

1 **The Genetic Architecture of DNA Replication Timing in Human Pluripotent  
2 Stem Cells**

3  
4 Qiliang Ding<sup>1</sup>, Matthew M. Edwards<sup>1</sup>, Michelle L. Hulke<sup>1</sup>, Alexa N. Bracci<sup>1</sup>, Ya Hu<sup>1,2</sup>, Yao  
5 Tong<sup>1</sup>, Xiang Zhu<sup>3</sup>, Joyce Hsiao<sup>4</sup>, Christine J. Charvet<sup>1</sup>, Sulagna Ghosh<sup>5,6,7</sup>, Robert E.  
6 Handsaker<sup>5,6</sup>, Kevin Eggan<sup>5,7,8</sup>, Florian T. Merkle<sup>9</sup>, Jeannine Gerhardt<sup>10,11</sup>, Dieter Egli<sup>12</sup>,  
7 Andrew G. Clark<sup>1</sup>, Amnon Koren<sup>1\*</sup>.

8  
9 <sup>1</sup> Department of Molecular Biology and Genetics, Cornell University, Ithaca NY 14853, USA.

10 <sup>2</sup> New York Genome Center, New York NY 10013, USA.

11 <sup>3</sup> Department of Statistics, Stanford University, Stanford CA 94305, USA.

12 <sup>4</sup> Department of Human Genetics, University of Chicago, Chicago IL 60637, USA.

13 <sup>5</sup> Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge MA  
14 02142, USA.

15 <sup>6</sup> Department of Genetics, Harvard Medical School, Boston MA 02115, USA.

16 <sup>7</sup> Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge MA  
17 02138, USA.

18 <sup>8</sup> Howard Hughes Medical Institute, Harvard University, Cambridge MA 02138, USA.

19 <sup>9</sup> Wellcome Trust - Medical Research Council Institute of Metabolic Science, University of  
20 Cambridge, Cambridge, United Kingdom.

21 <sup>10</sup> Ronald O. Perelman and Claudia Cohen Center for Reproductive Medicine, Weill Cornell  
22 Medicine, New York NY 10065, USA.

23 <sup>11</sup> Department of Obstetrics and Gynecology, Weill Cornell Medicine, New York NY 10065,  
24 USA.

25 <sup>12</sup> Department of Pediatrics, Columbia University, New York NY 10032, USA.

26  
27 \* Correspondence to Amnon Koren, [koren@cornell.edu](mailto:koren@cornell.edu).

28

29

## 1 Abstract

2 DNA replication follows a strict spatiotemporal program that intersects with chromatin structure  
3 and gene regulation. However, the genetic basis of the mammalian DNA replication timing  
4 program is poorly understood<sup>1-3</sup>. To systematically identify genetic regulators of DNA  
5 replication timing, we exploited inter-individual variation in 457 human pluripotent stem cell  
6 lines from 349 individuals. We show that the human genome's replication program is broadly  
7 encoded in DNA and identify 1,617 *cis*-acting replication timing quantitative trait loci (rtQTLs<sup>4</sup>)  
8 – base-pair-resolution sequence determinants of replication initiation. rtQTLs function  
9 individually, or in combinations of proximal and distal regulators, to affect replication timing.  
10 Analysis of rtQTL locations reveals a histone code for replication initiation, composed of  
11 bivalent histone H3 trimethylation marks on a background of histone hyperacetylation. The H3  
12 trimethylation marks are individually repressive yet synergize to promote early replication. We  
13 further identify novel positive and negative regulators of DNA replication timing, the former  
14 comprised of pluripotency-related transcription factors while the latter involve boundary  
15 elements. Human replication timing is controlled by a multi-layered mechanism that operates  
16 on target DNA sequences, is composed of dozens of effectors working combinatorially, and  
17 follows principles analogous to transcription regulation: a histone code, activators and  
18 repressors, and a promoter-enhancer logic.

19

## 20 Main

21 Eukaryotic genomes are replicated according to a strict spatiotemporal program, in which  
22 replication initiates from specific locations along chromosomes and at reproducible times. The  
23 replication timing program is a fundamental property of chromosome organization, interfaces  
24 with gene regulation and shapes the mutational landscape of the genome. Efforts to  
25 understand the locations and nature of initiation sites and the factors that regulate DNA  
26 replication timing in mammalian cells have been ongoing for decades, with limited success<sup>1-3</sup>.  
27 Specifically, it is still unclear to what extent the DNA replication timing program is determined  
28 by local DNA sequences, by epigenetic factors, or by a combination thereof. Earlier studies  
29 suggested that specific sequence elements control replication initiation in human cells, with  
30 several distal and proximal elements often acting in concert<sup>5-11</sup>. More recently, CRISPR/Cas9-  
31 mediated deletions have suggested that several DNA sequences locally interact to control  
32 early replication in mice<sup>12</sup>.

33 Numerous lines of evidence link replication regulation to epigenetic states, in particular histone  
34 acetylations and methylations marking open chromatin<sup>3,13-17</sup>. However, no single epigenetic  
35 mark appears to be absolutely required nor sufficient for replication origin function. This has  
36 led to suggestions that a combination of histone marks may be required for specifying patterns  
37 of DNA replication<sup>18</sup>. Similarly, it has been proposed that indiscriminate DNA-binding patterns  
38 of the replication machinery may translate into a consistent, organized replication program by

1 means of combinatorial chromatin modifications influencing subsets of replication initiation  
2 sites<sup>3</sup>.  
3 The nature of such modular, combinatorial regulation of DNA replication at the genetic and  
4 epigenetic levels remains to be revealed. Previous studies applied stepwise reverse  
5 engineering approaches to probe for mechanisms controlling replication timing. However, such  
6 a complex system may be best studied with an unbiased and comprehensive interrogation of  
7 genetic and epigenetic factors and their interactions. While such an approach is currently  
8 challenging experimentally, an alternative is to take advantage of natural genetic variation. We  
9 previously showed that replication timing is variable among individuals, that it can be studied at  
10 fine-scale on a population level by sequencing the genomes of proliferating cells, and that  
11 genotype information from the same genome sequences can be used to associate replication  
12 timing variation with specific genetic polymorphisms. This results in the identification of  
13 replication timing quantitative trait loci (rtQTLs), DNA sequences that act in *cis* to affect  
14 replication initiation<sup>4</sup>. Leveraging human genetic variation enables the equivalent of numerous  
15 surgical genetic manipulations and their association with DNA replication timing alterations.  
16 Here, we apply this approach to hundreds of human embryonic stem cell (hESC) and induced  
17 pluripotent stem cell (iPSC) lines. Pluripotent stem cells are particularly useful for this analysis,  
18 since they are non-transformed, karyotypically stable and highly proliferative, and have a  
19 wealth of epigenetic data available for multi-omic analyses. We identify 1,617 *cis*-rtQTLs and  
20 analyze their locations and allelic differences. These analyses delineate the architecture of  
21 human replication timing as a quantitative trait involving combinatorial regulation by several  
22 layers of epigenetic mechanisms rooted in *cis*-acting DNA sequences.

23  
24 **High-resolution population-scale replication timing profiles**

25 To comprehensively characterize human inter-individual replication timing variation and its  
26 genetic basis, we analyzed deep (~30x) whole-genome sequences of 121 hESC lines and 326  
27 iPSC lines<sup>19</sup> and sequenced another 24 hESCs and 17 iPSCs for a total of 488 cell lines  
28 (Methods). ES and iPS cultures are highly proliferative, containing 35–55% cells in S phase.  
29 DNA replication timing leads to variation in DNA copy number along chromosomes among S  
30 phase cells (e.g., early-replicating regions are duplicated in most cells), causing read depth  
31 fluctuations in the sequencing data<sup>4</sup>. Indeed, we were able to generate high-resolution  
32 replication timing profiles for a total of 140 hESCs and 317 iPSCs (Methods). ES and iPS cells  
33 had similar replication profiles, as expected.

34 Replication timing profiles were continuous along chromosomes, highly reproducible among  
35 samples (median  $r = 0.93$ ), and consistent with previous replication timing measurements by  
36 Repli-Seq (median  $r = 0.86$ ; Fig. 1, A–D). The replication profiles were exceptionally sharp, in  
37 line with recent high-resolution Repli-Seq data<sup>20</sup>, with discrete peaks and valleys (local maxima  
38 and minima) that were themselves highly reproducible among individuals. Replication timing  
39 peaks represent prominent initiation sites containing one or more replication origins. We further

1 improved data resolution using principal component (PC)-based correction across cell lines  
2 (Fig. 1, C and D; Methods).

3

4 **DNA replication timing is broadly influenced by *cis*-acting sequences**

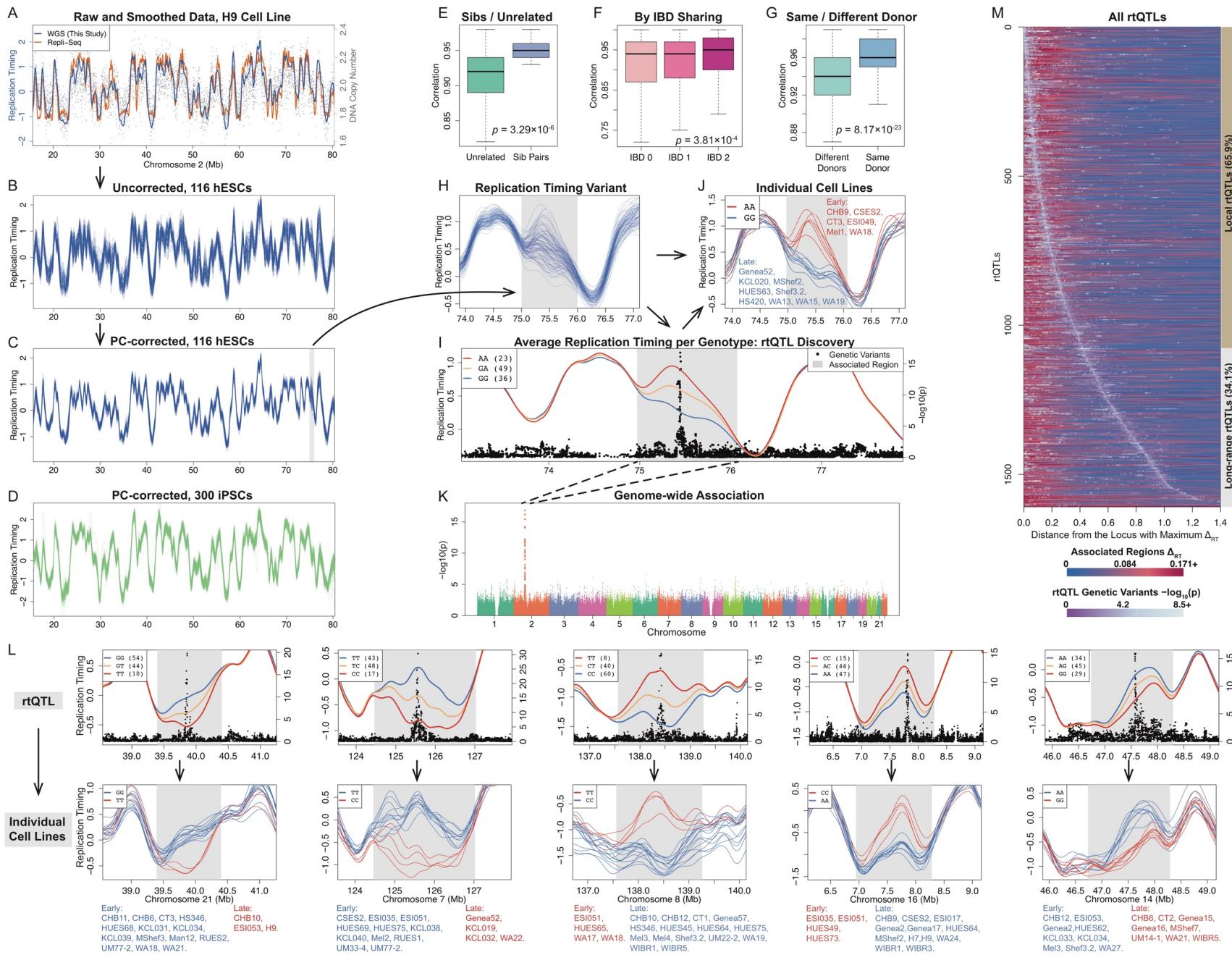
5 While replication timing profiles were highly reproducible among individuals, we nonetheless  
6 observed genomic regions with substantial inter-individual variation. We identified 1,489  
7 autosomal replication timing variants in hESCs and 1,837 in iPSCs, cumulatively  
8 encompassing 795 Mb (34%) and 980 Mb (40.8%), respectively, of the analyzable genome  
9 (Fig. 1, C and D). We hypothesized that at least some of this variation is due to genetic  
10 polymorphism. To test this, we first compared replication timing variation between 24 pairs of  
11 hESC lines that are genetic siblings, versus unrelated cell lines; between genomic regions that  
12 are identical by descent (IBD), half-identical or non-identical between sibling cell lines; and  
13 between 108 pairs of iPSC lines derived from the same donor, compared to different donors  
14 (Methods). Consistent with a significant genetic contribution to replication timing variation,  
15 samples or genomic intervals that are genetically related consistently showed greater  
16 replication timing similarity than unrelated comparisons (Fig. 1, E–G).

17 To further dissect genetic contributions to replication timing variation, we used our previously  
18 described rtQTL mapping approach<sup>4</sup> to associate replication timing with specific genetic  
19 polymorphisms. This approach was applied here at larger scale, to deeper-sequenced data,  
20 and with refined algorithms than before (Methods). We limited this analysis to 108 hESCs of  
21 European ancestry and to 192 iPSCs from different individuals.

22 We identified 1,617 *cis*-rtQTLs (FDR 0.1; 1,012 were identified with FDR 0.05; Fig. 1, I–M;  
23 Table S1), two orders of magnitude more than previous associations of replication timing with  
24 *cis*-acting sequences<sup>4,12</sup>. We used CAVIAR<sup>21</sup> to fine-map (90% credible level) a median of 33  
25 SNPs per rtQTL, with 316 rtQTLs mapped to within 10 SNPs and 36 rtQTLs mapped to no  
26 more than three SNPs. rtQTL mapping was cross-validated between ES and iPS cells and  
27 further confirmed using additionally-sequenced cell lines and with a locus-specific single-  
28 molecule assay (Fig. S1).

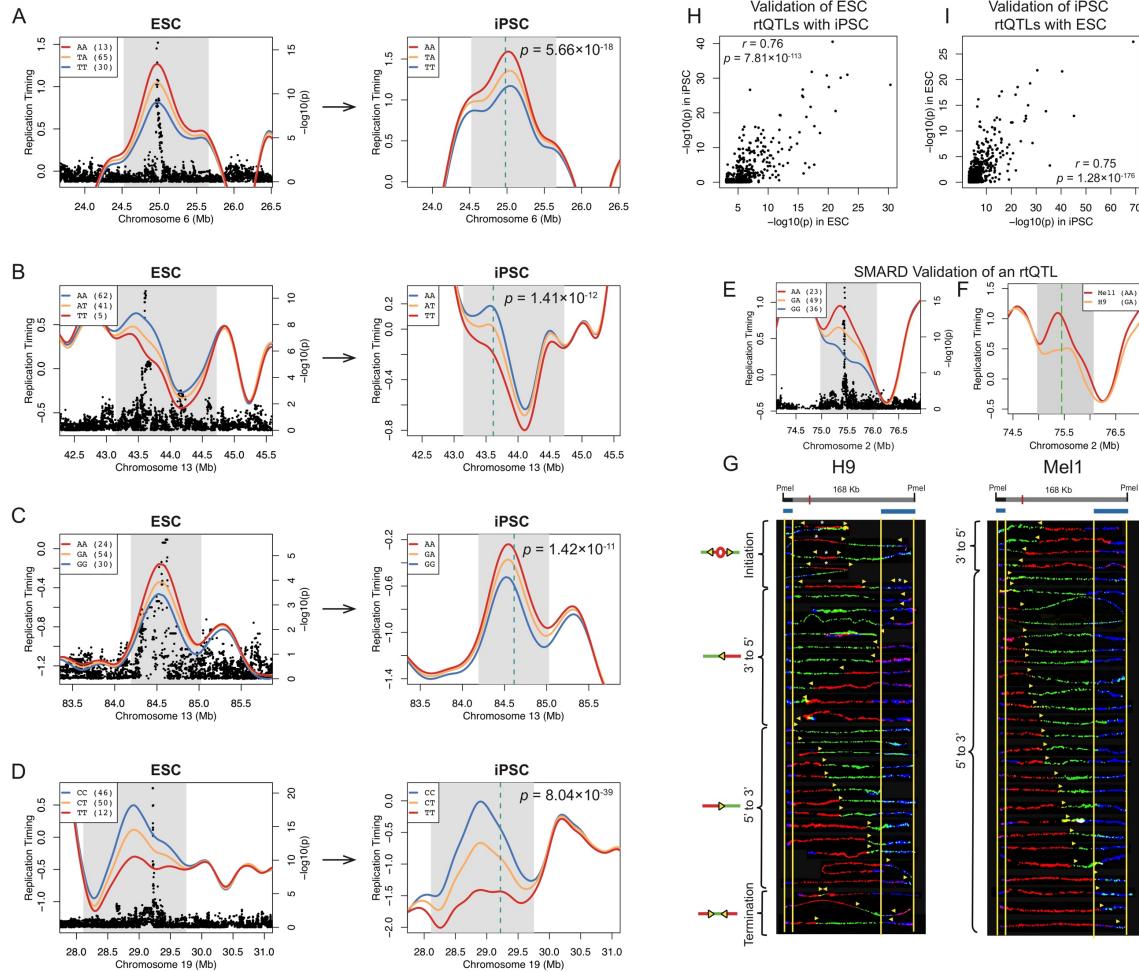
29 rtQTLs influenced the replication timing of regions spanning 858 kb on average and a total of  
30 741.8 Mb of genomic sequence (31.8% of the genome, Fig. 1M). This is a lower bound  
31 estimate of the extent to which human replication timing is influenced by DNA sequence, since  
32 our approach will not detect weaker rtQTLs, invariant sequences or rare variant rtQTLs.  
33 Intriguingly, 67.9% (1,098) of rtQTLs coincided with sharp peaks in the replication profiles  
34 (binomial  $p = 2.24 \times 10^{-25}$ ; Fig. 1, I–L), and rtQTL SNPs were significantly closer to peaks than  
35 expected (Wilcoxon rank-sum  $p = 1.77 \times 10^{-16}$ ). This suggests that rtQTLs may influence  
36 replication initiation, as previously reported<sup>4,22</sup>, and that most rtQTLs can be used as fine-scale  
37 markers of replication initiation regions. The identification of rtQTLs as precise genetic

- 1 determinants of replication timing provides a unique opportunity to fine-map molecular
- 2 mechanisms controlling replication initiation and timing.



1 Figure 1. The Human Genome's Replication Timing Program is Extensively Encoded in DNA.  
2 (A) Replication timing inferred from read depth fluctuations. Read depth (gray dots) and replication timing profile  
3 (blue line; Z-score) of the H9 cell line. Green line: Repli-Seq data of the same cell line.<sup>23</sup> (B) Replication timing  
4 profiles are highly reproducible among samples. (C, D) Leveraging the population-scale of the data, PC-based  
5 correction greatly improves replication profile accuracy. (E–G) Genetic relatedness is associated with replication  
6 timing similarity. (E) Comparison of sibling vs. unrelated hESC lines. (F) Genomic regions stratified by increasing  
7 identity-by-descent. (G) iPSCs from the same or different donors. (H) A genomic region (gray) with substantial  
8 inter-individual replication timing variation. (I–K) Genetic association reveals rtQTLs. (I) A SNP haplotype strongly  
9 associates in *cis* with the replication timing variant from panel H (panel K shows the genome-wide association).  
10 Mean replication timing profiles (left Y axis) for individuals with different genotypes at rs12713840, the top SNP,  
11 demonstrates that SNPs in *cis* (right Y axis) associate with replication initiation activity. Gray shaded area  
12 represents the affected genomic region. (J) Replication timing at the variant from panel H, stratifying individuals  
13 by rs12713840 genotype, demonstrates that genotype is the main determinant of replication timing variation. (L)  
14 Additional rtQTL examples. Similar to panels I and J. Most rtQTLs affect peaks (replication initiation regions). (M)  
15 All rtQTLs. Each horizontal line is an rtQTL, oriented from the replication timing locus with maximum difference  
16 between early- and late-replicating genotypes ( $\Delta_{RT}$ ) and showing the averaged replication timing difference on  
17 both sides of that locus (i.e., the rtQTL-associated region spans twice the distance shown; refer to panel I).  
18 Foreground (gray-purple) shades are the rtQTL SNPs, color-coded by *p*-values, and placed according to their  
19 distance to the locus of maximal  $\Delta_{RT}$ . rtQTLs are encoded in localized haplotypes yet influence extended genomic  
20 regions up to 5.6 Mb. Most rtQTLs influence surrounding genomic region (“local”), while a subset show long-range  
21 effects.

22



1

2 Figure S1. rtQTL Validation.

3 (A–D) Validation of rtQTLs in 192 iPSC lines (Methods). The left panels are examples of rtQTLs in hESCs. The  
4 right panels show replication timing in the same regions in iPSCs, stratified by the genotype of the top rtQTL SNP  
5 discovered in the hESCs (vertical line). Association  $p$ -values in iPSCs are indicated. Excellent agreement  
6 between hESCs and iPSCs demonstrate that the rtQTLs discovered in hESCs are reproducible in an independent  
7 cohort. (E–G) SMARD (single-molecule analysis of replicated DNA<sup>22</sup>) analysis of an rtQTL on chromosome 2  
8 (Fig. 1I) in Mel1 and H9 cell lines confirms variation in initiation site activity consistent with rtQTL genotypes. (F)  
9 Replication timing flanking the rtQTL locus (gray region); green line: the region analyzed by SMARD. The initiation  
10 site on the left side of the green line is an rtQTL (panel E), at which Mel1 and H9 carry the early-replicating and  
11 heterozygous genotype, respectively. (G) SMARD results, where each line indicates one DNA molecule, and the  
12 shift from red to green reveals the location and direction of replication forks (yellow arrows). Significantly more  
13 forks are progressing from 5' to 3' in Mel1 when compared with H9 ( $p = 0.027$ , Fisher's exact test), indicating that  
14 the upstream initiation site is much stronger in Mel1 than H9, consistent with the rtQTL analysis. (H, I) rtQTLs are  
15 highly reproducible between the ESCs and iPSCs. When directly testing ESC rtQTLs using iPSCs (H) or vice  
16 versa (I), the  $p$ -values show strong positive correlation. Among the 602 ESC rtQTLs tested, 38.7% (233/602) were  
17 validated ( $p < 0.05$  and the same direction of effect) in at least one dataset (HipSci iPSC or ESC/iPSC additionally  
18 sequenced), much greater than expected ( $p = 1.15 \times 10^{-80}$ , binomial test). For rtQTLs with  $p \leq 5 \times 10^{-8}$ , 85.6%  
19 (89/104) were validated ( $p = 3.75 \times 10^{-74}$ ). Among the iPSC rtQTLs tested, 31.7% (303/955) were validated in ESC  
20 ( $p \ll 2.2 \times 10^{-16}$ ). For iPSC rtQTLs with  $p \leq 5 \times 10^{-8}$ , 82.3% (149/181) were validated ( $p \ll 2.2 \times 10^{-16}$ ).

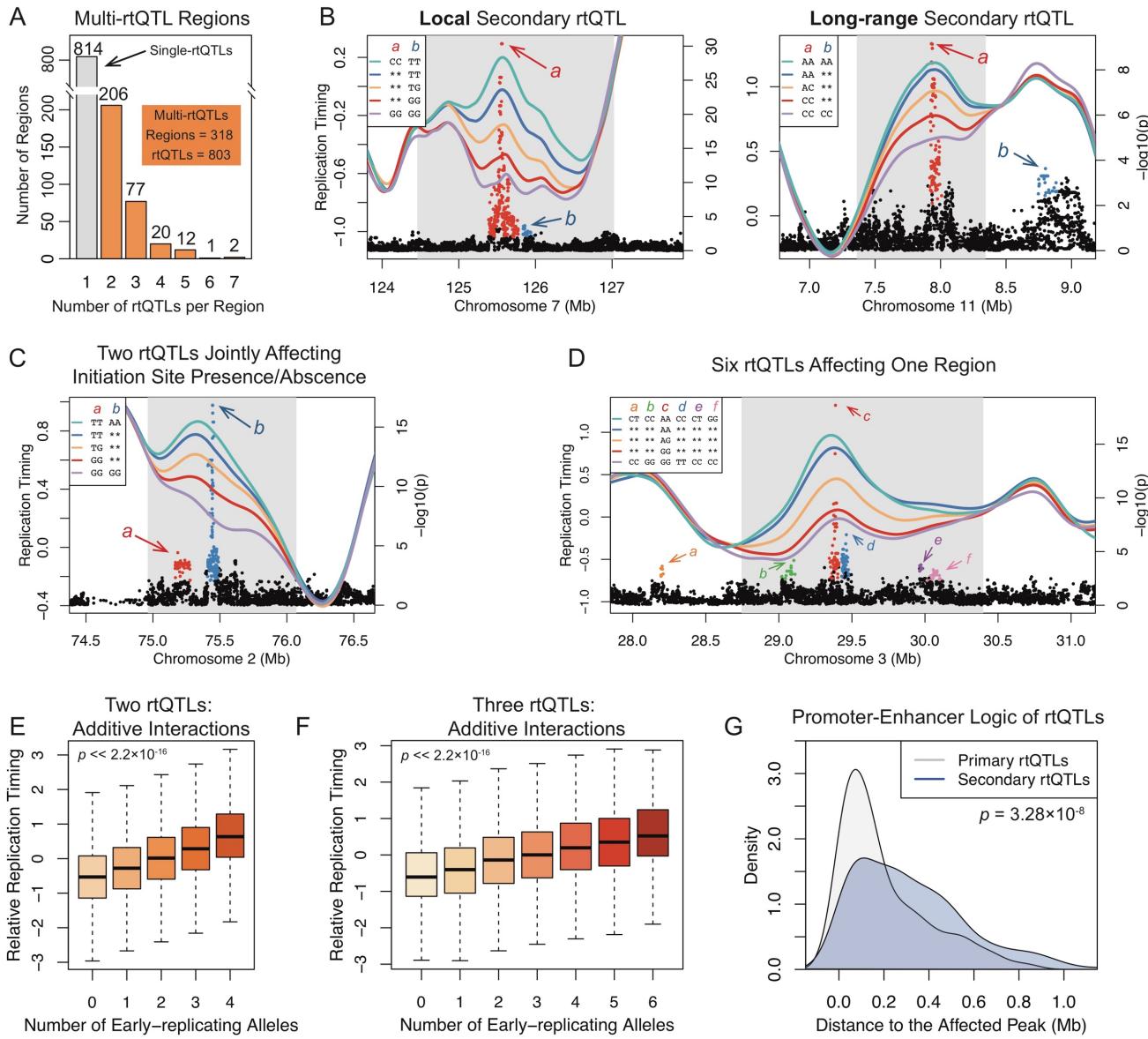
## 1 A promoter-enhancer logic of replication timing regulation

2 We first used rtQTLs to examine the *cis*-regulatory logic of human replication timing. We  
3 observed that a subset of rtQTLs are distal to their associated genomic region (Fig. 1M), and  
4 that in many regions, separate rtQTLs clustered in close proximity. This suggests that multiple  
5 DNA sequences, local and distal, could interact to affect the replication timing of a given locus.  
6 We identified 318 cases, encompassing 803 rtQTLs, where at least two, and strikingly, up to  
7 seven rtQTLs were associated with the same region, each providing additional explanatory  
8 power (Fig. 2, A–D). We call these “multi-rtQTL” regions and refer to the strongest rtQTL as  
9 the “primary”, while all other rtQTLs are “secondary”. In some cases, one rtQTL quantitatively  
10 influenced replication timing, while several rtQTLs together explained the actual presence of  
11 active initiation (Fig. 2C). Thus, replication initiation is regulated along a continuum, one  
12 extreme of which is no activity at all despite the presence of a potential initiation site.

13 We directly tested for interactions between primary and secondary rtQTLs at regions that  
14 harbored two rtQTLs, hence between zero and four early-replicating alleles. We further pooled  
15 all genomic regions containing three or four rtQTLs and evaluated the relationship between the  
16 number of early-replicating alleles and replication timing of the associated regions. Replication  
17 timing showed a linear relationship with the number of early-replicating alleles (linear  
18 regression  $p << 2.2 \times 10^{-16}$ ; Fig. 2, E and F), and none of the individual regions showed  
19 evidence for synergistic interactions between rtQTLs. This suggests that primary and  
20 secondary rtQTLs additively affect local replication timing.

21 Of the 318 multi-rtQTL regions, 176 were associated with replication timing peaks. In 115 of  
22 these cases (65.3%), primary rtQTLs were closer to the peak than secondary rtQTLs (Fig. 2G,  
23  $p = 3.28 \times 10^{-8}$ ). This resembles eQTLs (expression QTLs), in which primary eQTLs show  
24 stronger enrichment at promoters, while weaker eQTLs are enriched at enhancers<sup>24</sup>. Also in  
25 resemblance to enhancers and promoters, primary and secondary rtQTLs tended to cluster in  
26 nuclear space (based on Hi-C data) more than expected by chance ( $p = 9.73 \times 10^{-3}$ , Z-test).  
27 Drawing from this analogy, we propose that rtQTLs may follow a logic akin to promoters and  
28 enhancers, in which primary rtQTLs function as main *cis*-acting regulators of replication  
29 initiation, while other sequences, marked by secondary rtQTLs, serve as distal regulatory  
30 elements that fine-tune the replication dynamics of a given region.

31



1

2 Figure 2. Multiple DNA Sequences Interact to Regulate Replication Timing.

3 (A) Hundreds of regions are controlled by multi-rtQTLs. (B, C) Two rtQTLs affecting the same region. Blue,  
4 yellow, and red lines represent one rtQTL. Purple and green lines represent the mean replication timing of  
5 individuals carrying the late- or early-replicating genotypes, respectively, at both rtQTLs. Considering both rtQTLs  
6 explains a larger fraction of variation (green lines are higher than blue lines; conversely for purple/red lines).  
7 Asterisks (in legends): any genotype at this rtQTL. In panel C, the GG/GG combination of alleles is associated  
8 with complete loss of initiation activity. (D) A replication initiation site associated with six rtQTLs. Each rtQTL was  
9 significant even after conditioning on all other five rtQTLs in the region. (E, F) rtQTLs exert additive effects. All  
10 regions with two (E) or three (F) rtQTLs were pooled; replication timing is linearly correlated to the number of  
11 early-replicating alleles. (G) Multi-rtQTLs conform to a “promoter-enhancer” logic, primary rtQTLs being closer to  
12 the affected replication timing peak than secondary rtQTLs.

13

## 1 A histone code for DNA replication initiation

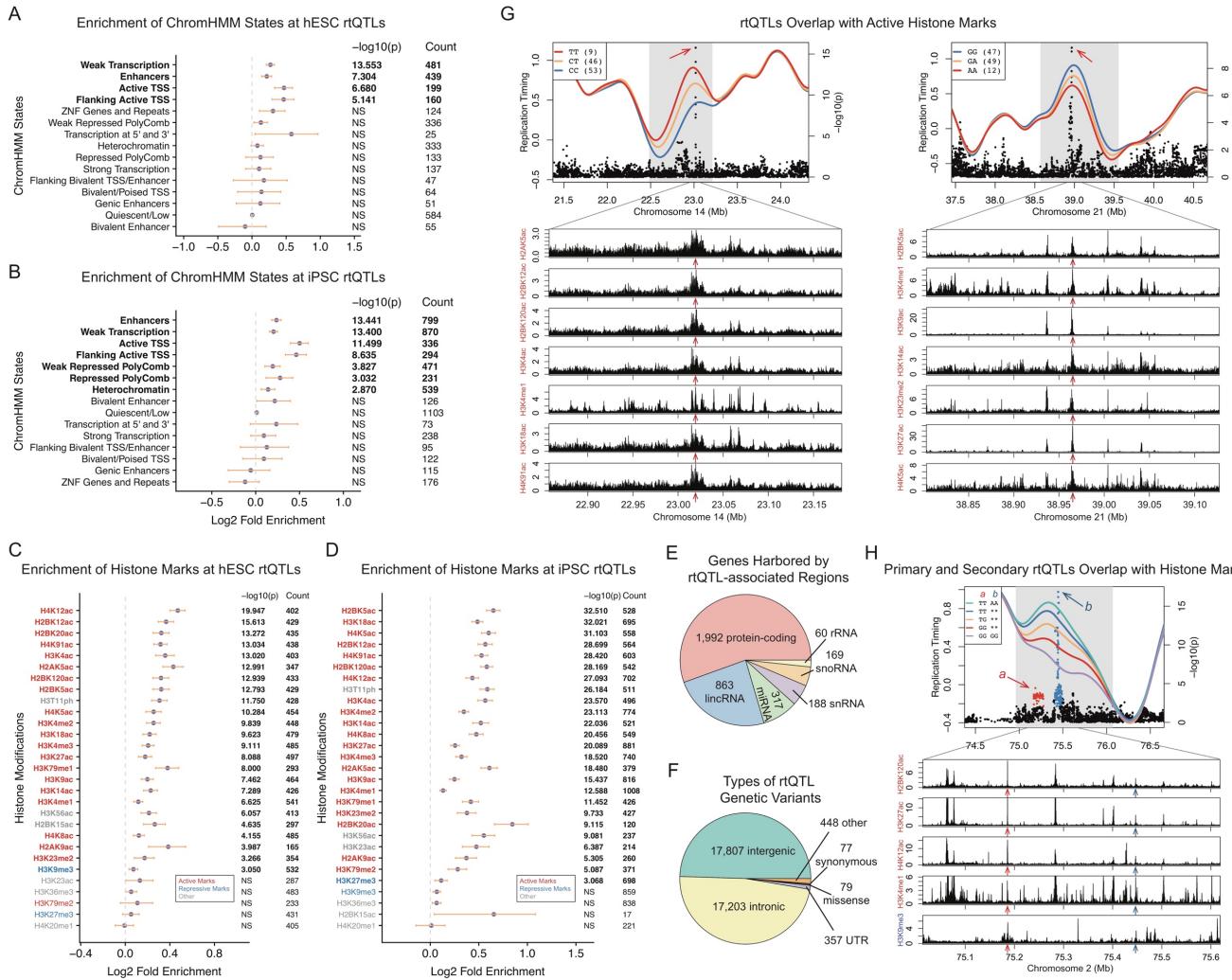
2 We next utilized the basepair-resolution sequence-specificity of rtQTLs to investigate the  
3 molecular determinants of DNA replication timing. We initially considered rtQTL locations *per*  
4 *se*, independently of allelic variation. Since extensive epigenetic data was available for seven  
5 of the hESC lines in our dataset, we focused this analysis on hESCs and used iPSCs for  
6 validation. Consistent with previously described correlations between early replication and  
7 open chromatin<sup>3</sup>, rtQTLs were enriched for active chromHMM states including enhancers and  
8 transcription start sites (although they were not specifically associated with genes; Fig. S2),  
9 DNase I hypersensitivity sites ( $p = 2.62 \times 10^{-8}$ ,  $4.11 \times 10^{-19}$  in iPSCs), and H2A.Z sites<sup>25</sup> ( $p =$   
10  $6.69 \times 10^{-4}$ ;  $p = 3.20 \times 10^{-17}$  in iPSCs). rtQTLs also significantly overlapped with 24 histone marks  
11 (25 in iPSCs), of which 20 were active marks (Fig. S2). The majority of these histone marks  
12 were acetylations, including several not linked to replication timing before, for example,  
13 H2BK120ac, H2BK12ac and H2BK20ac. H3T11ph was also consistently enriched at hESC  
14 and iPSC rtQTL sites, as so were, modestly, methylated forms of H3K4.

15 Of note, the histone mark enrichments were modest, and each present at between 165 to 541  
16 (median: 429) of 608 hESC rtQTLs (median: 542 of 1,167 iPSC rtQTLs), while each rtQTL  
17 overlapped 20 histone marks on average. We surmised that this abundance of histone  
18 modifications may be suggestive of combinatorial regulation. To test this, we systematically  
19 searched for combinations of histone marks with stronger enrichments at rtQTLs when  
20 considered jointly (Methods). We identified 152 combinations of two overlapping histone marks  
21 that were more enriched than the individual marks. We further identified 128 co-enriched three-  
22 mark combinations, 72 four-mark combinations, and 13 five-mark combinations (enrichment *p*-  
23 values:  $2.42 \times 10^{-37}$ – $1.09 \times 10^{-45}$ ), at which point no further improvements in enrichment were  
24 obtained (Fig. 3A). Importantly, since these enrichments controlled for replication timing, they  
25 were not identified because they mark early-replicating regions, but because they specifically  
26 mark rtQTL locations, and, by inference, replication initiation sites.

27 Strikingly, all 13 combinations of five histone marks contained the trimethylation marks  
28 H3K9me3 and H3K36me3, and 12 of the combinations also contained H3K4me3. In addition,  
29 all 13 combinations included at least one histone acetylation mark. H3K56ac was included in  
30 11 of the combinations, while the additional acetylations occurred on variable histone residues  
31 (Fig. 3B). Further analysis indicated that various acetylation marks often coincided with the five  
32 histone mark combinations, e.g., in 70.8% of the cases, 11 or more acetylation marks co-  
33 occurred at the location of a five-mark combination. We term this combination of three H3  
34 trimethylations together with hyperacetylation the “ $\text{me}^3\text{ac}^{\text{hyper}}$  histone code”. Genome-wide,  
35 there were 6,670 such locations in hESCs. They covered a median of 635 bp and cumulatively  
36 encompassed 0.24% of the genome, thus they represent specific, localized genomic sites.

37 Importantly, when considered individually, the implicated histone modifications only showed  
38 weak enrichments (Fig. S2C). H3K9me3 and H3K36me3, in particular, showed marginal or no  
39 enrichment at rtQTLs. H3K9me3 is a marker of heterochromatin (although has been observed

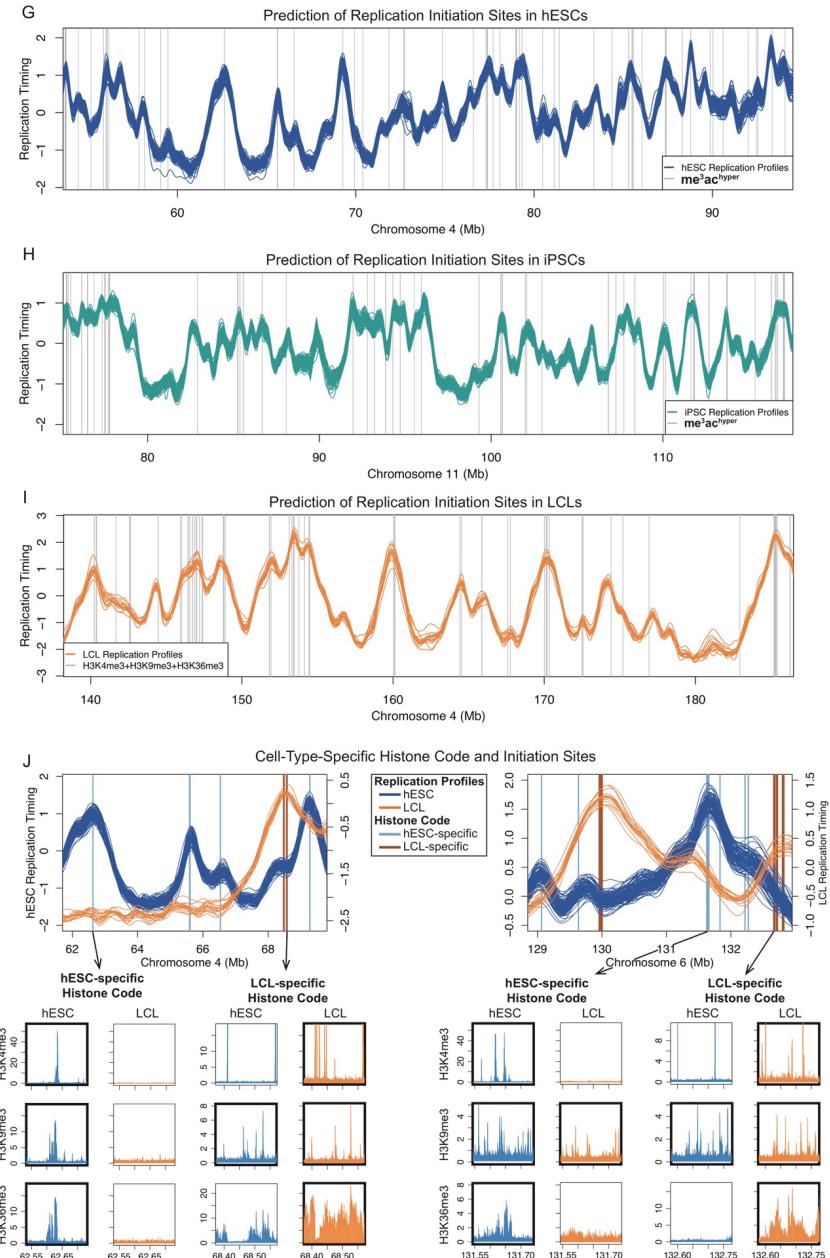
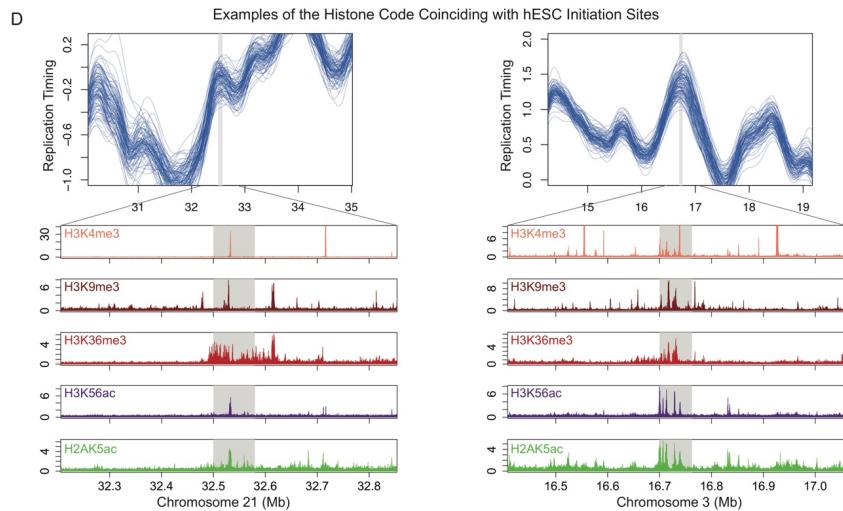
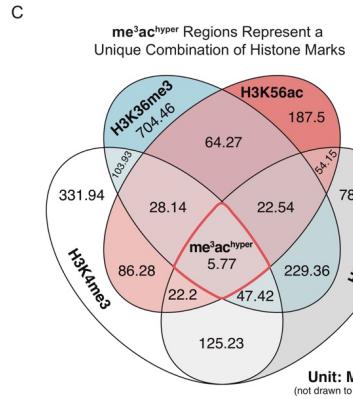
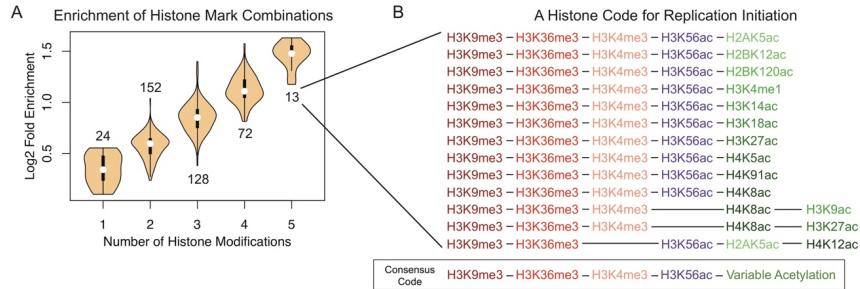
1 in the bodies of actively transcribed genes)<sup>26</sup>, while H3K4me3 marks gene promoters and  
2 H3K36me3 is typically present in gene bodies. These histone trimethylations are largely  
3 mutually exclusive. However, in rare cases, they coincide in the same genomic locations (Fig.  
4 3, C and D) to form a previously undescribed bivalent chromatin that is not specifically  
5 associated with genes or gene expression (e.g., only 5.2% of me<sup>3</sup>ac<sup>hyper</sup> regions overlapped  
6 the TSS of an active gene). It is in these rare locations that rtQTLs tend to be present.  
7 The identified histone mark combinations have been previously linked to the recruitment of  
8 components of the replication machinery to DNA. Histone H3 trimethylations on lysines 4, 9  
9 and 36 have been shown to exert a cross-talk that serves as an “epigenetic addressing  
10 system” for site-specific replication initiation<sup>27,28</sup>. They recruit KDM4 and KDM5 family histone  
11 demethylases that directly interact with, and/or are required for recruitment to DNA of MCM,  
12 PCNA, DNA polymerases and other replication factors<sup>27-31</sup>. H3K4me3 also synergizes with  
13 flanking H3K9ac and H3K14ac (both identified as part of the me<sup>3</sup>ac<sup>hyper</sup> histone code) to recruit  
14 chromatin readers to DNA<sup>32</sup>. Another study showed that histone hyperacetylation synergizes  
15 with H3K9me3 to promote early replication of otherwise late-replicating mouse  
16 chromocenters<sup>33</sup>. In turn, acetylated histones have been shown to recruit replication initiation  
17 factors including TICRR/TRESLIN, ORC and MCM, via mediators such as BRD2, BRD4 and  
18 the histone acetyltransferase HBO1 (histone acetyltransferase binding to ORC)<sup>14,34-36</sup>. In  
19 particular, HBO1 promotes MCM loading by acetylating H4 on lysines 5, 8 and 12, and  
20 subsequently promotes origin activation by acetylating H3K14<sup>37</sup>; we identified all of these  
21 acetylations as part of the me<sup>3</sup>ac<sup>hyper</sup> combinations. Moreover, H4K12ac, the most strongly  
22 enriched mark at rtQTLs, is a preferred target of HBO1 at replication origins<sup>34,36</sup>. These  
23 biochemical evidence provide a plausible explanation for the combination of histone marks  
24 being associated with replication initiation activity.  
25 Taken together, we identified a combination of histone marks, consisting of three trimethylated  
26 H3 residues (H3K4me3, H3K9me3, H3K36me3) together with H3K56ac and broadly  
27 hyperacetylated chromatin that consistently coincide with rtQTLs. To further test the  
28 involvement of this histone “code” in replication initiation, we analyze below its association  
29 with: (1) replication timing peaks in general (independent of rtQTLs); (2) replication timing  
30 peaks in other cell types; (3) replication timing peaks that vary between cell types; and (4)  
31 replication timing variation among individuals at rtQTLs.  
32



1

2 Figure S2. rtQTLs are Enriched for Active Chromatin States and Histone Marks.

3 (A, B) Enrichment of chromHMM chromatin states at rtQTLs identified in hESCs (A) or iPSCs (B). Orange bars: 4 95% confidence intervals. NS: not significant at Bonferroni-corrected  $p = 0.05$ . (C, D) Enrichment of histone marks 5 at hESC (C) and iPSC (D) rtQTLs. Similar to panels A and B. (E) Breakdown of gene types located within rtQTL- 6 associated regions. The number of genes in rtQTL-associated regions was significantly lower than expected ( $p = 7 4.85 \times 10^{-17}$ , Z-test) and these genes were not enriched for any gene ontology term<sup>38</sup>. (F) Breakdown of functional 8 annotations of rtQTL genetic variants. (G) rtQTLs colocalize with active histone modifications. The bottom panels 9 show ChIP-seq tracks of various active histone modifications in hESC. Imputed histone tracks<sup>39,40</sup> from the 10 Roadmap Epigenomics Project were used for plotting. Red arrows: locations of the rtQTL variants indicated in the 11 top panels. (H) A multi-rtQTL region (same as Fig. 2C) at which both the primary and secondary rtQTLs overlap 12 with various active histone marks.



1 Figure 3. A Histone Code for Replication Initiation.

2 (A) Iterative identification of histone mark combinations enriched at rtQTLs. Shown are enrichment distributions;  
3 the number of combinations in each category is indicated. Fold-enrichment increases gradually and is maximal for  
4 five-mark combinations. (B) A histone code for human replication initiation. The 13 combinations of five histone  
5 marks converged to a consensus “code”. (C) The histone code represents a rare combination of both active and  
6 repressive histone marks.  $\text{me}^3\text{ac}^{\text{hyper}}$  regions comprised 0.7–3% of the regions that carry the individual histone  
7 marks. (D) Examples of histone mark combinations (Roadmap Epigenomics imputation)<sup>39,40</sup> coinciding with  
8 replication timing peaks not identified as rtQTLs. (E, F) Distribution (after subtraction of permutations) of physical  
9 (E) and fractional distances (F) of the  $\text{me}^3\text{ac}^{\text{hyper}}$  locations to the nearest replication timing peak. (G) Combination  
10 of histone marks (gray,  $\text{me}^3\text{ac}^{\text{hyper}}$  locations) predict replication initiation sites in hESCs. (H, I) Histone code  
11 locations (gray vertical lines) correspond to replication timing peaks in iPSCs (H) and LCLs (I). (J) Cell-type-  
12 specific histone code locations mark cell-type-specific replication initiation sites. At regions with distinct replication  
13 timing profiles for hESCs and LCLs, LCL (hESC)-specific replication timing peaks are predicted by LCL (hESC)-  
14 specific histone code locations. Lower panels: initiation sites coincide (thick borders) with all three histone  
15 trimethylation marks in the cell type in which they are active, but with one or none of the marks in the cell type in  
16 which they are inactive.

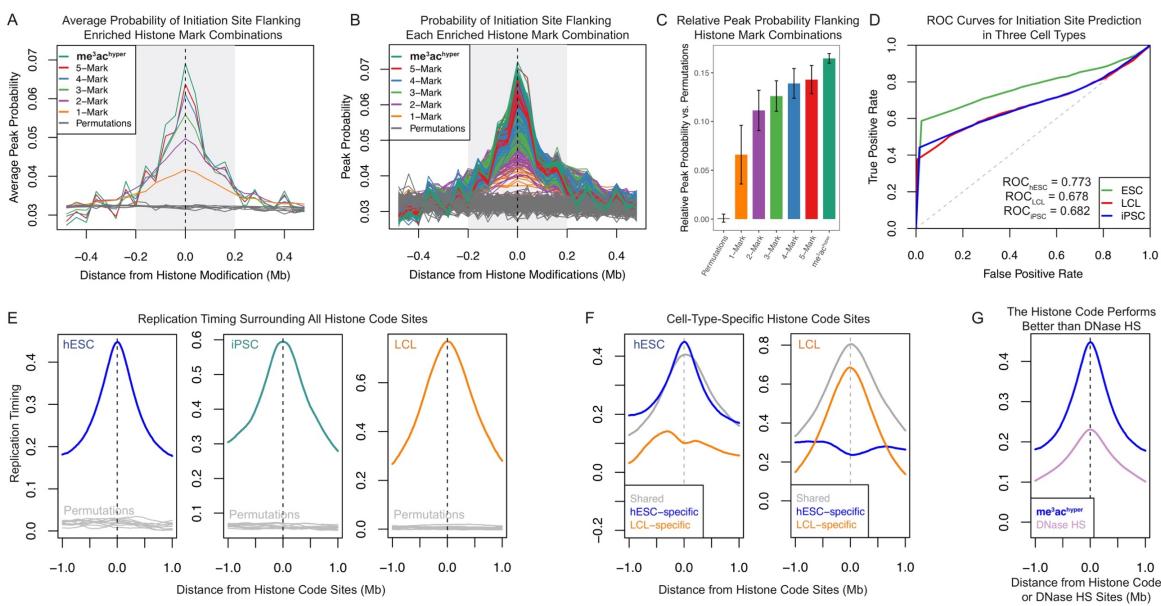
17

## 18 The histone code predicts replication initiation sites across cell types

19 We considered whether a histone code could be a general property of replication initiation  
20 sites, revealed by leveraging the base-pair resolution of rtQTLs, but not limited to rtQTLs. We  
21 therefore tested whether the histone code also associated with the larger number of replication  
22 timing peaks (found in > 10% of the samples) not identified as rtQTLs (81.5% of all peaks).  
23 While the probabilities of having a peak near the 24 individually-enriched histone marks were  
24 significantly greater than expected (one-tailed Wilcoxon rank-sum  $p = 4.96 \times 10^{-17}$ ) and were  
25 greatest at the actual histone mark sites (Fig. S3, A and B), individual histone marks are very  
26 common in the genome and insufficient for predicting peaks. Combining histone marks  
27 gradually increased their association with peaks, up to the five-mark combinations, which were  
28 significantly more likely than expected to coincide with peaks ( $p = 4.10 \times 10^{-10}$ ).  $\text{me}^3\text{ac}^{\text{hyper}}$  sites  
29 had an even higher likelihood of overlapping peaks (Fig. S3, A–C). The distances of  $\text{me}^3\text{ac}^{\text{hyper}}$   
30 regions to the nearest peak were significantly shorter than permutations (Fig. 3, E and F). Of  
31 all  $\text{me}^3\text{ac}^{\text{hyper}}$  sites, 57.3% corresponded to replication timing peaks within 100 kb (positive  
32 predictive value; Z-test  $p < 2.2 \times 10^{-16}$ ; the median inter-peak distance was 971.2 kb); 41.7%  
33 were less than 10 kb from a peak. Conversely, 70.8% of peaks were located within 100 kb of  
34 predicted regions (sensitivity;  $p = 1.03 \times 10^{-93}$ ); 60.3% were less than 10 kb from predicted  
35 regions. We further evaluated prediction performance of the  $\text{me}^3\text{ac}^{\text{hyper}}$  regions visually (Fig.  
36 3G) and with ROC curves (Fig. S3D). Peaks predicted by histone marks replicated earlier than  
37 other peaks (median: 0.61 vs. 0.14, Wilcoxon rank-sum  $p = 6.58 \times 10^{-53}$ ) and were locally more  
38 prominent (timing difference compared to flanking valleys, median: 0.32 vs. 0.18,  $p = 1.19 \times 10^{-17}$ ). Consistently, the replication profiles surrounding  $\text{me}^3\text{ac}^{\text{hyper}}$  sites formed a sharp peak (Fig.  
39 S3E). The histone code was substantially more specific and matched replication timing profiles  
40 much better than DNase I hypersensitivity (Fig. S3G), which was previously suggested to  
41 explain 87% of replication timing profiles<sup>41</sup>. Taken together, the combinations of histone marks  
42

1 that are enriched at rtQTLs predict ~70% of initiation site locations across the genome, even  
 2 for those sites without rtQTLs, and particularly for the early and most prominent initiation sites.  
 3 These histone mark combinations may thus promote replication initiation not just at specific  
 4 genomic loci, as previously proposed<sup>27,28,42</sup>, but across a large fraction of the genome. We  
 5 note, though, that some replication timing peaks did not co-localize with histone code  
 6 locations, thus there must be additional mechanisms independently specifying replication  
 7 initiation sites, underscoring the complexity of mammalian replication initiation.  
 8 An even more rigorous test of the five-mark combinations being indicators of replication  
 9 initiation is whether they could predict the location of replication timing peaks in other cell  
 10 types. Examining both iPSCs and lymphoblastoid cell lines (LCLs)<sup>43-45</sup>, we found that the  
 11 histone code can predict initiation sites as accurately and specifically as in hESC (Fig. 3, H and  
 12 I), and similarly associates with early replication (Fig. S3E). In particular, LCLs have epigenetic  
 13 and replication timing landscapes that are distinct from those of hESC (and iPSCs). In genomic  
 14 regions at which LCL and hESC replication timing differed, LCL-specific histone code locations  
 15 corresponded to LCL-specific initiation sites, and *vice versa* for hESCs (Fig. 3J). Predicted cell-  
 16 type-specific initiation sites resided in early-replicating genomic regions in the corresponding  
 17 cell type, but not in other cell types (Fig. S3F). Thus, the histone code characterizes and  
 18 predicts cell-type-specific replication initiation.

19



20

21 Figure S3. Further Support for a Histone Code for Human DNA Replication Initiation Sites.

22 (A) Histone mark combinations correspond to replication initiation sites. The probability of having an initiation site  
 23 increases with proximity to the histone mark combinations (gray shade), peaks at the actual histone mark sites,  
 24 and scales with the number of marks. (B) Probability of having an initiation site as a function of distance from  
 25 histone marks (in 40 kb bins), similar to panel A but for each individual histone mark combination (as opposed to  
 26 the averages of all combinations of a given number of marks). (C) Normalized cumulative probability of initiation

1 sites being present within 200 kb (i.e., area under the curve, gray shade in panel A) of individual histone marks or  
2 combinations thereof. The probabilities were normalized based on permutations by subtracting the permutation  
3 mean. Replication initiation sites are increasingly enriched as the number of histone marks increases. Error bar:  
4 standard deviation. Error bars: standard deviation. (D) ROC curves show the strength of the histone code for  
5 predicting replication initiation sites in various cell types. Diagonal lines represent random guesses. For all three  
6 ROC curves, the area under the ROC curve ( $AUC_{ROC}$ ) is significantly larger than random permutations (all Z-test  
7  $p << 2.2 \times 10^{-16}$ ). (E) Cumulative replication timing profiles surrounding histone code locations suggest that they  
8 coincide with locally early replication across cell types. For LCLs, only methylation marks were available. Gray  
9 lines: ten permutations. (F) Cumulative replication timing profiles in hESCs and LCLs surrounding histone code  
10 locations found in both cell types (gray), LCLs only (orange), or hESCs only (blue). Histone code locations predict  
11 replication initiation patterns in a cell-type-specific manner. (G) The histone code performs better at predicting  
12 replication timing peaks than DNase hypersensitivity (HS) sites. Cumulative replication profile centered at histone  
13 code locations (blue) is sharper and higher than that centered at DNase HS sites (purple). In addition, there are  $> 99,000$  DNase HS sites in the genome, totaling  $> 304$  Mb of sequence (i.e.,  $\sim 10\%$  of the genome; in contrast to  
14 the histone code covering 0.24% of the genome), which provides very low positive predictive value and resolution  
15 for predicting individual replication initiation sites.  
16

17

18 **Co-variation of replication timing and histone modifications reveals combinatorial  
19 control of replication timing**

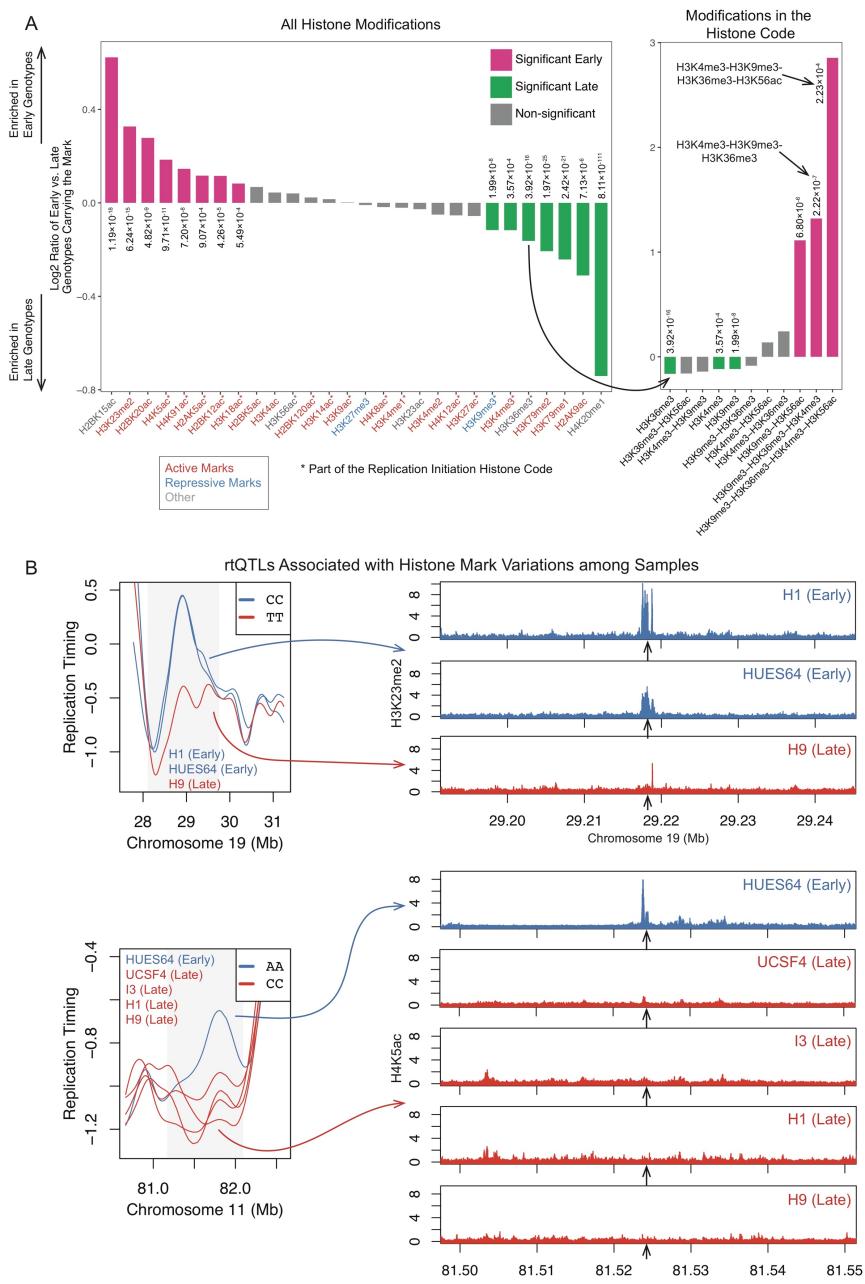
20 The previous analyses considered rtQTL locations *per se*. However, since rtQTLs represent  
21 replication timing variation among individuals, their allelic differences provide a powerful  
22 opportunity to investigate molecular mechanisms controlling replication timing. In particular,  
23 given that specific histone marks associate with replication initiation, we predicted that rtQTL  
24 SNP alleles will be associated with variation in the abundance of these marks among  
25 individuals.

26 We took an unbiased approach using seven hESC lines with both replication timing and  
27 histone modification data (Methods). Cell lines carrying early-replicating genotypes at rtQTLs  
28 were more likely than individuals with late-replicating genotypes to harbor active histone marks  
29 and chromHMM states at those rtQTL sites (Fig. 4 and S4). Across individuals and genomic  
30 sites, eight histone modifications were consistently present in individuals with rtQTL alleles  
31 indicative of early replication. Of those, seven were acetylations, consistent with histone  
32 acetylation promoting early replication<sup>3,13-17,34,36</sup>. Of the 12 acetylation marks that are part of  
33 the replication initiation histone code, nine individually associated with early-replicating rtQTL  
34 genotypes (five of which reached statistical significance). We also identified seven  
35 modifications that consistently coincided with late replicating alleles, of which six were  
36 methylation marks (Fig. 4A); Thus, histone methylation emerges as being generally repressive  
37 for replication.

38 Counter-intuitively, the histone code trimethylation marks (H3K4me3, H3K9me3 and  
39 H3K36me3) were individually more likely to be associated with late-replicating genotypes (Fig.  
40 4A). In contrast, the combination of all three trimethylation marks was 2.5-times more likely to  
41 be carried by early-replicating than by late-replicating genotypes. Furthermore, a combination

1 that also included H3K56ac was 7.24-times more likely to be carried by early-replicating  
 2 genotypes (Fig. 4A). Thus, these marks appear to individually act as weak repressors of  
 3 replication but act synergistically, in non-canonical ways, to strongly promote early replication.  
 4 Taken together, the involvement of  $\text{me}^3\text{ac}^{\text{hyper}}$  in replication initiation is supported by several  
 5 lines of evidence: enrichment at rtQTLs (Fig. 3A); correspondence with replication timing  
 6 peaks in general, and across several cell types (Fig. 3, D–I, Fig. S3, A–E); co-variation with  
 7 cell-type-specific replication initiation patterns (Fig. 3J and Fig. S3F); and correlation with inter-  
 8 individual replication timing variation (Fig. 4).

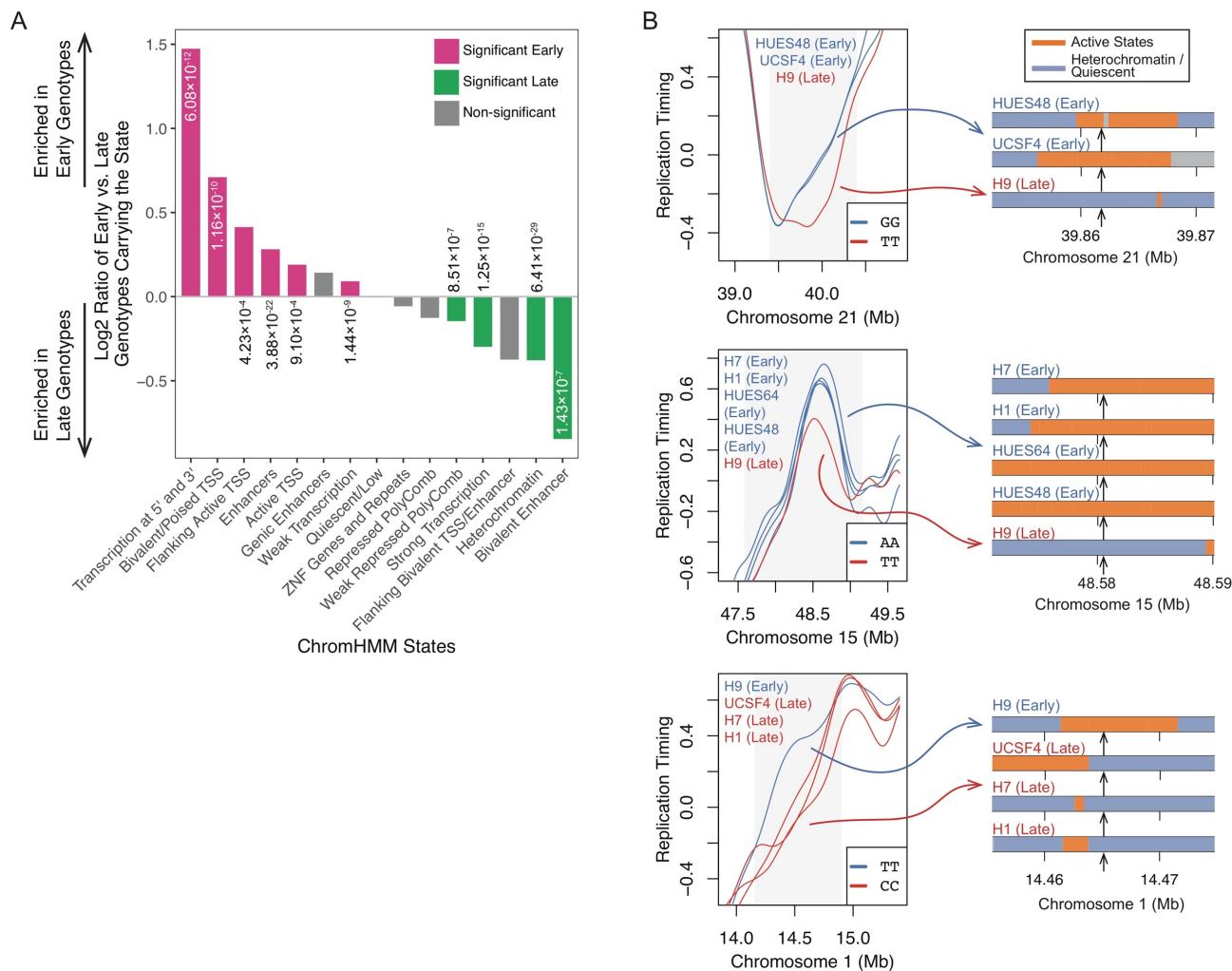
9



1 Figure 4. Histone Marks Affect DNA Replication Timing.

2 (A) Association of rtQTL genotypes with individual (left panel) or combinations (right panel) of histone marks.  
 3 Positive (negative) values indicate that individuals with early (late)-replicating genotypes are more likely to carry a  
 4 histone mark at those rtQTL sites. Right panel: while individual H3 methylation marks associate with late  
 5 replication, the H3K4me3-H3K9me3-H3K36me3 combination is strongly associated with early replication, and  
 6 even more so when combined with H3K56ac. Note the different Y scale. (B) Examples of rtQTLs associated with  
 7 histone mark variations. Replication timing and corresponding histone ChIP-seq tracks for individual cell lines  
 8 homozygous for the early- or late-replicating alleles. Early replication correlates with the presence of the specified  
 9 histone marks.

10



11

12 Figure S4. rtQTLs Impact Replication Timing by Affecting Chromatin States.

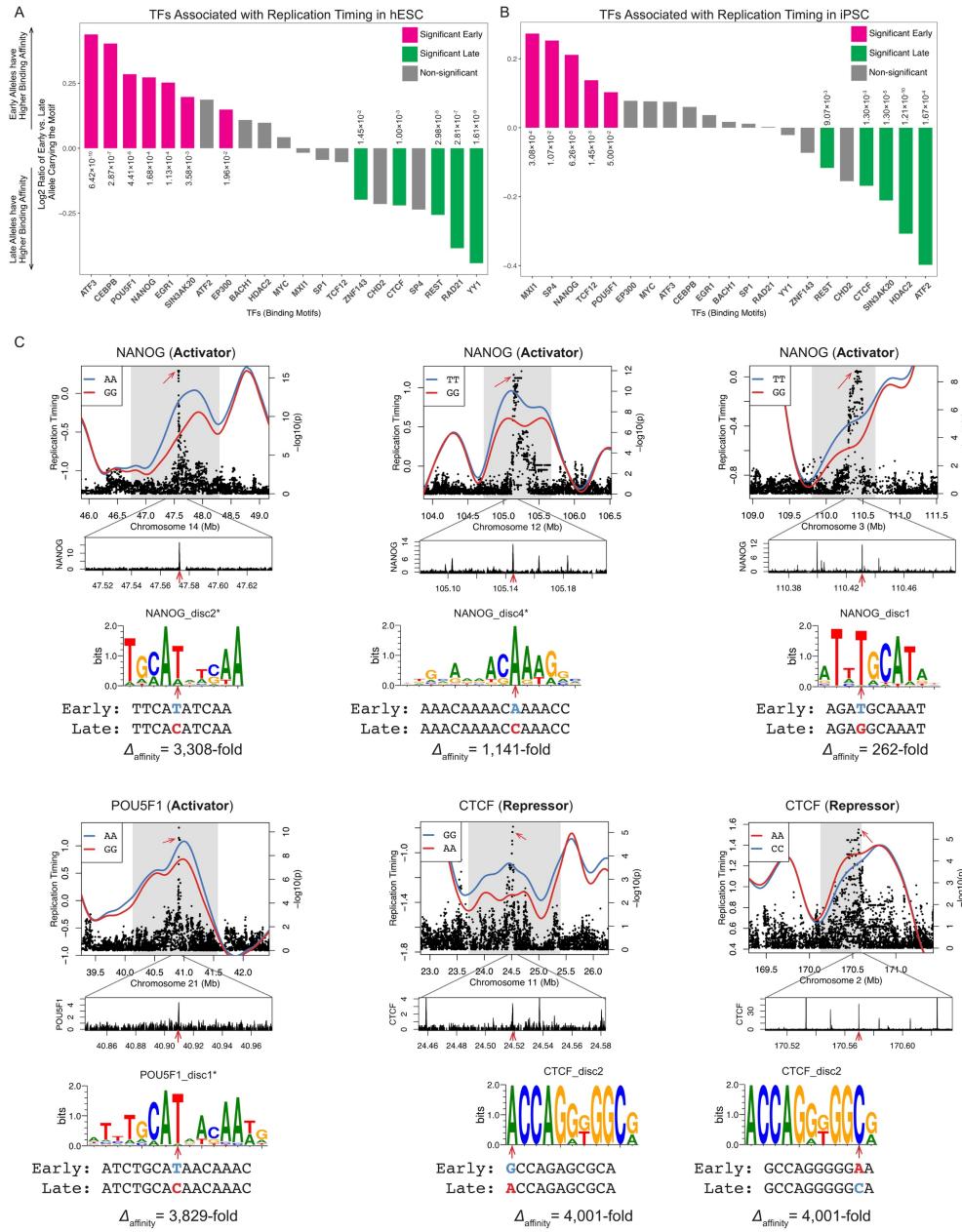
13 (A) Associations of rtQTL genotypes with chromHMM states. Positive values indicate that the early-replicating  
 14 genotypes are more likely to carry a given chromatin state, and vice versa for negative / late genotypes. (B)  
 15 Examples of rtQTLs associated with chromatin states. The right panels show chromatin states flanking the rtQTL  
 16 in the same cell lines. Orange: active states (TSS, enhancer, or weak transcription), blue: heterochromatin or  
 17 quiescent states, gray: other states.

1

2 **DNA-binding factors modulate DNA replication timing**

3 The above results indicate that *cis*-acting sequences, manifesting as rtQTLs, influence the  
4 positions and timing of replication initiation by associating with histone modifications. To  
5 identify additional factors that influence replication timing via *cis*-acting sequences, we  
6 analyzed the binding sites of 51 DNA binding factors in hESCs<sup>43,46</sup>. Binding of eight factors  
7 was significantly enriched at rtQTLs, including the main pluripotency factors SOX2, POU5F1  
8 (OCT4) and NANOG, the latter two reproducible with available data in iPSCs (Fig. S5). Three  
9 chromatin remodelers, EP300 (P300), SP1, and RBBP5, were also enriched at rtQTLs. EP300  
10 is a histone acetyltransferase that catalyzes at least six acetylation marks in the replication  
11 initiation histone code, including H3K56ac<sup>47</sup>.

12 Transcription factors (TFs) bind DNA in a sequence-specific manner at characteristic motifs.  
13 This offers an opportunity to test, at base pair resolution, whether TF binding affects replication  
14 timing at rtQTLs (Methods). Strikingly, OCT4 and NANOG had significantly higher binding  
15 affinity for early- compared to late-replicating alleles in both hESCs and iPSCs, while EP300  
16 and ATF3 (Activating Transcription Factor 3, which is enriched at EP300 sites<sup>48</sup>), were linked  
17 to early replication at least in hESCs (Fig. 6, A and B). These associations appeared to be  
18 independent from gene expression, as they were reproduced for rtQTLs > 250 kb away from  
19 expressed genes. For these early-replication-associated TFs, the rtQTLs fell within the TF  
20 binding motifs such that a single base-pair change disrupted or even abolished binding; this  
21 was associated with delayed replication of the rtQTL-affected initiation site (Fig. 6C). An  
22 unexpected finding was rtQTL alleles with the opposite effect, i.e., higher binding affinity  
23 associated with later-replication. We infer that in these cases protein binding suppresses  
24 replication initiation (Fig. 6). These included CTCF, an insulator of topologically associated  
25 domains (TADs); REST(NRSF), a repressor of transcription<sup>49</sup>; ZNF143, which associates with  
26 the CTCF-cohesin cluster<sup>50</sup>; and at least in hESCs also RAD21 (part of the cohesin complex)  
27 and YY1, which co-localize with CTCF at TAD boundaries<sup>51-54</sup>. These associations were yet  
28 stronger when considering only motifs with biochemically confirmed TF binding when data was  
29 available (Methods). Taken together, we conclude that some rtQTL alleles alter DNA binding  
30 protein motifs, abolish a DNA binding site or generate a new one, and consequently alter DNA  
31 replication timing through specific protein binding. This analysis uncovers several new factors  
32 that can thus regulate DNA replication timing. In addition, different factors influence subsets of  
33 replication initiation sites, further illuminating the complex combinatorial landscape that  
34 controls human DNA replication timing. Finally, these results demonstrate how a single base-  
35 pair alteration could affect the replication timing of an extended genomic region.

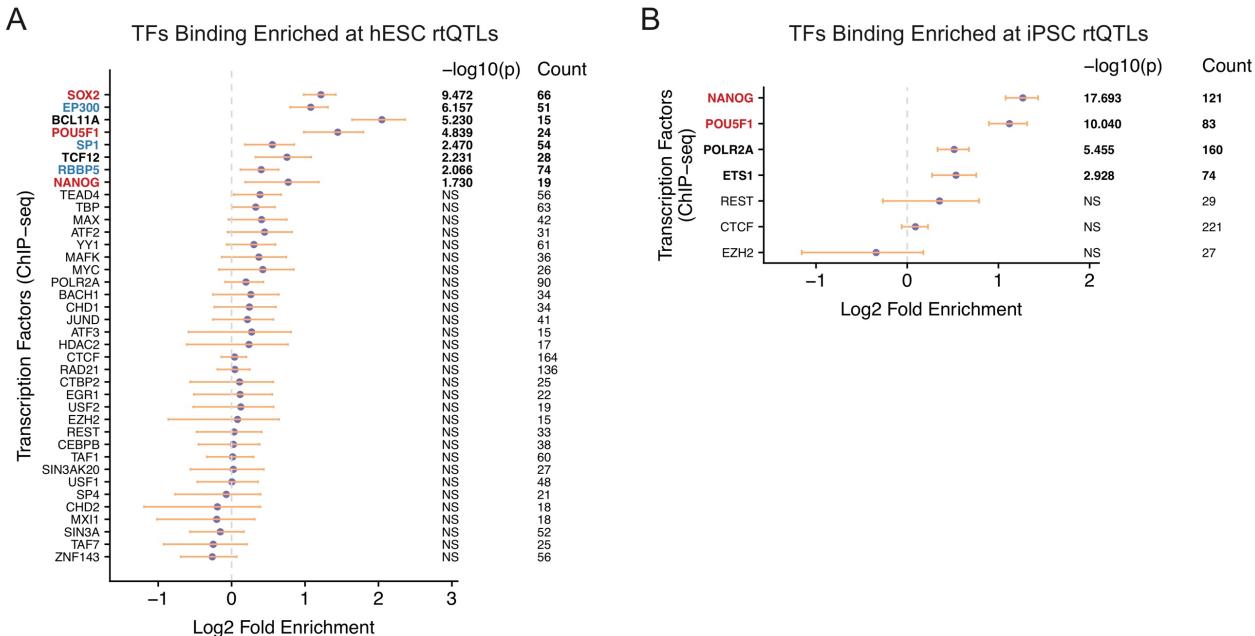


1

2 Figure 5. rtQTLs Affect Replication Timing by Altering TF Binding Motifs.

3 (A, B) Binding of TFs such as OCT4 and NANOG promotes earlier replication, while binding of CTCF, REST and  
4 other factors is associated with later replication in hESCs (A) and iPSCs (B). Chi-squared test, FDR <10%. (C)  
5 Examples of rtQTLs altering binding affinity of TFs that function as replication activators or repressors.  
6 Heterozygous profiles are not plotted. Center panels: ChIP-seq tracks. Lower panels: sequence logos of the  
7 motifs containing the rtQTL SNPs, motif names, and changes in binding affinity (calculated based on motif  
8 scores). Asterisk indicates that the motif was on the negative strand and the sequence shown is the reverse  
9 complement. Red arrows: locations of the rtQTL SNPs. For activators, the rtQTL allele associated with early  
10 replication encodes an intact binding motif, while the allele associated with late replication abolishes the motif.  
11 Repressors have the opposite pattern: the early allele abolishes the motif.

12



1

2 Figure S5. Enrichment of TFs at hESC (A) and iPSC (B) rtQTLs. rtQTLs are enriched at binding sites of central  
3 pluripotency factors (red) and chromatin remodelers (blue). NS: not significant at 10% FDR. Only TFs overlapping  
4 with at least 15 rtQTLs are plotted.

5

## 6 Discussion

7 The spatiotemporal regulation of DNA replication, and its dependence on regulatory DNA  
8 sequences, are poorly understood. Here, we leveraged population-scale replication timing and  
9 genetic polymorphism data to perform the equivalent of millions of surgical genetic  
10 interrogations of replication timing determinants. This approach enabled us to identify an  
11 unprecedented number of precise sequence determinants of replication timing.

12 Studying chromatin structure at rtQTL sites revealed a histone code that accurately predicts  
13 replication initiation across cell types. This code represents non-canonical functions of histone  
14 H3 lysine methylations that form a previously undescribed bivalent chromatin state<sup>55</sup> present at  
15 specific sites throughout the genome. Prior biochemical evidence supports an involvement of  
16 these histone marks in DNA replication initiation<sup>14,27-32,34-37</sup>. We propose that this histone code  
17 promotes local replication activity, although we do not necessarily imply that it marks the  
18 locations of replication origins *per se*.

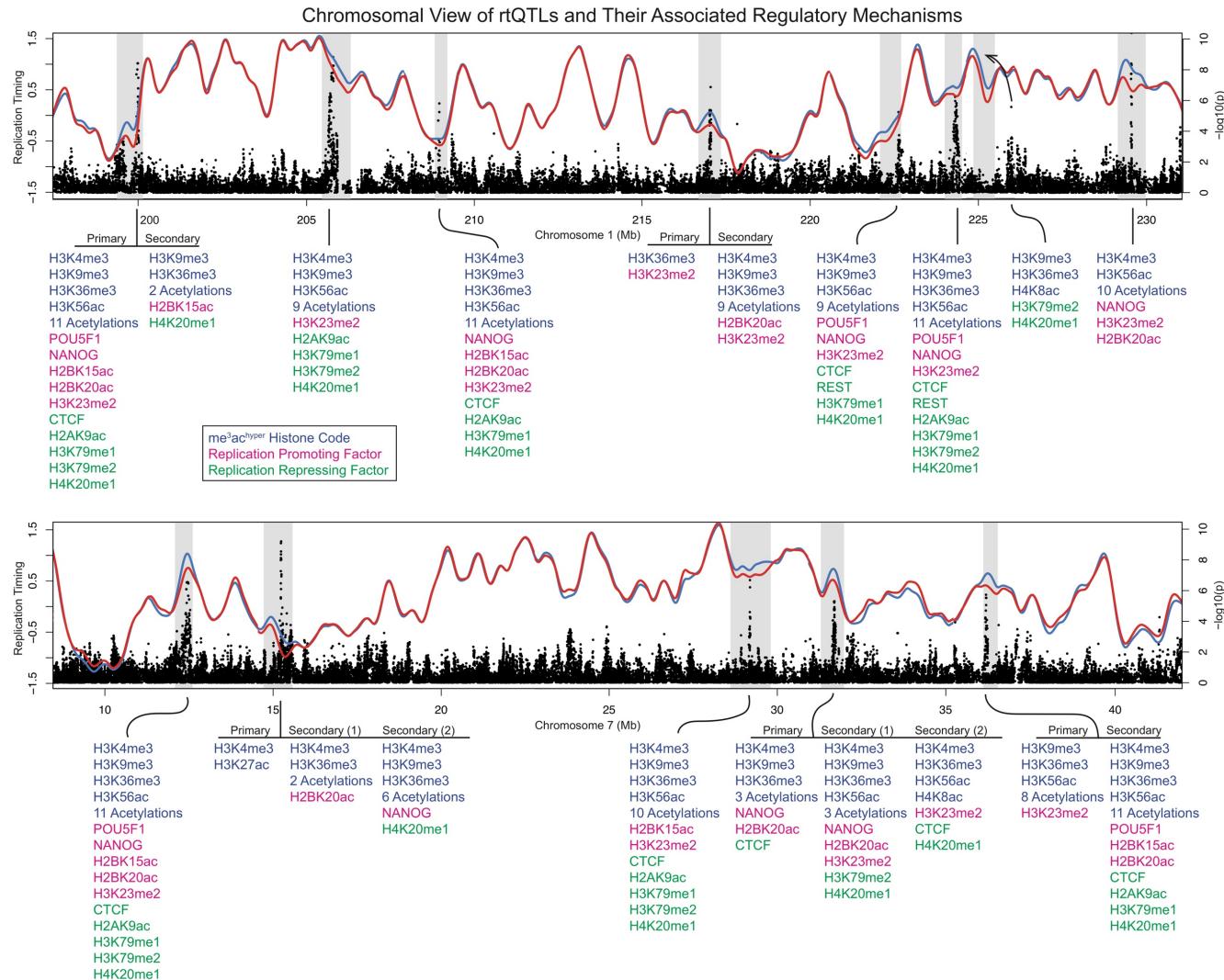
19 rtQTLs further associated with inter-individual variation in histone marks and TF binding  
20 affinity. In many cases, several *cis*-acting sequences affected a region's replication timing both  
21 proximally as well as distally. Altogether, we were able to assign at least one molecular  
22 determinant to 98.8% of rtQTLs, while two or more determinants were implicated in 93.9% of  
23 rtQTLs (Fig. S6). Replication timing determinants acted additively among nearby sequences,  
24 synergistically between histone modifications, and combinatorially across transcription factors.

1 This system generates a continuum of replication activities: some epigenetic marks may  
2 contribute only modestly to replication activity, or even suppress it, yet can interact with other  
3 factors to ultimately promote robust early replication. Taken together, this study systematically  
4 reveals a complex, combinatorial landscape of genetic regulation of human DNA replication  
5 timing.

6 A recent study using CRISPR/Cas9-mediated deletions in mouse ESCs identified several  
7 interacting sequence elements responsible for early replication (“early replicating control  
8 elements”<sup>12</sup>). Consistent with our results, the identified elements bound P300 and pluripotency-  
9 related TFs. However, the specific features identified with deletions represented the properties  
10 of only 1.5% of rtQTLs. Instead, rtQTLs associated with replication throughout S phase (not  
11 just with early replication); some interacted with others while many did not; and there was no  
12 single DNA-binding factor that was always bound to them. rtQTL mapping reveals a much  
13 more complex picture of replication timing regulation than previous approaches were powered  
14 to uncover. Replication timing regulation emerges as a quantitative trait, requiring a  
15 quantitative genetics approach to elucidate its complex sequence underpinnings. rtQTL  
16 mapping in larger sample sets and additional cell types will further refine the details of  
17 replication timing regulation and reveal additional *cis*-acting sequences and their mode of  
18 action. In addition, rtQTL mapping refines the relationship between DNA replication timing and  
19 gene expression and reveals influences of replication timing on personalized mutational  
20 landscapes and on human phenotypes including disease risk<sup>4,22</sup> (our unpublished results).

21 Our findings draw corollaries between replication timing regulation and classical concepts of  
22 gene expression regulation: promoter/enhancer logic, activators and repressors, and a histone  
23 code. Thus, replication and transcription regulation appear to be based at least in part on  
24 similar principles and building blocks. Replication timing is robustly encoded in DNA, yet  
25 multiple DNA sequences dictate DNA replication combinatorially via chromatin effectors. The  
26 replication timing program of the human genome emerges as being sequence-dependent,  
27 without being sequence-specific.

28



1

2 Figure S6. rtQTLs Regulate Replication Timing via Numerous Activating and Repressing Effectors.

3 Different combinations of TFs and histone marks exert positive and negative effects on subsets of replication  
4 initiation sites. Both examples show 10 ESC rtQTLs spanning a ~30-Mb region (on chromosomes 1 and 7). The  
5 blue and red lines are mean replication profiles of individuals carrying the early- and late-replicating genotypes,  
6 respectively. The rtQTL at 225 Mb of chromosome 1 exerts a long-range effect (arrow). Histone marks and TFs  
7 overlapping rtQTL genetic variants are shown below. They include positive (magenta) and negative (green)  
8 determinants of replication timing (Fig. 4 and 5), and instances of the replication initiation histone code (blue, Fig.  
9 3).

1 **Data Availability Statement**

2 Data of hESC and iPSC lines sequenced in this study were deposited in dbGaP (accession  
3 number: phs001957).

4

5 **Code Availability Statement**

6 Computer codes used in this study are available from the corresponding author upon request.

7

## 1 References and Notes:

2 1 Prioleau, M.-N. & MacAlpine, D. M. DNA replication origins—where do we begin? *Genes Dev.*  
3 **30**, 1683-1697, doi:10.1101/gad.285114 (2016).

4 2 Hamlin, J. L., Mesner, L. D. & Dijkwel, P. A. A winding road to origin discovery. *Chromosome Res.*  
5 **18**, 45-61, doi:10.1007/s10577-009-9089-z (2010).

6 3 Aladjem, M. I. & Redon, C. E. Order from clutter: selective interactions at mammalian  
7 replication origins. *Nat. Rev. Genet.* **18**, 101, doi:10.1038/nrg.2016.141 (2016).

8 4 Koren, A. *et al.* Genetic variation in human DNA replication timing. *Cell* **159**, 1015-1026,  
9 doi:10.1016/j.cell.2014.10.025 (2014).

10 5 Forrester, W. C. *et al.* A deletion of the human beta-globin locus activation region causes a  
11 major alteration in chromatin structure and replication across the entire beta-globin locus.  
12 *Genes Dev.* **4**, 1637-1649 (1990).

13 6 Liu, G., Malott, M. & Leffak, M. Multiple Functional Elements Comprise a Mammalian  
14 Chromosomal Replicator. *Mol. Cell. Biol.* **23**, 1832, doi:10.1128/MCB.23.5.1832-1842.2003  
15 (2003).

16 7 Aladjem, M. I., Rodewald, L. W., Kolman, J. L. & Wahl, G. M. Genetic Dissection of a Mammalian  
17 Replicator in the Human  $\beta$ -Globin Locus. *Science* **281**, 1005, doi:10.1126/science.281.5379.1005  
18 (1998).

19 8 Altman, A. L. & Fanning, E. The Chinese Hamster Dihydrofolate Reductase Replication Origin  
20 Beta Is Active at Multiple Ectopic Chromosomal Locations and Requires Specific DNA Sequence  
21 Elements for Activity. *Mol. Cell. Biol.* **21**, 1098, doi:10.1128/MCB.21.4.1098-1110.2001 (2001).

22 9 Aladjem, M. I. *et al.* Participation of the Human  $\beta$ -Globin Locus Control Region in Initiation of  
23 DNA Replication. *Science* **270**, 815, doi:10.1126/science.270.5237.815 (1995).

24 10 Kalejta, R. F. *et al.* Distal Sequences, but Not ori- $\beta$ /OBR-1, Are Essential for Initiation of DNA  
25 Replication in the Chinese Hamster DHFR Origin. *Mol. Cell* **2**, 797-806,  
26 doi:[https://doi.org/10.1016/S1097-2765\(00\)80294-4](https://doi.org/10.1016/S1097-2765(00)80294-4) (1998).

27 11 Wang, L. *et al.* The Human  $\beta$ -Globin Replication Initiation Region Consists of Two Modular  
28 Independent Replicators. *Mol. Cell. Biol.* **24**, 3373, doi:10.1128/MCB.24.8.3373-3386.2004  
29 (2004).

30 12 Sima, J. *et al.* Identifying cis Elements for Spatiotemporal Control of Mammalian DNA  
31 Replication. *Cell* **176**, 1-15, doi:10.1016/j.cell.2018.11.036 (2019).

32 13 Conner, A. L. & Aladjem, M. I. The chromatin backdrop of DNA replication: Lessons from  
33 genetics and genome-scale analyses. *Biochim. Biophys. Acta* **1819**, 794-801,  
34 doi:<https://doi.org/10.1016/j.bbagen.2012.01.017> (2012).

35 14 Sansam, C. G. *et al.* A mechanism for epigenetic control of DNA replication. *Genes Dev.* **32**, 224-  
36 229 (2018).

37 15 Goren, A., Tabib, A., Hecht, M. & Cedar, H. DNA replication timing of the human beta-globin  
38 domain is controlled by histone modification at the origin. *Genes Dev.* **22**, 1319-1324,  
39 doi:10.1101/gad.468308 (2008).

40 16 Aggarwal, B. D. & Calvi, B. R. Chromatin regulates origin activity in Drosophila follicle cells.  
41 *Nature* **430**, 372-376, doi:10.1038/nature02694 (2004).

1 17 Vogelauer, M., Rubbi, L., Lucas, I., Brewer, B. J. & Grunstein, M. Histone Acetylation Regulates  
2 the Time of Replication Origin Firing. *Mol. Cell* **10**, 1223-1233,  
3 doi:[https://doi.org/10.1016/S1097-2765\(02\)00702-5](https://doi.org/10.1016/S1097-2765(02)00702-5) (2002).

4 18 Rizzardi, L. F., Dorn, E. S., Strahl, B. D. & Cook, J. G. DNA Replication Origin Function Is Promoted  
5 by H3K4 Di-methylation in *Saccharomyces cerevisiae*. *Genetics* **192**, 371,  
6 doi:10.1534/genetics.112.142349 (2012).

7 19 Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human iPSCs.  
8 *Nature* **546**, 370-375, doi:10.1038/nature22403 (2017).

9 20 Zhao, P. A., Sasaki, T. & Gilbert, D. M. High-resolution Repli-Seq defines the temporal  
10 choreography of initiation, elongation and termination of replication in mammalian cells.  
11 *Genome Biol.* **21**, 76, doi:10.1186/s13059-020-01983-8 (2020).

12 21 Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci  
13 with multiple signals of association. *Genetics* **198**, 497-508, doi:10.1534/genetics.114.167908  
14 (2014).

15 22 Gerhardt, J. *et al.* Cis-acting DNA sequence at a replication origin promotes repeat expansion to  
16 fragile X full mutation. *J. Cell Biol.* **206**, 599-607, doi:10.1083/jcb.201404157 (2014).

17 23 Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin  
18 interactions and distinguish closely related cell types. *Genome Res.* **20**, 761-770,  
19 doi:10.1101/gr.099655.109 (2010).

20 24 GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-  
21 213, doi:10.1038/nature24277 (2017).

22 25 Long, H. *et al.* H2A.Z facilitates licensing and activation of early replication origins. *Nature* **577**,  
23 576-581, doi:10.1038/s41586-019-1877-9 (2020).

24 26 Vakoc, C. R., Mandat, S. A., Olenchock, B. A. & Blobel, G. A. Histone H3 Lysine 9 Methylation and  
25 HP1 $\alpha$ ; Are Associated with Transcription Elongation through Mammalian Chromatin. *Mol.*  
26 *Cell* **19**, 381-391, doi:10.1016/j.molcel.2005.06.011 (2005).

27 27 Black, Joshua C. *et al.* KDM4A Lysine Demethylase Induces Site-Specific Copy Gain and  
28 Rereplication of Regions Amplified in Tumors. *Cell* **154**, 541-555,  
29 doi:<https://doi.org/10.1016/j.cell.2013.06.051> (2013).

30 28 Mishra, S. *et al.* Cross-talk between Lysine-Modifying Enzymes Controls Site-Specific DNA  
31 Amplifications. *Cell* **174**, 803-817, doi:<https://doi.org/10.1016/j.cell.2018.06.018> (2018).

32 29 Liang, Z., Diamond, M., Smith, J. A., Schnell, M. & Daniel, R. Proliferating cell nuclear antigen is  
33 required for loading of the SMCX/KMD5C histone demethylase onto chromatin. *Epigenetics &*  
34 *Chromatin* **4**, 18, doi:10.1186/1756-8935-4-18 (2011).

35 30 Rondinelli, B. *et al.* H3K4me3 demethylation by the histone demethylase KDM5C/JARID1C  
36 promotes DNA replication origin firing. *Nucleic Acids Res.* **43**, 2560-2574,  
37 doi:10.1093/nar/gkv090 (2015).

38 31 Wu, R., Wang, Z., Gan, H., Zhang, H. & Zhang, Z. H3K9me3 demethylase Kdm4d facilitates the  
39 formation of pre-initiative complex and regulates DNA replication. *Nucleic Acids Res.* **45**, 169-  
40 180, doi:10.1093/nar/gkw848 (2016).

41 32 Vermeulen, M. *et al.* Quantitative Interaction Proteomics and Genome-wide Profiling of  
42 Epigenetic Histone Marks and Their Readers. *Cell* **142**, 967-980,  
43 doi:<https://doi.org/10.1016/j.cell.2010.08.020> (2010).

1 33 Casas-Delucchi, C. S. *et al.* Histone hypoacetylation is required to maintain late replication  
2 timing of constitutive heterochromatin. *Nucleic Acids Res.* **40**, 159-169, doi:10.1093/nar/gkr723  
3 (2012).

4 34 Iizuka, M. *et al.* Histone acetyltransferase Hbo1: catalytic activity, cellular abundance, and links  
5 to primary cancers. *Gene* **436**, 108-114, doi:10.1016/j.gene.2009.01.020 (2009).

6 35 Yan, K. *et al.* The Chromatin Regulator BRPF3 Preferentially Activates the HBO1  
7 Acetyltransferase but Is Dispensable for Mouse Development and Survival. *J. Biol. Chem.* **291**,  
8 2647-2663, doi:10.1074/jbc.M115.703041 (2016).

9 36 Miotto, B. & Struhl, K. HBO1 Histone Acetylase Activity Is Essential for DNA Replication  
10 Licensing and Inhibited by Geminin. *Mol. Cell* **37**, 57-66,  
11 doi:<https://doi.org/10.1016/j.molcel.2009.12.012> (2010).

12 37 Feng, Y. *et al.* BRPF3-HBO1 regulates replication origin activation and histone H3K14  
13 acetylation. *EMBO J.* **35**, 176-192, doi:10.15252/embj.201591293 (2016).

14 38 Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more  
15 genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic  
16 Acids Res.* **47**, D419–D426, doi:10.1093/nar/gky1038 (2019).

17 39 Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human  
18 epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

19 40 Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of  
20 diverse human tissues. *Nat. Biotechnol.* **33**, 364-376, doi:10.1038/nbt.3157 (2015).

21 41 Gindin, Y., Valenzuela, M. S., Aladjem, M. I., Meltzer, P. S. & Bilke, S. A chromatin structure-  
22 based model accurately predicts DNA replication timing in human cells. *Mol. Syst. Biol.* **10**, 722-  
23 722, doi:10.1002/msb.134859 (2014).

24 42 Black, J. C. *et al.* Conserved Antagonism between JMJD2A/KDM4A and HP1y during Cell Cycle  
25 Progression. *Mol. Cell* **40**, 736-748, doi:<https://doi.org/10.1016/j.molcel.2010.11.008> (2010).

26 43 The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human  
27 genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

28 44 Kasowski, M. *et al.* Extensive Variation in Chromatin States Across Humans. *Science* **342**, 750-  
29 752, doi:10.1126/science.1242510 %J Science (2013).

30 45 Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by  
31 genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27**,  
32 157-164, doi:10.1101/gr.210500.116 (2017).

33 46 Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ES cell differentiation.  
34 *Nature* **518**, 344-349, doi:10.1038/nature14233 (2015).

35 47 Khare, S. P. *et al.* Histome--a relational knowledgebase of human histone proteins and histone  
36 modifying enzymes. *Nucleic Acids Res.* **40**, D337-D342, doi:10.1093/nar/gkr1125 (2012).

37 48 Zhao, J., Li, X., Guo, M., Yu, J. & Yan, C. The common stress responsive transcription factor ATF3  
38 binds genomic sites enriched with p300 and H3K27ac for transcriptional regulation. *BMC  
39 Genomics* **17**, 335, doi:10.1186/s12864-016-2664-8 (2016).

40 49 Hwang, J.-Y. & Zukin, R. S. REST, a master transcriptional regulator in neurodegenerative  
41 disease. *Curr. Opin. Neurobiol.* **48**, 193-200, doi:<https://doi.org/10.1016/j.conb.2017.12.008>  
42 (2018).

43 50 Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at  
44 gene promoters. *Nature Communications* **6**, 6186, doi:10.1038/ncomms7186 (2015).

1 51 Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome  
2 Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922,  
3 doi:10.1016/j.cell.2017.05.004 (2017).

4 52 Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin  
5 removal. *Nature* **551**, 51-56, doi:10.1038/nature24281 (2017).

6 53 Atchison, M. L. Function of YY1 in Long-Distance DNA Interactions. *Front. Immunol.* **5**, 45-45,  
7 doi:10.3389/fimmu.2014.00045 (2014).

8 54 Schwalie, P. C. *et al.* Co-binding by YY1 identifies the transcriptionally active, highly conserved  
9 set of CTCF-bound regions in primate genomes. *Genome Biol.* **14**, R148-R148, doi:10.1186/gb-  
10 2013-14-12-r148 (2013).

11 55 Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in  
12 Embryonic Stem Cells. *Cell* **125**, 315-326, doi:<https://doi.org/10.1016/j.cell.2006.02.041> (2006).

13 56 International HapMap Consortium *et al.* Integrating common and rare genetic variation in  
14 diverse human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).

15 57 Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human  
16 genome. *Genome Res.* **16**, 1182-1190, doi:10.1101/gr.4565806 (2006).

17 58 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature*  
18 **526**, 75, doi:10.1038/nature15394  
19 <https://www.nature.com/articles/nature15394#supplementary-information> (2015).

20 59 Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome  
21 structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269-276,  
22 doi:10.1038/ng.768 (2011).

23 60 Dimitromanolakis, A., Paterson, A. D. & Sun, L. Fast and Accurate Shared Segment Detection  
24 and Relatedness Estimation in Un-phased Genetic Data via TRUFFLE. *The American Journal of  
25 Human Genetics* **105**, 78-88, doi:<https://doi.org/10.1016/j.ajhg.2019.05.007> (2019).

26 61 Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component  
27 analysis of SNP data. *Bioinformatics* **28**, 3326-3328, doi:10.1093/bioinformatics/bts606 (2012).

28 62 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method  
29 for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529-  
30 e1000529, doi:10.1371/journal.pgen.1000529 (2009).

31 63 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad.  
32 Sci. USA* **100**, 9440-9445, doi:10.1073/pnas.1530509100 (2003).

33 64 Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL  
34 mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-1485,  
35 doi:10.1093/bioinformatics/btv722 (2016).

36 65 Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55**, 997-1004,  
37 doi:10.1111/j.0006-341X.1999.00997.x (1999).

38 66 Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with  
39 Unlimited Zoom. *Cell Syst.* **3**, 99-101, doi:10.1016/j.cels.2015.07.012 (2016).

40 67 Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human  
41 replication timing. *Proc. Natl. Acad. Sci. USA* **107**, 139-144, doi:10.1073/pnas.0912402107  
42 (2010).

1 68 Yaffe, E. *et al.* Comparative analysis of DNA replication timing reveals conserved large-scale  
2 chromosomal architecture. *PLoS Genet.* **6**, e1001011, doi:10.1371/journal.pgen.1001011  
3 (2010).

4 69 Bock, C. *et al.* Reference Maps of human ES and iPS cell variation enable high-throughput  
5 characterization of pluripotent cell lines. *Cell* **144**, 439-452, doi:10.1016/j.cell.2010.12.032  
6 (2011).

7 70 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and  
8 regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**,  
9 D930-D934, doi:10.1093/nar/gkr917 %J Nucleic Acids Research (2012).

10 71 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator.  
11 *Genome Res.* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).

12

1

2

3

4

## Supplementary Materials for The Genetic Architecture of DNA Replication Timing in Human Pluripotent Stem Cells

5

Qiliang Ding<sup>1</sup>, Matthew M. Edwards<sup>1</sup>, Michelle L. Hulke<sup>1</sup>, Alexa N. Bracci<sup>1</sup>, Ya Hu<sup>1,2</sup>, Yao Tong<sup>1</sup>, Xiang Zhu<sup>3</sup>, Joyce Hsiao<sup>4</sup>, Christine J. Charvet<sup>1</sup>, Sulagna Ghosh<sup>5,6,7</sup>, Robert E. Handsaker<sup>5,6</sup>, Kevin Eggan<sup>5,7,8</sup>, Florian T. Merkle<sup>9</sup>, Jeannine Gerhardt<sup>10,11</sup>, Dieter Egli<sup>12</sup>, Andrew G. Clark<sup>1</sup>, Amnon Koren<sup>1\*</sup>.

6

Correspondence to: [koren@cornell.edu](mailto:koren@cornell.edu)

7

8

9

10

11

12

13

14

15

16

### This PDF file includes:

17

Materials and Methods  
Figs. S1 to S6

18

19

20

1 **Materials and Methods**

2 Whole-genome sequence data

3 Whole-genome sequence data and genotype calls for 121 hESC lines were obtained from  
4 Merkle et al. (submitted). We denote these as “Merkle\_batch1”. Nine additional hESCs were  
5 sequenced in another batch from Merkle et al. (submitted), denoted as “Merkle\_batch2”. We  
6 further used whole-genome sequence data from 326 iPSC lines from the HipSci Project<sup>19</sup>  
7 (ENA accession number: PRJEB15299), denoted as “HipSci”. We sequenced an additional 15  
8 hESCs and 17 iPSCs (dbGaP accession number: phs001957), denoted as “in\_house\_hESC”  
9 and “in\_house\_iPSC”, respectively.

10 For the in-house datasets, DNA was extracted using the MasterPure Complete DNA and RNA  
11 Purification Kit (Lucigen). Sequencing libraries were prepared using the Illumina TruSeq PCR-  
12 free kit and sequenced on an Illumina HiSeq X Ten to ~16-fold coverage with 150×2 paired-  
13 end reads. Sequencing was performed at GeneWiz (South Plainfield, NJ). No approval was  
14 needed for sequencing. Reads were aligned to GRCh37 using BWA, and genetic variants  
15 were called following the GATK Best Practices. Variants were filtered using GATK’s variant  
16 quality score recalibration, such that SNPs had a 99.9% sensitivity to true variants (HapMap  
17 3.3 and Omni 2.5M)<sup>56</sup> and a 99.0% sensitivity to true indels (Mills / 1000 Genomes indels)<sup>57,58</sup>.

18

19 Inference of DNA replication timing

20 DNA replication timing was inferred by analyzing sequence read depth (corrected for GC  
21 content bias) in non-overlapping windows of 10 kb of uniquely alignable sequence using  
22 GenomeSTRiP<sup>4,59</sup>. Among the 121 hESC lines from Merkle\_batch1, five did not optimally  
23 thrive in culture, resulting in read depth profiles with low correlations to other samples; these  
24 cell lines were excluded from further analysis. We excluded 26 of the 326 iPSC lines from the  
25 HipSci dataset for the same reasons. As described below, further filtering were performed for  
26 Merkle\_batch1 and HipSci datasets. Replication timing inference for the in-house datasets is  
27 described separately (see the “validation of rtQTLs” section below).

28 Replication timing windows falling under any of the following categories were filtered out in all  
29 cell lines: (1) spanning GRCh37 gaps; (2) overlapping structural variants (SV) with  $\geq 1\%$  MAF  
30 in the 1000 Genomes European individuals; (3) overlapping short CNVs (median size: 3.51 kb)  
31 identified directly in the analyzed cell lines (applicable for Merkle\_batch1 only, Merkle et al.,  
32 submitted); (4) having a median window copy number (across samples) at least 0.4 copies  
33 away from the median (across windows) of median copy number of all autosomal windows  
34 (across samples); and (5) the 25% / 75% percentile (across samples) copy number of the  
35 window was at least 0.4 copies away from the median (across windows) of 25% / 75%  
36 percentile copy number of all autosomal windows (across samples). Criteria 2 and 3 remove  
37 windows that are possibly affected by SVs or CNVs, while criteria 4 and 5 remove windows  
38 that had outlying copy number values across a significant portion of samples. Specific

1 parameters for criteria 4 and 5 (as well as the filtering steps described below) were chosen  
2 after evaluation of data values that were consistent across chromosomal replication profiles  
3 and across samples. Altogether, 28,769 data windows (11.0% of all windows) were removed,  
4 leaving 232,027 windows after filtering for Merkle\_batch1. For the HipSci dataset, 239,516  
5 windows remained.

6 Replication timing windows falling under any of the following categories were filtered out in  
7 individual cell lines: (1) at least 0.6 copies away from the median (across windows) of median  
8 copy number of all windows (across samples); (2) at least 0.25 copies away from the median  
9 copy number of that replication timing window; (3) in a large CNV (median size: 3.02Mb)  
10 identified in that individual (applicable for Merkle\_batch1 only); and (4) in a region suspected to  
11 be a large subclonal CNV (sub-integer change in copy number over a large region, usually an  
12 entire chromosome or a chromosome arm). These criteria were implemented to further remove  
13 outlier data points. Data after the above filtering steps is referred to as “filtered raw data”.

14 Processing of the X chromosome data was performed separately for males and females. For  
15 males, because they only carry one X chromosome, all the thresholds above were divided by  
16 two.

17 The filtered raw data was further normalized to Z-score (i.e., autosomal mean of zero and  
18 standard deviation of one) by subtracting the mean then dividing by the standard deviation of  
19 all data points, and smoothed using a penalized smoothing spline using the R *pspline*  
20 package with smoothing parameter  $10^{-16}$ . For each chromosome, we smoothed across gaps  
21 only if the gaps were shorter than 300 kb. Continuous genomic segments (between gaps) that  
22 were smaller than 300 kb were removed from further analysis. Data after the above  
23 normalization and smoothing is referred to as “smoothed data” (Fig. 1A) and was used in  
24 further analyses. The total length of replication timing windows in the smoothed data was  
25 2,330.66 Mb for autosomes (referred to as the “analyzable genome”), 121.15 Mb for the X  
26 chromosome in females, and 121.19 Mb for the X chromosome in males. For analyses  
27 involving the analyzable genome, only autosomal rtQTLs were counted.

28 For correlation calculations involving sib pairs vs. non-sib pairs (Merkle\_batch1) and cell lines  
29 derived from the same donor vs. different donors (HipSci), we used replication timing data from  
30 chromosomes 1 to 5. The Wilcoxon rank-sum test was used to assess significance. For the  
31 analysis regarding IBD segments in sib pairs, we first inferred pairwise IBD using TRUFFLE<sup>60</sup>,  
32 then binned the IBD segments into 2.5 Mb regions. The purpose was to minimize bias in  
33 correlation estimation because of variable IBD segment sizes. We calculated pairwise  
34 correlation in these regions, and assigned the estimate to one of three groups (IBD 0/1/2).  
35 ANOVA was used to assess significance of difference in average correlation among IBD 0/1/2  
36 groups. For all box plots in this study, the center line represents median, box limits represent  
37 the first and third quartile, and the whiskers represent the maximum and minimum. Outliers as  
38 determined by the R *boxplot* function were not plotted.

39

1 Identification of replication timing peaks

2 We identified peaks in the Merkle\_batch1 dataset. For each sample, peaks were identified in  
3 the replication timing profile as local maxima. Peaks across all samples were then clustered  
4 using agglomerative hierarchical clustering in MATLAB (functions `linkage` and `cluster`)  
5 with a distance threshold of 200 kb, which yields a list of peak clusters, each containing one or  
6 more peak locations. When a cluster contained multiple peaks from the same sample, the  
7 peak closest to the cluster center was retained and all other peaks from the sample were  
8 dropped. For each peak cluster, the boundary was defined as the full range of peak locations  
9 in this cluster. We only used peak clusters that contained peaks from more than 10% of the  
10 samples.

11

12 Identification of replication timing variants

13 We searched for replication timing variants using the Merkle\_batch1 and HipSci datasets. We  
14 expect genomic regions with strong replication timing variation to have greater standard  
15 deviation (SD) across samples, compared to average genomic regions. We calculated SD  
16 across samples for each replication timing window across the genome. Since local maxima will  
17 indicate the highest regional SD values, we called peaks in SD across the genome. To prevent  
18 calling peaks at single outlier data points, we first smoothed the SD curve. Then, we removed  
19 peaks that were below a SD threshold equal to the mean of the genome-wide SD distribution.  
20 We performed pairwise *t*-tests on pairs of samples for replication timing difference on 500 kb  
21 windows centered at the remaining SD peaks. For example, in the Merkle\_batch1 dataset,  
22 2,154 windows were tested, of which 1,785 (82.7%) were significant at a Bonferroni-corrected  
23 significance threshold of  $p = 4 \times 10^{-9}$ . These significant windows were extended by testing  
24 adjacent 200 kb windows, sliding 100 kb at a time, until there were no longer any significantly  
25 different cell line pairs. After the extension step, SD peaks in close proximity occasionally  
26 resulted in overlapping replication timing variants. In these cases, if the correlation of  
27 replication timing across samples at the SD peaks was greater than 0.9, we merged these  
28 variants. Otherwise, adjacent variants were separated at the valley between SD peaks. Last,  
29 replication timing variants driven by less or equal than 1% of the samples were removed. This  
30 resulted in a total of 1,489 and 1,837 replication timing variants in the Merkle\_batch1 and  
31 HipSci dataset, respectively.

32

33 Data processing prior to rtQTL mapping

34 Sample selection

35 We performed principal component analysis (PCA) on the genotypes of the hESC lines, using  
36 the 1000 Genomes Phase 3 European, East Asian, and African samples as references. Eight  
37 samples appearing to have non-European ancestry (admixed or East Asian) were removed

1 from rtQTL mapping, leaving 108 individuals for further analysis. PCA was performed using the  
2 SNPRelate package in R<sup>61</sup>. We also performed PCA with the HipSci dataset, and confirmed  
3 that all samples were of European ancestry. A total of 192 unrelated samples in the HipSci  
4 dataset were used for rtQTL mapping. While we kept sib pairs in the ESC dataset, all rtQTLs in  
5 ESC were reproducible (at nominal  $p < 0.05$ ) when using only unrelated samples.

6

7 *Genotype imputation*

8 Imputation was performed with IMPUTE2<sup>62</sup> using the 1000 Genomes Project Phase 3  
9 reference panel and default parameters. Variants with minor allele frequency (MAF)  $\leq 1\%$  in  
10 Europeans or Americans were not used for imputation. Imputed variants with average  
11 genotype probability  $\geq 80\%$  were used in subsequent analyses.

12 Prior to rtQTL mapping, we filtered out variants that had MAF  $< 5\%$ , were non-biallelic, or that  
13 deviated from Hardy-Weinberg equilibrium ( $p < 1 \times 10^{-3}$ ). In addition, we required that variants  
14 should have all three genotypes (homozygous reference allele, homozygous alternative allele,  
15 and heterozygous genotypes) observed in the samples.

16

17 *PCA of replication timing data*

18 To account for potential batch effects and other unknown systematic biases in the replication  
19 timing data, we performed PCA using the filtered raw data with R function `prcomp`. Principal  
20 components (PCs) of the filtered raw data (“phenotype PCs”), along with the genotype PCs  
21 calculated above, were used as covariates in rtQTL mapping.

22

23 rtQTL mapping

24 *Selection of phenotype PCs in rtQTL mapping*

25 We followed the eQTL mapping framework used in the GTEx Project<sup>24</sup>  
26 (<https://gtexportal.org/home/documentationPage>) to map rtQTLs. We included the genotype  
27 (first three, similar to GTEx) and phenotype (first  $k$ ) PCs in rtQTL mapping to account for non-  
28 genetic confounding factors. To find the optimal  $k$ , we tested each integer from 1 to 40. We  
29 consider the optimal  $k$  as the one leading to the highest number of windows harboring rtQTLs  
30 identified in rtQTL mapping. In this analysis, permutation parameter “`permut 50 500`” was  
31 used in fastQTL. Window level  $p$ -values were calculated, and the R package `qvalue`<sup>63</sup> was  
32 used to identify windows harboring rtQTLs at 10% FDR. This resulted in 24 and 22 selected as  
33 the optimal  $k$  for ESC and iPSC rtQTL mapping analysis, respectively, which was used in all  
34 subsequent rtQTL mapping analyses.

35

36 *Cis-rtQTL mapping using fastQTL*

1 We implemented a two-step approach to map rtQTLs using fastQTL<sup>64</sup>. We generally restricted  
2 our analysis to *cis*-rtQTLs, defined as 1 Mb upstream or downstream of the center of each  
3 tested replication timing window. The first three genotype PCs and first 24 or 22 (for ESC and  
4 iPSC, respectively) phenotype PCs were included as covariates.

5 In the first step, we calculated a window-level *p*-value for each replication timing window using  
6 fastQTL, and then identified “significant windows”, i.e., windows with at least one significant  
7 rtQTL at 10% FDR, using the R package *qvalue*. This step is analogous to the identification  
8 of “eGenes” in eQTL mapping. For each window, fastQTL computes the lowest variant-level *p*-  
9 value and uses permutations to calculate the probability of observing a variant with equal or  
10 lower *p*-value under the scenario of no association, followed by beta approximation. Adaptive  
11 permutation parameter “*permut 1000 10000*” was used (similar to GTEx). We also repeated  
12 this step at 5% FDR.

13 In the second step, we identified genetic variants (referred to as SNPs for simplicity)  
14 associated with the “significant windows” identified in step 1, at 10% FDR. Here, we used a  
15 permutation-based strategy to determine the significance threshold for each tested window. By  
16 definition, FDR is the ratio of false positives (FP) to the sum of FP and true positives (TP). At a  
17 given *p*-value threshold  $p_t$ , variants passing  $p_t$  are composed of both TP and FP. However, if  
18 we permute the phenotype, all variants with *p*-values lower than  $p_t$  are FP. Therefore, for a  
19 given window, FDR for a given  $p_t$  could be estimated as the mean number of variants passing  
20  $p_t$  in permutations (i.e., all FP) divided by the number of variants passing  $p_t$  in the true  
21 association test (FP+TP). We then consider the maximum  $p_t$  with  $\text{FDR} \leq 10\%$  as the  
22 significance threshold of the window. The mean number of variants passing  $p_t$  in permutations  
23 was computed based on 500 permutations.

24

### 25 *Evaluation of inflation of rtQTL mapping*

26 To ensure that the computed variant-level *p*-values were not inflated, we calculated inflation  
27 factor with the Genomic Control method<sup>65</sup>. We selected 200 windows (100 selected from  
28 windows carrying putative rtQTLs, and the other 100 randomly selected from the rest of the  
29 genome) and computed their association with genome-wide variants. We obtained variant-  
30 level statistics (which follows  $\chi^2_1$  distribution under the null hypothesis) and computed the ratio  
31 of their median to the median of  $\chi^2_1$  (0.456) as the genomic inflation factor. We calculated a  
32 genomic inflation factor ( $\lambda$ ) as 1.03 and 1.00 for the ESC and iPSC dataset, respectively, thus  
33 the nominal *p*-values were not inflated; this was also supported by quantile-quantile plots.

34

### 35 *Identification of rtQTLs*

36 The following procedure was used to identify discrete rtQTLs, i.e., independent (not in LD)  
37 association signals, based on the significant SNPs mapped using the aforementioned two-step

1 approach. For clarity, we denote independent association signals as rtQTLs, each of which  
2 contains multiple SNPs that are part of the association signal.

3 For each window, we identified all SNPs (if any) that passed the significance threshold. We  
4 selected the SNP with the lowest  $p$ -value as the “tag” variant of an rtQTL and assigned SNPs  
5 in LD ( $r^2 \geq 0.2$ ) with the tag variant to the rtQTL. If there were any SNPs remaining that passed  
6 the significance threshold, we selected the SNP with the lowest  $p$ -value among the remaining  
7 SNPs as the tag variant of a new rtQTL and assigned all variants in LD with the new tag  
8 variant to the new rtQTL. This step was repeated until no variants passing the significance  
9 threshold were left. For the rtQTLs identified above, we kept only those with at least 10  
10 variants and for which the  $p$ -value of the tag variant was less than  $10^{-3}$ .

11 For all calculations involving LD, data from the 1000 Genomes Phase 3 Europeans was used  
12 whenever available. For SNPs not called in the 1000 Genomes dataset, the current dataset  
13 was used for LD calculation.

14 Since nearby replication timing windows are highly correlated, the same rtQTL can be detected  
15 across multiple windows. We consolidated association signals detected in different windows if  
16 they satisfy all of the following three criteria: (1) the tag variants are in LD ( $r^2 \geq 0.2$ ), (2) the  
17 replication timing windows are correlated ( $R^2 \geq 0.1$ ), and (3) the distance between the windows  
18 is less than 2Mb.

19 In addition to separating rtQTLs by LD, we performed conditional association for each  
20 identified rtQTL. We conditioned on the top variant of each rtQTL and examined whether any  
21 SNPs that belong to this rtQTL still have significant association  $p$ -value (at  $p = 0.05$  after  
22 Bonferroni correction). If so, this rtQTL was divided into multiple rtQTLs, each representing an  
23 independent association signal.

24

25 *Filtering of rtQTLs*

26 The putative rtQTLs identified were subjected to further filtering. First, we determined the  
27 boundaries of regions that significantly associated with each putative rtQTL. Starting at the  
28 window that most strongly associated with the tag variant (i.e., with the lowest  $p$ -value) of an  
29 rtQTL, we extended the region bi-directionally until the association was no longer significant  
30 ( $p > 0.05$ ). We refer to this region as the “associated region”.

31 Next, we filtered false positives suspected to be potentially caused by short CNVs. During data  
32 processing (described above), we removed windows in which copy number measurement are  
33 potentially influenced by CNVs. However, short CNVs, spanning only one or two windows, may  
34 not have been detected and filtered and could lead to false positive rtQTLs (if they are in LD  
35 with SNPs). This type of false positive was identified by utilizing the raw unsmoothed data as  
36 follows: if a putative rtQTL is a false positive caused by a CNV, it would be (1) only observed in  
37 a small number of unsmoothed raw windows (overlapping with the CNV), and (2) will be more

1 strongly associated with the raw data than with the smoothed data (in which the CNV will be  
2 smoothed within a broader region, thus decreasing association). Furthermore, it may have  
3 much stronger association with windows removed during replication timing data processing.  
4 We computed the association *p*-values of the tag variant of each rtQTL with the (1) smoothed  
5 data within the associated region, (2) filtered raw data within the associated region, and (3)  
6 data that were removed during data processing within 1 Mb upstream or downstream of the  
7 associated region (referred to as “removed data” below).  
8 Putative rtQTLs must satisfy all of the following criteria to be included in the final list of rtQTLs:  
9 (1) In the raw data, the tag variant must be associated ( $p < 0.05$ ) with at least five windows.  
10 (2) The minimum *p*-value of the raw data must be higher (i.e., less significant), or no more  
11 than one order of magnitude lower, than that of the smoothed data.  
12 (3) The minimum *p*-value of the removed data must be higher, or no more than one order of  
13 magnitude lower, than that of the raw data. This criterion is relaxed to two or four orders  
14 of magnitude for rtQTLs with top *p*-value  $\leq 5 \times 10^{-6}$  and  $\leq 5 \times 10^{-8}$ , respectively.  
15 (4) No more than two windows in the removed data have *p*-values lower than the minimum  
16 *p*-value for the raw data. This criterion is relaxed to three windows for rtQTLs with top *p*-  
17 value  $\leq 5 \times 10^{-8}$ .  
18 (5) The minimum *p*-value from the raw data must be less than 0.01.  
19 (6) The associated region must be larger than one replication timing window.  
20 In total, we identified 608 ESC rtQTLs, among which 603 were on autosomes and five were on  
21 the X chromosome in males. No rtQTLs were found on the X chromosome in females. This  
22 was not due to the reduced number of individuals tested, but likely resulted from the less  
23 structured replication timing profiles attributed to the female inactive X chromosomes: the  
24 similar-sized chromosome 7 had ten rtQTLs in the 50 male samples, not significantly different  
25 than the male X chromosome ( $p = 0.31$ , Fisher’s exact test), while there were fifteen rtQTLs on  
26 chromosome 7 in 66 female samples, significantly more than the none found on the female X  
27 chromosome ( $p = 7.41 \times 10^{-5}$ ). We identified 1,167 iPSC rtQTLs. The nominal *p*-value of rtQTLs  
28 ranged from  $1.02 \times 10^{-69}$  to  $9.63 \times 10^{-4}$  (106 and 218 rtQTLs [17.4% and 18.7%] had  $p \leq 5 \times 10^{-8}$  in  
29 the ESC and iPSC dataset, respectively). The early- and late-replicating alleles were equally  
30 likely to be the reference allele (binomial  $p = 0.55$ ), thus rtQTL mapping was not influenced by  
31 reference mappability bias.  
32  
33 *Prioritizing causal genetic variants*  
34 For each rtQTL, CAVIAR<sup>21</sup> was used to produce a shortlist of possible causal SNPs at 90%  
35 probability, from all SNPs in LD with the tag variant of the rtQTL ( $r^2 \geq 0.2$ ). The shortlisted  
36 SNPs were used in all enrichment analyses.  
37

1 *rtQTL classification*

2 We classified each rtQTL as affecting peak (initiation site), valley (terminus), or slope  
3 (transition region). For each rtQTL, we identified the replication timing loci that have large  
4 difference in replication timing (at least 90% of the maximum difference) between the early-  
5 replicating and late-replicating individuals (denoted as the “most variable replication timing  
6 loci”). We then calculated “fractional distance” of these loci along the peak-to-valley interval in  
7 which they reside. If a replication timing locus, with position  $a$ , resides in the interval between a  
8 peak (with position  $b$ ) and a valley (with position  $c$ ), its fractional distance was calculated as  $a$   
9 minus  $b$ , divided by  $c$  minus  $b$ . We considered an rtQTL as affecting an initiation site if the  
10 fractional distance of at least one of the most variable replication timing loci was less than 0.3.  
11 Conversely, we considered an rtQTL as affecting a valley if the fractional distances of all of the  
12 most variable replication timing loci were greater than 0.7. rtQTLs that did not fall into either of  
13 these two categories were categorized as affecting slopes.

14 We further classified rtQTLs that affect peaks based on whether the top rtQTL SNP was  
15 located proximal or distal to the peak. Specifically, we calculated fractional distance of the top  
16 rtQTL SNP for each rtQTL that affect peaks, using the same approach as described above.  
17 The top rtQTL SNP was considered proximal to the peak if its fractional distance was less than  
18 0.3 and was considered distal to the peak otherwise.

19

20 *Merging ESC and iPSC rtQTLs*

21 We combined ESC and iPSC rtQTLs for a number of analyses. To minimize double counting of  
22 rtQTLs discovered in both datasets, we generated a merged rtQTL list for these analyses. This  
23 list excluded iPSC rtQTLs that met the following criteria: (1) a genetic variant that belongs to  
24 the given iPSC rtQTL and has a  $p$ -value no more than two orders of magnitude higher than the  
25 top  $p$ -value of the iPSC rtQTL also belongs to a ESC rtQTL, and (2) the direction of effect of  
26 the given genetic variant is the same in the iPSC and ESC datasets. We merged the 608 ESC  
27 rtQTLs and 1,167 iPSC rtQTLs into a list of 1,617 combined rtQTLs.

28

29 Validation of rtQTLs

30 To validate the iPSC rtQTLs, we examined their reproducibility in the Merkle\_batch1 ESC  
31 dataset (108 European ancestry samples only). Validation was performed using fastQTL<sup>64</sup> by  
32 testing the association between the strongest rtQTL SNP and the replication timing locus  
33 closest to the locus with the strongest association in the discovery set (HipSci iPSCs). Three  
34 genotype PCs and 24 phenotype PCs were included as covariates. When the strongest rtQTL  
35 SNP was not available in the validation dataset (Merkle\_batch1 ESCs), an rtQTL SNP from the  
36 same rtQTL that has  $p$ -value less than two orders of magnitude higher than that of the  
37 strongest rtQTL SNP was used instead. We found that the  $-\log_{10}(p\text{-values})$  of rtQTLs are highly

1 correlated between the discovery and validation datasets ( $r = 0.75, p = 1.28 \times 10^{-176}$ ). We then  
2 repeated this analysis in the opposite direction (validate ESC rtQTLs using HipSci iPSCs) and  
3 obtained similar results ( $r = 0.76, p = 7.81 \times 10^{-113}$ ). These observations support that the rtQTLs  
4 identified in this study are highly reproducible.

5 We also used three additional datasets to validate ESC rtQTLs. The first dataset contains 9  
6 hESCs in Merkle\_batch2 and the 8 hESCs in Merkle\_batch1 that were excluded in rtQTL  
7 mapping due to ancestry. The second and third datasets are the in-house hESC and iPSC  
8 dataset, respectively.

9 For the first dataset, validation was performed in fastQTL. Validation using the second and  
10 third datasets were performed in MATLAB by calculating the Pearson correlation  $p$ -value  
11 between the strongest rtQTL genetic variant and the replication timing locus with the strongest  
12 association in the discovery set. We tested rtQTLs of which the top genetic variant was  
13 polymorphic and had all three genotypes in the validation dataset. rtQTLs were excluded if the  
14 alternative allele of the top genetic variant in the validation dataset was not consistent with that  
15 of in the discovery set. This left 427 regions that could be tested in the third dataset, and 396  
16 regions in the fourth dataset. Replication timing of these two datasets were inferred using  
17 GenomeSTRiP (as described above) in 2.5Kb windows of uniquely alignable sequence<sup>59</sup>. For  
18 each sample, windows with copy number  $>3$  or  $<1$  were removed. We used a segmentation  
19 algorithm (segment in MATLAB) to further remove outlier data points (segments with  
20 mean  $>2.45$  or  $<1.55$  were removed). The data was then smoothed using cubic smoothing  
21 spline with parameter  $10^{-17}$ .

22 We considered an rtQTL as “validated” if it was associated with replication timing with nominal  
23  $p < 0.05$  and had the same direction of effect in at least one of the validation datasets. The  
24 binomial test was used to assess significance of the number of validated rtQTLs, with binomial  
25 parameter calculated as  $1 - (1 - 0.05/2)^4 = 0.0963$  (i.e., the probability under random chance that  
26 an rtQTL will be validated in at least one dataset).

27

## 28 SMARD

29 SMARD analysis was carried out as previously described<sup>22</sup>. Briefly, cells were pulse labeled  
30 sequentially with 25  $\mu$ M IdU and CldU. The cells were then embedded in 1% InCert agarose  
31 and lysed. The remaining embedded genomic DNA was digested with restriction  
32 endonucleases. Pulsed field gel electrophoresis (PFGE) was used to separate DNA according  
33 to size. The segment containing the locus-of-interest was identified with Southern blot and the  
34 gel slice was excised. Agarose was then melted, and individual DNA strands were stretched  
35 on silanized glass slides. Immunostaining was employed to detect the halogenated nucleotides  
36 in the replicated DNA. Biotinylated FISH probes were used to identify DNA molecules  
37 containing the locus-of-interest.

38

1 Multi-rtQTLs

2 To identify multi-rtQTL regions, we considered separate rtQTLs to be associated with the same  
3 region if the replication timing loci most strongly associated with them were correlated ( $R^2 \geq$   
4 0.2) across individuals, were in physical proximity (< 2 Mb apart), and each provided additional  
5 explanatory power for replication timing. Secondary rtQTLs were either not in LD with the  
6 primary ones (130 and 265 multi-rtQTL regions in the ESC and iPSC dataset, respectively), or  
7 provided additional explanatory power despite being in LD (5 cases in ESC and 10 cases in  
8 iPSC).

9 Some analyses were performed with ESC and iPSC multi-rtQTL regions combined. To avoid  
10 double-counting in these analyses, we excluded iPSC multi-rtQTL regions that has at least one  
11 rtQTL that was also found in the ESC dataset. We combined 135 ESC and 275 iPSC multi-  
12 rtQTL regions into 318 multi-rtQTL regions.

13 We examined the possible interaction between primary and secondary rtQTLs in regions with  
14 two, three, and four rtQTLs. If there was no interaction, we expect that the replication timing in  
15 these regions will be positively linearly correlated with the dosage of early-replicating alleles.  
16 To enable pooling of multi-rtQTL regions for Fig. 2E and 2F, we normalized replication timing  
17 for the loci with strongest association with the primary rtQTL of each multi-rtQTL region to Z-  
18 score (by subtracting the mean and dividing by the standard deviation of replication timing of  
19 this locus among samples) and denoted them as relative replication timing. They were pooled  
20 and linear regression analysis was performed using the R `lm` function.

21 We used a likelihood-ratio test to assess whether the additive or synergistic models better  
22 explained replication timing at multi-rtQTL regions. We tested the null hypothesis by which  
23 replication timing is proportional to the number of early-replicating rtQTL alleles carried by an  
24 individual at a multi-rtQTL region (additive effect), against the alternative, by which replication  
25 timing is more extremely biased in individuals carrying multiple early (or late) rtQTL alleles  
26 (synergistic interaction). We used 58 regions that harbored two rtQTLs and had at least one  
27 individual with zero and one with four early-replicating alleles. We fitted two linear models, with  
28 the response variable being replication timing and explanatory variable being genotype  
29 dosage. In the null (additive) model, genotype dosage was between zero to four, matching the  
30 number of early-replicating alleles that individual carried. In the alternative (synergistic) model,  
31 genotype dosages of individuals carrying zero or four early-replicating alleles were estimated  
32 from actual data. We then compared  $-2 \times (\log \text{likelihood ratio})$  with the chi-squared distribution  
33 with two degrees of freedom to obtain a *p*-value.

34 We examined whether the primary and secondary rtQTLs in ESC were in close spatial  
35 proximity in nuclear space. We obtained Hi-C contact matrix of the H1 cell line from Juicebox<sup>66</sup>  
36 and computed contact score between each pair of primary and secondary rtQTLs. We  
37 compared the median of these scores with 100 permutations, in which the distances between  
38 primary and secondary rtQTLs were preserved but actual genomic locations were randomly

1 shifted between 1 and 2 Mb up- or downstream. *P*-value was computed using *Z* score, with  
2 mean and standard deviation estimated from the permutations.

3

4 Epigenetic enrichment analyses

5 *Data sources*

6 Chromatin state and histone mark data for eight human ESC lines (seven of which are  
7 included in our primary replication timing data) and five human iPSC lines were obtained from  
8 the Roadmap Epigenomics Project<sup>39</sup>. For the analyses of overall enrichment of epigenetic  
9 features at rtQTL locations, we combined (i.e., took the union of) histone peaks and chromatin  
10 state calls from the eight cell lines. For histone marks, observed data was used when  
11 available, and imputed data (from ChromImpute<sup>40</sup>) was used when observed data was not  
12 available. Imputed data were used for plotting of histone tracks. Binding site information for 51  
13 TFs was obtained from the ENCODE Project<sup>43</sup>. SOX2 binding site information was obtained<sup>46</sup>.  
14 TFs with binding sites that overlapped < 15 rtQTLs were excluded from this analysis.

15

16 *Enrichment calculations*

17 For each feature (chromatin state, histone marks, TF, etc.), we are interested in the number of  
18 rtQTLs that have at least one SNP overlapping with the feature, and whether this is more or  
19 less likely (i.e., enriched or depleted) than expected by chance. Statistical significance was  
20 assessed with one-tailed binomial test. The binomial parameter *p* was estimated from 100  
21 random permutations, from which we estimated the probability of random SNPs (matched with  
22 the rtQTLs, see details below) overlapping with the feature. Correction for multiple testing was  
23 applied when multiple features from the same category (e.g., histone marks) were tested.

24 For each rtQTL, we searched for random SNPs that match the characteristics of the tag variant  
25 of the rtQTL (denoted as “actual tag variant”) and used the matched variants (“matched tag  
26 variants”) to tag the random sets of SNPs used in permutations. We required that the matched  
27 tag variants must be at least 2 Mb away from the actual tag variant. The matched tag variants  
28 must also have satisfied all three following criteria: (1) have similar minor allele frequency (<  
29 5% difference), (2) have similar distances to the nearest replication initiation site and terminus  
30 (< 50 kb difference), and (3) have similar replication timing (< 0.5 units difference) with the  
31 actual tag variant. We require the matched tag variants to have the same number, or more,  
32 SNPs in LD ( $r^2 \geq 0.2$ ) than the actual tag variant.

33 In each permutation, and for each rtQTL, we constructed a set of random SNPs using SNPs in  
34 LD with a randomly selected matched tag variant. The number of variants in the set is the  
35 same number of variants included in the actual rtQTL. Eleven (1.82%) rtQTLs in hESC and 41  
36 (3.51%) rtQTLs in iPSC that had less than 200 matched tag variants genome-wide were  
37 excluded from the analysis.

1 Enrichment analyses were also repeated using epigenetic data from one hESC line only (as  
2 opposed to combining data from eight hESC lines), with consistent results.

3  
4 Using epigenetic features to predict replication initiation site locations  
5 *Identification of epigenetic feature combinations*

6 To identify combinations of chromatin marks enriched at rtQTLs, we used a stepwise, iterative  
7 approach. The hESC rtQTLs and epigenetic data were used. We considered 29 histone marks  
8 (Fig. S2C) and also included H2A.Z, DNase I hypersensitivity, and binding sites of 51 TFs and  
9 other DNA binding factors (referred to as TFs for simplicity).

10 First, we tested each individual epigenetic feature (histone mark or TF) to identify features that  
11 are enriched at rtQTL SNPs. Enrichment was examined using the same permutation-based  
12 approach described above. The only difference was that each rtQTL individual SNP was  
13 considered independently (as opposed to being considered together with other SNPs assigned  
14 to the same rtQTL), as our goal was to identify co-occurrence of epigenetic features at the  
15 same exact genomic locations. Statistical significance was assessed using Fisher's exact test.  
16 We corrected for multiple testing at 5% FDR using the Benjamini-Hochberg correction.

17 Next, for each enriched feature identified in the first step, we examined whether the pairwise  
18 combination of this feature and any of the other epigenetic features has stronger enrichment.  
19 Specifically, we restricted the enrichment analysis to the rtQTL SNPs that carry the enriched  
20 feature and tested whether the additional epigenetic feature is enriched in the set of restricted  
21 rtQTL SNPs. This step was repeated iteratively, each round restricting the analysis to the  
22 enriched combinations of epigenetic features identified in the previous round, until no further  
23 enrichment was found. In Fig. 3A, combinations containing TFs were not included for simplicity  
24 and since they were not carried through to the four- and five-mark combinations.

25 To identify “me<sup>3</sup>ac<sup>hyper</sup>” regions, we first identified regions that carry one of the 13 five-mark  
26 combinations and kept regions that overlap with peaks from at least 11 variable acetylation  
27 marks. We merged me<sup>3</sup>ac<sup>hyper</sup> regions that co-occurred within 10 kb. In Fig. S4A–C, the  
28 position of initiation sites found in >10% of the samples were determined based on local  
29 maxima in the averaged replication timing profile. When calculating distances (fractional and  
30 physical), distance was set to zero for me<sup>3</sup>ac<sup>hyper</sup> regions that overlap with an initiation site (i.e.,  
31 the interval between boundaries of the initiation sites). If a me<sup>3</sup>ac<sup>hyper</sup> region does not overlap  
32 with any initiation site, its physical distance was calculated as the distance to the nearest  
33 initiation site boundary. To explore the independence of the replication initiation histone code  
34 from gene expression, we divided me<sup>3</sup>ac<sup>hyper</sup> regions into two classes, based on whether there  
35 were TSS of expressed gene(s) (mean RPKM in ES cell lines > 0.5) within a given me<sup>3</sup>ac<sup>hyper</sup>  
36 region. RNA-seq data was obtained from the Roadmap Epigenomics Project<sup>39</sup>. We then  
37 compared the positive predictive value for these two classes of me<sup>3</sup>ac<sup>hyper</sup> regions.

1

2 *Receiver Operating Characteristic (ROC) curves*

3 To obtain ROC curves, we used various thresholds (see below) to predict whether a replication  
4 timing window corresponds to a replication initiation site. Specifically, we predict a window as  
5 being an initiation site if it was located within  $k$  kb of a  $\text{me}^3\text{ac}^{\text{hyper}}$  region (in hESCs and iPSCs)  
6 or a region that carries the H3K4me3-H3K9me3-H3K36me3 combination (in LCLs). We used  
7 various values for  $k$ , from 0 to 2,000. We then compared the prediction with actual data  
8 (whether the replication timing windows fell within the boundaries of the identified initiation  
9 sites) to calculate true and false positive rates. For permutations, we randomly shifted the  
10 locations of the  $\text{me}^3\text{ac}^{\text{hyper}}$  regions between 1 Mb and 2Mb and obtained ROC curves and  
11  $\text{AUC}_{\text{ROC}}$  based on these random intervals.

12

13 *Replication initiation site prediction in LCLs and iPSCs*

14 We assessed the generalizability of the replication initiation histone code in LCLs and iPSCs.  
15 LCL is a cell type that has a distinct epigenetic and replication landscape from hESC  
16 lines<sup>23,67,68</sup>, and iPSCs have similar but not identical to replication timing profiles to hESCs ( $r =$   
17 0.90). Replication timing profile for the GM12878 LCL and 192 unrelated iPSCs were inferred  
18 from whole-genome sequencing data<sup>19,45</sup>. For iPSCs, initiation site locations were identified  
19 based on the averaged iPSC replication timing profile. When calculating physical distance of  
20 predicted initiation sites to actual initiation sites, we defined initiation site boundaries as 100 kb  
21 upstream and downstream of the local maxima in the replication timing profiles. Data for  
22 H3K4me3, H3K9me3, and H3K36me3 for the GM12878 LCL was from the ENCODE Project<sup>43</sup>.  
23 Additional data of H3K4me3 and H3K36me3 for 18 LCLs was obtained from<sup>44</sup>, and merged  
24 with the ENCODE data. Histone mark data for five iPSCs was from the Roadmap Epigenomics  
25 Project<sup>39</sup>. If a histone code location was found in one cell type (either hESC or LCL), but no  
26 histone code location was found within 100 kb in the other cell type, we denoted this region as  
27 cell-type-specific. Otherwise, this region was denoted as “shared” between the two cell types.

28

29 Identification of features associated with replication timing

30 *Chromatin states and histone marks*

31 Replication timing data was available for seven of the eight hESC lines that were analyzed in  
32 the Roadmap Epigenomics Project. Using rtQTL and epigenetic data from these seven cell  
33 lines, we designed an analysis to identify chromatin states and histone marks associated with  
34 replication timing. The rationale is that epigenetic features promoting earlier replication would  
35 be more likely to be carried by early-replicating-associated rtQTL genotypes, and *vice versa* for  
36 late replication. We were only able to perform this analysis for hESC rtQTLs because we did  
37 not have replication timing, genotype, and epigenetic data for the same iPSC lines.

1 We aggregated information from all rtQTL SNPs, except those that are monomorphic in the  
2 seven cell lines. We assigned each cell line by genotype to one of three categories, i.e., early-  
3 replicating, heterozygous, and late-replicating, at each rtQTL SNP. For each epigenetic  
4 feature, we tested whether the cell lines carrying the early-replicating genotypes are more (or  
5 less) likely to harbor it than the cell lines carrying the late-replicating genotypes, using the two-  
6 tailed binomial test. The binomial parameter  $p$  was calculated as  $p_{late} \times (p_{perm\_early} / p_{perm\_late})$ ,  
7 where  $p_{late}$  is the proportion of late-replicating genotypes carrying this feature, and  $p_{perm\_early}$   
8 and  $p_{perm\_late}$  are estimated from ten permutations (described below). Bonferroni correction was  
9 used to correct for multiple testing at the 0.05 level.

10 In each permutation, we used random SNPs matched for rtQTLs (for details see the  
11 “enrichment analyses” section), and randomly designated one genotype as the early-  
12 replicating genotype. We obtained genotype and epigenetic information from the seven cell  
13 lines at these random SNPs and calculated the proportion of early- and late-replicating  
14 genotypes carrying the feature in ten permutations ( $p_{perm\_early}$  and  $p_{perm\_late}$ ).

15 We examined the relationship between early-replicating genotypes and expression of nearby  
16 genes (within 200 kb of rtQTLs). Array-based expression data was obtained for nine ES cell  
17 lines<sup>69</sup> for which replication timing data was also available. Genes with mean expression  
18 level > 1 were used, and expression level was normalized within each gene. We aggregated  
19 the expression levels of all genes near all rtQTLs for the nine ES cell lines (except for rtQTLs  
20 that were monomorphic in these cell lines), and tested the correlation between expression  
21 level and the number of early-replicating alleles.

22

### 23 *Transcription factors*

24 To identify TFs that regulate replication timing, we tested whether rtQTL alleles (in the CAVIAR  
25 90% causal set) influence the binding affinity (motif score) of 21 TFs<sup>70</sup>. Under the hypothesis  
26 that some rtQTLs function by altering sequence motifs of TFs that promote or repress  
27 replication, early-replicating alleles will be more likely to have higher binding affinities than late-  
28 replicating alleles to the TFs that promote earlier replication, and *vice versa* for late-replicating  
29 alleles. We used this principle to identify TFs associated with replication timing. We tested the  
30 motifs of all TFs studied in Fig. S5A, if available. Of note, SOX2 was not included in this  
31 analysis because its motif information was not available. This analysis was repeated with iPSC  
32 rtQTLs. We were not able to perform the analysis described above for chromatin states and  
33 histone marks with TFs because TF ChIP-seq data was only available for one hESC or iPSC  
34 line.

35 TF binding affinity data, measured by motif score, was obtained from HaploReg<sup>70</sup>. Sequence  
36 logos for TF binding motifs were created using WebLogo 3<sup>71</sup>. For each rtQTL SNP, motif  
37 scores of the two alleles were obtained for the TFs, and their difference is the  $\log_2$  fold

1 difference in probability that the allele is in a binding motif of the given TF. Higher difference in  
2 motif scores means that this SNP can more substantially alter the binding affinity of this TF.  
3 For each TF, we counted the weighted number of rtQTLs for which the early-replicating (or the  
4 late-replicating) allele had higher predicted binding affinity, weighted by the difference in motif  
5 scores between the two alleles, i.e., rtQTLs with a higher motif score difference will have  
6 heavier weight. This weighting scheme assigns heavier weights to those rtQTLs for which the  
7 changes in allele state will result in more substantial change in TF binding affinity. If there were  
8 more than one potential causal SNPs of an rtQTL located within binding motifs of a given TF,  
9 the SNP with the lowest  $p$ -value was used. We compared the numbers to permutations, in  
10 which SNPs matched for rtQTLs were randomly selected and the early-replicating alleles were  
11 randomly assigned, using the chi-squared test for a 2×2 contingency table. This test assesses  
12 whether the early-replicating alleles are more (or less) likely to have higher TF binding affinity  
13 than late-replicating alleles than expected by chance. Benjamini-Hochberg correction at 10%  
14 FDR was used to correct for multiple testing.  
15 For OCT4, NANOG, and CTCF (for which there are abundant ChIP-seq data available in  
16 hESC), we repeated this analysis using only motifs that overlap with TF ChIP-seq peaks (i.e.,  
17 confirmed TF binding). Consistent with the results in Fig. 6A, we found that OCT4 and NANOG  
18 were significantly more likely to bind early-replicating alleles ( $p = 5.97 \times 10^{-7}$  and  $2.62 \times 10^{-15}$ ;  $\log_2$   
19 ratio improved from 0.27 and 0.29 to infinity and 2.58, respectively), while CTCF was  
20 significantly more likely to bind late-replicating alleles ( $p = 0.02$ ,  $\log_2$  ratio improved from -0.22  
21 to -1.19).  
22  
23 Supplementary Figures were included in the main text near where they were mentioned.  
24  
25 Table S1. List of rtQTLs Identified in This Study.  
26 For the last column (“classification”), “VALLEY” or “SLOPE” denotes that this rtQTL affects valley or slope,  
27 respectively. “PEAK (SNP proximal)” or “PEAK (SNP distal)” denotes that this rtQTL affects peak, and the top  
28 rtQTL SNP is proximal or distal to the peak, respectively.  
29 This table is provided in a separate Excel spreadsheet.  
30