THE EVOLUTIONARY DYNAMICS AND IMPACT OF 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

Alexa Nicole Bracci

December 2022

THE EVOLUTIONARY DYNAMICS AND IMPACT OF DNA REPLICATION TIMING

Alexa Nicole Bracci, Ph. D.

Cornell University 2022

DNA replication occurs in a defined spatiotemporal order, which is mediated by the firing of origins across the genome at different times during S phase. Replication timing is associated with gene regulation and influences mutation rates across the genome, but the evolutionary forces shaping replication timing are largely unknown. In this dissertation, I profiled DNA replication timing from humans, chimpanzees, and rhesus macaques to study the causes and consequences of replication timing evolution. I found that hundreds of regions vary in replication timing within and between the genomes of humans and chimpanzees, and over a hundred regions were classified as changes that occurred specifically in the human lineage. Importantly, human-chimpanzee variants highly overlap regions of within species variation, which points to the presence of ongoing evolution. Replication timing variation was also found to be correlated with regulatory evolution (e.g. gene expression and chromatin structure) and had elevated levels of sequence divergence. Finally, linking genetic variation to replication timing variation within humans and chimpanzees facilitated the identification of sequence determinants of replication timing evolution. Overall, DNA replication timing shows ongoing evolution in the human lineage, at least in part driven by sequence alterations, and with important implications for regulatory and sequence evolution.

BIOGRAPHICAL SKETCH

Alexa Nicole Bracci was born in Buffalo, NY. She attended St. Christopher Roman Catholic School from 2000-2009, where she first recalls learning about genetics in seventh grade science. She was intrigued by the DNA that is packaged into each one of our cells and codes for physical traits, like attached earlobes and cleft chin. She then attended Clarence High School from 2009-2013, where her interest in biology remained.

She was then accepted into the State University of New York at Buffalo Honors College with a major in Biological Sciences. Her first research experience was in the lab of Dr. Omer Gokcumen, where she contributed to two research projects over two years studying structural genomic variation and its role in human evolution. The first investigated the evolutionary genomics of duplicated histatin genes, and the second involved implementing a target capture protocol to sequence repetitive genomic regions of interest. Alexa also participated in the Research Experience for Undergraduates program at the University of Cincinnati during the summer of 2016. There, she worked with Dr. Stephanie Rollmann on a project studying olfactory behavioral genetics with the model organism, *Drosophila*. Alexa also participated in the Biological Sciences honors program during her final year at the University at Buffalo, and graduated Summa Cum Laude with highest honors in May 2017.

Alexa was then accepted to the Genetics, Genomics and Development Ph.D. program at Cornell University starting in August 2017. There, she worked in the lab of Dr. Amnon Koren studying the evolution of DNA replication timing across humans and chimpanzees.

To my family.

ACKNOWLEDGMENTS

I would first like to thank my Ph.D. advisor, Dr. Amnon Koren, for supporting and guiding me throughout my time at Cornell. Thank you for training me in computational research, and for constructive and thorough feedback on analysis and writing. I would also like to acknowledge the rest of my dissertation committee, Dr. Andrew Clark, Dr. Charles Danko and Dr. Cedric Feschotte. Thank you for serving on my committee and providing valuable feedback on my research.

I would also like to thank current and previous Koren lab members for research feedback, brainstorming sessions, and coffee runs. Thank you also to those who contributed to this project: Madison Caballero, Andy Ding, Melissa Hubisz, and especially Anissa Dallmann, a previous Cornell undergraduate student who worked with me for three years.

I would not have decided to do my Ph.D. without the guidance of Dr. Omer Gokcumen and Dr. Jessica Poulin from the University at Buffalo. Thank you for your support and mentorship when I was first learning about research.

Thank you to my friends, both from Buffalo and the ones that I made at Cornell. The many lunches, walks, and trips were valued breaks from work. I am also thankful for the support from my family, including my parents, siblings, and future in-laws. Finally, thank you to my fiancé, Tom Siskar, who has been my best friend for over ten years, and has supported and encouraged me throughout this Ph.D.

My dissertation work was partially supported by the Cornell Center for Vertebrate Genomics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 2:

LIST OF ABBREVIATIONS

1KG	1000 Genomes
AFR	African super-population
ARS	Autonomously replicating sequence
ATAC-seq	Assay for transposase accessible chromatin followed with sequencing
BrdU	Bromodeoxyuridine
ChIP-seq	Chromatin immunoprecipitation followed with sequencing
Chr	Chromosome
CMG	Cdc45-MCM2-7-GINS complex
CNV	Copy number variant
df	Degrees of freedom
DNM	De novo mutation
EAS	East Asian super-population
EPO	Enredo-Pecan-Ortheus
eQTL	Gene expression quantitative trait locus
ESC	Embryonic stem cell line
EUR	European super-population
FC	Fold change
FDR	False discovery rate
GATK	Genome Analysis Toolkit
GWAS	Genome-wide association study
HAR	Human accelerated region
Indels	Insertions or deletions
iPSC	Induced pluripotent stem cell line
Kb	Kilobase (1,000 base pairs)

LCL	Lymphoblastoid cell line
LD	Linkage disequilibrium
lincRNA	Long intergenic non-coding RNA
LoF	Loss of function
Mb	Megabase (1,000,000 base pairs)
MCM	Minichromosome maintenance complex
OK-seq	Okazaki fragment sequencing
ORC	Origin recognition complex
PC	Principal component
PCA	Principal component analysis
Pos	Position
Pre-RC	Pre-replication complex
RT	Replication timing
rtQTL	Replication timing quantitative trait locus
SD	Standard deviation
SNP	Single nucleotide polymorphism
SNS-seq	Small nascent strand sequencing
SV	Structural variant
TIGER	Timing Inferred from Genome Replication
TPM	Transcripts per million
UCE	Ultra-conserved element
WGS	Whole genome sequencing
YRI	Yoruba in Idadan, Nigeria

LIST OF SYMBOLS

χ^2	Chi-square
Δ RT	Change in replication timing

CHAPTER 1

INTRODUCTION

Eukaryotic DNA replication

DNA replication is an essential process for all multicellular organisms to generate new copies of DNA prior to cell division. During this process, the genome must replicate completely and accurately to maintain cellular integrity. Mistakes in DNA replication, if not repaired, can lead to mutations. The consequences of these mutations depend on the cell type that they occur in. Mutations in somatic cells have the potential to lead to cancer, while mutations in germ cells have the potential to be passed to offspring and subject to evolutionary processes. Thus, DNA replication is an important process linked to both disease and evolution.

DNA replication initiation is tightly regulated and begins at origins of replication across the genome (Fragkos et al., 2015; Parker et al., 2017). During G1 of the cell cycle, origins are specified with binding of the origin recognition complex (ORC) (Bell and Stillman, 1992; Gavin et al., 1995) and Cdc6. Together with Cdt1, ORC-Cdc6 then recruits two copies of the minichromosome maintenance complex hexamer (MCM2-7), which is the replicative DNA helicase (Evrin et al., 2009; Remus et al., 2009). This step forms the pre-replication complex (pre-RC), and is also known as origin licensing. In the transition to S phase, the addition of Cdc45 and GINS form a Cdc45-MCM2-7-GINS (CMG) complex with active helicase activity (Ilves et al., 2010; Moyer et al., 2006). The helicase can then unwind DNA and start fork progression in a bidirectional manner (i.e. origin firing). Only a subset of origins normally fire, and thus there are a number of dormant origins that will fire only in situations of replicative stress (e.g. fork stalling, failure to initiate pre-RCs) (Ge et al., 2007). DNA synthesis,

accomplished with DNA polymerases, proceeds following each active helicase until there are two full copies of the genome by G2 of the cell cycle. Epigenetic information is also duplicated during DNA replication; histones are extruded prior to the replication fork and reassembled along with newly synthesized histones into nucleosomes on the newly replicated strands (Groth et al., 2007).

In yeast (*Saccharomyces cerevisiae*), origin locations are specified by a consensus DNA sequence. These autonomously replicating sequences (ARS) are start sites of replication and are composed of several sequence elements; most importantly, the 11 base pair A element (i.e. ARS consensus sequence) specifies the location of ORC binding (Bell and Stillman, 1992; Marahrens and Stillman, 1992). However, in higher eukaryotes there is not a specific consensus sequence that defines origin locations, so finding them has been more difficult. Many different methods have been developed to map origin locations (reviewed in: (Ganier et al., 2019; Hulke et al., 2020)), including sequencing of chromatin immunoprecipitated ORC (ChIP-seq; (Dellino et al., 2013)), small nascent strands of RNA-primed DNA at the replication fork (SNS-seq; (Cayrou et al., 2011)), the replication bubble (bubble-seq; (Mesner et al., 2013)), and okazaki fragments (OK-seq; (Smith and Whitehouse, 2012)). Unfortunately, the overlap of origins identified across these methods is surprisingly low and the total number of origins identified varies widely (e.g. several thousand to a hundred thousand in humans), which may at least partially be due to technical differences in resolution or methodology (Hulke et al., 2020).

DNA replication timing

Another method that gives us general information about replication origin locations is DNA replication timing. Replication timing is the spatiotemporal pattern of replication across the

genome that results from initiation at origins at different times during S phase. Genomic regions near the earliest firing origins will have the earliest replication timing, while regions between origins or near late firing origins will be later replicating. Replication timing is highly conserved across individuals of the same species, and fairly conserved across species as well (Ryba et al., 2010; Yaffe et al., 2010). However, replication timing changes across development and varies between differentiated cell types of the same species; it was estimated that 30.5% of the replication timing program changes across 26 human cell types (Rivera-Mulia et al., 2015). This suggests that replication timing is important to cell type specific regulation.

Accordingly, early DNA replication is associated with active transcription, open chromatin, and higher gene density (Rhind and Gilbert, 2013). In particular, the association with transcription has been thoroughly studied (Blin et al., 2019; Klein et al., 2021; Marchal et al., 2019; Müller and Nieduszynski, 2017; Rivera-Mulia et al., 2015; Zhang et al., 2002), but it is still unclear whether there is a direct association between transcription and replication timing or if they are both affected through a shared mechanism (e.g. chromatin). For example, one study found that the deletion of origins in a region with histone genes (in yeast) decreased their expression (Müller and Nieduszynski, 2017), suggesting an impact of replication timing on transcription. While other studies found that inducing transcription advanced replication timing both in *Drosophila* and mouse embryonic stem cell lines (ESCs) (Koryakov et al., 2012; Therizols et al., 2014). Even other studies proposed that replication timing and transcription are regulated independently. For example, tethering a histone deacetylase to the human β globin locus caused the region to be late replicating, but did not highly impact transcription (Goren et al., 2008). The connection between replication timing and chromatin is also not completely understood, as decondensation of chromatin in mouse ESCs was unable to advance replication

timing (Therizols et al., 2014). Therefore, replication timing is linked with genome regulation and function, but the directionality and cause of the association is still uncertain.

In contrast, late replicating regions of the genome have higher frequency of many mutation types, including copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) (Koren et al., 2012). An early study found that human-chimpanzee sequence divergence and human SNP density are both enriched in late replicating regions of the human genome (Stamatoyannopoulos et al., 2009). A more recent study looked at the correlation of human replication timing with several different types of mutations, finding that transversions are generally later replicating than transitions, which in turn are later replicating than CNVs (Koren et al., 2012). The mechanism driving variation in mutation distribution is still unclear but has been shown to in part be due to differences in repair (Supek and Lehner, 2015) and levels of available nucleotides throughout S phase (Kenigsberg et al., 2016). Overall, replication timing bridges an interesting connection between genome regulation and evolution.

Global regulators of replication timing have been difficult to identify. Many gene knock-outs do not disrupt replication timing, suggesting that the program is highly robust and essential. Rif1 is the one well-known global regulator of replication timing, as depletion in yeast, mice, and humans have all shown genome-wide disruption of replication timing (Cornacchia et al., 2012; Klein et al., 2021; Yamazaki et al., 2012). Recently, complete knock-out of Rif1 in human cells caused alterations in both replication timing and epigenetics. Using a degron inducible system they were able to specify Rif1 as the regulator of replication timing, which then impacts chromatin by the first S phase after Rif1 depletion (Klein et al., 2021). Another suggested global regulator of replication timing is MCM10, a protein involved in the initiation of replication; a MCM10 knock-out cell line produced changes in replication timing across nearly 50% of the

human genome (Caballero et al., 2022). Additionally, Xist has been identified as a regulator of X chromosome replication timing, as is marks the inactive X chromosome with heterochromatin, causing it to be late and randomly replicating (Koren and McCarroll, 2014). Overall, replication timing regulation appears to involve both genetic and epigenetic mechanisms.

Evolution of DNA replication timing

Comparing replication timing across species can help identify mechanisms of replication timing regulation and evolution; however, few studies have taken this approach. Early studies comparing *Saccharomyces* species found that active origins were generally the most conserved, both in location and activation time across species, while dormant origins were less conserved (Müller and Nieduszynski, 2012). Another study found that histone genes were generally the earliest replicating and most conserved across divergent budding yeast species, and this was required for proper packaging of DNA after genome replication (Müller and Nieduszynski, 2017). The most recent evolutionary study in yeast focused on changes that occur in replication timing between *Lachnacea* species, which were more divergent than species used in previous yeast studies, finding that evolution takes form in dynamic gain and loss of origins across species (Agier et al., 2018).

There have also been limited evolutionary studies of replication timing in mammals. A few early studies compared human and mouse, finding that syntenic regions were generally conserved in replication timing across species (Ryba et al., 2010; Yaffe et al., 2010). Specifically, Yaffe and colleagues found that ‘domains’ of replication were conserved and just shuffled in position (i.e. rearrangements) through evolution, and that fusions preferentially occurred in regions with similar time of replication. Additionally, one study to date has

compared primate replication timing across single human, chimpanzee, orangutan, gibbon, and green monkey samples (Yang et al., 2018). This study identified genomic regions with conserved and lineage-specific replication timing patterns and associated these patterns with sequence and regulatory elements and higher order genome organization.

Overall, these studies have primarily focused on the question of how replication timing evolved across species. They have been limited by small sample sizes, and in some cases low resolution, and thus did not have the power to link genetic causes of replication timing evolution. Also, very few studied the potential impacts of replication timing evolution (Müller and Nieduszynski, 2017; Yang et al., 2018). Considering the strong link of replication timing to transcription and chromatin structure, we may expect co-variation in these genomic features across species.

In multicellular organisms, it is also important to consider that evolution of replication timing may be different across cell types, as replication timing changes across development (Rivera-Mulia et al., 2015). One study compared human and mice across two cell types – fibroblasts and lymphoblasts – and found that replication timing was more conserved between species of the same cell type, than between the two cell types from the same species (Yaffe et al., 2010). Thus, replication timing evolution may occur in a cell-type specific manner, and reflect genome regulation important to the species per cell type.

Methods for generating replication timing across populations and species

It is possible that a lack of resolution has hindered the ability to find fine-scale variation of replication timing between species. Replication timing patterns can be visualized with a DNA replication timing profile, where peaks in the profile represent the potential locations of origins

or origin clusters (Figure 1.1). Replication timing values are presented as a z-score, where positive values are early replicating and negative values are late. Traditional methods of generating these profiles are time consuming and low-throughput as they require cell sorting of G1 and S phase cells prior to sequencing and/or nucleotide analog incorporation (Hulke et al., 2020).

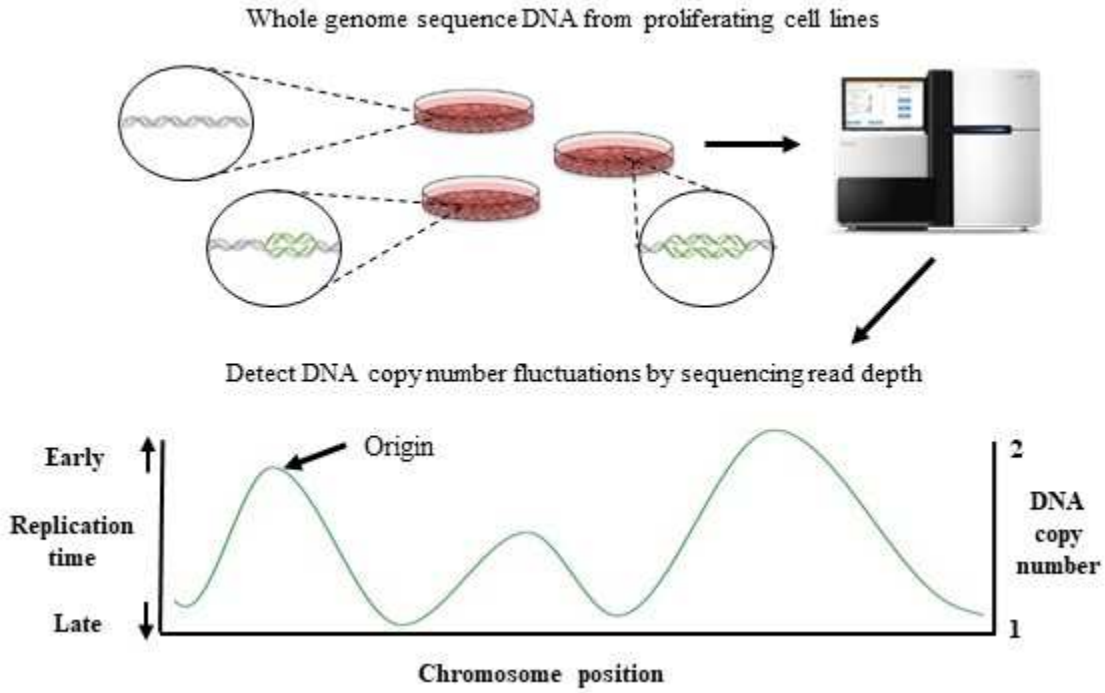


Figure 1.1. DNA replication timing.
Replication timing profiles can be generated from whole genome sequencing of DNA from proliferating cell lines. Some of the sequenced cells will not be actively replicating, while others will be at a variety of time points during S phase or completely replicated. Sequencing read depth is used to calculate DNA copy number fluctuations across the genome which are then used to infer replication timing. Regions of the genome that have already replicated for the majority of sequenced cells will have higher read depth, thus higher DNA copy number. An example replication timing profile is shown, where higher replication timing values on the y-axis are early replicating while lower values are late replicating. Peaks in the profiles represent the locations of origins (or origin clusters).

One of the most common methods of generating replication timing, Repli-seq, involves BrdU (i.e. thymidine nucleotide analog) labelling of newly synthesized DNA strands, followed by cell sorting into S-phase fractions and sequencing of immuno-precipitated BrdU labelled

DNA (Hansen et al., 2010). The most recent, high-resolution version of Repli-seq enables sorting of up to 16 S-phase fractions (Zhao et al., 2020). Repli-seq has facilitated generation of replication timing profiles for a number of species, including humans, non-human primates, and mice (Hansen et al., 2010; Yang et al., 2018; Zhao et al., 2020).

Our lab and others frequently utilize another method for profiling replication timing, that requires sorting of G1 and S phase cells but not nucleotide analog incorporation (Koren et al., 2012). In this method, DNA is sequenced from the sorted cells, followed by a comparison of S-phase sequencing reads to G1 reads in bins across the genome. This generates genomic copy number fluctuations that are corrected for base-line genome copy number without replication (e.g. non-CNV regions should be diploid copy number of two). This method has been applied to many species as well, including humans, mice, and zebrafish (Koren et al., 2012; Siefert et al., 2017; Yehuda et al., 2018).

Recently, our lab has developed a high-throughput and high-resolution method to infer replication timing, without cell sorting or DNA labelling. This method infers replication timing from proliferating cell lines using whole-genome sequencing read depth as a proxy for genomic copy number (Koren et al., 2014; Koren et al., 2021) (Figure 1.1). Therefore, in a population of proliferating cells, regions of the genome that have already replicated will have a higher abundance of DNA and thus more sequencing reads than regions that have not yet replicated. We recently published a comprehensive computational pipeline that generates replication timing profiles from whole-genome sequencing, called TIGER (Timing Inferred from Genome Replication) (Koren et al., 2021). These profiles are corrected for GC-content biases and generated based on binning read counts across the genome, excluding regions that are not uniquely alignable. Copy number variants and other outliers are removed before smoothing and

normalizing the profiles to a z-score. For this method to work, only 10% of sequenced cells are required to be in S-phase, but high sequencing read depth is preferred (e.g. 10-30x) (Koren et al., 2014; Koren et al., 2021). This method can be applied to any species with a contiguous reference genome, making it an optimal method to utilize in evolutionary studies of replication timing.

Replication timing quantitative trait loci

As with identifying genetic determinants of origins, finding cis-acting genetic modifiers of replication timing has also been a challenge. There is some early evidence that replication timing is influenced by sequence elements, but these studies are mostly locus specific (Altman and Fanning, 2001; Liu et al., 2003; Wang et al., 2004). A more recent study used CRISPR-Cas9 deletions in mouse ESCs to identify "early replicating control elements", which when deleted, caused changes in replication timing, transcription, and local chromatin structure (Sima et al., 2019). These studies suggest that replication timing is at least in part dependent on cis-acting sequences, but there is some technical difficulty with scaling these types of analyses genome-wide.

Our high-throughput method has allowed us to profile replication timing for a large number of human individuals using population sequencing data from cell lines (Ding et al., 2021; Koren et al., 2014). With these profiles, we can identify regions of the genome that vary in replication timing across individuals and find genetic variants associated with these differences through quantitative trait locus mapping (rtQTLs). For example, in 108 human ESCs, there are 1,489 regions that vary in replication timing across the autosomes, spanning over 30% of the genome, and 1,837 variant regions in 192 human induced pluripotent stem cell lines (iPSCs) (Ding et al., 2021). In total, 1,617 rtQTLs were mapped across the genome in these cell types,

many of which were associated with regions of variation across individuals. rtQTLs have been found to primarily impact sites with peaks in the replication timing profiles, thus potentially impacting origin initiation (Ding et al., 2021; Koren et al., 2014). The pluripotent stem cell rtQTLs have been subsequently used to identify a combination of histone modifications that predict replication origins. Thus, rtQTLs have been an important method to understanding replication timing regulation.

Variation in replication timing has not only been linked to chromatin structure but also transcription. Genes that fall into rtQTL associated regions have been shown to display variation in expression as well (Ding et al., 2021; Koren et al., 2014). A relatively small number of the iPSC rtQTLs (13.7%) were shared with variations in gene expression across individuals (Ding et al., 2021). This suggests that genetic regulation of replication timing variation is primarily separate from transcription, but in a small number of cases may have shared genetic causes.

Inter-individual replication timing variation and rtQTLs may be important considerations in studies of replication timing evolution, as regions with large amounts of variation may reflect different patterns of evolution if only looking at a few samples. Knowledge of replication timing variation and rtQTLs may also provide a mechanistic way to understand ongoing replication timing evolution by using a closely related species as an outgroup. All previous studies of replication timing evolution did not consider inter-individual replication timing variation and how this could affect perceived evolution of replication timing across species.

Human molecular and regulatory evolution

The aforementioned connection between replication timing and mutation rate supports the possibility that replication timing has an impact on sequence evolution and thus evolution of

phenotype. To understand human specific phenotype and evolution, many studies have compared single genes, specific sequence elements, and genome regulation in human to closely related species (Olson and Varki, 2003). Humans are a part of the great ape family that includes chimpanzees, bonobos, gorillas, and orangutans. We are more distantly related to other primate species including Old and New World monkeys, tarsiers, and lemurs. Humans and our closest living relatives, chimpanzees, are nearly 99% identical at the single nucleotide level with an estimated divergence time of six million years ago (Consortium, 2005; Kronenberg et al., 2018) (Figure 1.2). Rhesus macaques (an Old World monkey species) are approximately 93% identical to humans with a divergence time of approximately 25 million years ago (Gibbs et al., 2007; Warren et al., 2020) (Figure 1.2). Despite this high sequence similarity, especially between humans and chimpanzees, these species exhibit extensive phenotypic divergence.

Comparative sequence data is important for identification of genes and genomic regions under selection in the human lineage, which could be important contributors to human specific traits. Identifying selection generally involves testing single genes or genome-wide scans (Fu and Akey, 2013). A common test of selection compares the number of protein coding changes to non-protein coding changes in orthologous genes across species (dN/dS; non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site) (Nielsen et al., 2005). More non-synonymous substitutions suggests positive (i.e. directional) selection, while more synonymous changes suggests negative (i.e. purifying) selection. Other methods to identify more recent selection utilize human population variation to identify selective sweeps (Fu and Akey, 2013; Voight et al., 2006); haplotypes with the favorable allele display an increase in frequency and would be in strong linkage disequilibrium (LD) with surrounding SNPs. These methods, among others, have identified many genes under selection in humans,

including several that appear to impact human specific traits; for example, language (FOXP2; (Enard et al., 2002)), lactase persistence (LCT; (Tishkoff et al., 2007)), and brain size (ASPM; (Mekel-Bobrov et al., 2005); *microcephalin;* (Evans et al., 2005)). Another comparative approach identified sequences important to human evolution by finding those that are conserved in mammals but diverged in humans, known as human accelerated regions (HARs; (Hubisz and Pollard, 2014)). Interestingly, HARs primarily occurred in regulatory regions, rather than in genes, and many were proposed to function as developmental enhancers (e.g. some in brain and limbs).

In addition to single nucleotide changes, structural variation has also been shown to play a role in human evolution. There are over 120,000 deletions and insertions between humans and chimpanzees spanning approximately 80 Mb, in addition to several large inversions and fusion of chromosomes 2A and 2B in the human lineage (Ijdo et al., 1991; Kronenberg et al., 2018). Some studies identified sequences shared among non-human primates, but that were completely absent in humans (Kronenberg et al., 2018; McLean et al., 2011). These conserved human specific deletions were found to primarily impact regulatory DNA and were associated with human specific phenotypes, such as loss of penile spines and the expansion of the neocortex (Kronenberg et al., 2018; McLean et al., 2011).

Phenotypic divergence between humans and chimpanzees has increasingly been attributed to regulatory evolution between human and chimpanzee genomes (Fraser, 2013; King and Wilson, 1975). Regulatory evolution, including gene expression, histone modifications and other marks of genomic regulation have shown divergence across humans and chimpanzees in multiple cell types (García-Pérez et al., 2021; Khan et al., 2013; Romero et al., 2012; Zhou et al., 2014). Primary samples from chimpanzees and other great apes are difficult to obtain, so the use

of induced pluripotent stem cells have proven very important in this area of research as they can be differentiated into multiple cell types (Blake et al., 2018; Eres et al., 2019; Housman et al., 2022; Pavlovic et al., 2018; Prescott et al., 2015; Romero et al., 2015). Most recently, fused human-chimpanzee iPSCs were created to separate cis- and trans-acting factors contributing to regulatory evolution (Agoglia et al., 2021; Gokhman et al., 2021). Overall, these studies have identified changes in gene regulation that contribute to human-specific craniofacial morphology (Gokhman et al., 2021; Prescott et al., 2015), embryonic limb development (Cotney et al., 2013), and neurodevelopment (Agoglia et al., 2021). Therefore, genome regulation plays a significant role in human evolution.

Human evolution and DNA replication timing

DNA replication timing is a method of genome regulation that has been understudied in terms of its potential impact on human evolution by affecting gene regulation and/or mutation rate. Previous studies linked replication timing to sequence divergence, where human-chimpanzee and human-macaque divergent sites were enriched in late replicating genomic regions (Stamatoyannopoulos et al., 2009). Similarly, fixed deletions in the human lineage were enriched in late replicating genomic regions (Koren et al., 2012; McLean et al., 2011). Most interestingly, mutation density correlates with replication timing more highly in germ cells than in somatic cells, further supporting that replication timing may have evolutionary consequences (Yehuda et al., 2018). A very recent study concluded that the contribution of replication timing to mutation rate is stable across the great apes lineage and doesn't appear to contribute to species specific mutation rate, but this analysis only considered human replication timing (Goldberg and Harris, 2022). Thus, replication timing is associated with mutation rate and sequence evolution,

but an open question is whether difference in replication timing across species contributes to mutation rate divergence. A thorough analysis of DNA replication timing in great apes would be needed to address this question.

Research questions

Based on previous research, there are two key questions in the field of DNA replication timing evolution. First, how mechanistically does replication timing evolve across species and how it is that regulated? Second, what are the genomic impacts of replication timing evolution?

In Chapter 2, I utilize multiple cell types and multiple individual replication timing profiles from human, chimpanzee, and rhesus macaque, to identify how replication timing has evolved (e.g. origins gained/lost, change in origin firing times) in human and chimpanzee lineages (Figure 1.2). I also identify regions with replication timing variation across chimpanzee individuals, search for rtQTLs associated with this variation, and compare these to our previous results in human. These analyses address the evolution of rtQTLs and inter-individual replication timing variation both within and between species. To understand factors that could be driving replication timing regulation and evolution, I test for differential enrichment of genomic features (e.g. gene expression, chromatin accessibility) in replication timing variable regions between species. Finally, I address the impact of replication timing on human evolution though association of replication timing with local mutation rate, sequence divergence, and genes/other genomic regions important to human evolution.

In Chapter 3, I highlight the key findings and contributions of the research from Chapter 2, and elaborate on limitations and potential future research.

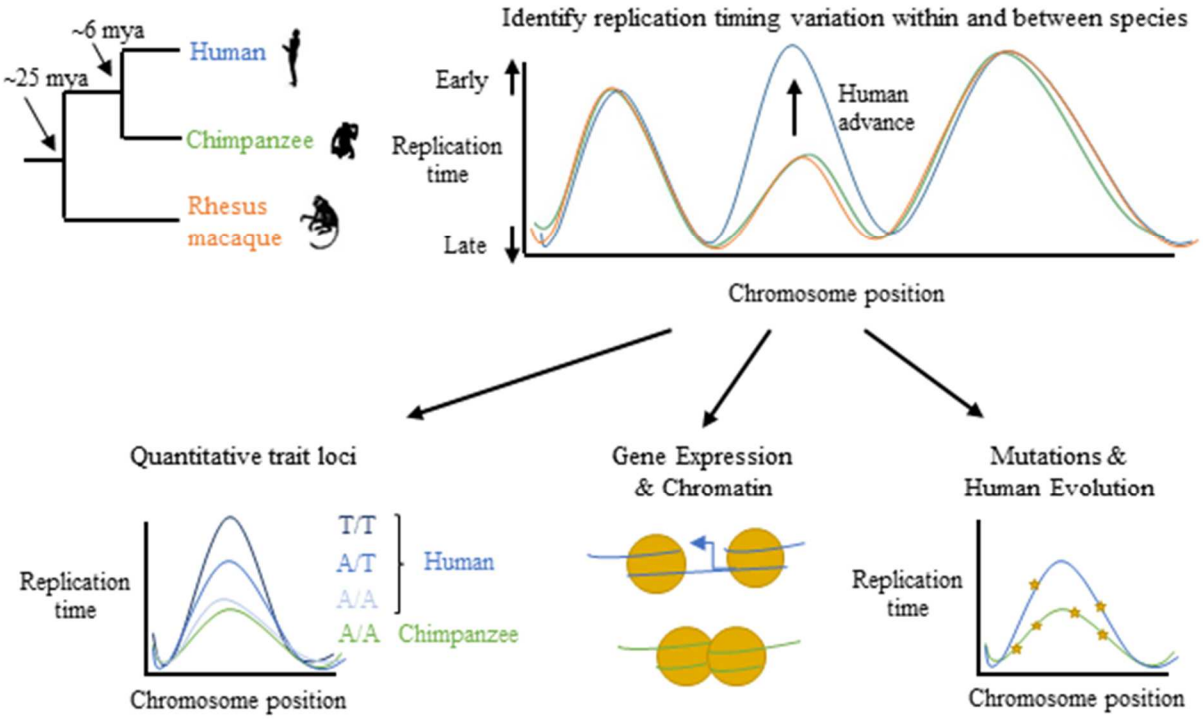


Figure 1.2. Research approach.
In this dissertation, I generate replication timing profiles for humans, chimpanzees and rhesus macaques and identify replication timing variation within and between species. I then identify causes and consequences of replication timing evolution. I use quantitative trait loci to link sequence changes to replication timing changes between species. In the example, a replication timing profile is shown for each human genotype. Humans with the same genotype as chimpanzees share similar replication timing. I also study evolutionary impacts of replication timing changes on gene expression and chromatin structure. In the example, humans are earlier replicating, so they have more open chromatin and active transcription, while chimpanzee has closed chromatin and no transcription. Lastly, I study the impact of replication timing evolution on mutations and human evolution. In the example, mutations (indicated with stars) occur more frequently in chimpanzees since they are later replicating than humans.

CHAPTER 2

THE EVOLUTION OF THE HUMAN DNA REPLICATION TIMING PROGRAM

This chapter is published on bioRxiv as: Bracci, A.N., Dallmann, A., Ding, Q., Hubisz, M.J., Caballero, M., and Koren, A. (2022). The evolution of the human DNA replication timing program. bioRxiv, 2022.2008.2009.503365. doi: https://doi.org/10.1101/2022.08.09.503365 (Bracci et al., 2022)

Abstract

DNA is replicated according to a defined spatiotemporal program that is linked to both gene regulation and genome stability. The evolutionary forces that have shaped replication timing programs in eukaryotic species are largely unknown. Here, we studied the molecular causes and consequences of replication timing evolution across 94 humans, 95 chimpanzees, and 23 rhesus macaques. Replication timing differences recapitulated the species' phylogenetic tree, suggesting continuous evolution of the DNA replication timing program in primates. Hundreds of genomic regions had significant replication timing variation between humans and chimpanzees, of which 66 showed advances in replication origin firing in humans while 57 were delayed. Genes overlapping these regions displayed correlated changes in expression levels and chromatin structure. Many human-chimpanzee variants also exhibited inter-individual replication timing variation, pointing to ongoing evolution of replication timing at these loci. Association of replication timing variation with genetic variation revealed that DNA sequence evolution can explain replication timing variation between species. Taken together, DNA replication timing shows substantial and ongoing evolution in the human lineage that is driven by sequence alterations and impacts regulatory evolution.

Introduction

Understanding of human specific phenotypes and their evolution has primarily focused on the comparison of individual genes or sequence elements, and their regulation, between humans and closely related species (Olson and Varki, 2003). Humans and chimpanzees are approximately 99% identical at the single nucleotide level, yet have undergone extensive phenotypic divergence (Kronenberg et al., 2018). This has increasingly been attributed to regulatory evolution, including gene expression, which has been associated with brain, skeletal, and other phenotypes (Agoglia et al., 2021; Fraser, 2013; Gokhman et al., 2021; King and Wilson, 1975). An understudied form of genome regulation, with potential impact on regulatory and sequence evolution, is the spatiotemporal program of DNA replication.

Genome replication is accomplished by replication origins that fire at different times during S phase, resulting in a defined pattern of DNA replication timing. Early DNA replication is associated with high gene density, open chromatin, and active transcription (Rhind and Gilbert, 2013), while later replicating regions typically exhibit higher frequencies of single nucleotide mutations and polymorphisms (Francioli et al., 2015; Koren et al., 2012; Stamatoyannopoulos et al., 2009). Replication timing thus bridges between genome regulation and evolution. As a corollary, understanding the evolution of replication timing can reveal the selective forces that have shaped particular replication programs, inform mechanisms of replication timing regulation, and uncover impacts of replication timing on sequence, molecular and phenotypic evolution.

Only a handful of studies have compared replication timing across species. Studies in yeast suggested that replication origins dynamically gain and lose activity during evolution (Agier et al., 2018) and that conserved early replication, in particular of histone genes, is

required for high gene expression levels (Müller and Nieduszynski, 2017). In contrast, replication timing has been shown to be highly conserved between corresponding cell types of humans and mice despite extensive genome rearrangements (Ryba et al., 2010; Yaffe et al., 2010), while a more recent study suggested the presence of both conserved and species-specific replication timing regions among five primate species (Yang et al., 2018). Importantly, previous studies have been under-powered to identify the genetic changes that drive replication timing evolution nor its potential impacts on regulatory and sequence evolution.

Here, we address the causes and consequences of replication timing evolution by profiling a large number of humans, chimpanzees, and rhesus macaques. We find that replication timing has continuously evolved across these species at hundreds of locations. Comparison to intra-species variation and sequence polymorphisms within species and divergence between species revealed the genetic basis of a subset of replication origins that have gained or lost activity during evolution. On the other hand, analysis of gene expression and chromatin structure suggests a complex relationship between the evolution of replication timing and gene regulation. Overall, this study advances our knowledge of how replication timing evolves, the association of replication timing with genome regulation and transcription, and the determinants of replication timing evolution.

Results

High resolution DNA replication timing profiles across humans, chimpanzees, and rhesus macaques

To study the evolution of DNA replication timing across primates, we sequenced the genomes of 90 chimpanzee lymphoblastoid cell lines (LCL), 23 rhesus macaque LCLs, and

seven chimpanzee induced pluripotent stem cell lines (iPSC), along with 88 human LCLs and eight human iPSCs. We aligned each species' sequencing reads to its own reference genome and inferred DNA replication timing from read depth fluctuations across chromosomes (Ding et al., 2021; Koren et al., 2014; Koren et al., 2021) (see Methods). Our method of inferring replication timing from whole genome sequence data was particularly suited to this task, as chimpanzee material is scarcely available for the experimental manipulations required by other approaches (e.g. Repli-seq; (Hansen et al., 2010)). One chimpanzee LCL, one chimpanzee iPSC and two human iPSC were filtered due to low data quality. Read depth fluctuations showed long-range continuity along chromosomes (autocorrelation) consistent with DNA replication, and LCL data resolution was further improved using principal component (PC)-regression (Ding et al., 2021) (Figure 2.1 A-E). The resulting replication timing profiles were highly consistent across samples within each species (human LCLs r = 0.94—0.99; chimpanzee LCLs r = 0.84–1; rhesus LCLs r = 0.97–1; human iPSCs r = 0.91–0.97; chimpanzee iPSCs r = 0.96–0.97) (Figure 2.2 A-E; 2.3 A-D).

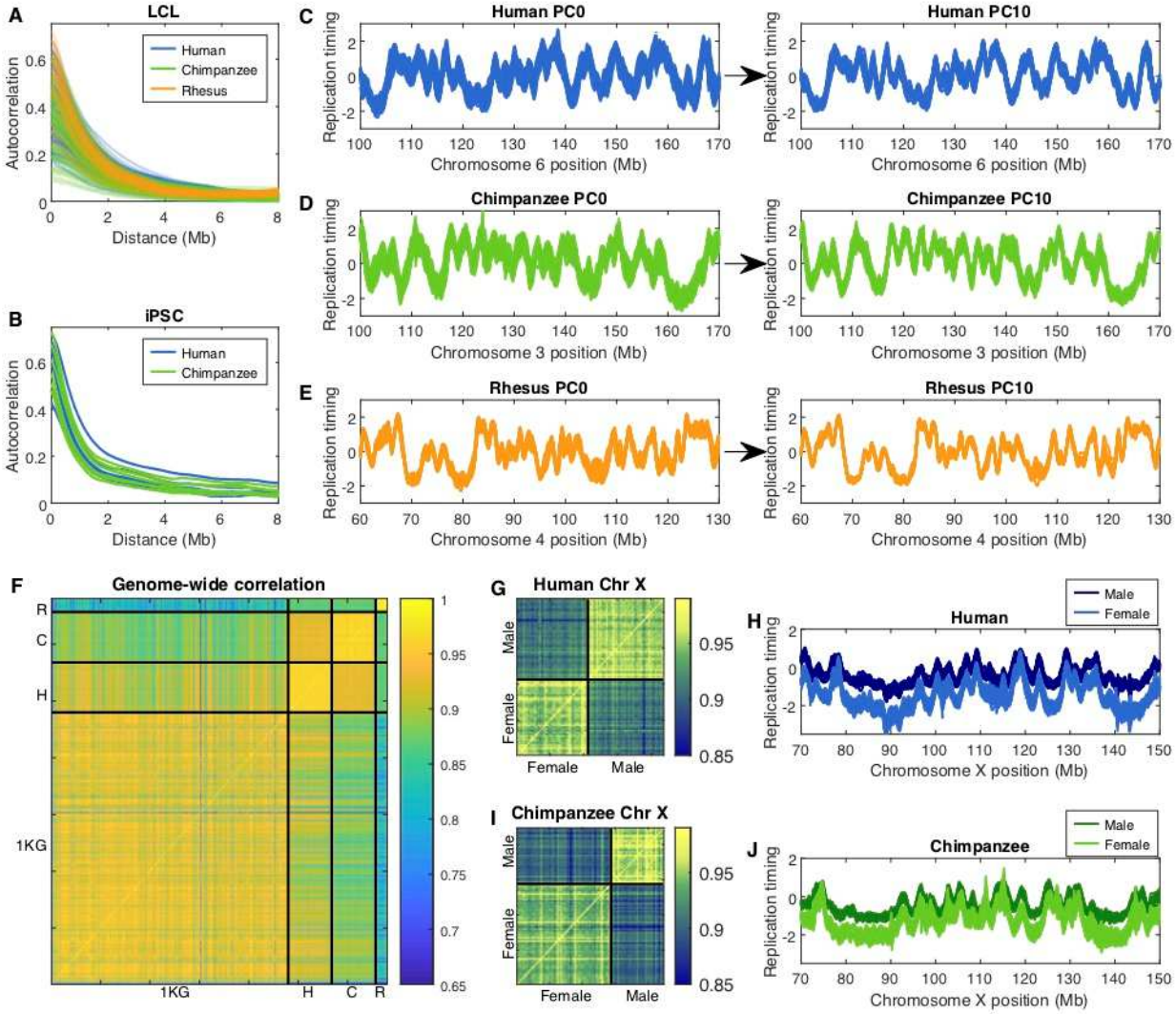


Figure 2.1. Quality control and data processing.
(A-B) Autocorrelation of raw autosomal replication timing data for LCLs (A) and iPSCs (B). (C-E) Smoothed replication timing profiles before principal component (PC)-correction (PC0; left) and after correction for 10 PCs (PC10; right) for human, chimpanzee and rhesus macaque LCLs. (F) Pairwise Pearson correlation of autosomal replication timing across all 1000 Genomes (1KG) African samples (n=480) to the human (H), chimpanzee (C), and rhesus macaque (R) samples from this study. Correlations are higher between 1000 Genomes samples and humans compared to chimpanzees or rhesus macaques from this study. (G, I) Pairwise Pearson correlation of human and chimpanzee LCL X chromosome replication timing values, confirming the conservation of sex differences. (H, J) Human and chimpanzee LCL X chromosome replication timing profiles for males and females. Female X chromosome profiles are later replicating than male, and also appear noisier (more diffuse).

We further validated the replication timing profiles in three ways. First, we measured replication timing by sorting and sequencing G1 and S phase cells of select samples (Koren et

al., 2012), which provided replication timing profiles highly correlated to those generated without cell sorting (human LCL mean r = 0.97; chimpanzee iPSC mean r = 0.87; rhesus LCL mean r = 0.95) (Figure 2.2 A, 2.2 C; 2.3 B). Second, we compared the chimpanzee LCL samples to a previously published chimpanzee replication timing profile generated using Repli-seq (Yang et al., 2018) (mean r = 0.90) (Figure 2.2 B). Finally, as we showed previously in humans (Koren and McCarroll, 2014), replication of the X chromosomes was delayed and less structured in chimpanzee LCL females compared to males (Figure 2.1 G-J). Together, these results demonstrate that the replication timing profiles of all three species are of high quality and reproducibility.

Next, we compared genome-wide replication timing to human-chimpanzee divergence, single nucleotide polymorphisms (human dbSNP 153 common, n=9,585,612; chimpanzee dbSNP, n=1,468,866), and somatic cell line mutations (identified as de novo mutations in chimpanzee trios; see Methods) and found that all were enriched at late-replicating genomic regions in both humans and chimpanzees (Figure 2.2 F, G). On the other hand, gene density, gene expression levels (Khan et al., 2013; Romero et al., 2015; Soto et al., 2020; Zhou et al., 2014), ATAC-seq (García-Pérez et al., 2021), and H3K27ac ChIP-seq (Zhou et al., 2014) were all enriched at, or correlated with, early-replicating genomic regions in both humans and chimpanzees (Figure 2.2 F, G; 2.3 J). These genome-wide trends were also replicated in human and chimpanzee iPSCs (Romero et al., 2015; Soto et al., 2020) (Figure 2.3 H-J), overall supporting the conservation of genomic features associated with replication timing across cell types and species.

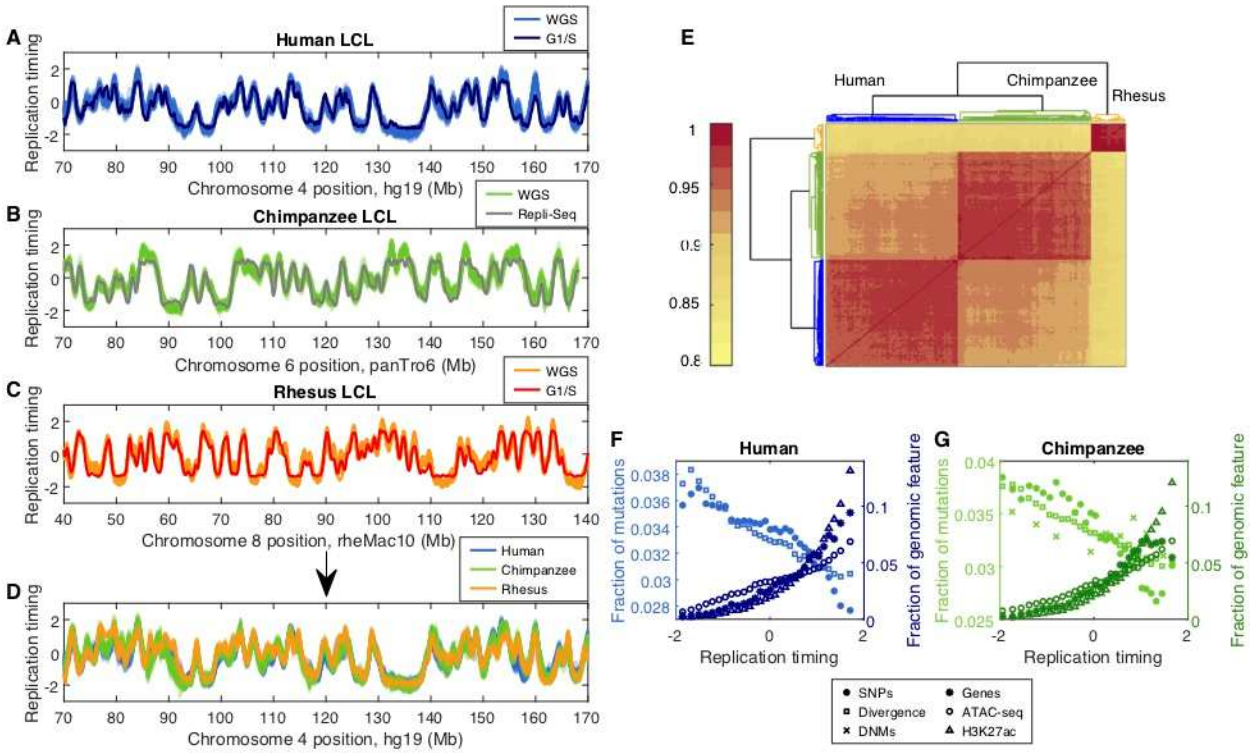


Figure 2.2. Replication timing evolution in primate species.
(A-C) Replication timing was inferred from read depth fluctuations in whole genome sequencing (WGS) data of human, chimpanzee, and rhesus macaque LCLs. Units are standard deviation from an autosome-wide mean of 0. Shown for comparison are a consensus G1/S profile for human LCLs (A) (Koren et al., 2012), a G1/S rhesus macaque LCL profile (generated in this study; C), and a chimpanzee LCL replication profile generated using Repli-Seq (B) (Yang et al., 2018). (D) Human, chimpanzee, and rhesus macaque replication timing profiles, plotted on human genomic coordinates (hg19; see also Figure 2.4), show conservation of the replication timing program. (E) Hierarchical clustering of human, chimpanzee, and rhesus macaque LCL Pearson correlation values. Replication timing is highly consistent within species and, while is largely conserved among species, exhibits significant inter-species variation that corresponds to the evolutionary divergence of primates. (F-G) SNPs, human-chimpanzee divergent sites, and de novo mutations (DNMs) are enriched at late-replicating DNA, while protein coding genes and marks of accessible chromatin (ATAC-seq peaks and H3K27ac ChIP-seq peaks) are enriched at early-replicating DNA. Fraction of human (F) or chimpanzee (G) genomic features in 30 replication timing bins per species. DNM rate was calculated in 10 replication timing bins (fraction is 3x).

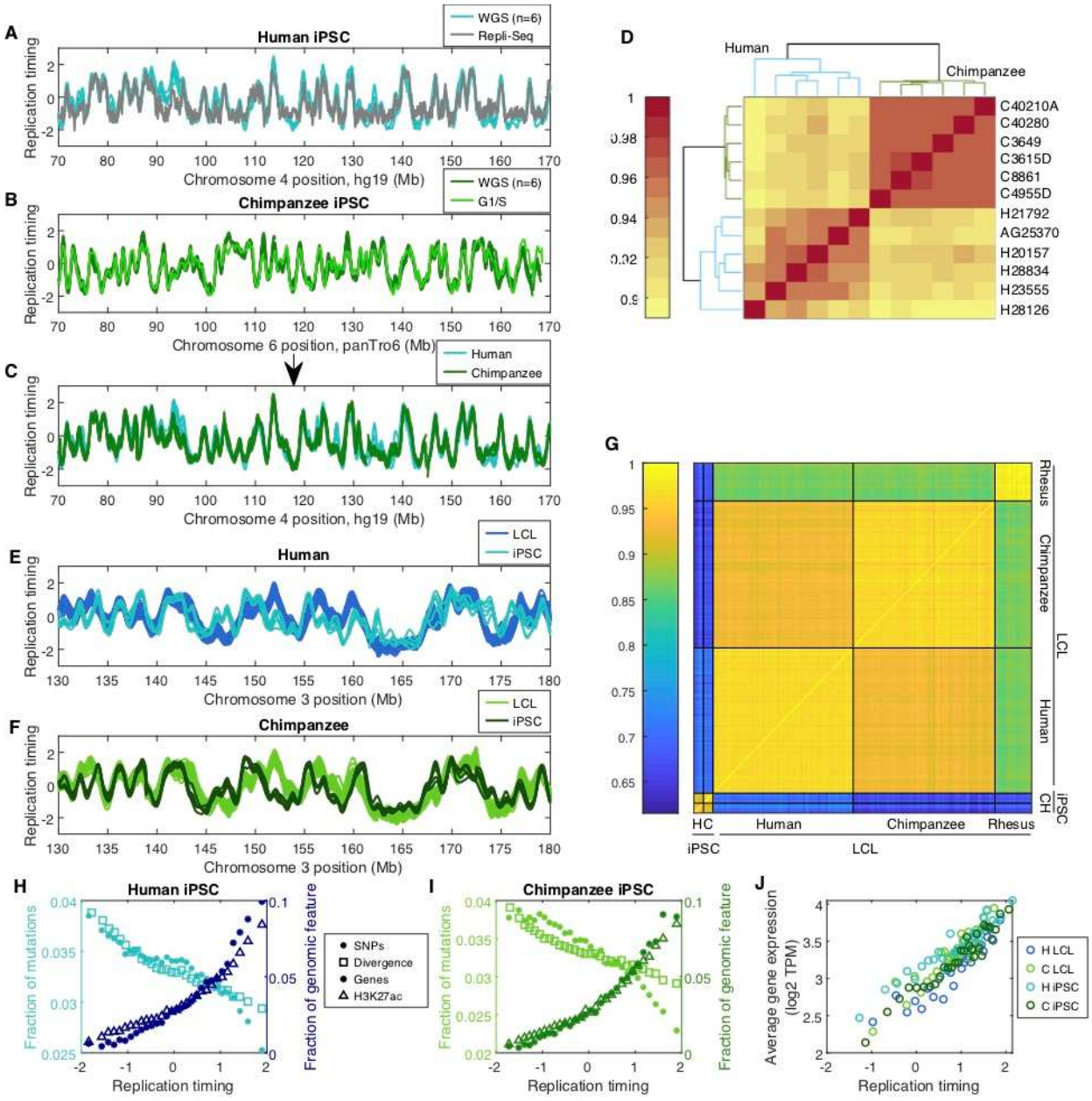


Figure 2.3. Replication timing varies between cell types more than between species.
(A-D, H, I) As in Figure 1, for human and chimpanzee iPSC data. Chimpanzee G1/S was generated in this study; human Repli-seq was obtained from Replication Domain (Accession: Ext30484475). (E-F) Overlaid comparison of human (E) and chimpanzee (F) LCL and iPSC replication timing profiles. (G) Pairwise Pearson correlation of autosomal replication timing across all iPSC and LCL samples. (J) Gene expression (averaged across cell lines for each cell type and species; data obtained from (Soto et al., 2020)) compared to replication timing for human and chimpanzee LCLs and iPSCs, at human-chimpanzee orthologous genes (with average TPM>0.1) in 30 replication timing bins.

Substantial variation in replication timing between species

To compare replication timing profiles between species, we converted the chimpanzee and rhesus macaque replication timing data to human genome coordinates (see Methods); these conversions had a minimal effect on the structures of the replication profiles (Figure 2.4 A, B). We found that replication timing was highly conserved across species (Figure 2.2 D, E; 2.3 C, D) and that replication timing variation was greatest between cell types (LCL and iPSC; mean r = 0.69) (Figure 2.3 E-G). Nonetheless, there were also clear inter-species differences within the same cell type. Hierarchical clustering of replication timing values across samples recapitulated the phylogenetic tree for these three species (Figure 2.2 E; 2.3 D), suggesting that replication timing has evolved continuously across the primate lineage, primarily in a cell-type-specific manner.

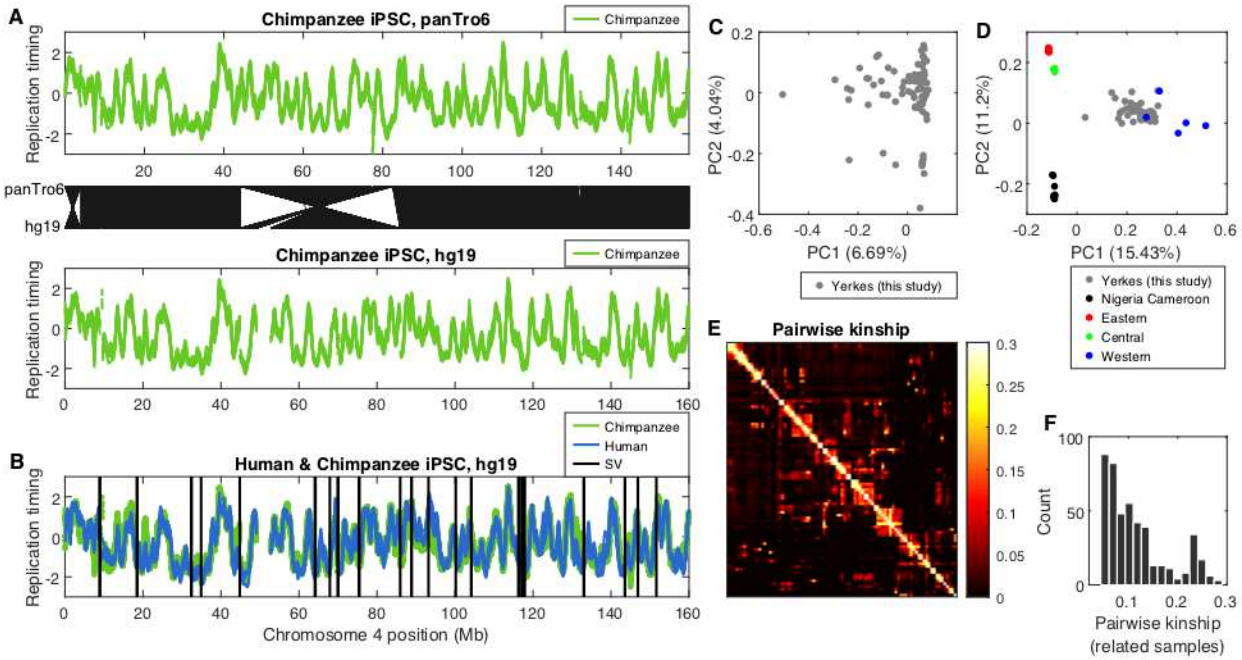


Figure 2.4. Chimpanzee relatedness, population structure, and replication timing coordinate conversion. (A) Chimpanzee iPSC chromosome 4 replication timing profiles in panTro6 (top) and human hg19 (bottom) coordinates. Black lines between the top and bottom panels indicate syntenic loci between panTro6 and hg19 genome builds. Two large inversions are apparent. (B) Human and chimpanzee iPSC

replication timing profiles for chromosome 4 (hg19 coordinates) along with previously mapped structural variants (SVs; inversions, deletions, and insertions) (Kronenberg et al., 2018; Soto et al., 2020) larger than 100 Kb between human and chimpanzee. (C) Genotype principal component analysis (PCA) for the chimpanzee LCLs in this study. PC2 separates the samples into two groups (which does not correspond to sequencing batch). (D) Genotype PCA for chimpanzee LCLs in this study ("Yerkes LCLs") together with samples from known chimpanzee sub-populations (Prado-Martinez et al., 2013), indicates that the chimpanzee samples were primarily from the western chimpanzee sub-population (*Pan troglodytes versus*). (E) Pairwise kinship values across all samples. Values were clustered using hierarchical clustering. (F) Distribution of pairwise kinship for related samples (3rd degree relationship or closer, kinship>0.04) (463 pairs).

We systematically searched for specific differences in replication timing between humans and chimpanzees, separately for LCLs and iPSCs, using sliding ANOVA tests with a Bonferroni corrected p-value threshold of 8.7×10^{-7} (see Methods). For the X chromosome, males and females were considered separately. We identified 858 autosomal regions where human and chimpanzee LCLs significantly differed in replication timing. These regions covered 1.1 Mb on average and cumulatively spanned 980 Mb (36.6% of the analyzable autosomes). Similarly, we identified 47 variant regions on the X chromosome in females and 39 in males (1.4 and 1.2 Mb on average, spanning a total of 64 (42.9% of the analyzable X chromosome) and 45 Mb (29.9%) in females and males, respectively). In iPSCs, we identified 704 autosomal variant regions covering 1.1 Mb on average and cumulatively spanning 797 Mb (29.8% of the autosomes), likely less than in LCLs due to the more limited sample size.

A majority of the human-chimpanzee variant regions occurred at peaks in the replication timing profiles (LCL: 620/944, 65.7%; iPSC: 476/704, 67.6%), suggesting that a major source of replication timing variation is changes in replication origin (or origin cluster) activity. Extending from this observation, and under the assumption that changes in origin activity are the most likely explanation for the large replication timing variants that we observed, we designated the center of the peak as the most likely source of replication timing variation within each region (see Methods). This was only applied to variant regions that contained replication timing peaks,

while 134 LCL variants with more than one peak were separated into several independent variant regions. Overall, we called 731 LCL and 557 iPSC replication timing variants that each contained one putative source site (Figure 2.5; 2.6; 2.7; Table S1), and utilized them for downstream analyses. These variants covered on average 1.2 Mb in LCLs and 1.1 Mb in iPSCs, and had an average magnitude of replication timing difference between humans and chimpanzees of 0.4 standard deviations in LCLs (Figure 2.6 B, C) and 0.5 in iPSCs. The majority of these variant regions were earlier replicating in humans compared to chimpanzees, in both LCLs (57.0%; Binomial test $p=1.2x10^{-4}$; Figure 2.5 E; 2.6 B) and iPSCs (53.9%; p=0.08). Consistent with these variants containing replication profile peaks, the distribution of replication timing at variants was skewed towards early replication in both humans and chimpanzees (Figure 2.6 A). The fraction of replication timing peaks that varied between humans and chimpanzees – 32.5% – was comparable to the fraction of the genome with replication timing variation; thus, the widespread evolution of replication timing is not an inflated estimate due to the broad effect of individual replication origins.

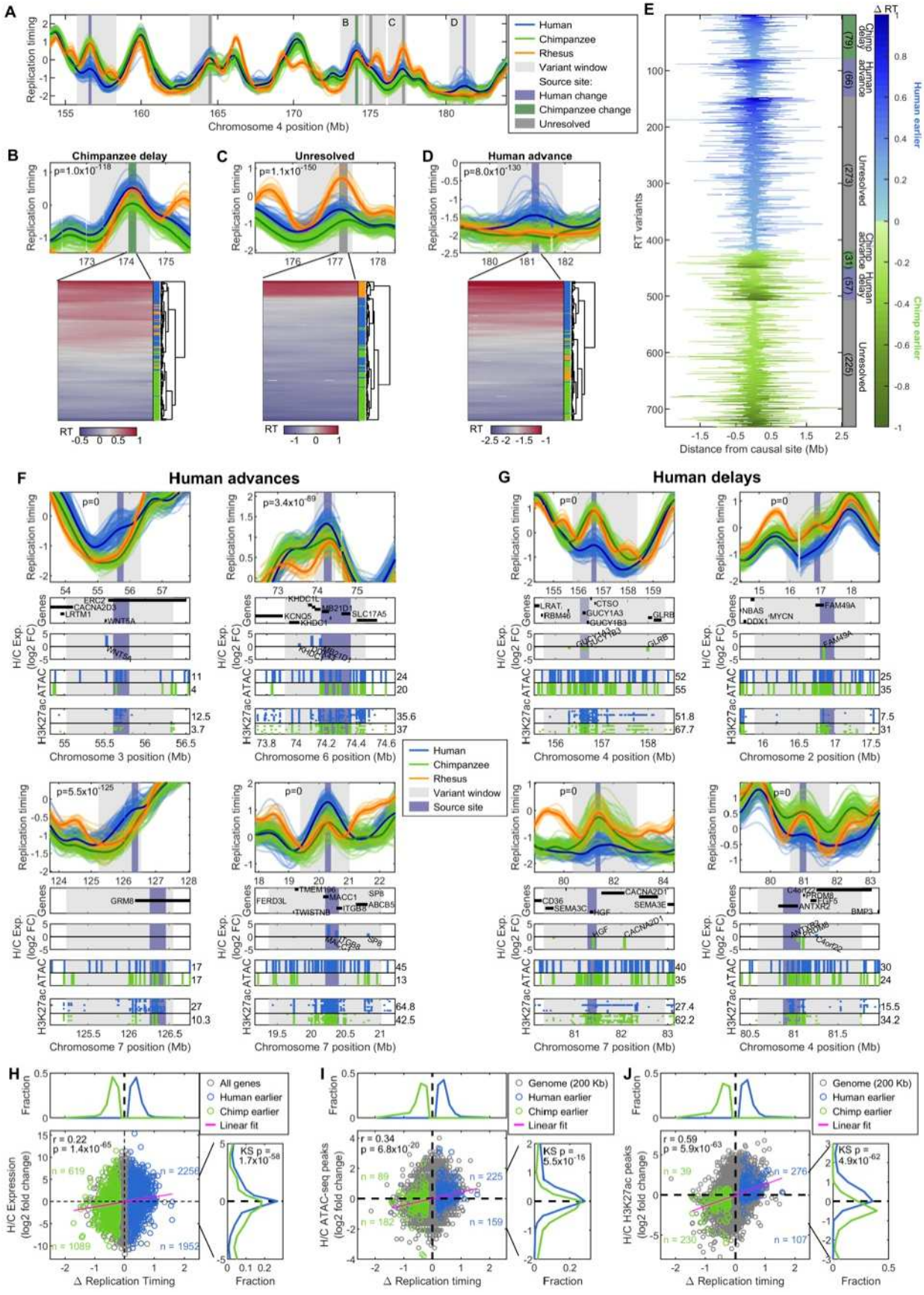
In order to infer whether each replication timing change occurred in the human or chimpanzee lineage, we compared the average LCL replication timing profile for each species to the average profile of rhesus macaque as an outgroup. Specifically, we calculated the pairwise Euclidean distance between each pair of species within each variant region source site (see Methods). Human-specific replication timing changes were defined as regions in which chimpanzees and rhesus macaques were closer to each other than either were to humans, and chimpanzee-specific changes were similarly defined as cases in which humans and rhesus macaques were the most similar (see Methods). Regions with significantly different replication timing among all three species were considered unresolved for evolutionary direction. In total,

we resolved 233 replication timing variant regions (out of the 731 LCL variants containing a replication timing peak), of which 123 and 110 were changes in the human and chimpanzee lineages, respectively (similar number of changes in each lineage expected based on the molecular evolutionary clock; χ^2=0.73, df=1, p=0.39 (Tajima, 1993)). Of these, we inferred 66 to be human advances and 57 to be human delays, while another 31 and 79 were inferred to be replication timing advances or delays, respectively, in chimpanzees (Figure 2.5; 2.6 D, E). Of the 66 human advances, 55 represented earlier activation of a shared origin while another 11 regions appeared to represent de novo evolutionary emergence of novel replication origins in the human lineage. Similarly, 55 replication origins appeared to have been delayed in their firing time in humans compared to chimpanzees, with evidence for two replication origins being entirely lost in some humans. Importantly, all human origin gains and losses were polymorphic, present in 31-72% and 7-13% of individuals, respectively (Figure 2.8). This suggests that these origins have been recently gained or lost and are subject to ongoing evolution in the human lineage. In comparison, we identified seven and 56 origins that have been putatively gained or lost, respectively, in virtually all human and chimpanzee samples compared to macaques (allowing up to 5% technical variation of samples). Thus, on a broader evolutionary timescale, we see compelling evidence of more substantial restructuring of the replication program in primates.

Of 731 human-chimpanzee LCL replication timing variants, 47 (6.4%) were shared in iPSCs; of these, 30 had a similar shape of replication timing profiles (by correlation; see Methods) between cell types of the same species (Figure 2.7 E). Thus, although variants shared across cell types have greater potential to impact species-specific phenotypic differences, most human-chimpanzee replication timing differences are cell-type-specific.

Figure 2.5. Replication timing evolution and its co-variation with gene expression and chromatin accessibility.

(A) Replication timing profiles for a region of chromosome 4 for humans (n=88, blue), chimpanzees (n=89, green), and rhesus macaques (n=23, orange) along with identified human-chimpanzee replication timing variant regions (light gray) and their called source sites. Source site color indicates the lineage in which replication timing was inferred to have evolved. (B-D) Three example regions indicated in (A) are shown at greater resolution. Replication timing of each sample within the source site shown as heat maps; dendrograms: hierarchical clustering of sample replication timing similarity. The clustering demonstrates the separation of the majority (or all) of human from chimpanzee samples as well as the clustering of one (or none) of them to rhesus macaque replication timing. P-values: significance (ANOVA) of human-chimpanzee differences within the variant region. (E) Mean difference in replication timing (Δ RT) between humans and chimpanzees across each replication timing variant, centered at the source sites. Variants are sorted by being earlier in humans or chimpanzees, then by the species in which the change was inferred to have happened, and last by the magnitude of inter-species replication timing difference (Δ RT). (F-G) Examples of replication timing advances (F) and delays (G) inferred to have occurred in the human lineage. Genes, expression, ATAC-seq peaks and H3K27ac ChIP-seq data shown beneath the replication timing profiles for each variant and flanking 200 Kb. Numbers next to ATAC-seq track: the number of human and chimpanzee ATAC-seq peaks within the variant window. Numbers next to H3K27ac track: average number of human and chimpanzee H3K27ac ChIP-seq peaks within the variant window. Some gene names were removed from the Genes track for readability. (H) Differences between human and chimpanzee LCL replication timing compared to differences in gene expression, for all genes as well as genes within replication timing variants with either earlier replication timing in humans (blue) or in chimpanzees (green). Correlation coefficient (r) and p-value indicated for variant regions. Number of genes in each quadrant further demonstrates the correlation between gene expression and replication timing variation. Top and right histograms: distributions of replication timing and gene expression differences, respectively, within replication timing variant regions. (I-J) As in H, using ATAC-seq (I) or H3K27ac ChIP-seq (J) data.


A
Replication timing
Chromosome 4 position (Mb)
Human
Chimpanzee
Rhesus
Variant window
Source site:
Human change
Chimpanzee change
Unresolved
B
Chimpanzee delay
C
Unresolved
D
Human advance
E
RT variants
Distance from causal site (Mb)
Δ RT
Human earlier
Chimp earlier
F
Human advances
G
Human delays
H/C Exp. (log2 FC)
H3K27ac
ATAC
Genes
Chromosome 3 position (Mb)
Chromosome 6 position (Mb)
Chromosome 4 position (Mb)
Chromosome 2 position (Mb)
Chromosome 7 position (Mb)
H
I
J
All genes
Genome (200 Kb)
Human earlier
Chimp earlier
Linear fit
H/C Expression (log2 fold change)
H/C ATAC-seq peaks (log2 fold change)
H/C H3K27ac peaks (log2 fold change)
Δ Replication Timing
Δ Replication timing
Fraction

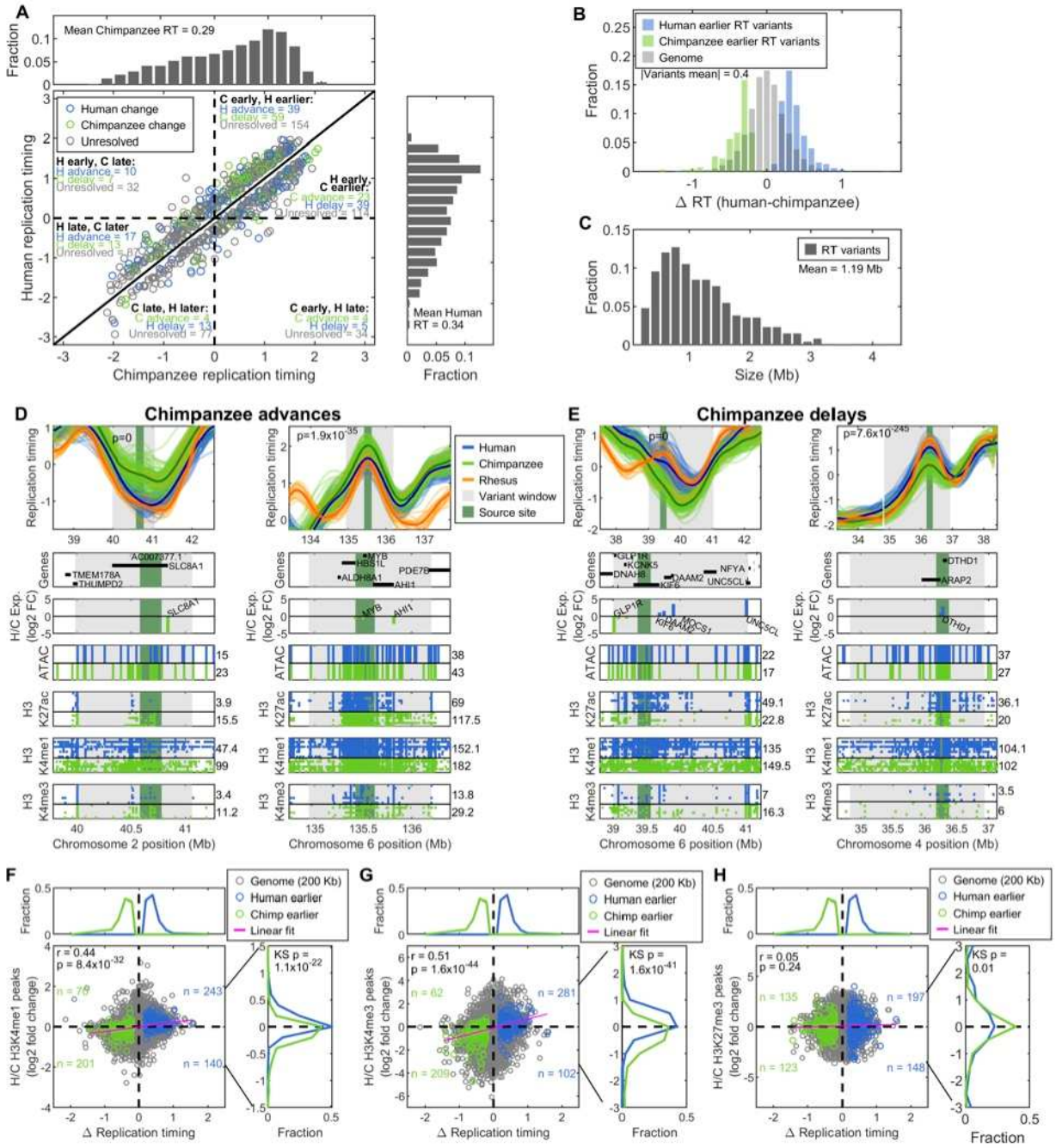


Figure 2.6. Further characterization of human-chimpanzee replication timing variants.
(A) Comparison of human and chimpanzee replication timing at variant regions (averaged data within each source site). Blue, green and gray data points represent variants resolved as having evolved in the human or chimpanzee lineages, or being unresolved, respectively. Top and right histograms: distributions of chimpanzee and human replication timing, respectively, within replication timing variant regions. (B) The distribution of the magnitudes of human-chimpanzee replication timing differences (averaged data within source sites of each variant). Background genome replication timing differences were calculated for each replication timing window across the genome. (C) Size distribution of human-chimpanzee replication timing variant regions. (D, E) Similar to Figure 2 F-G, for chimpanzee replication timing advances and delays. (F-H) As in Figure 2 J, but using human and chimpanzee LCL H3K4me1 (F),

H3K4me3 (G), and H3K27me3 (H) ChIP-seq data.

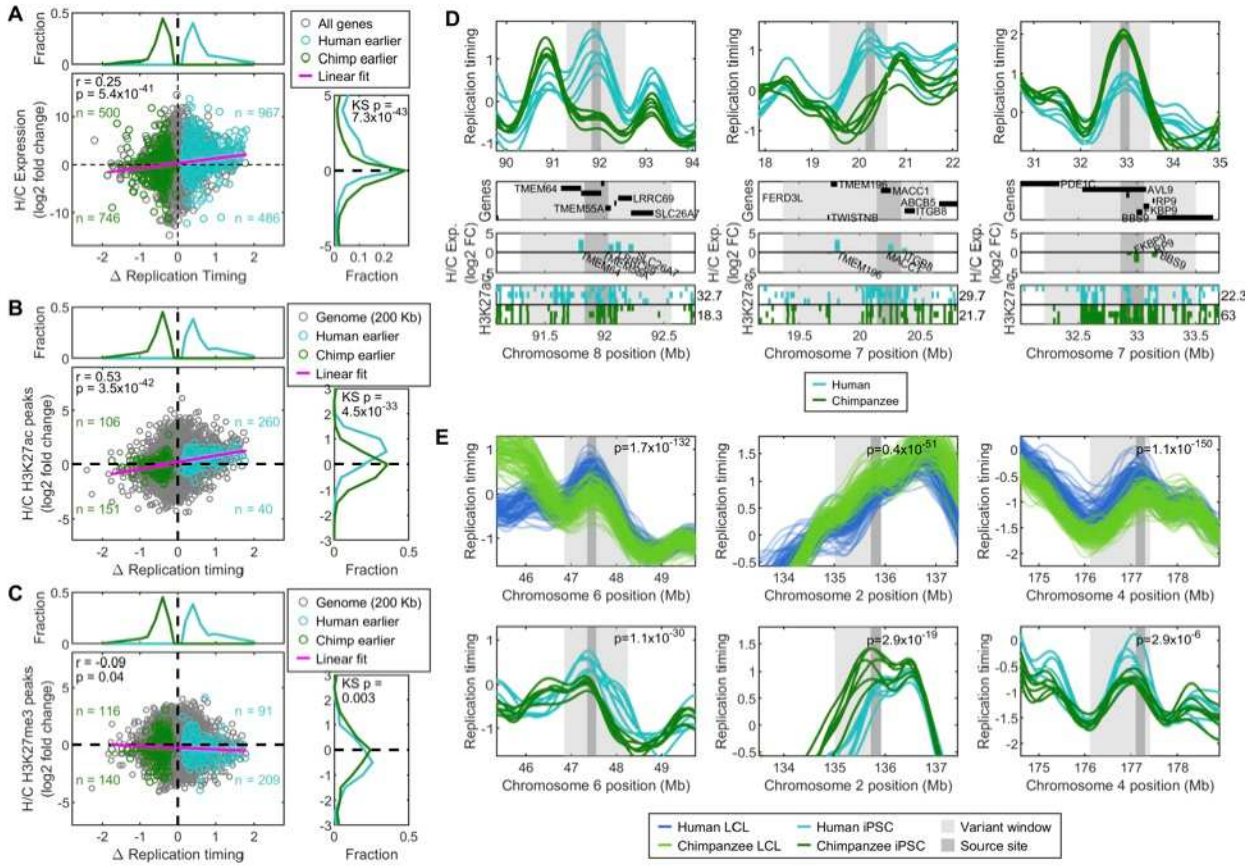


Figure 2.7. iPSC replication timing variants.
(A) Differences between human and chimpanzee iPSC replication timing compared to differences in gene expression, as in Figure 2H. (B-C) As in Figure 2J, but using human and chimpanzee iPSC H3K27ac (B) or H3K27me3 (C) ChIP-seq data. (D) Examples of iPSC human-chimpanzee replication timing variant regions. (E) Examples of human-chimpanzee LCL replication timing variant regions observed also in iPSCs. Top subplots show the LCL replication timing data and indicate the LCL variant window with p-value. Bottom subplots show the iPSC data for the same variant regions, p-values indicate the significance (ANOVA) of human-chimpanzee iPSC differences within the LCL variant region.

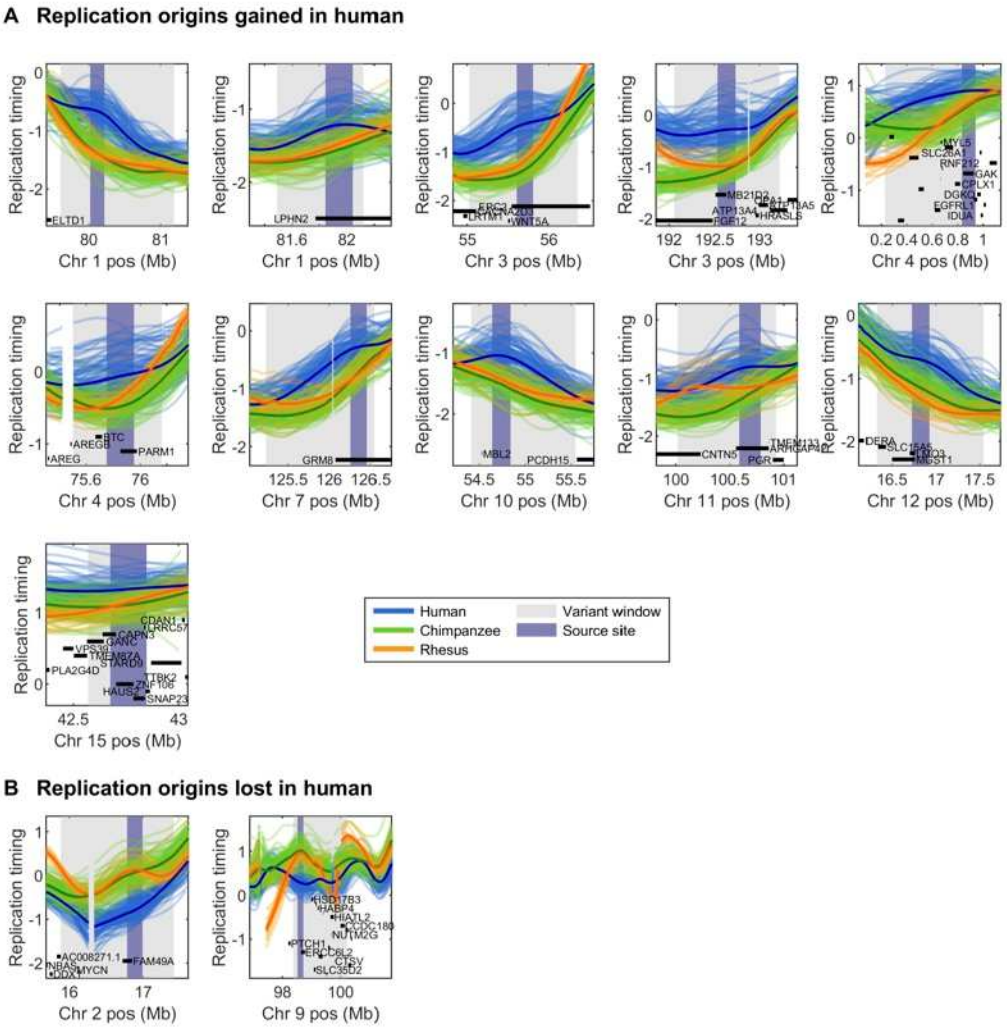


Figure 2.8. Gained and lost replication origins in humans.
Example regions with replication origins gained (A) and lost (B) in humans. Human-chimpanzee replication timing variant window and source site are indicated for each region along with genes that fall into each region. Some gene names removed for readability.

Association of replication timing variation with gene evolution

iPSC variant regions spanned 2,801 protein coding genes (mean of 5 genes per variant), while LCL replication timing variants spanned a total of 6,156 protein coding genes (mean of 8.5 genes per variant). iPSC replication timing variant regions were enriched for genes involved in immunity and development (Table S2), some of which are known to be under adaptive evolution

in humans (e.g. AKAP11, GLB1L2, SYBU, CD59, PYHIN1, PYDC2, SIGLEC9/L1, ADAM2, OVGP1, SEMG1, SEMG2, ANG) (Gayà-Vidal and Albà, 2014; Vallender and Lahn, 2004). Similarly, several genes inferred to be under positive selection in humans fell into LCL replication timing variant regions. These genes included several with roles in cell cycle progression (TLE6; Figure 2.9 H), Wnt signaling (TLE4), and sperm motility (CATSPER1, SEMG1, SEMG2), and several associated with human diseases or conditions including Usher Syndrome (USHBP1; Figure 2.9 I), glaucoma (RMDN2; Figure 2.9 J), intellectual disability (KPTN) and microcephaly (ASPM; Figure 2.9 K). One notable LCL variant region spanned the APOBEC cluster that includes APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H; these genes play a role in antiviral activity and most have been under positive selection in primates (Sawyer et al., 2004) (Figure 2.9 G). APOBEC genes replicated earlier in humans, and most fell within the source site of replication timing variation.

We identified 877 protein coding genes with variable replication timing between human and chimpanzee in both LCL and iPS cells. One notable gene under positive selection in humans, PYHIN1 (IFIX), replicated later in humans compared to chimpanzees (and macaques; Figure 2.9 L). This gene is a known tumor suppressor, down regulation of which is associated with breast cancer (Ding et al., 2004).

As a complementary analysis, we examined replication timing evolution at various genomic elements previously described to undergo atypical rates of evolution. These included human-chimpanzee divergent sites; more specifically human accelerated regions (HARs), which are conserved in mammals yet have undergone many sequence changes in humans (Hubisz and Pollard, 2014); and regions identified as under ancient positive selection in humans (selective

sweeps (Peyrégne et al., 2017)). Conversely, we analyzed regions under evolutionary constraint: loss of function intolerant genes (gnomAD) (Karczewski et al., 2020)), and ultra-conserved elements (UCEs) that are completely conserved in sequence across human, mouse, and rat (Bejerano et al., 2004). As expected, sites of sequence divergence (Figure 2.2 F-G) and HARs (although not sites of selective sweeps; Figure 2.9 C-D) were biased to late replication, while UCEs and loss of function intolerant genes replicated earlier than expected (compared to genes in general in the latter case; Figure 2.9 A-B). More significantly, we also found that divergent sites and regions under ancient positive selection in humans (but not HARs) were enriched in replication variant regions (focusing on variants in iPSC – the cell type better reflecting the germline; Figure 2.9 C, D, F). On the other hand, loss of function intolerant genes, as well as all protein coding genes, were found to be significantly depleted in iPSC replication timing variant regions (Figure 2.9 B, E). Taken together, these results suggest that replication timing alterations are unfavorable at conserved regions, possibly because they have an impact on genome function. Conversely, sequence divergence appears to be associated with replication timing differences between species.

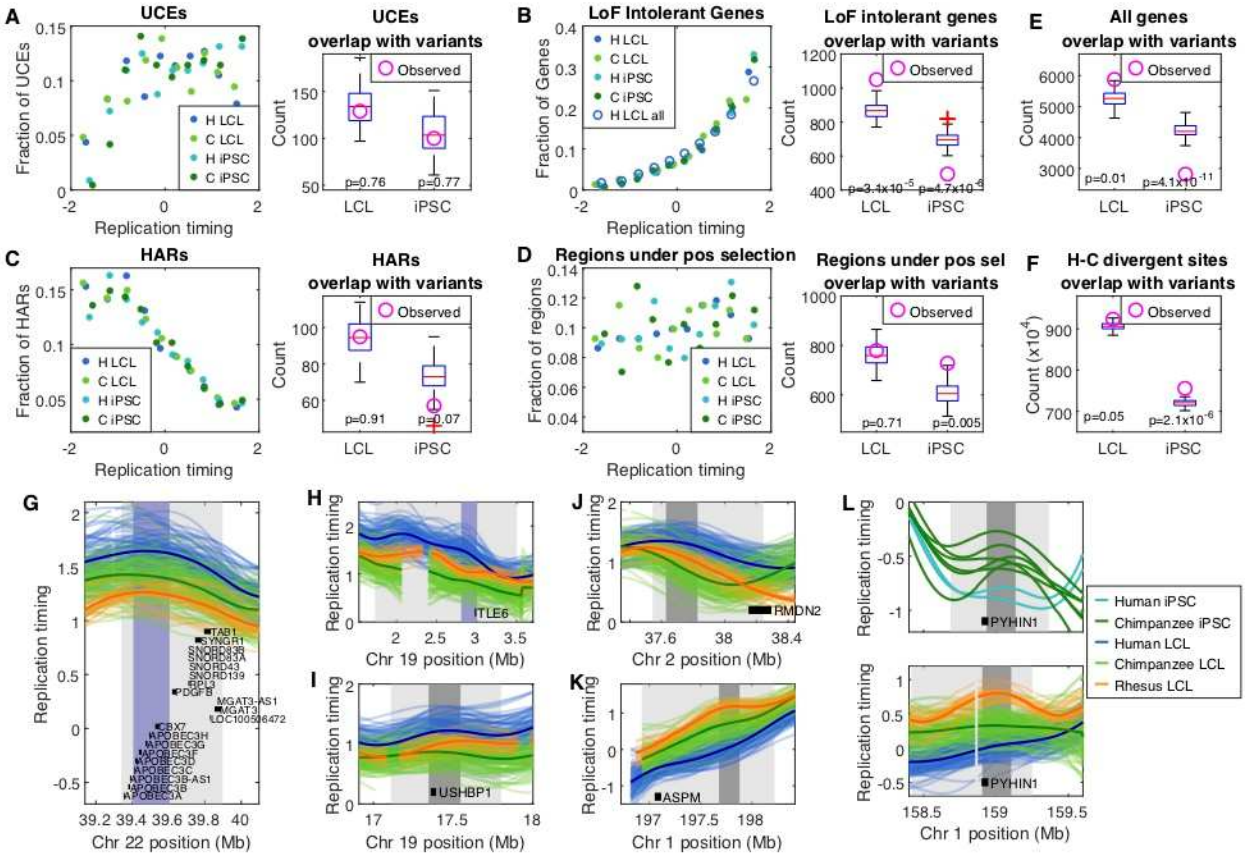


Figure 2.9. Replication timing at regions under constraint or adaptive evolution.
(A-D) Left subplots: fraction of ultra-conserved elements (UCEs, A), loss of function (LoF) intolerant genes (B), human accelerated regions (HARs, C), or regions under positive selection in human (D) in 10 replication timing bins compared to the mean LCL or iPSC replication timing of that bin per species. (B) Includes human LCL replication timing at all protein-coding genes for comparison to LoF intolerant genes. (A-D) Right subplots and (E-F): the overlap of each genomic feature with LCL and iPSC replication timing variants. Boxplots indicate overlap of the genomic feature with random genomic regions (number, size, and replication timing matched to the variants; 100 randomizations); magenta circle indicates the actual number of each genomic feature that overlaps the replication timing variants. P-values were calculated with a Z-test. (E-F) See Figure 1 for replication timing at all protein-coding genes (E) or human-chimpanzee divergent sites (F). (G) Human-chimpanzee LCL replication timing variant spanning the APOBEC gene cluster. (H-K) Examples of genes under adaptive evolution in humans that fall into human-chimpanzee LCL replication timing variants. (L) PYHIN1 falls within a shared LCL-iPSC human-chimpanzee replication timing variant.

A complex association between DNA replication timing and gene regulation

Since replication timing is correlated with genome regulation (e.g. gene expression, chromatin accessibility; Figure 2.2), we tested whether replication timing variation was itself

correlated with differences in gene expression or chromatin accessibility. Indeed, replication timing differences were positively correlated with gene expression variation (LCL: r=0.22, iPSC: r=0.25) and most replication timing variants (LCLs: 407/731, 56%, z-test $p=7.0x10^{-4}$; iPSCs: 312/557, 56%, z-test $p=9.2x10^{-4}$) contained predominantly genes with inter-species gene expression variation that corresponded to the direction of replication timing variation (i.e. earlier replication associated with elevated gene expression, later replication with reduced gene expression) (Figure 2.5 H; 2.7 A). Similarly, we observed a positive correlation between replication timing variation and chromatin accessibility, assessed using ATAC-seq (LCL: r=0.35) and the histone modifications H3K27ac (LCL: r=0.59; iPSC: 0.53), H3K4me1 (LCL: r=0.44), and H3K4me3 (LCL: 0.51): the earlier replicating species had a relatively higher density of the open chromatin marks compared to the later replicating species (Figure 2.5 I-J; 2.6 F-G; 2.7 B). In contrast, density of the repressive chromatin mark H3K27me3 was not significantly correlated with replication timing variation (LCL: 0.05; iPSC: -0.09; Figure 2.6 H; 2.7 C). Overall, 90% of autosomal human-chimpanzee replication timing variant regions had concordant changes in replication timing and either gene expression or chromatin structure (based on H3K27ac, the histone mark most correlated to replication timing), and 52% (343/656) had concordant changes in all three.

To get a better understanding of the cause-and-effect relationships between DNA replication timing and chromatin, we analyzed their spatial co-variation. In some variant regions, chromatin structure differed between species primarily at the source site of replication timing variation (Figure 2.5 F, top left; 2.5 G, top right), suggesting that chromatin structure could be a determinant of the observed replication timing variation. In contrast, in other instances, differential chromatin structure/accessibility was present across the entire variant region (e.g.

Figure 2.5 G, top left and bottom left), suggesting that instead, replication timing could be exerting long range effects on chromatin structure. Similarly, there was no consistent spatial relationship between replication timing and gene expression variation; in some cases, gene expression varied concordantly primarily at the source site of replication timing variation (Figure 2.5 F, bottom right; 2.5 G, top right), while in other cases, concordant replication timing-gene expression variation extended across the entire variant (Figure 2.5 F top right; 2.5 G, bottom left). The evidence for each of these patterns across numerous genomic regions suggests that the interaction between replication timing and gene expression regulation is complex and locus-specific. As an extension, gene expression variation was not generally higher for genes in replication timing variant regions compared to non-variant regions (mean log FC of genes in variants=1.3, non-variants=1.4), together indicating that gene expression and replication timing variation, while often linked, are neither sufficient nor necessary drivers of one another.

The genetic basis of replication timing evolution

The differences between species described above are suggestive of past and/or ongoing evolution of replication timing. As an extension of this observation, ongoing evolution is expected to manifest as inter-individual variation within a given species. Indeed, we have previously shown that replication timing varies among humans at hundreds of genomic locations (Ding et al., 2021; Koren et al., 2014). Consistently, in the current LCL sample set we identified 185 human and 195 chimpanzee genomic regions with significant variation among individuals (Methods; Figure 2.10). Of those, 73 regions varied among individuals in both species, significantly more than expected by chance (18 expected; z-test $p=6.5x10^{-40}$) (Figure 2.10 C, F, G), while 112 regions were variable only among humans and 122 only among chimpanzees

(Figure 2.10 D, E, G). More than half of the intra-species variants were also identified as inter-species variants (Figure 2.10 D-G), including for variants that were shared across species (40/73; 22 expected; z-test $p=7.8x10^{-6}$) or those that were species-specific (63/112 human; 34 expected; z-test $p= 4.8x10^{-9}$; 57/122 chimpanzee; 34 expected; z-test $p=1.1x10^{-6}$). When directly testing human-chimpanzee variants for within species variation, 239 variants were also polymorphic in at least one of the species. Notably, 20 resolved human-evolved variants were also variable among humans, suggesting ongoing evolution of human replication timing in these regions. Taken together, we find significant evidence for replication timing polymorphism within both humans and chimpanzees, a substantial fraction of which appears to represent deep evolutionary processes that manifest as either conserved replication timing variation (Figure 2.10 C) or concomitant intra- and inter-species variation (Figure 2.10 D-F).

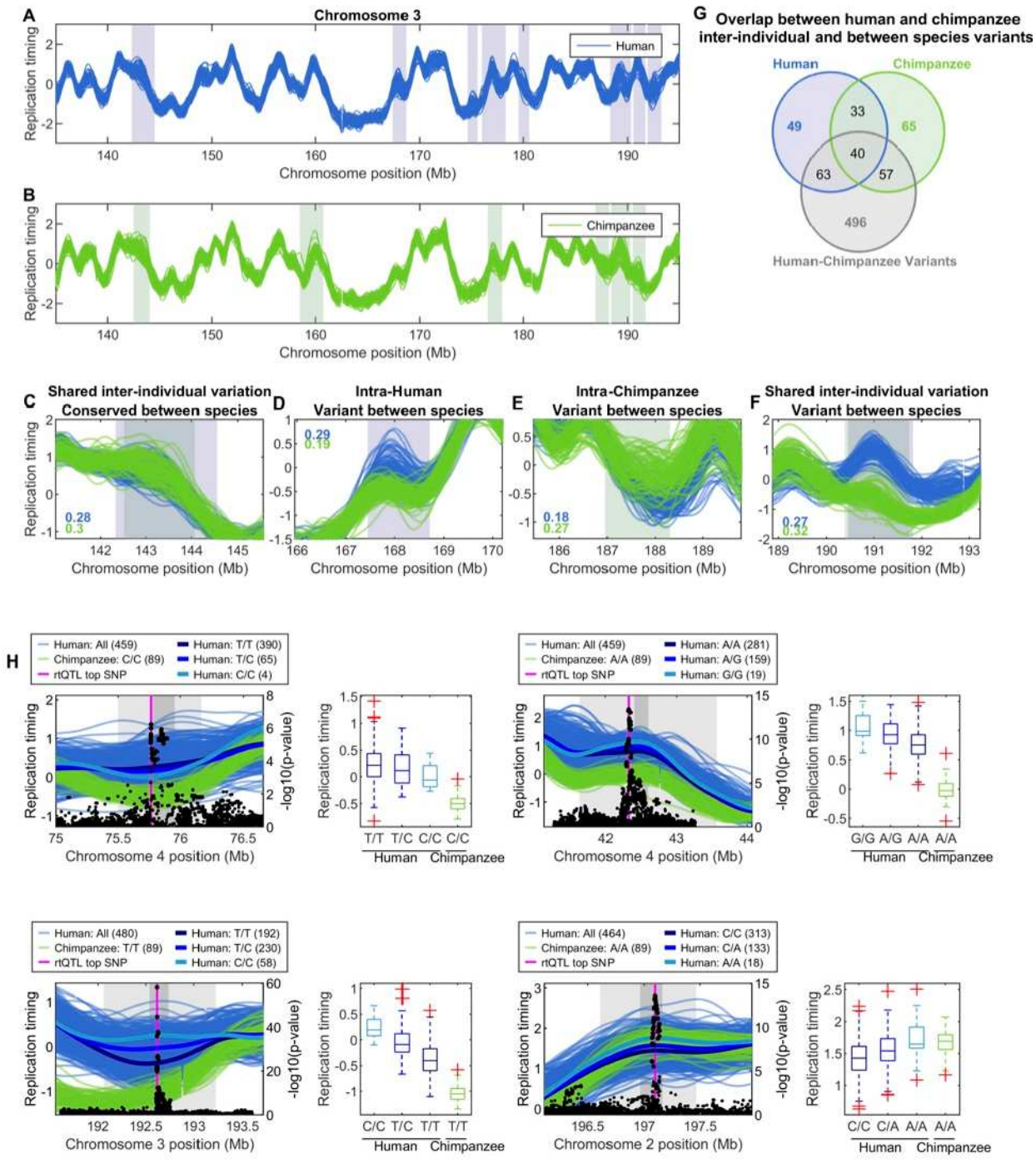


Figure 2.10. Genetic variation underlying inter-individual and inter-species replication timing variation. (A, B) Regions of inter-individual replication timing variation in human (A) and chimpanzee LCLs (B) for a section of chromosome 3. (C-F) Examples of shared and species-specific inter-individual replication timing variant regions. The numbers within each plot indicate the maximum standard deviation of replication timing values within the variant region for human and chimpanzee (blue and green, respectively). (G) Sharing of inter-individual replication timing variant regions between humans and chimpanzees and with between species (human-chimpanzee) variant regions. (H) Examples of human

rtQTLs that overlap human-chimpanzee replication timing variant regions. The top rtQTL SNP (magenta) falls within, or near, the source site of replication timing variation. Human 1000 Genomes data was used for replication timing profiles and boxplots. In three of the examples, replication was earlier in humans, and the chimpanzee allele at the top associated rtQTL SNP matches the late replicating human allele. The opposite direction, i.e., derived late replication in humans, is observed in the bottom right example.

Identifying intra-specific replication timing variation is particularly relevant in the context of this study, since such variation can be used to map replication timing quantitative trait loci (rtQTLs; (Ding et al., 2021; Koren et al., 2014)) which can then be tested for association with inter-species variation. Our population-level measurement of replication timing across species thus lends itself to the identification of the genetic basis of replication timing evolution.

To map rtQTLs in chimpanzees, we used fastQTL as recently described ((Ding et al., 2021); see Methods), controlling for the population structure and relatedness of our sample set (Figure 2.4 C-F). This unbiased genome-wide analysis identified 21 rtQTLs – a relatively small number which we ascribe to the limited sample size and relatedness of the 89 chimpanzees. To increase rtQTL discovery power, we further mapped chimpanzee rtQTLs directly in regions of chimpanzee inter-individual replication timing variation and human-chimpanzee replication timing variation. This identified an additional 31 rtQTLs among the 195 chimpanzee inter-individual variants, and a further 33 in the 656 autosomal human-chimpanzee variant regions (Figure 2.11).

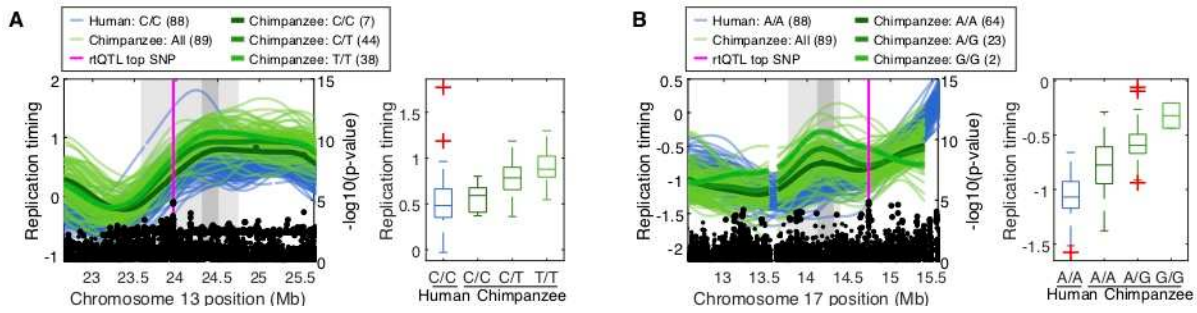


Figure 2.11. Chimpanzee rtQTLs.
(A-B) Examples of chimpanzee rtQTLs that overlap human-chimpanzee replication timing variant regions. Human data from this study was used for replication timing profiles and boxplots. In both examples, replication was later in humans, and the human allele at the top associated rtQTL SNP matches the late replicating chimpanzee allele.

To compare replication timing polymorphisms to genetic variation in the current human LCL samples, we took advantage of the much larger number of 1,775 rtQTLs that we previously mapped in human stem cells (Ding et al., 2021) and 3,752 rtQTLs that we independently mapped in LCLs from the 1000 Genomes project (our unpublished results). We validated 276 of the 1000 Genomes rtQTLs directly in the current human samples (out of 2,793 rtQTLs for which we had all three genotypes at the top associated SNP location) and also verified that most human inter-individual variants from the current study overlapped rtQTLs in the 1000 Genomes dataset (141/185, Z-test $p=8.3x10^{-7}$; 83 overlapped human rtQTLs validated in current samples). Thus, rtQTLs reflect the genetic basis of replication timing variation at a substantial fraction of the sites we mapped in this study.

Since we observed a high concordance of within species variation with between species variation (Figure 2.10 A-G), we predicted that human rtQTLs will also be associated with replication timing variation between humans and chimpanzees and could thus shed light on the genetic evolution of this form of variation. Indeed, 187 LCL and 57 iPSC human-chimpanzee variant source sites overlapped an rtQTL top associated SNP in the respective cell type,

significantly more than expected based on randomizations (LCL: z-test $p=3.5x10^{-33}$; iPSC z-test p=0.004). Since some rtQTL associated SNPs affect replication timing at a distance, we also confirmed that source sites were enriched in rtQTL affected regions in addition to rtQTL SNPs per se (LCL: z-test $p=7.5x10^{-4}$).

To test whether rtQTL sequences, at least in part, stand at the basis of replication timing evolution, we asked whether the derived allele matched the evolved replication timing state. For example, we would predict that humans carrying the ancestral allele for an rtQTL would have the ancestral replication timing (i.e. similar to chimpanzees), while humans with the derived allele would have the derived (i.e. different) replication timing state. We tested this prediction on human rtQTLs that spanned an inter-species variant source site and used the top associated rtQTL SNP and strongly linked SNPs (LD>0.8). We found a strong enrichment of rtQTLs where the human-derived allele was associated with the evolved replication timing state, while the ancestral allele was associated more closely with the chimpanzee replication timing (at least one tested SNP for 1,249/1,605 rtQTLs, 78%, permutations p=0.0072; >50% of SNPs for 741/1,605, 46%, p=0.0012; see examples in Figure 2.10H). Of these rtQTLs, 227 spanned human-chimpanzee variant regions that were resolved as changes in the human lineage. In 215 of these rtQTLs (95%), the chimpanzee allele of at least one tested SNP in high LD matched the macaque allele, suggesting that the genetic association may be sustained throughout the primate lineage as well.

The same analysis for the chimpanzee rtQTLs revealed that the chimpanzee derived allele matched the evolved replication timing state for 18 out of the 44 chimpanzee rtQTLs mapped in human-chimpanzee variants, suggesting that the derived chimpanzee allele was contributing to the difference in replication timing between humans and chimpanzees in these regions.

Importantly, 30 of the 44 chimpanzee rtQTL associated regions (mapped in human-chimpanzee replication timing variants) were also shared with a human rtQTL (expected 21, z-test p=0.005). This was not the result of ancient polymorphisms with conserved effects on DNA replication timing, as humans and chimpanzees did not share the associated rtQTL SNPs. Instead, this suggests that independent genetic contributions influence the replication timing of a given region across species, while rtQTL sharing further reflects either evolutionary pressures to maintain replication timing polymorphisms, or relaxed selective constraints to fix replication timing at these loci.

Shared genetic causes of replication timing and gene expression evolution

We showed above that the evolution of DNA replication timing can be ascribed to sequence evolution while it also impacts regulatory evolution. Considered jointly, and further with the sequence determinants of gene regulation, these observations could potentially reveal how gene regulation and DNA replication timing have co-evolved. We previously showed that, across humans, replication timing and gene expression variation often share genetic causes (Ding et al., 2021). We thus took advantage of comprehensive mapping of gene expression QTLs (eQTLs) in LCLs by the GTEx consortium (Lonsdale et al., 2013), and compared them to the top associated SNPs (and/or SNPs in LD>0.8 to that top SNP) of the rtQTLs we found to be associated with replication timing variation between humans and chimpanzees. We found 488 rtQTLs (out of 1,605 that overlap human-chimpanzee variants) were also significant eQTLs (q-value<0.05; 192 unique variant regions). At these eQTLs, 64% of the involved genes (194 out of 301 for which expression data was available) had concordant changes in gene expression and replication timing, suggesting shared genetic causes of replication timing and gene expression

evolution.

A notable example was a human-chimpanzee variant region that was both an rtQTL (Figure 2.12 A) and an eQTL for two protein coding genes (Figure 2.12 C) and one lincRNA. The rtQTL top SNP was the same as the top eQTL SNP (rs7806550), and there were no SNPs within 10 Kb of the variant with LD>0.4 (Ensembl 1000 Genomes YRI LD). Among human populations, the ancestral allele frequency for rs7806550 was highest in African populations (19%) and much lower in out of Africa populations (0-4%) (Figure 2.12 B). This suggests that the derived allele emerged in the common ancestor of humans and increased in frequency to become the major allele in modern day humans. The region impacted by this shared rtQTL-eQTL was earlier replicating in humans than chimpanzees, and the two protein coding genes associated with the eQTL, ITGB8 (integrin complex subunit that mediates cellular interactions) and MACC1 (regulator of hepatocyte growth factor receptor involved in cell growth and motility), were also more highly expressed in humans (Figure 2.12 A). Although rs7806550 is not known to be associated with any human phenotype (GWAS catalog; (Buniello et al., 2019)), it fell within a strong LCL enhancer, and was predicted to affect two transcription factor binding motifs – for GATA and for HDAC2 – where the alternate allele (T; also ancestral allele) has higher binding affinity (Figure 2.12 D). HDAC2 catalyzes deacetylation of lysine residues at the N-terminal regions of core histones (H2A, H2B, H3 and H4) and we previously showed that HDAC2 binding is associated with late replicating rtQTL alleles (Ding et al., 2021). In this specific example, the ancestral allele (with higher HDAC2 binding affinity) was later replicating and matched chimpanzee replication timing. These observations can be explained if the human derived allele interrupts the HDAC2 binding site, decreases its ability to bind and deacetylate histones in the area, thus leading to greater histone acetylation, greater chromatin accessibility,

and ultimately earlier replication and higher expression levels of genes in the immediate area. Interestingly, we also identified this region as variant between human and chimpanzee iPSCs (Figure 2.7 D, middle), and previously showed it to be the location of a human iPSC rtQTL (Ding et al., 2021). Overall, this indicates that sequence changes may coordinate the concomitant evolution of replication timing and gene expression, through a chromatin intermediate.

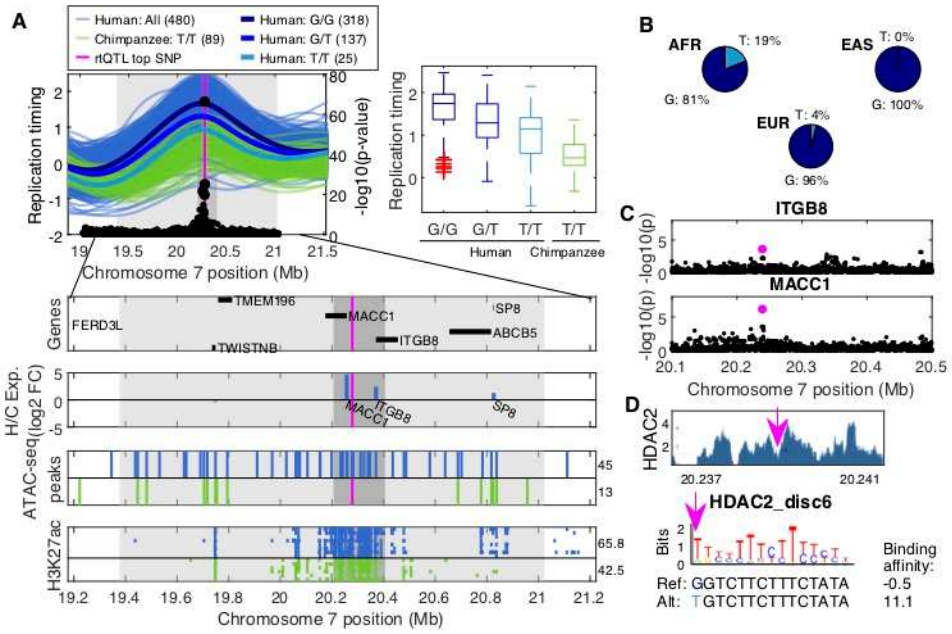


Figure 2.12. A genetic variant affecting HDAC2 binding, DNA replication timing and regional gene expression.

(A) Chimpanzee replication timing profiles (this study) together with all replication timing profiles from the 1000 Genomes African super-population (AFR) and averaged replication profiles per genotype at the top associated rtQTL SNP (rs7806550). Genes, expression levels, ATAC-seq and H3K27ac density shown below. Magenta: rs7806550, located between the genes ITGB8 and MACC1. (B) AFR, East Asian (EAS), and European (EUR) 1000 Genomes phase 3 allele frequencies for rs7806550. (C) ITGB8 and MACC1 LCL eQTLs (GTEx). Magenta: rs7806550. (D) A HDAC2 regulatory motif is altered by the rtQTL-eQTL top SNP (rs7806550), with the alternate allele (T) having higher binding affinity. Top: Encode HDAC2 ChIP-seq data for GM12878 (2012). Bottom: Sequence logo of HDAC2 regulatory motif that is altered by rs7806550 (magenta arrow). Binding affinities are from HaploReg (Ward and Kellis, 2011) (Δ affinity = 3,104-fold).

Discussion

A long-standing question in human biology is what are the genetic changes that distinguish us from other species? The significant sequence similarity between humans and our ape relatives has pointed to regulatory evolution as a likely explanation of our unique phenotypes (Fraser, 2013; King and Wilson, 1975). However, studies of the evolution of gene expression and other epigenetic features has fallen short of fully explaining the complex adaptations in the human lineage. An understudied biological process from an evolutionary standpoint has been DNA replication timing, a fundamental genomic process that bridges genome regulation and maintenance. Previous studies of replication timing evolution have been restricted by small sample sizes, limiting their ability to fully describe evolutionary alterations and reveal the genetic drivers and the impacts of replication timing evolution. This has also limited the understanding of the forces that drive replication timing evolution and thus understanding of its functional significance. Here, we utilized population-scale replication timing profiling of humans, chimpanzees, and rhesus macaques to identify hundreds of genomic locations that vary in replication timing within and between these species, regulatory features that co-vary with replication timing across species, and sequence variants associated with replication timing evolution.

Notwithstanding local variation, the majority of the genome exhibits highly conserved replication timing both between and within species. This is unlikely even for lack of input sequence variation, as we have previously shown that multiple sequence determinants, spread over areas spanning several megabases, can influence the activity of any given replication origin (Ding et al., 2021). For the replication timing variants that we do observe, most are quantitative (changes in replication origin firing times), and none can be considered of very large magnitude

(e.g., >half of S phase). Thus, it appears that DNA replication timing is largely under evolutionary constraint and thus likely harbors an essential function(s), consistent with previous studies (Müller and Nieduszynski, 2012; Ryba et al., 2010; Yaffe et al., 2010). On the background of this overall conservation, we provide evidence for replication timing evolution across more than 30% of the human genome, including 123 genomic regions that have specifically evolved in the human lineage. This extent of inter-species replication timing differences is on par with gene expression evolution (e.g., 24% of genes in LCLs; (Khan et al., 2013)) and far exceeds the ~1% sequence divergence between humans and chimpanzees. This illuminates the functional potential of DNA replication timing in human evolution, with sequence alterations that affect replication timing carrying a disproportional effect on the genome compared to other regulatory sequence adaptations. The null assumption should be that these alterations are evolutionary neutral, which would be consistent with the observation that replication timing evolution mimics the phylogenetic tree of the same studied species. Shared variation between and within species, and shared rtQTLs (similar to previous observations of shared eQTLs; (Fair et al., 2020; Jasinska et al., 2017; Tung et al., 2015)), could also be a reflection of neutral drift and lack of local constraint, although an intriguing alternative possibility in this case is the action of balancing selection. Nonetheless, we did identify numerous genes with evidence of having undergone positive selection in humans as implicated in inter-species replication timing variant regions. This, and the general correlations with variation in gene regulation and potentially with mutation rates (see further below), points to the possibility that a subset of replication timing-evolved regions carry important functional implications for human evolution.

Studying a large number of individuals provided the unique ability to detect replication

timing variation concomitantly between and within species. We observed extensive overlap of variation within and between species, pointing to deep and ongoing evolutionary processes impacting replication timing. This overlap also highlights the value of large sample sizes in evolutionary studies of replication timing, since variation between species is likely obscured in studies using smaller sample sizes. Furthermore, identifying and comparing variation within and between species enabled us, for the first time (to our knowledge), to reveal some of the genetic determinants of human replication timing evolution. We did this using an rtQTL mapping approach, which we applied to each species separately and then combined by considering co-occurring inter-species replication timing variation and linking derived rtQTL alleles to derived replication timing states. We anticipated that a much larger fraction of replication timing variants is determined by sequence evolution and could have been revealed with a larger samples size and hence greater power to detect rtQTLs.

Identification of genetic determinants of replication timing evolution provides a means for revealing mechanisms of replication timing control as well as the population genetic and evolutionary dynamics of replication timing. A notable example we highlighted pertains to the role of histone acetylation. HDAC binding has been described previously as a repressor of replication timing (Ding et al., 2021; Goren et al., 2008), likely by promoting a more repressive chromatin state. Here, we showed that a sequence polymorphism impacting an HDAC2 binding site within a human LCL enhancer has likely led to variation in both replication timing and gene expression within humans as well as between humans and chimpanzees. The high frequency of the derived allele in humans suggests a likely directional evolution of this variant and therefore of replication timing in this case.

More generally, and consistent with previous studies, we observed correlated co-variation

of DNA replication timing, chromatin accessibility and gene expression. A notable advantage of our study, however, is the ability to interrogate these relationships across hundreds of genomic regions harboring natural variation in DNA replication timing. While previous studies typically described these correlations as indicative of unidirectional causality relationships (in either direction; (Blin et al., 2019; Klein et al., 2021; Marchal et al., 2019; Müller and Nieduszynski, 2017; Rivera-Mulia et al., 2015; Zhang et al., 2002)), our study suggests a more complex picture of replication timing and gene expression co-varying and potentially affecting each other (with chromatin structure being a likely intermediary) in a cell type-, locus- and context-specific manner.

Another important functional aspect of replication timing is its influence on the mutational landscape and therefore on local sequence evolution. Beyond validating the correlation between replication timing and rates of mutation and sequence variation, we found that sequence divergence between human and chimpanzee was elevated specifically in replication timing variant regions (in iPSC in particular). This suggests that evolutionary changes in replication timing could potentially alter local mutation rates and patterns. This would also be consistent with the observation that protein-coding genes (especially loss of function intolerant genes) were generally depleted in the same variant regions; thus, replication timing alterations in regions with highly conserved genomic function may be unfavorable, possibly due to the dual relationship of replication timing with genome regulation and genome maintenance.

More comprehensive mutational data would be required in order to test with sufficient statistical power how replication timing affects sequence evolution. The ability to obtain additional replication timing and mutation data for chimpanzees is, however, notably limited by the scarcity and regulatory limitations of using chimpanzee material. An alternative is to study

replication timing variation within the more limited evolutionary timescale of human populations, or the broader timescale of more diverged mammalian species such as rodents compared to primates.

Another critical observation is that replication timing evolution is predominantly cell-type-specific. iPSCs and LCLs had very little overlap of inter-species replication timing variants. By inference, other cell types can be expected to show replication timing evolution at yet other genomic locations, and therefore the full functional impact of replication timing evolution would only be possible to evaluate in a larger number of cell types.

Overall, our findings highlight the importance of replication timing evolution as both a driver and consequence of sequence and regulatory evolution. DNA replication timing may thus carry an important yet previously under-considered role in human evolution. As such, it would be highly informative to incorporate replication timing in future studies of sequence and epigenetic evolution.

Methods

Sample preparation and whole genome sequencing

Genomic DNA from ninety chimpanzee lymphoblastoid cell lines (LCLs) (Table S3) was sourced from the Coriell Institute (Camden, NJ). These cell lines were originally derived at the Yerkes Primate Center in Atlanta, GA (samples from animals received prior to 2015). Genomic DNA from 85 human LCLs was sourced from the Coriell Institute and three LCLs from Grant Stewart (Table S3). Twenty-three rhesus macaque LCLs and a panel of seven human and seven chimpanzee induced pluripotent stem cell lines (iPSCs) were obtained from the Gilad Lab at the University of Chicago (Table S3) (Romero et al., 2015). LCLs were cultured in Roswell Park

Memorial Institute 1640 media with 15% fetal bovine serum, and iPSCs were cultured in mTeSR Plus media. Approximately one million cells of each cell lines were flash-frozen, and genomic DNA was extracted from them using the MasterPure DNA Purification kit (Epicentre, Madison, WI, USA). Genomic DNA from one additional human iPSC (AG25370) was obtained from the Coriell Institute.

All DNA samples were sequenced to approximately 20x coverage using the Illumina HiSeq X Ten with 2x150 paired-end reads (GENEWIZ, Inc., South Plainfield, NJ, USA). Chimpanzee LCLs were sequenced across two separate batches with two samples re-sequenced across batches as a control (NS03621 r=0.99; NS03639 r=0.97); clustering of replication timing values did not show evidence of a significant batch effect (batch 1 vs. 2 mean r=0.92). Human LCLs were sequenced in either batch primarily for other purposes, but used as batch-matched controls for this project. Rhesus macaque LCLs were sequenced in a third batch, with two chimpanzee LCL samples re-sequenced across batches as controls. All iPSC samples were sequenced in the same batch.

We were not able to generate reliable replication timing profiles for two chimpanzee LCLs, one chimpanzee iPSC, and two human iPSC samples. These samples had much lower mean correlation to the rest of the samples (chimp LCLs r=-0.03, -0.60; human iPSCs r=0.76, 0.83; chimpanzee iPSC r=0.8) and were removed from further analysis. One of the low-quality chimpanzee LCL samples was sequenced separately to approximately 40x coverage and yielded good quality data, thus we included these additional sequence data for this sample in our analyses; this led to a total of 89 chimpanzee LCL samples that were used for further analysis.

Chimpanzee relatedness and population structure

We used a modified genome analysis toolkit (GATK; v4.1.4.0) best practices pipeline (Van der Auwera et al., 2013) to call SNPs and indels across the 89 chimpanzee LCL samples. We recalibrated base quality scores using chimpanzee dbSNP locations (prior to recalibration, we converted the dbSNP locations from the panTro5 to the panTro6 reference genome using the GATK tool LiftoverVcf (Picard; http://broadinstitute.github.io/picard/)). After recalibration, we genotyped each sample separately, then joint-genotyped all samples, as per the best practices. This resulted in a total of 28,476,465 variants including 23,911,740 SNPs and 4,125,969 indels. Next, we filtered the resulting variants with hard filtering thresholds based on recommendations from GATK (SNPs: QD<2.0, MQ<40.0, FS>60.0, SOR>3.0, MQRankSum<-12.5, ReadPosRankSum<-8.0; indels: QD<2.0, ReadPosRankSum<-20.0, InbreedingCoeff<-0.8, FS>200.0, SOR>10.0). This resulted in 26,492,303 total variants including 22,442,083 SNPs and 4,050,220 indels.

We used vcftools to evaluate the overlap of our called genetic variants with other datasets. Our variants overlapped 817,964/1,034,979 (79%) dbSNP variants, 9,375,559/25,923,958 (36%) variants from Prado-Martinez et al. 2013 (Prado-Martinez et al., 2013), and 7,921,456/24,469,855 (32%) variants from de Manuel et al. 2016 (De Manuel et al., 2016).

Genotype PCAs were generated using the R package SNPRelate (Zheng et al., 2012). We pruned genotypes for minor allele frequency (<0.05), missing rate (>0.1) and linkage disequilibrium (>0.1) using snpgdsLDpruning before generating a PCA with snpgdsPCA.

We used KING (Manichaikul et al., 2010) to calculate pairwise kinship values and infer relationships between chimpanzee LCL samples. Kinship values indicated a large number of first-degree relatives (77 pairs), thus we directly tested for the presence of first-degree familial

relationships within our sample set. We first identified chimpanzee trios where two individuals had a first-degree relationship with the same third individual. A trio could represent several configurations such as two parents and their shared offspring, a three-generation family, or one parent and their two offspring. We isolated 18 two-parent and shared offspring configurations by assessing Mendelian error rates among high population frequency SNPs: two parents homozygous for the same SNP should not produce a heterozygous offspring; however, this pattern can occur for the three-generation and parent with two offspring configurations.

Chimpanzee de novo mutation calling

We utilized the chimpanzee trios to call de novo (germline, or somatic cell line) mutations in the offspring of the trios and evaluate their relationship with replication timing. Genotyping, mutation identification, and candidate mutation filtering followed a pipeline we describe in detail elsewhere (Caballero et al., in preparation), modified for chimpanzee genomic resources. Briefly, BAM files were recalibrated with GATK using chimpanzee dbSNP. We did not recalibrate genotypes due to the lack of training resources. Candidate mutations were removed around the HLA locus (chr6:28,000,000–33,600,000 in panTro6). To remove inherited variants where the genotype was miscalled in a parent, we removed candidate mutations where any other LCL chimpanzee sample in this study contained reads matching the mutant allele. After all filtering steps, 14,774 autosomal mutations remained in the 18 LCL offspring (mean: 820.77, range: 273-1,439). We expect the majority of these to be cell line mutations, and a small number to be germline mutations (Koren et al., 2012).

Generation of DNA replication timing profiles

Human, chimpanzee and rhesus macaque whole genome sequencing data were aligned to their respective reference genomes (hg19, panTro6 and rheMac10) using BWA-MEM. We calculated GC-corrected sequencing read depth in 1 Kb uniquely alignable windows across each species' genome, as previously described (Koren et al., 2021). We merged 1 Kb to 10 Kb windows and then filtered the data as follows: for iPSC samples, we filtered out windows with CNVs and outlier data points using segmentation (MATLAB function *segment*) as previously described (Koren et al., 2021). For the LCL samples, instead of segmentation we used a "population" filtering method similar to the one we described previously (Ding et al., 2021). We calculated the median value (across samples) of each genomic window. We then used the median of these values across windows to represent the "common" copy number of the genome. Any window with a median of more than 0.4 copies (0.2 copies for male X and Y chromosomes) above or below this common number was removed. We repeated this "population" filtering method using the 25% percentile and separately the 75% percentile instead of median, which allowed to better capture outliers.

We further removed, in individual cell lines, genomic windows that were copy number outliers in specific samples (rather than across all samples). We removed data points that were at least 0.6 copies (0.4 copies for male X and Y chromosomes) above or below the common copy number (see above), or at least 0.35 copies (0.25 copies for male X and Y chromosomes) above or below the median copy number across samples of that specific replication timing window, in any particular sample. Together, these two parameters ensured the efficient filtering of absolute or relative outliers, respectively. For all samples (LCLs and iPSCs), we removed large (mostly chromosome level) copy number alterations, short (<500 Kb) segments between genome gaps and short (<100 Kb) segments between runs of missing data.

The filtered data was then smoothed between gaps ≥50 Kb and regions separated by ≥100 Kb with a cubic smoothing spline (MATLAB function csaps; parameter=10^{-17}) and subsequently normalized to a mean of zero and standard deviation of one.

We generated consensus replication timing profiles for each species per cell type by averaging the filtered data across samples (before smoothing). We then smoothed and normalized the averaged filtered data with the same parameters as described per sample (see above).

PC-correction of LCL profiles

We performed principal component analysis of raw filtered replication timing data for human, chimpanzee, and rhesus macaque LCLs separately and corrected each species' data for 10 principal components (PC10) using linear regression (Ding et al., 2021). The X chromosome for each species was corrected for male and female samples separately. We did not PC-correct the female macaque X chromosome since there were only three samples. We then smoothed and normalized the data as described above. We used the PC10 smoothed data for all analyses. We did not PC-correct the iPSC samples due to their low number.

Generation of G1/S replication timing profiles

G1/S replication timing profiles were generated as previously described (Koren et al., 2012). Briefly, approximately one million G1 and S phase cells for macaque LCL sample 76-06 and chimpanzee iPSC sample C3649 were sorted using a FACSAria Fusion (BD Biosciences, San Jose, CA, USA), DNA was extracted and sequenced as above. Following sequence alignment, we defined genomic windows of varying size, each encompassing 200 reads in the G1

data. We then counted the number of S phase reads that fell into those windows. Data was normalized to a mean of zero and standard deviation of 1. Windows of 100 Kb with standard deviation greater than 1.1 were removed as well as data points greater or less than 3.5 standard deviations. Filtered data was smoothed with a cubic smoothing spline (MATLAB function *csaps*; parameter=10^{-17}).

Comparison of WGS with Repli-seq

Chimpanzee lymphocyte and H2 Human iPSC Repli-seq data was downloaded from Replication Domain (https://www2.replicationdomain.com/database.php; Accessions: Int10455570, Ext30484475). The chimpanzee data was smoothed (MATLAB function *csaps*; parameter=10^{-17}) and normalized to a mean of 0 and standard deviation of 1. We then used the UCSC liftOver tool to convert genomic coordinates from hg38 to panTro6 (hg38.panTro6.rbest.chain) for comparison with our data. During this conversion, 6,533/419,622 (1.6%) of windows were lost. Linear interpolation was used to match genomic window coordinates prior to calculating the correlations with the chimpanzee LCLs in this study.

Replication timing window lift over

To compare replication timing profiles between species, we used the UCSC genome browser liftOver tool with the reciprocal best mapping chains to convert the center coordinate of the 10 Kb chimpanzee (panTro6) and rhesus macaque (rheMac10) replication timing windows to human (hg19) coordinates (panTro6.hg19.rbest.chain and rheMac10.hg19.rbest.chain, respectively). During this conversion, 6,615/270,666 (2.4%) of chimpanzee replication timing windows and 37,713/274,457 (13.7%) of rhesus macaque windows were lost.

Replication origin prediction

We predicted the most likely locations of replication origins based on the sharing among samples of peaks in the replication timing profiles. We called local maxima in each sample, then used hierarchical clustering with average linkage and a distance threshold of 300 Kb to identify clusters of recurrent nearby peaks across all samples of a given cell type (e.g. human, chimpanzee, rhesus macaque LCLs). We removed replication origins that were present in less than 10% of samples in at least one of the species. We further removed origin calls that occurred within 100 Kb of a mapped structural variant (SV) (Kronenberg et al., 2018; Soto et al., 2020) or a gap in the human genome.

Replication timing variation between humans and chimpanzees

To identify genomic regions with significant replication timing variation between species, we performed ANOVA tests comparing all samples from one species with all samples from the other species, in 200 Kb windows, sliding by 50 Kb, across all autosomes and the male and female X chromosome separately. Tested windows were considered to be significant if they passed a Bonferroni-corrected p-value threshold of 8.7×10^{-7}. Overlapping significant windows were then merged, and p-values were recalculated. These were considered as "initial variant regions". We excluded individual replication timing windows (10 Kb) from within these initial variant regions if they spanned genome gaps or that had a mean difference in replication timing between species of less than 0.2 standard deviations. These filters resulted in some of the initial variant regions being split or removed completely, yielding "filtered variant regions". Filtered variant regions that were less than 200 Kb long were removed. In cases in which adjacent filtered

variant regions had intervening replication timing windows with a mean replication timing difference between species greater than 0.2 (even if they were not significant in the initial ANOVA scan), we extended and merged these filtered variant regions, then recalculated the p-values of the merged variants. These were considered as "extended variant windows" and used for downstream analyses.

Since there is only one copy of the male X chromosome, we divided the filtering and extension thresholds above by 2 (i.e. used 0.1 standard deviation).

To classify the likely molecular type of replication timing variants, we tested each variant for overlap with predicted replication origin locations (peaks in the replication profiles) in each species. Variants harboring peaks (in >25% of samples) in both tested species were considered to be alterations in replication origin activation time, while variants with a peak in only one species (in >25% of samples) were considered to be an evolutionary gain or loss of a replication origin (more details below).

Next, we utilized the predicted origin locations to identify the most likely "source" sites of replication timing variation within each variant. We identified the called origins within each replication timing variant and considered only those that were present in >25% of samples of the species with earlier replication (including shared origins, in which both species had an origin in >25% of samples). Most variant regions contained only one origin, which was then considered as the source site. In cases with more than one origin within a variant region, if there was a valley in either the human or chimpanzee consensus profiles (or both, in which case the two valley locations were averaged) between the two origins, or otherwise if the origins were separated by more than 500 Kb, we split the variant region at the valley (or middle location, respectively) between the two origins and considered each origin to be a source site for its own variant.

Otherwise, we considered the source site to be the middle location between the origins. In either case, the source sites were regarded as 200 Kb regions centered at the origin locations, but were not allowed to extend beyond the bounds of the variant region.

Inference of directionality of evolutionary changes

To identify the specific lineage (human or chimpanzee) in which replication timing has likely evolved at replication timing variants, we compared each variant to the replication timing in rhesus macaques. This was done only for LCLs, for which we had data for all three species. For each variant, we calculated the pairwise Euclidean distance of consensus replication timing values between each species pair (i.e. human to chimpanzee, chimpanzee to macaque, human to macaque) at the source site and identified the species pair with the smallest distance. Variants for which the smallest distance was between humans and macaques were preliminarily considered to be evolutionary changes that occurred in chimpanzees, while replication timing variants that had the smallest distance between chimpanzees and macaques were preliminarily called as changes in humans. Replication timing variants for which the smallest distance was between humans and chimpanzees were considered unresolved in the absence of additional outgroups.

We then subjected the preliminary human and chimpanzee resolved changes to two quality filters. First, we required that the species pair with the smallest distance (see above) had replication timing profiles of similar shape. We assessed this for each preliminary resolved variant by calculating the correlation of consensus replication timing values within a 500 Kb window centered at the source site for the species pair with the smallest distance. If the correlation was less than 0.1, the region was re-categorized as unresolved. Second, we filtered the preliminary resolved changes where the species pair with the smallest distance (see above)

was high (>1, or >1.5 for regions where the human to chimpanzee distance was greater than 3) or where the macaque consensus profile was equally distant to the human and chimpanzee consensus profiles (the difference between the human to macaque and chimpanzee to macaque distances was <0.35).

We manually filtered an additional 48 regions that had a visually different replication timing profile shape in macaque (despite passing the correlation threshold), or where the macaque profile visually looked equidistant from human/chimpanzee as they were from each other.

Preliminary resolved variants that passed the filters comprised the final resolved variants, which we then categorized into advances and delays. If the species in which the evolutionary change was inferred to have occurred was earlier replicating that the other two, the variant was called as an advance in that species, while delays were considered to be cases in which the species that underwent replication timing evolution was later replicating than the others. We then used the previously classified molecular type of replication timing variants (see above) to subset the advances and delays into gains and losses of origins versus changes in origin activation time. In addition, gains required that there was no origin in greater than 25% of macaque samples, while losses required an origin in more than 25% of macaque samples. Hierarchically-clustered heat maps of replication timing values across samples were generated with the MATLAB function *clustergram*.

Replication timing variants shared between LCLs and iPSCs

To identify replication timing variants shared across cell types, we analyzed, for each human-chimpanzee LCL replication timing variant, the replication timing in human and

chimpanzee iPSCs. We calculated the pairwise Euclidean distances of human and chimpanzee LCL and iPSC consensus replication timing values at each replication timing variant source site. Variants were considered to be shared across the two cell types if the distances within species (e.g. human LCL to human iPSC; chimpanzee LCL to chimpanzee iPSC) were lower than the distances within cell types (e.g. human LCL to chimpanzee LCL; human iPSC to chimpanzee iPSC). Of the shared variants, we also asked if the shape of the profiles was consistent within species by calculating the Pearson correlation of consensus replication timing values within species (e.g. human LCL to human iPSC; chimpanzee LCL to chimpanzee iPSC). Shared variants were considered to be of similar shape if the within-species correlations were greater than 0.5.

Association with gene expression, chromatin accessibility, and sequence variation

Association between DNA replication timing and gene density, gene expression, chromatin accessibility (ATAC-seq and H3K27ac ChIP-seq), and sequence variation (SNP density, human-chimpanzee divergence, de novo mutations) was performed for human and chimpanzee LCLs and, separately, iPSCs when data was available. Data was obtained and prepared for analysis as follows:

We used the center coordinates (hg19) of Ensembl protein-coding genes. Gene expression analyses were based on published data for LCLs and iPSCs (Khan et al., 2013; Romero et al., 2015; Soto et al., 2020; Zhou et al., 2014). Chromatin data was obtained from the following sources: LCL ATAC-seq peak data (García-Pérez et al., 2021), LCL histone modification ChIP-seq peak data (Zhou et al., 2014), and iPSC ChIP-seq peak data (Romero et al., 2015). All chimpanzee ChIP-seq data was originally in the panTro3 reference genome while

ATAC-seq data was in panTro5; both were lifted-over to hg19.

Human and chimpanzee SNPs were obtained from dbSNP and filtered for coding sites. Divergent sites between human and chimpanzee were inferred from the Ensembl Enredo-Pecan-Ortheus (EPO) 12 primate multiple alignments (release 104) (n=32,301,278). We used liftOver to convert these sites from hg38 to hg19 (n=32,253,773 lifted). Replication timing windows with zero divergent sites were not considered in the following analyses.

For each of the genomic features described above, the following genome-wide analysis was performed: we binned replication timing windows into 30 equally portioned bins and counted the number of genomic features within each bin. Ten bins were used for de novo mutations, due to their small number. We subsequently compared the number of genomic features in each bin to the average replication timing value for that bin.

The following replication timing variant analysis was performed on each of the following data types (ATAC-seq, ChIP-seq H3K27ac, ChIP-seq H3K27me3, ChIP-seq H3K4me1, and ChIP-seq H3K4me3) separately, but will collectively be referred to as chromatin accessibility peaks. Prior to this analysis, we mapped human accessibility peaks back to panTro3 (ChIP-seq) or panTro5 (ATAC-seq) and removed human peaks that were not successfully lifted-over. We counted the number of peaks within each LCL or iPSC variant region and calculated log2 fold change in peak density normalized by the total number of human and chimpanzee peaks [log2((# human peaks in variant region/total # human peaks) / (# chimp peaks in variant region/total # chimp peaks))]. For comparison, we also calculated log2 fold change in peak density for 200 Kb windows across the genome. Log2 fold change in peak density was compared to the change in replication timing of the windows.

Gene ontology, constraint, and selection

Gene ontology enrichment was performed using the PANTHER Overrepresentation Test (Released 2022-02-02) (2021; Ashburner et al., 2000; Mi et al., 2019), separately on protein-coding genes that fell into LCL human-chimpanzee variants, iPSC variants, and protein-coding genes that were shared across the LCL-iPSC variants. Gene enrichment was calculated with the Fisher's Exact Test and evaluated with 5% FDR.

We analyzed LCL and iPSC replication timing at 481 ultra-conserved elements (UCEs) across humans, mice, and rats (Bejerano et al., 2004), loss of function intolerant genes (gnomAD (Karczewski et al., 2020)), 2,701 non-coding human accelerated regions (Hubisz and Pollard, 2014), human-chimpanzee divergent sites, all protein-coding genes, and regions under ancient positive selection in humans (Peyrégne et al., 2017). For each of these genomic features, we binned replication timing windows into 10 equally portioned bins and counted the number of genomic features within each bin. We then compared the number of genomic features in each bin to the average replication timing value for the bin. To evaluate if each genomic feature was enriched or depleted in replication timing variants, we randomized the locations of LCL and iPSC autosomal variant regions 100 times (size and replication timing matched, +/- 0.25 standard deviation) and calculated the total number of genomic features that each random region overlaps. The distribution of the total number of overlaps for each iteration was compared to the observed number of overlaps for each feature with a Z-test.

Inter-individual replication timing variation

We identified genomic regions with significant replication timing variation among individuals of a given species as regions with a relatively high standard deviation (SD) of the

replication timing data across individuals. We calculated SD across samples for each replication timing window across the genome (using replication timing data for each species own reference genome). To identify regions with the greatest regional SD, we first smoothed the SD values (MATLAB function csaps; parameter=10^{-14}) and then called peaks in the smoothed SD profiles. We removed peaks with SD lower than the mean of all autosomal SD peaks. The remaining SD peaks were considered to be centers of inter-individual variant regions. We then extended these regions until the closest local SD minima (identified from the smoothed SD profiles similar to peaks) or until the SD equaled the genome-wide mean SD. We filtered any extended variant regions that spanned gaps or that were shorter than 200 Kb and performed a pairwise t-test on the remaining variant regions. For each tested variant region, we identified significant sample pairs using a Bonferroni corrected p-value threshold. We removed variant regions that resulted from a single sample causing the observed variation. Chimpanzee variants were lifted-over from panTro6 to human hg19 coordinates; 10/195 regions failed to be fully lifted-over. We considered the intra-species variants as shared between humans and chimpanzees if a variant from one species overlapped the center of a variant from the other species or vice versa. We additionally tested human-chimpanzee replication timing variants for intra-species variants directly by the pairwise t-test and following filtering steps as described above.

rtQTL mapping and validation

Chimpanzee rtQTLs were mapped genome-wide, in human-chimpanzee replication timing variants, and within chimpanzee replication timing variants using fastQTL as in (Ding et al., 2021). Briefly, smoothed replication timing data was used with 10 phenotype principal components (PCs) and 3 genotype PCs as covariates. In the genome-wide analysis, phenotype

windows were chosen with a FDR<0.1, while phenotype windows for human-chimpanzee replication timing variants and within chimpanzee variants were chosen as the center variant window coordinate. For all three analyses, a SNP was identified as significantly associated with the phenotype if it belonged to a group of at least three consecutive SNPs with FDR<0.1. Merging and filtering of rtQTLs was performed as in (Ding et al., 2021).

1000 Genomes rtQTLs were mapped in six populations separately as a part of a separate study, using fastQTL as in (Ding et al., 2021).

Validation of 1000 Genomes rtQTLs in the human samples from this study was performed as in (Ding et al., 2021). Briefly, we calculated Pearson correlation between the top associated rtQTL SNP and the replication timing value at the location with strongest association with the rtQTL. We only tested rtQTLs where we had all three genotypes of the top associated SNP in the human samples from this study (2,793/3,752 rtQTLs). rtQTLs were classified as validated if the p-value was less than 0.05 and had the same direction of effect.

We obtained the chimpanzee and rhesus macaque allele at each rtQTL top associated SNP and SNPs in LD>0.8 to the top SNP from the Ensembl EPO primate alignments. We performed the following allele directionality analysis only on human rtQTLs where the associated region spanned a source site of human-chimpanzee replication timing variation. For each rtQTL, we evaluated the allele effect (i.e. whether the human derived or chimpanzee allele was earlier replicating) for the top associated rtQTL SNP and SNPs in LD>0.8 to that top SNP. We then counted the number of rtQTLs where the allele effect was consistent with the direction of replication timing difference between species for at least one of the evaluated SNPs. For example, if chimpanzees were earlier replicating we asked if the chimpanzee allele matched the early replicating human allele, while if humans were earlier replicating we asked if the

chimpanzee allele matched the late replicating human allele. To assess significance, we permuted the number of rtQTLs that we tested for allele direction. For each rtQTL tested, we compared the direction of replication timing difference between species to permuted rtQTL haplotypes (top associated SNP and SNPs in LD>0.8), then asked if the permuted SNP allele direction was consistent with the change in replication timing of the tested rtQTL. We counted the number of rtQTLs with at least one tested SNP with a consistent allele direction. We repeated these steps for 1000 permutations and used a z-test to calculate p-value.

The allele directionality analysis was also repeated on chimpanzee rtQTLs that were called in human-chimpanzee replication timing variant regions. Allele effect (i.e. whether the chimpanzee derived or human allele was earlier replicating) was only evaluated at the top chimpanzee rtQTL SNP.

Chimpanzee rtQTLs were classified as shared with human rtQTLs if the top associated human rtQTL location fell within the chimpanzee rtQTL associated region. Significance was evaluated with 100 randomizations and a z-test.

Identifying shared eQTLs

To identify shared genetic causes of replication timing and expression variation, we located GTEx LCL eQTLs that shared a top associated rtQTL SNP (and/or one in LD>0.8 to the top associated SNP). Significant GTEx eQTLs had a q-value less than 0.05. For the shared rtQTL-eQTL example given, we used HaploReg (Ward and Kellis, 2011) to identify if rs7806550 interrupted any regulatory motifs. Sequence logo and binding affinities for HDAC2 were obtained from HaploReg (Ward and Kellis, 2011) and ChIP-seq data for HDAC2 was obtained from ENCODE (2012).

Acknowledgements

We thank Ana Rita Rebelo, Bronte Zhang, Lauren Mei, Sean Kim and Tiffany Ge for technical assistance, and members of our lab for critical reading of the manuscript. We thank Yoav Gilad for sharing iPS and LCL samples and for fruitful discussions. Chimpanzee LCL samples were sourced from the Yerkes Center (Grant No. P51OD011132); DONSON human LCL cell lines were a gift from Grant Stewart. This work was supported by the National Science Foundation (MCB-1921341 to A.K.). A.N.B. was partially supported by the Cornell Center for Vertebrate Genomics. A.D. is a Hunter R. Rawlings III Cornell Presidential Research Scholar.

Data availability

Raw chimpanzee and rhesus macaque whole-genome sequencing data is available under SRA accession PRJNA856315, and human data is available under dbGaP accession phs002597. Three human LCL and six human iPSC samples were not consented for release of raw genomic sequence data. Processed replication timing data is available for all samples (Supplemental files 1 and 2).

Supplemental Tables and Files

Available with the published manuscript

Table S1. Human-chimpanzee replication timing variant regions

Table S2. Gene ontology enrichment analysis

Table S3. Sample information

File S1. LCL replication timing data

File S2. iPSC replication timing data

CHAPTER 3

CONCLUSIONS AND FUTURE DIRECTIONS

Evolution of replication timing

Evolutionary studies are crucial for understanding the functional importance of biological features. Replication timing, the spatiotemporal order that genomic segments are replicated in during S phase, has been the focus of few evolutionary studies. In these studies, replication timing has generally been found to be highly conserved across species, and thus inferred to have a conserved, essential function (Müller and Nieduszynski, 2012; Ryba et al., 2010; Yaffe et al., 2010). But the exact function of replication timing is generally unknown. Replication timing is positively correlated with many genomic and epigenomic features, such as gene density, transcription and chromatin accessibility, while negatively correlated with sequence divergence and mutation density. Previous studies have been underpowered to assess potential variation in replication timing within and between species, what drives/regulates replication timing variation, and potential impacts of replication timing evolution.

In this dissertation, I was able to extend findings from previous evolutionary studies, by analyzing the evolutionary dynamics and impacts of replication timing in a large number of humans, chimpanzees, and rhesus macaques. This was accomplished using a population based method of generating replication timing profiles from unsorted, whole-genome sequencing data (Koren et al., 2021). With this large number of samples, I evaluated variation both within and between humans and chimpanzees, linked replication timing variation to sequence evolution using replication timing quantitative trait loci (rtQTLs), and correlated replication timing changes with regulatory evolution.

Consistent with previous studies, replication timing was generally conserved in structure across species pointing to a level of evolutionary constraint on this process. However, I also observed fine-scale variation in replication timing at hundreds of locations between humans and chimpanzees, which primarily occurred at origins of replication and were cell-type specific. I

was also able to resolve over one hundred human-chimpanzee variants as changes in the human lineage using rhesus macaque as an outgroup. Importantly, all identified new and lost origins in human were polymorphic, pointing to ongoing evolution of replication timing at these loci.

Replication timing also varied within humans and chimpanzees at more than one hundred regions, which significantly overlapped between species and with human-chimpanzee variants. Sequence determinants of replication timing evolution were identified by mapping rtQTLs in humans and chimpanzees and linking derived alleles to derived replication timing states. A few rtQTLs were shared across humans and chimpanzees, but the underlying rtQTL SNPs were often different in these cases, which could either suggest a lack of strong evolutionary constraint or alternatively balancing selection on these loci.

Additionally, replication timing co-varied with gene expression and chromatin accessibility across humans and chimpanzees. Evaluating this on the level of individual human-chimpanzee variants indicated that the direction of effect is complex and likely locus specific. Replication timing and transcription may also co-vary though connection with a chromatin intermediate. For example, I highlighted a shared human rtQTL-eQTL region where the human derived allele impacted a HDAC2 binding site. The expression of two genes in the immediate area had higher expression in humans than chimpanzees, and were also earlier replicating in humans.

Sequence divergence between human and chimpanzee was elevated in iPSC replication timing variant regions, suggesting that a change in replication timing could potentially alter local mutation rate. Additionally, I found that protein-coding genes were generally depleted in the same variant regions, which may suggest that replication timing changes in regions with highly conserved genomic function are unfavorable. This could be due to the coordinated changes in mutation rate or gene expression that can occur with a change in replication timing.

Overall, this dissertation reveals that replication timing varies substantially more than previously thought between closely related species, variation can be controlled by sequence evolution, and is highly correlated to regulatory evolution.

Future research and preliminary results

Application to additional cell types and species

Since replication timing variation between humans and chimpanzees was primarily cell type specific, one avenue of future research should focus on profiling replication timing from additional cell types. This can inform species-specific cell type regulation of replication timing and impacts of replication timing changes specific to these cell types. With limitations on primary samples from chimpanzees and other great apes, the best way to accomplish this is by differentiating induced pluripotent stem cells (iPSCs) into a variety of cell types (Romero et al., 2015). Previous studies have utilized iPSC differentiation to understand variation in gene expression between humans and chimpanzees in different cell types (Blake et al., 2018; Eres et al., 2019; Pavlovic et al., 2018; Prescott et al., 2015). As such, human and chimpanzee iPSC differentiation protocols have been generated for many cell types, including neuronal and skeletal cell types (Agoglia et al., 2021; Gokhman et al., 2021; Housman et al., 2022), which may be the most interesting and important to address first due to their known phenotypic variation between humans and chimpanzees.

This research also resolved a number of replication timing variants as changes in the human or chimpanzee lineages; however, this number may have been limited by using only one outgroup species (rhesus macaque). Future studies should focus on expanding the number of species evaluated for replication timing. In addition to identifying replication timing evolution specific to other lineages, this would also allow us to better understand our unresolved regions. For example, are these regions variable along the whole lineage or variable only in human and chimpanzee but conserved further in the primate lineage? In a preliminary study, I profiled replication timing from available whole genome sequencing of gorilla and orangutan LCLs (García-Pérez et al., 2021). Clustering replication timing of these samples, in addition to a subset of our human, chimpanzee, and rhesus macaque data, across the genome replicated the species tree (Figure 3.1), further supporting that replication timing evolved continuously across the primate lineage. Future work should evaluate pairwise differences between species to further

pinpoint locations of replication timing evolution across the lineage. Ideally, there should be multiple samples from each species for replication timing comparisons, as we showed that replication timing varies within species and could potentially impact the inference of differences between species.

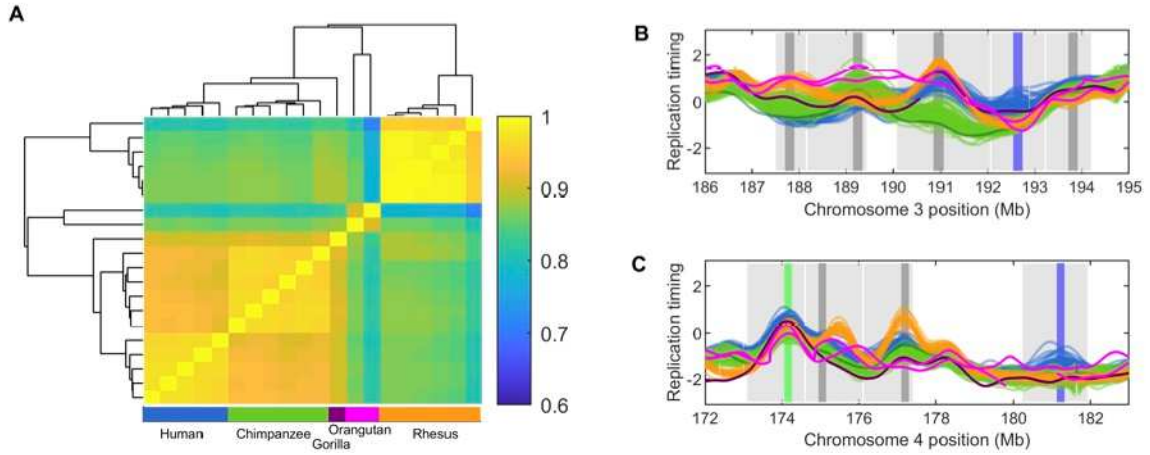


Figure 3.1. Replication timing of orangutan and gorilla.
(A) Subset of five human, five chimpanzee, and five rhesus macaque samples from Chapter 2. Whole genome sequencing of one chimpanzee, one gorilla, one orangutan, and one rhesus macaque were obtained from (García-Pérez et al., 2021) and replication timing profiles were generated. Repli-seq of one orangutan sample obtained from (Yang et al., 2018). Hierarchical clustering of replication timing values from chromosomes 1-16, excluding 4 and 10. Samples cluster by known phylogenetic tree. (B) Replication timing profiles from Chapter 2 with additional chimpanzee, gorilla, orangutan, and rhesus macaque samples described in (A).

Impact on local mutation rate

Another open question is whether replication timing differences could impact local mutation rate differences between species. I found that human-chimpanzee sequence divergence was elevated in iPSC replication timing variant regions, but was not able to associate differences in replication timing between human and chimpanzees with differences in mutation rate on the level of individual variants (Figure 3.2). This could be because humans and chimpanzees don't vary at enough locations across the genome to give us power for the association or that the change in replication timing happened too recently to have a substantial impact on mutation rate.

When I sub-sampled the genome to equal length as the replication timing variants, I found that the association between mutation density and replication timing disappears (Figure 3.2 E), suggesting this is at least partly a power issue. If there was power for this association, it would be important to use cell type matched mutation and replication timing data, since replication timing variation is cell type specific, or use an undifferentiated cell type (e.g. iPSC). Another option to increase power is to compare more divergent species (e.g. human and mouse). Previous studies of mutational signatures across great apes identified replication timing as a consistent contributor across the great ape lineage (Goldberg and Harris, 2022); incorporation of chimpanzee and other great ape replication timing data into these types of studies may be able to help resolve if replication timing contributes to local mutation rate differences across species.

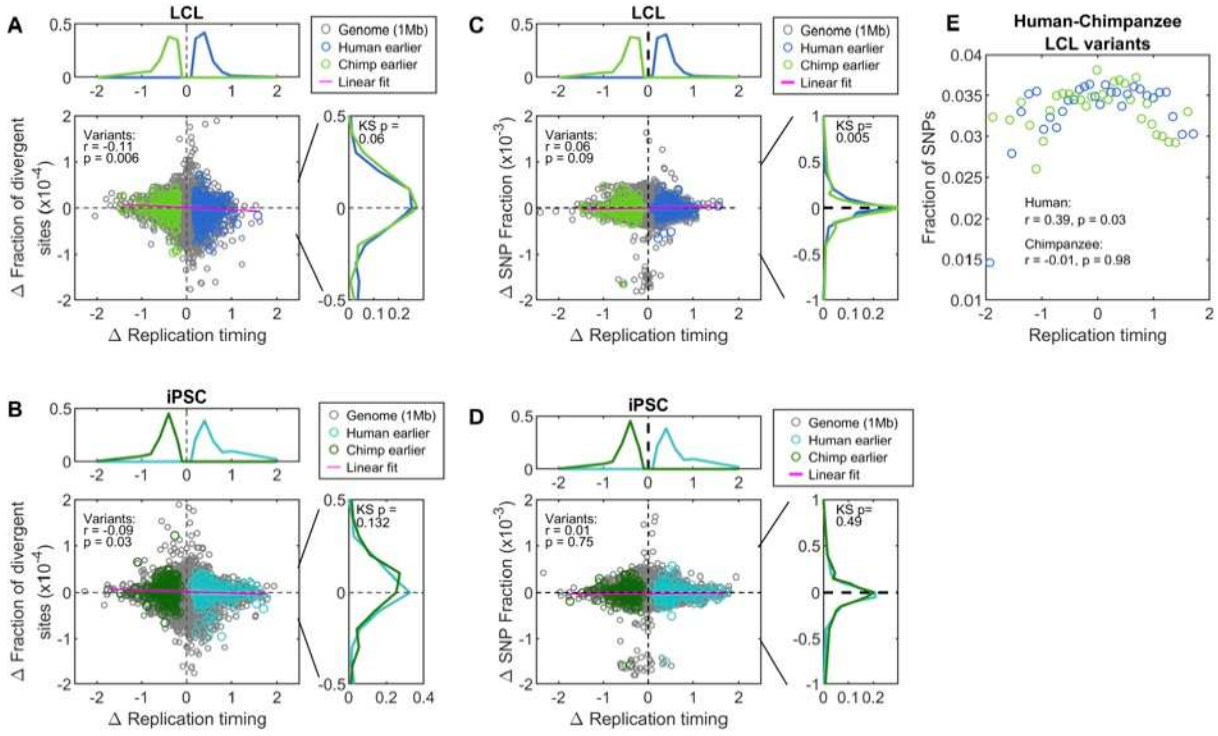


Figure 3.2. Differences in mutation density do not correlate with replication timing variation.
(A-B) The difference in fraction of lineage specific divergent sites (# human sites / Mb / total number of human sites in the genome - # chimpanzee sites / Mb / total number chimpanzee sites in the genome) at replication timing variant regions (blue and green) and across the genome (gray). Histograms on top show the distribution of replication timing differences between the species. Histograms on the right show the distribution of differences in divergent site densities. Shown for LCLs (A) and iPSCs (B). (C-D) Differences in replication timing between humans and chimpanzees compared to differences in SNP

densities (# SNPs / Mb / total # SNPs in the genome), for replication timing variants and across the genome. As in A-B, shown for LCLs (C) and iPSCs (D). (E) Relationship between replication timing and SNP density disappears when only look at variant regions. Fraction of SNPs (from human-chimpanzee replication timing variant regions only) in 30 replication timing bins per species.

Sequence determinants of replication timing evolution

Replication timing QTLs are a powerful method to link replication timing evolution to sequence determinants and co-variation with gene expression. Ideally, more unrelated chimpanzee samples would be needed to have the power to further map rtQTLs in this species. This would allow further comparison with human rtQTLs to identify shared and species specific genetic determinants of replication timing evolution. As mentioned above, with restrictions on chimpanzee research, iPSCs may be the most realistic way of obtaining samples. Otherwise, using a species that is more thoroughly studied in higher quantities (e.g. rhesus macaque, mouse) may be preferable. Another future direction with rtQTLs would be to narrow down causal SNPs with fine mapping tools and experimentally validate the sequence changes with functional assays. One potential option is to use CRISPR-Cas9 to alter the sequence at rtQTLs in human or chimpanzee cell lines. For example, if one was to alter the chimpanzee sequence to match the derived allele of a human rtQTL, does that create a corresponding change in replication timing, gene expression and/or chromatin accessibility? This could help further disentangle the regulatory mechanism linking the evolution of replication timing to sequence, chromatin, and gene expression evolution.

Conclusions

Overall, studying replication timing in the context of evolution is a powerful tool to understand the regulation and function of replication timing. In this dissertation, I described that replication timing varies substantially within and between humans and chimpanzees, and that this variation is overwhelmingly coordinated with changes in gene expression and chromatin accessibility between species. I further took advantage of the large number of samples to analyze rtQTLs in humans and chimpanzees, and linked replication timing evolution to sequence

changes. In some cases, both replication timing and gene expression changes could be explained by identical sequence changes, suggesting a mechanism by which evolution of replication timing and gene regulation are coordinated. Future research will focus on expanding the number of cell types and species studied as well as functional validation to further resolve replication timing changes between species, and understand their sequence determinants and evolutionary impacts.

REFERENCES

(2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

(2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res *49*, D325-d334.

Agier, N., Delmas, S., Zhang, Q., Fleiss, A., Jaszczyszyn, Y., van Dijk, E., Thermes, C., Weigt, M., Cosentino-Lagomarsino, M., and Fischer, G. (2018). The evolution of the temporal program of genome replication. Nature communications *9*, 1-12.

Agoglia, R.M., Sun, D., Birey, F., Yoon, S.J., Miura, Y., Sabatini, K., Pașca, S.P., and Fraser, H.B. (2021). Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. Nature *592*, 421-427.

Altman, A.L., and Fanning, E. (2001). The Chinese hamster dihydrofolate reductase replication origin beta is active at multiple ectopic chromosomal locations and requires specific DNA sequence elements for activity. Molecular and Cellular Biology *21*, 1098-1110.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene Ontology: tool for the unification of biology. Nature Genetics *25*, 25-29.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

Bell, S.P., and Stillman, B. (1992). ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. Nature *357*, 128-134.

Blake, L.E., Thomas, S.M., Blischak, J.D., Hsiao, C.J., Chavarria, C., Myrthil, M., Gilad, Y., and Pavlovic, B.J. (2018). A comparative study of endoderm differentiation in humans and chimpanzees. Genome biology *19*, 1-18.

Blin, M., Le Tallec, B., Nähse, V., Schmidt, M., Brossas, C., Millot, G.A., Prioleau, M.-N., and Debatisse, M. (2019). Transcription-dependent regulation of replication dynamics modulates genome stability. Nature structural & molecular biology *26*, 58-66.

Bracci, A.N., Dallmann, A., Ding, Q., Hubisz, M.J., Caballero, M., and Koren, A. (2022). The evolution of the human DNA replication timing program. bioRxiv, 2022.2008.2009.503365.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., and Sollis, E. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research *47*, D1005-D1012.

Caballero, M., Ge, T., Rebelo, A.R., Seo, S., Kim, S., Brooks, K., Zuccaro, M., Kanagaraj, R., Vershkov, D., Kim, D.*, et al.* (2022). Comprehensive analysis of DNA replication timing across 184 cell lines suggests a role for MCM10 in replication timing regulation. Human Molecular Genetics.

Cayrou, C., Coulombe, P., Vigneron, A., Stanojcic, S., Ganier, O., Peiffer, I., Rivals, E., Puy, A., Laurent-Chabalier, S., and Desprat, R. (2011). Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome research *21*, 1438-1449.

Consortium, C.S.a.A. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69-87.

Cornacchia, D., Dileep, V., Quivy, J.P., Foti, R., Tili, F., Santarella□Mellwig, R., Antony, C., Almouzni, G., Gilbert, D.M., and Buonomo, S.B. (2012). Mouse Rif1 is a key regulator of the replication□timing programme in mammalian cells. The EMBO journal *31*, 3678-3690.

Cotney, J., Leng, J., Yin, J., Reilly, Steven K., DeMare, Laura E., Emera, D., Ayoub, Albert E., Rakic, P., and Noonan, James P. (2013). The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb. Cell *154*, 185-196.

De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V.C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., and Hallast, P. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science *354*, 477-481.

Dellino, G.I., Cittaro, D., Piccioni, R., Luzi, L., Banfi, S., Segalla, S., Cesaroni, M., Mendoza-Maldonado, R., Giacca, M., and Pelicci, P.G. (2013). Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. Genome research *23*, 1-11.

Ding, Q., Edwards, M.M., Wang, N., Zhu, X., Bracci, A.N., Hulke, M.L., Hu, Y., Tong, Y., Hsiao, J., and Charvet, C.J. (2021). The genetic architecture of DNA replication timing in human pluripotent stem cells. Nature communications *12*, 1-18.

Ding, Y., Wang, L., Su, L.K., Frey, J.A., Shao, R., Hunt, K.K., and Yan, D.H. (2004). Antitumor

activity of IFIX, a novel interferon-inducible HIN-200 gene, in breast cancer. Oncogene *23*, 4556-4566.

Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P., and Pääbo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. Nature *418*, 869-872.

Eres, I.E., Luo, K., Hsiao, C.J., Blake, L.E., and Gilad, Y. (2019). Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. PLoS genetics *15*, e1008278.

Evans, P.D., Gilbert, S.L., Mekel-Bobrov, N., Vallender, E.J., Anderson, J.R., Vaez-Azizi, L.M., Tishkoff, S.A., Hudson, R.R., and Lahn, B.T. (2005). Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. science *309*, 1717-1720.

Evrin, C., Clarke, P., Zech, J., Lurz, R., Sun, J., Uhle, S., Li, H., Stillman, B., and Speck, C. (2009). A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. Proceedings of the National Academy of Sciences *106*, 20240-20245.

Fair, B.J., Blake, L.E., Sarkar, A., Pavlovic, B.J., Cuevas, C., and Gilad, Y. (2020). Gene expression variability in human and chimpanzee populations share common determinants. Elife *9*, e59929.

Fragkos, M., Ganier, O., Coulombe, P., and Méchali, M. (2015). DNA replication origin activation in space and time. Nature Reviews Molecular Cell Biology *16*, 360-374.

Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., Van Duijn, C.M., Swertz, M., Wijmenga, C., and Van Ommen, G. (2015). Genome-wide patterns and properties of de novo mutations in humans. Nature genetics *47*, 822-826.

Fraser, H.B. (2013). Gene expression drives local adaptation in humans. Genome research *23*, 1089-1096.

Fu, W., and Akey, J.M. (2013). Selection and adaptation in the human genome. Annual review of genomics and human genetics *14*, 467-489.

Ganier, O., Prorok, P., Akerman, I., and Méchali, M. (2019). Metazoan DNA replication origins. Current opinion in cell biology *58*, 134-141.

García-Pérez, R., Esteller-Cucala, P., Mas, G., Lobón, I., Di Carlo, V., Riera, M., Kuhlwilm, M., Navarro, A., Blancher, A., and Di Croce, L. (2021). Epigenomic profiling of primate lymphoblastoid cell lines reveals the evolutionary patterns of epigenetic activities in gene regulatory architectures. Nature communications *12*, 1-17.

Gavin, K.A., Hidaka, M., and Stillman, B. (1995). Conserved Initiator Proteins in Eukaryotes. Science *270*, 1667-1671.

Gayà-Vidal, M., and Albà, M.M. (2014). Uncovering adaptive evolution in the human lineage. BMC Genomics *15*, 599.

Ge, X.Q., Jackson, D.A., and Blow, J.J. (2007). Dormant origins licensed by excess Mcm2–7 are required for human cells to survive replicative stress. Genes & development *21*, 3331-3341.

Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., Venter, J.C., and Wilson, R.K. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. science *316*, 222-234.

Gokhman, D., Agoglia, R.M., Kinnebrew, M., Gordon, W., Sun, D., Bajpai, V.K., Naqvi, S., Chen, C., Chan, A., and Chen, C. (2021). Human–chimpanzee fused cells reveal cis-regulatory divergence underlying skeletal evolution. Nature genetics *53*, 467-476.

Goldberg, M.E., and Harris, K. (2022). Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny. Genome Biol Evol *14*, evab104.

Goren, A., Tabib, A., Hecht, M., and Cedar, H. (2008). DNA replication timing of the human β-globin domain is controlled by histone modification at the origin. Genes & development *22*, 1319-1324.

Groth, A., Rocha, W., Verreault, A., and Almouzni, G. (2007). Chromatin challenges during DNA replication and repair. Cell *128*, 721-733.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proceedings of the National Academy of Sciences of the United States of America *107*, 139-144.

Housman, G., Briscoe, E., and Gilad, Y. (2022). Evolutionary insights into primate skeletal gene

regulation using a comparative cell culture model. PLoS genetics *18*, e1010073.

Hubisz, M.J., and Pollard, K.S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. Current opinion in genetics & development *29*, 15-21.

Hulke, M.L., Massey, D.J., and Koren, A. (2020). Genomic methods for measuring DNA replication dynamics. Chromosome Res *28*, 49-67.

Ijdo, J., Baldini, A., Ward, D., Reeders, S., and Wells, R. (1991). Origin of human chromosome 2: an ancestral telomere-telomere fusion. Proceedings of the National Academy of Sciences *88*, 9051-9055.

Ilves, I., Petojevic, T., Pesavento, J.J., and Botchan, M.R. (2010). Activation of the MCM2-7 helicase by association with Cdc45 and GINS proteins. Molecular cell *37*, 247-258.

Jasinska, A.J., Zelaya, I., Service, S.K., Peterson, C.B., Cantor, R.M., Choi, O.-W., DeYoung, J., Eskin, E., Fairbanks, L.A., Fears, S.*, et al.* (2017). Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate. Nature Genetics *49*, 1714-1721.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P.*, et al.* (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434-443.

Kenigsberg, E., Yehuda, Y., Marjavaara, L., Keszthelyi, A., Chabes, A., Tanay, A., and Simon, I. (2016). The mutation spectrum in genomic late replication domains shapes mammalian GC content. Nucleic acids research *44*, 4222-4232.

Khan, Z., Ford, M.J., Cusanovich, D.A., Mitrano, A., Pritchard, J.K., and Gilad, Y. (2013). Primate transcript and protein expression levels evolve under compensatory selection pressures. Science *342*, 1100-1104.

King, M.-C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences. science *188*, 107-116.

Klein, K.N., Zhao, P.A., Lyu, X., Sasaki, T., Bartlett, D.A., Singh, A.M., Tasan, I., Zhang, M., Watts, L.P., Hiraga, S.-i.*, et al.* (2021). Replication timing maintains the global epigenetic state

in human cells. Science *372*, 371-378.

Koren, A., Handsaker, R.E., Kamitaki, N., Karlić, R., Ghosh, S., Polak, P., Eggan, K., and McCarroll, S.A. (2014). Genetic variation in human DNA replication timing. Cell *159*, 1015-1026.

Koren, A., Massey, D.J., and Bracci, A.N. (2021). TIGER: inferring DNA replication timing from whole-genome sequence data. Bioinformatics.

Koren, A., and McCarroll, S.A. (2014). Random replication of the inactive X chromosome. Genome research *24*, 64-69.

Koren, A., Polak, P., Nemesh, J., Michaelson, J.J., Sebat, J., Sunyaev, S.R., and McCarroll, S.A. (2012). Differential relationship of DNA replication timing to different forms of human mutation and variation. The American Journal of Human Genetics *91*, 1033-1040.

Koryakov, D.E., Pokholkova, G.V., Maksimov, D.A., Belyakin, S.N., Belyaeva, E.S., and Zhimulev, I.F. (2012). Induced transcription results in local changes in chromatin structure, replication timing, and DNA polytenization in a site of intercalary heterochromatin. Chromosoma *121*, 573-583.

Kronenberg, Z.N., Fiddes, I.T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O.S., Underwood, J.G., Nelson, B.J., Chaisson, M.J.P., Dougherty, M.L.*, et al.* (2018). High-resolution comparative analysis of great ape genomes. Science *360*.

Liu, G., Malott, M., and Leffak, M. (2003). Multiple functional elements comprise a mammalian chromosomal replicator. Molecular and cellular biology *23*, 1832-1842.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N.*, et al.* (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics *45*, 580-585.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867-2873.

Marahrens, Y., and Stillman, B. (1992). A yeast chromosomal origin of DNA replication defined by multiple functional elements. Science *255*, 817-823.

Marchal, C., Sima, J., and Gilbert, D.M. (2019). Control of DNA replication timing in the 3D genome. Nature Reviews Molecular Cell Biology *20*, 721-737.

McLean, C.Y., Reno, P.L., Pollen, A.A., Bassan, A.I., Capellini, T.D., Guenther, C., Indjeian, V.B., Lim, X., Menke, D.B., and Schaar, B.T. (2011). Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature *471*, 216-219.

Mekel-Bobrov, N., Gilbert, S.L., Evans, P.D., Vallender, E.J., Anderson, J.R., Hudson, R.R., Tishkoff, S.A., and Lahn, B.T. (2005). Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. Science *309*, 1720-1722.

Mesner, L.D., Valsakumar, V., Cieślik, M., Pickin, R., Hamlin, J.L., and Bekiranov, S. (2013). Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early-and late-firing origins. Genome research *23*, 1774-1788.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res *47*, D419-d426.

Moyer, S.E., Lewis, P.W., and Botchan, M.R. (2006). Isolation of the Cdc45/Mcm2–7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. Proceedings of the National Academy of Sciences *103*, 10236-10241.

Müller, C.A., and Nieduszynski, C.A. (2012). Conservation of replication timing reveals global and local regulation of replication origin activity. Genome research *22*, 1953-1962.

Müller, C.A., and Nieduszynski, C.A. (2017). DNA replication timing influences gene expression level. Journal of Cell Biology *216*, 1907-1914.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J.*, et al.* (2005). A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLOS Biology *3*, e170.

Olson, M.V., and Varki, A. (2003). Sequencing the chimpanzee genome: insights into human evolution and disease. Nature Reviews Genetics *4*, 20-28.

Parker, M.W., Botchan, M.R., and Berger, J.M. (2017). Mechanisms and regulation of DNA replication initiation in eukaryotes. Critical Reviews in Biochemistry and Molecular Biology *52*, 107-144.

Pavlovic, B.J., Blake, L.E., Roux, J., Chavarria, C., and Gilad, Y. (2018). A comparative assessment of human and chimpanzee iPSC-derived cardiomyocytes with primary heart tissues. Scientific reports *8*, 1-14.

Peyrégne, S., Boyle, M.J., Dannemann, M., and Prüfer, K. (2017). Detecting ancient positive selection in humans using extended lineage sorting. Genome research *27*, 1563-1572.

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G.*, et al.* (2013). Great ape genetic diversity and population history. Nature *499*, 471.

Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T., and Wysocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. Cell *163*, 68-83.

Remus, D., Beuron, F., Tolun, G., Griffith, J.D., Morris, E.P., and Diffley, J.F. (2009). Concerted loading of Mcm2–7 double hexamers around DNA during DNA replication origin licensing. Cell *139*, 719-730.

Rhind, N., and Gilbert, D.M. (2013). DNA replication timing. Cold Spring Harbor perspectives in biology *5*, a010132.

Rivera-Mulia, J.C., Buckley, Q., Sasaki, T., Zimmerman, J., Didier, R.A., Nazor, K., Loring, J.F., Lian, Z., Weissman, S., and Robins, A.J. (2015). Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. Genome research *25*, 1091-1103.

Romero, I.G., Pavlovic, B.J., Hernando-Herraez, I., Zhou, X., Ward, M.C., Banovich, N.E., Kagan, C.L., Burnett, J.E., Huang, C.H., and Mitrano, A. (2015). A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. Elife *4*, e07103.

Romero, I.G., Ruvinsky, I., and Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics *13*, 505-516.

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

Sawyer, S.L., Emerman, M., Malik, H.S., and Harvey, P. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS biology *2*, e275.

Siefert, J.C., Georgescu, C., Wren, J.D., Koren, A., and Sansam, C.L. (2017). DNA replication timing during development anticipates transcriptional programs and parallels enhancer activation. Genome research *27*, 1406-1416.

Sima, J., Chakraborty, A., Dileep, V., Michalski, M., Klein, K.N., Holcomb, N.P., Turner, J.L., Paulsen, M.T., Rivera-Mulia, J.C., and Trevilla-Garcia, C. (2019). Identifying cis elements for spatiotemporal control of mammalian DNA replication. Cell *176*, 816-830. e818.

Smith, D.J., and Whitehouse, I. (2012). Intrinsic coupling of lagging-strand synthesis to chromatin assembly. Nature *483*, 434-438.

Soto, D.C., Shew, C., Mastoras, M., Schmidt, J.M., Sahasrabudhe, R., Kaya, G., Andrés, A.M., and Dennis, M.Y. (2020). Identification of Structural Variation in Chimpanzees Using Optical Mapping and Nanopore Sequencing. Genes *11*, 276.

Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M., and Sunyaev, S.R. (2009). Human mutation rate associated with DNA replication timing. Nature genetics *41*, 393-395.

Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature *521*, 81-84.

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. Genetics *135*, 599-607.

Therizols, P., Illingworth, R.S., Courilleau, C., Boyle, S., Wood, A.J., and Bickmore, W.A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science *346*, 1238-1242.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M.*, et al.* (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics *39*, 31-40.

Tung, J., Zhou, X., Alberts, S.C., Stephens, M., and Gilad, Y. (2015). The genetic architecture of gene expression levels in wild baboons. eLife *4*, e04729.

Vallender, E.J., and Lahn, B.T. (2004). Positive selection on the human genome. Human Molecular Genetics *13*, R245-R254.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy□Moonshine, A., Jordan, T., Shakir, K., Roazen, D., and Thibault, J. (2013). From FastQ data to high□confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics *43*, 11.10. 11-11.10. 33.

Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A Map of Recent Positive Selection in the Human Genome. PLOS Biology *4*, e72.

Wang, L., Lin, C.-M., Brooks, S., Cimbora, D., Groudine, M., and Aladjem, M.I. (2004). The human β-globin replication initiation region consists of two modular independent replicators. Molecular and cellular biology *24*, 3373-3386.

Ward, L.D., and Kellis, M. (2011). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Research *40*, D930-D934.

Warren, W.C., Harris, R.A., Haukness, M., Fiddes, I.T., Murali, S.C., Fernandes, J., Dishuck, P.C., Storer, J.M., Raveendran, M., and Hillier, L.W. (2020). Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. Science *370*, eabc6617.

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

Yamazaki, S., Ishii, A., Kanoh, Y., Oda, M., Nishito, Y., and Masai, H. (2012). Rif1 regulates the replication timing domains on the human genome. The EMBO journal *31*, 3667-3677.

Yang, Y., Gu, Q., Zhang, Y., Sasaki, T., Crivello, J., O'Neill, R.J., Gilbert, D.M., and Ma, J. (2018). Continuous-trait probabilistic model for comparing multi-species functional genomic data. Cell systems *7*, 208-218. e211.

Yehuda, Y., Blumenfeld, B., Mayorek, N., Makedonski, K., Vardi, O., Cohen-Daniel, L., Mansour, Y., Baror-Sebban, S., Masika, H., and Farago, M. (2018). Germline DNA replication timing shapes mammalian genome composition. Nucleic acids research *46*, 8299-8310.

Zhang, J., Xu, F., Hashimshony, T., Keshet, I., and Cedar, H. (2002). Establishment of

transcriptional competence in early and late S phase. Nature *420*, 198-202.

Zhao, P.A., Sasaki, T., and Gilbert, D.M. (2020). High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. Genome biology *21*, 1-20.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326-3328.

Zhou, X., Cain, C.E., Myrthil, M., Lewellen, N., Michelini, K., Davenport, E.R., Stephens, M., Pritchard, J.K., and Gilad, Y. (2014). Epigenetic modifications are associated with inter-species gene expression variation in primates. Genome biology *15*, 547.