure 4.1A), and would exhibit a skewed proportion of resampled forks (orange and skyblue bars in Figure 4.1B). Therefore, RFD > 0 indicates rightward-moving TTRs (orange, Figure 4.1C) and, likewise, RFD < 0 indicates leftward-moving TTRs (skyblue). Along the RFD plots, flat horizontal segments (RFD = 1 or RFD = -1) indicate near-consistent resampling results in each 1 kb window (Figure 4.1C, bottom panel).

A confidence interval |RFD| >= 0.9 was applied to define timing transition region (TTR) across the genomes. RFD >= 0.9 (orange, Figure 4.1E) represents predominant (>900) rightward-moving forks upon 1,000 resampling per kb window, and vice versa for RFD <= -0.9 (skyblue).

4.2.4 Mapping constant timing region (CTR) using RFD values

CTR domains, including initiation and termination zones, are flanked by bi-directional forks (IZs and TZs; Figure 4.1A), and would exhibit a random proportion of near 500:500 resampled rightward and leftward forks in each 1 kb window (RFD = 0 in Figure 4.1B).

A confidence interval |RFD| < 0.9 was accordingly applied to define constant timing region (CTR) across the genomes (white bars in Figure 4.1D). Along the RFD plots, the slope of each 1 kb window was estimated within a 20-kb sliding window (stepped by 1 kb) using the rollapply function from the zoo package in R (Reijns et al. 2015). It is worth noting that the slope of RFD profile described here (Figure 4.1C, bottom panel) is a different measurement compared to the slope of timing profile mentioned earlier (Figure 4.1C, top panel).

After applying a 20-kb sliding window, only around 0.2%~0.3% RFD domains were undefined across the autosomes in the respective genomes (green in Figure 4.3). Notably, the choice of 15-kb sliding window used by OK-seq (Petryk et al. 2016; Tubbs et al. 2018) resulted in around 0.2~0.6% undefined RFD domains across the autosomes. Therefore, the difference between the two is marginal (green in Figure A.8).

4.3 Results

4.3.1 A novel bootstrap-based replication fork directionality (RFD)

To infer the direction of replication fork movement, I leveraged the stochastic but symmetrical nature of bi-directional replication (Figure 4.1A, top schematic) with a mathematical model. Given that each RT profile represents a possible combination of replication origins and termini from a tumour population, I envisioned that I could therefore model every possible origin and terminus usage in a tumour population through bootstrap resampling (Figure 4.1D). To do so, I repeated random sampling with replacement 1,000 times in the respective cancer cell populations, followed by

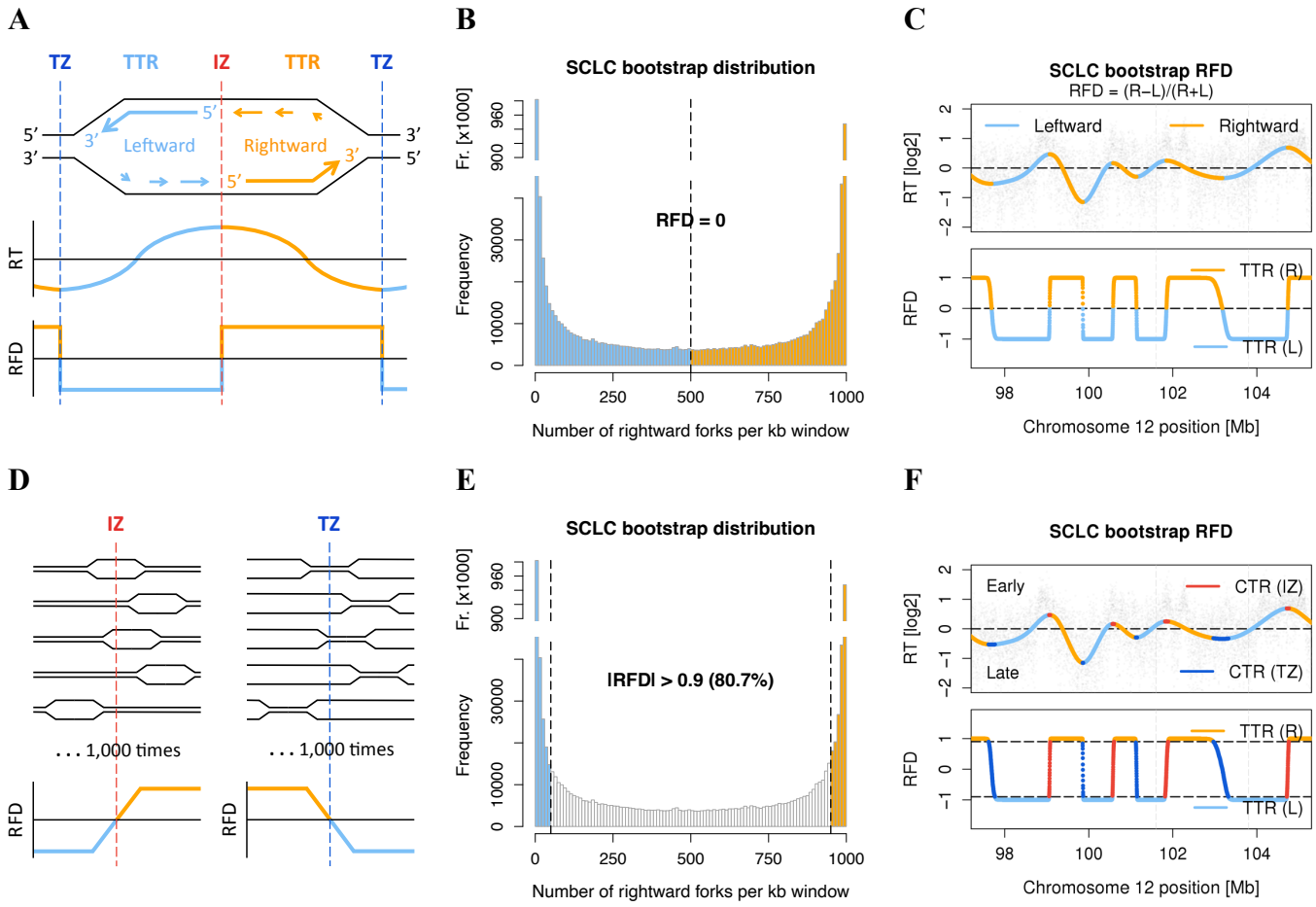


Figure 4.1: A novel bootstrap-based replication fork directionality (RFD). (A) Schematic depiction of the replication fork movement centring on a replication origin. An initiation zone (IZ; red) is flanked by bi-directional rightward (orange) and leftward (skyblue) forks, which elongate and stop at two termination zones (TZs; blue) on each side. Forks between IZ and TZ also called timing transition regions (TTRs). (B) Distribution of the number of resampled rightward forks per kb window after 1,000 bootstrap resampling. Bootstrap-based RFD = 0 indicates a random proportion (500:500) of resampled rightward (orange) and leftward (skyblue) forks. (C) (Top panel) RFD plot showing the direction of replication forks is color-coded by RFD values along the RT profile for part of chromosome 12. (Bottom panel) Flat horizontal segments (RFD = 1 or RFD = -1) indicate near-constant resampling results. (D) (Left) An inefficient, broad IZ is surrounded by delocalized origin firings (depicted as bubbles), resulting in a tilted jump with a positive slope along the RFD profile. (Right) An inefficient, broad TZ is surrounded by delocalized termination events (depicted as closed zippers), resulting in a tilted jump with a negative slope. (E) TTR domains (|RFD| >= 0.9) account for 80.7% of the SCLC genome. RFD > 0.9 (orange) represents predominant (>90%) rightward forks upon 1,000 resampling per kb window, and vice versa for the RFD < -0.9 (skyblue). (F) (Top panel) RFD plot showing the replication domains are color-coded by RFD values along the RT profile. (Bottom panel) IZ is defined as a vertical jump with positive slope (red; as in D, left), and TZ a vertical jumps with negative slope along the RFD profile (blue; as in D, right).

reconstruction of replication timing profiles for each resampled population. Finally, bootstrap-based replication fork directionality (RFD) values were calculated for each 1 kb window as the difference between the proportions of resampled rightward- (orange) and leftward-moving (skyblue) forks (Figure 4.1B; see Methods), in line with the principle of directional sequencing of Okazaki fragments (OK-seq) (Smith and Whitehouse 2012; McGuffee, Smith, and Whitehouse 2013; Petryk et al. 2016).

Owing to the stochastic but symmetrical nature of bi-directional replication, constant timing regions (CTRs; red and blue dotted lines in Figure 4.1A), which are flanked by bi-directional rightward and leftward forks, exhibited a random distribution of near 500:500 fork directions per kb window across 1,000 resampled profiles (RFD = 0 in Figure 4.1B). In contrast, timing transition regions (TTRs; orange and skyblue forks in Figure 4.1A), which are elongated by unidirectional forks, exhibited a skewed distribution towards 0:1,000 or 1,000:0 fork directions per kb window (Figure 4.1B). Therefore, RFD > 0 indicates rightward TTRs (orange bars in Figure 4.1B) and, likewise, RFD < 0 indicates leftward TTRs (skyblue). Not surprisingly, I observed that flat horizontal segments appeared near |RFD| = 1 along the bootstrap-based RFD plots (Figure 4.1C, bottom panel). This indicates near-constant resampling results (i.e. in total 0 or 1,000 rightward forks per kb window in Figure 4.1B), which is consistent with the Okazaki fragment strand bias measured by OK-seq *in vitro* (Petryk et al. 2016; Chen et al. 2019).

To map TTR domains from the genome, I applied a confidence interval of |RFD| >= 0.9 in the respective genomes (Figure 4.1E), in line with the OK-seq threshold (Petryk et al. 2016). I observed that TTR domains dominated 80.7% of the SCLC autosomes (orange and skyblue bars in Figure 4.1E). In total, TTR accounted for 87.8%, 80.7%, 76.9% and 75.5% of the SCLC-NL, SCLC, NBL and CLL autosomal genomes, respectively (black line in Figure 4.2B). Conversely, to map replication CTR domains, which comprise of initiation and termination zones (IZ and TZ; Figure 4.1A), I applied a confidence interval of |RFD| < 0.9 in the respective genomes (white bars in Figure 4.1E). Together, not only do I simultaneously perform replication profiling and RFD mappings from the same tumour populations, I also establish a novel bootstrap-based approach to determine RFD domains using sequences from both the leading and lagging strands.

4.3.2 Distribution of termination events coordinates with initiation activities in both normal and cancer genomes

To further fine map replication initiation zones (IZs) from the CTR domains and quantify their efficiency (Figure 4.1D, left), I identified the vertical jumps with a positive slope along the RFD profiles (red segments in Figure 4.1F, bottom panel; see Methods) (Petryk et al. 2016). In total, IZs accounted for 6.1%, 9.5%, 11.6% and 11.9% of the SCLC-NL, SCLC, NBL and CLL autosomes, respectively (red line in Figure 4.2B). Likewise, to fine map replication termination zones (TZs) and quantify their efficiency (Figure 4.1D, right), I identified the vertical jumps with a negative slope along the RFD profiles (blue segments in Figure 4.1F, bottom panel) (Petryk et al. 2016). In total, TZ accounted for 6.1%, 9.7%, 11.5% and 12.6% of the SCLC-NL, SCLC, NBL and CLL autosomes (blue line in Figure 4.2B).

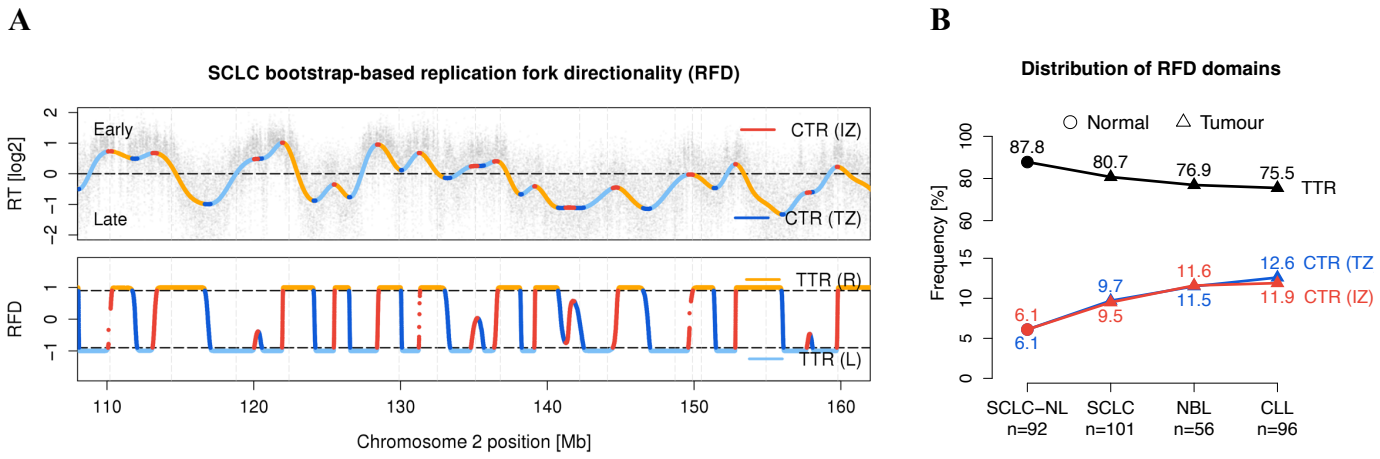


Figure 4.2: (A) RT and RFD profiles for chromosome 2 showing IZ and TZ domains appear alternately (red and blue jumps by turns, bottom panel). (B) Genome-wide distribution of termination events is closely coordinated with initiation activities in both normal and tumour genomes.

Previous OK-seq RFD mappings has reported seemingly over-represented termination zones between two human cell lines (39.4% and 49.5%) (Petryk et al. 2016), compared to the initiation zones (7.1% and 11.5%). Unexpectedly, however, I observed that the genome-wide distribution of TZ domains is closely coordinated with the IZ domains in

both normal and cancer genomes (blue and red lines in Figure 4.2B). Furthermore, I also found that IZ and TZ domains appeared alternately along the RFD profiles (red and blue jumps by turns in Figure 4.1F and 4.2A, bottom panel), which is in line with recent reports that termination zones are determined by and located between two activating origin firings (Hawkins et al. 2013; Petryk et al. 2016).

To determine whether this coordinated distribution observed above in our data is reproducible, I conducted random downsampling analysis. In this analysis, I equally and randomly divided M2 and M1 tumours into two test populations in the respective

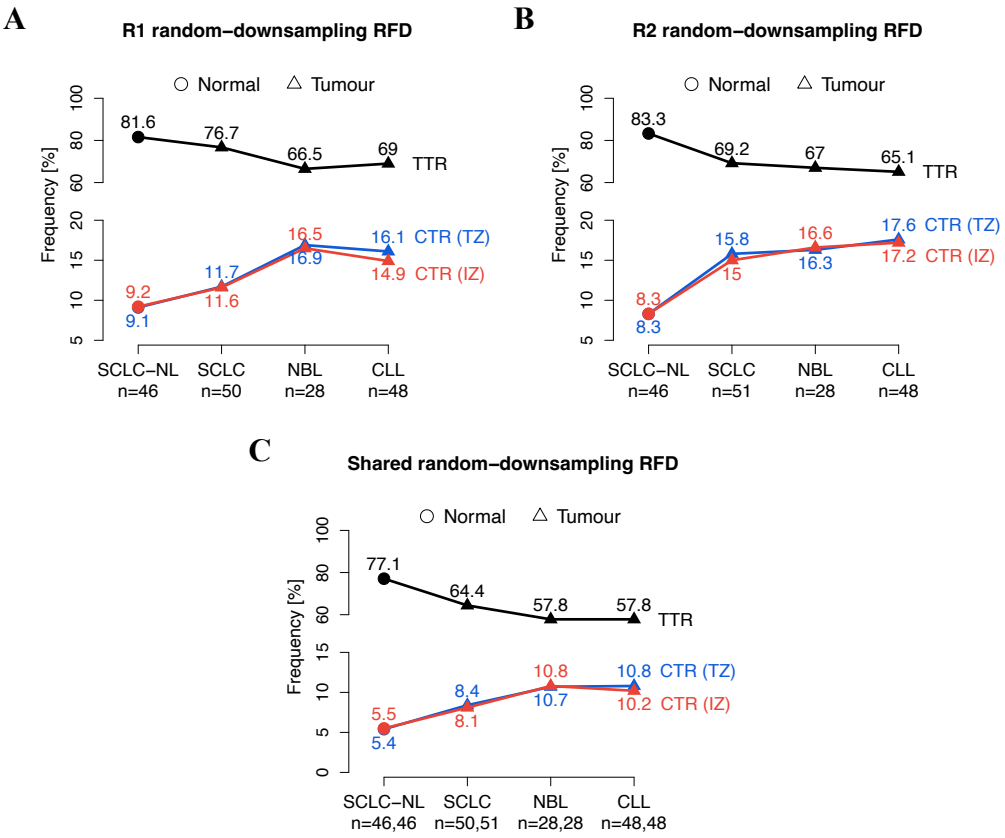


Figure 4.3: Random downsampling analysis by equally and randomly dividing tumour samples into two independent test populations in the respective normal and cancer genomes, followed by bootstrap-based RFD mapping separately in the two independent, randomlised test population 1 (R1) and 2 (R2). (A and B) Despite the randomised replication landscape slightly varied between the two independent test populations, the distribution of TZ domains is consistently coordinated with that of IZ domains in each test population. (C) The distribution of shared TZs between the two test populations is consistently coordinated with shared IZs, in line with our initial observations (as in Figure 4.2B).

normal and cancer genomes (Figure 4.3). By doing so, I equally distributed every possible origin and terminus usage in a population into two independent test sets, and at the same time maintained the same composition between M2 and M1 tumours in each test population. Finally, I performed bootstrap-based RFD to map the replication domains in the respective test populations. Intriguingly, despite the randomised replication landscape slightly varied between the two independent test populations owing to independent origins and termini usage, I constantly observed this coordinated distribution between TZ and IZ domains in both test populations (Figure 4.3A and 4.3B). Furthermore, by identifying shared domains between the two test populations, the genome-wide distribution of shared TZs is consistently coordinated with the shared IZs (Figure 4.3C), in line with our initial observations (Figure 4.2B). Altogether, I conclude that replication-domain landscape can be directly mapped from the primary tumour samples using cancer whole genomes, and the genome-wide distribution of termination events is closely coordinated with the initiation activities in both normal and cancer genomes.

4.3.3 Tumour replication-domain landscape is preserved in closely related normal tissues and lineage-specific cancer cell lines

To determine whether reconstructed replication domains were cell type-specific in the respective cancer genomes, I further reconstructed RFD domains from 92 SCLC-NL normal lung tissues. I identified that SCLC-NL normal genome shared more RFD domains with its SCLC cancer counterpart (85.8%) than the NBL (76.9%) and CLL (74.5%) cancer genomes (Figure 4.4A-C, right panels). Besides, the number of shared domains was 6-fold more than that of specific ones between SCLC-NL and SCLC genomes, compared to only 3.3-fold and 2.9-fold changes between SCLC-NL and the rest of two cancer genomes (ratios in Figure 4.4A-C, right panels). This suggests that the cellular plasticity of replication domains is preserved and shared between SCLC tumours and their closely related normal lung tissues, consistent with our earlier observations in the replication timing profiles (as in Figure 3.2A). Moreover, the number of shared IZ and TZ domains was roughly 2-fold more than that of specific

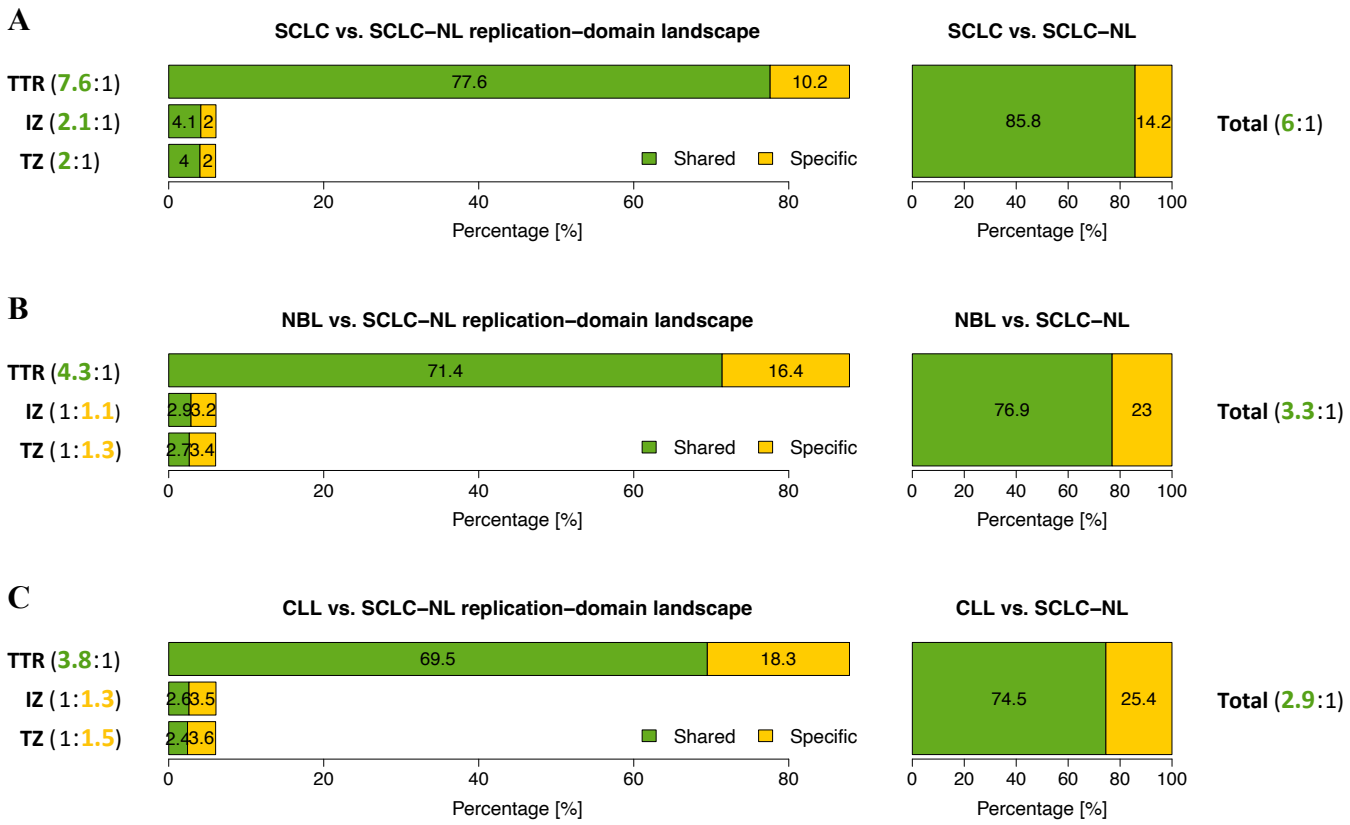


Figure 4.4: The landscape of SCLC tumour replication domains is topologically preserved in closely related SCLC-NL normal lung tissues**.** (A, B and C) (Right panels) Comparisons between RFD domains showing SCLC-NL normal genome shared more replication domains with its SCLC cancer counterpart (85.8%) than NBL (76.9%) and CLL (53.2.5%) cancer genomes. (Left panels) (A) Shared IZ and TZ domains were roughly 2-fold higher than those of specific ones between SCLC-NL and SCLC. (B and C) In contrast, shared IZ and TZ domains were around 1.3-fold lower than those of specific ones between SCLC-NL and the other two cancer genomes.

ones between SCLC-NL and SCLC (ratios in Figure 4.4A, left panel), compared to around 1.3-fold lower between SCLC-NL and the other two cancer genomes (Figure 4.4B-C), further demonstrating that IZ and TZ domains are among the most cell type-specific genomics properties in the human genome.

To confirm the reproducibility, I further mapped RFD domains in the validation cohort of 8 neuroblastoma cell lines (NBL-CLs). Consistently, I identified that NBL-CL genome shared more replication domains with its NBL cancer counterpart (59.3%) than

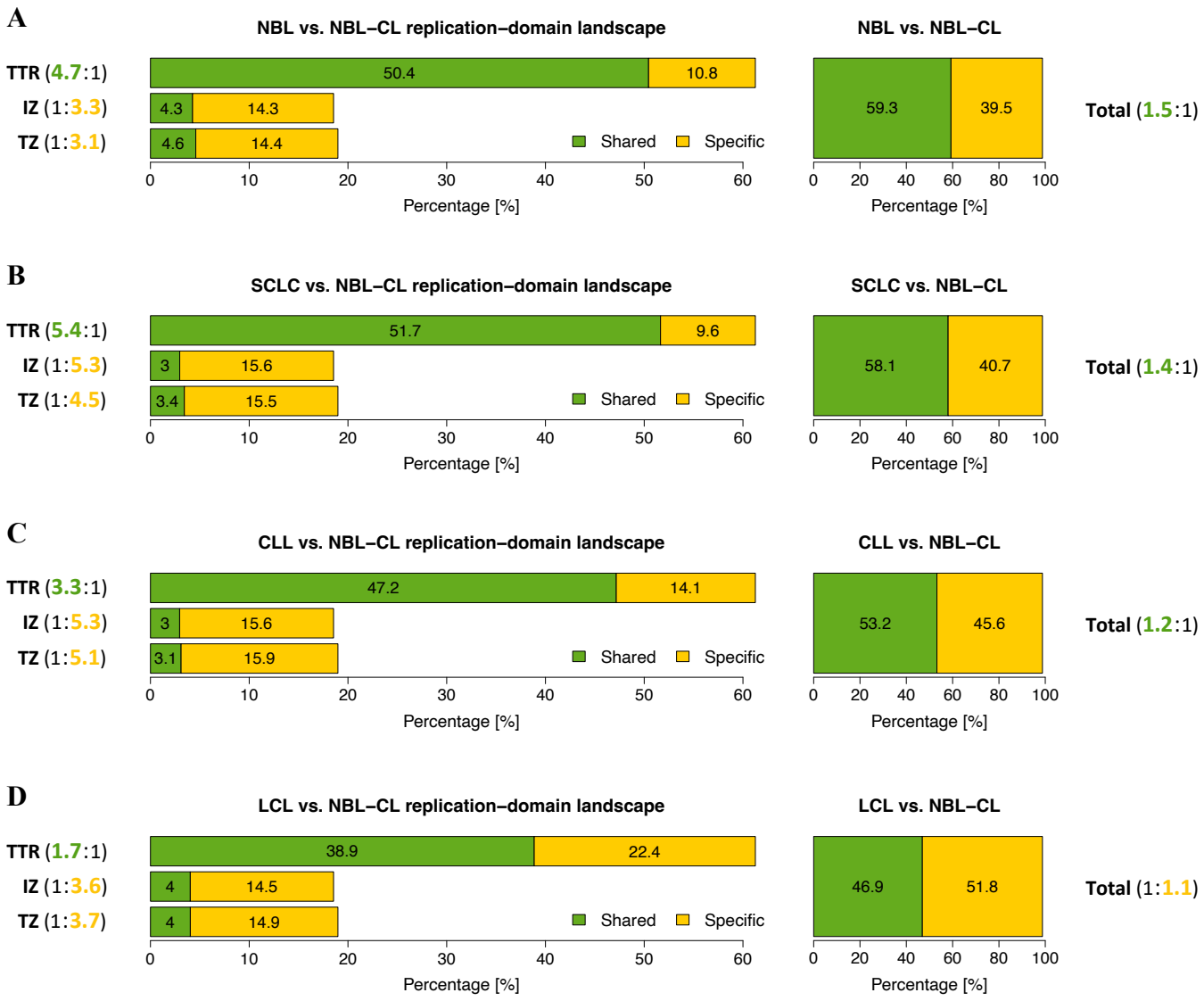


Figure 4.5: The landscape of NBL tumour replication domains is topologically preserved in lineage-specific NBL-CL cancer cell lines. (A, B, C and D) (Right panels) NBL-CL cancer genome shared more replication domains with its NBL cancer counterpart (59.3%) than the SCLC (58.1%) and CLL (53.2.5%) cancer genomes. Moreover, the number of shared domains (green bars) is 1.5-fold higher than that of specific ones (yellow) between NBL-CL and NBL genomes, compared to 1.4-fold and 1.3-fold changes between NBL-CL and the other two cancer genomes. Notably, the NBL-CL cancer genome shared the fewest replication domains with the LCL reference genome (46.9%).

the SCLC (58.1%) and CLL (53.2%) cancer genomes (green bars in Figure 4.5A-C, right panels). Moreover, the number of shared domains is 1.5-fold more than that of specific ones between NBL-CL and NBL genomes, compared to 1.4-fold and 1.2-fold changes between NBL-CL and the other two cancer genomes (ratios in Figure 4.5A-C;

right panels). Finally, NBL-CL genome shared the least replication domains with LCL reference genome (46.9%; green bar in Figure 4.5D, right panel) despite they were both derived from the similar cell culture environment. This is consistently in line with our earlier observations of low correlation between NBL-CL and LCL timing profiles (Figure 3.3B), when compared with other tumour timing profiles (Figure 3.3C). All together, I demonstrate that the plasticity of our bootstrap-based replication-domain landscape is preserved and shared between SCLC and its matched normal lung tissues, as well as between NBL and its cancer cell lines that originate from the same cell type of origin.

4.4 Discussion

In this chapter, I extended my *in silico* framework to assess the spatial landscapes of the tumour replication domains. Not only did I introduce a novel bootstrap-based RFD method to model the bi-directional replication using sequences from both the leading and lagging strands, I also simultaneously reconstructed replication timing and RFD domains from the same primary cancer genomes. By adapting the principle of OK-seq, my novel bootstrap-based RFD approaches allow one to mathematically model every possible origin and terminus usage in a cell population (Figure 4.1D) and, thus, to simultaneously infer the direction of replication fork movement (Figure 4.1F). As illustrated above, CTR domains, either an initialization zone or a termination region, are generally flanked by two bi-directional TTR forks along the RFD profile (Figure 4.1A).

However, I also observed a minority of CTRs that was flanked by two uni-directional TTR forks. For examples, some small but notable peaks around 120 Mb, 135 Mb and 158 Mb of chromosome 2 from the SCLC RFD profile were found to be flanked by two leftward-moving TTR forks (Figure 4.2A, lower panel). Furthermore, these particular CTRs are a composition of an initialization zone adjacent to a termination region (as in Figure 4.2A). Most recently, a similar pattern has been suggested to be a breakage at the TTR regions (Zhao, Sasaki, and Gilbert 2020). Therefore, I presume that these particular domains may be prone to replication stress, leading to a random termination after the stalling of the replication fork (Hawkins et al. 2013; Petryk et al. 2016). Based

on my high-resolution RFD mappings, a putative dormant origin may then be fired immediately downstream of the random terminus in order to complete the fork progression.

Unexpectedly, my data suggests that the genome-wide distribution of termination events is closely coordinated with the initiation activities in both the normal and cancer genomes (Figure 4.2B), which is consistent with the consensus notion reported in yeast (Hawkins et al. 2013), but has not been reported in humans using OK-seq (Petryk et al. 2016; Zhao, Sasaki, and Gilbert 2020). Notably, I consistently reproduced this coordinated distribution in two randomly down-sampled test populations, as well as in shared domains between the two test populations (Figure 4.3), thus supporting a *bona fide* mapping on replication initiation, progression, and termination by our bootstrap-based RFD method.

My results reaffirm that replication origins and termini are among the most cell type-specific genomic properties in the human genome (Figure 4.4 and 4.5), despite the megabase-sized early/late replicating compartments are intrinsic and conserved across different type of cells (Figure 3.4D). Together, the cellular plasticity of tumour replication domains recapitulated by my novel *in silico* framework adds a new spatiotemporal perspective to the three-dimensional human cancer genome.

Chapter 5

Transcription-replication interference

5.1 Introduction

It is well known that highly expressed genes are enriched at early-replicating regions of the human genome (Pope et al. 2014; Polak et al. 2015; Haradhvala et al. 2016; Marchal, Sima, and Gilbert 2019). Most recently, OK-seq mappings have further revealed that origin firing preferentially initiates at the transcription start site (TSS) of highly transcribed genes (Petryk et al. 2016; Chen et al. 2019). Therefore, I reasoned that I could use this relationship between transcriptional activities and replication-domain landscape to evaluate whether our proposed bootstrap-based RFD mappings also reflected this biological insight.

5.2 Material and methods

5.2.1 Datasets

RNA sequencing (RNA-seq) of primary tumour samples derived from a subset of the same cancer patients were used in this analysis, including 70 small cell lung cancer (SCLC) (George et al. 2015), 53 neuroblastoma (NBL) (Peifer et al. 2015), and 71 chronic lymphocytic leukemia (CLL) (Puente et al. 2011).

Table 5.1: Transcriptome sequencing data (RNA-seq)

Published data	Source	Identifier
Small cell lung cancer (SCLC)	George et al. 2015	EGAS00001000925
Neuroblastomas (NBL)	Peifer et al. 2015	EGAS00001001308
Chronic lymphocytic leukemia (CLL)	Puente et al. 2011	EGAS00000000092

5.2.2 Transcriptome sequencing quantification

All paired reads were aligned to Ensembl GRCh37.74 cDNA sequences using kallisto 0.43.1 (Bray et al. 2016). Additional 100 bootstraps per sample were performed to correct for sequence bias. Transcript-level abundance was subsequently estimated, followed by transcripts per million (TPM) normalisation. Aggregated gene-level TPMs were then calculated using sleuth 0.29.0 with default quality-control filters (Pimentel et al. 2017). To align each Ensembl gene to a RFD domain, I used the position of the TSS to retrieve RFD values from the allocated 1 kb window along the RFD profile in the respective cancer genomes. In total, out of 34,908 Ensembl genes, there were 31,612 genes in the SCLC, 31,694 genes in the NBL, and 30,320 genes in the CLL cancer transcriptomes were mappable for a RFD value for further analysis.

5.2.3 Statistical analysis

Wilcoxon rank-sum test was used to test for differential analysis.

5.3 Results

5.3.1 Transcriptional activity strongly correlates with the landscape of replication domains across three cancer types

In total, genes mapped to the termination zones accounted for 9.5%, 11.7% and 11.8% of the respective SCLC, NBL and CLL transcriptomes (blue line in Figure 5.1A), which

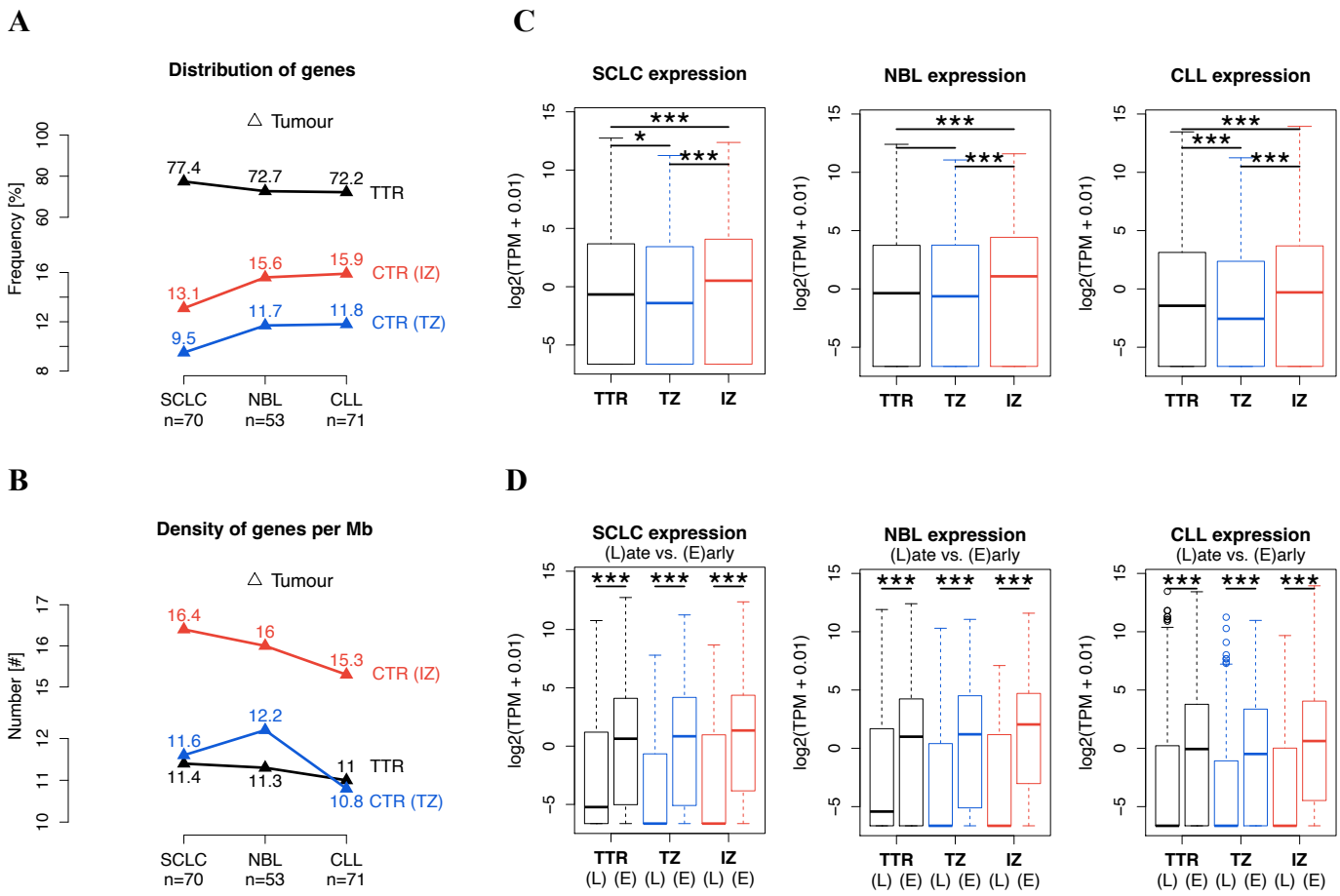


Figure 5.1: Transcriptional activity strongly correlates with the replication domains across three cancer types**.** (A) Distribution of genes mapped to the RFD domains in the indicated cancer genomes, showing that the proportion of IZ genes (red line) is much higher than the proportion of IZ domains (as in 4.2B). (B) Density of genes per Mb plot showing human genes are unevenly clustered at initiation zones. (C) Boxplots of differential gene expression analyses between three replication domains in the indicated cancer transcriptomes, showing IZ genes are more highly expressed than their TZ and TTR counterparts (P < 2.27E-15, as in Figure A.9). (D) Boxplots of differential gene expression analyses between early and late replicating regions in the indicated replication domains, showing early-replicating genes are more highly expressed than their late-replicating counterparts across IZ, TZ and TTR genes (P < 4.4E-58, as in Figure A.10) (Wilcoxon rank-sum test *P < 1E-03, **P < 1E-06, ***P < 1E-09).

is similarly to the proportion of termination zones at the genome level (as in Figure 4.2B). In contrast, genes mapped to the initiation zones accounted for 13.1%, 15.6% and 15.9% of the respective transcriptomes (red line in Figure 5.1A), which is much higher than the proportion of initiation zones at the genome level (as in Figure 4.2B). This indicates that human genes are unevenly and constitutively clustered near replication

origins. On average, the density of IZ genes per Mb is roughly 1.4-fold more than that of TZ and TTR genes across three cancer types (Figure 5.1B). Subsequently, differential gene expression analyses further revealed that IZ genes were ubiquitously and significantly (P < 2.27E-15, Wilcoxon rank-sum test), highly expressed than their TZ and TTR counterparts in the respective cancer transcriptomes (Figure 5.1C; Figure A.9A-B). Therefore, I can confirm that not only human genes are constitutively clustered at initiation zones; genes near initiation zones are also highly transcribed.

To determine whether transcription activities were also associated with earlier replication, I further separated early- and late-replicating regions in the respective cancer genomes (Figure A.10A) and transcriptomes (Figure A.10B). Indeed, early-replicating genes were ubiquitously and significantly (P < 2.44E-58, Wilcoxon rank-sum test), highly expressed than their late-replicating counterparts across IZ, TZ and TTR domains in the respective cancer transcriptomes (Figure 5.1D; Figure A10.D-F). Having earlier demonstrated that I can simultaneously perform replication profiling and RFD mappings from the same cancer cell population, this reaffirms the biological insights underlying our reconstructed replication timing landscapes. As expected, around two thirds (65.3% and 70.3%) of TZ and TTR genes were constitutively located in the early-replicating regions (Table 5.1), in line with the overall average in the human genome (Marchal, Sima, and Gilbert 2019). Intriguingly, I identified that up to 82.3% of IZ genes were located in the early-replicating regions (Table 5.2). Therefore, I can further confirm that DNA replication preferentially initiates at early-replicating IZ genes, in line with the recent OK-seq findings (Petryk et al. 2016; Chen et al. 2019).

Table 5.2: Proportion of genes located in early-replicated regions

RFD (Early)	SCLC	NBL	CLL	Average
IZ	84.1%	82.3%	80.5%	82.3%
TTR	73.6%	69.7%	67.5%	70.3%
TZ	67.9%	66.2%	61.7%	65.3%

Table 5.3: Proportion of genes shared between three cancer types

RFD (Shared)	Early	Late	Total
TZ	15.2%	16.2%	18.4%
IZ	19.9%	13.3%	20.1%
TTR	53.1%	63.8%	64.6%

To further determine whether IZ and TZ genes were cell type-specific genomics properties in the human genome, I interrogated genes that were shared and constitutively conserved between three cancer types in the respective RFD domains. As expected, only 20.1% and 18.4% of IZ and TZ genes, compared to 64.6% of the TTR genes were shared between three cancers types (Table 5.3). This indicates that there are more cell type-specific genes topologically preserved in the initiation and termination zones. Most notably, this repeatedly explains why it should be prudent to use a single reference timing profile originated from an unrelated cell line (Koren et al. 2012, 2014) to study replication timing in multiple cancer types, especially when investigating IZ and TZ genes in recent large-scale cancer sequencing projects (Haradhvala et al. 2016; Li et al. 2019). Therefore, the differential gene expression analyses between the replication initiation and termination domains have not been previously reported in a cell type-specific manner in the human cancer genomes (Figure 5.1C and 5.1D). Collectively, I conclude that my tumour replication-domain landscape is strongly correlated with the transcriptional activities, and this correlation is robust and reproducible across different cancer genomes and transcriptomes from independent studies, thus providing strong support for my bootstrap-based RFD mappings on replication initiation, fork elongation, and termination as *bona fide*.

5.4 Discussion

In this chapter, I show that the landscape of tumour replication domains is strongly coupled with the transcriptional activities in the respective cancer types. My data confirms that human genes are not only unevenly clustered near IZ domains (Figure 5.1A), but IZ genes are also more highly transcribed (Figure 5.1C). In addition, my data

further suggests that more IZ genes are located in the early-replicating regions, compared to their TZ and TTR counterparts (Table 5.2), reaffirming that origin firing preferentially initiates at the highly transcribed genes (Petryk et al. 2016; Chen et al. 2019). Furthermore, our data show that fewer IZ and TZ genes were shared between the three cancer types (Table 5.3), indicating that human genes are topologically preserved in the initiation and termination zones in a cell type-specific manner.

This again highlights that why we must be cautious to use a single reference timing profile to investigate genes near the IZ and TZ regions (Haradhvala et al. 2016). I now show that my tumour replication domains are significantly coupled with the transcriptional activities across three different cancer types, thus providing strong support for our bootstrap-based RFD mappings on where exactly, and how differently, DNA replication is initiated and terminated across human genomes (Figure 5.1A). Therefore, I now report for the first time where exactly transcriptional activities occur near the replication initiation and termination domains, and how differently this occurs in a cell type-specific manner across the three human cancer genomes and transcriptomes (Figure 5.1C and 5.1D). This could provide new insights into the identification of potential cancer target genes that is topologically preserved near the replication origins and termini of specific cancer types (Haradhvala et al. 2016; Chen et al. 2019; Marchal, Sima, and Gilbert 2019).

Chapter 6

Future perspectives

During the S phase of the cell cycle, DNA sequences are replicated in a temporal and spatial order known as the DNA replication timing programme. DNA replication timing implicates in the transcriptional and mutational landscape of the cancer genome, however, where exactly, and how differently, DNA replication is initiated and terminated across different cellular types remains poorly understood in the human genome. Understanding the cell type specificity of replication timing in cancers is essentially important to identify the underlying cellular mechanisms that give rise to cancer and help cancer cells survive. Although sequencing data from cell lines can reflect the replication programme in time, in space, and in disease, appropriate cancer cell lines for the cells of origin of specific primary cancers are limited. This has hampered the possibility to investigate the most cell type-specific genomic properties; that is, replication initiation and termination domains of the cancer genome.

DNA sequences that underwent somatic alterations in primary cancers have been comprehensively studied in recent large-scale cancer sequencing projects. However, the rest of DNA sequences that accurately replicated in the same primary cancer genomes have not been widely explored, and almost nothing known about their roles in replication timing.

In this study, I present a novel *in silico* framework to assess both the temporal dynamics and the spatial landscape of the replication timing programme directly from primary tumour samples rather than from cancer cell lines. I show that the cellular plasticity of tumour replication timing is topologically preserved in closely related normal tissues, as

well as in lineage-specific cancer cell lines. Unexpectedly, I find that the genome-wide distribution of termination events is closely coordinated with the initiation activities in both the normal and cancer genomes. Importantly, I demonstrate that the landscape of tumour replication domains is significantly coupled with the transcriptional activities in the respective cancer types.

In Chapter 2, my results highlight that the majority of primary cancer genomes are accurately replicated during cancer cell proliferation, thus provide a snapshot in time and space of the replication programme. Based on my *in silico* sorting predictions of primary tumours, their overall sequencing read depth patterns can also be directly used to infer S phase cell fraction in a cell population. However, my data suggest that it is the G1 phase reads in a cell population correlate more strongly, but negatively with the replication timing profiles (Section 2.3.1), thus challenging the direct use of S phase reads as *de facto* timing profiles (Koren et al. 2014; Marchal et al. 2018), and further emphasizing the indispensible role of G1 phase reads when profiling DNA replication. Upon adapting the canonical S/G1 read depth ratio method in Chapter 3, my direct replication profiling approach allows one to average out the individual- or sequence-specific replication timing in a cell population (Ryba et al. 2012; Koren et al. 2014; Sasaki et al. 2017), as well as an unknown background signal, both of which have not been properly addressed in the alternative profiling method, as noted earlier. Together, the replication timing programme recapitulated by my direct profiling approach is not only intrinsically conserved across different normal and cancer genomes (Section 3.3.3), but is also topologically preserved in distinct cancer types, e.g. small cell lung cancers and neuroblastomas (Section 3.3.2).

Moving on to Chapter 4, my results also highlight that replication termini are determined by, and located between two activating origin firings in the human genome. Previous analyses of nascent Okazaki fragments (OK-seq) from the lagging strands may lead to seemingly over-represented termination events compared to the initiation activities (Petryk et al. 2016). However, my novel bootstrap-based RFD mapping is advantageous, since it allows one to mathematically model the bi-directional replication using sequences from both the leading and lagging strands. Unexpectedly, my data suggest that the genome-wide distribution of termination events is closely coordinated

with the initiation activities in both the normal and cancer genomes (Section 4.3.2), which is consistent with the consensus notion reported in yeast (Hawkins et al. 2013), but has not been previously reported in humans using only nascent lagging strands by the directional sequencing of Okazaki fragments (OK-seq) *in vitro* (Petryk et al. 2016; Zhao, Sasaki, and Gilbert 2020). Notably, I consistently reproduced this coordinated distribution in two randomly down-sampled test populations, as well as in shared domains between the two test populations (Section 4.3.2), thus supporting a *bona fide* mapping on replication initiation, progression, and termination by our bootstrap-based RFD method.

Finally, my findings highlight that replication origins and termini are among the most cell type-specific genomic properties in the human genome. Furthermore, my data show that fewer IZ and TZ genes were shared between the three cancer types (Table 5.2), indicating that human genes are topologically preserved in the initiation and termination zones in a cell type-specific manner. This again highlights that why we must be cautious to use a single one-size-fits-all reference timing profile to study multiple cancer types (Koren et al. 2014; Polak et al. 2015; Li et al. 2019), especially when investigating IZ and TZ genes (Haradhvala et al. 2016). Importantly, I now report for the first time where exactly transcriptional activities occur near the replication initiation and termination domains, and how differently this occurs in a cell type-specific manner across the three human cancer genomes and transcriptomes (Figure 5.1C and 5.1D).

All together, my novel *in silico* framework allows one to assess the tumour replication timing programme directly from primary cancer genomes without the need for *in vitro* sorting procedures, and therefore opens up opportunities for the increasing numbers of whole-genome sequencing studies published by the broader cancer genomics community. The cellular plasticity of the tumour replication programme recapitulated by my novel *in silico* framework adds a new spatiotemporal perspective to the three-dimensional human cancer genome, thus could provide new insights into the identification of potential cancer target genes that is topologically preserved near the replication origins and termini of specific cancer types (Haradhvala et al. 2016; Chen et al. 2019; Marchal, Sima, and Gilbert 2019).

6.1 Code availability

R scripts purpose-written for this project are available in the public domain at https://github.com/tsunpo/R under the GNU General Public License v3.0. Further details on the computational and statistical approaches can be found in code comments.

6.2 Data availability

Tumour replication timing (RT) profiles and RFD domain mppings for each chromosome in the respective cancer types are available at Mendeley Data https://data.mendeley.com/datasets/cj3gt6fz7y/draft?a=7b2e4996-2269-4846-91f5-1750eb3d5f6a

Appendix A

Supplementary Figures

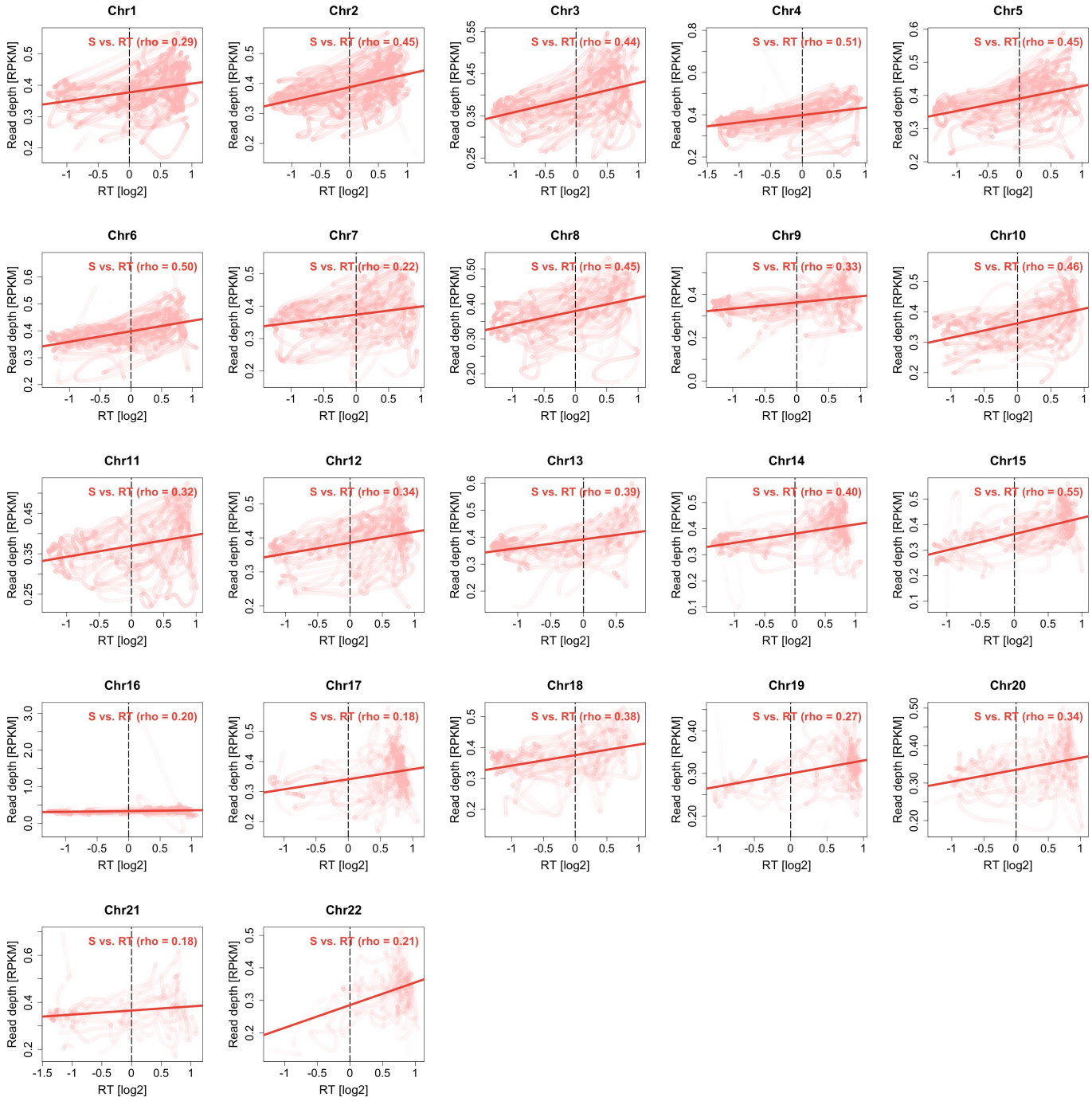


Figure A.1: A positive correlation trend between S phase read depth pattern and the reference timing profile across 22 human autosomes. Scatterplots demonstrating relationship between smoothed S phase read depth (y-axis) and smoothed replication timing (RT) profile (log2 S/G1 ratio; x-axis) across seven, non-synchronized lymphoblastoid cell lines (LCLs; with one repetition) from Koren et al. 2012 (rho = Spearman's rank correlation coefficient; solid regression line = Linear regression)

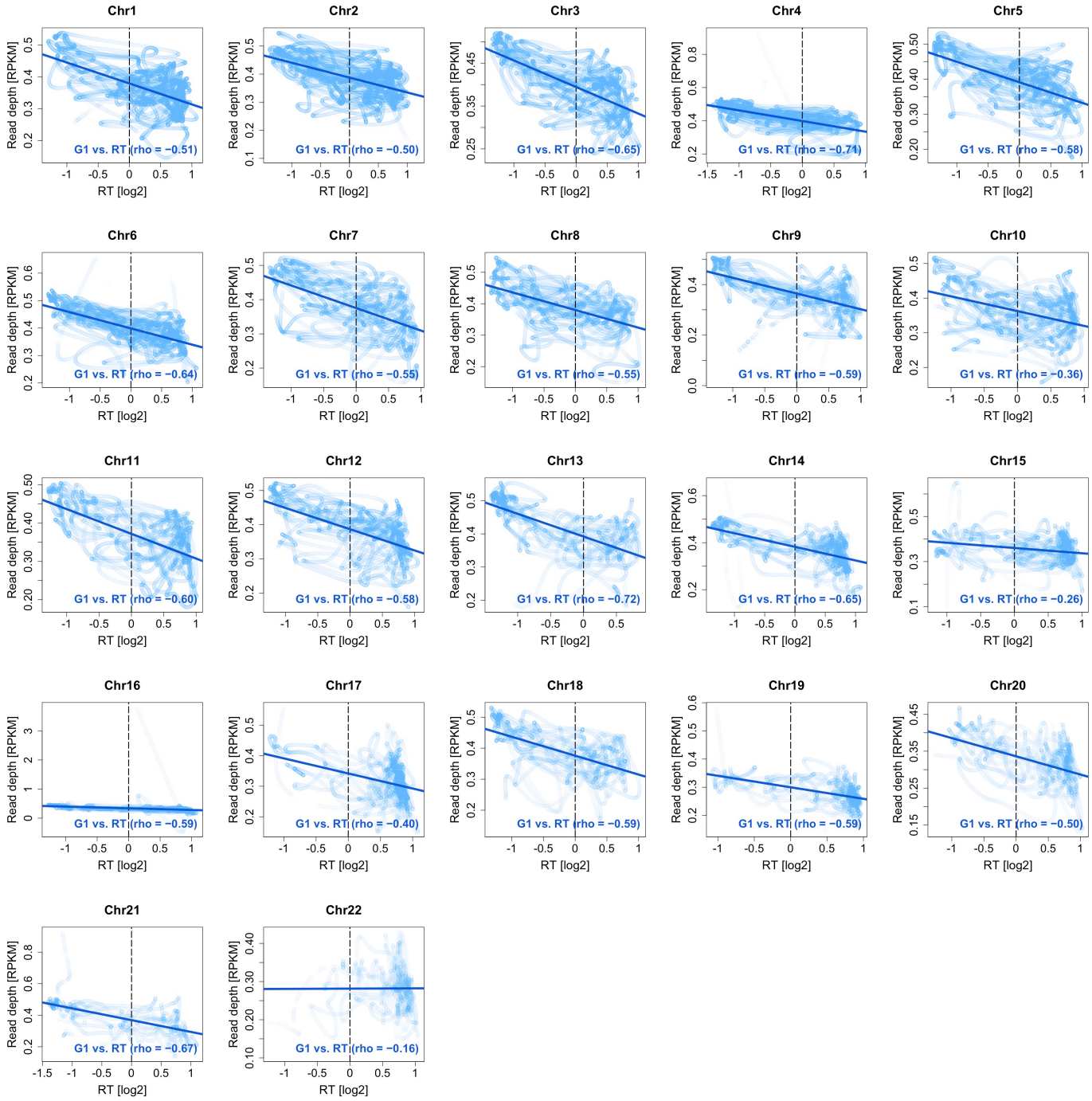


Figure A.2: A negative correlation trend between G1 phase read depth pattern and the reference timing profile across 22 human autosomes. Scatterplots demonstrating relationship between smoothed G1-phase read depth (y-axis) and smoothed RT profile (log2 S/G1 ratio; x-axis) across seven, non-synchronized LCLs from Koren et al. 2012 (rho = Spearman's rank correlation coefficient; solid regression line = Linear regression)

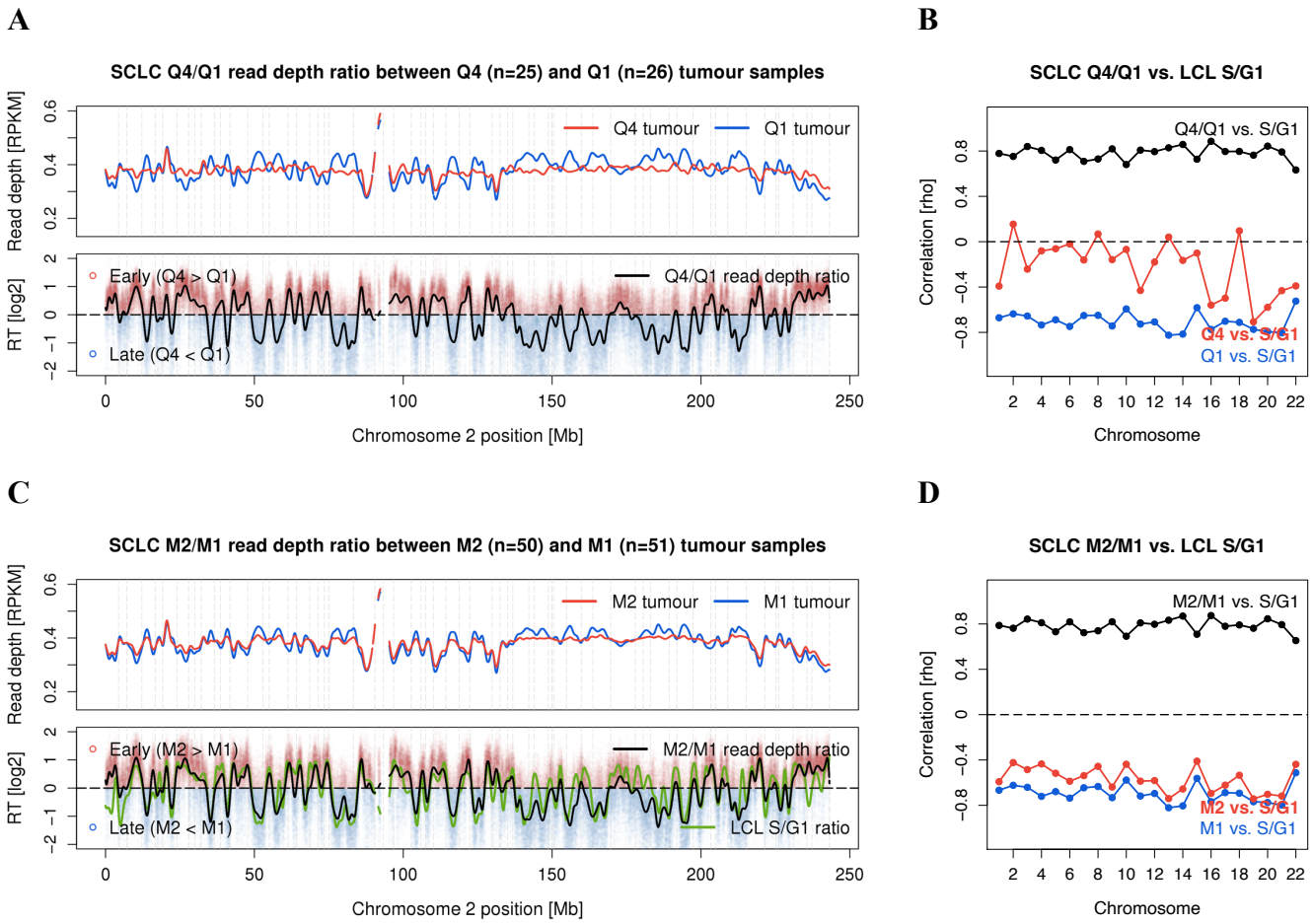


Figure A.3: Direct profiling of tumour replication timing from 101 SCLC tumour samples using primary cancer whole genomes. (A and B) (A) SCLC tumour timing profile for chromosome 2 as inferred between Q4 (the forth quartile) and Q1 (the first quartile) tumour samples (as in Figure 2.4C). (B) Correlations between SCLC subgroup reads and the LCL reference timing profile across 22 autosomes. (C and D) (C) SCLC tumour timing profile for chromosome 2 as inferred between M2 (the second median; Q4+Q3) and M1 (the first median; Q2+Q1) tumour samples. (D) Correlation analyses as in (B)

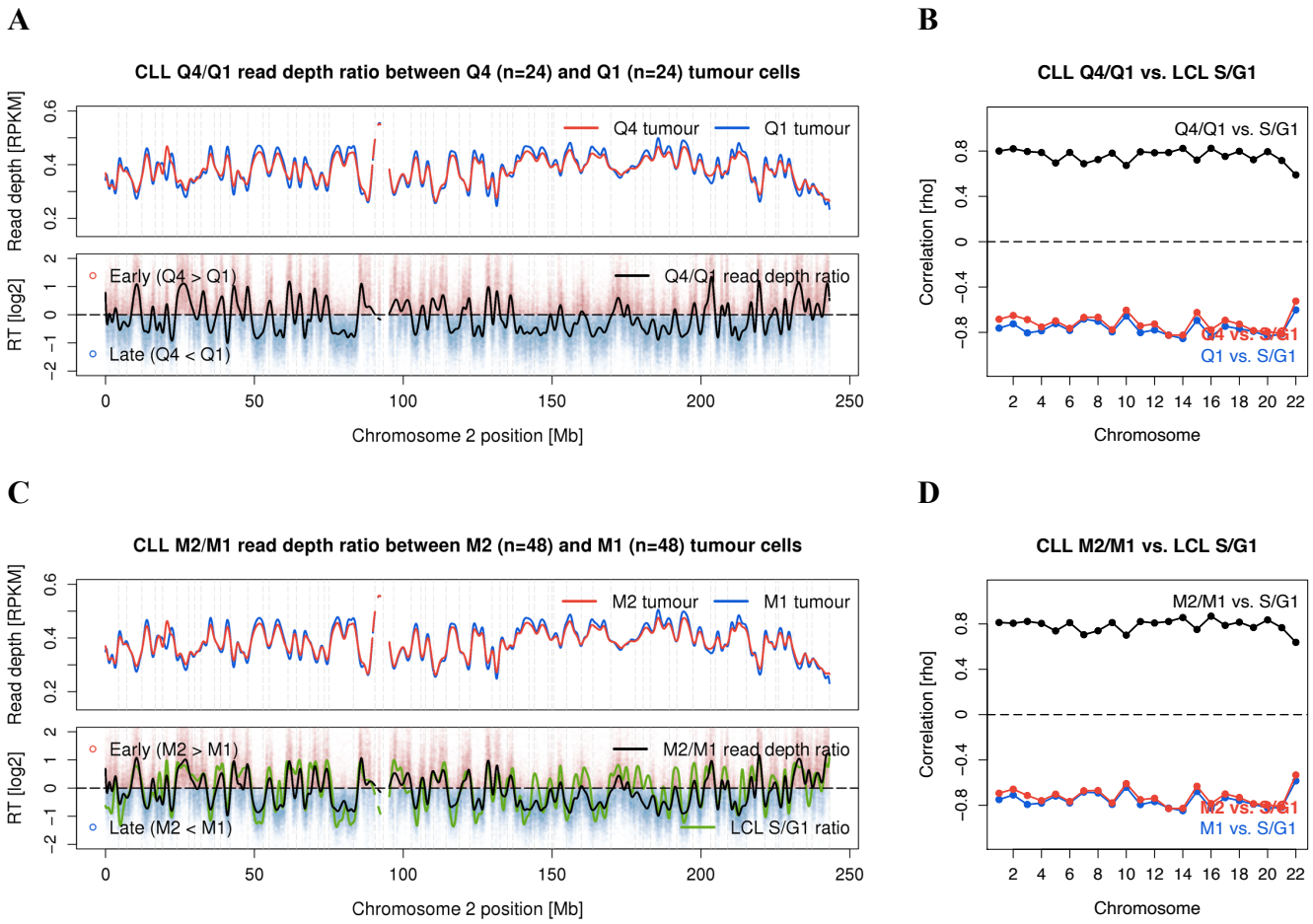


Figure A.4: Direct profiling of tumour replication timing from 96 CLL tumour cells using primary cancer whole genomes. (A and B) (A) CLL tumour timing profile for chromosome 2 as inferred between Q4 (the forth quartile) and Q1 (the first quartile) tumour cells (as in Figure 2.4C). (B) Correlations between CLL subgroup reads and the LCL reference timing profile across 22 autosomes. (C and D) (C) CLL tumour timing profile for chromosome 2 as inferred between M2 (the second median; Q4+Q3) and M1 (the first median; Q2+Q1) tumour cells. (D) Correlation analyses as in (B)

Figure 2.4A

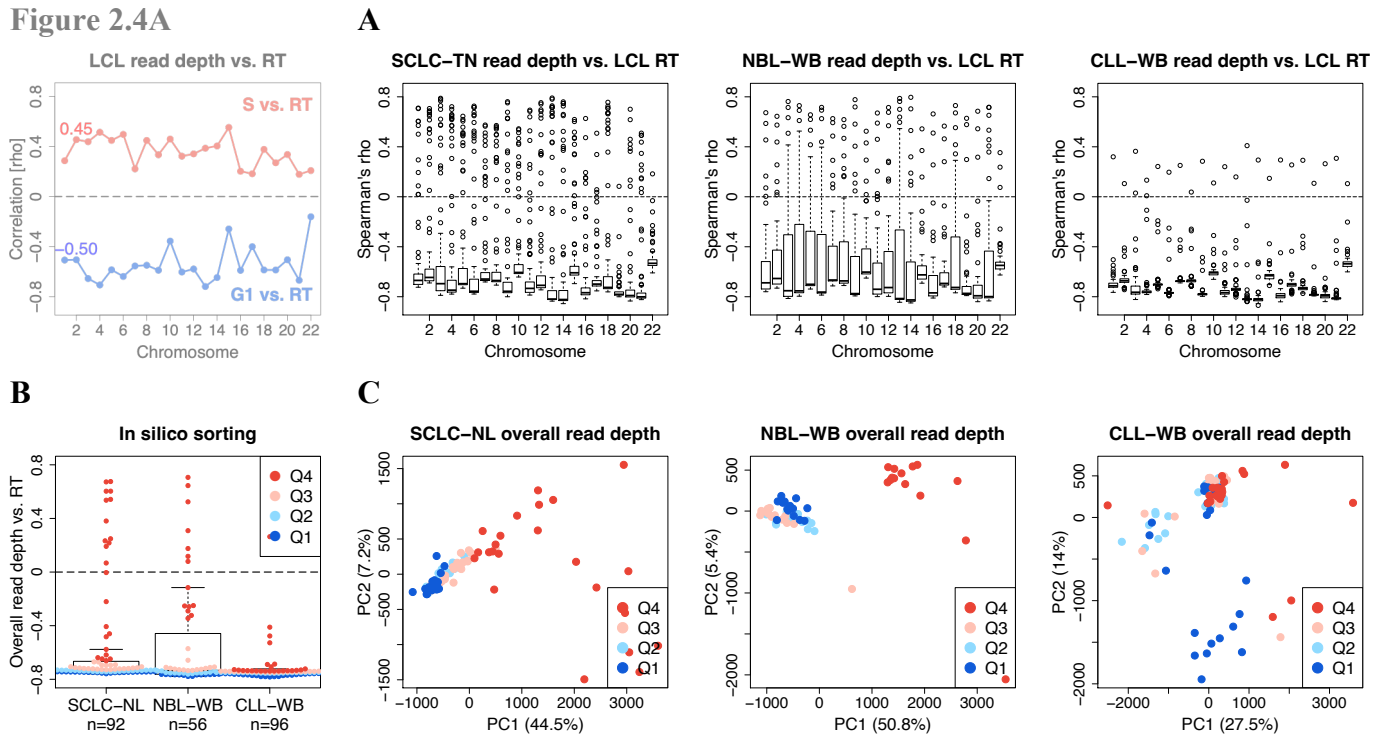


Figure A.5: *In silico* sorting of matched normal samples using whole-genome sequences. (A) Correlations between read depth patterns and the reference RT profile across 22 autosomes in three matched normal samples, consistent with the adversarial correlation trend in the LCL reference genome (as in Figure 2.4A). SCLC-NL: adjacent non-neoplastic lung tissues derived from the same 92 SCLC patients; NBL-WB: whole blood samples from the same 56 NBL patients; CLL-WB: whole blood samples from the same 96 CLL patients. (B and C) (B) Boxplot showing proposed *in silico* sorting predictions across the three matched normals. (C) Unsupervised principal component analysis (PCA) on normal samples' overall read depth patterns, showing the direction of the first principal component (PC1; x-axis) is generally in parallel to the distribution of quartile subgroups in the SCLC-NL normal lung tissues. However, matched whole bloods from both 56 NBL-WB and 96 CLL-WB samples were shown unpredictable in our *in silico* sorting procedure, and were therefore precluded from further analyses. This is because peripheral blood cells are well known to contain multiple cell types (Tsaprouni et al. 2014), and generally leave the cell cycle upon differentiation (Koren et al. 2014).

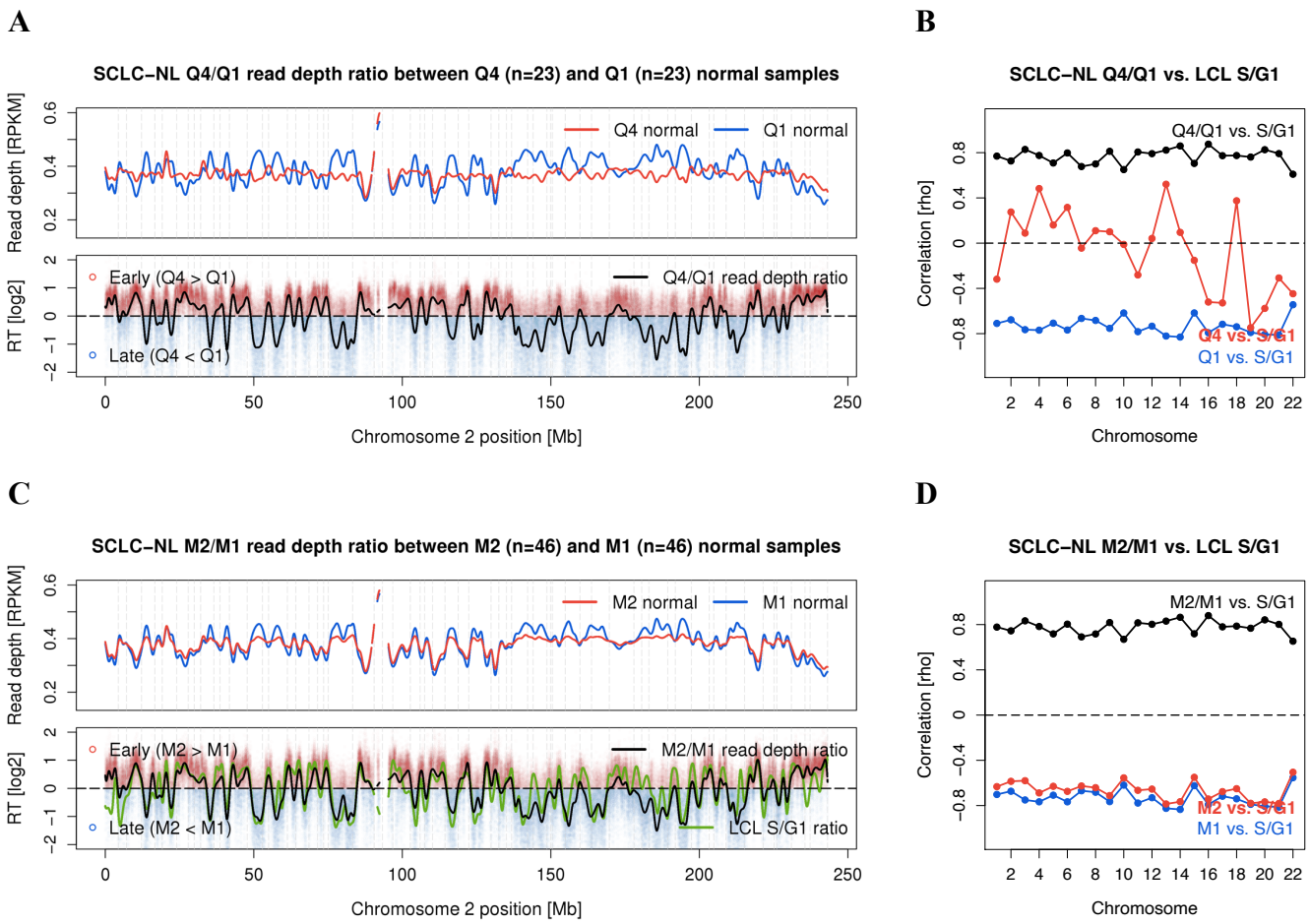


Figure A.6: Direct profiling of normal replication timing from 92 SCLC-NL matched normal samples using normal whole genomes. (A and B) (A) SCLC-NL normal timing profile for chromosome 2 as inferred between Q4 (the forth quartile) and Q1 (the first quartile) tumour samples (as in Figure 2.4C). (B) Correlations between SCLC-NL subgroup reads and the LCL reference timing profile across 22 autosomes. (C and D) (C) SCLC-NL normal timing profile for chromosome 2 as inferred between M2 (the second median; Q4+Q3) and M1 (the first median; Q2+Q1) normal samples. (D) Correlation analyses as in (B)

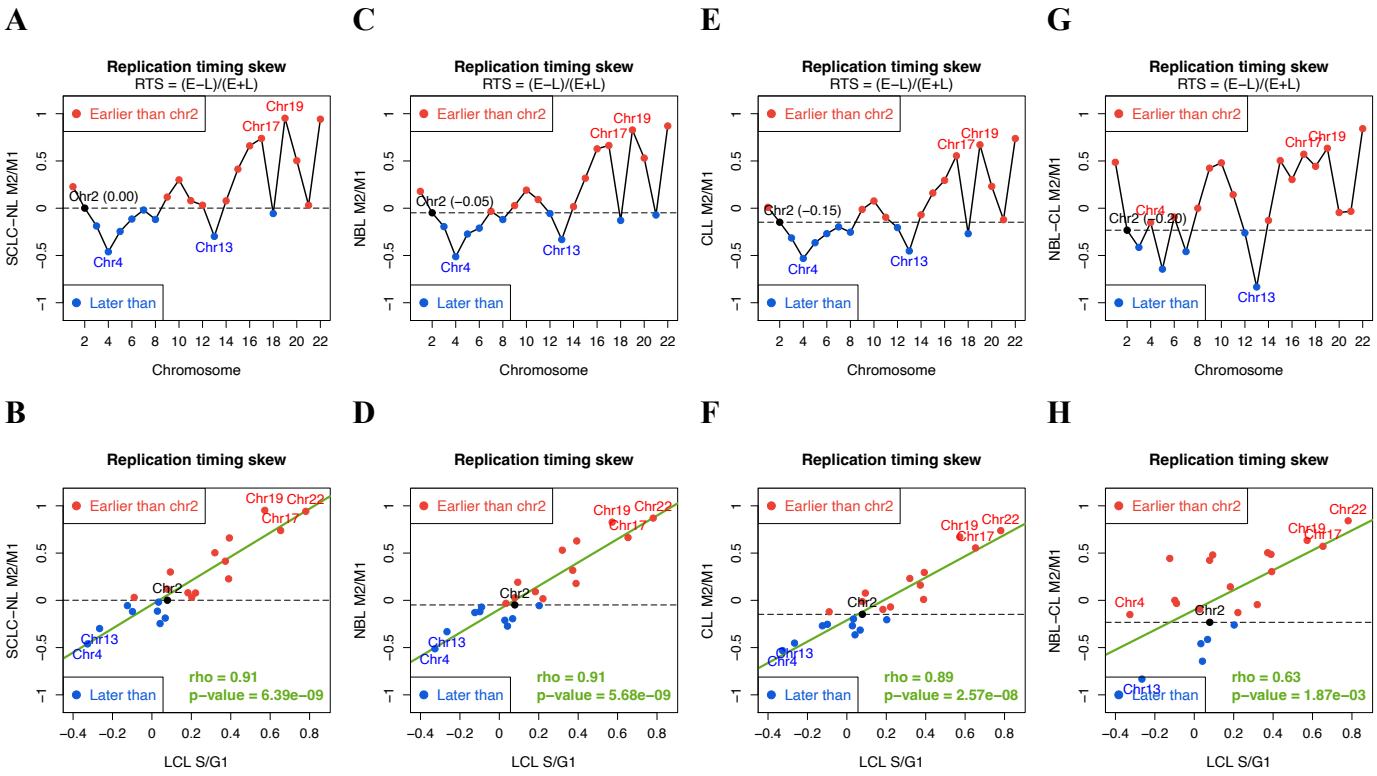


Figure A.7: Universal early-and-late replication division identified across different normal and cancer genomes, related to Figure 3.3D. Replication timing skew (RTS) values derived from (A and B) SCLC-NL normal genome, (C and D) NBL, and (E and F) CLL cancer genomes were significantly correlated with those from the LCL reference genome. It is worth noting that, the RTS values of chromosome 2 in SCLC, NBL and CLL (0, -0.05, -0.15) also reflected our *in silico* sorting predictions as described earlier, and further suggested that this early-to-late ratio is correlated with the proportions of S phase cell in a cell population.

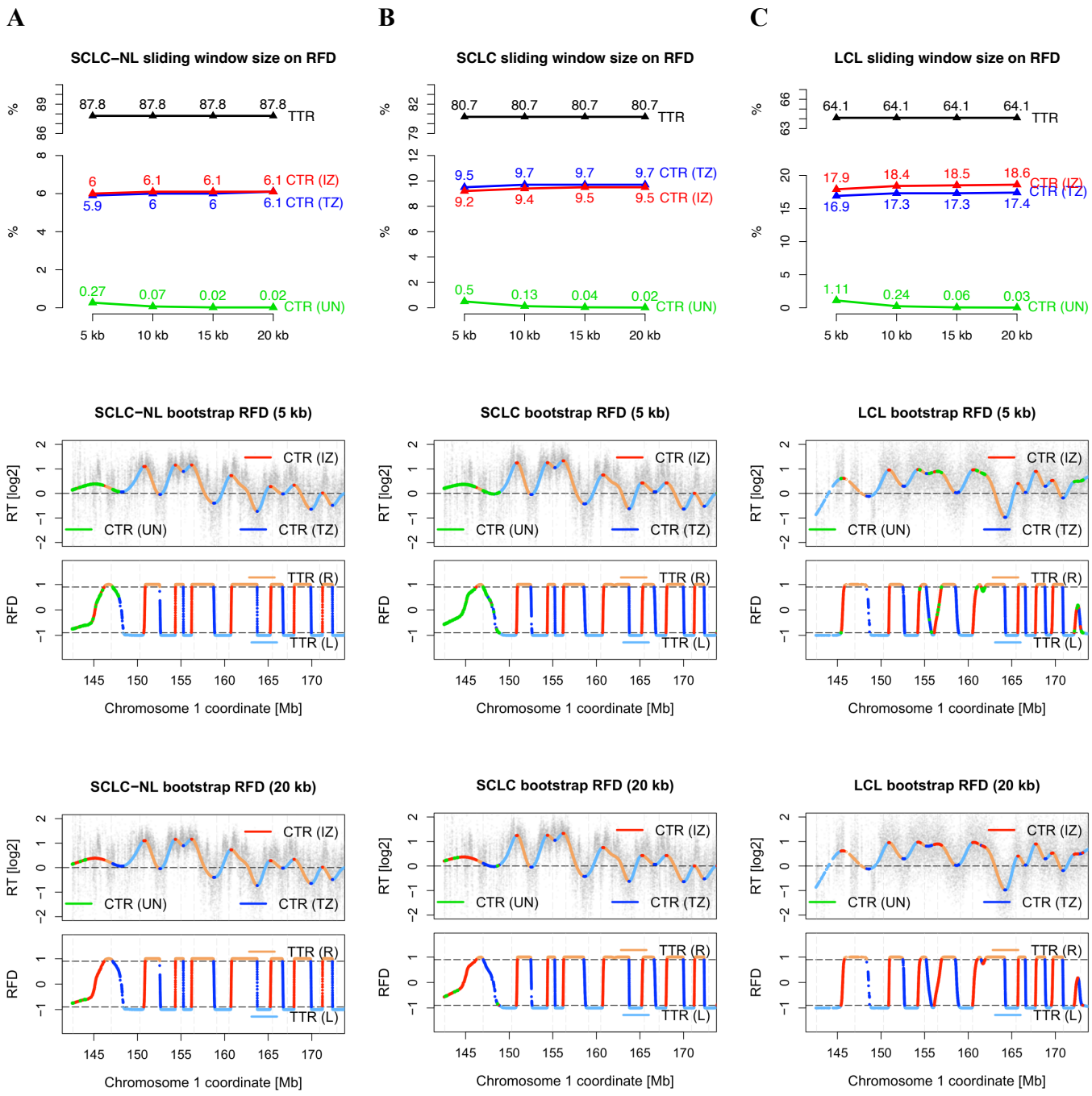


Figure A.8: Only around 0.02% of the genome is unmappable across the autosomes, related to Figure 6B

Figure 4.2B

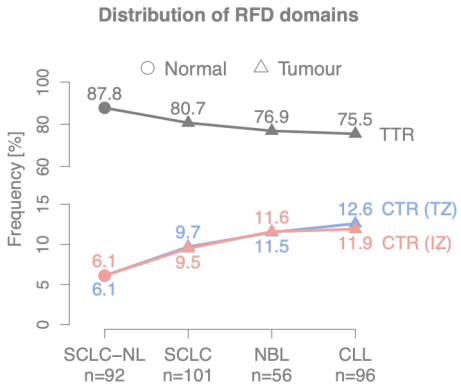


Figure 5.1A

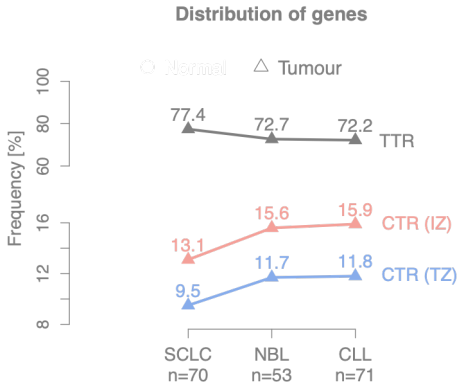
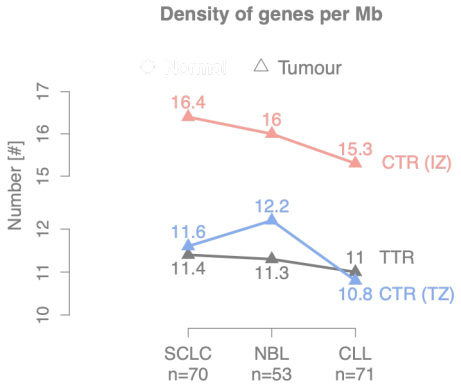


Figure 5.1B



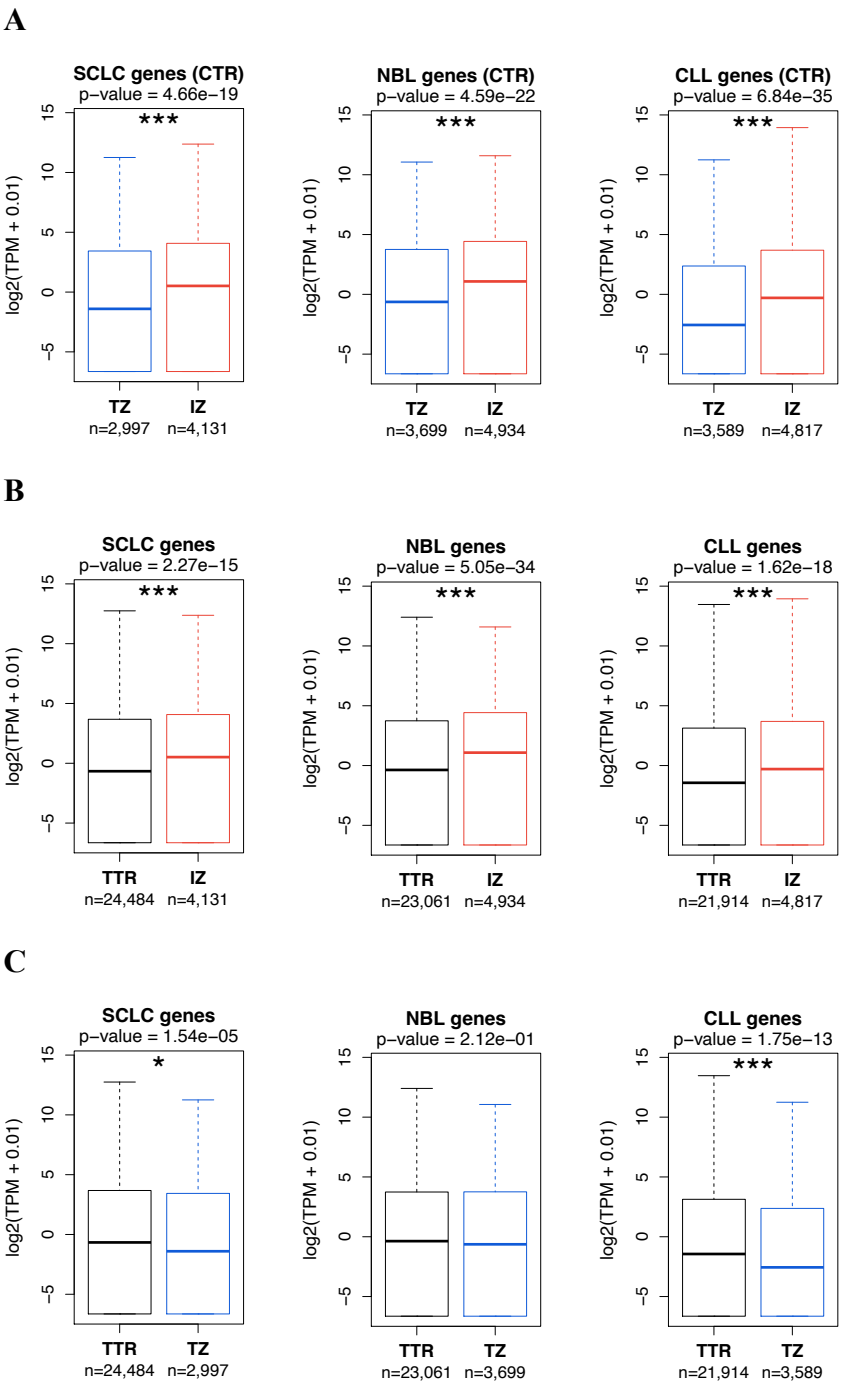


Figure A.9: Differential gene expression between three replication domains, related to Figure 5.1C. (A and B) Pairwise differential gene expression analyses between replication domains revealed that IZ genes are ubiquitously and significantly, highly expressed than TZ and TTR genes. (A and C) TZ genes ubiquitously have the lowest expression level in the respective transcriptomes, when compared to the IZ and TTR genes (Wilcoxon rank-sum test **P* < 1E-03, ***P* < 1E-06, ****P* < 1E-9).

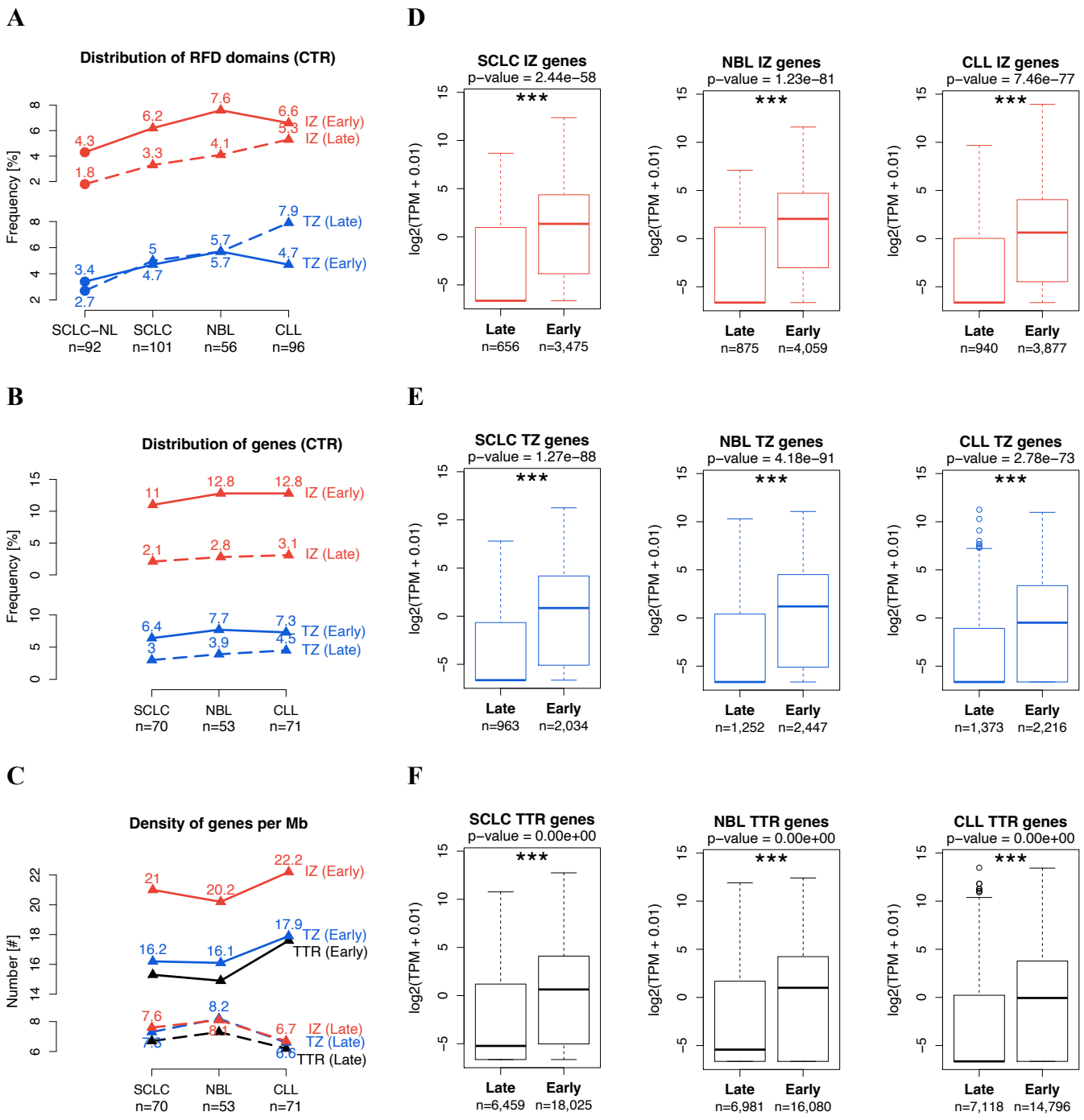


Figure A.10: Differential gene expression between early and late replicating regions, related to Figure 5.1D. (A and B) Distribution of early- and late-replicating regions in the respective (A) cancer genomes and (B) transcriptomes. (C) Intriguingly, we identified that the density of genes at late-replicating regions exhibited no difference across three RFD domains (dashed lines), but showed stark difference in IZ genes at early-replicating regions (solid lines). (D, E and F) Pairwise differential gene expression analyses between early- and late-replication regions in the respective RFD domains (Wilcoxon rank-sum test *P < 1E-03, **P < 1E-06, ***P < 1E-9).

List of abbreviations

CLL	chronic lymphocytic leukemia
CTR	constant timing region
G1 phase	gap 1 phase
IZ	initiation zone
kb	kilobase
LCL	lymphoblastoid cell lines
M2	second median
M1	first median
Mb	megabase
NBL	neuroblastomas
OK-seq	Okazaki fragment sequencing
PCA	principle component analysis
Q4	forth quartile
Q1	first quartile
RPKM	reads per kilobase, per million
RT	replication timing
RTS	replication timing skew
RFD	replication fork directionality
SCLC	small cell lung cancer
SNR	signal-to-noise ratio
S phase	synthesis phase
TPM	transcripts per million
TTR	timing transition region
TZ	termination zone

Bibliography

Aria, Valentina, and Joseph T.P. Yeeles. 2019. “Mechanism of Bidirectional Leading-Strand Synthesis Establishment at Eukaryotic DNA Replication Origins.” *Molecular Cell* 73 (2): 199-211.e10. https://doi.org/10.1016/j.molcel.2018.10.019.

Audit, Benjamin, Antoine Baker, Chun Long Chen, Aurélien Rappailles, Guillaume Guilbaud, Hanna Julienne, Arach Goldar, et al. 2013. “Multiscale Analysis of Genome-Wide Replication Timing Profiles Using a Wavelet-Based Signal-Processing Algorithm.” *Nature Protocols* 8 (1): 98–110. https://doi.org/10.1038/nprot.2012.145.

Bartholdy, Boris, Rituparna Mukhopadhyay, Julien Lajugie, Mirit I Aladjem, and Eric E Bouhassira. 2015. “Allele-Specific Analysis of DNA Replication Origins in Mammalian Cells.” *Nature Communications* 6 (May): 1–12. https://doi.org/10.1038/ncomms8051.

Besnard, Emilie, Amélie Babled, Laure Lapasset, Ollivier Milhavet, Hugues Parrinello, Christelle Dantec, Jean Michel Marin, and Jean Marc Lemaitre. 2012. “Unraveling Cell Type-Specific and Reprogrammable Human Replication Origin Signatures Associated with G-Quadruplex Consensus Motifs.” *Nature Structural and Molecular Biology* 19 (8): 837–44. https://doi.org/10.1038/nsmb.2339.

Bray, Nicolas L, Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Burkhart, Deborah L, and Julien Sage. 2008. “Cellular Mechanisms of Tumour Suppression by the Retinoblastoma Gene.” *Nature Reviews. Cancer* 8 (9): 671–82. https://doi.org/10.1038/nrc2399.

Campbell, Peter J., Gad Getz, Joshua M. Stuart, Jan O. Korbel, Lincoln D. Stein, and - ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net. 2020. “Pan-Cancer Analysis of Whole Genomes Consortium.” *Nature* 578 (June): 82–93. https://doi.org/10.1101/162784.

Chen, Yu Hung, Sarah Keegan, Malik Kahli, Peter Tonzi, David Fenyö, Tony T. Huang, and Duncan J. Smith. 2019. “Transcription Shapes DNA Replication Initiation and Termination in Human Cells.” *Nature Structural and Molecular Biology* 26 (1): 67–77. https://doi.org/10.1038/s41594-018-0171-0.

Cun, Yupeng, Tsun Po Yang, Viktor Achter, Ulrich Lang, and Martin Peifer. 2018. “Copy-Number Analysis and Inference of Subclonal Populations in Cancer Genomes Using Sclust.” *Nature Protocols* 13 (6): 1488–1501. https://doi.org/10.1038/nprot.2018.033.

Dietlein, Felix, Donate Weghorn, Amaro Taylor-Weiner, André Richters, Brendan Reardon, David Liu, Eric S Lander, Eliezer M Van Allen, and Shamil R Sunyaev. 2018. “Discovery of Cancer Driver Genes Based on Nucleotide Context.” *BioRxiv*

52 (FebruarY): 485292. https://doi.org/10.1038/s41588-019-0572-y.
Du, Qian, Saul A. Bert, Nicola J. Armstrong, C. Elizabeth Caldon, Jenny Z. Song, Shalima S. Nair, Cathryn M. Gould, et al. 2019. "Replication Timing and Epigenome Remodelling Are Associated with the Nature of Chromosomal Rearrangements in Cancer." *Nature Communications* 10 (1): 1–15. https://doi.org/10.1038/s41467-019-08302-1.
Fragkos, Michalis, Olivier Ganier, Philippe Coulombe, and Marcel Méchali. 2015. "DNA Replication Origin Activation in Space and Time." *Nature Reviews. Molecular Cell Biology* 16 (6): 360–74. https://doi.org/10.1038/nrm4002.
George, Julie, Jing Shan Lim, Se Jin Jang, Yupeng Cun, Luka Ozretić, Gu Kong, Frauke Leenders, et al. 2015. "Comprehensive Genomic Profiles of Small Cell Lung Cancer." *Nature*. https://doi.org/10.1038/nature14664.
Hahn, William C, Joel S Bader, Theodore P Braun, Andrea Califano, Paul A Clemons, Brian J Druker, Andrew J Ewald, et al. 2021. "An Expanded Universe of Cancer Targets." *Cell* 184 (5): 1142–55. https://doi.org/10.1016/j.cell.2021.02.020.
Hamperl, Stephan, and Karlene A Cimprich. 2016. "Conflict Resolution in the Genome: How Transcription and Replication Make It Work." *Cell*. Elsevier. https://doi.org/10.1016/j.cell.2016.09.053.
Hanahan, Douglas, and Robert a. Weinberg. 2011. "Hallmarks of Cancer: The next Generation." *Cell* 144 (5): 646–74. https://doi.org/10.1016/j.cell.2011.02.013.
Hansen, R. S., S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, and J. A. Stamatoyannopoulos. 2010. "Sequencing Newly Replicated DNA Reveals Widespread Plasticity in Human Replication Timing." *Proceedings of the National Academy of Sciences* 107 (1): 139–44. https://doi.org/10.1073/pnas.0912402107.
Haradhvala, Nicholas J., Paz Polak, Petar Stojanov, Kyle R. Covington, Eve Shinbrot, Julian M. Hess, Esther Rheinbay, et al. 2016. "Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair." *Cell* 164 (3): 538–49. https://doi.org/10.1016/j.cell.2015.12.050.
Hawkins, Michelle, Renata Retkute, Carolin A. Müller, Nazan Saner, Tomoyuki U. Tanaka, Alessandro P.S. deMoura, and Conrad A. Nieduszynski. 2013. "High-Resolution Replication Profiles Define the Stochastic Nature of Genome Replication Initiation and Termination." *Cell Reports* 5 (4): 1132–41. https://doi.org/10.1016/j.celrep.2013.10.014.
Jolly, Clemency, and Peter Van Loo. 2018. "Timing Somatic Events in the Evolution of Cancer." *Genome Biology* 19 (1): 1–9. https://doi.org/10.1186/s13059-018-1476-3.
Klein, Kyle N., Peiyao A. Zhao, Xiaowen Lyu, Daniel A. Bartlett, Amar Singh, Ipek Tasan, Lotte P. Watts, et al. 2019. "Replication Timing Maintains the Global Epigenetic State in Human Cells." *BioRxiv* 378 (April): 371–78. https://doi.org/10.1101/2019.12.28.890020.
Koren, Amnon, Robert E. Handsaker, Nolan Kamitaki, Rosa Karlić, Sulagna Ghosh, Paz Polak, Kevin Eggan, and Steven A. McCarroll. 2014. "Genetic Variation in Human DNA Replication Timing." *Cell* 159 (5): 1015–26. https://doi.org/10.1016/j.cell.2014.10.025.
Koren, Amnon, Paz Polak, James Nemesh, Jacob J Michaelson, Jonathan Sebat, Shamil R Sunyaev, and Steven A. McCarroll. 2012. "Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation." *American Journal of Human Genetics* 91 (6): 1033–40.

https://doi.org/10.1016/j.ajhg.2012.10.018.
Lawrence, Michael S, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, et al. 2013. “Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes.” *Nature* 499 (7457): 214–18. https://doi.org/10.1038/nature12213.
Li, Yilong, Nicola D Roberts, Joachim Weischenfeldt, Jeremiah A Wala, Ofer Shapira, Steven E Schumacher, Ekta Khurana, et al. 2019. “Patterns of Somatic Structural Variation in Human Cancer Genomes.” *Nature* 578 (February).
Malone, Clare F., Neekesh V. Dharia, Guillaume Kugener, Alexandra B. Forman, Michael V. Rothberg, Mai Abdusamad, Alfredo Gonzalez, et al. 2021. “Selective Modulation of a Pan-Essential Protein as a Therapeutic Strategy in Cancer.” *Cancer Discovery*. https://doi.org/10.1158/2159-8290.cd-20-1213.
Mangel, Marc, and Francisco J. Samaniego. 1984. “Abraham Wald’s Work on Aircraft Survivability.” *Journal of the American Statistical Association* 79 (386): 259–67. https://doi.org/10.1080/01621459.1984.10478038.
Marchal, Claire, Takayo Sasaki, Daniel Vera, Korey Wilson, Jiao Sima, Juan Carlos Rivera-Mulia, Claudia Trevilla-García, Coralin Nogues, Ebtesam Nafie, and David M. Gilbert. 2018. “Genome-Wide Analysis of Replication Timing by next-Generation Sequencing with E/L Repli-Seq.” *Nature Protocols* 13 (5): 819–39. https://doi.org/10.1038/nprot.2017.148.
Marchal, Claire, Jiao Sima, and David M. Gilbert. 2019. “Control of DNA Replication Timing in the 3D Genome.” *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/s41580-019-0162-y.
Martincorena, Iñigo, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, and Peter J. Campbell. 2017. “Universal Patterns of Selection in Cancer and Somatic Tissues.” *Cell* 171 (5): 1029-1041.e21. https://doi.org/10.1016/j.cell.2017.09.042.
Masters, John R.W. 2000. “Human Cancer Cell Lines: Fact and Fantasy.” *Nature Reviews Molecular Cell Biology* 1 (3): 233–36. https://doi.org/10.1038/35043102.
McGuffee, Sean R., Duncan J. Smith, and Iestyn Whitehouse. 2013. “Quantitative, Genome-Wide Analysis of Eukaryotic Replication Initiation and Termination.” *Molecular Cell* 50 (1): 123–35. https://doi.org/10.1016/j.molcel.2013.03.004.
Miga, Karen H., Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, et al. 2020. “Telomere-to-Telomere Assembly of a Complete Human X Chromosome.” *Nature* 585 (7823): 79–84. https://doi.org/10.1038/s41586-020-2547-7.
Nik-Zainal, Serena, and Benjamin A. Hall. 2019. “Cellular Survival over Genomic Perfection.” *Science* 366 (6467): 802–3. https://doi.org/10.1126/science.aax8046.
Nik-Zainal, Serena, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, et al. 2012. “The Life History of 21 Breast Cancers.” *Cell* 149 (5): 994–1007. https://doi.org/10.1016/j.cell.2012.04.023.
Peifer, Martin, Lynnette Fernández-Cuesta, Martin L. Sos, Julie George, Danila Seidel, Lawryn H. Kasper, Dennis Plenker, et al. 2012. “Integrative Genome Analyses Identify Key Somatic Driver Mutations of Small-Cell Lung Cancer.” *Nature Genetics* 44 (10): 1104–10. https://doi.org/10.1038/ng.2396.
Peifer, Martin, Falk Hertwig, Frederik Roels, Daniel Dreidax, Moritz Gartlgruber, Roopika Menon, Andrea Krämer, et al. 2015. “Telomerase Activation by Genomic

Rearrangements in High-Risk Neuroblastoma." *Nature*. https://doi.org/10.1038/nature14980.
Petryk, Nataliya, Malik Kahli, Yves D'Aubenton-Carafa, Yan Jaszczyszyn, Yimin Shen, Maud Silvain, Claude Thermes, Chun Long Chen, and Olivier Hyrien. 2016. "Replication Landscape of the Human Genome." *Nature Communications* 7: 1–13. https://doi.org/10.1038/ncomms10208.
Pimentel, Harold, Nicolas L. Bray, Suzette Puente, Páll Melsted, and Lior Pachter. 2017. "Differential Analysis of RNA-Seq Incorporating Quantification Uncertainty." *Nature Methods* 14 (7): 687–90. https://doi.org/10.1038/nmeth.4324.
Polak, Paz, Rosa Karlić, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S. Lawrence, Alex Reynolds, et al. 2015. "Cell-of-Origin Chromatin Organization Shapes the Mutational Landscape of Cancer." *Nature* 518 (7539): 360–64. https://doi.org/10.1038/nature14221.
Pope, Benjamin D., Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L. Vera, et al. 2014. "Topologically Associating Domains Are Stable Units of Replication-Timing Regulation." *Nature* 515 (7527): 402–5. https://doi.org/10.1038/nature13986.
Pourkarimi, Ehsan, James M Bellush, and Iestyn Whitehouse. 2016. "Spatiotemporal Coupling and Decoupling of Gene Transcription with DNA Replication Origins during Embryogenesis in *C. Elegans*." *ELife* 5: 1–12. https://doi.org/10.7554/eLife.21728.
Puente, Xose S., Magda Pinyol, Víctor Quesada, Laura Conde, Gonzalo R. Ordóñez, Neus Villamor, Georgia Escaramis, et al. 2011. "Whole-Genome Sequencing Identifies Recurrent Mutations in Chronic Lymphocytic Leukaemia." *Nature* 475 (7354): 101–5. https://doi.org/10.1038/nature10113.
Pyatnitskiy, Mikhail, Dmitriy Karpov, Ekaterina Poverennaya, Andrey Lisitsa, and Sergei Moshkovskii. 2015. "Bringing down Cancer Aircraft: Searching for Essential Hypomutated Proteins in Skin Melanoma." *PLoS ONE* 10 (11): 1–14. https://doi.org/10.1371/journal.pone.0142819.
Reijns, Martin A M, Harriet Kemp, James Ding, Sophie Marion de Procé, Andrew P Jackson, and Martin S Taylor. 2015. "Lagging-Strand Replication Shapes the Mutational Landscape of the Genome." *Nature* 518 (7540): 1–17. https://doi.org/10.1038/nature14183.
Rhind, Nicholas, and David M. Gilbert. 2013. "DNA Replication Timing." *Cold Spring Harbor Perspectives in Biology* 5 (8): 1–26. https://doi.org/10.1101/cshperspect.a010132.
Rosswog, Carolina, Christoph Bartenhagen, Anne Welte, Yvonne Kahlert, Nadine Hemstedt, Witali Lorenz, Maria Cartolano, et al. 2021. "Chromothripsis Followed by Circular Recombination Drives Oncogene Amplification in Human Cancer," 1–30. https://doi.org/10.1038/s41588-021-00951-7.
Ryba, Tyrone, Dana Battaglia, Benjamin D Pope, Ichiro Hiratani, and David M Gilbert. 2011. "Genome-Scale Analysis of Replication Timing: From Bench to Bioinformatics." *Nature Protocols* 6 (6): 870–95. https://doi.org/10.1038/nprot.2011.328.
Ryba, Tyrone, David K Crockett, D Battaglia, Perry G Ridge, B H Chang, Andrew R Wilson, J W Shirley, et al. 2012. "Abnormal Developmental Control of Replication-Timing Domains in Pediatric Acute Lymphoblastic Leukemia." *Genome Research* 22: 1833–44. https://doi.org/10.1101/gr.138511.112.22.

Ryba, Tyrone, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C. Schulz, Allan J. Robins, Stephen Dalton, and David M. Gilbert. 2010. "Evolutionarily Conserved Replication Timing Profiles Predict Long-Range Chromatin Interactions and Distinguish Closely Related Cell Types." *Genome Research* 20 (6): 761–70. https://doi.org/10.1101/gr.099655.109.

Sasaki, Takayo, Juan Carlos Rivera-Mulia, Daniel Vera, Jared Zimmerman, Sunny Das, Michelle Padget, Naoto Nakamichi, et al. 2017. "Stability of Patient-Specific Features of Altered DNA Replication Timing in Xenografts of Primary Human Acute Lymphoblastic Leukemia." *Experimental Hematology* 51: 71-82.e3. https://doi.org/10.1016/j.exphem.2017.04.004.

Siefert, Joseph C., Constantin Georgescu, Jonathan D. Wren, Amnon Koren, and Christopher L. Sansam. 2017. "DNA Replication Timing during Development Anticipates Transcriptional Programs and Parallels Enhancer Activation." *Genome Research* 27 (8): 1406–16. https://doi.org/10.1101/gr.218602.116.

Smith, Duncan J., and Iestyn Whitehouse. 2012. "Intrinsic Coupling of Lagging-Strand Synthesis to Chromatin Assembly." *Nature* 483 (7390): 434–38. https://doi.org/10.1038/nature10895.

Takahashi, Saori, Hisashi Miura, Takahiro Shibata, Koji Nagao, Katsuzumi Okumura, Masato Ogata, Chikashi Obuse, Shin ichiro Takebayashi, and Ichiro Hiratani. 2019. "Genome-Wide Stability of the DNA Replication Program in Single Mammalian Cells." *Nature Genetics* 51 (March). https://doi.org/10.1038/s41588-019-0347-5.

Ticau, Simina, Larry J. Friedman, Nikola A. Ivica, Jeff Gelles, and Stephen P. Bell. 2015. "Single-Molecule Studies of Origin Licensing Reveal Mechanisms Ensuring Bidirectional Helicase Loading." *Cell* 161 (3): 513–25. https://doi.org/10.1016/j.cell.2015.03.012.

Tsaprouni, Loukia G, Tsun-po Yang, Jordana Bell, Katherine J Dick, Stavroula Kanoni, James Nisbet, Ana Viñuela, et al. 2014. "Cigarette Smoking Reduces DNA Methylation Levels at Multiple Genomic Loci but the Effect Is Partially Reversible upon Cessation." *Epigenetics* 9 (10): 1382–96. https://doi.org/10.4161/15592294.2014.969637.

Tubbs, Anthony, and André Nussenzweig. 2017. "Endogenous DNA Damage as a Source of Genomic Instability in Cancer." *Cell* 168 (4): 644–56. https://doi.org/10.1016/j.cell.2017.01.002.

Tubbs, Anthony, Sriram Sridharan, Niek van Wietmarschen, Yaakov Maman, Elsa Callen, Andre Stanlie, Wei Wu, et al. 2018. "Dual Roles of Poly(DA:DT) Tracts in Replication Initiation and Fork Collapse." *Cell* 174 (5): 1127-1142.e19. https://doi.org/10.1016/j.cell.2018.07.011.

Woodfine, Kathryn, Heike Fiegler, David M. Beare, John E. Collins, Owen T. McCann, Bryan D. Young, Silvana Debernardi, Richard Mott, Ian Dunham, and Nigel P. Carter. 2004. "Replication Timing of the Human Genome." *Human Molecular Genetics* 13 (2): 191–202. https://doi.org/10.1093/hmg/ddh016.

Yates, Lucy R., and Peter J. Campbell. 2012. "Evolution of the Cancer Genome." *Nature Reviews Genetics* 13: 795–806. https://doi.org/10.1016/j.tig.2012.01.003.

Zhao, Peiyao A., Takayo Sasaki, and David M. Gilbert. 2020. "High-Resolution Repli-Seq Defines the Temporal Choreography of Initiation, Elongation and Termination of Replication in Mammalian Cells." *Genome Biology* 21 (1): 1–20. https://doi.org/10.1186/s13059-020-01983-8.

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor and mentor Prof. Dr. Martin Peifer for giving me the opportunity to work with him on this exciting project. After several years working as a bioinformatician, I hoped to work on my own project. Martin gave me the freedom to design my own project from scratch, let me organically grow this research idea that arose from my twin profession of bio and informatics, and guided me through uncertainty and setbacks across overarching research goals. All this is possible because of his openness to different/funny ideas, critical thinking, and most importantly his curiosity and insight to science. My tireless experience also taught me that if I cannot convince him of the worth of my research directions or results, I should probability incline toward his gut feeling.

I am also grateful for the company and support from all the members of the Peifer Lab, in particular Dr. Maria Cartolano for illuminating debate/discussion, and especially during the solitude state of my writing process in an empty office space over a pandemic.

Additionally, I would like to thank Prof. Dr. Roman Thomas for creating such a collaborative research environment within and outside the Department of Translational Genomics. I was very fortunate to benefit from great discussion and advice especially through multiple joint-lab and floor meetings in the intellectual development of this project.

I would also like to thank Prof. Dr. Matthias Fischer for insightful advice and continued wet-bench support for this computational project, and my thesis committee Prof. Dr. Björn Schumacher for critical suggestion on the direction of this project.

Finally, I would like to thank my family in Taiwan, especially my parents for their understanding and support throughout these years.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbstständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit -einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Martin Peifer betreut worden.

Übersicht der Publikationen:

Yang TP, Etich J, Acht T, George J, Thomas RK, Fischer M, Peifer M. "Cell type-specific landscapes of DNA replication initiation and termination in primary cancer whole genomes." *In preparation.*

Cun Y*, **Yang TP***, Achter V, Lang U, Peifer M. 2018. “Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust.” *Nature Protocols* 13 (6): 1488–1501.

Ich versichere, dass ich alle Angaben wahrheitsgemäß nach bestem Wissen und Gewissen gemacht habe und verpflichte mich, jedmögliche, die obigen Angaben betreffenden Veränderungen, dem Promotionsausschuss unverzüglich mitzuteilen.

....

Ort, Datum Unterschrift