PROBING UNDEREXPLORED AXES OF VARIATION IN HUMAN DNA REPLICATION TIMING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Dashiell Massey

May 2022

PROBING UNDEREXPLORED AXES OF VARIATION IN HUMAN DNA REPLICATION TIMING

Dashiell Massey, Ph.D.

Cornell University 2022

Faithful replication of the DNA is critical for cell proliferation and genome stability. In eukaryotes, initiation and progression of DNA replication is highly organized in both space and time. The spatiotemporal DNA replication timing program can be observed as fluctuations in local DNA copy number measured by whole-genome sequencing, and is associated with genomic features including gene expression, nucleotide composition, and chromatin state. However, it remains unclear what regulatory mechanisms underlie the high reproducibility of replication timing assays. Any such mechanism must be able to explain the range of replication timing variation observed; this dissertation characterizes such variation in human replication timing in two distinct contexts.

First, I use two new human reference genome assemblies to characterize replication timing variation in genomic regions with high satellite DNA content. These repeat-rich regions are difficult to assemble and have historically been excluded from replication timing analysis. I focus initially on centromeres, and then expand the analysis to constitutive heterochromatin. I find that these regions are biased toward replication in late S phase and appear to have similar replication timing structure to other better-characterized regions of the genome. Of particular interest, I find that some human cell lines replicate the centromeric regions earlier in S phase than

others and that this trend is consistent across all chromosomes for a given cell line, suggesting that centromeric replication may be coordinate genome-wide.

Second, I examine replication timing variation at the single-cell level, across cell lines and cell types. I describe an *in silico* cell sorting strategy for identifying replicating cells based on single-cell DNA sequencing and analyze replication timing for up to 2,437 cells from a single cell line. I find that sites of replication initiation are shared across cells. The timing of initiation at those sites is predicted by their ensemble replication timing, although all initiation sites fire at an unexpected time during S phase in some fraction of cells. In particular, late-replicating regions contain previously unappreciated heterogeneity in initiation behavior, including initiation sites that are early-replicating but infrequently used.

Together, these studies describe underexplored aspects of variation that ought to be accounted for in models of DNA replication regulation.

BIOGRAPHICAL SKETCH

Dashiell Massey grew up in Cambridge, Massachusetts. He attended the Cambridgeport School from 1999 to 2006 and graduated from the Cambridge Rindge and Latin School in 2010. He received a Bachelor of Arts degree in biology from Swarthmore College, where he also read a lot of Plato, in 2014. It was at Swarthmore that Dashiell fell in love with teaching, through his work as a Writing Associate at the Swarthmore College Writing Center and a Teaching Assistant for the BIO 1 and BIO 2 courses.

After graduating from Swarthmore, Dashiell moved to Washington DC to work at Georgetown University. He was the Laboratory Coordinator for the Department of Human Science in the School of Nursing and Health Studies there from 2014 to 2016, and helped to teach the lab sections for genetics, molecular biology, microbiology, and anatomy and physiology. At Georgetown, Dashiell also conducted research on the impacts of aging on DNA double-strand break repair pathway choice in *Drosophila melanogaster*, under the guidance of Dr. Jan LaRocque.

Dashiell moved to Ithaca, New York, in August 2016, to be a Ph.D. student in the Graduate Field of Genetics, Genomics, and Development (GGD) at Cornell University. He joined the lab of Dr. Amnon Koren in March 2017, where he has worked on various projects relating to biological and technical variation in DNA replication timing and had the opportunity to mentor a talented undergraduate researcher, Sneha Sharma. Dashiell was a Teaching Assistant for the undergraduate Genetics lab course four times, served on the GGD Admissions Committee, and was an active participant in the Department of Molecular Biology and Genetics' Diversity Council.

For my grandma Adina,

who was my staunchest supporter

and my most challenging interlocutor for 29 years

ACKNOWLEDGMENTS

The work in this dissertation would not have been possible without the support and guidance of my doctoral advisor, Dr. Amnon Koren. Thank you, Amnon, for pushing me to grow every day of the past six years, and for trusting that I had useful points to make, even when it took me a few drafts to articulate them. Thank you, also, to Dr. Robert Weiss and Dr. Charles Danko for serving on my thesis committee and for your suggestions, feedback, and encouragement after many seminars and during many committee meetings.

I am tremendously grateful to have had the opportunity to work with amazing colleagues at Cornell. Thank you to my labmates – Alexa Bracci, Madison Caballero, Andy Ding, Matt Edwards, Ya Hu, Michelle Hulke, Tiffany Ge, Sean Kim, Rita Rebelo, and Sneha Sharma – for listening to my rambling half-formulated ideas and for always being up for coffee. I am especially thankful to Kayla Brooks for her friendship at the very beginning. Grad school would not have been nearly as fun without my cohort: Jawaher Al Zahrani, Sylvia Chang, Dawn Chen, Mike DeBerardine, Jullien Flynn, Hui Ji, Yeonui Kwak, Felicia New, Keaton Phillips, Jens Sannerud, Albert Vill, Miwa Wenzel, and Marquita Winters. They were important sources of support, as were Sofie Delbare, Carolyn Castillo, and Rachel Sandman.

The MBG Diversity Council has been an integral part of my sense of purpose and belonging at Cornell. Thank you especially to Irma Fernandez, Jawuanna McAllister, Mariela Núñez Santos, Oriana Teran Pumar, and Kara Zielinski for being my frequent collaborators in these endeavors.

I would not have started down the path to getting a Ph.D. if it were not for Dr. Parul Matani, nor stuck with it without the guidance of Dr. Liz Vallen, Stacey Miller, and Jocelyne Mattei Noveral. I took my first foray into research

with the support of Denise Simmons and Dr. Carolyn Turk. Dr. Pablo Irusta, Dr. Jan LaRocque, Dr. Ted Nelson, Dr. Alex Theos, and Dr. Ronit Yarden have remained important mentors and friends since leaving Georgetown. Thank you to Dr. Shana Minkin and Dr. Grace Ledbetter for encouraging me outside of science. I am grateful to Dr. Eric Alani and Dr. Mariana Wolfner for being there to listen, at my seminars and when I needed advice. Dr. Kristina Blake-Hodek, Dr. Mike Goldberg, and Dr. Kimberly Williams gave me opportunities to teach, think about teaching, and talk about teaching. Nothing logistical would have been possible without the incomparable Diane Fritz at Swarthmore, and Casey Moore, Vic Shaff, and Ginger Tomassini at Cornell. I owe the greatest thanks to teacher-extraordinaire Rosalind O'Sullivan.

Life, like science, is a collaborative enterprise. I am fortunate to have walked through life for 25+ years alongside friends like Olivia MacLennan, Sula Watermulder, Lucy Flamm, and Hannah Firestone. Suryani Dewa Ayu taught me to care about aesthetics. Sierra O'Mara Schwartz and Courtney Murray cheer me on every day. Hannah Grunwald should be my lab partner in every context. Caroline Batten, Natalie Campen, Anna Cha, Peter Daniels, Brenna Hilferty, Kiera James, Shivani Mantha, Elèna Ruyter, Christine Song, and Paloma Villareyes Perez have been there for the many highs and lows.

I am so grateful for the love and encouragement of my family: thank you to Dita Obler, Kevin Massey, Adina Obler, Martin Obler, Robyn Obler, Doris Massey Ruud, Cindy Wishengrad, Gil Obler, Carla Morgenstern, Wilson Merrell, the Walkers (Gretchen, Luke, Sonja, Ivan), Bobbie and Gudmund Iversen, and Harriette Crawford. Thank you to Andrew St. James for being my friend above everything else. Most of all, thank you to my sister, Kassia, for knowing when I have had a bad day before I have even said anything and for having my back always. You really are my favorite person in the world.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

2N	unreplicated state; copy-number 2
4N	replicated state; copy number 4
ARS	autonomously replicating sequence
bp	base pairs
BrdU	bromodeoxyuridine
BWA	Burrows-Wheeler short-read alignment algorithm
CDC6	cell division cycle protein 6
CDC7	cell division cycle protein 7
CDC45	cell division cycle protein 45
CDT1	chromatin licensing and DNA replication factor 1
CENP-A	centromere protein A; histone H3 variant
ChIP-seq	chromatin immunoprecipitation sequencing
CMG	CDC45-MCM-GINS complex
CN	copy number
CNA	copy-number aberration
CNV	copy-number variant
CTR	constant replication timing region
DBF4	dumbbell former 4 protein
DBP11	DNA polymerase binding protein 11
DDK	DBF4-dependent kinase; also known as DBF4-CDC7
DLP+	direct DNA transposition single-cell library preparation
DNA	deoxyribonucleic acid
DOP-PCR	degenerate oligonucleotide-primed polymerase chain reaction
EdUseq-HU	sequencing of 5-ethynyl-2'-deoxyuridine-labeled nascent DNA after hydroxyurea-induced cell cycle arrest

ESC	embryonic stem cell line
FACS	fluorescence-activated cell sorting
FBS	fetal bovine serum
FISH	fluorescence *in situ* hybridization
G_1 phase	growth 1 phase of the cell cycle
GenomeSTRiP	Genome Structure in Populations pipeline
GINS	*Go-Ichi-Ni-San* complex; comprised of Sld5, Psf1-3
GRCh38/hg38	genome reference consortium human build 38
HMM	hidden Markov model
HOR	α-satellite higher-order repeat DNA
HP1	heterochromatin protein 1
HSat1	human classical DNA satellite 1
HSat2	human classical DNA satellite 2
HSat3	human classical DNA satellite 3
IR	initiation region
kb	kilobase; 1,000 base pairs
LCL	lymphoblastoid cell line
LIANTI	linear amplification via transposon insertion
MALBAC	multiple annealing and looping-based amplification cycles
MAPD	median absolute deviation of pairwise differences between adjacent genomic windows
MAPQ	mapping quality score
Mb	megabase; 1,000,000 base pairs
MCM10	mini-chromosome maintenance protein 10
MDA	multiple displacement amplification
OK-seq	Okazaki fragment sequencing
ORC	origin recognition complex; comprised of ORC1-6

PCR	polymerase chain reaction
pre-RC	pre-replication complex
R1	first sequencing read in a mate-pair
RECQL4	RecQ-like helicase 4
RIF1	replication timing regulatory factor 1
S phase	DNA synthesis phase of the cell cycle
S-CDK	S-phase-specific cyclin-dependent kinase
SLD2	synthetic lethal with DBP11 protein 2
SLD3	synthetic lethal with DBP11 protein 3
SMARD	single molecule analysis of replicated DNA
SRA	National Center for Biotechnology Information Sequence Read Archive
T2T-CHM13	telomere-to-telomere genome assembly of CHM13-hTERT
TIGER	Timing of Genome Replication pipeline
TOPBP1	DNA topoisomerase 2-binding protein 1
WGA	whole-genome amplification

CHAPTER 1: INTRODUCTION

Author Contributions: D.J.M. and M.L.H. contributed equally to the manuscript. D.J.M. wrote the initial draft of all text included in this chapter.

Eukaryotic DNA replication initiation

Proper replication of the entire DNA content of the genome once and only once per cell cycle is an essential process for ensuring genome stability and continued cellular viability. Therefore, initiation of DNA replication by the helicase activity of the mini-chromosome maintenance complex (MCM; MCM2-7) is tightly regulated. Key to this regulation is temporal separation of MCM loading and MCM activation –the conditions necessary for "licensing" replication origins do not coincide during the cell cycle with those necessary for "firing" them (for a thorough review, see [2,3]).

During late mitosis and early G_1 phase of the cell cycle, replication origins are licensed by *de novo* MCM loading, through the assembly of the pre-replication complex (pre-RC) on the chromatin. Pre-RC formation is nucleated by the heterohexameric origin recognition complex (ORC; ORC1-6)[4,5]. This initial ORC-DNA interaction is subsequently stabilized by binding of a related AAA^+ ATPase, CDC6[6-8]. Next, the licensing factor CDT1 associates with the replication origin[9,10]. MCM is loaded last, through a mechanism that requires both CDC6[11-13] and CDT1[9,10]. Two pre-RCs are coordinately assembled per replication origin, resulting in a fully licensed origin bound by two inactive MCM complexes in a head-to-head arrangement[14-16].

In the subsequent S phase, assembly of the CDC45-MCM-GINS (CMG) complex activates the helicase activity of MCM[17-19]. CMG assembly is limited to S phase by reliance on the Dbf4-dependent (DDK) and cyclin B/cyclin-dependent (S-CDK) kinases, which increase in activity at the G_1-to-S transition[20-22]. Namely, the CDC45-MCM interaction requires DDK-mediated phosphorylation of MCM and Sld3[23,24]. Likewise, phosphorylation of Dbp11 and Sld2 by S-CDK is necessary to recruit GINS and DNA polymerase ε[25]. (In metazoa, Treslin, TopBP1, and RECQL4 fulfill the roles of Sld3, Dbp11, and Sld2, respectively[26-29].) Simultaneously, S-CDK prevents new MCM loading during S phase via inhibition of the pre-RC components ORC2, ORC6, and CDC6, and exclusion of CTD1 and MCM2-7 from the nucleus, by a variety of mechanisms (reviewed in [30]). Thus, during any given cell cycle, only those origins licensed during the preceding G_1 phase can be fired in S phase.

In the budding yeast *Saccharomyces cerevisiae*, replication origins were initially isolated as sequences of several hundred base pairs capable of autonomous replication initiation of extrachromosomal DNA (autonomously replicating sequences; ARS)[31-33]. Subsequent analyses revealed that ARS are composed of modular sequence elements[34], with sequence specificity of ORC binding conferred by the "A" element, an 11-bp AT-rich consensus sequence[35-37]. However, efforts to identify analogous mechanisms in other eukaryotes have proven unsuccessful: in humans, for instance, autonomous replication assays have demonstrated that any fragment of human DNA longer than 15kb is capable of promoting replication, including a tandem array of a sequence that cannot efficiently replicate autonomously as a monomer[38,39]. This suggests that replication initiation in humans may require a large stretch of DNA sequence content favorable to ORC binding, with low specificity for individual loci. Consistent with this hypothesis, chromatin immuno-

precipitation sequencing (ChIP-seq) in the human lymphoblastoid Raji cell line demonstrates dispersed binding of ORC and MCM across the genome[40]. However, complete genome replication within a consistent duration of time would seem to require a particular number and spacing of replication origins; it is unclear how stochastic origin designation in each cell cycle would accomplish this. Furthermore, the ability of a fragment of DNA to act as an origin *in vitro* does not necessarily mean that it is used as an origin *in vivo*. Along these lines, human cells appear to license a large excess of "dormant" origins, which only fire under conditions of replication stress[41,42]. Without a thorough mechanistic understanding of how human replication origins are designated, it remains challenging to elucidate their regulation.

DNA replication timing

One clue to this puzzle is the DNA replication timing profile. Measurements of local DNA copy number in cultured cells produce a robust and reproducible wave pattern across the length of chromosomes[43-45]. This was first observed in *S. cerevisiae*, where sharp local maxima in copy number correspond to the locations of known replication origins, and the relative amplitudes corresponded to previous observations of origin firing early *vs*. late in S phase[43]. Replication timing profiles for human cell lines also reveal distinct peaks in local copy number at consistent locations and with consistent amplitudes across replicates, suggesting that human replication origins are primarily in the same locations across cells, which fire at specific times in S phase[44-46]. Profiling of ensemble DNA replication timing thus represents a powerful tool for inquiry into the mechanisms by which human replication origins are licensed and fired.

Several other properties of replication timing are notable and may offer insight into its regulation. First, replication timing is cell-type-specific: the profiles for cell lines of the same cell type are highly correlated ($r > 0.95$)[46], whereas profiles between cell types differ on average at ~50% of loci across the genome[44,47]. *In vitro* differentiation experiments with embryonic stem cell lines substantiate the notion that cell lineage trajectories involve a continuum of replication timing changes[48-50], although these results are based on a relatively small number of cell types. Furthermore, cell-type replication profiles are shared across species, *e.g.*, within syntenic regions between human and mouse[47,51].

Second, replication timing is associated with numerous genomic features relevant to genome stability and disease. Early-replicating regions tend to be gene-rich and transcriptionally active, while late-replicating regions tend to be gene-poor and more heterochromatic[52-55]. Some cell-type differences in replication timing have been associated with coordinate changes in gene expression, such that the cell type with higher gene expression is also earlier-replicating in the vicinity of that gene[48,56]. Additionally, late-replicating regions are enriched for a higher density of repetitive sequence elements and mutations[45,55,57,58]. The associations of replication timing with gene density and expression and with mutation density suggest that replication timing may bear a causal relationship to these genomic features – although the direction of causality is unknown. On the other hand, it is also possible that each of these associations is driven by a shared mechanism that shapes all of these genomic features as a consequence.

Interestingly, global replication timing defects are rare. While tumor-derived and other disease-associated cell lines display local alterations in replication timing, mutations in only two genes (*RIF1* and *MCM10*) have been

associated with large-scale or global disruption[59,60]. This suggests that the range of observable replication timing profiles is highly constrained. One interpretation of this observation is that replication timing is tightly regulated and/or under strong purifying selection, such that almost any deviation from the expected program is highly deleterious or incompatible with viability. However, an alternative interpretation is that replication timing is not regulated at all, but rather is an emergent property of genome architecture: perhaps the chromatin landscape (to give one example) restricts the accessibility of possible replication initiation sites, resulting in the consistent use of the same set of replication origins without any direct regulation of origins themselves.

These unresolved questions about whether replication timing is actively regulated (and if so, by what mechanism) make it difficult to understand its relationship to the genomic landscape of mutations, its role in cancer and other diseases, and its evolutionary history. By examining variation in replication timing more closely, this dissertation provides new insight into the types of mechanisms that could explain replication timing.

Methods for studying DNA replication timing

Two primary methods are used to assay replication timing by whole-genome DNA sequencing of a population of proliferating cells. Both use sequencing read depth as a proxy for local DNA copy number. By virtue of being replicated in a larger proportion of cells, early-replicating regions are expected to have higher relative copy number and thus be overrepresented among sequencing reads. Conversely, late-replicating regions have been

replicated in a smaller proportion of cells, have a lower relative copy number, and are underrepresented among reads.

The first method specifically sequences nascent DNA labeled with a nucleotide analogue (*e.g.*, BrdU). After collecting cells along a time course, it is possible to infer the replication timing profile by comparing read depth in late-S-phase cells to early-S-phase cells[44]. This approach has been performed with varying numbers of time points, ranging from two[61] to sixteen[62], and typically fluorescence-activated cell sorting (FACS) has been used to isolate a series of "time points" from an unsynchronized population.

The second method instead directly uses sequencing read depth from unlabeled DNA[45]. This method is typically performed using a single S-phase fraction isolated by FACS and normalized by a G_1-phase control fraction to account for the confounding effects of GC-content, mappability, and copy-number variants. More recently, methods for computational simulation of a control sample have proved useful for inferring replication timing from a single asynchronous population without any experimental manipulations[46,63-65].

Difficult to sequence regions

The methods outlined above rely on whole-genome sequencing to assay replication timing. They also account for the fact that not all regions of the genome are equally amenable to analysis by sequencing due to their nucleotide content. GC-content influences the thermostability of DNA, meaning that the efficiency of molecular biology technologies that rely on DNA synthesis – namely, PCR and Illumina sequencing – will not be uniform across the genome at a constant melting temperature[66]. Furthermore, short-read alignment algorithms depend on sufficiently unambiguous matches

between sequencing reads and the reference genome, such that reads arising from repetitive sequence elements will be difficult (if not impossible) to map. These biases influence sequencing read depth and confound replication timing inference. Thus, both methods rely on a control sample with similar bias to specifically isolate the signal of DNA replication.

However, neither method can resolve regions of the human genome where repetitive satellite DNA sequence content is essentially high, which are often represented as gaps in the reference genome. This includes the centromeric regions of all chromosomes, as well as large spans of constitutive heterochromatin, most notably in the pericentromeric regions of chromosomes 1, 9, and 16, and the entire p-arms of chromosomes 13, 14, 15, 21, and 22. These regions have traditionally been excluded from genomic analyses. Thus, there remains ~8% of the human genome for which replication timing has not been profiled.

In the past decade, long-read DNA sequencing has been used to assemble satellite DNA arrays in order to close gaps in the human reference genome[67-70]. This has opened the possibility of studying the replication timing of these regions with the same methods we use to study other genomic regions. In **Chapter 2**, I present preliminary replication timing profiles for 18 human centromeres using human reference genome build 38 (GRCh38/hg38)[71]. Centromeres are represented in hg38 by computationally designed sequence models, which leveraged variation within individual repeat elements to construct a chromosome-specific alignment decoy[68]. Thus, analysis is limited to centromeres as a category of genomic region. In **Chapter 3**, I present telomere-to-telomere replication timing profiles inferred using the T2T-CHM13 genome assembly[70]. This assembly includes linear sequences for the entire human genome, including the centromeres of all 22 autosomes and the

X chromosome and the p-arms of the acrocentric chromosomes. Using these profiles, I replicate the results from **Chapter 2** and analyze the replication timing of satellite DNA and heterochromatin more broadly.

Together, these two chapters reduce an important knowledge gap about replication timing of the human genome. I expand the proportion of the genome that has been assayed for replication timing, particularly for gene-poor regions that serve important structural roles in replication. Differences in centromeric replication timing between cell lines raise new questions about if and how centromeric replication is coordinated across chromosomes and with other cell cycle checkpoints.

The value of single-cell replication timing•

While genome-wide replication timing assays have proven highly reproducible and informative, a clear limitation of these assays is that they rely on ensemble population analyses and do not provide information about replication progression in individual cells. Similarly, replication origin mapping techniques rely on many cells, largely masking heterogeneity that might be present among cells.

It has long been argued that DNA replication is stochastic, in the sense that different cells (or the descendants of the same cell in different cell cycles) activate different subsets of replication origins. From the perspective of studying replication origins, the manifestation of this stochasticity is that a given origin may or may not fire in a given cell cycle. Thus, origins are characterized by a genomic location and a preferred time of activation, but also by the probability of being activated. Importantly, many DNA replication assays measure a combination of these factors, and it may be difficult to

disentangle what is actually being measured. For instance, an origin that fires frequently in late S phase may appear similar in an ensemble analysis to an origin that fires early but only in a subset of cells. In addition, if origins are clustered along chromosomes, the firing of different origins within the same region in different cells will give the impression of an extended initiation zone, masking the activity of the individual origins.

Existing methods for single-cell analysis•

The inefficiency of replication origins is supported by DNA combing experiments, which assay nascent DNA replication in single molecules[72-74] and are extensions of earlier DNA fiber autoradiography analyses[75]. In these experiments, replication is allowed to proceed in the presence of a succession of labeled nucleotide analogues, followed by stretching of individual DNA fibers on glass slides. The location and orientation of replication forks is inferred from immunofluorescent detection of the incorporated nucleotides. Combing-based analyses of DNA replication have been performed in frog cell extracts[76-78], yeasts[79,80], flies[81], mice[81-83] and human cells[84-87]. These studies have shown that firing of an origin in one cell is not correlated to its firing in other cells[79] or in a subsequent cell cycle[78,80]. However, at scales larger than an individual origin, single-molecule replication *is* conserved: averaging data across all single molecules recapitulates the population-level replication profile[79], and broad regions appear to have consistent replication timing across cell cycles[78,81]. Thus, stochastic firing of individual replication origins, parameterized by differential firing potential[88,89], still predicts overall replication timing profiles that are consistent with observed ensemble measurements.

A main limitation of DNA combing, however, is throughput: a typical experiment involves two nucleotide analogue pulses, each of which requires a distinct antibody detection step. In addition, combing alone cannot determine where the observed replication initiation events are located within the genome. Approaches that combine DNA fiber analysis with fluorescence *in situ* hybridization (FISH) of sequence-specific probes[90,91], most notably a technique called SMARD (single molecule analysis of replicating DNA)[74,82], overcome this last restriction yet introduce further experimental challenges.

Recently, two strategies have been put forward for single-molecule replication timing analysis on a genomic scale. The first, optical mapping with microfluidic nanochannels, attempts to alleviate the bottlenecks that reduce the throughput of traditional DNA combing experiments. The use of fluorescently tagged nucleotides instead of nucleotide analogues dramatically speeds up detection of nucleotide incorporation (*i.e.*, nascent DNA replication), obviating the need for antibody detection steps prior to imaging[92,93]. In addition, fluorescent tag locations can be mapped to the genome by treating stretched DNA molecules with a nicking endonuclease and identifying restriction patterns[92,94]. Together, these innovations make it practical to automate DNA combing for higher throughput: DNA fibers can be imaged as they are flowed (and stretched) through a microfluidic nanochannel, and retrospectively mapped back to the reference genome by their endonuclease fingerprints[93,95]. This strategy has been employed to develop a high-coverage single-molecule map of origins in *Xenopus* egg extracts[95] and to study early-firing origins in synchronized, aphidicolin-treated HeLa S3 cells[96]. The vast majority of origins detected in single-molecules by optical mapping overlap with origins called by OK-seq and are enriched for ORC1 binding. Additionally, origins can be identified that are used by as few

as 1% of the cells in a population[96]. However, this approach requires specialized equipment and, more fundamentally, resolution is inherently limited by the distribution of endonuclease cleavage sites in the genome.

The second strategy for single molecule origin detection uses DNA sequencing, which enables direct genomic mapping. The Oxford Nanopore Technologies sequencer produces reads that can reach upwards of 100kb in length from single DNA molecules without the need for DNA amplification. Within the sequencer, a single strand of DNA is translocated through a protein nanopore by electrophoresis. As the DNA traverses the nanopore, characteristic changes in ionic current can be detected and interpreted as sequence readout[97]. It has been shown that nanopore sequencing can distinguish BrdU from thymidine and thus, replicated from unreplicated DNA. This strategy has been used to map BrdU tracks to early-replicating origins in yeast[98,99]. Intriguingly, ~20% of origins observed at the single-molecule level were *not* detected in population-scale sequencing[99]. The long-read approach generates large amounts of contextual information, making it easier to map reads to the genome and providing information on how neighboring origins may impact one another. However, it remains technically challenging to distinguish nucleotide analogs from canonical nucleotides. For instance, nanopore base calling relies on the shifts in ionic current characteristic of *k*-mers, rather than single nucleotides – and BrdU can only be distinguished from thymidine in certain 6-mer contexts[99]. Another concern is that this strategy compares BrdU-labelled DNA to non-labelled DNA; biases in base-calling and mappability between BrdU and the native thymidine must be accounted for[99]. The recent demonstration that 11 different thymidine analogs can be detected using Oxford Nanopore Technologies sequencing

provides a promising avenue for using analog combinations to identify the location and direction of DNA synthesis events[100].

Finally, recent advances have extended DNA copy number-based replication timing assays (that use short-read sequencing depth) to single cells. Measuring genome-wide replication timing in single cells is appealing: direct measurements can be made without cell synchronization or other perturbations, and the set of possible copy number values is theoretically limited to integers. Single-cell replication timing requires isolation of individual cells, which has so far been done using flow cytometry. A greater challenge is that, unlike with nanopore sequencing, short-read DNA sequencing requires whole-genome amplification (WGA), and there is rightful concern that even small biases in amplification will be exponentially magnified given the paucity of starting material. To date, several WGA strategies have been developed (reviewed in [101,102]). Some, like degenerate-oligonucleotide-primed PCR (DOP-PCR)[103,104]), amplify the genome exponentially. Other methods like Multiple Annealing and Looping-Based Amplification Cycles (MALBAC)[105] combine reduced amplification bias from the low-temperature isothermal φ29 polymerase with the higher amplification efficiency of traditional PCR once the amount of input template has been increased sufficiently. More recently, an approach called Linear Amplification via Transposon Insertion (LIANTI) was developed, which uses *in vitro* RNA transcription to linearly amplify the genome[106].

The possibility that single cells contain sufficient copy number variation between early- and late-replicating regions to be detected on a genomic level was first proposed by analyzing microarray data of single lymphoblastoid cells in S phase[107]. Much higher resolution, however, has been obtained using LIANTI and DNA sequencing of 11 single human BJ fibroblast

cells in early S phase[106]. Although replication profiles were correlated between cells and the average replication timing correlated well with a Repli-seq ensemble profile, there was also strong evidence of discrepant local copy number between cells, suggesting stochasticity of replication timing. In contrast, more recent studies that used DOP-PCR to amplify DNA from hundreds of mouse embryonic stem cells found limited heterogeneity between cells[50,108,109]. This heterogeneity may be non-uniformly distributed in the cell cycle: Dileep and Gilbert[108] reported that variability in replication timing was comparable between early- and late-firing regions, while Takahashi *et al.*[109] claim that heterogeneity peaks in mid S phase with lower levels observed in early and late S phase. These studies also showed that single-cell replication timing profiling can distinguish between different cell types and even between the two copies of each chromosome. It cannot, however, identify individual replication initiation events. Future improvements in data resolution, which will likely require minimally biased amplification regimes such as LIANTI, may enable high-resolution single-cell replication timing analysis that could explicitly examine the activity of replication origins.

At present, the primary drawback of short-read sequencing for single-cell replication analysis is scalability: applying standard library preparation methods to single cells is laborious. This limits the size of available datasets, constraining the conclusions that can be drawn from them. However, these limitations can be overcome with improvements in barcoding strategies (*e.g.*, as in [110-112]) to allow larger numbers of cells to be pooled during time-consuming steps of library preparation, and increased accessibility of microfluidic devices to enable further automation of this process[112-114].

In **Chapter 4**, I demonstrate the value of microfluidic platforms for scaling up replication timing analysis in single cells. Using the 10x Genomics

Single-Cell CNV platform, I present analysis of replication timing variation across > 5,000 cells, ten cell lines, and three cell types. These results suggest that the locations of replication initiation are primarily shared across cells, and that the order of origin firing is highly structured but not entirely predictable.

Further applications of single-cell replication timing

In the concluding chapter, I include a demonstration of the value of the single-cell methods developed in **Chapter 4** to two new lines of inquiry. First, in addition to replication initiation variation among single cells, the methods I have developed can be applied to sequencing of micronuclei in individual cells, complementing ensemble analyses to ask not only whether genomic regions differ in their propensity to form micronuclei but also whether micronuclei tend to harbor fragments of one or multiple chromosomes. Second, my *in silico* sorting method can be used to generate ensemble replication timing profiles when only single-cell data are available, as is often the case for primary tissue. This will allow us to study the replication timing of rare samples, *e.g.*, healthy and diseased tissue biopsies, as well as cell lines like fibroblasts that proliferate too slowly to profile by traditional methods. Single-cell tumor datasets, in particular, are often collected to study clonal expansion and tumor evolution[115-117], meaning that more and more data will become available for replication timing analysis going forward.

CHAPTER 2: NEXT-GENERATION SEQUENCING ENABLES SPATIOTEMPORAL RESOLUTION OF HUMAN CENTROMERE REPLICATION TIMING

This chapter is published as: Massey DJ, Kim D, Brooks KE, Smolka MB & Koren, A. Next-generation sequencing enables spatiotemporal resolution of human centromere replication timing. *Genes* 10, 269 (2019).[118]

Author Contributions: D.J.M. and D.K. contributed equally to the manuscript. D.K. and K.E.B performed experiments; D.J.M. analyzed data; M.B.S. and A.K. supervised the study; D.J.M. and A.K. wrote the manuscript with input from D.K. and M.B.S. D.J.M. wrote the initial draft of all text in this chapter.

Abstract

Centromeres serve a critical function in preserving genome integrity across sequential cell divisions, by mediating symmetric chromosome segregation. The repetitive, heterochromatic nature of centromeres is thought to be inhibitory to DNA replication, but has also led to their underrepresentation in human reference genome assemblies. Consequently, centromeres have been excluded from genomic replication timing analyses, leaving their time of replication unresolved. However, the most recent human reference genome, hg38, included models of centromere sequences. To establish the experimental requirements for achieving replication timing profiles for centromeres, we sequenced G_1- and S-phase cells from five human cell lines, and aligned the sequence reads to hg38. We were able to infer DNA replication timing profiles for the centromeres in each of the five cell lines, which showed that centromere replication occurs in mid-to-late S phase. Furthermore, we found that replication timing was more variable between cell lines in the centromere regions than expected, given the distribution of variation in replication timing genome-wide. These results suggest the

potential of these, and future, sequence models to enable high-resolution studies of replication in centromeres and other heterochromatic regions.

Introduction

DNA replication during the S phase of the cell cycle initiates at replication origin loci, which are both spatially dispersed across the genome and asynchronously activated. The resultant spatiotemporal pattern of DNA replication timing is highly reproducible and largely conserved, producing consistent early- and late-replicating regions (reviewed in [3]). In general, early-replicating regions show greater transcriptional activity[48,52,53,119] and higher gene density[48,52-54,119], while late-replicating regions tend to accumulate more mutations[45,57,58,120]. Constitutive heterochromatin is widely accepted as a prime example of the relationship between closed chromatin state and late replication timing[121-124]. Late replication of heterochromatic regions has been linked to telomeric proximity[125,126] and transcriptional silencing[126,127] in the budding yeast *Saccharomyces cerevisiae*, and to histone hypoacetylation[128] and distance from the nuclear periphery[129] in mouse.

Centromeres are an intriguing potential outlier to the late-replicating heterochromatin paradigm: centromeres replicate early in multiple yeast species, including *S. cerevisiae*[43], the fission yeast *Schizosaccharomyces pombe*[130,131], and the pathogenic yeast *Candida albicans*[132]. This presents an opportunity for insight into the mechanisms that promote replication origin activity as well as the mechanisms that dictate late replication in other heterochromatic regions. Indeed, in *S. pombe*, early centromeric replication has been explained by interactions of heterochromatin protein 1 (HP1) with the replication initiation factors CDC6[133] and DDK[134]. Ablation of either of these interactions results in

the centromere replicating with other heterochromatin in late S phase[133,134], thus giving further support to the model that a closed chromatin state is generally repressive to origin firing.

The time at which centromeres replicate is less clear in higher eukaryotes: centromere replication in early S phase has been reported in the *Drosophila Kc* cell line[135], in mid S phase in multiple human cell lines[136], in late S phase in *Drosophila* larvae[137], and even throughout the full S phase in mouse cell lines[138]. However, the general consensus is that human centromeres replicate late in S phase[139-141], consistent with the timing of heterochromatin replication. Studies across species have also suggested that centromeres on different chromosomes may replicate at different times[137-139,141] and that the neighboring pericentromeric heterochromatin replicates earlier than the centromeres themselves[138,141].

Next-generation sequencing provides a high-resolution assay for replication timing at genome-scale[44,45]. However, in most eukaryotes, centromeres are satellite-rich constitutive heterochromatic regions ranging in size from hundreds of kilobases to several megabases. The high repetitive-sequence content of centromeres renders them difficult to sequence and assemble. As a result, centromeres have historically been gaps in reference genomes[68] and thus excluded from newer, sequencing-based analyses. However, the most recent human reference genome, hg38, includes sequence models for all 24 centromeres[71]. These constructed sequences take advantage of subtle variation within related centromeric satellites to build localized assemblies that are then arranged by a second-order Markov chain modeled on the frequency of these variants[68]. Although the sequence models do not necessarily reflect the accurate linear DNA sequence within the centromere,

they do allow sequencing reads originating from the centromeres to be aligned.

Here, we report that the centromere sequence models in hg38 enable measuring replication timing of human centromeres. We reveal their timing and variation in five cell lines, and detail the experimental conditions required to obtain this type of information. Our results demonstrate that high-throughput sequencing of human cell lines is both a feasible and a fruitful methodology to clarify a more detailed understanding of the human centromere and its time of replication during the S phase of the cell cycle.

Results

Genome-wide replication timing profiles for five human cell lines

To assess the feasibility of studying centromere replication by whole-genome sequencing, we generated replication timing profiles for five human cell lines: an apparently healthy lymphoblastoid cell line (GM12878; [142]), an embryonic kidney cell line (HEK293T), an ovarian carcinoma cell line (A2780), and two breast cancer cell lines (HCC1143 and HCC1954; [143]). For each cell line, an asynchronous population was flow-sorted to isolate 1 million cells from the G_1 (pre-replicative) and S (replicative) phases of the cell cycle. Pairs of G_1- and S-phase fractions were sequenced and aligned to hg38.

For each cell line, replication timing was inferred for the S-phase fraction in variable-size windows determined by the G_1-phase fraction (**Figure 2.1a**), as previously described[45]. Briefly, early-replicating regions are expected to be overrepresented (*i.e.*, have high sequencing read depth) in the S-phase fraction, while late-replicating regions will be underrepresented. The G_1-phase

fraction, for which all genomic regions are expected to be present in uniform copy number, was used as a baseline to account for mappability and sequencing biases, as well as copy-number variants (**Figure 2.2**).

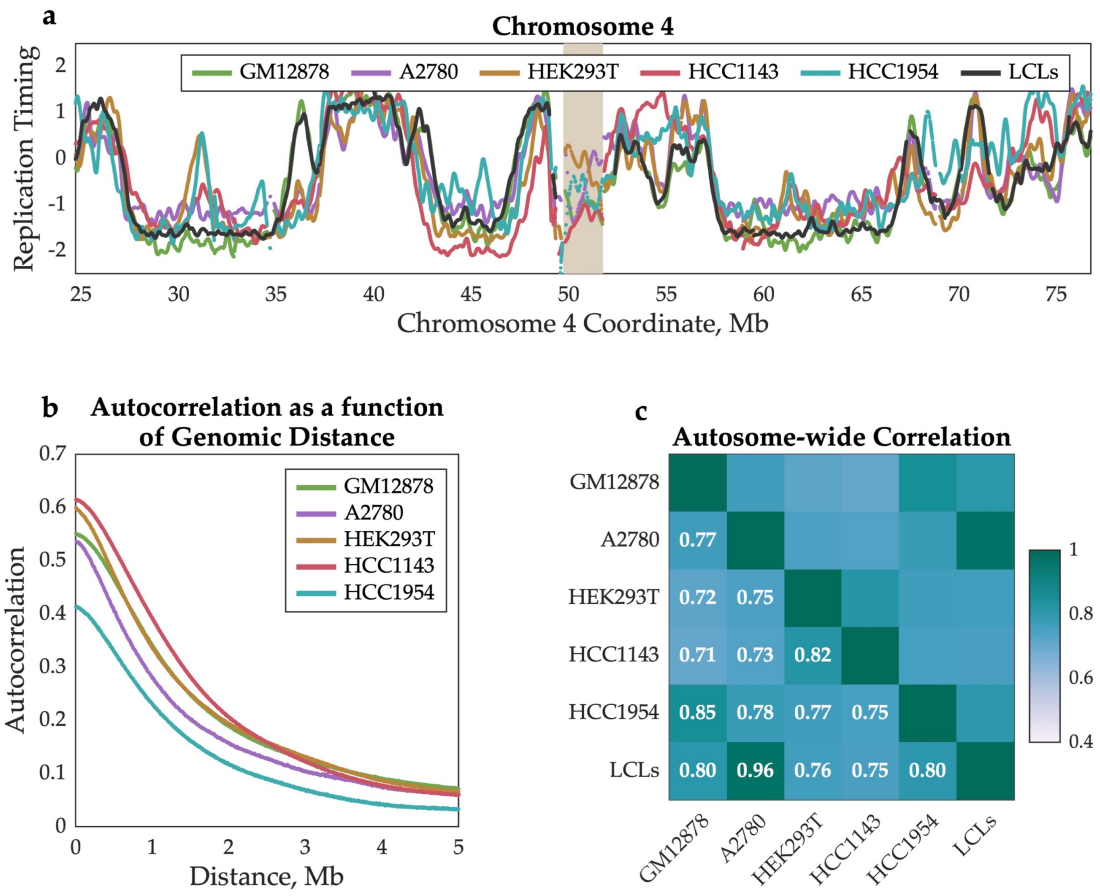


Figure 2.1. **Replication timing profiles of five human cell lines. a** Replication timing for a region of chromosome 4, centered on the centromere (*tan*). All samples were compared to an average replication timing profile of six lymphoblastoid cell lines (LCLs, from [45]; *black*). **b** The replication timing profiles for all five cell lines displayed strong spatiotemporal structure, as measured by autocorrelation. **c** Pearson correlations among cell lines. These are within the expected range for different cell types, given previous studies[44,47].

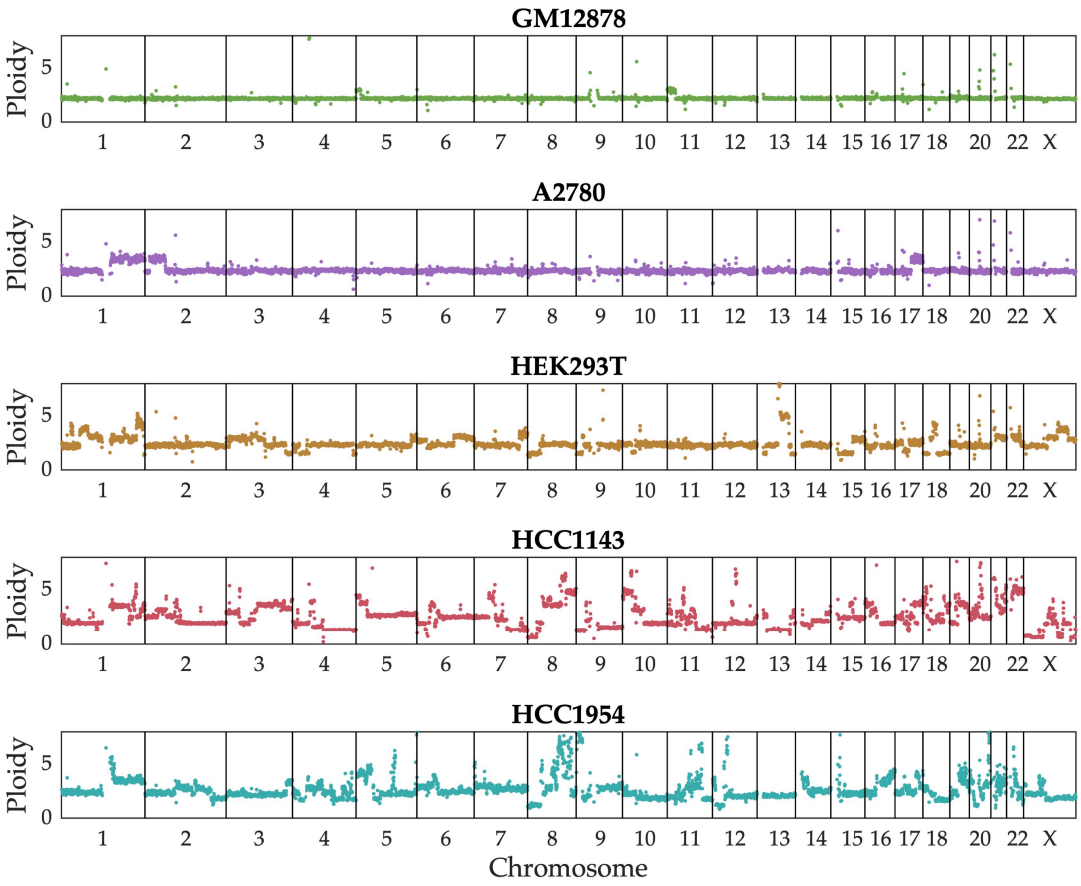


Figure 2.2. **Copy number in 1Mb windows for the G1-phase fractions, following mappability- and GC-bias correction using GenomeSTRiP[63].** GM12878 is diploid and A2780 is near-diploid, while HCC1143, HCC1954, and HEK293T display many chromosome gains and losses. Normalization of the S-phase fractions against these G1-phase backgrounds was necessary to account for these changes in sequencing read depth that do not reflect DNA replication.

To assess the quality of these replication timing profiles, we considered the autocorrelation of replication timing as a function of genomic distance. Consistent with the spatiotemporal dynamics of replication, each profile demonstrated high autocorrelation along the chromosome on the scale of several megabases (**Figure 2.1b**). Autosome-wide replication timing profiles were strongly correlated across samples (r = 0.70–0.85) and with our previously-published measurements in lymphoblastoid cell lines[45] (r = 0.75–

0.96; **Figure 2.1c**), consistent with previous reports that at least 50% of the replication timing is conserved between cell types[44,47].

Replication timing can be profiled in centromeric regions by paired-end sequencing

We next focused our attention on the centromeres, requiring that any centromere contain on average at least 10 G_1-defined windows (1100 reads) per megabase to be included in our analysis. For A2780, 13 centromeres were successfully profiled, while 17–18 centromeres passed this threshold in the other four cell lines (**Figure 2.3a**). The identity of the centromeres that were successfully profiled was consistent across samples, suggesting that this is a property of the sequence models of individual centromeres (**Figure 2.3b**).

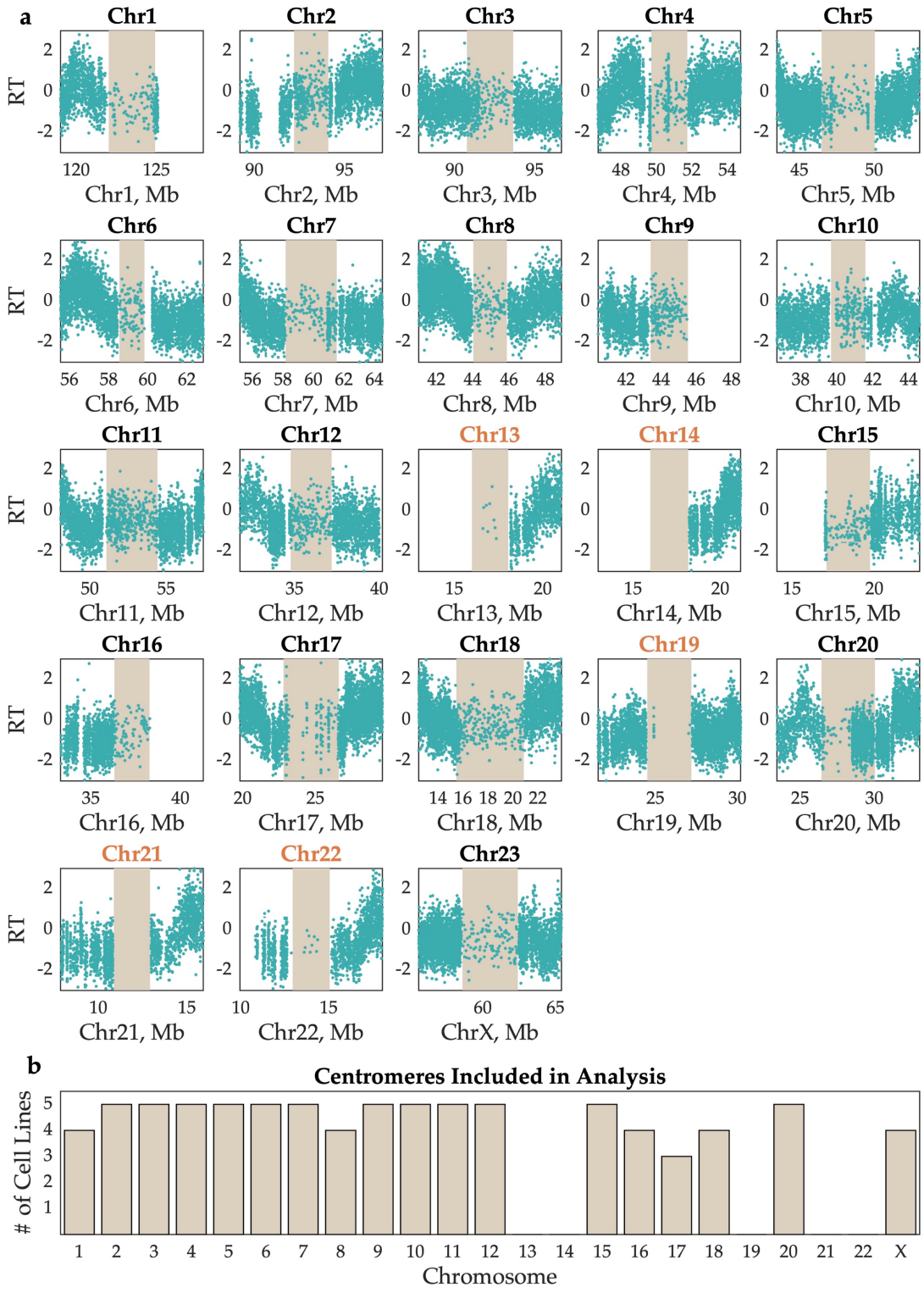


Figure 2.3. **Centromere replication timing can be consistently measured in human cell lines for most chromosomes. a** Unsmoothed replication timing data for the breast cancer cell line HCC1954 across all centromeres (*tan*) and flanking regions. Each dot represents a single

window, defined by 200 reads in the G_1-phase sample. Chromosomes labeled in *black* contain at least 10 centromeric windows and were included in the analyses for **Figure 2.7**, **Figure 2.9**, and **Figure 2.10**. **b** Replication timing inference is successful in the same subset of centromeres across cell lines. Bars represent the number of cell lines in which that chromosome's centromere contained enough windows to be included in further analyses.

The smaller number of successful centromere profiles in A2780 prompted us to consider the effect of sequencing read depth on the ability to infer replication timing profiles: this cell line was sequenced to approximately half the coverage of the others (~80 million filtered read pairs *vs*. ~140–165 million in the others; **Figure 2.4a**, *blue bars*). We hypothesized that because centromeres are highly repetitive, reads derived from those regions would be disproportionately likely to be flagged as poorly-mapped or as PCR duplicates during alignment. Indeed, a large proportion (~85%) of centromeric reads for all samples were flagged as "poorly mapped" and excluded (**Figure 2.5**).

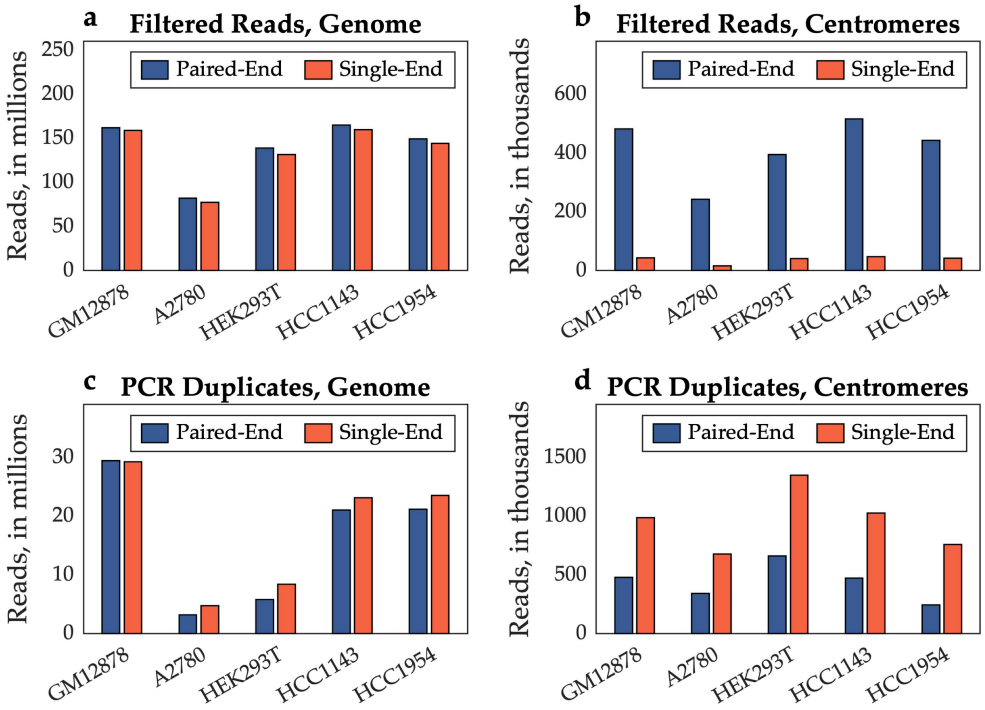


Figure 2.4. **Paired-end sequencing is critical for obtaining centromere replication timing.** Single-end sequencing was generated by considering only the first read of each pair. Read (or read-pair) counts were averaged across the G_1- and S-phase fractions for each cell line. **a, b** Single-end alignment had a negligible effect on read depth genome-wide but eliminated almost all of the reads in the centromeres. **c, d** The difference between single- and paired-end sequencing is largely driven by the difficulty in discriminating true sequence repeats from PCR and optical duplicates with single-end reads. All chromosomes/centromeres were considered for this analysis.

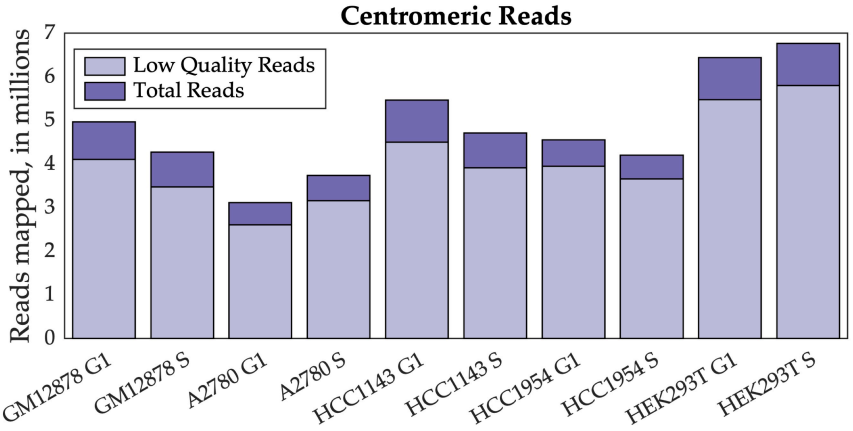


Figure 2.5. **Approximately 85% of read pairs mapped to centromeres are flagged as low-quality and removed prior to analysis**. This reflects the repetitive nature of the centromere

reference sequences. Low- quality reads were defined as those reads with a BWA mapping quality (MAPQ) score < 10.

Strikingly, while total read depth was an important factor in obtaining sufficient usable centromere sequence reads, paired-end sequencing proved to be even more crucial for the success of our approach. We re-aligned the sequencing data, considering only the first read of each read pair, and found that there was a negligible difference genome-wide in the number of reads passing quality filtering when using single-end sequencing (**Figure 2.4a**, *red bars*). In contrast, there was a roughly ten-fold reduction in the number of centromeric reads passing quality filtering (**Figure 2.4b**, *red bars*). In addition, there was a disproportionate loss of reads in A2780: there was on average a 2.6-fold difference in the number of single-end centromeric reads in A2780 (16,367 reads) relative to the other cell lines (HEK293T: 40,408 reads; HCC1954: 41,585; GM12878: 42,854 reads; HCC1143: 47,861 reads). The importance of paired-end sequencing was largely due to the ability to discriminate technical repeats from true repetitive centromeric sequences (**Figure 2.4c, d**). With single-end reads, identical reads are likely to be falsely flagged as PCR or optical duplicates. Paired-end sequencing ameliorates this issue because the probability of observing a non-unique read-pair is much lower than the probability of a non-unique single read. Together, these results establish the technical requirements for mapping DNA replication timing in human centromeres: an order of 100 million or more total reads, and, most importantly, paired-end sequencing.

Centromere replication occurs in mid-to-late S phase and varies among cell lines

Given the ability to measure centromeric replication timing, we next compared these profiles among the five cell lines. We first noted a relatively large variability in centromere replication timing among cell lines. The centromere of a given chromosome replicated earlier than the genome average in some cell lines, but later than the genome average in other cell lines (**Figure 2.6**). While chromosome-wide correlations were relatively high (r = 0.49–0.92; **Figure 2.7**, *blue circles*), the correlations within the centromeres were much less consistent, ranging from r = −0.98 to r = 0.89 (**Figure 2.7**, *gold circles*). Some centromeres, for instance, on chromosome 5, appeared to be more similar across samples (r > 0.5 in six pairwise comparisons). In contrast, other centromeres, such as on chromosome 8, were highly similar between some pairs (r = 0.83) but highly dissimilar between others (r = −0.86). This broad distribution of correlation coefficients was significantly different than would be expected for random genomic regions, controlling for the size of the centromeres (**Figure 2.8a**) and for the small number of centromeric windows relative to genomic windows (**Figure 2.8b**).

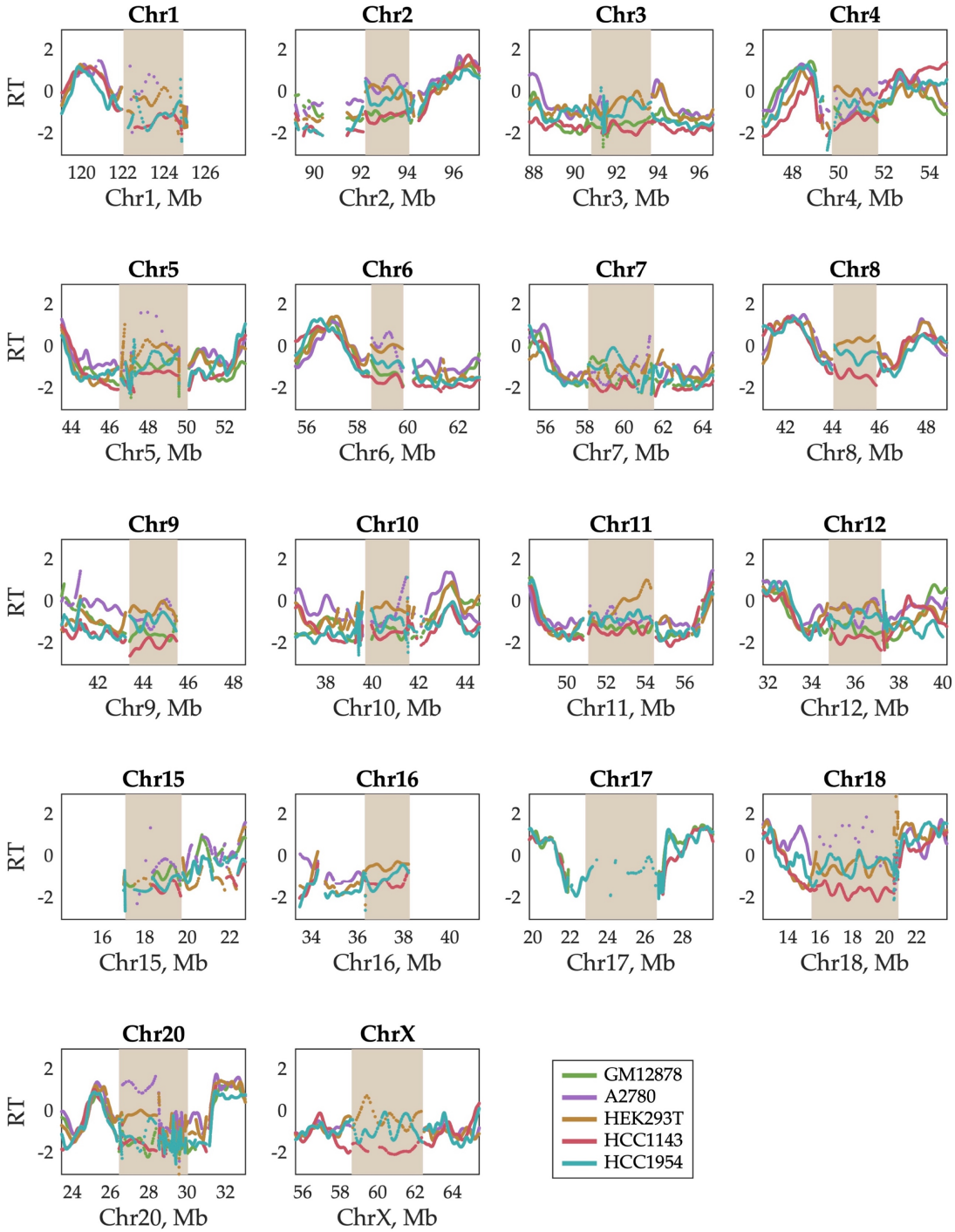


Figure 2.6. **Cell lines display variation in centromeric replication timing across all chromosomes.** Smoothed replication profiles for mappable (see **Figure 2.3**) centromeres in all five cell lines. HEK293T and A2780 tend to have earlier centromeric replication timing than the other cell lines.

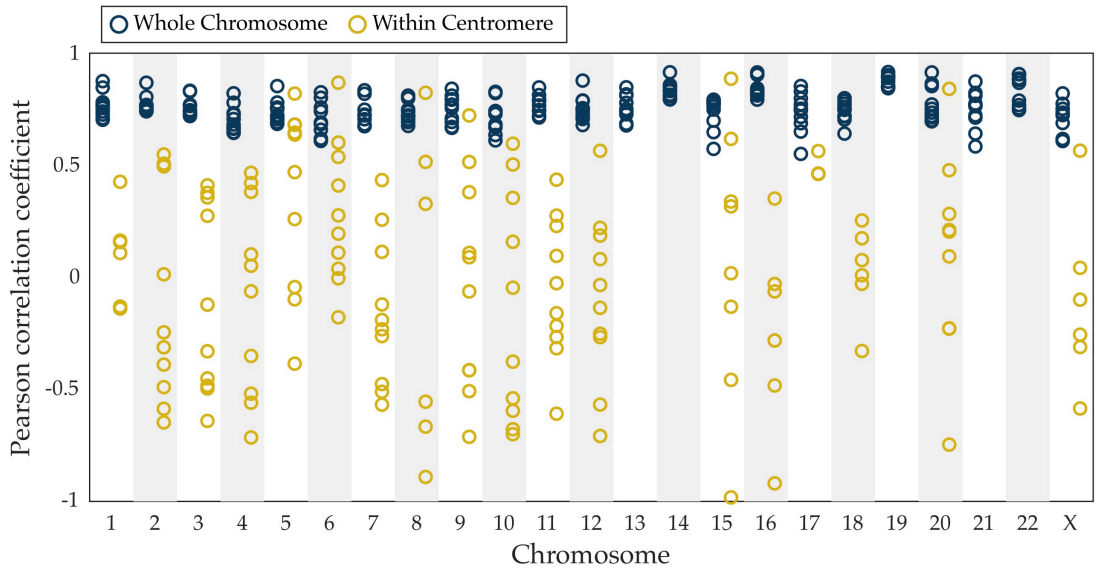


Figure 2.7. **Centromere replication timing (*gold*) is more variable between cell lines than chromosome-wide (*blue*) replication timing.** Pearson correlation was calculated for each mappable centromere (see **Figure 2.3**) and for each chromosome (excluding the centromere) for each pair of cell lines. Each circle represents an individual pairwise comparison. The difference in the distribution of correlation coefficients between the centromeres and whole chromosomes is robust when controlling for the size of the centromeres (**Figure 2.8**).

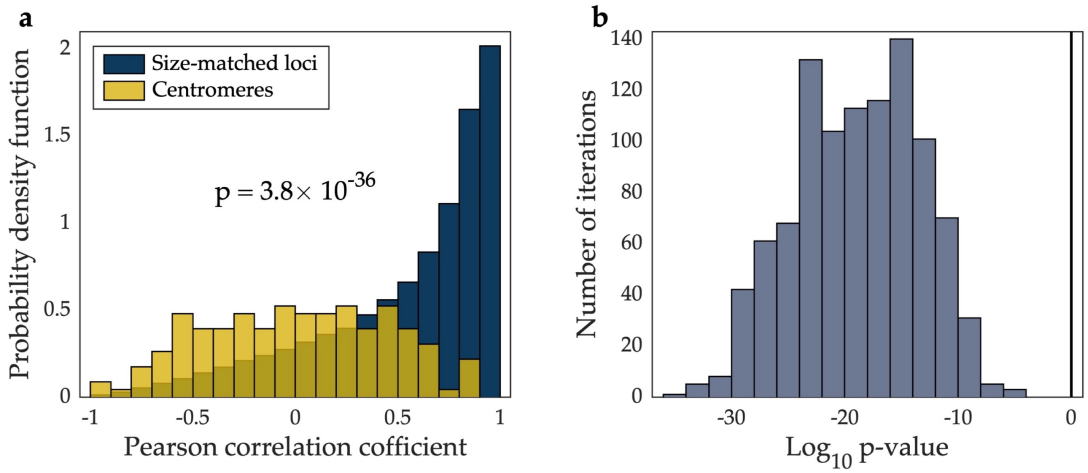


Figure 2.8. **The broad distribution of pairwise correlations for centromeric regions is significantly different than expected by chance. a** The distribution of pairwise correlation coefficients between cell lines for centromeric regions (*gold*) is significantly different from the distribution of pairwise correlation coefficients between cell lines for all possible size-matched windows (*blue*; two-sample Kolmogorov-Smirnov test, p = 3.8 x 10-36). **b** To account for the effects of sampling, we designated random size-matched genomic regions as the "centromeres" and calculated the correlation for these regions between cell lines, for 1,000 iterations. In 999 of

1,000 iterations, the distribution of correlation coefficients for an equivalent number of size-matched random loci was greater and significantly different from the observed distribution of centromeric correlation coefficients. Bars display the $\log_{10}$ p-value from a two-sample Kolmogorov-Smirnov test for each comparison (range: 10^{-34} – 10^{-4}). The black line indicates the Bonferroni-corrected p-value threshold of 5×10^{-5} for 1,000 tests.

We next analyzed the pattern and timing of centromere replication by aggregating centromeres across all chromosomes within each cell line. Although the centromeres of individual chromosomes did not display consistent replication timing across cell lines, centromeres within a given cell line were notably similar, particularly in the cell lines with later average centromere replication (**Figure 2.9**). This trend did not extend to the pericentromeric regions (or genome-wide), which showed much more variable replication timing values within individual cell lines. This observation potentially points to suppression of centromere replication in these cell lines, or alternatively less stringent control of centromeric replication timing in other cell lines.

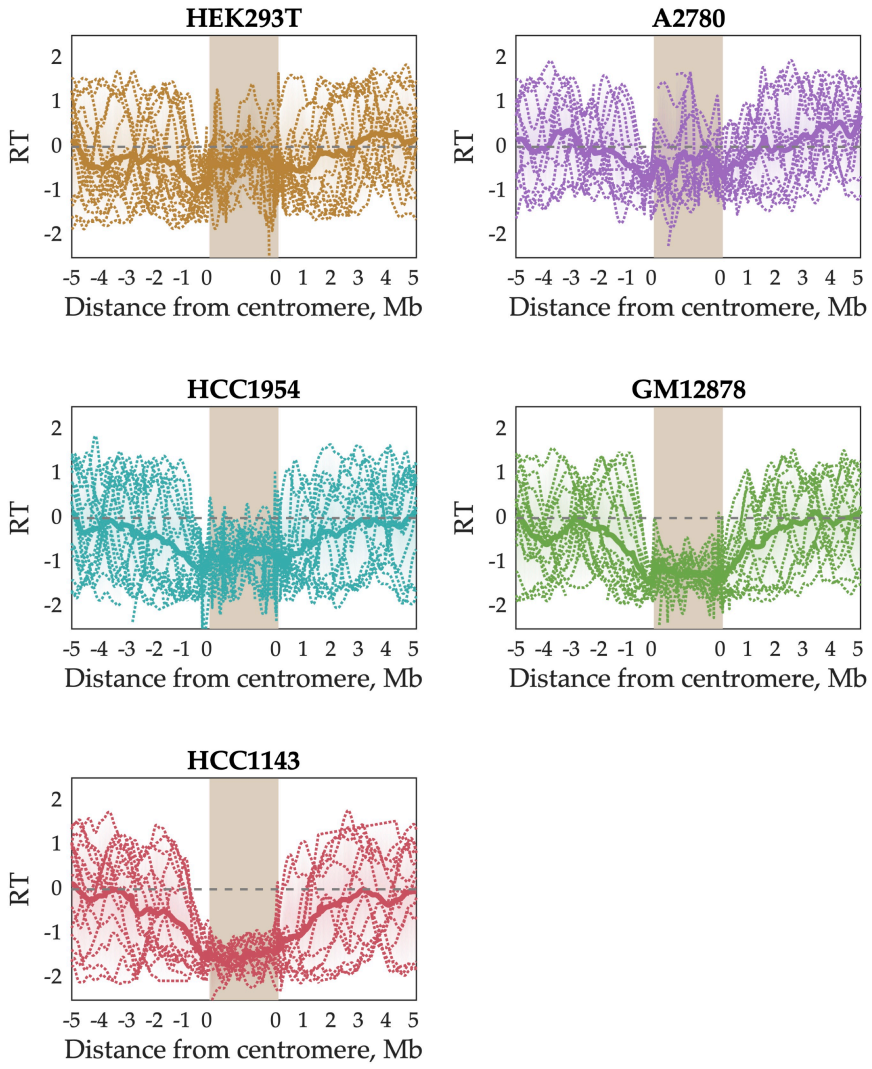


Figure 2.9. **Average replication timing is more consistent within the centromeres of a given cell line than in the surrounding pericentromeres (or the whole genome)**. For each cell line, the replication timing profile for each mappable centromere (see **Figure 2.3**) is shown, overlaid with an averaged "aggregate" profile for that cell line's centromeres. The shaded area indicates the minimum and maximum values, and the dashed line indicates the genome average. Each centromere was divided into 100 bins for the purpose of aggregation.

We also found that the pericentromeric regions replicated progressively later towards the centromeres, indicating that centromeres are embedded within a local area of relatively late replication (**Figure 2.9**). This was even more noticeable when overlaying the aggregate profiles to compare

centromeric replication between cell lines (**Figure 2.10**). However, within the centromeres themselves, this trend either plateaued (GM12878 and HCC1143), or even seemed to reverse (A2780, HCC1954, and HEK293T) such that the centromeres were not later-replicating (or were even earlier-replicating) than their surrounding pericentromeres. Strikingly, the centromeres in A2780 and HEK293T appeared to replicate very close to the genome average, while the other three cell lines showed centromeric replication in mid-to-late S phase. Based on these data, we suggest that human centromeres are not late-replicating, but instead replicate close to mid S phase and often earlier than their surrounding pericentromeric regions.

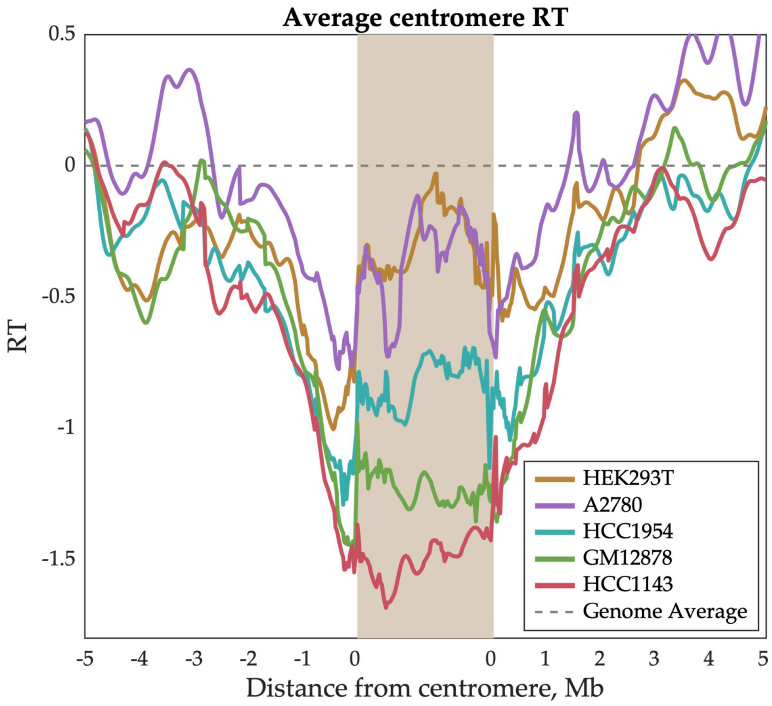


Figure 2.10. **Centromere replication timing is variable between cell lines, occurring between mid and mid-late S phase.** Each line represents the average replication timing of mappable centromeres (see **Figure 2.3**) in the indicated cell line.

Discussion

How and when human centromeres replicate remains an open question. While most studies in mammalian systems have concluded that centromeres replicate in mid-to-late S phase[139-141], several previous reports from human and mouse cell lines have also found evidence that centromeres replicate earlier than other heterochromatic regions[136,138,141,144]. In addition, human centromeres are transcriptionally active[145], which is generally associated with early replication[48,52,53,119]. Thus, despite the well-accepted notion that heterochromatin replicates late in S phase, centromeres appear to comprise a specialized type of chromatin with its own unique biology.

Using the latest human reference genome, hg38, we find that centromeres replicate in mid-to-late S phase, while the neighboring heterochromatin replicates markedly later, in agreement with previous reports that the centromeres replicate earlier than their surroundings[135,138,141]. Intriguingly, two of the cell lines in our study—A2780 and HEK293T—replicated their centromeres close to the genome average, *i.e.*, firmly in the middle of S phase.

We measure DNA replication timing without cell synchronization or fractioning S phase, capturing S phase as a continuous process. In addition, this is the first study to our knowledge to assay DNA replication timing of mammalian centromeres without nucleotide analogue incorporation. These methodological advancements enable us to generate high-resolution replication timing data for centromeres in the context of the whole genome. Furthermore, we are able to assay almost all of the centromeres in a chromosome-specific way, rather than using antibodies against centromeric histones[138], centromere-specific probes[136,139,140] or a pan-centromere consensus

probe[141]. These advantages allow us to observe that centromere replication timing was more variable between the studied cell lines than other regions of the genome. Inter-chromosome variability in centromere replication has previously been reported[137-139,141] but these studies have lacked the resolution to ascribe that variation to particular chromosomes.

The apparently early replication timing of the centromeres relative to the surrounding pericentromere is compatible with evidence from previous reports that centromeres contain active replication origins. Molecular combing has suggested that replication initiation sites are observed at the same density in centromeric regions compared to other genomic regions, and that α-satellite monomers bind the origin recognition complex *in vitro*—both of which imply the existence of active replication origins within centromeres[141]. These origins likely interact with specialized chromatin modifiers to promote origin firing within a generally repressive chromatin context. However, because the centromere reference models we used were assembled probabilistically and not from linear sequencing reads through the centromere, we cannot be confident that the centromere-wide sequence is in the correct order. Thus, at present, the centromere sequence models are not sufficient to generate contiguous replication profiles, from which the locations of these replication origins could be predicted.

We demonstrate here for the first time that human centromeric replication timing can be inferred by high-throughput sequencing and establish the technical requirements and the importance of paired-end sequencing for assaying centromere replication. As newer linear centromere reference sequences become available[67], this approach will prove to be valuable in identifying the specific locations of centromeric replication origins and characterizing variation among cell lines. In establishing a

straightforward method for detecting changes in replication timing of centromeres, we open the door to genetic assays which will help to better characterize the chromatin modifiers that are important for replication activity within these heterochromatin domains.

Methods

Tissue culture

HEK293T and A2780 cells were cultured in Dulbecco's modified Eagle medium (Corning Life Sciences, Tewksbury, MA, USA) supplemented with 10% fetal bovine serum (FBS; Corning). GM12878, HCC1143, and HCC1954 were grown in Roswell Park Memorial Institute 1640 medium (Corning) supplemented with 15% FBS. All cell lines were obtained from the American Type Culture Collection or the Coriell Institute and grown at 37°C in a 5% CO_2 atmosphere.

Fluorescence-activated cell sorting

Asynchronous populations of ~50 million cells were fixed in 70% ethanol, treated with RNase A (10 mg/mL) for 30 minutes at 37°C, and stained with propidium iodide (1 mg/mL) in the dark for 30 minutes at room temperature. Stained cells were flow-sorted on a FACSAria II (BD Biosciences, San Jose, CA, USA) to isolate 1 million G_1- and 1 million S-phase cells.

Library preparation and sequencing

DNA was isolated using the MasterPure™ DNA Purification Kit (Epicentre, Madison, WI, USA) and libraries were prepared with the TruSeq DNA PCR-Free Library Prep Kit (Illumina, Inc., San Diego, CA, USA). Paired-end sequencing was performed for 75 cycles with the Illumina NextSeq 500 (A2780 and HEK293T; Cornell University Biotechnology Resource Center, Ithaca, NY) or for 150 cycles with the Illumina HiSeq X Ten (GM12878, HCC1143, and HCC1954; GENEWIZ, Inc., South Plainfield, NJ, USA).

Sequence alignment

Sequence reads were aligned to the human reference genome hg38 using the Burrows–Wheeler Aligner maximal exact matches (BWA-MEM) algorithm (bwa v0.7.13). For HEK293T and A2780, quality-filtered reads were combined from two independent genomic libraries (HEK293T) or two independent sequencing runs (A2780) to enhance read depth. Centromere coordinates were obtained from the UCSC Genome Browser (University of California, Santa Cruz), genome build GRCh38/hg38. To account for repetitive sequences that might be represented as single copies in the reference genome (thus inflating estimates of copy number), reads were binned in 100kb windows, and the 99th percentile of windows with the highest read coverage were excluded. Similarly, regions with low mappability were filtered by binning reads in 500kb windows and excluding the bottom 0.5% lowest coverage windows. Ploidy was estimated using GenomeSTRiP[63].

Replication timing profiles

Replication timing profiles were generated as in [45]. Briefly, the G_1-phase cells were used to define sliding chromosome windows of equal read depth (200 reads), which were then used to bin the reads from the S-phase cells. Outlier read depth values were filtered using a piecewise segmentation model (MATLAB function `segment`, with assumed variance 0.04). Contiguously-mapped segments between gaps in the reference genome were smoothed with a cubic smoothing spline (MATLAB function `csaps`, with smoothing parameter 1×10^{-16}). Data were then normalized to an autosomal mean of 0 and standard deviation of 1.

Data availability

Sequence data reported in this study have been submitted to the Sequence Read Archive (SRA) under accession number PRJNA419407.
Smoothed replication timing profiles are available at:
http://amnonkoren.com/data.

Acknowledgments

This work was funded by grants DP2GM123495 (to A.K.) and R01GM123018 (to M.B.S.) from the National Institutes of Health. We thank Alexander Nikitin and Alexander Gimelbrant for sharing reagents, Linda (Yu-Ling) Lan and Sean Kim for assistance, David MacAlpine for useful suggestions on the manuscript, and members of our labs for helpful discussions.

CHAPTER 3: TELOMERE-TO-TELOMERE HUMAN DNA REPLICATION TIMING PROFILES

This chapter is published on bioRxiv as: Massey DJ & Koren, A. Telomere-to-telomere human DNA replication timing profiles. *bioRxiv,* 2022.2003.2028.486072, doi:10.1101/2022.03.28.486072 (2022).[146]

Abstract

The spatiotemporal organization of DNA replication produces a highly robust and reproducible replication timing profile. Sequencing-based methods for assaying replication timing genome-wide have become commonplace, but regions of high repeat content in the human genome have remained refractory to analysis. Here, we report the first telomere-to-telomere replication timing profiles in human, using the T2T-CHM13 genome assembly and sequencing data for five cell lines. We find that replication timing can be successfully assayed in centromeres and large blocks of heterochromatin. Centromeric regions replicate in mid-to-late S phase and contain replication timing peaks at a similar density to other genomic regions, while distinct families of heterochromatic satellite DNA differ in their bias for replicating in late S phase. The high degree of consistency in centromeric replication timing across chromosomes within each cell line prompts further investigation into the mechanisms dictating that some cell lines replicate their centromeres earlier than others, and what the consequences of this variation are.

Introduction

Eukaryotic DNA replication initiation is organized in space and time, reflecting a reproducible DNA replication timing program[3]. In general, late

replication appears to be associated with a more repressive chromatin state: late-replicating regions tend to localize to the nuclear periphery[136,147] and to broadly associate with the condensed "B" compartment in chromatin conformation capture assays[47,148]. Likewise, genes in late-replicating regions often have lower expression[48,119], with corresponding histone methylation[46,149] and deacetylation[46,150], than genes in early-replicating regions. Constitutive heterochromatin, which is gene-poor and highly condensed, is often described to be late-replicating[121,122,124], although direct visualization of nascent DNA by microscopy indicates that there are five distinct waves of replication initiation during S phase, with euchromatic replication primarily occurring during the first wave[136]. This suggests that heterochromatin replication timing is likely more complicated than currently appreciated, and potentially points to the existence of distinct heterochromatin subtypes that differ in their replication timing.

Existing methods for measuring replication timing at genome scale[1] are sequencing-based, making them reliant on the quality of reference genome assemblies. Notably, the current human reference genome (GRCh38/hg38) contains 151Mb of unresolved gaps, represented as multi-megabase arrays of unknown sequence[70]. Thus, these regions – which include large pericentromeric regions on chromosomes 1, 9, and 16 and the entire p-arms of the five acrocentric chromosomes (chr13, chr14, chr15, chr21, chr22) – have been refractory to whole-genome analyses, including those of replication timing. In addition, hg38 contains statistically modeled sequences for the centromeric α-satellite DNA, which were designed as decoys for sequence alignment rather than to reflect the true linear sequence of these arrays[68].

We previously reported[118] that these centromeric sequence models in hg38 enabled preliminary analysis of replication timing for the majority of

human centromeres. We found consistent evidence of replication timing peaks within centromeric regions, suggesting that centromeres contain replication origins. We further demonstrated that centromeric replication occurs during mid-to-late S phase and that its timing is highly divergent among cell lines. However, because the decoy sequences in hg38 were not linear assemblies of the centromeres, we were unable to analyze the precise locations of these peaks.

Here, we report telomere-to-telomere replication timing profiles across all autosomes and the X chromosome. Using the telomere-to-telomere human genome assembly T2T-CHM13, recently published by the Telomere-to-Telomere Consortium[70], we provide the first report of replication timing of constitutive heterochromatin in the context of the whole genome. The linear sequences for the centromeres in this genome assembly further enabled us to revisit and reaffirm our previously conclusions based on hg38, while also analyzing the locations of centromeric replication initiation sites.

Results and Discussion

Telomere-to-telomere replication timing profiles

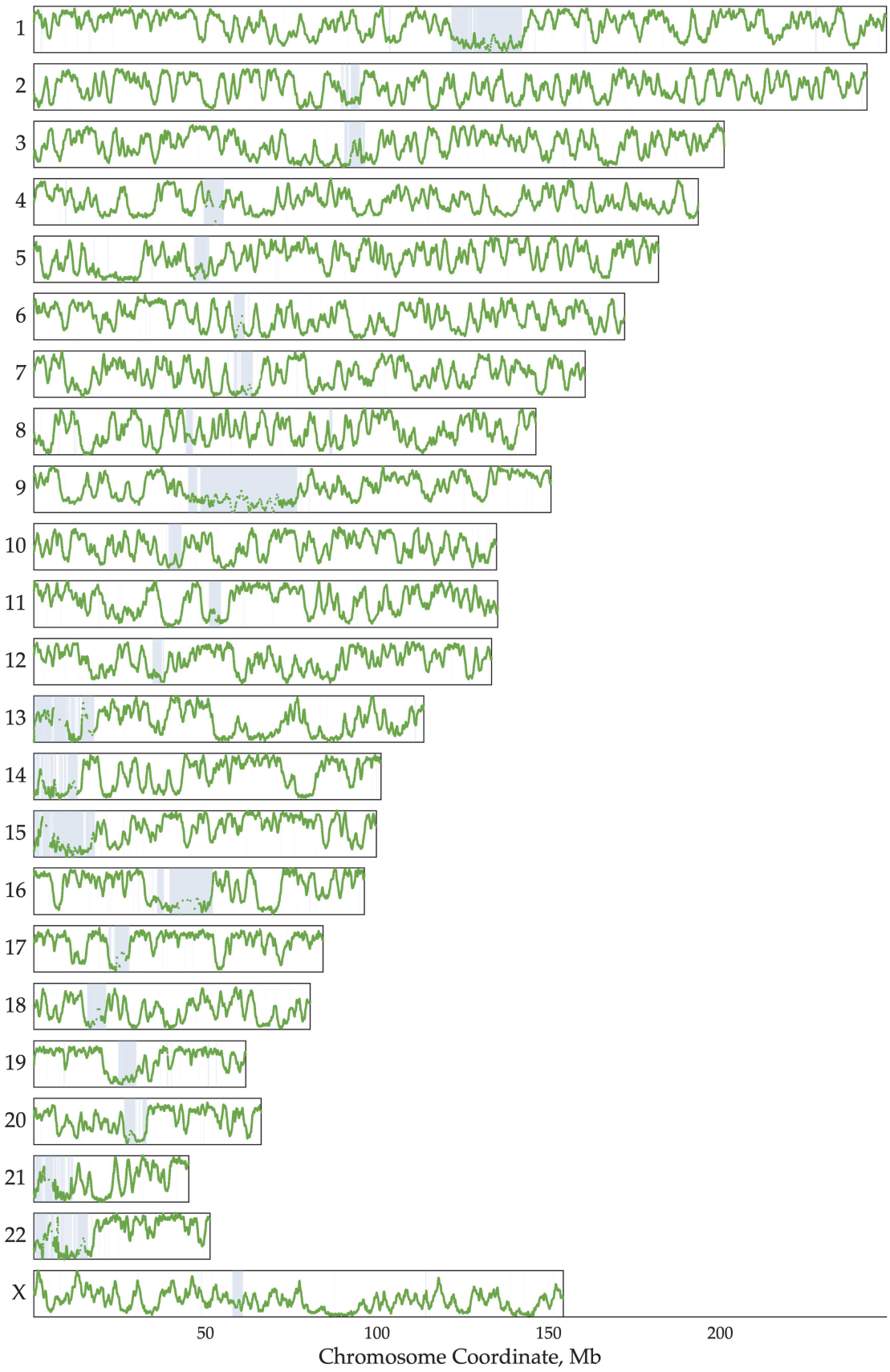
In our prior analysis[118], we generated replication timing profiles for five cell lines – the apparently healthy lymphoblastoid cell line GM12878, the embryonic kidney cell line HEK293T, the ovarian carcinoma cell line A2780, and the breast cancer cell lines HCC1143 and HCC1954 – by whole-genome sequencing of G_1- and S-phase populations isolated by fluorescence-activated cell sorting (FACS). The G_1-phase fraction was used to define variable-size

uniform-coverage genomic windows, accounting for sequencing biases and copy-number variants, and then sequencing read depth was assessed for the S-phase fraction. After S/G_1 normalization, fluctuations in S-phase read depth reflect only the effects of replication timing, such that early-replicating regions are more highly represented relative to late-replicating regions[45].

T2T-CHM13 is a gapless human genome assembly for CHM13-hTERT, a telomerase reverse transcriptase-transformed cell line derived from a complete hydatidiform mole with a stable 46, XX karyotype[70]. Hydatidiform moles are formed during fertilization and contain only DNA from the sperm; thus CHM13-hTERT is homozygous, reducing the complexity of genome assembly. T2T-CHM13 was assembled from long-read PacBio circular consensus sequencing and polished with a combination of other short- and long-read sequencing methods. To assess whether this new assembly could be used to study the replication timing of heterochromatin, we generated replication timing profiles from the same sequencing libraries as in **Chapter 2**, re-aligning the sequencing reads for each cell line to T2T-CHM13.

The resulting replication timing profiles were nearly gapless, with only the rDNA loci remaining as unresolved (**Figure 3.1**). (We note that CHM13-hTERT has an XX karyotype, as do all five cell lines studied. Thus, we did not consider the Y chromosome.) We validated these replication timing profiles by comparison to the hg38-based replication timing profiles, using the UCSC Genome Browser liftOver tool to convert between hg38 and T2T-CHM13 coordinates. The profile for each cell line was virtually identical ($r > 0.999$) between genome builds for regions that could by successfully "lifted over". Notably, this approach was not amenable to FACS-free inference of replication timing from genome sequence data[65] (**Figure 3.2**).

Figure 3.1. **Telomere-to-telomere replication timing profiles for all autosomes and chromosome X.** Regions larger than 5kb that are new in T2T-CHM13 are indicated with blue boxes. The replication timing profile for GM12878 is shown.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
50
100
150
200
Chromosome Coordinate, Mb

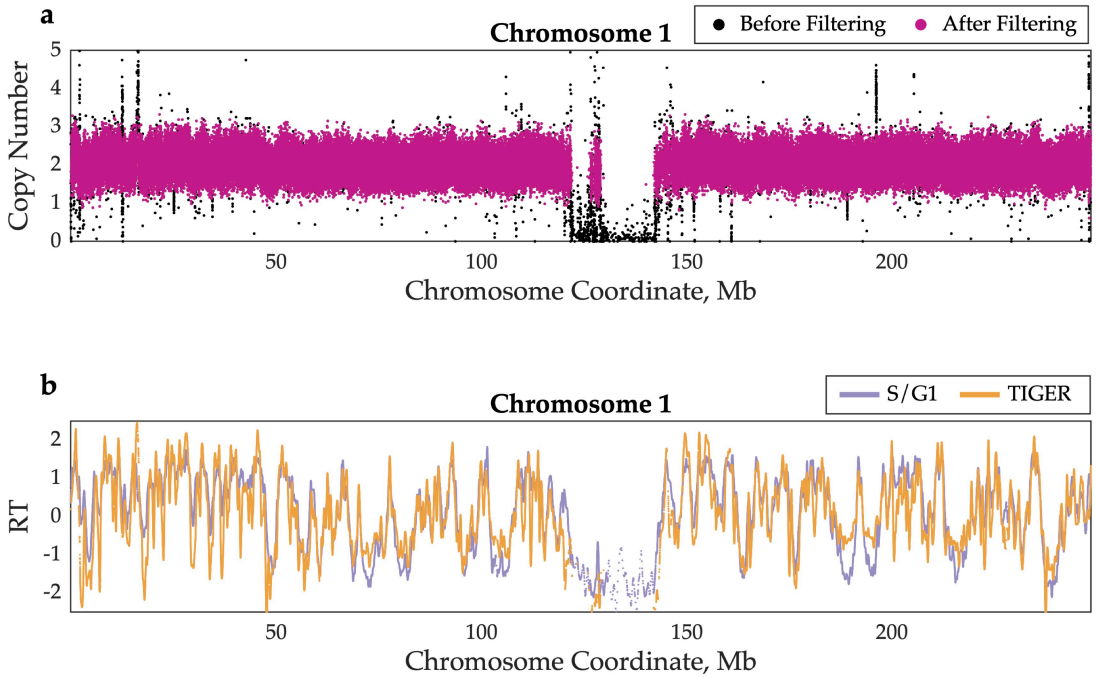


Figure 3.2. **Replication timing analysis of highly repetitive regions requires a G_1-phase control sample. a** Mappability and GC-content corrected read counts for an asynchronous population of GM12878 cells before (*black*) and after (*pink*) filtering regions with abnormal copy-number estimates. The low coverage in the centromeric region is inadequately corrected even after accounting for these sequencing biases using TIGER[65]. **b** Similar replication timing profiles are obtained in non-repetitive regions of the genome between the S/G_1 and TIGER methods.

Our telomere-to-telomere profiles revealed the replication timing of several large regions previously excluded from genomic analysis. This included the entire p-arms of the acrocentric chromosomes (except for the rDNA loci) and the large pericentromeric satellite arrays on chromosomes 1, 9, and 16. The replication timing profiles in each of these regions showed similar structure to the profiles for other genomic regions, with distinct local maxima and minima of varying amplitudes (**Figure 3.3**; **Figure 3.4**). Annotation of these new sequences[69] indicated that these regions include several multi-megabase repeat arrays of distinct satellite sequences, including human

satellite 1 (HSat1; 4.9Mb on chr13p), human satellite 2 (HSat2; 13.2Mb on chr1q, 12.7Mb on chr16q), human satellite 3 (HSat3; 27.6Mb on chr9, 8Mb on chr15p), and β-satellite (1.9Mb on chr22p). Within these larger satellite arrays, HSat1 appeared to replicate in mid S phase, while HSat2 and HSat3 were later-replicating; we further characterize the replication timing of each satellite family, across all family members genome-wide, below.

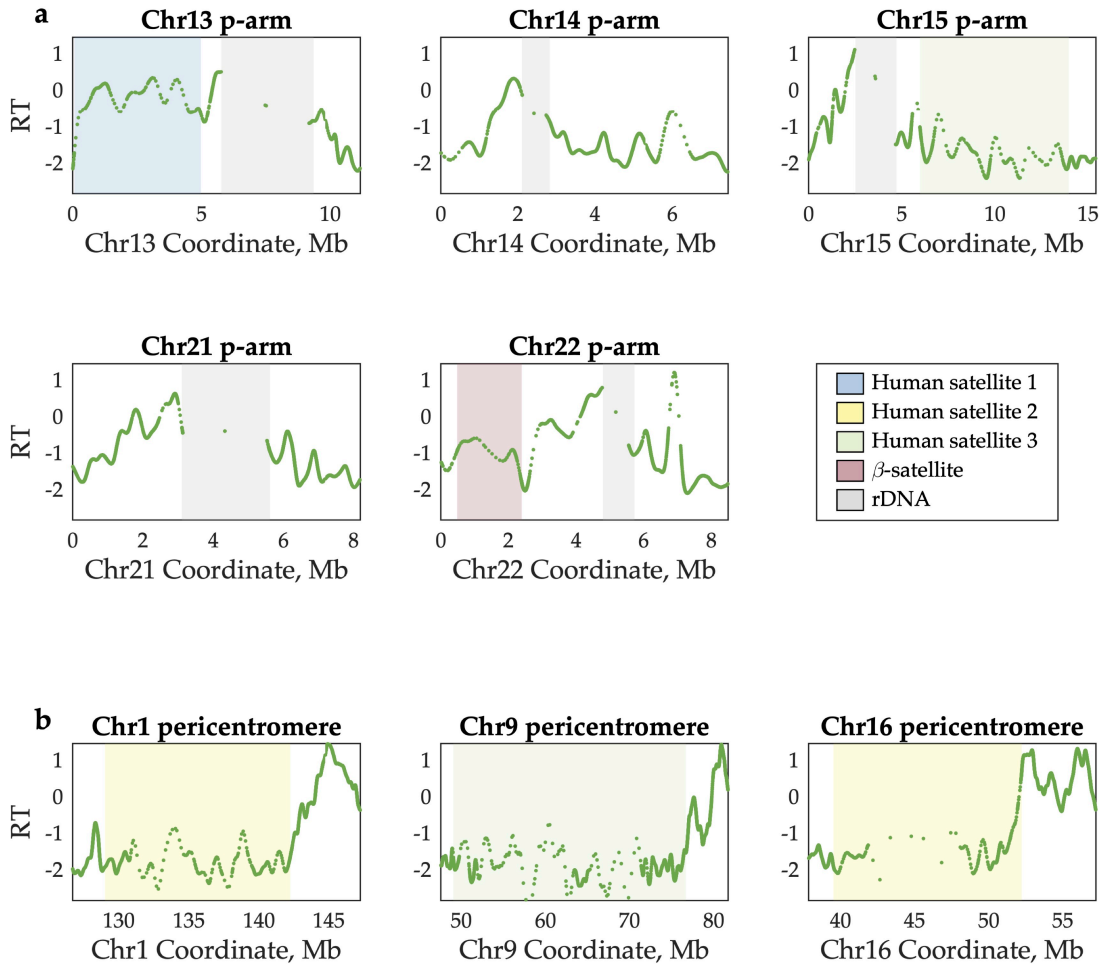


Figure 3.3. **Replication timing (RT) of previously unresolved regions of the human genome.** **a** RT profiles for the six acrocentric p-arms. rDNA arrays (*gray*) remain as gaps in the profile. **b** RT profiles for the large heterochromatin arrays neighboring the centromeres on the q-arms of chromosomes 1, 9, and 16. The RT profile for the lymphoblastoid cell line GM12878 is shown for each region.

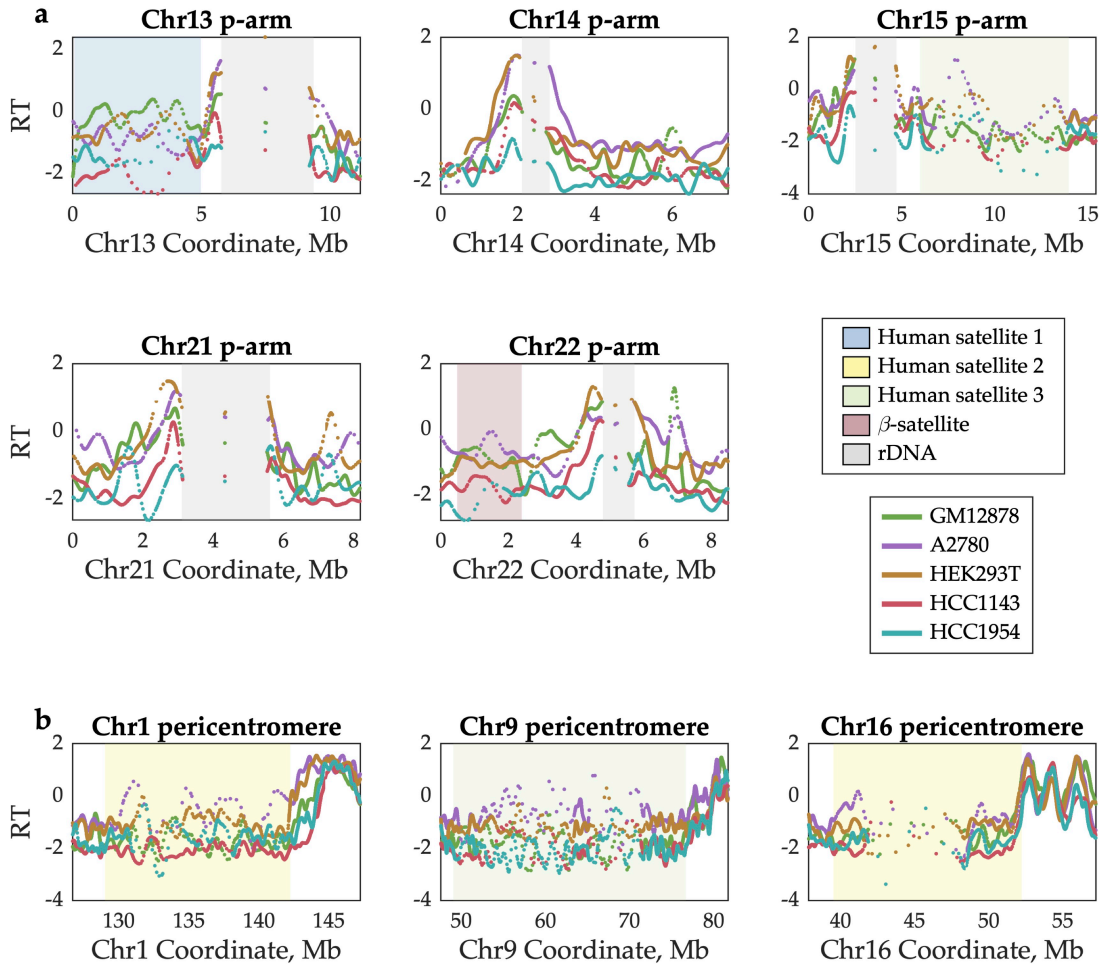


Figure 3.4. **Replication timing (RT) of previously unresolved regions of the human genome for five cell lines.** Compare to **Figure 3.3**.

Next, we visualized the centromeric regions. Using hg38, we previously reported that each centromeric region contains multiple replication timing peaks and that centromeric replication timing peaks were not particularly late relative to the rest of the genome[118]. Although the linear centromeric sequences in T2T-CHM13 completely replace the decoy sequences in hg38, these results were reproduced here (**Figure 3.5**; **Figure 3.6**). Additionally, we were able to meaningfully identify the locations of these local maxima within centromeric regions and to analyze their timing, as we present below.

Furthermore, satellite repeat elements within T2T-CHM13 centromeric regions are well-annotated[69], enabling us to characterize the replication timing of the rapidly-evolving centromere-specific α-satellite DNA, which is present as canonical higher-order repeat arrays (HORs), divergent higher-order repeat arrays, and α-satellite monomers (**Figure 3.7**). Although many of the centromeric regions contain multiple HORs, only a subset is observed to bind kinetochore proteins and function in active centromere assembly[151].

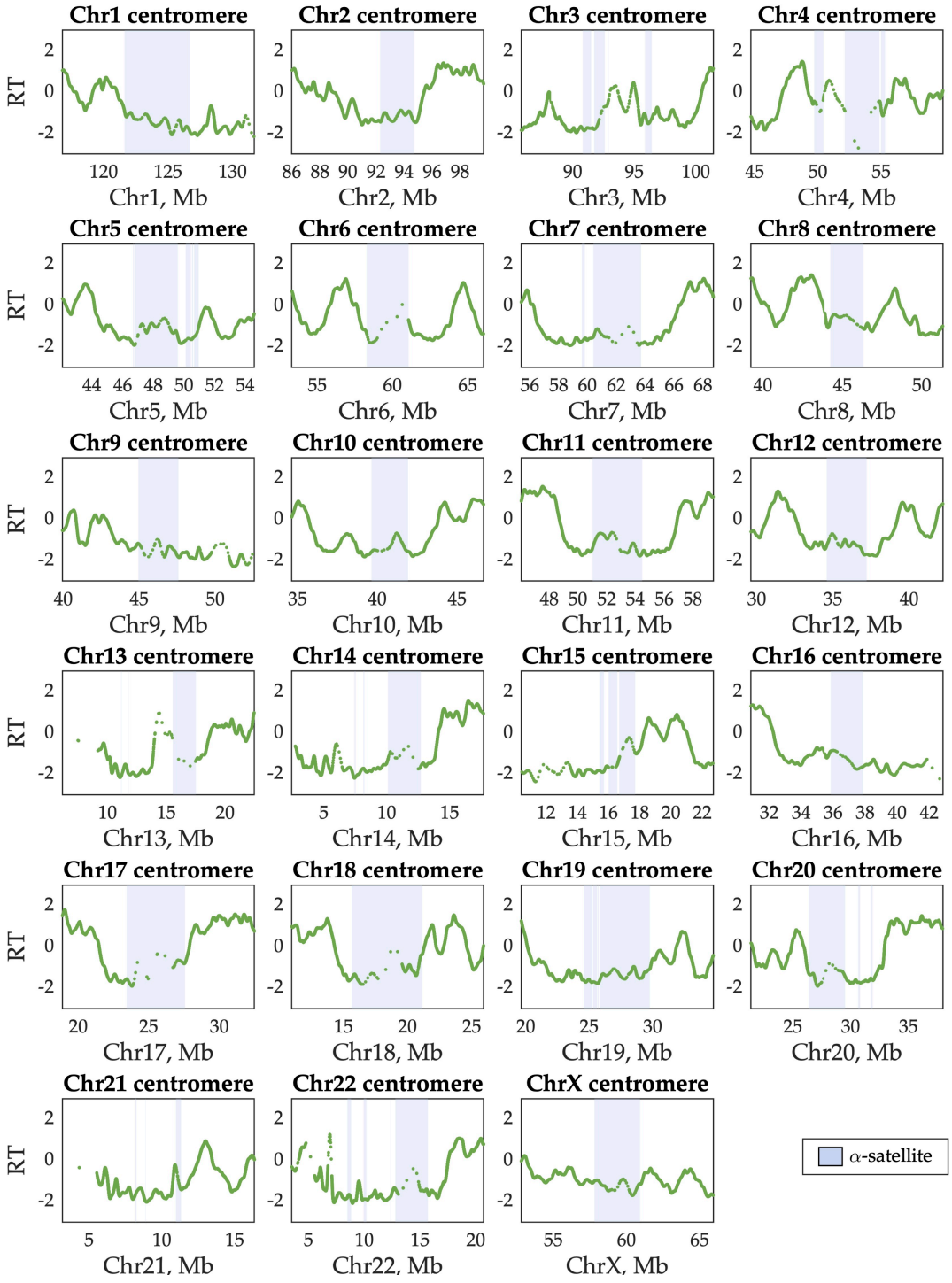


Figure 3.5. **Centromeric replication timing (RT) of all human autosomes and chromosome X.** The locations of α-satellite higher-order repeats on each chromosome, which scaffold active centromere assembly, are indicated in blue. For each chromosome, the entire region shown is annotated as centromeric. The RT profile for the lymphoblastoid cell line GM12878 is shown for each region.

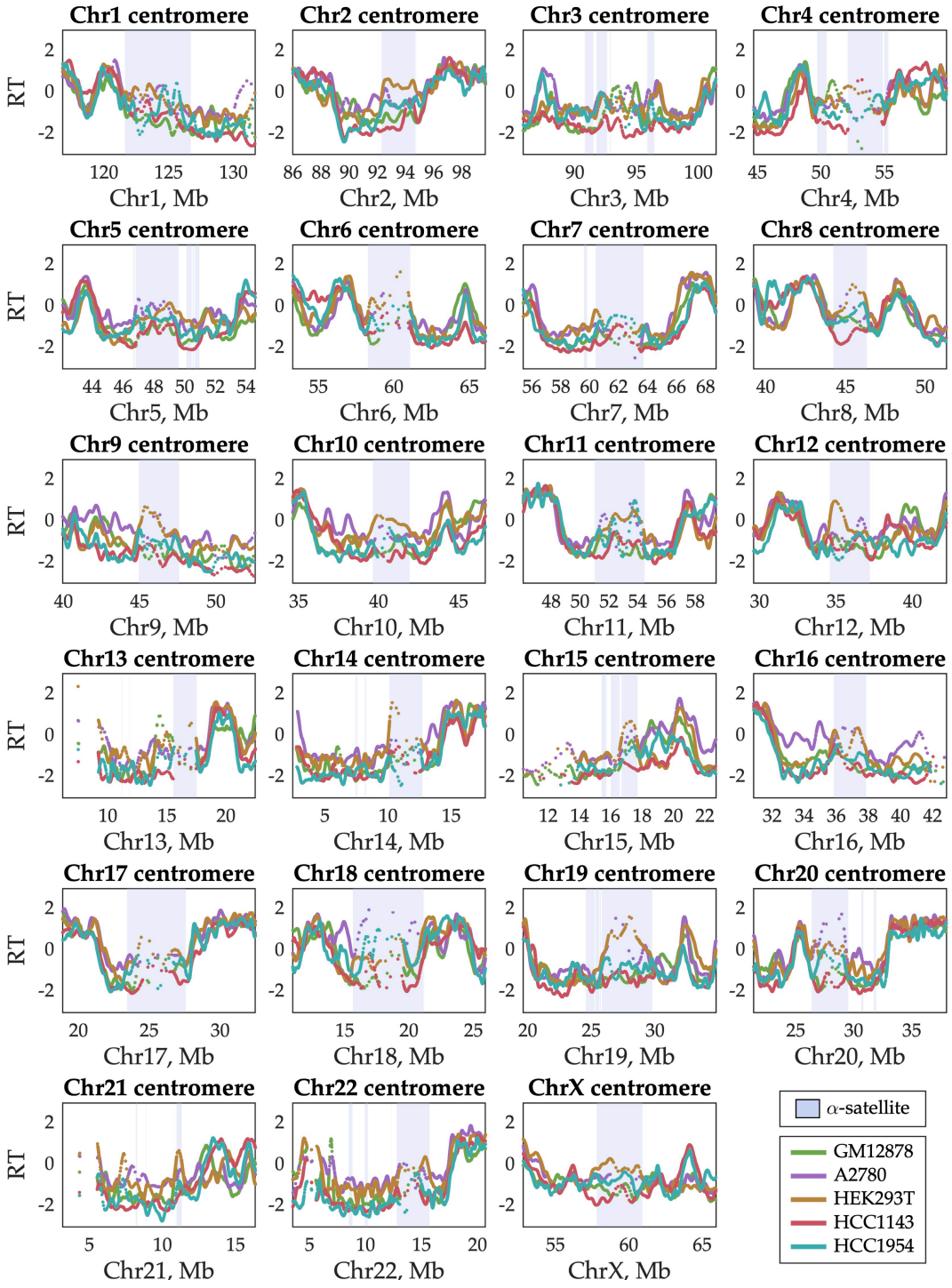


Figure 3.6. **Centromere replication timing (RT) of all human autosomes and chromosome X for five cell lines.** Compare to **Figure 3.5**.

Replication timing bias of repetitive sequence elements

Between the acrocentric p-arms and the centromeric regions, T2T-CHM13 adds 395Mb of densely annotated repeat-rich sequence whose replication timing has not been analyzed. Many of the annotated satellite sequences are relatively short (median: 7.25kb) and neighbored by sequences of other satellite families (**Figure 3.7a**). Thus, we were interested to know whether these satellite families differed from one another in their replication timing: persistent patterns in replication timing of a family across multiple chromosome contexts could reflect some underlying property that controls when it replicates.

Indeed, satellite families did differ in both the median and range of replication timing values observed (**Figure 3.7b**). Replication timing values for non-repetitive sequence in these regions (annotated as "ct") ranged from very early to very late, with a median somewhat later (RT = -0.25) than the genome average (RT = -0.03). In contrast, each of the satellite sequence families was biased toward late replication – although none were exclusively late-replicating. Notably, α-satellite HORs replicated earlier on average than HSat2 and HSat3, but later than HSat1. This is consistent with the notion that the active centromere is earlier replicating than its surrounding context, potentially to facilitate kinetochore loading onto both sister chromatids, at the appropriate time during S phase. Furthermore, late replication of HSat2 and HSat3, evolutionarily related satellites that form large blocks of constitutive heterochromatin, suggests that they may comprise the later waves of replication observed by microscopy[136].

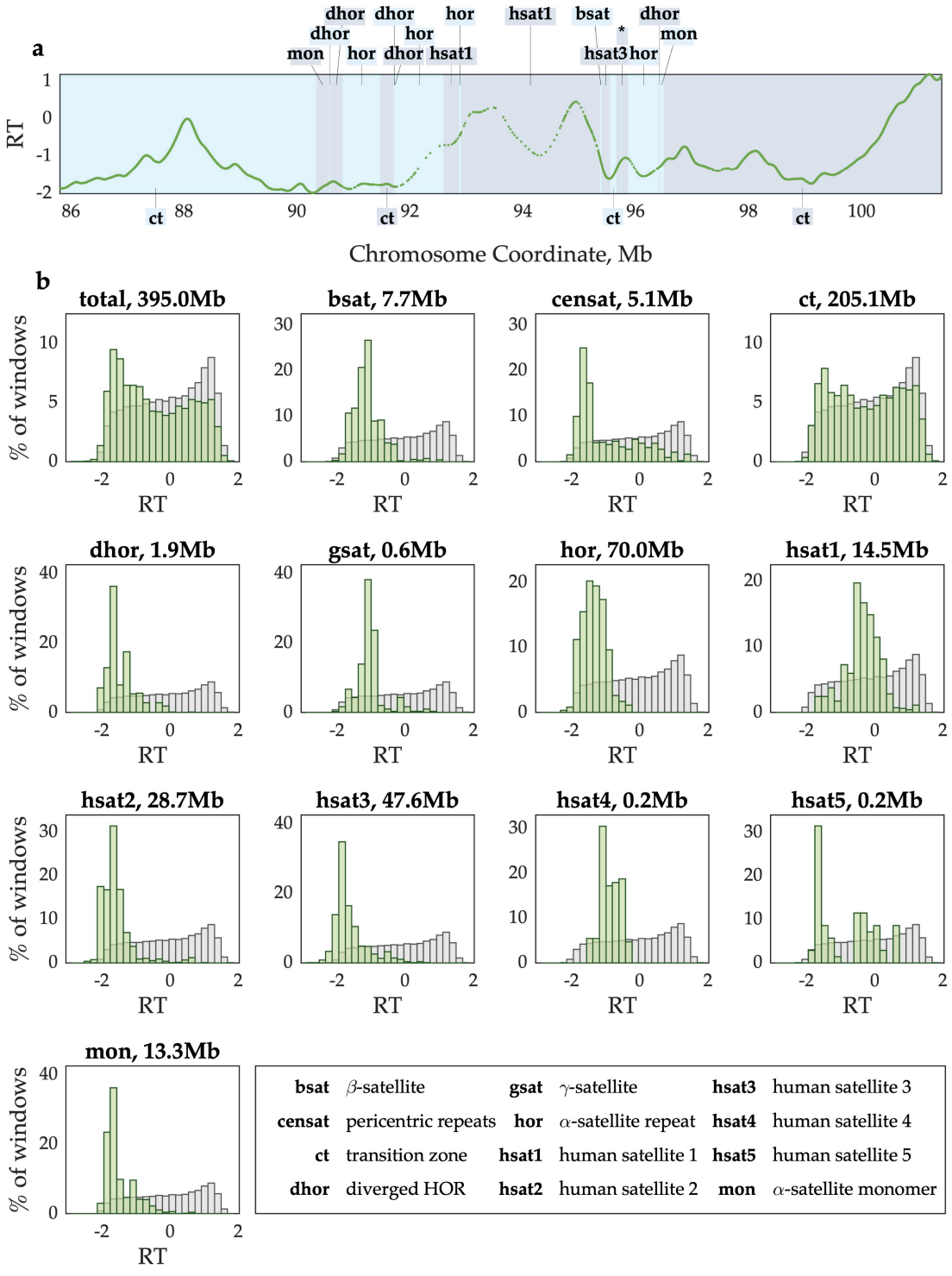


Figure 3.7. **Replication timing (RT) bias of different satellite sequence elements. a** The centromeric region of chromosome 3 is shown. Neighboring sequence elements are denoted in alternating colors. The 200kb region indicated with an asterisk contains 11 sequence elements. **b** For each sequence element category, the distribution of RT values (*green*) is compared to all non-centromeric regions of the genome (*gray*). Apart from transition zones ("ct"), which

include ~5Mb of the p- and q-arms flanking each centromeric region, all satellite families are biased toward late replication timing. However, the α-satellite higher-order repeats ("hor") are earlier-replicating than the large heterochromatic arrays (HSat2 and HSat3). RT values are for the lymphoblastoid cell line GM12878.

Replication dynamics within centromeric regions

Identifying the locations of replication timing peaks within centromeric regions allowed us to next ask about replication dynamics within these regions. We used two metrics to assess replication dynamics: the distance between consecutive replication timing peaks as a proxy for inter-origin distance, and the slope between replication timing peaks and valleys as a proxy for replication fork speed. We observed that inter-origin distances were slightly longer in centromeric regions relative to the rest of the genome (**Figure 3.8a**) and replication timing slopes were slightly shallower (**Figure 3.8b**). While looking specifically within α-satellite HORs, these trends were more pronounced (**Figure 3.8c, d**). This could suggest that the active centromere poses a barrier to replication initiation and/or elongation, resulting in fewer origins firing and/or slower replication progression through these satellite arrays. However, there was substantial overlap between the distributions in all comparisons, indicating that many individual origins have similar dynamics in centromeric and non-centromeric regions. Thus, we favor the explanation that these differences are an artifact of the relatively sparser sequencing coverage of centromeric regions, resulting in an undercounting of centromeric peaks.

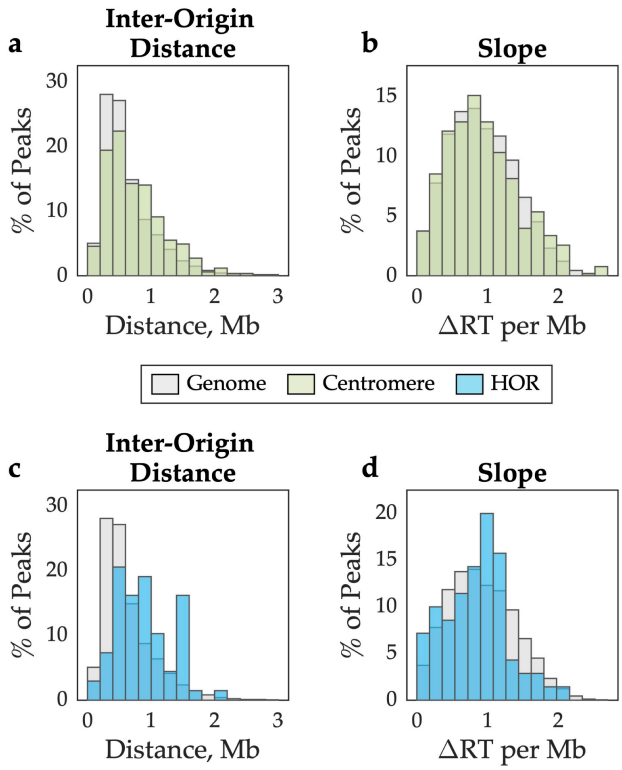


Figure 3.8. **Replication timing (RT) peaks are not substantially different in centromeric regions than in the rest of the genome**. **a, c** The distance between RT peaks was used as a metric of inter-origin distance. Inter-origin distances were slightly larger in centromeric regions (*green*, **a**) and α-satellite higher-origin repeats (*blue*, **c**), relative to the rest of the genome (*gray*). **b, d** RT profile slope was used as a proxy for replication fork speed. For each peak, the ascending and descending slopes are averaged. RT slopes were slightly shallower in centromeric regions (*green*, **b**) and α-satellite higher-origin repeats (*blue*, **d**), relative to the rest of the genome (*gray*). RT values are for the lymphoblastoid cell line GM12878.

Centromeric replication timing varies consistently among cell lines

Finally, we considered differences between the five cell lines analyzed. Replication timing biases of individual satellite repeat families were consistent across cell lines (**Figure 3.9a**). Likewise, inter-origin distances (**Figure 3.9b**) and replication timing slopes (**Figure 3.9c**) were comparable. We had previously observed that there were differences in average centromeric replication timing between these cell lines, such that the average centromeric

region in A2780 and HEK293T was early-replicating and the average centromeric region in HCC1954 and HCC143 was late-replicating (**Figure 2.10**)[118]. Even though the replication timing profiles in these regions could not be "lifted over" between hg38 and T2T-CHM13, this trend was again observed in the T2T-CHM13 profiles (**Figure 3.9d**). Using T2T-CHM13, we were further able to analyze replication timing of individual centromeric regions in each cell line. We found that the trend observed on average reflected a persistent pattern across chromosomes within each cell line, rather than being driven by the replication timing of the larger centromeres (**Figure 3.9e**).

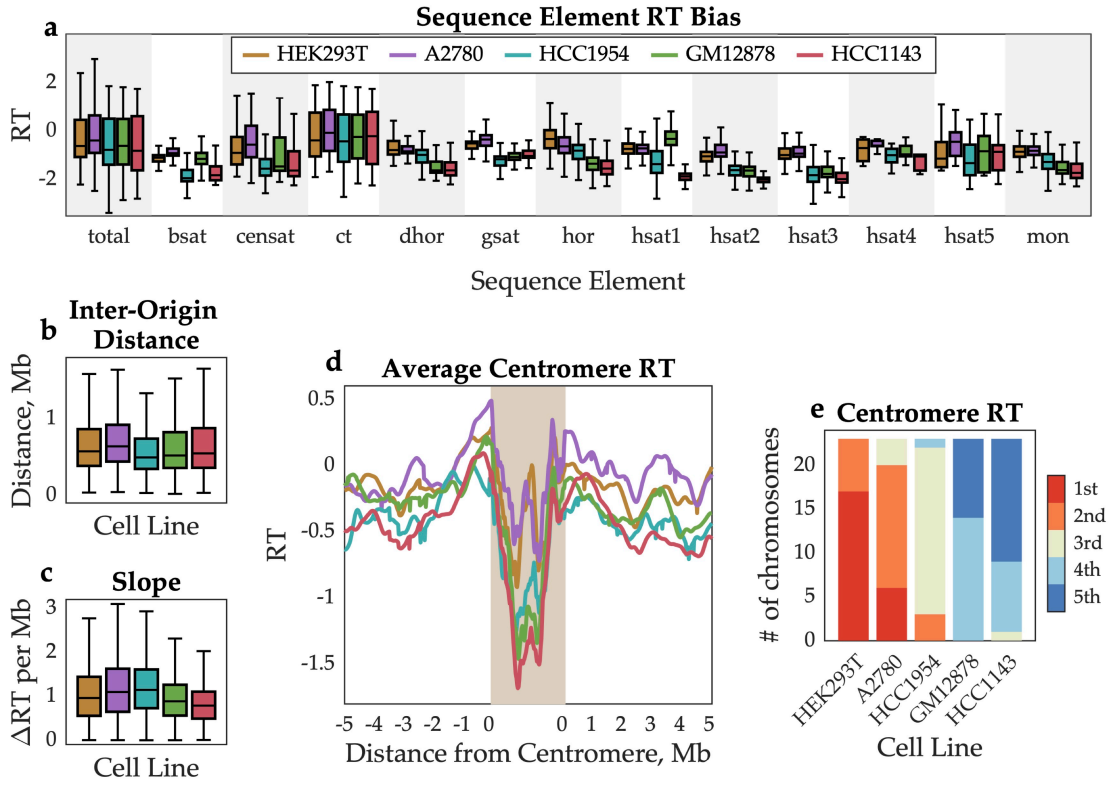


Figure 3.9. **Variability in centromeric regions among cell lines persists across sequence elements and chromosomes. a** The replication timing bias for each centromeric sequence element type is compared across five cell lines. HEK293T and A2780, which have, on average, the earliest centromeric replication timing, are earlier replicating across many different sequence elements. Compare to **Figure 3.7**. **b, c** Inter-origin distance and RT slope are similar across cell lines. Compare to **Figure 3.8**. **d** Average replication timing within centromeric

regions and the flanking 5Mb on either side. For each chromosome, the centromeric region was divided into 100 equally spaced bins. HEK293T and A2780 have the earliest average centromeric replication, while GM12878 and HCC1143 have the latest. Compare to **Figure 2.10**. **e** Differences in centromere replication timing among cell lines are consistent across chromosomes. Each bar represents the number of times that a given cell line is the earliest, 2nd earliest, 3rd earliest, etc. HEK293T and A2780 are consistently the earliest replicating, while GM12878 and HCC1143 are consistently the latest replicating, and HCC1954 is consistently in between.

Taken together, our results indicate that the T2T-CHM13 genome assembly provides a reliable tool for inference of nearly gapless telomere-to-telomere human replication timing profiles. These newly profiled regions confirm that heterochromatin is typically (but not exclusively) late-replicating and reveal differences in replication timing biases of satellite repeat families. Linear centromeric reference sequences enabled us to further confirm our prior findings that centromeres replicate in mid-to-late S phase, are not unusually late-replicating relative to the rest of the genome, and that their timing of replication differs between cell lines. One biological mechanism that could potentially shape differences between cell lines is differential recruitment of the centromere-specific histone H3 variant CENP-A. Variation in HOR array length and sequence divergence has been shown to influence the competency of centromeric regions to recruit CENP-A[152], and *in vitro* experiments suggest that depletion of CENP-A during S phase results in replication fork stalling specifically at centromeres[153]. Thus, sequence and copy-number variation at centromeric regions among cell lines may alter the replication timing of individual chromosomes. However, by comparing centromeric regions within the same cell line, we demonstrate that earlier centromeric replication timing appears to be a global phenomenon impacting all chromosomes. An intriguing possibility is that centromeric replication is coordinated across chromosomes, perhaps by their nuclear localization:

centromeres are strongly enriched for intrachromosomal interactions in budding yeast[154] and centromere location within the nucleus has been implicated in the maintenance of pluripotency in human embryonic stem cell lines[155]. In that scenario, advancing the replication timing of one centromere could have the impact of altering global centromeric replication timing. To our knowledge, such a mechanism has yet to be described. Likewise, the consequences of divergent centromeric replication timing between cell lines remain unclear.

Methods

Preparation of whole genome sequence data

All sequence data analyzed in this study were previously published in Massey *et al.*[118]. Tissue culture, fluorescence-activated cell sorting, library preparation, and sequencing are detailed in **Chapter 2**.

Sequencing reads were re-aligned to the human genome assembly T2T-CHM13 v1.1 with the Burrows-Wheeler Aligner maximal exact matches (BWA-MEM) algorithm (bwa v0.7.13). Sequence annotations are from Altemose *et al.*[69] and were downloaded from the UCSC Genome Browser (University of California, Santa Cruz; "cenSatAnnotation" track). For acrocentric chromosomes, the p-arm boundary of the centromere was defined as 5Mb from the p-most HOR element. For chromosomes 1, 9, and 16, the q-arm boundary of the centromere was defined as 5Mb from the q-most HOR element.

Replication timing profiles

Replication timing profiles were inferred by the S/G_1 method described in Koren *et al*. (2012)[45]. Briefly, variable-size genomic bins were defined such that each bin had uniform coverage (200 reads) in the G_1-phase library for a given cell line. Per-bin coverage was calculated for the corresponding S-phase library. The resulting profile was smoothed using a cubic smoothing spline (MATLAB function `csaps`, with smoothing parameter 1×10^{-16}), and normalized to an autosomal mean of 0 and standard deviation of 1.

Data availability

Sequence data analyzed in this study are available from the Sequence Read Archive (SRA) under accession number PRJNA419407.

Acknowledgements

This work was supported by the National Institutes of Health (DP2-GM123495 to A.K.) and the National Science Foundation (MCB-1921341 to A.K.).

CHAPTER 4: HIGH-THROUGHPUT ANALYSIS OF SINGLE HUMAN CELLS REVEALS THE COMPLEX NATURE OF DNA REPLICATION TIMING CONTROL

This chapter is published on bioRxiv as: Massey DJ & Koren, A. High-throughput analysis of single human cells reveals the complex nature of DNA replication timing control. *bioRxiv,* 2021.2005.2014.443897, doi:10.1101/2021.05.14.443897 (2022).[156]

Abstract

DNA replication initiates from replication origins firing throughout S phase. Debate remains about whether origins are a fixed set of loci, or a loose agglomeration of potential sites used stochastically in individual cells, and about how consistent their firing time is. We developed an approach to profile DNA replication from whole-genome sequencing of thousands of single cells, which includes "*in silico* flow cytometry", a method for discriminating replicating and non-replicating cells. Using two microfluidic platforms, we analyzed up to 2,437 replicating cells from a single sample. The resolution and scale of the data allow focused analysis of replication initiation sites, demonstrating that most occur in confined genomic regions. While initiation order is remarkably similar across cells, we unexpectedly identify several subtypes of initiation regions in late-replicating regions. Taken together, high throughput, high resolution sequencing of individual cells reveals previously underappreciated variability in replication initiation and progression.

Introduction

Faithful duplication of the genome is a critical prerequisite to successful cell division. Eukaryotic DNA replication initiates at replication origin loci,

which are licensed in the G_1 phase of the cell cycle and fired at different times during the S phase. In many eukaryotes, sequencing of cells at different stages of the cell cycle has been used to profile DNA replication timing, which measures the relative time that different genomic regions are replicated during S phase (reviewed in[157]). This replication timing program is highly reproducible across experiments[3], suggesting strict regulatory control; and conserved across phylogeny[47,51], suggesting selection under evolutionary constraint. However, the molecular mechanisms that determine the locations and preferred activation times of replication origins in mammalian genomes remain unclear. Furthermore, debate persists over whether the reproducible nature of the replication timing program reflects the consistent activity across cells of specific individual replication origins or stochastic firing of different origins in different cells within a given region. Ensemble replication timing measurements have been interpreted to indicate that replication is organized in broad "domains", spanning hundreds of kilobases to several megabases with consistent replication timing governed by the activity of clusters of replication origins[48,158]. Furthermore, some recent replication origin-mapping methods have indicated that replication origins are highly abundant and highly dispersed throughout the human genome[96,157], suggesting that many sites may function as origins used in a subset of cell cycles. In contrast, high-resolution measurements of hundreds of human replication timing profiles[46,64], or replication timing across multiple S-phase fractions[62], support initiation of replication from more localized genomic regions. While these replication timing methods reveal genomic regions that reproducibly replicate at characteristic times during S phase, it remains contested whether these represent a conserved pattern across cells or reflect the average behavior of single cells. Previous work has modeled how the stochastic firing of

replication origins could be sufficient to explain the replication timing profile[88,89], and single-molecule experiments (*e.g.*, with DNA combing) have suggested that cells may use different subsets of origins in each cell cycle[78,79].

Recently, replication timing has been analyzed by single-cell sequencing of several hundred mouse or human cells[106,108,109]. These studies focused on cells in the middle of S phase and analyzed replication at the level of domains, concluding that stochastic variation exists in replication timing and is highest in the middle of S phase. However, single-molecule and single-cell studies have been limited in their throughput and biased toward early S phase or mid S phase, respectively. Analyzing many cells is particularly important given that even when the whole genome is captured, a single cell provides only a snapshot of DNA replication at a single moment in time. By assaying many cells at different stages of S phase, it is possible to string these snapshots together to construct a picture of replication states over time. However, the resolution of this picture will be dependent both on capturing cells at many stages of S phase and on assaying a large number of cells.

Here, we report the analysis of whole-genome sequencing of thousands of single replicating cells across ten human cell lines. We developed an *in silico* approach to sort cells by cell cycle state, allowing us to capture cells throughout the full duration of S phase, and to analyze them in any number of sub-S-phase fractions down to single-cell resolution. We found that single cells within a given cell line largely used a consistent set of replication initiation regions, which were discrete genomic loci rather than megabase-scale domains. Furthermore, these initiation regions fired in a predictable, albeit not fixed, order. Some initiation regions were consistently fired early in S phase across cells, while others were fired consistently late. However, we also identified a subset of rarely fired initiation regions with a preference for early

firing and another subset that fired throughout S phase. We conclude that a consistent set of replication origins explains the vast majority of replication initiation events in single cells, and that existing models of replication timing fall short of explaining the diversity of firing time patterns.

Results

High-throughput measurement of single-cell replication

Previous sequencing-based studies measured DNA replication timing in a relatively small number of cells, mostly limited to mid S phase cells[108,109]. To analyze single cells, these studies performed DNA amplification using degenerate oligonucleotide-primed PCR (DOP-PCR), with one reaction per cell. Due to technical noise introduced during amplification, these studies were limited to analyzing replication timing at the level of large chromosomal domains (typically on the order of megabases). As an alternative approach, we devised a method to study DNA replication timing across the entire span of S phase, in hundreds to thousands of cells, and with higher spatial resolution than previous methods. Specifically, we used two microfluidic systems that isolate and barcode single-cell DNA: the 10x Genomics Single Cell CNV platform, which performs multiple-displacement amplification (MDA) on pooled barcoded cells, and direct DNA transposition single-cell library preparation (DLP+)[114], which is an amplification-free method. Both library preparation methods were followed by whole-genome sequencing of single cells. With each platform, automation of labor-intensive steps allows for dramatic increases in throughput. In addition, recent studies suggest that

improved MDA protocols have reduced noise relative to previous single-cell amplification methods like DOP-PCR[116].

As an initial proof-of-principle, we analyzed 5,793 cells from the human lymphoblastoid cell line (LCL) GM12878 isolated with the 10x Genomics system, following fluorescence-activated cell sorting (FACS) of G_1-, G_2-, and several fractions of S-phase cells. The resulting sequencing data were sufficient to distinguish replicating cells from non-replicating cells across a five-fold range of sequencing read depths (50-250 reads per Mb). Specifically, local read depth fluctuated more in replicating cells relative to non-replicating cells of similar coverage (**Figure 4.1a**). To validate that these fluctuations could be used to computationally distinguish replicating cells from non-replicating cells within an unsorted population, we quantified them using MAPD (median absolute deviation of pairwise differences between adjacent genomic windows[159]), which scales proportionally to read depth (**Methods**). Indeed, FACS-sorted G_1- and S-phase cells had distinct linear relationships between scaled MAPD and average read depth (**Figure 4.1b**). Therefore, we were able to computationally assign each cell as "G_1" or "S" (Figure 4.1b) and compare the resulting fractions to the FACS labels. *In silico* sorting was highly concordant with FACS labels (**Figure 4.1b**; **Figure 4.2**), allowing us to perform additional experiments without FACS. Accordingly, we sequenced an additional three GM12878 samples without cell sorting, recovering an additional 3,787 cells in total. We analyzed these cells together with the sorted cell libraries, as described below.

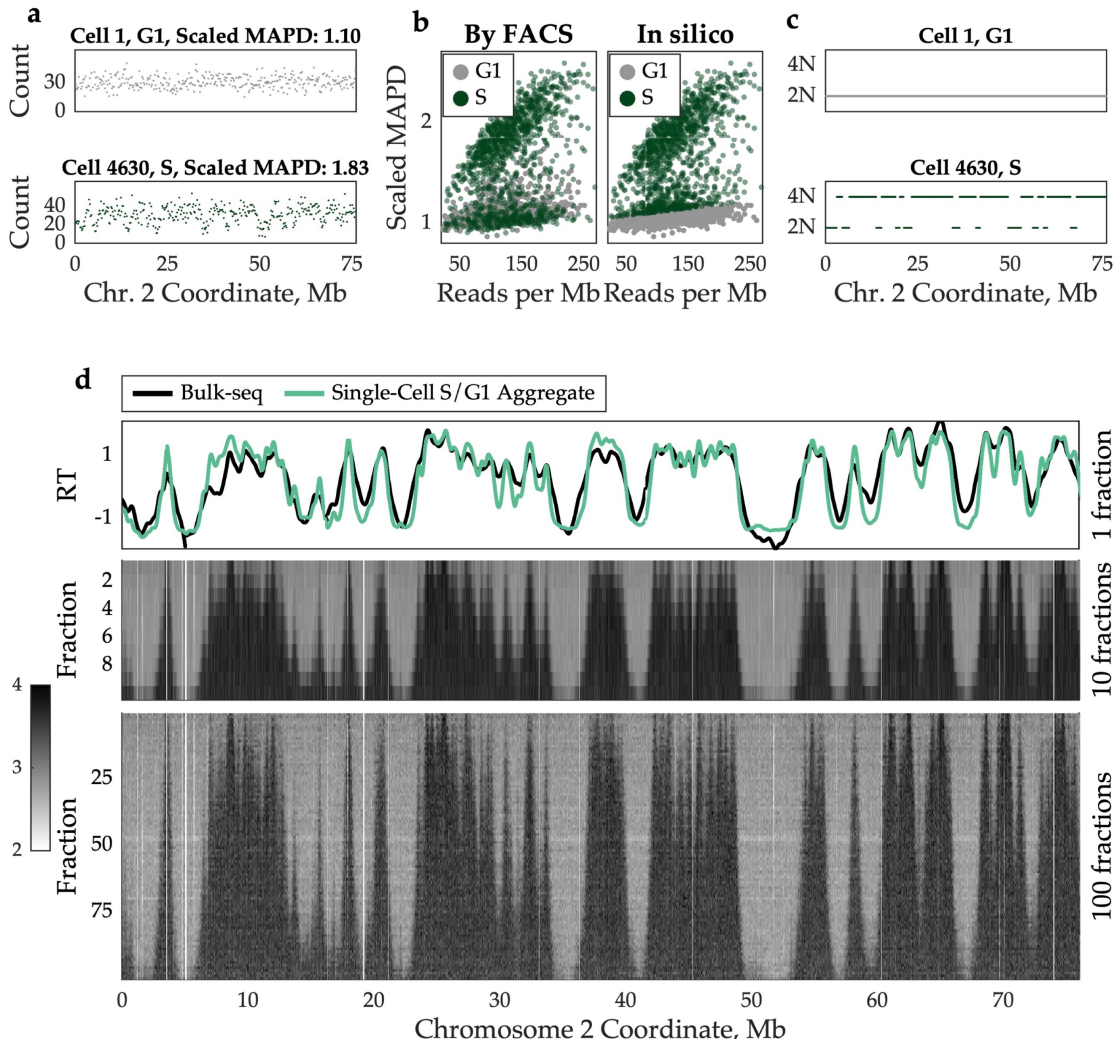


Figure 4.1. **Discrimination of replicating and non-replicating cells by *in silico* flow cytometry. a** Non-replicating G1 cells (*e.g.,* Cell 1, *top*) have a relatively uniform sequencing read depth across the genome, whereas S-phase cells (*e.g.,* Cell 4630, *bottom*) display fluctuations in read depth, consistent with the presence of two underlying copy number states. Each dot represents raw read count in a 200kb window. **b** Flow-sorted single cells (*left*) can be accurately sorted *in silico* (*right*). Replicating S-phase cells display a higher degree of read-depth fluctuation relative to non-replicating G1-phase cells sequenced to equivalent coverage (quantified by scaled MAPD; median absolute pairwise difference between adjacent genomic windows divided by the square root of mean coverage-per-Mb). *Left*: cells are labeled as G1- (*gray*) or S-phase (*green*) based on FACS sorting. Only the G1- and S-phase fractions are shown. *Right*: the same cells are labeled as G1- or S-phase based on scaled MAPD, revealing widespread S-phase contamination in the G1 FACS sample. **c** Replication profiles were inferred for each single cell, using a two-state hidden Markov model. Non-replicating cells (*e.g.,* Cell 1, *top*) display a single copy number (2N), while replicating cells (*e.g.,* Cell 4630, *bottom*) display two distinct copy number states (2N and 4N). Each dot represents the inferred replication state in a 20kb window. The same region is shown from **a**. **d** Aggregating data across S-phase cells into one or more fractions reveals a consistent structure of replication

progression at different times in S phase. *Top*: an ensemble replication timing profile inferred from all S-phase cells together (*green*) was highly correlated with a bulk-sequencing replication timing profile for the same cell line (*black*). *Middle, bottom*: single cells were aggregated into 10 or 100 fractions based on S-phase progression. Pileups of high read depth (caused by replication in most/all cells in the fraction) are observed in discrete locations across the chromosome. The triangular structure of these pileups suggests that replication initiation occurs from fixed loci and proceeds symmetrically in both directions. Each row represents one fraction (containing multiple cells), and each column represents a fixed-size window of 20kb. Low-mappability regions (*white*) have been removed.

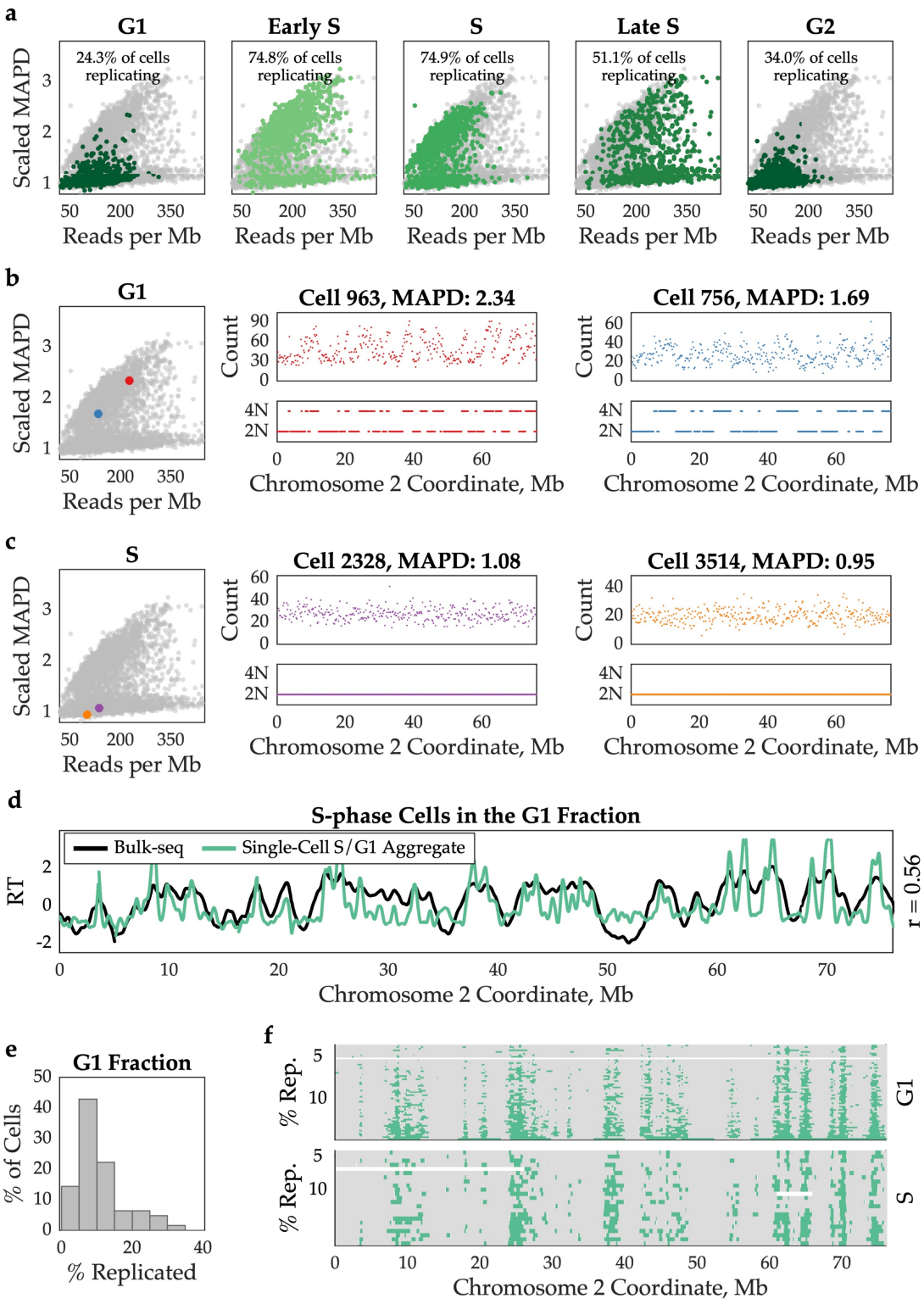


Figure 4.2. ***In silico* sorting of cells recapitulates fluorescence activated cell sorting (FACS).** **a** GM12878 cells were sorted into five fractions by FACS prior to sequencing. For each fraction,

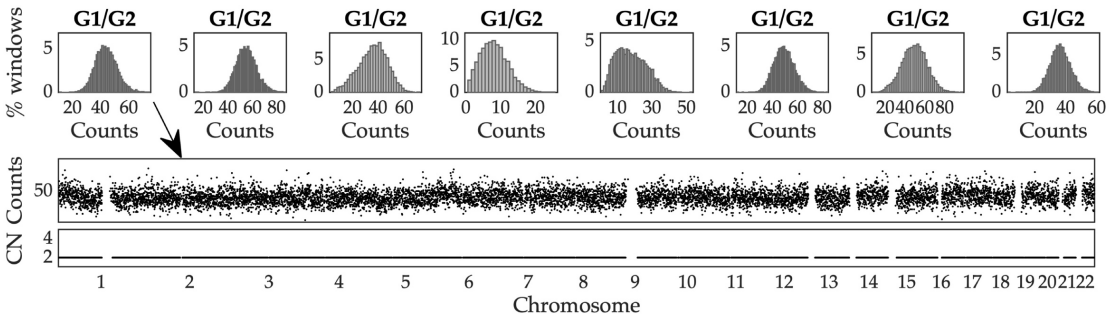
the corresponding cells are highlighted in green. Each fraction contains a mixture of G_1/G_2- and S-phase cells according to *in silico* sorting (as in **Figure 4.1b**). **b** The FACS G_1 fraction contained several cells with higher scaled MAPD than expected for a non-replicating cell. Two copy-number states were inferred across each chromosome, suggesting that these cells were replicating. **c** The FACS S-phase fraction contained many cells with lower scaled MAPD than expected for a replicating cell. A sole copy-number state was inferred across each chromosome, suggesting that these cells were not replicating. **d** An S/G_1 aggregate replication timing profile inferred from the G_1 FACS fraction library, based on preliminary *in silico* assignment of 323 cells as "S", recapitulated the bulk-sequencing profile for this cell line. **e** Cells within the G_1 FACS fraction assigned as replicating after copy-number inference (n = 63) were less than 35% replicated (*i.e.*, in early S phase). There was one outlier (87% replicated). **f** Replication profiles for early-S-phase cells in the G_1 and S FACS fractions were visually indistinguishable.

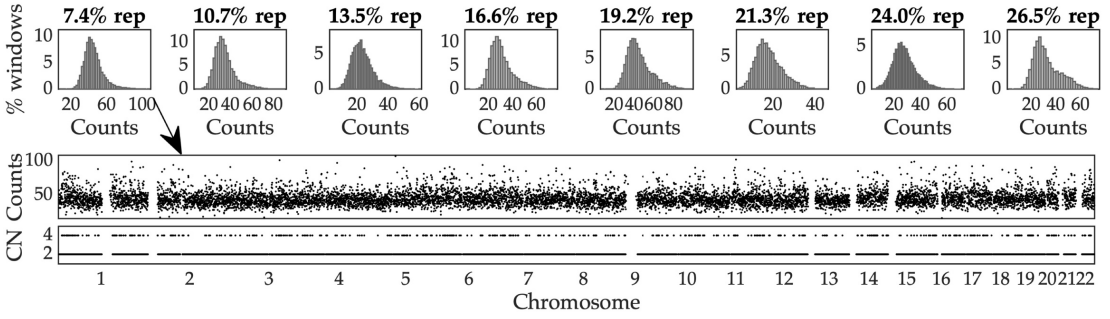
Post hoc in silico cell sorting from single-cell sequencing has two major benefits over sequencing single cells isolated from multiple flow cytometry-sorted populations. First, sequencing biases (particularly, GC-content bias[66]) are known to vary between sequencing libraries, a concern alleviated by using control cells from within the same library as the cells of interest. Second, this approach minimizes experimental manipulations, does not require DNA staining, and reduces inter-experimental variation, for instance, in defining FACS gates. However, other strategies may be more cost-effective: typical mammalian cell cultures contain up to ~30% of cells in S phase at any given time.

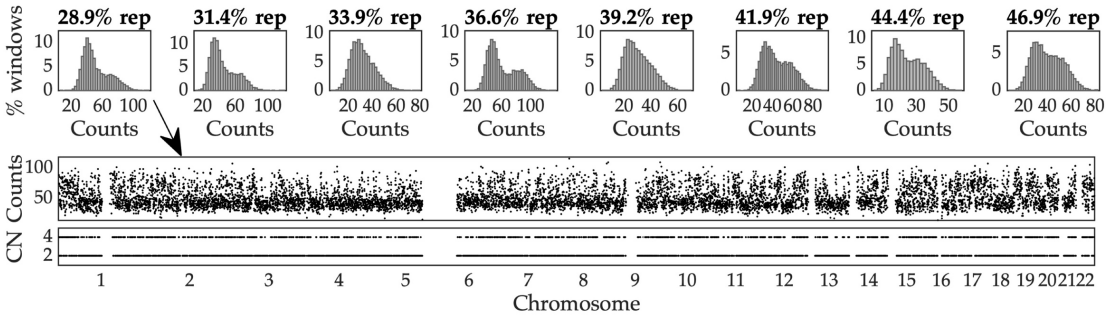
Using a conservatively-defined subset of non-replicating cells identified by this "*in silico* cell sorting" approach, we defined variable-size, uniform-coverage genomic windows that accounted for the effects of mappability and GC-content biases, as well copy number variations, on sequencing read depth[45]. (We note that *in silico* sorting cannot distinguish G_2 cells from G_1 cells because, in principle, both have a uniform copy number genome-wide. We will therefore refer to these cells as "G_1/G_2 cells" throughout.) We counted the number of sequencing reads in each window for each cell (**Figure 4.3**), and

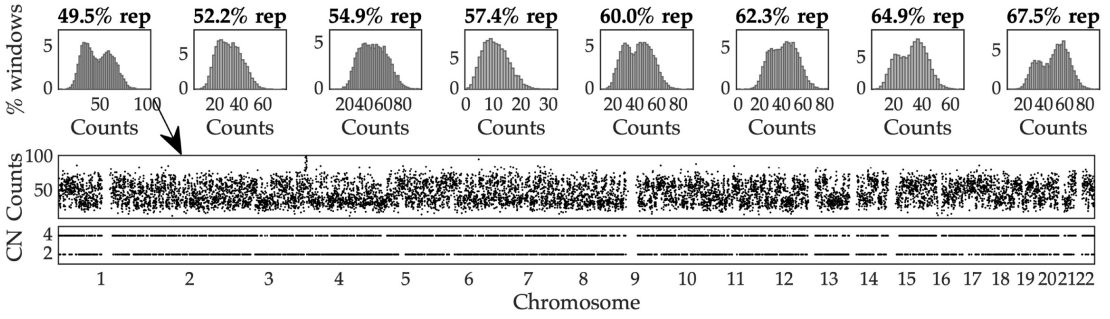
then used a two-state hidden Markov model (HMM) to infer whether each window contained replicated or unreplicated DNA (**Methods**). This confirmed the uniform DNA copy number across the genome in G_1/G_2 cells, and fluctuating regions of replicated and unreplicated DNA in S-phase cells (**Figure 4.1c**). We further validated the HMM by simulation, estimating that, on average, 96.4% of 20kb windows were called accurately in each cell (**Figure 4.4**).

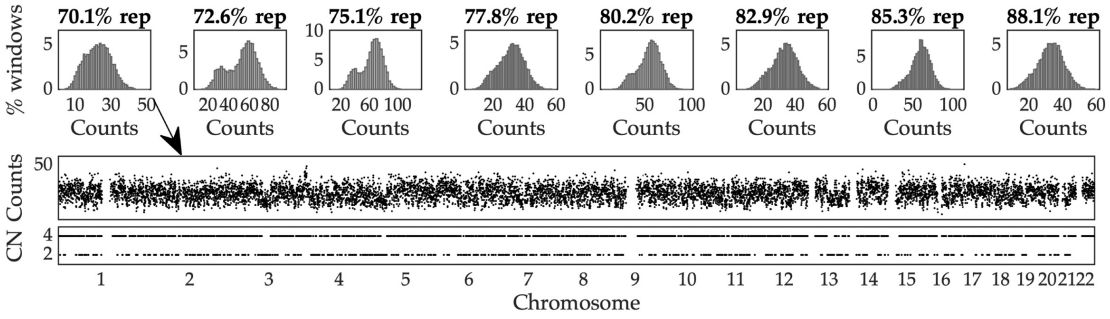
Figure 4.3. **Distribution of single-cell read counts in 200kb windows depends on S-phase progression.** Forty single cells from the S-phase FACS fraction library are shown. Read counts are distributed around a single mode in cells in G_1/G_2 (*top row*). The bimodal distribution of read counts in replicating cells is most apparent in mid S phase. For the left-most cell in each row, read counts and corresponding copy-number inferences (CN) are displayed for the whole genome. Representative examples were selected to reflect patterns across S-phase progression.


G1/G2
% windows
Counts
CN Counts
Chromosome


7.4% rep
10.7% rep
13.5% rep
16.6% rep
19.2% rep
21.3% rep
24.0% rep
26.5% rep
% windows
Counts
CN Counts
Chromosome


28.9% rep
31.4% rep
33.9% rep
36.6% rep
39.2% rep
41.9% rep
44.4% rep
46.9% rep
% windows
Counts
CN Counts
Chromosome


49.5% rep
52.2% rep
54.9% rep
57.4% rep
60.0% rep
62.3% rep
64.9% rep
67.5% rep
% windows
Counts
CN Counts
Chromosome


70.1% rep
72.6% rep
75.1% rep
77.8% rep
80.2% rep
82.9% rep
85.3% rep
88.1% rep
% windows
Counts
CN Counts
Chromosome

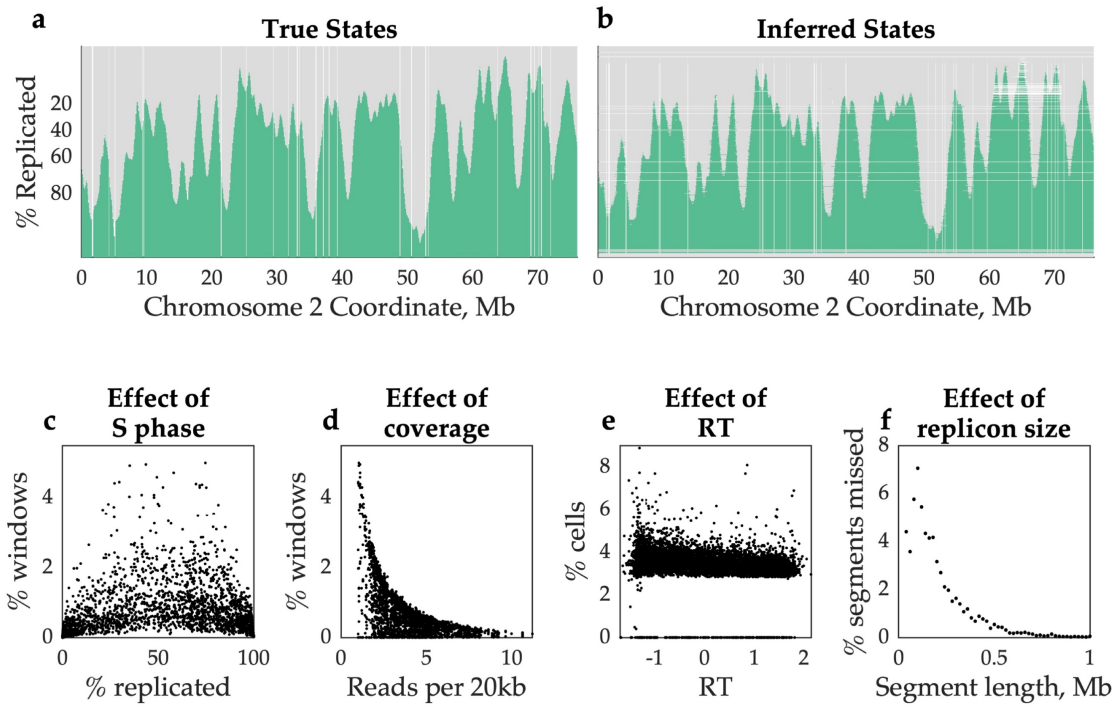


Figure 4.4. **Assessment of replication state inference by hidden Markov model (HMM) with simulated data. a** 2,500 cells were simulated with known "true" replication states at each genomic window, distributed across S phase. Each row represents one simulated cell. **b** Raw read counts were drawn from a Poisson process for each "true" replication state, with rate coefficients chosen to match the coverage of the GM12878 unsorted library. These raw counts were then processed with the same pipeline as the observed data. Each row represents the corresponding cell in **a**. Incorrect copy-number inferences were most frequent in mid-S-phase cells (**c**) and in lower-coverage cells (**d**), while the effects of replication timing were mild (**e**). **f** Copy-number inference allowed identification of 98.6% of "true" replicons. Replicons comprised of a single 20kb window were frequently missed (26.8% not successfully detected). Replicons comprised of at least two windows were detected in > 90% of cases.

Our proof-of-principle FACS experiment also revealed cross-contamination between fractions: *in silico* sorting labeled ~24.3% of cells in the G_1-phase FACS fraction as "S" and reciprocally ~25.1% of cells in the S-phase fractions as "G_1/G_2" (**Figure 4.2a**). However, because the objective of *in silico* sorting is to identify high-confidence G_1/G_2 cells to use as controls, it is designed to be conservative in labeling cells as "G_1/G_2". We therefore suspected that this estimate of S-phase cells in the G_1 FACS fraction was

inflated. Indeed, after HMM processing, 260/323 cells in the G_1 fraction initially called as "S" were re-assigned as G_1/G_2. The remaining 63 cells (4.7% of the G_1 fraction) were confirmed S-phase cells, displaying copy-number profiles consistent with early S phase and indistinguishable from early-S-phase cells in the S-phase fraction (**Figure 4.2f**). We further analyzed a published dataset of 1,475 cells from the human melanoma cell line COLO-829, for which FACS was used to isolate exclusively G_1 cells[160]. In this dataset, *in silico* sorting labeled 233 cells (15.8%) as "S", of which 32 cells were confirmed to be in S phase (cross-contamination: 2.2%; **Figure 4.5**). Thus, *in silico* sorting using single-cell sequence DNA is a viable strategy for identifying control (G_1/G_2) cells within an unsorted library, and when used in combination with the HMM, provide greater sensitivity to FACS.

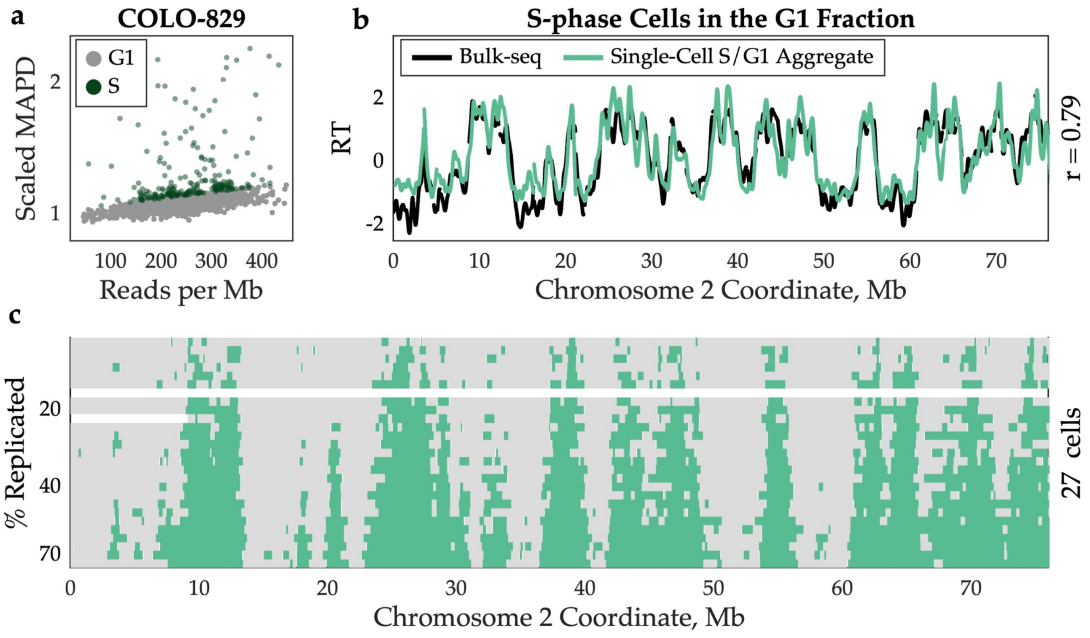


Figure 4.5. **S-phase contamination was observed in published G_1 FACS data**. **a** *In silico* sorting of 1,475 cells from the human melanoma cell line COLO-829 identified 233 potentially replicating cells in an experiment where FACS was used to isolated only G_1 cells for copy-number aberration analysis[160]. **b** The aggregate S/G_1 replication timing profile inferred using *in silico* sorting assignments recapitulated the ensemble replication timing profile measured in

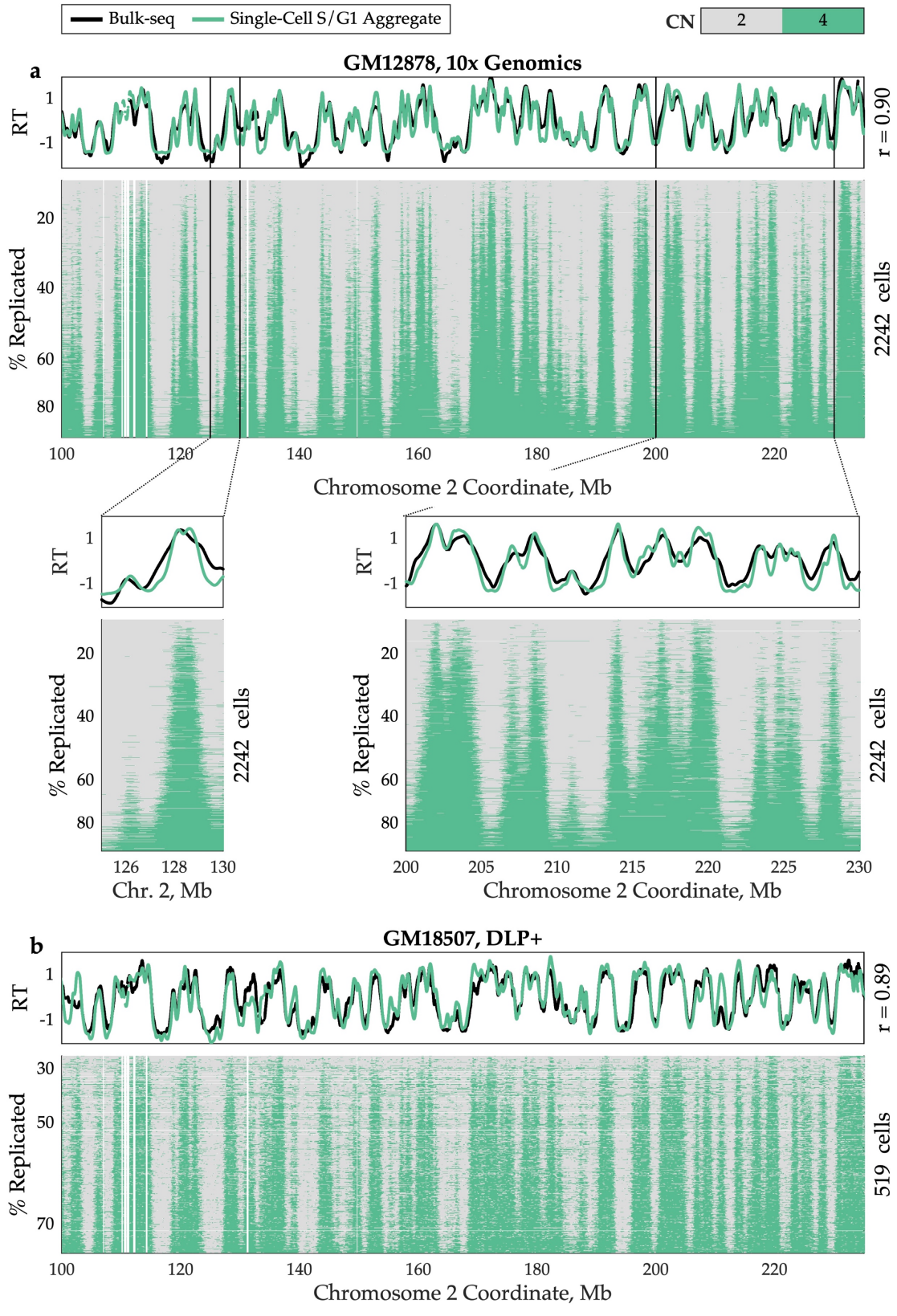
COLO-829. **c** Replication profiles for 32 cells were consistent with being in S phase and demonstrated a uniform replication progression pattern. Five cells are not shown due to chromosome-specific copy number anomalies.

A discrete benefit of single-cell data is the ability to aggregate similar cells together, effectively increasing the coverage without masking important heterogeneity between subsets of cells. Because the partitioning happens *in silico*, we can consider many different single-cell aggregates of the same data, from a single fraction (spanning all of S phase) down to single cells (wherein each cell is its own fraction). We generated several such aggregates, partitioning cells based on their progression through S phase (% of genome replicated) and summing per-window read counts across cells (**Figure 4.1d**). Validating this approach, the single fraction profile – analogous to an ensemble S/G_1 replication timing profile[45] – was highly correlated to a bulk replication timing profile for the same cell line ($r = 0.90$). Partitioning cells into 10 fractions, a structure emerged similar that seen in high-resolution Repli-seq[62]: triangular pileups of high read depth (corresponding to active replication) around peaks observed in bulk sequencing. Many of these regions of high read depth were evident in every fraction, although some (*e.g.*, **Figure 4.1d**, *~13.8Mb*) first appear later in S phase. This same structure was observed – at higher resolution – when cells were partitioned into 100 fractions. Thus, by this approach, we can capture sub-S-phase events across all of S phase without the risk of FACS cross-contamination, at a resolution for which FACS is infeasible (*i.e.*, 100 fractions), and with the ability to examine the same population of cells at multiple levels of resolution.

The logical extension of this partitioning approach is to consider each cell as comprising its own fraction. After filtering out cells that were not replicating or for which a two-fold relationship was not observed between

copy-number states, we analyzed 2,437 single GM12878 cells. At this single-cell resolution, we observed consistent pileups of distinct replicated and unreplicated segments across cells (**Figure 4.6a**). These pileups were in the same regions observed as peaks in the bulk-sequencing profile and sub-S-phase fractions, underscoring that these regions correspond to locations of active replication progression, centered at one or more replication origins. Even at single-cell resolution, these pileups were triangular (**Figure 4.6a**, *insets*), consistent with symmetric bidirectional replication fork progression from a common origin locus (or a tight cluster of replication origins), and appeared visually to be highly localized. Thus, we demonstrate the ability to measure single-cell replication timing in thousands of single cells, in an unbiased manner, and without the need for FACS. This represents roughly ten times more cells than have been reported in previous single-cell replication timing analyses, which have focused primarily on mid S phase cells[108,109].

Figure 4.6. **Single-cell replication state data, generated by multiple library preparation protocols. a** Single-cell replication profiles for 2,242 GM12878 cells (including both sorted and unsorted cells), following single-cell isolation and library reparation with the 10x Genomics Single-Cell CNV Solution. Consistency of the replication program is observed across cells at chromosome-scale and at the level of individual peaks (*inset*). Pileups reflect sharply defined and consistently replication regions, which overlap peaks in the bulk replication timing profile. Variation in activation time during S phase among initiation sites is also observed to mirror the replication timing profile. Each row represents a single cell, sorted by the percent of the genome replicated, and each column represents a fixed-size window of 20kb. 195 cells are not shown due to copy-number aberrations on this chromosome. Low-mappability regions and cell-specific copy-number alterations have been removed (*white*). Insets show smaller regions. **b** Single-cell replication profiles for 519 GM18507 cells, following amplification-free direct DNA transposition single-cell library preparation (DLP+). Due to noise, only 480-614 of the 759 S-phase cells were analyzed for any given chromosome. Raw data are from [114].


Bulk-seq
Single-Cell S/G1 Aggregate
CN
2
4
a
GM12878, 10x Genomics
RT
r = 0.90
% Replicated
2242 cells
Chromosome 2 Coordinate, Mb
Chr. 2, Mb
b
GM18507, DLP+
r = 0.89
519 cells

We repeated this analysis using single-cell data for 3,040 cells from the LCL GM18507, prepared using DLP+[114]. We identified 759 replicating cells within this dataset, and again observed pileups in consistent genomic regions, close to peaks in the S/G1 aggregate replication timing profile (**Figure 4.6b**). This dataset enabled us to benchmark our analysis strategy in the absence of amplification bias, ensuring that the observed single-cell pileups were not a persistent technical artifact of the 10x Genomics amplification method and validating the ability to accurately profile single-cell replication timing in hundreds to thousands of cells across multiple single-cell sequencing technologies.

Sites of replication initiation are consistent in single cells

The nature of DNA replication initiation events is among the most debated aspects of mammalian DNA replication, both regarding its spatial scale (specific loci[81,161-163], localized regions[164-166] or broad domains[48,158]) and the degree of spatial and temporal stochasticity across cells[88,89,157]. Our comprehensive single-cell DNA replication data enables us to rigorously address these subjects.

We focused first on the spatial dimension of variability among cells. As noted above, visual inspection of replicated region pileups revealed very little variation across single cells (**Figure 4.6**; **Figure 4.7a**). To analyze this axis of variation systematically, we began by identifying replicated segments in each single cell. Each replicated segment, which we termed a "track" (by analogy to single-molecule DNA combing tracks), represents the activity of at least one replication origin. Theoretically, if a replication track corresponds to a single replicon, initiating from one origin and expanded by symmetric progression

of sister replication forks, the origin of replication should be located at the center of that replication track. Thus, as a first approximation of origin locations, we assigned the center of each replication track as the most likely location of replication initiation for that track. (We excluded tracks longer than 1Mb in this initial analysis to reduce the likelihood of including tracks that reflected the activity of multiple independent origins that have converged.)

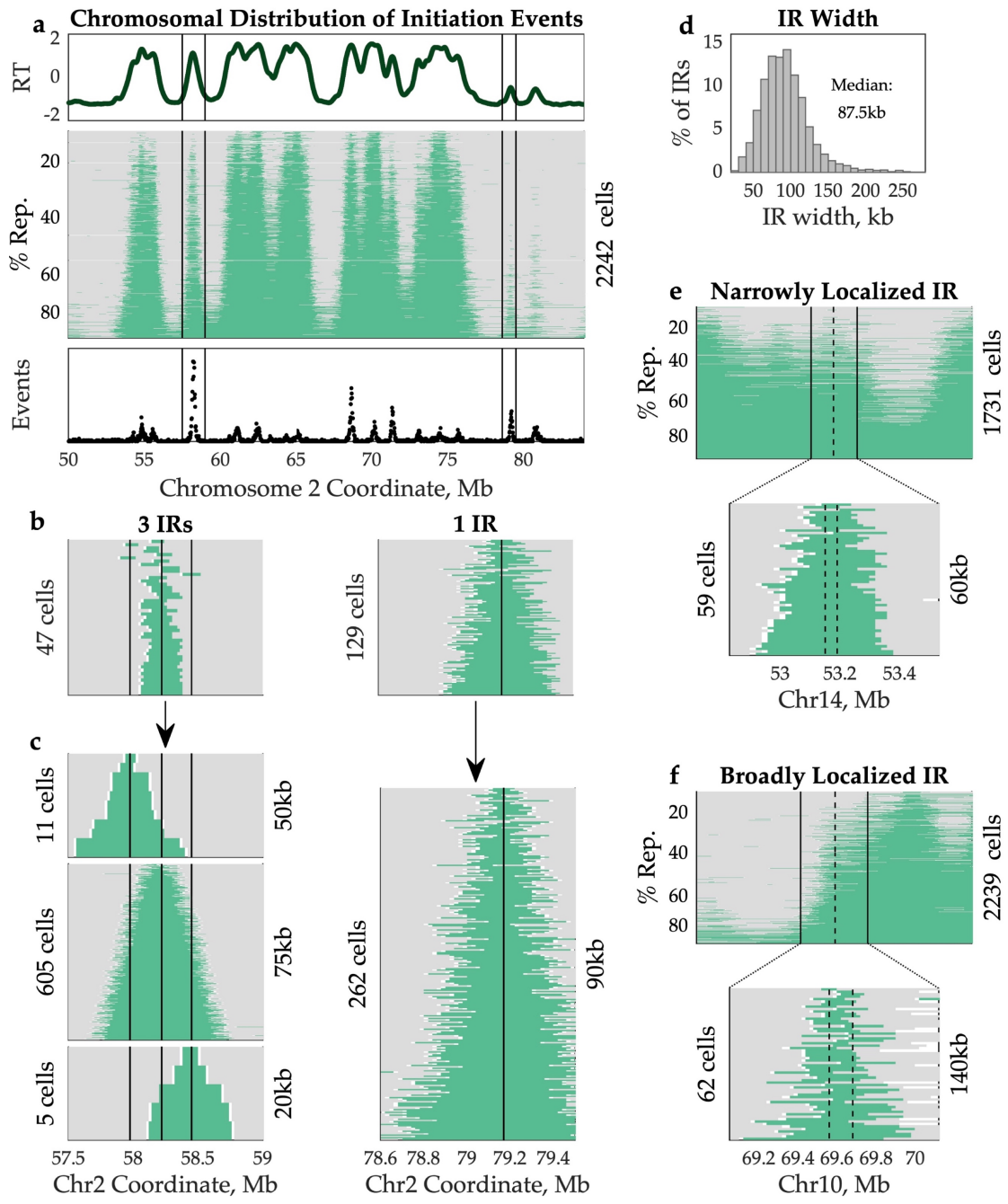


Figure 4.7. **Consistency of single-cell replication initiation sites. a** Peaks in the aggregate replication timing profile inferred from all GM12878 S-phase cells (*top*) correspond to segments that are consistently replicated across single cells (*middle*). These peaks also correspond to regions of dense initiation site calls (*bottom*). The indicated regions correspond to the full width of the insets in **b**. **b** Replicated regions in single cells are centered at consistent locations, termed initiation regions (IRs), overlapping peaks in the aggregate replication timing profile. For each IR (*black line*), a subset of cells was identified containing a replicated track (*green*) overlapping the IR center but not extending into either neighboring IR.

Some aggregate peaks corresponded to multiple IRs. **c** Assignment of replicated tracks < 1Mb to the nearest IR revealed a triangle around each IR center, consistent with symmetric replication fork progression. In contrast to **b**, some replication tracks centered at the indicated IR extend into a neighboring IR (likely reflecting passive replication). This larger set of replication tracks was used to determine the location of the IR: for each track, the center position was assigned as the location of replication initiation in that cell, and the IR was defined as the region between the 25th and 75th percentile of the range of initiation sites across cells. Black lines indicate the center (50th percentile) of the IR. The width of each IR is displayed on the right y-axis. **d** The location of each IR was identified at kilobase scale (median width: 87.5kb). IRs supported by < 5 replication tracks were excluded to avoid skewing the distribution to the left. **e** 78.9% of IRs could be localized to a region 100kb or narrower. In the example shown, 59 replication tracks overlapped the IR. The 25th to 75th percentile of midpoint locations for these tracks fell within a 60kb range (*dotted lines*). **f** Broad IRs may reflect the presence of multiple distinct initiation events that were not disambiguated, technical noise, or mild asymmetry in replication fork progression. In the example shown, 62 cells overlapped the IR. Visually there appear to be multiple distinct clusters of track midpoints. See **Figure 4.8**.

Consistent with previous work suggesting that replication initiation potential is diffuse throughout the genome[96], we found that 49.7% of mappable 20kb genomic windows were called as a probable initiation site in at least one cell. However, these probable initiation sites were not uniformly distributed across the genome. Rather, highly frequent initiation sites were neighbored by gradually less frequent initiation sites, creating peaks around these local maxima (**Figure 4.7a**). This structure suggests that a more limited group of genomic loci might give rise to replication initiation, as ambiguity in identifying the boundaries of replication tracks would result in slight shifts in the probable initiation site from the true midpoint to a neighboring locus and the observed gradual decrease in initiation frequency with increasing distance from that true midpoint.

Based on the conclusion that noise in individual cells was likely contributing substantially to variation in initiation site location, we devised an approach to cluster these sites into larger initiation regions (IRs) shared across cells, which did not rely on a 1Mb length cutoff to determine which replication

tracks were informative about individual origins and which represented the activity of multiple independent origins. Instead, replication tracks that overlapped multiple shorter replication tracks were treated as agnostic to IR location because they could plausibly be explained by firing of multiple of the overlapped origins or of a single central origin. These uninformative tracks were thus excluded from use in clustering. By this process, we identified a total of 7,522 IRs.

As noted above, single-cell pileups corresponded visually to peaks in the S/G_1 aggregate replication timing profile (**Figure 4.6**; **Figure 4.7a**). Indeed, 90.9% of peaks in the aggregate profile coincided with an IR. Of these aggregate peaks that overlapped an IR, 48.7% corresponded to multiple IRs (*e.g.*, **Figure 4.7b**, *left*), while the remaining 51.3% corresponded to a single IR (*e.g.*, **Figure 4.7b**, *right*). This suggests that origins are often clustered in hotspots along the chromosome; the replication timing peaks corresponding to single IRs could either be regions of lower origin density or, conversely, represent hotspots too dense for individual origins to be detected at this resolution. Thus, single-cell data are concordant with the ensemble replication timing profile, but also caution that smoothing of ensemble profiles likely removes information about distinct initiation sites.

We then assigned all replication tracks shorter than 1Mb to the IR whose center was closest to the midpoint of the track. This includes tracks that potentially overlap multiple fired IRs; however, when all replication tracks assigned to a given IR were sorted by length, a symmetric triangle was observed around the IR center (**Figure 4.7c**), consistent with sister replication forks progressing away from a single origin or tight cluster of origins at the IR center with similar processivity. For each IR, we calculated how tightly the midpoints of these replication tracks were clustered to assess how precisely

the most probable initiation site within the IR was identified. IRs were localized to a median width of 87.5kb (~4 windows; **Figure 4.7d**), which corresponds to an inter-IR distance of 50kb to 920kb (median: 260kb). Most IRs (78.9%) were 100kb or narrower (*e.g.*, **Figure 4.7e**). Visual inspection of broad IRs (> 120kb) suggested that many contain multiple initiation events that were grouped together because of overlap between replication tracks (**Figure 4.7f**; **Figure 4.8**). Thus, while we cannot determine whether IR width (and variability in IR width) reflects technical noise, inconsistency between cells in the precise location of initiation, or mild asymmetry in sister replication fork progression, we conclude that initiation events are relatively localized, and that at least some of IR widths are likely overestimated. Localized initiation regions are also apparent in the early-S fractions of the 10- and 100-fraction profiles (**Figure 4.1d**), where the impacts of noise are averaged across many cells.

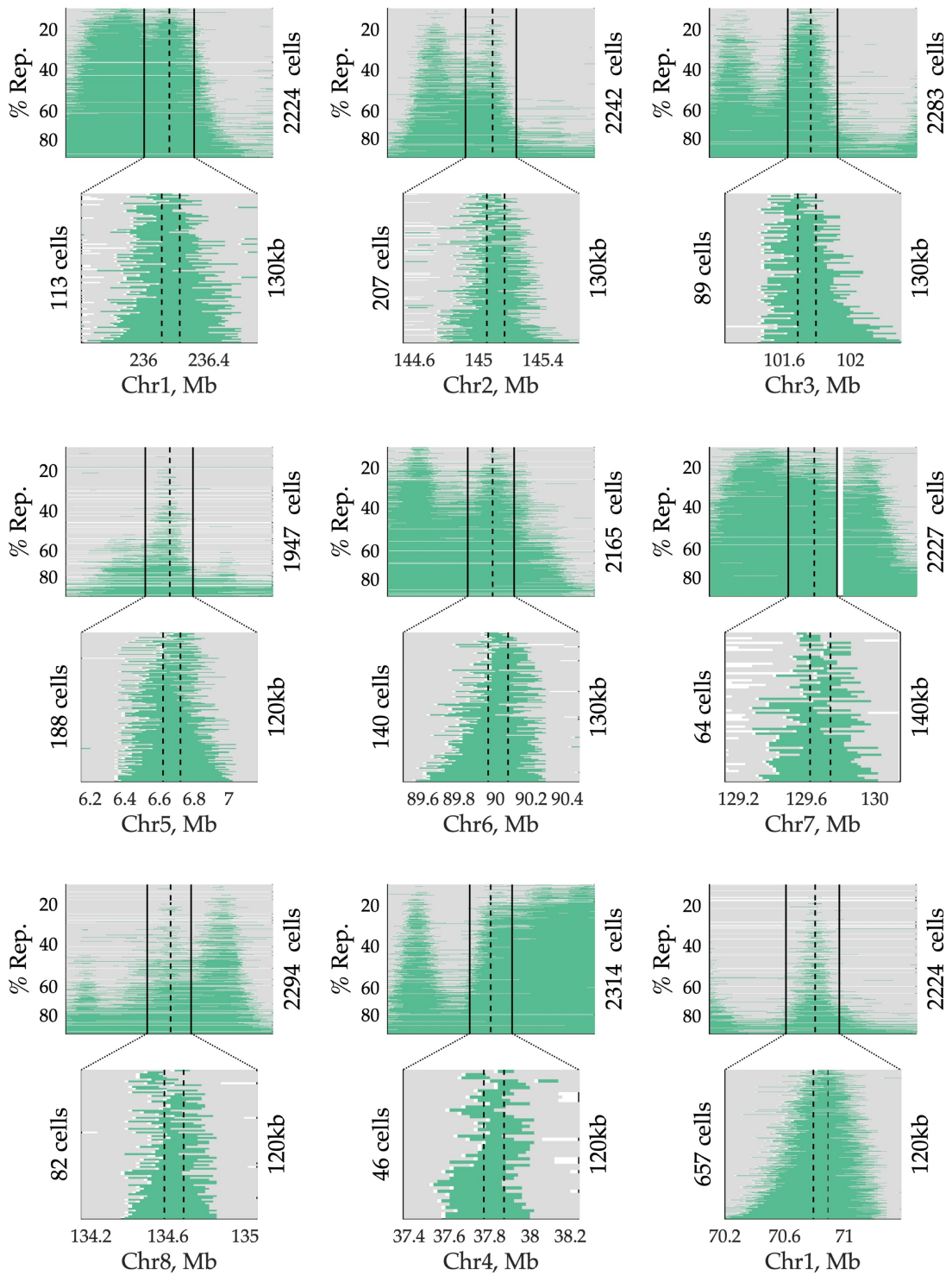


Figure 4.8. **Broad initiation regions (*i.e.*, those wider than 120kb) often appear visually to contain multiple non-overlapping replication tracks.** Eight representative examples are shown, in which a single IR was called, but visual inspection suggests multiple discrete

initiation sites that have not been disambiguated, supporting the notion that these IR widths are overestimated. One counterexample is shown in the bottom right (Chr1, ~70.8Mb) – this IR's width may be due to it encompassing a cluster of origins in close proximity or to replication fork asymmetry.

In our analysis of IRs, we did find evidence of ectopic replication initiation: only 29.2% of IRs contained a peak in the S/G_1 aggregate profile, and 31.5% of IRs were supported by a single replication track. However, these potentially ectopic events comprised a small fraction of all observed initiation events. Rather, 2,595 IRs (34.5%) accounted for 90% of all replication tracks, indicating that about a third of the IRs are used consistently across cells. Thus, contrary to previous studies that analyzed single-cell replication profiles at the level of large chromosomal "domains"[108,109], our data reveal localized initiation regions, which we assume correspond to individual, or tight clusters of, replication origins.

Consistent yet non-deterministic order of replication initiation

Given that single cells appear to initiate replication primarily from a consistent set of genomic locations, we turned our focus to the temporal axis of variation: how consistent is the order in which single cells initiate replication at these loci?

We first asked whether the single-cell data were compatible with strictly determined replication timing, such that every cell initiates replication at every IR in the same order. Strict determinism provides a straightforward prediction to test: the number of IRs replicated in any given cell should predict which IRs have been replicated in that cell. For example, a cell that has replicated one IR is predicted to have replicated the IR with the earliest replication timing; a cell that has replicated 100 IRs is predicted to have

replicated the 100 IRs with earliest replication timing; and so on. To test how well these predictions matched our data, we counted the number of IRs that were replicated in each cell and used that to assign each IR in that cell an "expected" state – either unreplicated or replicated – assuming that the firing order was fixed (**Figure 4.9a, b**). For a given IR, the observed replication state matched the predicted state in the vast majority of cells (**Figure 4.9c**), indicating that the firing of IRs in single cells follows a highly predictable order. However, we did observe that, on average, an IR differed from its expected state in 11.1% of cells (**Figure 4.9d**). Thus, we can formally rule out the hypothesis that replication timing is strictly determined; IR firing order at the single-cell level is orderly but not entirely predictable.

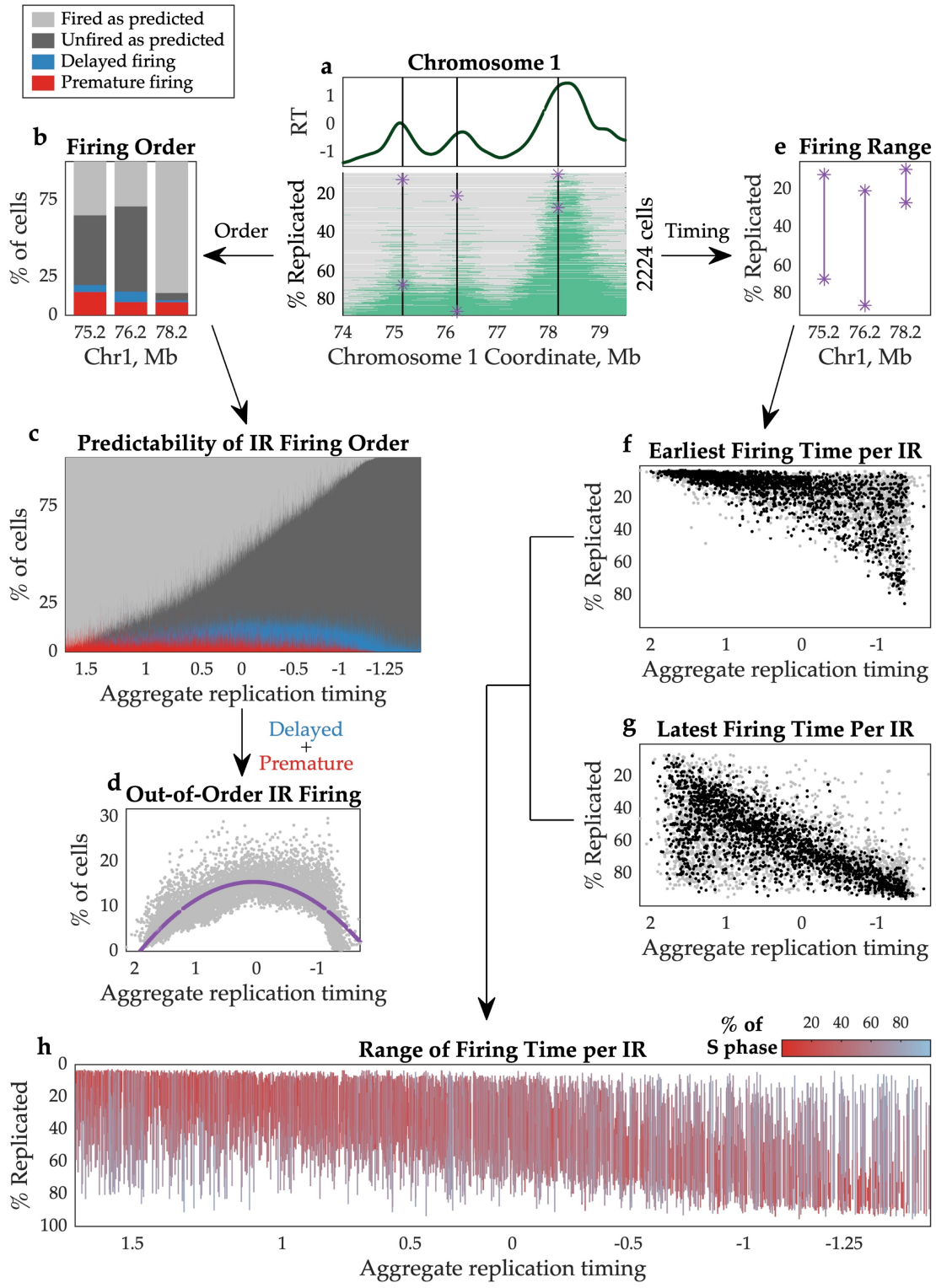


Figure 4.9. **Variation in the order and timing of replication initiation in single cells across S phase. a** An example region illustrates analyses of both IR firing order (**b-d**) and IR firing time (**e-h**). Black lines indicate the four IRs in **b** and **e**. Purple asterisks indicate the earliest cell in

which an IR was observed to fire and the latest cell in which it was observed to be unfired. **b** IRs differ in their degree of consistency across cells. IRs were ranked from earliest to latest, allowing prediction of which would have fired in each cell under a strict firing ordering. Cells that have replicated an IR not predicted to fire in that cell are considered "premature" (*red*), while those that have not replicated an IR predicted to have fired already are "delayed" (*blue*). **c** IRs are fired in the expected order in most cells. Each column represents an IR, from earliest (*left*) to latest (*right*). **d** IR firing varies most for IRs expected to fire in mid S phase. On average, IRs behaved differently than expected in 11.1% of cells (range: 0.3-29.7%). Each dot represents one IR. *Purple line*: second-order polynomial fit. **e** The range of IR firing spanned from the earliest observed fired cell to the latest observed unfired cell. S-phase progression was measured as the percent of the genome replicated. **f, g** For each IR, we identified the least replicated (*i.e.*, earliest) and most replicated (*i.e.*, latest) cell containing a replication track assigned to that IR. A second cell within 10% S-phase progression was used to "corroborate" the earliest and latest cell (*black dots*); all other dots are gray. **h** IRs with earlier aggregate replication timing tended to have narrower ranges of firing times than those with late aggregate replication timing. Each vertical line represents the range for one IR, color-coded by the % of S phase during which that IR fires (*i.e.*, the length of the line). A small number of constitutively late IRs (short red lines with late aggregate replication timing) can be observed. Only IRs whose earliest and latest values were corroborated by a second cell are shown.

Having observed variation across cells, we next asked if that variation was uniform across S phase or concentrated at specific times during S phase. We found a parabolic relationship between replication timing of an IR and the proportion of cells that fired that IR out of the strictly determined order. Thus, variability was lowest at the beginning and end of S phase and highest in the middle of S phase, such that 83.4% of the above-average variability occurred in the 53.7% of IRs with aggregate replication timing between 1 and -1. A similar parabolic trend was previously described by Takahashi *et al*.[109] and was robust in our larger sample size.

We next considered the extent of firing time variability, asking when in S phase IRs fire in the instances that they fire out of the predicted order. To answer this question, we identified the least-replicated (*i.e.*, earliest) cell in which an IR was observed to fire and the most-replicated (*i.e.*, latest) cell in which it had yet to fire (**Figure 4.9e**). We found that there was an association between the earliest time that an IR fired and its replication timing in the S/G_1

aggregate replication profile (r = -0.64; **Figure 4.9f**), indicating that IRs with late aggregate timing (hereafter, "ensemble-late IRs") tended to start replicating later in S phase than those with early aggregate timing ("ensemble-early IRs"). However, most IRs were observed to have fired in a subset of early-S-phase cells: 49% of GM12878 IRs fired at least once in a cell with < 10% of its genome replicated, 83% in a cell with < 25% replicated, and 96% in a cell < 50% replicated. Thus, many ensemble-late IRs were not restricted to firing in late S phase. There was also an association between how late into S phase an IR remained unfired and its aggregate replication timing (r = -0.66; **Figure 4.9g**). Thus, ensemble-early IRs tended to finish firing across all cells relatively early in S phase, while ensemble-late IRs tended to remain unfired into late S phase.

After determining these earliest and latest cells for each IR, we considered them in a paired manner to determine the range of firing times of each IR (**Figure 4.9h**). Given that range is sensitive to outliers (*i.e.*, a duplication called "replicated" or a deletion called as "unreplicated"), we focused on IRs for which the minimum and maximum values were "corroborated" by a second cell within 10% of S phase from the extreme. Ensemble-early IRs tended to first fire in early S phase and to complete their replication before the genome was 50% replicated. In contrast, ensemble-late IRs tended to also first fire in early S phase, but to remain unfired in some cells until the end of S phase. Therefore, the firing time of ensemble-early IRs was constrained to early S phase, while ensemble-late IRs appeared to be less constrained. However, we did observe a small number of IRs that fired exclusively in late S phase across cells; these had a more constrained range. We thus proceeded to further analyze these different behaviors in regions with late aggregate replication timing.

Ensemble-late initiation regions comprise multiple subtypes

Our analysis of single-cell replication timing indicated that IRs are fired in a consistent order across most cells, but that ensemble-late IRs fire across a larger portion of S phase relative to ensemble-early IRs (**Figure 4.9h**). We further dissected the nature of these IRs with large firing ranges to better understand whether we were capturing rare occasions of extremely premature firing or perhaps observing a capacity of IRs to fire throughout S phase. In other words: do these IRs fire substantially ahead of schedule in some cells, or do they not have a scheduled time to fire at all?

We found that each of these two explanations for large range of firing times were supported by a substantial fraction of IRs, and that neither behavior was sufficient to explain all cases on its own (**Figure 4.10a**). This indicates that some ensemble-late IRs tend to fire late but sometimes fire very early, while others fire at many different times in S phase. Specifically, 12% of IRs (27.3% of ensemble-late IRs) fired inconsistently throughout S phase (**Figure 4.10b, e**), with earlier aggregate timing corresponding to more cells firing the IR (compare **Figure 4.10b**, *top left vs. top right*). On the other hand, 27% of IRs (63.7% of ensemble-late IRs) fired rarely and almost all the replication tracks associated with these IRs were from cells < 50% replicated (**Figure 4.10c, e**). Finally, 4% of IRs (9.0% of ensemble-late IRs) were never observed to fire in a cell < 50% replicated (**Figure 4.10d, e**). Comparing these three classes, IRs that fired throughout S phase tended to have the earliest aggregate replication timing (median: -0.32 and as early as 0.54), while constitutively late IRs had the latest aggregate timing (median: -1.22; **Figure 4.10f**). These unexpected results demonstrate that the late-replicating regions observed in ensemble assays contain origins with heterogeneous firing

behavior; these results cannot be fully explained by either a deterministic timing model (which posits these regions contain constitutively late-firing origins) or existing stochastic firing models (which posit that these regions contain low-efficiency origins that become increasingly likely to fire as S phase progresses[88]).

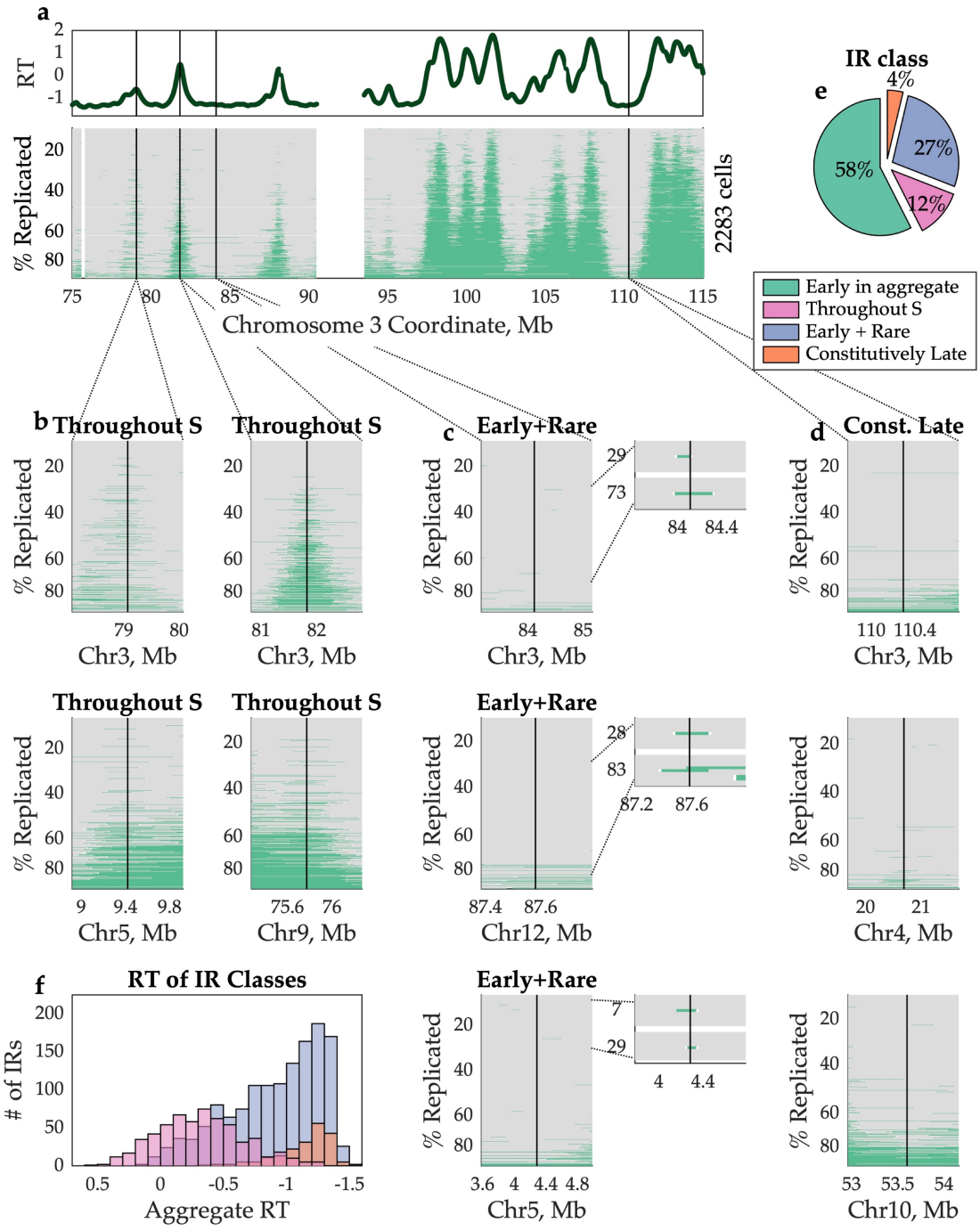


Figure 4.10. **Three distinct classes of IRs with late aggregate replication timing. a-d** "Late" IRs can be classified into three classes based on their behavior across single cells: some fire throughout S phase (**b**), some fire rarely but often fire early when they do fire (**c**), and some were never observed to fire early **(d)**. The IRs indicated with black lines in **a** are shown in the top row of **b**, **c**, and **d**. Additional examples are shown below. **e** 27% of IRs with late aggregate replication timing fire infrequently but with a preference for early S phase, while 12% fire throughout S phase. Constitutive late firing is rare (4% of IRs). **f** IRs that fire throughout S

phase (*pink*) tend to have earlier replication timing than the other two classes of IRs, while those that were constitutively late (*orange*) had the latest average replication timing.

Single-cell replication timing across cell lines throughout S phase

Having established a workflow for high-throughput replication analysis of unsorted cells, we performed whole-genome sequencing of 9,658 single cells across eight additional cells lines: two LCLs, three embryonic stem cell lines (ESCs), and three cancer cell lines. As with GM12878, we performed *in silico* cell sorting to distinguish replicating and non-replicating cells within each library (**Figure 4.11**). For each cell line, we generated an aggregate S/G_1 profile that was highly correlated to an S/G_1 bulk replication timing profile for the same cell line (r = 0.84-0.97; **Figure 4.12**). We then generated replication profiles for between 110 and 501 S-phase cells across the different cell lines (**Figure 4.12**; **Figure 4.13a, b**).

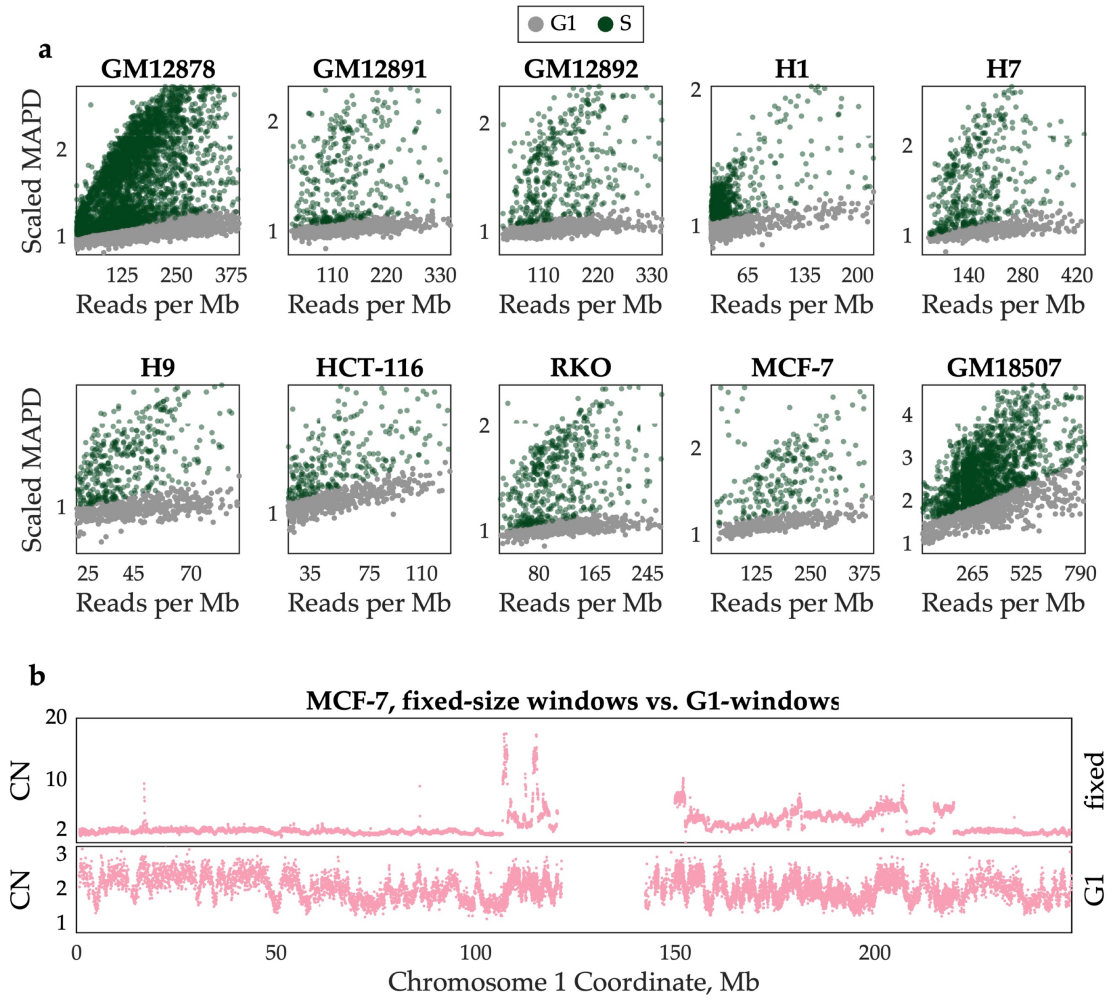


Figure 4.11. **Metrics of *in silico* cell sorting for all cell lines under study. a** *In silico* sorting identifies two distinct populations of cells based on the relationship between mean reads per Mb and scaled MAPD, corresponding to G_1/G_2- and S-phase cells. GM18507 cells were sequenced after DLP+ library preparation[114]. **b** *In silico* sorting is robust in aneuploid cancer cell lines (*e.g.*, MCF-7) and corrects for the effects of copy-number aberrations (CNAs). *Top*: MCF-7 contains numerous CNAs with a wide range of copy-number values. Each dot represents the sum of read counts across all *in silico* assigned S-phase cells in fixed 20kb windows. *Bottom*: Using *in silico* assigned G_1/G_2 cells to define genome windows reveals more subtle copy-number fluctuations corresponding to replication timing. Each dot represents the sum of read counts across all *in silico* assigned S-phase cells in G_1-defined windows.

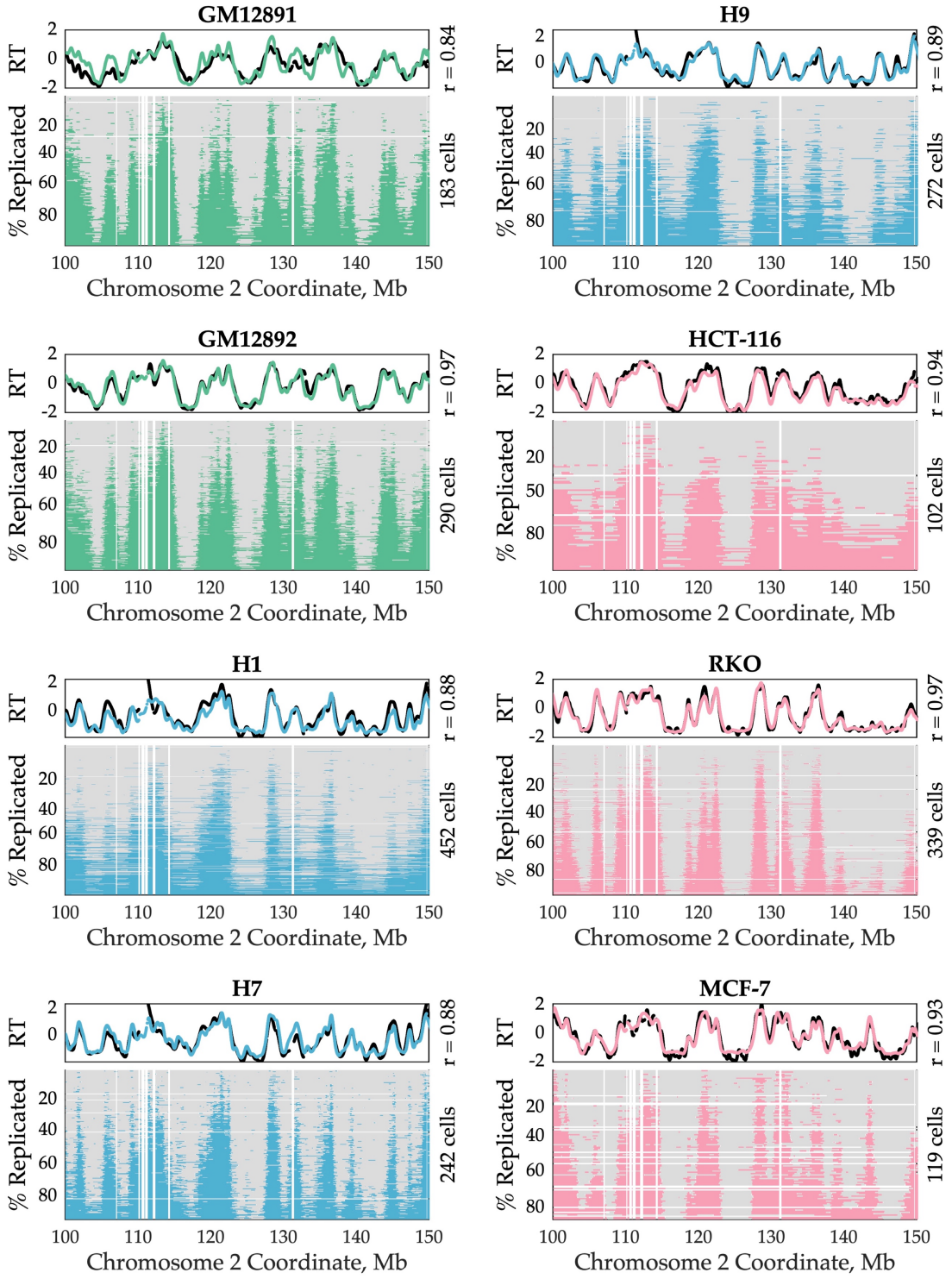


Figure 4.12. **Replication state inference for thousands of single cells across human lymphoblastoid cell (*green*), embryonic stem cell (*blue*), and cancer-derived cell (*pink*) lines.** As in **Figure 4.6a**, for all additional cell lines. *Top*: aggregate S/G1 replication timing profiles from *in silico* sorting of single cells recapitulates the bulk-sequencing replication timing profile

for the same cell line, in all cell lines. *Bottom*: single-cell replication state in fixed 20kb windows. Each row represents a single cell.

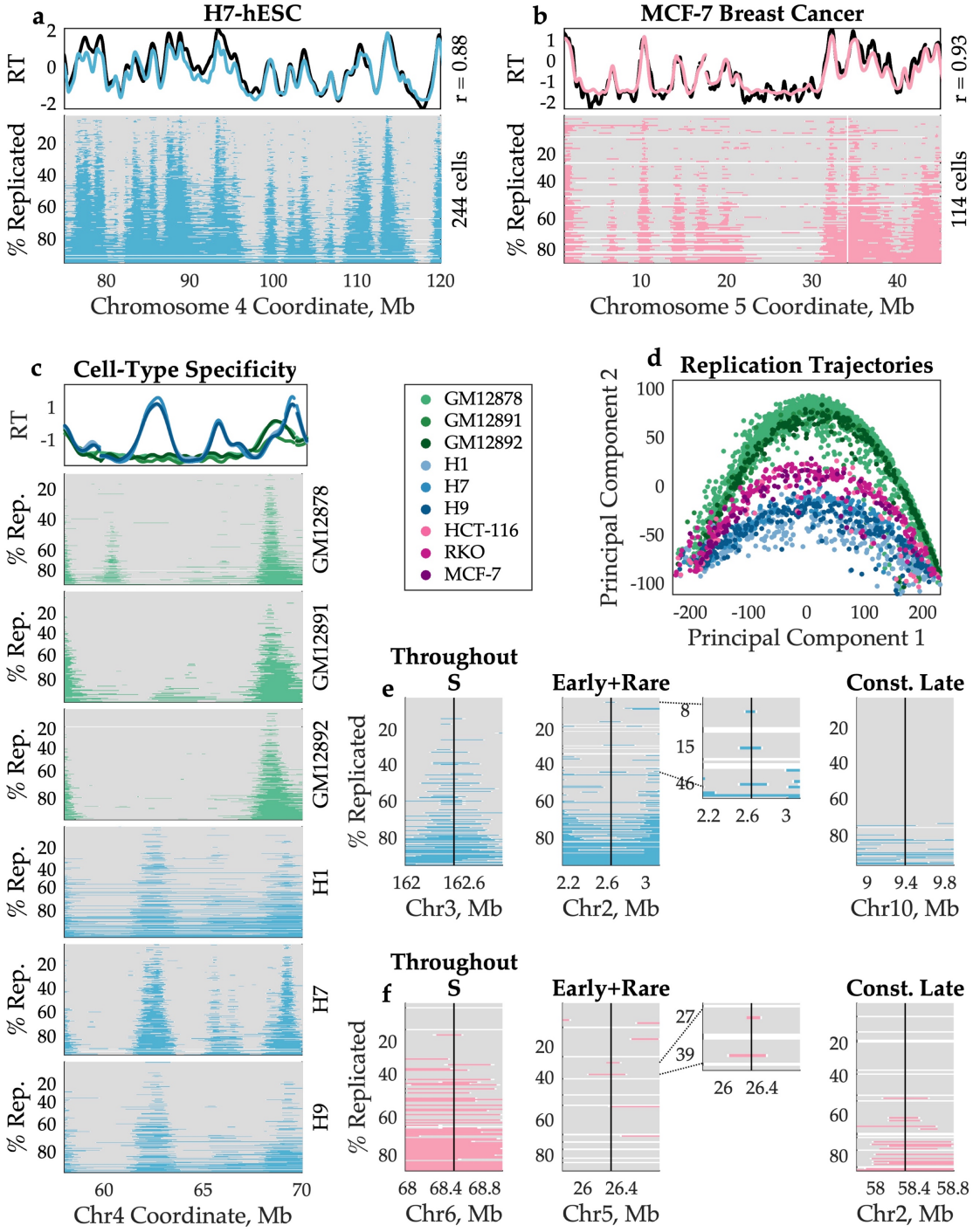


Figure 4.13. **Comprehensive measurement of single-cell replication timing across cell types**. **a, b** As in **Figure 4.6a**, for the embryonic stem cell line H7 (**a**) and for the breast cancer cell line MCF-7 (**b**). **c** Replication timing variation between cell types is observed at the single-cell level.

Top: bulk-sequencing consensus replication timing profiles for LCL (*green*) and hESC (*blue*). *Bottom*: single-cell data reveals that the bulk-sequencing peaks at ~62Mb and ~65.5Mb reflect the presence of hESC-specific initiation sites. **d** Single cells follow cell-type-specific trajectories of S-phase progression, as determined by principal component analysis (PCA). PCA was performed on replication states in all genomic windows across autosomes. PC1 corresponds to the % of the genome replicated (r = 0.99), with negative values of PC1 reflecting early S phase and positive values reflecting late S phase. Cell types segregate along PC2. Each dot represents a single cell. **e, f** All three categories of IRs with late aggregate replication timing described in **Figure 4.10** were also observed in H7 (**e**) and MCF-7 (**f**).

The aneuploid breast cancer cell line MCF-7 highlights the broader applicability of *in silico* sorting. While we apply this method to focusing our analysis only on replicating cells, it is also valuable in single-cell analysis of copy-number aberrations (CNAs) in cancer. In that context, it is necessary to remove replicating cells prior to CNA calling, since both replication and duplications/deletions affect copy number estimation. MAPD has previously been used to filter out "noisy" cells in this type of analysis[160]. However, aneuploidy inflates MAPD values (**Figure 4.11a**, compare MCF-7 to other cell lines), making it difficult to effectively set a threshold for filtering. In contrast, explicit modeling of G_1/G_2 and S cell populations with distinct linear relationships between read coverage and MAPD efficiently discriminates cells of interest (either for replication analysis or CNA analysis; **Figure 4.11b**).

It has been well demonstrated in ensemble experiments that cell types have distinct replication timing programs, which are shared by cell lines of the same cell type[44,47,56]. Thus, we asked whether cell-type differences among these nine cell lines were preserved at the single-cell level. Indeed, cell-type differences among the aggregate replication timing profiles were found to be consistent at the single-cell level (**Figure 4.13c**; **Figure 4.14**). These differences in replication state between cell types were sufficient to cluster single cells by cell line and cell type (**Figure 4.13d**), suggesting that individual cells of the

same cell type follow a similar trajectory through S phase. Two types of replication timing differences can be observed at the ensemble level: differences in peak locations (*i.e.*, in the location of fired origins) and differences in peak amplitude (*i.e.*, in the timing at which a shared origin is fired). We observe both of these classes of variation at the single-cell level: cell-type-specific peaks in the S/G1 aggregate profile that reflect the presence of a cell-type-specific initiation site (*e.g.*, **Figure 4.14b**, *right*) and peaks of different amplitude in the S/G1 aggregate that correspond to early *vs*. late firing of a shared initiation site (*e.g.*, **Figure 4.14b**, *left*). Most intriguingly, we also observe a novel type of cell-type difference invisible to ensemble profiling methods: a subset of cell-type differences that appears to be driven by inconsistent usage of an initiation site in one cell type (*e.g.*, **Figure 4.14**, *left ~196.1Mb*).

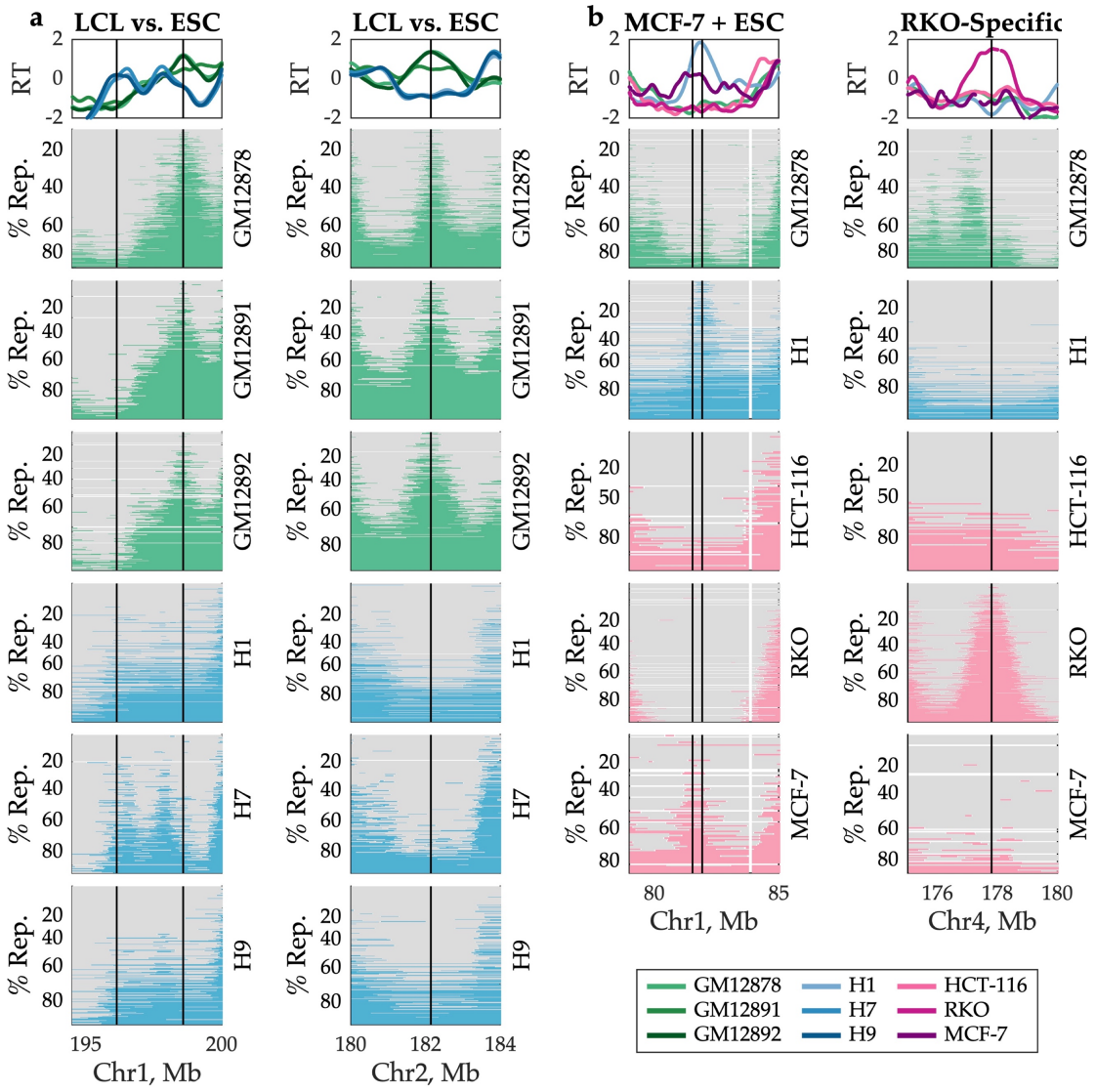


Figure 4.14. **Cell-type- and cell-line-specific differences in the aggregate replication timing profiles are reflected at the single-cell level.** The replication timing profiles shown are S/G_1 bulk-sequencing profiles. **a** Cell-type differences in replication timing between LCL and ESCs are observed at the single-cell level. *Left*: an ESC-specific peak in the bulk profile ~196.1Mb corresponds to a region of consistent replication initiation across all three ESCs, which is absent in GM1278 and GM12892 and fires sporadically in GM12891. Likewise, an LCL-specific peak in the bulk profile ~198.5Mb corresponded to a region of consistent replication initiation in LCL but not in ESC. *Right*: a second example of an LCL-specific region of replication initiation. **b** Differences in replication timing unique to individual cancer cell lines are consistent at the single-cell level. *Left*: an MCF-7-specific peak ~81.5Mb is observed at the single-cell level only in that cell line. To clarify that this initiation site is *not* observed in H1, a neighboring peak in ESC ~81.9Mb is also indicated with a vertical black line. *Right*: an RKO-specific peak in the bulk profile is observed to be unique to RKO single cells as well.

We proceeded to call IRs in each cell line and repeated the above analyses of IR order and timing variability. Despite having ~10 times fewer cells relative to GM12878, we were able to identify 1,811-5,055 IRs (compared to 7,522 in GM12878) per cell line in all cell lines except for HCT-116 (discussed below). To directly test the hypothesis that we identified fewer IRs because of the smaller number of cells, we performed down-sampling of the GM12878 cell line and confirmed that the number of IRs calls rapidly increases with increasing sample size (**Figure 4.15**). IRs called for other cell lines were slightly broader than the GM12878 IRs, but still localized (median: 110kb-220kb; **Figure 4.16a**). This suggests that increasing the number of cells analyzed will likely yield additional IRs in all cell lines and also further narrow their localization.

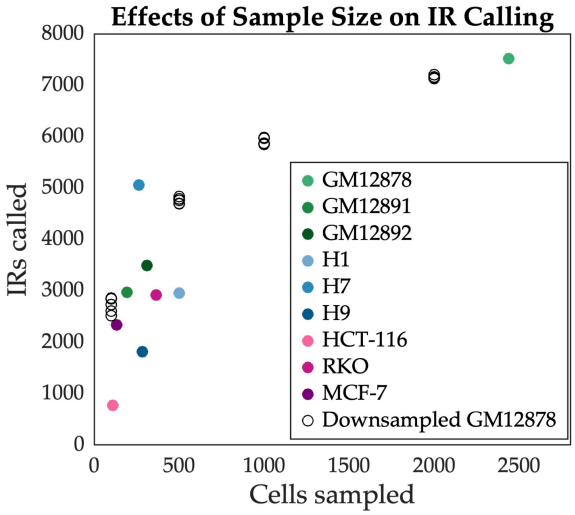


Figure 4.15. **The number of initiation region (IR) calls increases non-linearly with increasing number of cells analyzed.** GM12878 cells were down-sampled (n = 250; 500; 1,000; 2,000) and initiation regions were called from each sample (*black circles*). Down-sampling was performed five times for each sample size to account for stochastic effects of small sample sizes. The observed number of IRs in each cell line is shown for comparison (*colored dots*).

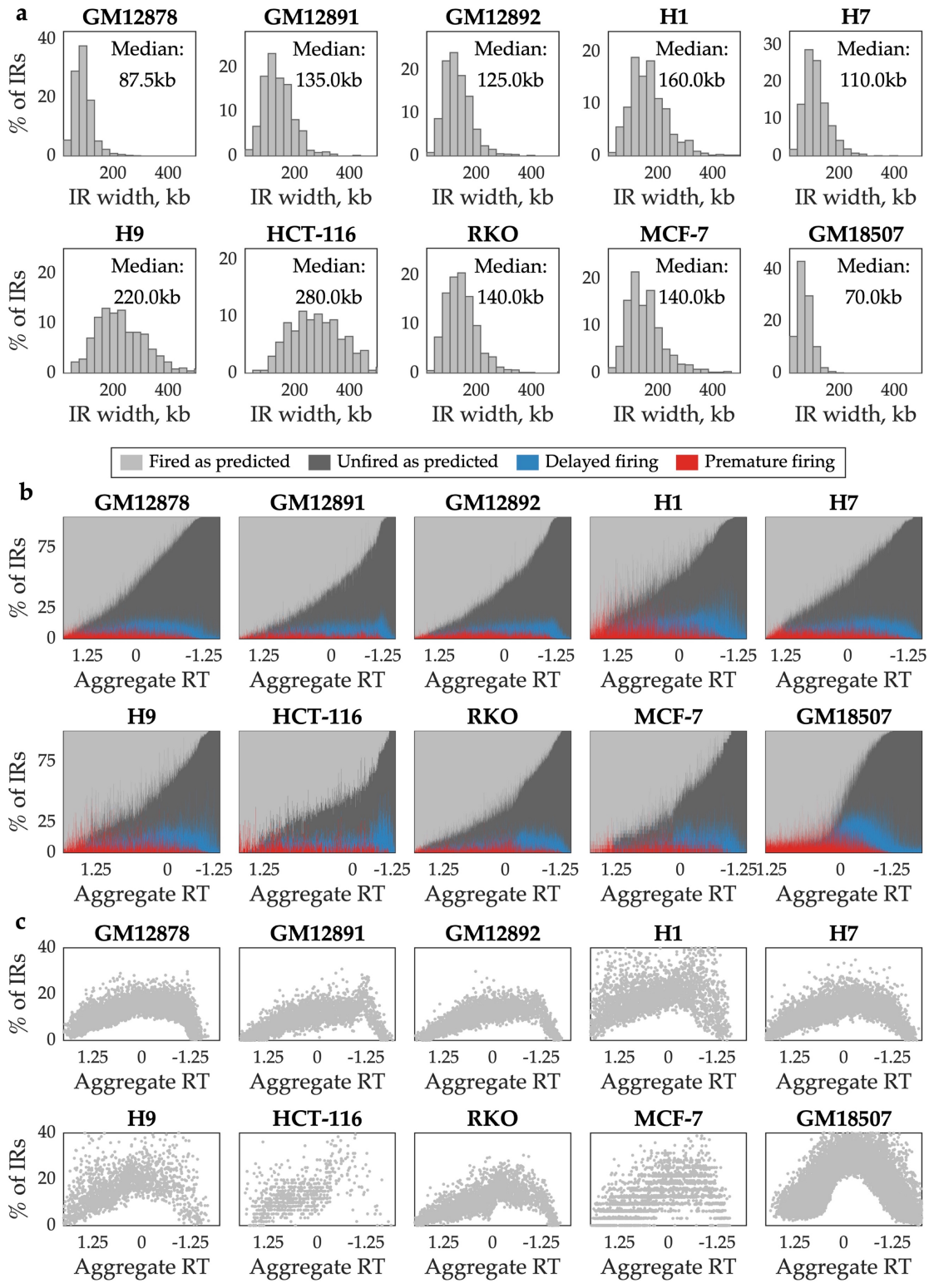


Figure 4.16. **Initiation regions (IRs) fire in a similar, but not fixed order across cell lines. a** The distribution of IR widths is shown for all IRs genome-wide in each cell line under study. With the exception of HCT-116, where only 758 larger IRs were called, the distribution of IR

widths was similar across cell lines and cell types, despite the noted cell-type-specific difference in IR location and the ~10x reduction in sample size relative to GM12878. **b** IRs primarily fire within the expected order (*light and dark gray*), but sometimes fire earlier (*red*) or later (*blue*) than expected by a strictly determined ordering. There are no observed cell-type-specific differences. Compare to **Figure 4.9c**. **c** Variability in IR firing order is not uniform across S phase, but least pronounced at the beginning and end of S phase. This trend is not apparent in HCT-116 where almost all called IRs were early firing. Compare to **Figure 4.9d**. GM18507 cells were sequenced after DLP+ library preparation[114].

Patterns of initiation site localization and timing variability across cell lines were broadly consistent with those observed in GM12878, even though the specific locations of IRs differed between cell types. IRs were fired in a predictable but not fixed order (**Figure 4.16b**) that was more disordered for those IRs with mid-S-phase aggregate timing (**Figure 4.16c**). With regards to firing time, ensemble-late IRs fired early in S phase in a subset of cells, although with the smaller sample size, fewer IRs had multiple cells corroborating this behavior (**Figure 4.17a, b**). This is consistent with how rarely these events occurred per IR in GM12878 and suggests that these events would be observed more frequently in other cell lines when looking across a larger number of cells. However, the fact that so many rare events are observed even in a sample size of ~200 cells suggests that the full scope of variability remains underestimated, including in GM12878.

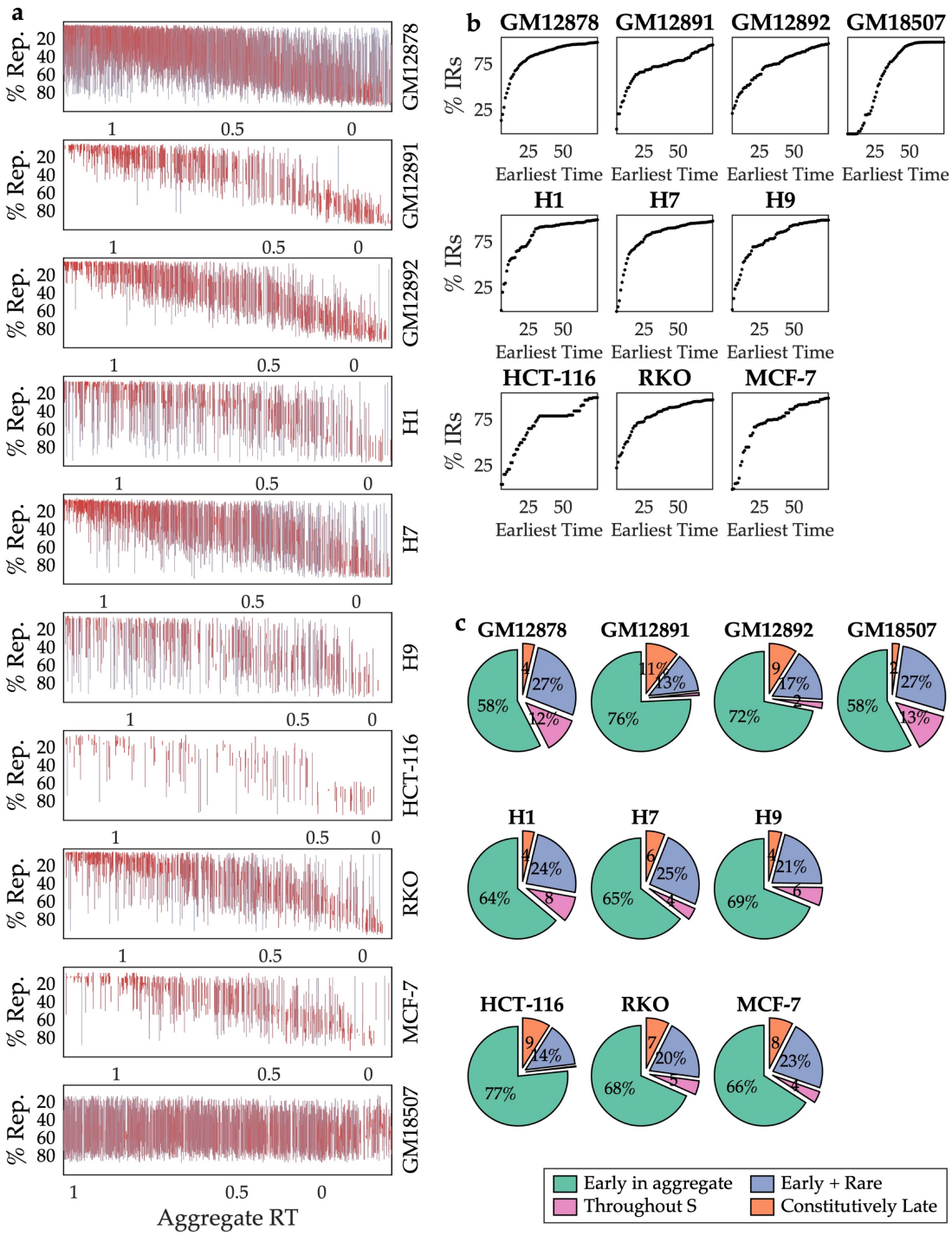


Figure 4.17. **Initiation regions (IRs) with early replication timing in aggregate tend to complete replication in early S phase, whereas IRs with late replication timing in aggregate tend to fire across a wider range of S phase. a** The range of IR firing was more constrained in IRs with early aggregate replication timing and less constrained in IRs with late aggregate replication timing. With a smaller sample size, we were able to call fewer IRs and we were able to corroborate the early-S-phase firing of many IRs in a second cell. This produces a more

constrained range in the additional cell lines (most pronounced in HCT-116). A random subsample of ¼ of the IRs is shown for GM18507, where the large number of IRs called obscures the trend. **b** Most IRs were observed to fire in early S phase in at least one cell. Each panel shows the cumulative percent of IRs that fire earliest before a given time in S phase. Across cell lines, 80-97% of IRs were observed to first fire in a cell < 50% replicated. **c** Distribution of late IR classes across the additional cell lines. Although the smaller sample sizes resulted in calling about half as many IRs as in GM12878, a class of IRs replicating throughout S phase was observed in almost every cell line. The effects of sampling are most pronounced in the inflated proportion of IRs categorized as "early". GM18507 cells were sequenced after DLP+ library preparation[114].

Finally, the three classes of ensemble-late IRs were present in each cell line (except HCT-116), and two features were common between GM12878 and other cell lines. First, rarely used IRs with a preference for early firing were more common than IRs that fired throughout S phase; second, a small fraction (4-11% in most cell lines) of IRs were constitutively late (**Figure 4.13e, f**; **Figure 4.17c**).

The outlier cell line was the colorectal cancer cell line HCT-116, for which we recovered only 110 replicating cells and identified only 768 IRs. In addition to wider IRs, with a median width of 280kb, 78% of IRs identified in HCT-116 were ensemble-early IRs. (In other cell lines, this value was close to 50%, in line with the genome-wide replication timing values.) This bias toward discovering IRs in early-replicating regions creates the impression that variability increases monotonically across S phase, particularly when examining HCT-116 alone. These results are presented alongside those of the other cell lines to illustrate how a low cell count can bias IR identification and conclusions drawn from subsequent analyses. However, while not particularly informative about IRs in late-replicating regions, the data from HCT-116 are not incompatible with the trends observed specifically for early IRs across cell lines.

Additionally, we repeated all analyses for the DLP+ GM18507 library. Single-cell replication profiles for this cell line were noisier (**Figure 4.6b**), and we called 12,952 IRs from 759 cells. GM18507 IRs recapitulated the results from the 10x Genomics cell lines, with the caveat that these cells were strongly skewed toward mid S phase (**Figure 4.16**; **Figure 4.17**).

In summary, data from ten human cell lines encompassing LCLs, ESCs, and cancer cell lines support the conclusion that replication initiation occurs in localized regions that are largely consistent across cells. Furthermore, patterns of heterogeneity in origin firing order and firing time appear to be generalizable across cell lines and cell types.

Discussion

While ensemble replication profiling methods cannot capture (and may be confounded by) cell-to-cell heterogeneity, previous single-molecule and single-cell methods have been largely limited in their throughput or accuracy. Here, we report a scalable method for analysis of thousands of single replicating cells, across multiple cell lines, and at kilobase resolution. We describe an *in silico* strategy to sort cells across S phase, analogous to and more accurate than traditional flow cytometry, and demonstrate how this method enables simultaneous analysis of replication initiation at population, subpopulation, and single-cell resolutions. In addition, by focusing specifically on replication initiation events called from single cells, we are able to identify which cells are informative about which replication initiation sites, capturing information that is analogous to that collected from lower-throughput single-molecule studies. In a parallel study, Gnan *et al*.[167] developed a similar

approach to use single-cell sequencing data to infer DNA replication timing at large scale.

We find that single cells primarily initiate replication at consistent loci, corresponding to peaks in the replication timing profile. Across cells, we are able to pinpoint the locations of 78.9% of these initiation events to regions no larger than 100kb (likely overestimated due to low coverage), challenging the model that there are megabase-long replication domains that are replicated simultaneously[48,158]. Analogously, our data do not support the existence of large constant replication regions (CTRs)[168], particularly in early-replicating regions, for which we have more data. While it is conceivably straightforward to envision how measurements with limited resolution would give the impression of domains or CTRs where none exist, it appears more difficult to reconcile the sharp and discrete initiation peaks in our single-cell data with the idea of large regions with constant replication timing. In contrast, our data is consistent with recent high-resolution studies that suggest that replication initiation is confined to regions of several tens of kilobases[62,96,166,169]. Our observation that even tight peaks in ensemble replication timing profiles often encompass multiple discrete single-cell initiation events lends further credence to the argument that initiation events are even more localized than measured here, with the caveat that we cannot distinguish in our data between single origins and tight clusters of nearby origins. We find evidence for ectopic initiation from regions outside these commonly used initiation regions (as in [96]), although these events comprise a small fraction of overall events. While many previous studies of mammalian replication origins relied on biochemical enrichments of DNA synthesis events and are therefore more prone to false-positive identification of apparent initiation events, single-cell

DNA sequencing more reliably represents productive and internally-validated DNA replication[157].

While spatial variability in replication initiation is rare, temporal variation is more common. In general, initiation regions (IRs) expected to fire in the middle of S phase are more variable than those expected to fire earlier or later, consistent with previous reports[109]. At the level of individual IRs, we find that many, particularly those with early aggregate replication timing, have a preferred time of firing that is captured by the aggregate replication timing profile. IRs with early aggregate replication timing tend to be fired in all cells early in S phase, while those with late aggregate replication timing fire across a broader range of S phase. We further find that late-replicating IRs can be divided into multiple classes, with only a small subset (< 10%) firing constitutively late. Instead, most late IRs can and do fire early – sometimes rarely and sometimes often.

Our data do not rule out the possibility of a global regulator (or regulators) that dictates replication timing in a semi-deterministic manner. However, they are also consistent with the more parsimonious model that origin-specific firing probabilities produce a relatively consistent replication timing landscape in single cells. IRs with late aggregate replication timing that occasionally fire early in S phase are consistent with this hypothesis: these rarely-early IRs could contain an inefficient origin (or clusters of inefficient origins) that rarely fires but can be early firing when it does fire, and this low efficiency is what is measured by the aggregate replication timing profile. The constitutively late IRs have even later aggregate replication timing; under this same hypothesis, they would be expected to fire early in S phase even less often. Thus, we cannot rule out the possibility that the IRs we observed to be

constitutively late do sometimes fire early, but at so low a frequency that it was not captured in our sample.

While our data are consistent with an important role for origin firing efficiency in determining replication timing, the distinct classes of initiation regions we describe also highlight a shortcoming of considering origin efficiency at the level of individual loci: while low-efficiency origins would be expected to rarely fire in early S phase, their probability of firing should remain constant or even increase as S phase progresses.[88] In other words, origins in late-replicating regions of the genome should fire throughout S phase. Instead, we see that the majority (63.7%) of the inefficient IRs have a low probability of firing in early S phase, and a negligible probability of firing later in S phase (**Figure 4.10e**), suggesting that the context of replication initiation changes across S phase in a manner that has not been previously characterized.

Our results suggest origin-specific firing efficiencies play a key role in producing the replication timing program; as such, they underscore the value of future work parsing out the contributions of DNA sequence, gene expression, chromatin accessibility, and doubtless other factors to these firing efficiencies. At the same time, a future model for replication timing must also explain why many origins appear to have their highest probability of firing at the beginning of S phase, rather than becoming increasingly likely to fire as S phase progresses – and also why that does not result in large regions of under-replication that persist into G_2 phase, as modeled in [88].

While single-cell sequencing provides insight into cell-to-cell variability that ensemble measurements cannot capture, such experiments are also more expensive, more time-consuming, and require more complex analysis methods. The approach we describe here is limited by the high cost of existing

commercial microfluidics-based library preparation methods, especially given that not every cell will be informative about every locus. Thus, there is a tradeoff between sample size and information content per cell. In this study, we have demonstrated that low per-cell sequencing coverage is sufficient for distinguishing the two-fold copy-number difference between replicated and unreplicated regions at ~20kb resolution. However, we are unable to reliably distinguish smaller differences in copy number, *i.e.*, between 2 and 3, or between 3 and 4 copies. Increased resolution and/or allele-specific mapping[108,109] would be required to identify cases of allelic asynchrony, especially those specific to individual cells. In addition, single-cell sequencing is, at least currently, not the optimal technology for identifying individual replication origins, and existing origin-mapping methods (*e.g.*, OK-seq[166], EdUseq-HU[169], high-resolution Repli-seq[62], or optical replication mapping[96]) are better suited to this purpose. Origins identified by these methods overlap well with ensemble replication timing profiles, as do the IRs we identified here.

Single-cell DNA sequencing of proliferating cell samples, without experimental manipulation (*e.g.*, cell synchronization or sorting), can reveal the dynamics of DNA replication in exquisite detail. Applying this approach across cell types, genetic backgrounds, and experimental conditions will reveal how replication is altered at the spatiotemporal levels in different physiological contexts. With constantly improving methods for high-throughput single-cell isolation and accurate whole-genome amplification[111,116,170], this approach promises to become ever more informative for the understanding of the DNA replication timing program.

Methods

Cell culture

Lymphoblastoid cell lines (GM12878, GM12891, and GM12892) were obtained from the Coriell Institute for Medical Research and cultured in Roswell Park Memorial Institute 1640 medium (Corning Life Sciences, Tewksbury, MA, USA), supplemented with 15% fetal bovine serum (FBS; Corning). Embryonic stem cell lines (H1, H7, and H9) were obtained from the WiCell Research Institute (Madison, WI, USA) and cultured feeder-free on Matrigel culture matrix in mTeSR™ 1 medium (WiCell). Tumor-derived cell lines (MCF-7, RKO, and HCT-116) were obtained from the American Type Culture Collection. MCF-7 and RKO cells were cultured in Eagle's Minimum Essential Medium (Corning), supplemented with 10% FBS. HCT-116 cells were cultured in McCoy's 5a medium (Corning), supplemented with 10% FBS. All cell lines were grown at 37°C in a 5% CO_2 atmosphere.

Library preparation and sequencing

For sorted libraries, GM12878 were stained with Vybrant™ DyeCycle™ Green Stain (ThermoFisher Scientific, Waltham, MA, USA) and sorted into five fractions (G_1, G_2, early-S, late-S, and full S phase) with a BD FACSMelody™ Cell Sorter (BD Biosciences, Franklin Lakes, NJ, USA).

For both sorted and unsorted libraries, isolation, barcoding, and amplification of single-cell genomic DNA was performed on the 10x Genomics Chromium Controller instrument, using the 10x Genomics Single Cell CNV Solution kit (10x Genomics, Pleasanton, CA, USA). Paired-end

sequencing was performed for 100 cycles with the Illumina NovaSeq 6000 (10x Genomics), 150 cycles with the Illumina HiSeq X Ten (GENEWIZ, Inc., South Plainfield, NJ, USA), or 36 or 75 cycles with the Illumina NextSeq 500 (Cornell University Biotechnology Resource Center, Ithaca, NY, USA). For libraries sequenced multiple times, FASTQ files were merged prior to downstream processing. See **Table 4.1** for details.

Table 4.1. **DNA sequencing details for each sequencing library generated in this chapter.** Replicates are derived from culture of independent cryogenic vials. Each library was sequenced one to three times, as indicated.

Cell Line	Fraction	Rep.	Instrument	Cycles	Center
GM12878	G_1	1	NovaSeq 6000	2 × 100	10x Genomics
GM12878	Early-S	1	NovaSeq 6000	2 × 100	10x Genomics
GM12878	Full S	1	NovaSeq 6000	2 × 100	10x Genomics
GM12878	Late-S	1	NovaSeq 6000	2 × 100	10x Genomics
GM12878	G_2	1	NovaSeq 6000	2 × 100	10x Genomics
GM12878	Unsorted	1	NovaSeq 6000	2 × 100	10x Genomics
GM12878	Unsorted	2	HiSeq X Ten	2 × 150	GENEWIZ
GM12878	Unsorted	3	HiSeq X Ten	2 × 150	GENEWIZ
GM12891	Unsorted	1	HiSeq X Ten	2 × 150	GENEWIZ
			HiSeq X Ten	2 × 150	GENEWIZ
GM12892	Unsorted	1	HiSeq X Ten	2 × 150	GENEWIZ
			HiSeq X Ten	2 × 150	GENEWIZ
H1	Unsorted	1	HiSeq X Ten	2 × 150	GENEWIZ
H1	Unsorted	2	NextSeq 500	2 × 36	Cornell BRC
			HiSeq X Ten	2 × 150	GENEWIZ
			HiSeq X Ten	2 × 150	GENEWIZ
H7	Unsorted	1	HiSeq X Ten	2 × 150	GENEWIZ
			HiSeq X Ten	2 × 150	GENEWIZ
H9	Unsorted	1	HiSeq X Ten	2 × 150	GENEWIZ
HCT-116	Unsorted	1	HiSeq X Ten	2 × 150	GENEWIZ
			HiSeq X Ten	2 × 150	GENEWIZ
RKO	Unsorted	1	HiSeq X Ten	2 × 150	GENEWIZ
			HiSeq X Ten	2 × 150	GENEWIZ
MCF-7	Unsorted	1	NextSeq 500	2 × 75	Cornell BRC
			HiSeq X Ten	2 × 150	GENEWIZ
			HiSeq X Ten	2 × 150	GENEWIZ

Processing of single-cell barcodes

The first 16bp of each R1 read (containing the cell-specific barcode) was trimmed with seqtk (v1.2-r102-dirty). Raw barcode sequences were compared to a list of 737,280 possible barcode sequences (10x Genomics) and filtered by abundance to produce a list of barcodes present in the library. Specifically, a set of "high count" barcodes was identified as those that were represented at least 1/10 as often as the highest abundance barcode. A minimum barcode abundance threshold was then set as 1/10 the 95th percentile of the high-count abundances.

Next, we attempted to correct barcode reads that were not found in the set of valid barcodes. To be corrected, we required that the barcode read contain no more than one base position with a quality score < 24 and that there was only one valid barcode with a Hamming distance of 1.

Processing of sequencing reads

After filtering out sequencing reads without a valid barcode, reads were aligned to the human reference genome hg37 using the Burrows-Wheeler maximal exact matches (BWA-MEM) algorithm (bwa v0.7.13). Barcodes were then merged into the aligned BAM files using a custom awk script, and barcode-aware duplicate marking was performed using Picard Tools (v2.9.0). High-quality (MAPQ ≥ 30) primary mate-pair alignments were included in further analysis. Members of a mate-pair were counted together if they were mapped within 20kb of one another (weight of 0.5/read), and separately (weight of 1/read) if not.

Identification of G_1/G_2 cells and definition of G_1 windows

Reads were counted in fixed size windows of 20kb. After removing low-mappability windows (in which fewer < 75% of nucleotide positions were uniquely mappable[63]), sets of 50 windows were aggregated together to calculate the median absolute deviation of pairwise differences between adjacent windows (MAPD)[159]. MAPD was then divided by the square root of the mean number of reads per aggregated window (mean coverage/Mb), to produce a linear relationship between coverage and scaled MAPD. For each sequencing library, an expectation-maximization procedure was used to fit the data as a mixture of two Gaussian functions. The linear fit predicting the smaller scaled MAPD value at the maximum observed coverage was assumed to model the G_1/G_2 relationship between coverage and scaled MAPD, and cells with a residual ≤ 0.05 from this model were assigned as G_1/G_2.

Next, we defined a set of variable-size, fixed-coverage windows using a G_1 control, along the lines of Koren *et al.*[45]. In this case, the G_1 control was created *in silico* by aggregating reads from a subset of G_1/G_2 cells, prioritizing high-coverage G_1/G_2 cells. (The number of cells used varied between libraries and was determined as the number of cells that would define windows of ~20kb.) This was performed independently for each sequencing library prepared from the same cell line. Per-cell read counts were calculated in these G_1-windows, to account for mappability and GC-content bias, as well as any copy-number variations that were common to many cells within the library.

Finally, we identified and filtered out cell-specific copy-number aberrations (CNA). To do this, we fit a two-component mixed Poisson model to aggregated read counts (15 windows, ~300kb), and searched for the genomic region with the lowest probability of being observed under either

rate coefficient, λ. If the median probability of each window within this region was less than the median probability of all windows genome-wide, we determined it to be a CNA and masked the read counts in that region. This process was performed iteratively until no new regions were discovered. Cells with an autocorrelation in read counts > 0.15 after filtering were assumed to have residual undetected CNA and were excluded from analysis.

Replication state inference

For each cell, we assigned each G_1-defined window as "replicated" or "unreplicated" using a two-state hidden Markov model (HMM). To initialize the model, we again fit a two-component mixed Poisson model to aggregated read counts (15 windows, ~300kb) and assigned each window to the mean to which it was closest. If this initial model did not converge, or if the ratio between the two mean copy numbers was not ~2 (between 1.5 and 2.5), the cell was excluded. Otherwise, we refined the initial window assignments using the HMM, which modeled read counts as the mixture of two Poisson processes.

Because the HMM does not model the expected two-fold relationship between replicated and unreplicated regions, we assessed the quality of the HMM output using this ratio. Specifically, we calculated the ratio between the average number of reads in windows assigned as replicated to the average number of reads in windows assigned as unreplicated. To be included in further analysis, this ratio was required to be between 1.5 and 2.5. This filter removed cells poorly modeled by the HMM, which could be explained by a variety of biological and technical factors, including large CNA, large

replication defects, ineffective selection of cell-specific initialization parameters, and atypical noise.

Additionally, to find any cells that contained uncorrected CNA, we performed three filtering steps. First, we calculated the average copy-number assigned to each chromosome and excluded cells for which the standard deviation between chromosomes was greater than 0.4. Second, any cell that contained both a fully unreplicated chromosome and a fully replicated chromosome was excluded. Third, we calculated the pairwise correlations between cells for each chromosome individually. If the mean pairwise correlation between a cell and all other cells was negative, or if the pairwise correlation between a cell and one of its 10 closest neighbors was a statistical outlier, that chromosome was excluded for that cell.

Finally, for the ease of analysis, we interpolated the data back onto fixed size 20kb windows. Interpolated values for fixed size windows that overlapped multiple G_1-defined windows were not always integers. Thus, windows assigned a non-integer copy number were masked, as were low-mappability windows.

Assessment of HMM resolution

To assess the resolution of HMM copy-number inferences, read counts were simulated for 2,500 single GM12878 cells following a strictly determined replication timing program. First, the bulk replication timing profile was divided into 1,000 equally-spaced bins (corresponding to ~0.046 replication timing units). For each vertical "time point" slice through the bulk profile, windows with earlier replication timing were assigned “4N” and windows with later replication timing were assigned “2N”. For each simulated cell, one

of these 1,000 possible time points was randomly selected, and an average coverage was drawn from the distribution of observed coverage values for the unsorted GM12878 library. Then, read counts for 2N and 4N windows were drawn from two Poisson distributions, with rate coefficients selected to produce the desired average coverage. Finally, the simulated read counts were run through the copy-number inference pipeline.

Bulk-sequencing replication timing profiles

Replication timing profiles from bulk sequencing assays were used to benchmark single-cell replication profiles. For GM18507, an LCL consensus profile[45] was used. For all other cell lines, a profile for the specific cell line was used. For Illumina Platinum LCLs (GM12878, GM12891, and GM12892)[171] and hESCs (H1, H7, and H9)[46], these data are previously published.

Aggregate replication timing profiles

For each cell line, we generated an aggregate S/G_1 profile, as in [45], except that we generated the G_1 and S fractions *in silico* by aggregating reads across all cells assigned to that fraction. Briefly, the G_1 fraction was used to generated variable-size windows with a fixed number of reads (n = 200), and the number of S-phase reads was then counted in the same windows. This profile was smoothed in a gap-aware fashion with a cubic smoothing spline (MATLAB function `csaps`, with smoothing parameter 1×10^{-16}), and normalized to a mean of 0 and standard deviation of 1.

Sub-S-phase fraction profiles

To generate a profile for 10 sub-S-phase fractions, we partitioned cells into 10 bins of equal cell population, based on the % of the genome replicated. We summed the read counts (in G_1-normalized windows) across all cells within each partition. To normalize read counts between fractions, we then scaled these values, setting the 1st percentile value as 2 and the 99.9th percentile value as 4. The same procedure was used to generate 100 fractions.

Identification of initiation regions

To identify single-cell replication initiation sites, we began by defining all replicated segments ("replication tracks") across the genome of each cell. These segments were defined as contiguous windows with inferred copy-number of 4, containing no more than 5 consecutive masked windows. As a first approximation of the locations of replication initiation, the midpoint of each replication track was assigned as the most likely site of initiation. (Replication tracks longer than 1Mb were excluded from this analysis.)

To cluster single-cell initiation sites, we grouped together replication tracks that overlapped one another. We considered three possible midpoints for each replication track: the observed midpoint as well as the midpoint if either the left or right boundary had been misplaced by 2.5 windows. Starting with the shortest replication tracks, we asked whether each replication track overlapped any previously defined initiation regions (IRs). Tracks overlapping a single IR were attributed to activity of that IR (as long as its midpoint overlapped at least one track already assigned to that IR), while tracks that did not overlap any IRs were used to define a novel IR. Tracks that

overlapped multiple IRs were inferred to reflect the activity of multiple initiation events and were not used to define IRs.

After defining IRs, we reconsidered any track less than 1Mb in length that had not been attributed to an IR (*i.e.*, tracks that overlapped multiple IRs). These tracks were then assigned to the IR closest to its midpoint. The width of each IR was calculated from the 25th percentile to the 75th percentile of the midpoints of replication tracks attributed to the IR, and the center was set at the 50th percentile. IRs supported by fewer than 5 tracks were not included when calculating the median IR width.

Variation in firing order across cells

To assess variation in the order in which IRs were fired across cells, we compared the data to a null model under which every cell fires the same IRs in the same order. Under this model, the number of IRs inferred to be replicated also dictates which IRs those are. Thus, we counted the number of replicated regions overlapping IRs in each cell, and then predicted which regions those would be under the null model. For each IR, we then calculated how many cells did not match our prediction.

Variation in firing time across cells

To determine the range of firing orders for each IR, we identified the earliest cell containing a replication track attributed to an IR, and the latest cell in which the center of the IR was inferred to be unreplicated (after excluding outlier cells that had not replicated any of the neighboring IRs). The percent of the genome replicated in each of these cells was used as a proxy for time

during S phase. Given that range is a metric extremely sensitive to outliers, we considered an IR's range to be "corroborated" if a second cell was observed within 10% of its earliest and latest firing time. We focused on these IRs with corroborated ranges in subsequent analyses.

Finally, we classified IRs that fired in fewer than 50% of cells into three groups based on their firing behavior throughout S phase. To do this, we considered the percent of the genome replicated in each cell containing a replication track attributed to that IR. IRs that were not associated with any cells < 50% replicated were considered constitutively late firing, while those associated with more than 5 cells > 50% replicated were considered to fire throughout S phase. The remaining IRs, which were associated with 1-5 cells in early S phase, were considered to be rarely fired with a preference for early firing.

Data availability

Sequencing data generated in this manuscript were deposited at the Sequence Read Archive under accessions PRJNA770772 (single-cell) and PRJNA419407 (bulk). Bulk-sequencing replication timing profiles used for comparison are available at http://www.thekorenlab.org/data.

Code availability

All scripts used in data processing, analysis, and visualization are available at: https://github.com/TheKorenLab/Single-cell-replication-timing.

Acknowledgements

We thank Claudia Catalanotti and Rajiv Bharadwaj for sharing data analyzed in this work, Peter Schweitzer and Jennifer Mosher for technical support and guidance with library preparation and sequencing, Kevin Massey for many fruitful discussions about algorithms, Sneha Sharma for help with various preliminary analyses, and members of the Koren lab for their feedback. This work was supported by the National Institutes of Health (grant DP2-GM123495 to A.K) and a seed grant from the Cornell Center for Vertebrate Genomics.

CHAPTER 5: CONCLUSIONS AND FUTURE DIRECTIONS

In the work presented in this dissertation, I have focused on two aspects of human DNA replication timing variability that have been underexplored in the literature to date: (1) replication timing of satellite DNA and (2) replication timing variability at the single-cell level.

Replication timing of satellite DNA

Satellite DNA has long been difficult to study because its high repeat content is a significant hurdle to short-read sequence alignment. In **Chapter 2**, I used computationally-modeled centromere sequence decoys in the current human reference genome, hg38, to measure replication timing of centromeric regions, specifically. Although these decoy sequences do not provide reliable local information, I was able to leverage them to draw general conclusions about the replication timing of centromeres as a class of genomic regions. Specifically, I showed that centromeric regions replicated in mid-to-late S phase, were often earlier replicating than their surrounding context (suggesting the presence of replication origins), and were unusually variable between cell lines relative to other genomic regions.

In **Chapter 3**, I took advantage of a newer human genome assembly to follow up on this earlier study. Using an almost-entirely homozygous cell line derived from a complete hydatidiform mole, the Telomere-to-Telomere Consortium released the first complete and gapless human genome assembly, T2T-CHM13, in 2021. Re-analysis of the sequencing data used in **Chapter 2** revealed replication timing information not only for centromeres but also for large arrays of constitutive heterochromatin that have previously been

represented in the human reference genome only as long stretches of unknown bases. I confirmed the prior results that centromeres replicated in mid-to-late S phase and were hyper-variable between cell lines relative to other genomic regions. Because T2T-CHM13 contains linear sequences for centromeric regions, I was further able to demonstrate the existence of replication timing peaks (likely, replication origins) within these regions, and to show that differences between cell lines were global across all centromeres, rather than driven by one or two chromosomes. In addition, I was able to show that while all families of satellite DNA within heterochromatin are biased toward replication in late S phase, a subset (namely, the human satellite 2 and 3 families) are more strongly late-replicating.

The T2T-CHM13 assembly has opened the door to much more in-depth analysis of replication timing of heterochromatin across cell lines; my work has just barely scratched the surface. One direction of particular interest for future work is inquiry into the causes and consequences of replication timing differences between cell lines. My finding that some cell lines have consistently earlier replication of centromeres than others raises several intriguing hypotheses about how centromeric replication is regulated but these possible explanations are impossible to rigorously assess based on only five cell lines. The most pressing methodological hurdle is the requirement for matched G_1- and S-phase samples, which precludes the use of most published sequence data, and limits analysis to samples that have been sequenced specifically for replication timing analysis. While methods exist to infer replication timing from unsorted cell populations, they do not adequately account for the low mappability of satellite DNA – at least, currently. Even if these methods can be adapted for the study of these regions, accounting for

centromeric copy-number variation between cell lines will likely remain a challenge without a matched control.

Single-cell replication timing variability

Although ensemble assays have demonstrated that replication timing profiles are remarkably reproducible, debate persists as to what this means at the single-cell level. Thus, the second major question my dissertation addresses is variability in replication timing among individual cells, when I discuss in **Chapter 4**. Analyzing DNA sequencing data from 22,278 single cells across ten cell lines, I first developed a strategy for *in silico* cell sorting to computationally discriminate replicating and non-replicating cells. This enabled me to use non-replicating cells to account for sequencing biases and copy-number variants when analyzing replicating cells within the same sequencing library. I next developed a hidden Markov model-based approach to assign each genomic window in a replicating cell as either "unreplicated" or "replicated".

Within each of the ten cell lines, I found remarkable consistency in replication initiation at the single-cell level. Across the genome, the same regions were found to be replicated in almost every cell, centered on peaks in ensemble replication timing. This suggests that the same set of genomic loci act as replication origins across all (or most) cell cycles. While single-cell variability was low along the spatial dimension, it was more variable in time: the order in which genomic regions were replicated was predicted by ensemble replication timing, but not strictly followed by every cell. Indeed, each region differed from the expected order in ~11% of cells. Stratifying by ensemble replication timing, I found that regions that are early-replicating in

ensemble assays tend to be consistently replicated early across cells, whereas many regions that are late-replicating in ensemble assays replicate early in a subset of cells. Most surprisingly, I identified a class of "ensemble-late" regions that appear to contain an early-firing origin that is infrequently used. This heterogeneity in late-replicating regions is not predicted by existing models of stochastic origin firing and indicates that some fundamental aspect of the stochastic dynamics of replication initiation is not being modeled.

I propose a three-pronged approach to following up on these results. First, it is worthwhile to continue development of the copy-number inference pipeline. Single-cell DNA sequencing data are extremely noisy and the influence of this noise on supposed heterogeneity is a persistent concern. In particular, the single-cell data generated using the 10x Genomics Single-Cell CNV platform suggest a high frequency of copy-number anomalies, including within presumed diploid cell lines. It is unclear whether this is biological or technical, but in either case, it hinders replication state inference. While I have been largely successful in removing these regions in each cell, further improvements are likely needed. In addition, this noise prevented me from successfully using a three-state model, which is necessary to assess asynchronous replication between homologous chromosomes.

Second, I was surprised to find that replication timing variability was similar between lymphoblastoid, embryonic stem cell, and tumor-derived cell lines. However, my analysis was largely restricted to genome-wide trends. Further analysis of the tumor-derived cell line data may uncover more localized changes in replication timing variability, *e.g.*, in the vicinity of structural variants or near overexpressed genes. Primary tumors and cancer cell lines are among the most common sources of single-cell DNA sequencing

data; thus, I forecast that datasets will become increasingly available to be repurposed for replication timing analysis.

Third, I propose the need for more sophisticated modeling of single-cell replication initiation and progression. Simulation-based modeling of DNA replication is an underutilized tool for understanding the expected distribution of single-cell variation under a variety of mechanistic hypotheses, which can then be used to rule out some potential mechanisms. Such models would be particularly useful for integrating multiple single-cell data types, including single-cell RNA sequencing, single-cell accessibility mapping, and single-cell chromatin conformation. It currently remains infeasible to collect these data from the same single cell, making it difficult to determine, for instance, whether cells with earlier replication of a region also have higher expression of the genes in that region, or if recycling of replication initiation machinery is influenced the distance between loci in three-dimensional space. While modeling cannot answer these questions, it does offer the prospect of understanding that the single-cell data would be expected to look like under a given hypothesis.

Additional applications of single-cell methods

The methods that I developed in **Chapter 4** were intended to be used for analysis of replication timing variability in single cells. However, preliminary results suggest that they can be useful for answering other questions. Below, I present two additional applications for these methods.

Single-micronucleus sequencing

Genomic instability can result in the incorporation of chromosomal fragments into micronuclei. Thus, sequencing of micronuclei can be used to measure the propensity of different genomic regions to breakage and rearrangement[172]. Using single-cell preparation and analysis methods described in **Chapter 4,** I have performed a preliminary analysis of micronuclei isolated from single mouse normochromatic erythrocytes (**Figure 5.1**). While the small sample size limits interpretation of these data, this proof-of-concept experiment demonstrates the utility of single-cell methods to studying micronuclei. These preliminary data concur with ensemble micronucleus sequencing that acentric (right-sided) fragments are more commonly found to reside in micronuclei than centric (left-sided) fragments, although entire chromosomes may also be observed (*e.g.*, **Figure 5.1**, *MCM4^{chaos3} cell 4*). These data may also suggest that some individual micronuclei do contain fragments from multiple chromosomes, a phenomenon that cannot be parsed out from ensemble sequencing. However, we do not have the resolution to determine if these cells (*e.g.*, **Figure 5.1**, *Rad9a^{SA} cell 5*) contain a single micronucleus bearing fragments of multiple chromosomes or multiple micronuclei.

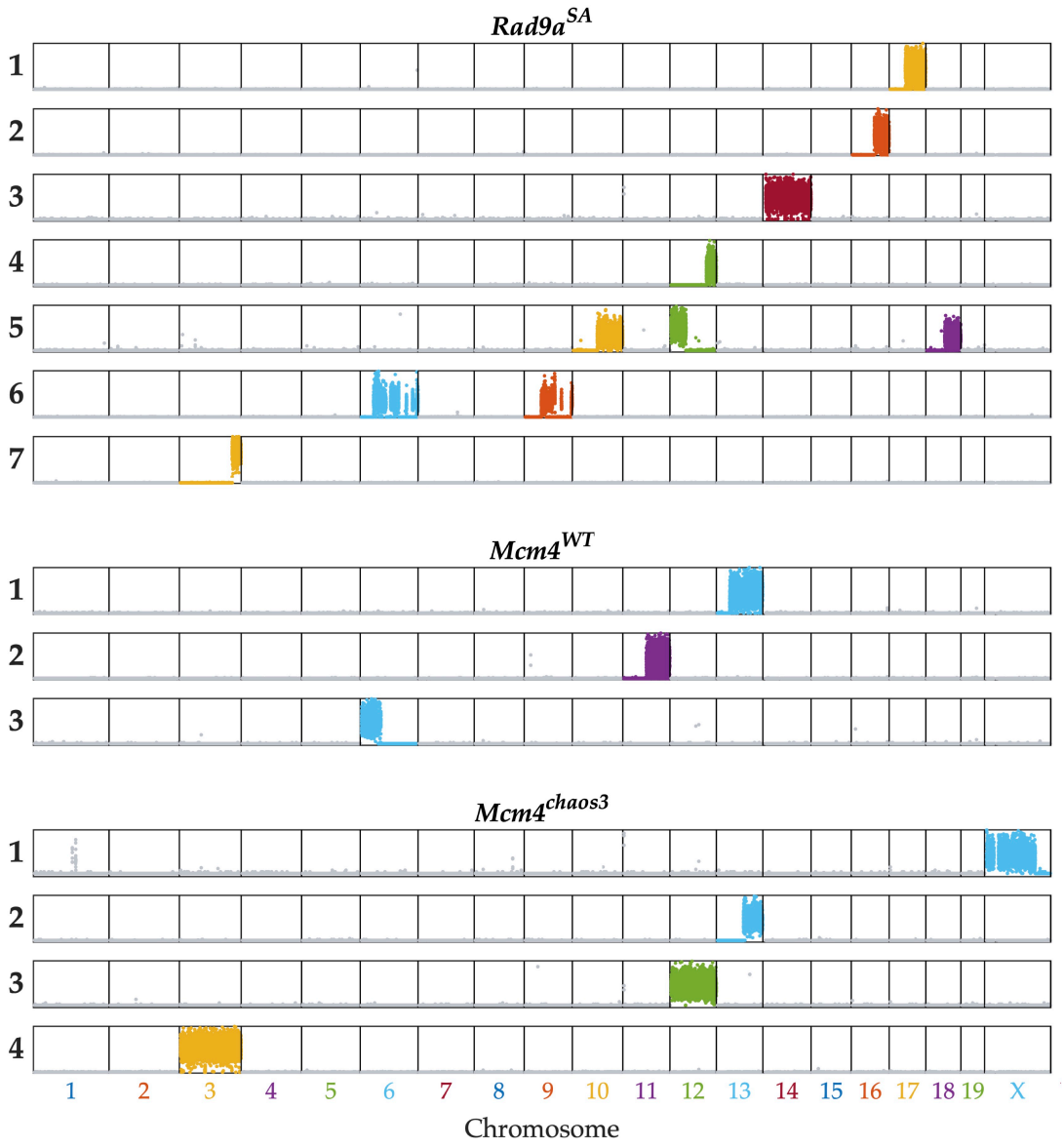


Figure 5.1. **Copy-number profiles for 14 individual micronuclei-containing cells.** Flow-cytometry was used to isolate enucleated mouse normochromatic erythrocytes containing micronuclei, which were then prepared for sequencing using the 10x Genomics Single Cell CNV platform. Each row shows the copy-number profile genome-wide of a single cell. Cells contained fragments of one to three chromosomes. This figure is reproduced from the preprint Pereira et al.[172], on which I am a co-author.

Replication timing for difficult-to-obtain samples

Ensemble replication timing profiles are typically generated from sequencing of large populations of cells (typically, ~1 million cells), and often use flow-cytometry. This has limited the types of samples for which replication timing has been analyzed – in mammals, replication timing has been studied almost exclusively in highly proliferative cell lines. Thus, it remains an open question how well these profiles capture DNA replication as it occurs in the context of primary tissue.

As demonstrated in **Chapter 4,** my *in silico* single-cell cell sorting method can be used to generate ensemble replication timing profiles by aggregating single-cell data (**Figure 4.1d**). In **Chapter 4**, I used this strategy to validate the single-cell replication timing data for cell lines with well-characterized ensemble replication timing profiles. However, this same approach can also be used generate ensemble replication timing data in cases where it is difficult to get sufficient sample for flow-sorting. I applied the *in silico* cell sorting approach to two single-cell datasets generated by 10x Genomics: one for the slow-growing BJ fibroblast cell line and one for a primary triple-negative ductal carcinoma *in situ* breast biopsy. In both contexts, the proportion of proliferating cells is expected to be quite low, making it necessary to use *in silico* sorting to identify the smalls subset of cells that are informative about replication timing.

After identifying S-phase cells, I generated aggregate replication timing profiles for the BJ fibroblast cell line and for the diploid primary cells within the biopsy (likely, healthy cells adjacent to the tumor). Each aggregate profile was positively correlated to an appropriate control profile generated by traditional methods (**Figure 5.2**). However, further refinement is likely

necessary before applying this approach to additional samples. In each sample, sequencing read depth was negatively correlated with GC-content in a substantial proportion of cells, suggesting that they may be undergoing apoptosis, and inclusion of these cells in the analysis confounded replication timing inference. Given the small proportion of cells in S phase, it is likely that additional filtering steps will improve the aggregate profiles further. In addition, in the tumor context, it is essential the G_1- and S-phase cells used to generate the profile have a shared copy-number background. Thus, it is necessary first to accurately assign individual cells to clonal populations, and only then to infer replication timing for each clone separately.

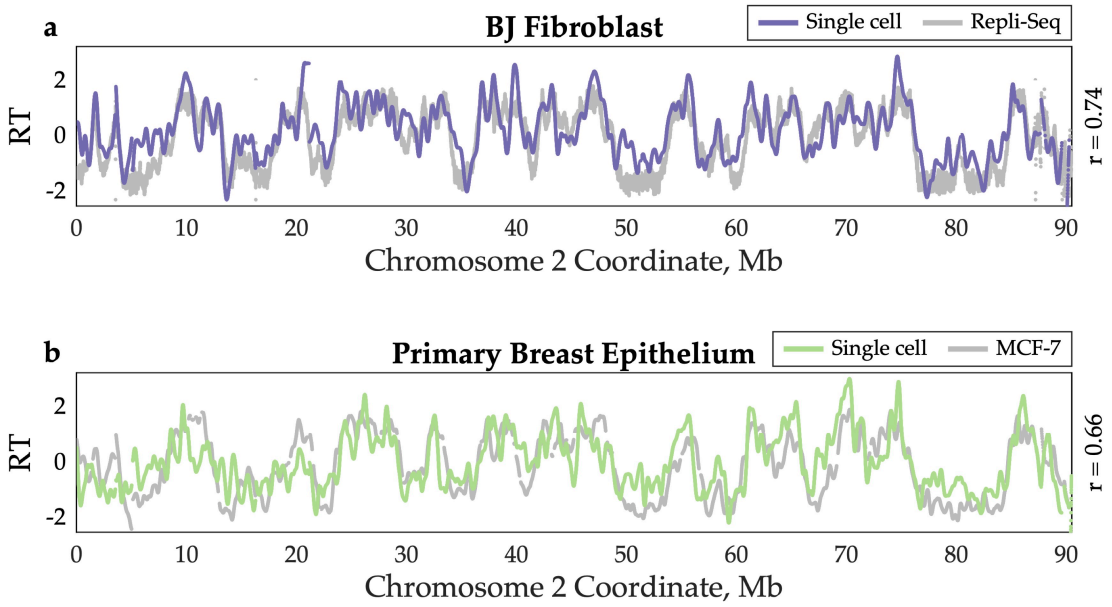


Figure 5.2. **Replication timing profiles can be inferred for difficult to obtain samples by aggregating data across single cells.** Single BJ fibroblast (**a**) and primary breast epithelial cells (**b**) were sequenced following library preparation on the 10x Genomics Single-Cell CNV platform. G_1- and S-phase cells were identified by *in silico* cell sorting and read were aggregated across cells to perform S/G_1 replication timing inference. BJ fibroblast Repli-seq data are from Hansen et al.[44]. MCF-7 S/G_1 data are from **Chapter 4**.

Conclusions

Together, the work presented in this dissertation extends our understanding of human replication timing variability in two directions and open up exciting possibilities for future research. In **Chapters 2** and **3**, I demonstrated the replication timing of centromeres consistently differs between cell lines, in a manner that may suggest coordinated replication of these regions. Further work with additional cell lines is needed to understand the causes and consequences of this variability. In **Chapter 4**, I developed novel methods for studying replication timing in single cells and showed that origin firing time, but not origin location, is likely shaped by stochasticity. However, existing models of stochastic origin firing fail to fully account for the heterogeneity of behaviors observed in late-replicating regions. Future work on this project should involve a more local analysis of the cancer cell lines, as well as the development of a new model of single-cell replication initiation that can explain more of the variability observed.

REFERENCES

1 Hulke, M. L., Massey, D. J. & Koren, A. Genomic methods for measuring DNA replication dynamics. *Chromosome Res* **28**, 49-67, doi:10.1007/s10577-019-09624-y (2020).

2 Masai, H., Matsumoto, S., You, Z., Yoshizawa-Sugata, N. & Oda, M. Eukaryotic chromosome DNA replication: where, when, and how? *Annu Rev Biochem* **79**, 89-130, doi:10.1146/annurev.biochem.052308.103205 (2010).

3 Fragkos, M., Ganier, O., Coulombe, P. & Mechali, M. DNA replication origin activation in space and time. *Nat Rev Mol Cell Biol* **16**, 360-374, doi:10.1038/nrm4002 (2015).

4 Bell, S. P. & Stillman, B. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* **357**, 128-134, doi:10.1038/357128a0 (1992).

5 Gavin, K. A., Hidaka, M. & Stillman, B. Conserved initiator proteins in eukaryotes. *Science* **270**, 1667-1671, doi:10.1126/science.270.5242.1667 (1995).

6 Hartwell, L. H. Sequential function of gene products relative to DNA synthesis in the yeast cell cycle. *J Mol Biol* **104**, 803-817, doi:10.1016/0022-2836(76)90183-2 (1976).

7 Cocker, J. H., Piatti, S., Santocanale, C., Nasmyth, K. & Diffley, J. F. An essential role for the Cdc6 protein in forming the pre-replicative complexes of budding yeast. *Nature* **379**, 180-182, doi:10.1038/379180a0 (1996).

8 Mizushima, T., Takahashi, N. & Stillman, B. Cdc6p modulates the structure and DNA binding activity of the origin recognition complex in vitro. *Genes Dev* **14**, 1631-1641 (2000).

9 Nishitani, H., Lygerou, Z., Nishimoto, T. & Nurse, P. The Cdt1 protein is required to license DNA for replication in fission yeast. *Nature* **404**, 625-628, doi:10.1038/35007110 (2000).

10 Maiorano, D., Moreau, J. & Mechali, M. XCDT1 is required for the assembly of pre-replicative complexes in Xenopus laevis. *Nature* **404**, 622-625, doi:10.1038/35007104 (2000).

11 Donovan, S., Harwood, J., Drury, L. S. & Diffley, J. F. Cdc6p-dependent loading of Mcm proteins onto pre-replicative chromatin in budding yeast. *Proc Natl Acad Sci U S A* **94**, 5611-5616, doi:10.1073/pnas.94.11.5611 (1997).

12 Tanaka, T., Knapp, D. & Nasmyth, K. Loading of an Mcm protein onto DNA replication origins is regulated by Cdc6p and CDKs. *Cell* **90**, 649-660, doi:10.1016/s0092-8674(00)80526-7 (1997).

13 Liang, C. & Stillman, B. Persistent initiation of DNA replication and chromatin-bound MCM proteins during the cell cycle in cdc6 mutants. *Genes Dev* **11**, 3375-3386, doi:10.1101/gad.11.24.3375 (1997).

14 Remus, D. *et al.* Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139**, 719-730, doi:10.1016/j.cell.2009.10.015 (2009).

15 Evrin, C. *et al.* A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proc Natl Acad Sci U S A* **106**, 20240-20245, doi:10.1073/pnas.0911500106 (2009).

16 Miller, T. C. R., Locke, J., Greiwe, J. F., Diffley, J. F. X. & Costa, A. Mechanism of head-to-head MCM double-hexamer formation revealed by cryo-EM. *Nature* **575**, 704-710, doi:10.1038/s41586-019-1768-0 (2019).

17 Moyer, S. E., Lewis, P. W. & Botchan, M. R. Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc Natl Acad Sci U S A* **103**, 10236-10241, doi:10.1073/pnas.0602400103 (2006).

18 Aparicio, T., Guillou, E., Coloma, J., Montoya, G. & Mendez, J. The human GINS complex associates with Cdc45 and MCM and is essential for DNA replication. *Nucleic Acids Res* **37**, 2087-2095, doi:10.1093/nar/gkp065 (2009).

19 Ilves, I., Petojevic, T., Pesavento, J. J. & Botchan, M. R. Activation of the MCM2-7 helicase by association with Cdc45 and GINS proteins. *Mol Cell* **37**, 247-258, doi:10.1016/j.molcel.2009.12.030 (2010).

20 Oshiro, G., Owens, J. C., Shellman, Y., Sclafani, R. A. & Li, J. J. Cell cycle control of Cdc7p kinase activity through regulation of Dbf4p stability. *Mol Cell Biol* **19**, 4888-4896, doi:10.1128/MCB.19.7.4888 (1999).

21 Cheng, L., Collyer, T. & Hardy, C. F. Cell cycle regulation of DNA replication initiator factor Dbf4p. *Mol Cell Biol* **19**, 4270-4278, doi:10.1128/MCB.19.6.4270 (1999).

22 Kuhne, C. & Linder, P. A new pair of B-type cyclins from Saccharomyces cerevisiae that function early in the cell cycle. *EMBO J* **12**, 3437-3447 (1993).

23 Heller, R. C. *et al.* Eukaryotic origin-dependent DNA replication in vitro reveals sequential action of DDK and S-CDK kinases. *Cell* **146**, 80-91, doi:10.1016/j.cell.2011.06.012 (2011).

24 Tanaka, S., Nakato, R., Katou, Y., Shirahige, K. & Araki, H. Origin association of Sld3, Sld7, and Cdc45 proteins is a key step for determination of origin-firing timing. *Curr Biol* **21**, 2055-2063, doi:10.1016/j.cub.2011.11.038 (2011).

25 Muramatsu, S., Hirai, K., Tak, Y. S., Kamimura, Y. & Araki, H. CDK-dependent complex formation between replication proteins Dpb11, Sld2, Pol (epsilon}, and GINS in budding yeast. *Genes Dev* **24**, 602-612, doi:10.1101/gad.1883410 (2010).

26 Boos, D. *et al.* Regulation of DNA replication through Sld3-Dpb11 interaction is conserved from yeast to humans. *Curr Biol* **21**, 1152-1157, doi:10.1016/j.cub.2011.05.057 (2011).

27 Schmidt, U. *et al.* Characterization of the interaction between the human DNA topoisomerase IIbeta-binding protein 1 (TopBP1) and the cell division cycle 45 (Cdc45) protein. *Biochem J* **409**, 169-177, doi:10.1042/BJ20070872 (2008).

28 Sangrithi, M. N. *et al.* Initiation of DNA replication requires the RECQL4 protein mutated in Rothmund-Thomson syndrome. *Cell* **121**, 887-898, doi:10.1016/j.cell.2005.05.015 (2005).

29 Xu, X., Rochette, P. J., Feyissa, E. A., Su, T. V. & Liu, Y. MCM10 mediates RECQ4 association with MCM2-7 helicase complex during DNA replication. *EMBO J* **28**, 3005-3014, doi:10.1038/emboj.2009.235 (2009).

30 Tanaka, S. & Araki, H. Regulation of the initiation step of DNA replication by cyclin-dependent kinases. *Chromosoma* **119**, 565-574, doi:10.1007/s00412-010-0291-8 (2010).

31 Stinchcomb, D. T., Struhl, K. & Davis, R. W. Isolation and characterisation of a yeast chromosomal replicator. *Nature* **282**, 39-43, doi:10.1038/282039a0 (1979).

32 Huberman, J. A., Spotila, L. D., Nawotka, K. A., el-Assouli, S. M. & Davis, L. R. The in vivo replication origin of the yeast 2 microns plasmid. *Cell* **51**, 473-481, doi:10.1016/0092-8674(87)90643-x (1987).

33 Brewer, B. J. & Fangman, W. L. The localization of replication origins on ARS plasmids in S. cerevisiae. *Cell* **51**, 463-471, doi:10.1016/0092-8674(87)90642-8 (1987).

34 Rao, H., Marahrens, Y. & Stillman, B. Functional conservation of multiple elements in yeast chromosomal replicators. *Mol Cell Biol* **14**, 7643-7651, doi:10.1128/mcb.14.11.7643-7651.1994 (1994).

35 Broach, J. R. *et al.* Localization and sequence analysis of yeast origins of DNA replication. *Cold Spring Harb Symp Quant Biol* **47 Pt 2**, 1165-1173, doi:10.1101/sqb.1983.047.01.132 (1983).

36 Deshpande, A. M. & Newlon, C. S. The ARS consensus sequence is required for chromosomal origin function in Saccharomyces cerevisiae. *Mol Cell Biol* **12**, 4305-4313, doi:10.1128/mcb.12.10.4305-4313.1992 (1992).

37 Li, N. *et al.* Structure of the origin recognition complex bound to DNA replication origin. *Nature* **559**, 217-222, doi:10.1038/s41586-018-0293-x (2018).

38 Heinzel, S. S., Krysan, P. J., Tran, C. T. & Calos, M. P. Autonomous DNA replication in human cells is affected by the size and the source of the DNA. *Mol Cell Biol* **11**, 2263-2272, doi:10.1128/mcb.11.4.2263-2272.1991 (1991).

39 Krysan, P. J., Smith, J. G. & Calos, M. P. Autonomous replication in human cells of multimers of specific human and bacterial DNA sequences. *Mol Cell Biol* **13**, 2688-2696, doi:10.1128/mcb.13.5.2688-2696.1993 (1993).

40 Kirstein, N. *et al.* Human ORC/MCM density is low in active genes and correlates with replication time but does not delimit initiation zones. *Elife* **10**, doi:10.7554/eLife.62161 (2021).

41 Ge, X. Q., Jackson, D. A. & Blow, J. J. Dormant origins licensed by excess Mcm2-7 are required for human cells to survive replicative stress. *Genes Dev* **21**, 3331-3341, doi:10.1101/gad.457807 (2007).

42 Ibarra, A., Schwob, E. & Mendez, J. Excess MCM proteins protect human cells from replicative stress by licensing backup origins of replication. *Proc Natl Acad Sci U S A* **105**, 8956-8961, doi:10.1073/pnas.0803978105 (2008).

43 Raghuraman, M. K. *et al.* Replication dynamics of the yeast genome. *Science* **294**, 115-121, doi:10.1126/science.294.5540.115 (2001).

44 Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**, 139-144, doi:10.1073/pnas.0912402107 (2010).

45 Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* **91**, 1033-1040, doi:10.1016/j.ajhg.2012.10.018 (2012).

46 Ding, Q. *et al.* The genetic architecture of DNA replication timing in human pluripotent stem cells. *Nat Commun* **12**, 6746, doi:10.1038/s41467-021-27115-9 (2021).

47 Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res* **20**, 761-770, doi:10.1101/gr.099655.109 (2010).

48 Hiratani, I. *et al.* Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol* **6**, e245, doi:10.1371/journal.pbio.0060245 (2008).

49 Hiratani, I. *et al.* Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res* **20**, 155-169, doi:10.1101/gr.099796.109 (2010).

50 Miura, H. *et al.* Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization. *Nat Genet* **51**, 1356-1368, doi:10.1038/s41588-019-0474-z (2019).

51 Yaffe, E. *et al.* Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet* **6**, e1001011, doi:10.1371/journal.pgen.1001011 (2010).

52 Woodfine, K. *et al.* Replication timing of the human genome. *Hum Mol Genet* **13**, 191-202, doi:10.1093/hmg/ddh016 (2004).

53 MacAlpine, D. M., Rodriguez, H. K. & Bell, S. P. Coordination of replication and transcription along a Drosophila chromosome. *Genes Dev* **18**, 3094-3105, doi:10.1101/gad.1246404 (2004).

54 Jeon, Y. *et al.* Temporal profile of replication of human chromosomes. *Proc Natl Acad Sci U S A* **102**, 6419-6424, doi:10.1073/pnas.0405088102 (2005).

55 Koren, A. DNA replication timing: Coordinating genome stability with genome regulation on the X chromosome and beyond. *Bioessays* **36**, 997-1004, doi:10.1002/bies.201400077 (2014).

56 Rivera-Mulia, J. C. *et al.* Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. *Genome Res* **25**, 1091-1103, doi:10.1101/gr.187989.114 (2015).

57 Stamatoyannopoulos, J. A. *et al.* Human mutation rate associated with DNA replication timing. *Nat Genet* **41**, 393-395, doi:10.1038/ng.363 (2009).

58 Yehuda, Y. *et al.* Germline DNA replication timing shapes mammalian genome composition. *Nucleic Acids Res* **46**, 8299-8310, doi:10.1093/nar/gky610 (2018).

59 Klein, K. N. *et al.* Replication timing maintains the global epigenetic state in human cells. *Science* **372**, 371-378, doi:10.1126/science.aba5545 (2021).

60 Caballero, M. *et al.* Comprehensive analysis of DNA replication timing in genetic diseases and gene knockouts identifies MCM10 as a novel regulator of the replication program. *bioRxiv*, 2021.2009.2008.459433, doi:10.1101/2021.09.08.459433 (2021).

61 Marchal, C. *et al.* Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nat Protoc* **13**, 819-839, doi:10.1038/nprot.2017.148 (2018).

62 Zhao, P. A., Sasaki, T. & Gilbert, D. M. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol* **21**, 76, doi:10.1186/s13059-020-01983-8 (2020).

63 Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat Genet* **47**, 296-303, doi:10.1038/ng.3200 (2015).

64 Koren, A. *et al.* Genetic variation in human DNA replication timing. *Cell* **159**, 1015-1026, doi:10.1016/j.cell.2014.10.025 (2014).

65 Koren, A., Massey, D. J. & Bracci, A. N. TIGER: inferring DNA replication timing from whole-genome sequence data. *Bioinformatics*, doi:10.1093/bioinformatics/btab166 (2021).

66 Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**, e72, doi:10.1093/nar/gks001 (2012).

67 Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* **36**, 321-323, doi:10.1038/nbt.4109 (2018).

68 Miga, K. H. *et al.* Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res* **24**, 697-707, doi:10.1101/gr.159624.113 (2014).

69 Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *bioRxiv*, 2021.2007.2012.452052, doi:10.1101/2021.07.12.452052 (2021).

70 Nurk, S. *et al.* The complete sequence of a human genome. *bioRxiv*, 2021.2005.2026.445798, doi:10.1101/2021.05.26.445798 (2021).

71 Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43**, D670-681, doi:10.1093/nar/gku1177 (2015).

72 Bensimon, A. *et al.* Alignment and sensitive detection of DNA by a moving interface. *Science* **265**, 2096-2098, doi:10.1126/science.7522347 (1994).

73 Michalet, X. *et al.* Dynamic molecular combing: stretching the whole human genome for high-resolution studies. *Science* **277**, 1518-1523, doi:10.1126/science.277.5331.1518 (1997).

74 Norio, P. & Schildkraut, C. L. Visualization of DNA replication on individual Epstein-Barr virus episomes. *Science* **294**, 2361-2364, doi:10.1126/science.1064603 (2001).

75 Huberman, J. A. & Riggs, A. D. On the mechanism of DNA replication in mammalian chromosomes. *J Mol Biol* **32**, 327-341, doi:10.1016/0022-2836(68)90013-2 (1968).

76 Blow, J. J., Gillespie, P. J., Francis, D. & Jackson, D. A. Replication origins in Xenopus egg extract Are 5-15 kilobases apart and are activated in clusters that fire at different times. *J Cell Biol* **152**, 15-25, doi:10.1083/jcb.152.1.15 (2001).

77 Herrick, J., Stanislawski, P., Hyrien, O. & Bensimon, A. Replication fork density increases during DNA synthesis in X. laevis egg extracts. *J Mol Biol* **300**, 1133-1142, doi:10.1006/jmbi.2000.3930 (2000).

78 Labit, H., Perewoska, I., Germe, T., Hyrien, O. & Marheineke, K. DNA replication timing is deterministic at the level of chromosomal domains but stochastic at the level of replicons in Xenopus egg extracts. *Nucleic Acids Res* **36**, 5623-5634, doi:10.1093/nar/gkn533 (2008).

79 Czajkowsky, D. M., Liu, J., Hamlin, J. L. & Shao, Z. DNA combing reveals intrinsic temporal disorder in the replication of yeast chromosome VI. *J Mol Biol* **375**, 12-19, doi:10.1016/j.jmb.2007.10.046 (2008).

80 Patel, P. K., Arcangioli, B., Baker, S. P., Bensimon, A. & Rhind, N. DNA replication origins fire stochastically in fission yeast. *Mol Biol Cell* **17**, 308-316, doi:10.1091/mbc.e05-07-0657 (2006).

81 Cayrou, C. *et al.* Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res* **21**, 1438-1449, doi:10.1101/gr.121830.111 (2011).

82 Norio, P. *et al.* Progressive activation of DNA replication initiation in large domains of the immunoglobulin heavy chain locus during B cell development. *Mol Cell* **20**, 575-587, doi:10.1016/j.molcel.2005.10.029 (2005).

83 Piunti, A. *et al.* Polycomb proteins control proliferation and transformation independently of cell cycle checkpoints by regulating DNA replication. *Nat Commun* **5**, 3649, doi:10.1038/ncomms4649 (2014).

84 Bester, A. C. *et al.* Nucleotide deficiency promotes genomic instability in early stages of cancer development. *Cell* **145**, 435-446, doi:10.1016/j.cell.2011.03.044 (2011).

85 Frum, R. A., Khondker, Z. S. & Kaufman, D. G. Temporal differences in DNA replication during the S phase using single fiber analysis of normal human fibroblasts and glioblastoma T98G cells. *Cell Cycle* **8**, 3133-3148, doi:10.4161/cc.8.19.9682 (2009).

86 Guilbaud, G. *et al.* Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. *PLoS Comput Biol* **7**, e1002322, doi:10.1371/journal.pcbi.1002322 (2011).

87 Lamm, N. *et al.* Folate levels modulate oncogene-induced replication stress and tumorigenicity. *EMBO Mol Med* **7**, 1138-1152, doi:10.15252/emmm.201404824 (2015).

88 Rhind, N., Yang, S. C. & Bechhoefer, J. Reconciling stochastic origin firing with defined replication timing. *Chromosome Res* **18**, 35-43, doi:10.1007/s10577-009-9093-3 (2010).

89 Yang, S. C., Rhind, N. & Bechhoefer, J. Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Mol Syst Biol* **6**, 404, doi:10.1038/msb.2010.61 (2010).

90 Lebofsky, R., Heilig, R., Sonnleitner, M., Weissenbach, J. & Bensimon, A. DNA replication origin interference increases the spacing between initiation events in human cells. *Mol Biol Cell* **17**, 5337-5345, doi:10.1091/mbc.e06-04-0298 (2006).

91 Pasero, P., Bensimon, A. & Schwob, E. Single-molecule analysis reveals clustering and epigenetic regulation of replication origins at the yeast rDNA locus. *Genes Dev* **16**, 2479-2484, doi:10.1101/gad.232902 (2002).

92 De Carli, F., Gaggioli, V., Millot, G. A. & Hyrien, O. Single-molecule, antibody-free fluorescent visualisation of replication tracts along barcoded DNA molecules. *Int J Dev Biol* **60**, 297-304, doi:10.1387/ijdb.160139oh (2016).

93 Lacroix, J. *et al.* Analysis of DNA Replication by Optical Mapping in Nanochannels. *Small* **12**, 5963-5970, doi:10.1002/smll.201503795 (2016).

94 Xiao, M. *et al.* Rapid DNA mapping by fluorescent single molecule detection. *Nucleic Acids Res* **35**, e16, doi:10.1093/nar/gkl1044 (2007).

95 De Carli, F. *et al.* High-Throughput Optical Mapping of Replicating DNA. *Small Methods* **2**, 1800146, doi:10.1002/smtd.201800146 (2018).

96 Wang, W. *et al.* Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. *Mol Cell* **81**, 2975-2988 e2976, doi:10.1016/j.molcel.2021.05.024 (2021).

97 Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**, 265-270, doi:10.1038/nnano.2009.12 (2009).

98 Hennion, M. *et al.* FORK-seq: replication landscape of the Saccharomyces cerevisiae genome by nanopore sequencing. *Genome Biol* **21**, 125, doi:10.1186/s13059-020-02013-3 (2020).

99 Muller, C. A. *et al.* Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nat Methods* **16**, 429-436, doi:10.1038/s41592-019-0394-y (2019).

100 Georgieva, D., Liu, Q., Wang, K. & Egli, D. Detection of base analogs incorporated during DNA replication by nanopore sequencing. *Nucleic Acids Res* **48**, e88, doi:10.1093/nar/gkaa517 (2020).

101 Blainey, P. C. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* **37**, 407-427, doi:10.1111/1574-6976.12015 (2013).

102 Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175-188, doi:10.1038/nrg.2015.16 (2016).

103 Arneson, N., Hughes, S., Houlston, R. & Done, S. Whole-Genome Amplification by Degenerate Oligonucleotide Primed PCR (DOP-PCR). *CSH Protoc* **2008**, pdb prot4919, doi:10.1101/pdb.prot4919 (2008).

104 Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718-725, doi:10.1016/0888-7543(92)90147-k (1992).

105 Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622-1626, doi:10.1126/science.1229164 (2012).

106 Chen, C. *et al.* Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **356**, 189-194, doi:10.1126/science.aak9787 (2017).

107 Van der Aa, N. *et al.* Genome-wide copy number profiling of single cells in S-phase reveals DNA-replication domains. *Nucleic Acids Res* **41**, e66, doi:10.1093/nar/gks1352 (2013).

108 Dileep, V. & Gilbert, D. M. Single-cell replication profiling to measure stochastic variation in mammalian replication timing. *Nat Commun* **9**, 427, doi:10.1038/s41467-017-02800-w (2018).

109 Takahashi, S. *et al.* Genome-wide stability of the DNA replication program in single mammalian cells. *Nat Genet* **51**, 529-540, doi:10.1038/s41588-019-0347-5 (2019).

110 Vitak, S. A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* **14**, 302-308, doi:10.1038/nmeth.4154 (2017).

111 Yin, Y. *et al.* High-Throughput Single-Cell Sequencing with Linear Amplification. *Mol Cell* **76**, 676-690 e610, doi:10.1016/j.molcel.2019.08.002 (2019).

112 Zahn, H. *et al.* Scalable whole-genome single-cell library preparation without preamplification. *Nat Methods* **14**, 167-173, doi:10.1038/nmeth.4140 (2017).

113 Bell, A. D. *et al.* Insights into variation in meiosis from 31,228 human sperm genomes. *Nature* **583**, 259-264, doi:10.1038/s41586-020-2347-0 (2020).

114 Laks, E. *et al.* Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell* **179**, 1207-1221 e1222, doi:10.1016/j.cell.2019.10.026 (2019).

115 Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94, doi:10.1038/nature09807 (2011).

116 Gonzalez-Pena, V. *et al.* Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2024176118 (2021).

117 Zaccaria, S. & Raphael, B. J. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat Biotechnol* **39**, 207-214, doi:10.1038/s41587-020-0661-6 (2021).

118 Massey, D. J., Kim, D., Brooks, K. E., Smolka, M. B. & Koren, A. Next-Generation Sequencing Enables Spatiotemporal Resolution of Human Centromere Replication Timing. *Genes (Basel)* **10**, doi:10.3390/genes10040269 (2019).

119 Farkash-Amar, S. *et al.* Global organization of replication time zones of the mouse genome. *Genome Res* **18**, 1562-1570, doi:10.1101/gr.079566.108 (2008).

120 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

121 Gilbert, D. M. Replication timing and transcriptional control: beyond cause and effect. *Curr Opin Cell Biol* **14**, 377-383, doi:10.1016/s0955-0674(02)00326-5 (2002).

122 Rhind, N. & Gilbert, D. M. DNA replication timing. *Cold Spring Harb Perspect Biol* **5**, a010132, doi:10.1101/cshperspect.a010132 (2013).

123 Sequeira-Mendes, J. & Gutierrez, C. Links between genome replication and chromatin landscapes. *Plant J* **83**, 38-51, doi:10.1111/tpj.12847 (2015).

124 Fu, H., Baris, A. & Aladjem, M. I. Replication timing and nuclear structure. *Curr Opin Cell Biol* **52**, 43-50, doi:10.1016/j.ceb.2018.01.004 (2018).

125 Ferguson, B. M. & Fangman, W. L. A position effect on the time of replication origin activation in yeast. *Cell* **68**, 333-339, doi:10.1016/0092-8674(92)90474-q (1992).

126 Stevenson, J. B. & Gottschling, D. E. Telomeric chromatin modulates replication timing near chromosome ends. *Genes Dev* **13**, 146-151, doi:10.1101/gad.13.2.146 (1999).

127 Zappulla, D. C., Sternglanz, R. & Leatherwood, J. Control of replication timing by a transcriptional silencer. *Curr Biol* **12**, 869-875, doi:10.1016/s0960-9822(02)00871-0 (2002).

128 Casas-Delucchi, C. S. *et al.* Histone hypoacetylation is required to maintain late replication timing of constitutive heterochromatin. *Nucleic Acids Res* **40**, 159-169, doi:10.1093/nar/gkr723 (2012).

129 Heinz, K. S. *et al.* Peripheral re-localization of constitutive heterochromatin advances its replication timing and impairs maintenance of silencing marks. *Nucleic Acids Res* **46**, 6112-6128, doi:10.1093/nar/gky368 (2018).

130 Kim, S. M. & Huberman, J. A. Regulation of replication timing in fission yeast. *EMBO J* **20**, 6115-6126, doi:10.1093/emboj/20.21.6115 (2001).

131 Kim, S. M., Dubey, D. D. & Huberman, J. A. Early-replicating heterochromatin. *Genes Dev* **17**, 330-335, doi:10.1101/gad.1046203 (2003).

132 Koren, A. *et al.* Epigenetically-inherited centromere and neocentromere DNA replicates earliest in S-phase. *PLoS Genet* **6**, e1001068, doi:10.1371/journal.pgen.1001068 (2010).

133 Li, P. C., Chretien, L., Cote, J., Kelly, T. J. & Forsburg, S. L. S. pombe replication protein Cdc18 (Cdc6) interacts with Swi6 (HP1) heterochromatin protein: region specific effects and replication timing in the centromere. *Cell Cycle* **10**, 323-336, doi:10.4161/cc.10.2.14552 (2011).

134 Hayashi, M. T., Takahashi, T. S., Nakagawa, T., Nakayama, J. & Masukata, H. The heterochromatin protein Swi6/HP1 activates replication origins at the pericentromeric region and silent mating-type locus. *Nat Cell Biol* **11**, 357-362, doi:10.1038/ncb1845 (2009).

135 Ahmad, K. & Henikoff, S. Centromeres are specialized replication domains in heterochromatin. *J Cell Biol* **153**, 101-110, doi:10.1083/jcb.153.1.101 (2001).

136 O'Keefe, R. T., Henderson, S. C. & Spector, D. L. Dynamic organization of DNA replication in mammalian cell nuclei: spatially and temporally defined replication of chromosome-specific alpha-satellite DNA sequences. *J Cell Biol* **116**, 1095-1110, doi:10.1083/jcb.116.5.1095 (1992).

137 Sullivan, B. & Karpen, G. Centromere identity in Drosophila is not determined in vivo by replication timing. *J Cell Biol* **154**, 683-690, doi:10.1083/jcb.200103001 (2001).

138 Weidtkamp-Peters, S., Rahn, H. P., Cardoso, M. C. & Hemmerich, P. Replication of centromeric heterochromatin in mouse fibroblasts takes place in early, middle, and late S phase. *Histochem Cell Biol* **125**, 91-102, doi:10.1007/s00418-005-0063-3 (2006).

139 Ten Hagen, K. G., Gilbert, D. M., Willard, H. F. & Cohen, S. N. Replication timing of DNA sequences associated with human centromeres and telomeres. *Mol Cell Biol* **10**, 6348-6355, doi:10.1128/mcb.10.12.6348-6355.1990 (1990).

140 Watanabe, Y., Kazuki, Y., Oshimura, M., Ikemura, T. & Maekawa, M. Replication timing in a single human chromosome 11 transferred into the Chinese hamster ovary (CHO) cell line. *Gene* **510**, 1-6, doi:10.1016/j.gene.2012.08.045 (2012).

141 Erliandri, I. *et al.* Replication of alpha-satellite DNA arrays in endogenous human centromeric regions and in human artificial chromosome. *Nucleic Acids Res* **42**, 11502-11516, doi:10.1093/nar/gku835 (2014).

142 Nag, A. *et al.* Chromatin signature of widespread monoallelic expression. *Elife* **2**, e01256, doi:10.7554/eLife.01256 (2013).

143 Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* **28**, 581-591, doi:10.1101/gr.221028.117 (2018).

144 Wear, E. E. *et al.* Genomic Analysis of the DNA Replication Timing Program during Mitotic S Phase in Maize (Zea mays) Root Tips. *Plant Cell* **29**, 2126-2149, doi:10.1105/tpc.17.00037 (2017).

145 McNulty, S. M., Sullivan, L. L. & Sullivan, B. A. Human Centromeres Produce Chromosome-Specific and Array-Specific Alpha Satellite Transcripts that Are Complexed with CENP-A and CENP-C. *Dev Cell* **42**, 226-240 e226, doi:10.1016/j.devcel.2017.07.001 (2017).

146 Massey, D. J. & Koren, A. Telomere-to-telomere human DNA replication timing profiles. *bioRxiv*, 2022.2003.2028.486072, doi:10.1101/2022.03.28.486072 (2022).

147 Dimitrova, D. S. & Gilbert, D. M. The spatial position and replication timing of chromosomal domains are both established in early G1 phase. *Mol Cell* **4**, 983-993, doi:10.1016/s1097-2765(00)80227-0 (1999).

148 Rivera-Mulia, J. C. *et al.* Allele-specific control of replication timing and genome organization during development. *Genome Res* **28**, 800-811, doi:10.1101/gr.232561.117 (2018).

149 Du, Q. *et al.* DNA methylation is required to maintain both DNA replication timing precision and 3D genome organization integrity. *Cell Rep* **36**, 109722, doi:10.1016/j.celrep.2021.109722 (2021).

150 Goren, A., Tabib, A., Hecht, M. & Cedar, H. DNA replication timing of the human beta-globin domain is controlled by histone modification at the origin. *Genes Dev* **22**, 1319-1324, doi:10.1101/gad.468308 (2008).

151 McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res* **26**, 115-138, doi:10.1007/s10577-018-9582-3 (2018).

152 Aldrup-MacDonald, M. E., Kuo, M. E., Sullivan, L. L., Chew, K. & Sullivan, B. A. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. *Genome Res* **26**, 1301-1311, doi:10.1101/gr.206706.116 (2016).

153 Giunta, S. *et al.* CENP-A chromatin prevents replication stress at centromeres to avoid structural aneuploidy. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2015634118 (2021).

154 Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363-367, doi:10.1038/nature08973 (2010).

155 Wiblin, A. E., Cui, W., Clark, A. J. & Bickmore, W. A. Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells. *J Cell Sci* **118**, 3861-3868, doi:10.1242/jcs.02500 (2005).

156 Massey, D. J. & Koren, A. High-throughput analysis of single human cells reveals the complex nature of DNA replication timing control. *bioRxiv*, 2021.2005.2014.443897, doi:10.1101/2021.05.14.443897 (2022).

157 Hulke, M. L., Massey, D. J. & Koren, A. Genomic methods for measuring DNA replication dynamics. *Chromosome Res*, doi:10.1007/s10577-019-09624-y (2019).

158 Pope, B. D., Hiratani, I. & Gilbert, D. M. Domain-wide regulation of DNA replication timing during mammalian development. *Chromosome Res* **18**, 127-136, doi:10.1007/s10577-009-9100-8 (2010).

159 Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* **12**, 1058-1060, doi:10.1038/nmeth.3578 (2015).

160 Velazquez-Villarreal, E. I. *et al.* Single-cell sequencing of genomic DNA resolves sub-clonal heterogeneity in a melanoma cell line. *Commun Biol* **3**, 318, doi:10.1038/s42003-020-1044-8 (2020).

161 Besnard, E. *et al.* Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat Struct Mol Biol* **19**, 837-844, doi:10.1038/nsmb.2339 (2012).

162 Cadoret, J. C. *et al.* Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci U S A* **105**, 15837-15842, doi:10.1073/pnas.0805208105 (2008).

163 Langley, A. R., Graf, S., Smith, J. C. & Krude, T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res* **44**, 10230-10247, doi:10.1093/nar/gkw760 (2016).

164 Mesner, L. D. *et al.* Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. *Genome Res* **23**, 1774-1788, doi:10.1101/gr.155218.113 (2013).

165 Chen, Y. H. *et al.* Transcription shapes DNA replication initiation and termination in human cells. *Nat Struct Mol Biol* **26**, 67-77, doi:10.1038/s41594-018-0171-0 (2019).

166 Petryk, N. *et al.* Replication landscape of the human genome. *Nat Commun* **7**, 10208, doi:10.1038/ncomms10208 (2016).

167 Gnan, S. *et al.* Kronos scRT: a uniform framework for single-cell replication timing analysis. *bioRxiv*, 2021.2009.2001.458599, doi:10.1101/2021.09.01.458599 (2021).

168 Boulos, R. E., Drillon, G., Argoul, F., Arneodo, A. & Audit, B. Structural organization of human replication timing domains. *FEBS Lett* **589**, 2944-2957, doi:10.1016/j.febslet.2015.04.015 (2015).

169 Tubbs, A. *et al.* Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. *Cell* **174**, 1127-1142 e1119, doi:10.1016/j.cell.2018.07.011 (2018).

170 Minussi, D. C. *et al.* Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature* **592**, 302-308, doi:10.1038/s41586-021-03357-x (2021).

171 Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* **27**, 157-164, doi:10.1101/gr.210500.116 (2017).

172 Pereira, C. *et al.* Sequencing Micronuclei Reveals the Landscape of Chromosomal Instability. *bioRxiv*, 2021.2010.2028.466311, doi:10.1101/2021.10.28.466311 (2021).