DNA replication follows a strict spatiotemporal program that intersects with chromatin structure but has a poorly understood genetic basis. To systematically identify genetic regulators of replication timing, we exploited inter-individual variation in human pluripotent stem cells from 349 individuals. We show that the human genome’s replication program is broadly encoded in DNA and identify 1,617 cis-acting replication timing quantitative trait loci (rtQTLs) – sequence determinants of replication initiation. rtQTLs function individually, or in combinations of proximal and distal regulators, and are enriched at sites of histone H3 trimethylation of lysines 4, 9, and 36 together with histone hyperacetylation. H3 trimethylation marks are individually repressive yet synergistically associate with early replication. We identify pluripotency-related transcription factors and boundary elements as positive and negative regulators of replication timing, respectively. Taken together, human replication timing is controlled by a multi-layered mechanism with dozens of effectors working combinatorially and following principles analogous to transcription regulation.
Eukaryotic genomes are replicated according to a strict spatiotemporal program, in which replication initiates from specific locations along chromosomes and at reproducible times. The replication timing program is a fundamental property of chromosome organization, interfaces with gene regulation, and shapes the mutational landscape of the genome. Efforts to understand the locations and nature of initiation sites and the factors that regulate DNA replication timing in mammalian cells have been ongoing for decades, with limited success\(^1\). Specifically, it is still unclear to what extent the DNA replication timing program is determined by local DNA sequences, by epigenetic factors, or by a combination thereof. Earlier studies suggested that specific sequence elements control replication initiation in human cells, with several distal and proximal elements often acting in concert\(^2\)–\(^10\). More recently, CRISPR/Cas9-mediated deletions have suggested that several DNA sequences locally interact to control early replication in mice\(^11\).

Numerous lines of evidence link replication regulation to epigenetic states, in particular histone acetylations and methylations marking open chromatin\(^1\)–\(^4\),\(^12\)–\(^16\). However, no single epigenetic mark appears to be absolutely required nor sufficient for replication origin function. This has led to suggestions that a combination of histone marks may be required for specifying patterns of DNA replication\(^17\). Similarly, it has been proposed that indiscriminate DNA-binding patterns of the replication machinery may translate into a consistent, organized replication program by means of combinatorial chromatin modifications influencing subsets of replication initiation sites\(^3\).

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

**DNA replication timing is broadly influenced by cis-acting sequences.** While replication timing profiles were highly reproducible among individuals, we nonetheless observed genomic regions with substantial inter-individual variation. We identified 1,489 autosomal replication timing variant loci in hESCs and 1,837 in iPSCs, cumulatively encompassing 795 Mb (34%) and 980 Mb (40.8%), respectively, of the analyzable genome (Fig. 1c, d). We hypothesized that at least some of this variation is due to genetic polymorphism. To test this, we first compared replication timing variation between 24 pairs of hESC lines that are genetic siblings, versus unrelated cell lines; between genomic regions that are identical by descent (IBD), half-identical or non-identical between sibling cell lines; and between 108 pairs of iPSC lines derived from the same donor, compared to different donors (Methods). Consistent with a significant genetic contribution to replication timing variation, samples or genomic intervals that are genetically related consistently showed greater replication timing similarity than unrelated comparisons (Fig. 1e–g). Using pairs of iPSC lines derived from the same donor, we observed substantial and consistent donor effects across replication timing loci genome-wide (median: 20.7%, maximum: 71.9% of the total variance). Donor effects accounted for the largest proportion of variance in 33.1% of replication timing loci. These results were comparable with previous observations of donor effects on iPSC gene expression\(^20\), and suggest that differences between donors substantially impact iPSC replication timing.

To further dissect genetic contributions to replication timing variation, we used our previously described rtQTL mapping approach\(^18\) to associate replication timing with specific genetic polymorphisms. We limited this analysis to 108 hESCs of European ancestry and to 192 iPSCs from different individuals. We identified 1,617 cis-rtQTLs (FDR 0.1; 1,012 were identified with FDR 0.05; Fig. 1i–m; Supplementary Data 1), two orders of magnitude more than previous associations of replication timing with cis-acting sequences\(^11\),\(^18\). The greater success in rtQTL identification was mainly driven by the more deeply sequenced genomes, the high proliferation rate of stem cells and by the improved analytical framework (Methods, Supplementary Fig. 1c). We used CAVIAR\(^22\) to fine-map (90% credible level; Supplementary Data 2) a median of 33 SNPs per rtQTL, with 316
rtQTLs mapped to within 10 SNPs and 36 rtQTLs mapped to no more than three SNPs. rtQTL mapping was cross-validated between ES and iPS cells and further confirmed using additionally-sequenced cell lines and with a locus-specific single-molecule assay (Supplementary Fig. 2). We found that replication timing loci that harbor rtQTLs had significantly greater donor effects than other loci (median: 20.8% vs. 18.8%, Wilcoxon rank-sum p << 2.2×10^{-16}). We further explicitly modelled rtQTL genotype in variance component analyses of replication timing at the loci with the strongest association of each iPSC rtQTL, and found that on average, 54.7% of donor effects (17.9% of the total variance) can be explained by rtQTL genotype. These findings indicate that genetic differences substantially contribute to the donor effects on replication timing.

rtQTLs influenced the replication timing of regions spanning 858 kb on average and a total of 741.8 Mb of genomic sequence (31.8% of the genome, Fig. 1m). This is a lower bound estimate of the extent to which human replication timing is influenced by DNA sequence, since our approach will not detect weaker rtQTLs, invariant sequences or rare variant rtQTLs. Intriguingly, 67.9% (1,098) of rtQTLs coincided with sharp peaks in the replication profile (binomial p = 2.24×10^{-25}; Fig. 1i–l), and rtQTL SNPs were significantly closer to peaks than expected (Wilcoxon rank-sum p = 1.77 × 10^{-16}). This suggests that rtQTLs may influence replication initiation, as previously reported.

**Fig. 1 The Human Genome’s Replication Timing Program is Extensively Encoded in DNA.**

- **a** Replication timing (blue line; Z-score) inferred from read depth fluctuations (gray) for the H9 cell line. Orange: Repli-Seq data for H9.
- **b** Replication timing profiles are highly reproducible among samples.
- **c** PC-correction improves replication profile accuracy.
- **d** Genetic relatedness associated with replication timing similarity.
- **e** Sibling vs. unrelated hESCs (two-sided Wilcoxon rank-sum test).
- **f** Genomic regions stratified by increasing identity-by-descent (ANOVA).
- **g** iPSCs from the same or different donors (two-sided Wilcoxon rank-sum test).
- **h** Analysis of replication timing divergence between genotypes at rs12713840, the top SNP.
- **i** Additional rtQTL examples.
- **j** RTQTLs and local effects.
- **k** rtQTLs extending beyond the replication timing locus.
- **l** rtQTLs influencing genome-wide replication timing.

Outliers removed using R boxplot option “outline=F”.

A genomic region (gray) with inter-individual replication timing variation.

i–k Genetic association reveals rtQTLs.

A haplotype strongly associated with the replication timing variant from panel h (panel k genome-wide association).

Mean replication timing (left Y axis) for individuals with different genotypes at rs12713840, the top SNP, demonstrates that SNPs in cis (right Y axis) associate with replication initiation.

Gray: affected region.

j Replication timing at the variant from panel h, stratifying individuals by rs12713840 genotype. Genotype is the main determinant of replication timing variation.

**Fig. 1** Additional rtQTL examples.
and that most rtQTLs can be used as fine-scale markers of replication initiation regions. We ruled out that rtQTLs are predominantly markers of gene expression variation: only 13.7% of iPSC rtQTLs were associated with variation in expression of genes overlapping rtQTL-affected regions (14.6% when extending an additional 1 Mb on each side). The identification of rtQTLs as precise genetic determinants of replication timing provides a unique opportunity to fine-map molecular mechanisms controlling replication initiation and timing.

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

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

A recent study suggested that synergistic interactions between cis-acting elements are required for controlling early replication. We directly tested for interactions between primary and secondary rtQTLs at regions that harbored two rtQTLs, hence between zero and four early-replicating alleles. We further pooled all genomic regions containing three or four rtQTLs and evaluated the relationship between the number of early-replicating alleles and the replication timing of the associated regions. Replication timing showed a linear relationship with the number of early-replicating alleles (linear regression \( p < 2.2 \times 10^{-16} \), Fig. 2e, f), and none of the individual regions showed evidence for synergistic interactions between rtQTLs. We also systematically searched for interactions between pairs of genetic variants in affecting replication timing, yet identified none (Methods). While these findings do not rule out synergistic relationships, they suggest that rtQTLs operate in a predominantly additive manner.

Of the 318 multi-rtQTL regions, 176 were associated with replication timing peaks. In 115 of these cases (65.3%), primary rtQTLs were closer to the peak than secondary rtQTLs (Fig. 2g, \( p = 3.28 \times 10^{-9} \)). This resembles eQTLs (expression QTLs), in which primary eQTLs show stronger enrichment at promoters, while weaker eQTLs are enriched at enhancers. Also in resemblance to enhancers and promoters, primary and secondary rtQTLs tended to cluster in nuclear space (based on Hi-C data) more than expected by chance (\( p = 9.73 \times 10^{-3}; Z \) test, Supplementary Fig. 3). Drawing from this analogy, we propose that rtQTLs may follow a logic akin to promoters and enhancers, in which primary rtQTLs function as main cis-acting regulators of replication initiation, while other sequences, marked by secondary rtQTLs, serve as distal regulatory elements that fine-tune the replication dynamics of a given region.

### Histone modifications associated with DNA replication initiation.

We next utilized the basepair-resolution sequence-specificity of rtQTLs to investigate the molecular mechanisms of DNA replication timing. We initially considered rtQTL locations per se, independently of allelic variation. Since extensive epigenetic data were available for seven of the hESC lines in our dataset, we focused this analysis on hESCs and used iPSCs for validation. Given that replication initiation in mammalian cells is thought to occur in diffuse regions encompassing at least several kbs, we initially considered inclusive chromatin data (“gapped-Peak”), combined from several cell lines when possible, and subsequently validated our findings in individual cell lines (Methods). Consistent with previously described correlations between early replication and open chromatin, rtQTLs were enriched for active chromHMM states including enhancers and transcription start sites (although they were not specifically associated with genes; Supplementary Fig. 4), DNase I hypersensitivity sites (\( p = 2.62 \times 10^{-8}; 4.11 \times 10^{-19} \) in iPSCs), and H2AZ sites (\( p = 6.69 \times 10^{-3}; p = 3.20 \times 10^{-17} \) in iPSCs). rtQTLs also significantly overlapped with 24 histone marks (25 in iPSCs), of which 20 were active marks (Supplementary Fig. 4). The majority of these histone marks were acetylations, including several not previously linked to replication timing, for example, H2BK120ac, H2BK12ac, and H2BK20ac. H3T11ph was also consistently enriched at hESC and iPSC rtQTL sites, and so were, modestly, methylated forms of H3K4. No significant enrichments were found when the analysis was repeated on genomic loci that were distant from rtQTLs or replication timing peaks, but that matched their timing distribution (Supplementary Fig. 1i and j). This suggests that replication timing per se is unlikely to explain the epigenetic marks associated with rtQTLs.

Of note, the histone mark enrichments were modest, with each present at between 165 to 541 (median: 429) of 608 hESC rtQTLs (median: 542 of 1167 iPSC rtQTLs), while each rtQTL overlapped 20 histone marks on average. We surmised that this abundance of histone modifications may be suggestive of combinatorial regulation. To test this, we systematically searched for combinations of histone marks with stronger enrichments at rtQTLs when considered jointly (Methods). We identified 152 combinations of two overlapping histone marks that were more enriched than the individual marks. We further identified 128 co-enriched three-mark combinations, 72 four-mark combinations, and 15 five-mark combinations (enrichment \( p \) values: 2.42\( \times 10^{-37} \text{–} 1.09 \times 10^{-45}\)), at which point no further improvements in enrichment were obtained (Fig. 3a; Supplementary Data 3). Although these enrichments were relatively modest, finding higher-order combinations of histone modifications at rtQTL sites suggested a potentially novel principle of replication timing regulation. Importantly, since these enrichments controlled for replication timing, they were not identified because they mark early replicating regions, but because they specifically mark rtQTL locations, and, by inference, replication initiation sites. Repeating this analysis using early vs. late-replicating peaks, or using peaks vs. valleys, only revealed individually enriched histone marks or pairs thereof, but not higher order combinations. Thus, our ability to identify the myriad epigenetic factors and combinations related to replication timing is directly related to our approach of rtQTL mapping in a population cohort.

Strikingly, all 13 combinations of five histone marks contained the trimethylation marks H3K9me3 and H3K36me3, and 12 of the combinations also contained H3K4me3. In addition, all 13 combinations included at least one histone acetylation mark. H3K56ac was included in 11 of the combinations, while the additional acetylations occurred on variable histone residues (Fig. 3b). The appearance of H3K56ac in multiple combinations was attributed to its more unique presence across the genome compared to the other histone acetylations, which tended to be redundant with each other (Supplementary Fig. 5a, b). Indeed, various acetylation marks often coincided with the five histone mark combinations, e.g., in 70.8% of the cases, 11 or more
Acetylation marks co-occurred at the location of a five-mark combination. We term this combination of three H3 trimethylations together with hyperacetylation the "me3achyper histone modifications." Both the enrichment of individual histone modifications and the me3achyper combinations were also observed in individual cell lines, the latter when extending the overlapping regions by 1 kb on each side of the modification peaks (Figs. S6 and S7). Thus, these modifications occurring in close proximity to each other may mark the locations of replication initiation sites. We note, however, that our data does not necessarily imply that all these modifications fall at the exact same genomic sites or on the same nucleosomes. Genome-wide, there were 6,670 such locations in hESCs (Supplementary Data 4) and a total of 5,120 identified in individual cell lines (median of 633 per cell line). They covered a median of 635 bp (1,597 bp with the extended overlap requirement) and cumulatively encompassed 0.24% (0.31%, respectively) of the genome, thus they represent specific, relatively localized genomic sites.

Importantly, when considered individually, the implicated histone modifications only showed weak enrichments (Supplementary Figure 4c). H3K9me3 and H3K36me3, in particular, showed marginal or no enrichment at rtQTLs. H3K9me3 is a...
marker of heterochromatin (although has been observed in the bodies of actively transcribed genes)\textsuperscript{26}, while H3K4me3 marks gene promoters and H3K36me3 is typically present in gene bodies. These histone trimethylations are largely mutually exclusive. However, in rare cases, they coincide in nearby genomic locations (Fig. 3c, d). These trimethylation marks with disparate effects on chromatin structure, together with histone hyperacetylation, comprise a unique chromatin state that is distinct from previously described bivalent chromatin (e.g., the combination of H3K4me3 and H3K27me3)\textsuperscript{27}. It is in these rare locations of unusual chromatin structure that rtQTLs tend to be present.

The me\textsuperscript{3ac}\textsuperscript{hyper} histone modifications appeared to be specific to DNA replication initiation rather than transcription, as only 5.2\% of histone modification regions overlapped with the transcription start site (TSS) of an active gene and the majority (71.2\%) of rtQTLs that overlapped the histone modifications did not associate with gene expression variation. On the other hand, G-quadruplexes were enriched at me\textsuperscript{3ac}\textsuperscript{hyper} histone modification sites (Supplementary Figure 4k–n), consistent with previous reports of replication origin mapping\textsuperscript{28–31}.

Subsets of the identified histone mark combinations have been previously linked to the recruitment of components of the replication machinery to specific chromosomal locations. Histone H3 trimethylation of lysines 4, 9, and 36 have been shown to exert a cross-talk that serves as an “epigenetic addressing system” for site-specific replication initiation\textsuperscript{32,33}. They recruit KDM4 and KDM5 family histone demethylases, which directly interact with, and/or are required for recruitment to DNA of MCM, PCNA, DNA polymerases, and other replication factors\textsuperscript{32–36}. H3K4me3
also synergizes with flanking H3K9ac and H3K14ac (both identified as part of the me\textsuperscript{ac\textsubscript{hyper}} histone modifications) to recruit chromatin readers to DNA\textsuperscript{35}. Another study showed that histone hyperacetylation synergizes with H3K9me3 to promote early replication of otherwise late-replicating mouse chromocenters\textsuperscript{36,38}. In turn, acetylated histones have been shown to recruit replication initiation factors including TICRR/TRESLIN, ORC, and MCM, via mediators such as BRD2, BRD4, and the histone acetyltransferase HBO1 (histone acetyltransferase binding to ORC)\textsuperscript{32,39–41}. In particular, HBO1 promotes MCM loading by acetylating H4 on lysines 5, 8, and 12, and subsequently promotes origin activation by acetylating H3K14\textsuperscript{42,43}; we identified all of these acetylations as part of the me\textsuperscript{ac\textsubscript{hyper}} combinations. Moreover, H4K12ac, the most strongly enriched mark at rtQTLs, is a preferred target of HBO1 at replication origins\textsuperscript{39,41}. These biochemical evidence provide a plausible explanation for the combination of histone marks being associated with replication initiation activity.

Taken together, we identified a combination of histone marks, consisting of three trimethylated H3 residues (H3K4me3, H3K9me3, H3K36me3) together with H3K56ac and broadly hyperacetylated chromatin that consistently coincide with rtQTLs. To further test the involvement of these histone modifications in replication initiation, we analyze below their association with: (1) replication timing peaks in general (independent of rtQTLs); (2) replication timing peaks in other cell types; (3) replication timing peaks that vary between cell types; (4) replication timing variation among individuals at rtQTLs.

Combinations of histone modifications predict replication initiation sites across cell types. We considered whether the histone modification combinations could be a general property of replication initiation sites, revealed by leveraging the base-pair resolution of rtQTLs, but not limited to rtQTLs. We therefore tested whether the histone modifications also associated with the larger number of replication timing peaks (found in >10% of the samples; Supplementary Data 5) not identified as rtQTLs (81.5% of all peaks). While the probabilities of having a peak near the 24 individually enriched histone marks were significantly greater than expected (one-tailed Wilcoxon rank-sum \( p = 4.96 \times 10^{-17} \)) and were greatest at the actual histone mark sites (Supplementary Fig. 5e–g), individual histone marks are very common in the genome and insufficient for predicting peaks. We refined prediction through combining histone marks (Supplementary Fig. 5h and i), demonstrating a gradual increase in association with peaks up to the five-mark combinations, which were significantly more likely than expected to coincide with peaks (Supplementary Fig. 5e–g, \( p = 4.10 \times 10^{-19} \)). me\textsuperscript{ac\textsubscript{hyper}} sites had an even higher likelihood of overlapping peaks (Supplementary Fig. 5e–g). Consistently, the distances of me\textsuperscript{ac\textsubscript{hyper}} regions to the nearest peak were substantially shorter than random permutations (42.57 kb; fold-improvement over random: 3.83-fold [physical], 3.17-fold [fractional]; Fig. 3e, f; Similar results were obtained for the histone modification combinations verified in individual cell lines: median distance to nearest peak of 42.10 kb). For example, 41.7% of the me\textsuperscript{ac\textsubscript{hyper}} sites corresponded to replication initiation timing peaks within 10 kb (positive predictive value; Z-test \( p < 2.2 \times 10^{-16}, 1.70\text{-fold improved over random} \)). Conversely, 60.3% of peaks were located within 10 kb of predicted regions (sensitivity; \( p < 2.2 \times 10^{-16}, 1.51\text{-fold enriched} \)). We further evaluated the prediction performance of the me\textsuperscript{ac\textsubscript{hyper}} regions visually (Fig. 3g) and with receiver operating characteristic (ROC) curves (Supplementary Fig. 5j–l, area under the ROC curve \( \text{AUC}_{\text{ROC}} \) for me\textsuperscript{ac\textsubscript{hyper}} regions: 0.776). Peaks predicted by histone marks replicated earlier than other peaks (median: 0.61 vs. 0.14, Wilcoxon rank-sum \( p = 6.58 \times 10^{-53} \)) and were locally more prominent (timing difference compared to flanking valleys, median: 0.32 vs. 0.18, \( p = 1.19 \times 10^{-17} \)). Consistently, the replication profiles surrounding me\textsuperscript{ac\textsubscript{hyper}} sites formed a sharp peak (Supplementary Fig. 5o).

The histone modification combinations were substantially more specific than DNase I hypersensitivity (Supplementary Fig. 5q), which was previously suggested to explain 87% of replication timing profiles\textsuperscript{34}. In addition, the ability to predict the locations of replication initiation sites was even greater for local clusters of the histone modifications. We found that genomic regions containing multiple nearby instances of the me\textsuperscript{ac\textsubscript{hyper}} histone modifications, and/or histone modification instances spanning longer regions, were closer to replication timing peaks than sparser occurrences of the modifications (Supplementary Fig. 5m). For example, local regions (10 kb windows) having at least 30% covered by the histone modifications had a median distance of 4.93 kb to peaks, an 8.75-fold improvement compared to all me\textsuperscript{ac\textsubscript{hyper}} histone modification regions and an 32.8-fold improvement compared to random. This is consistent with the idea that chromatin structure influences DNA replication initiation in a regional manner, and that clusters of these histone modifications function cooperatively to promote DNA replication initiation.

In order to test whether the histone modification combinations could be a consequence of early replication rather than a potential cause of replication initiation, we tested genomic loci with matched replication timing (difference < 0.5) that were distant (>500 kb away) from the center of replication timing peaks. The median distance from histone modification sites to these loci was 843.25 kb, compared to 43.16 kb for replication timing peaks (19.54-fold different) (Supplementary Fig. 5n). Thus, the histone modification combinations associate specifically with replication initiation.

Taken together, the combinations of nearby histone marks that are enriched at rtQTLs predict ~60% of initiation site locations across the genome at 10 kb resolution, even for those sites without rtQTLs, and particularly for the early and most prominent initiation sites. These histone mark combinations may thus promote replication initiation not just at specific genomic loci, as previously proposed\textsuperscript{32,33,45}, but across a large fraction of the genome. We note, though, that some replication timing peaks did not co-localize with histone modification locations, thus there must be additional mechanisms independently specifying replication initiation sites, underscoring the complexity of mammalian replication initiation.

An even more rigorous test of the five-mark combinations being indicators of replication initiation is whether they could predict the location of replication timing peaks in other cell types. Examining the histone modifications (identified in hESCs) in both iPSCs (Supplementary Data 6) and lymphoblastoid cell lines (LCLs; for which we only had experimentally measured data for the trimethylation marks)\textsuperscript{46–48}, we found that the same histone modifications can predict initiation sites (Supplementary Data 7 for iPSCs) as accurately and specifically as in hESC (Fig. 3h, i), and similarly associates with early replication (Supplementary Figure 5o). In particular, LCLs have epigenetic and replication timing landscapes that are distinct from those of hESC (and iPSCs). In genomic regions at which LCL and hESC replication timing differed, LCL-specific histone modification locations corresponded to LCL-specific initiation sites, and vice versa for hESCs (Fig. 3j). Predicted cell-type-specific initiation sites resided in early-replicating genomic regions in the corresponding cell type, but not in other cell types (Supplementary Fig. 5p).
Incorporating imputed (computationally predicted) LCL acetylation data further improved prediction performance from a median distance of 177.49 kb to the nearest replication timing peak for me³, to 130.69 kb for me³ + imputed H3K56ac, and down to 119.01 kb for me³ + imputed acH3K56ac. Thus, the histone modification combinations characterized and predicts cell-type-specific replication initiation.

Co-variation of replication timing and histone modifications reveals combinatorial control of replication timing. The previous analyses considered rtQTL locations per se. However, since rtQTLs represent replication timing variation among individuals, their allelic differences provide a powerful opportunity to investigate molecular mechanisms controlling replication timing. In particular, given that specific histone marks associate with replication initiation, we predicted that rtQTL SNP alleles will be associated with variation in the abundance of these marks among individuals.

We took an unbiased approach using seven hESC lines with both replication timing and histone modification data (Methods). Cell lines carrying early replicating genotypes at rtQTLs were more likely than individuals with late-replicating genotypes to harbor active histone marks and chromHMM states at those rtQTL sites (Fig. 4 and S8). Across individuals and genomic sites, eight histone modifications were consistently present in individuals with rtQTL alleles indicative of early replication. Of those, seven were acetylations, consistent with histone acetylation promoting early replication31,12–16,39,41. Of the 12 acetylation marks that are part of the replication initiation histone modification combinations, nine individually associated with early replicating rtQTL genotypes (five of which reached statistical significance). We also identified seven modifications that consistently coincided with late replicating alleles, of which six were methylation marks (Fig. 4a); Thus, histone methylation emerges as being generally repressive for replication.

Counter-intuitively, the histone trimethylation marks (H3K4me3, H3K9me3, and H3K36me3) were individually more likely to be associated with late-replicating genotypes (Fig. 4a). In contrast, the combination of all three trimethylation marks was 2.5-times more likely to be carried by early-replicating than by late-replicating genotypes. Furthermore, a combination that also included H3K56ac was 7.24-times more likely to be carried by early-replicating genotypes (Fig. 4a). Thus, these marks appear to individually act as weak repressors of replication but act synergistically, in noncanonical ways, to strongly promote early replication. Taken together, the involvement of me³acH3K56ac in replication initiation is supported by several lines of evidence: enrichment at rtQTLs (Fig. 3a); correspondence with replication timing peaks in general, and across several cell types (Fig. 3d–i, Supplementary Figure 5e–g, j, and o); co-variation with cell-type-specific replication initiation patterns (Fig. 3j and Supplementary Fig. 5p); and correlation with inter-individual replication timing variation (Fig. 4).

DNA-binding factors modulate DNA replication timing. The above results indicate that cis-acting sequences, manifesting as rtQTLs, influence the positions and timing of replication initiation by associating with histone modifications. To identify additional factors that influence replication timing via cis-acting sequences, we analyzed the binding sites of 51 DNA binding factors in hESCs46,49. The binding of eight factors was significantly enriched at rtQTLs, including the main pluripotency factors SOX2, Pou5F1 (OCT4), and NANOg, the latter two reproducible with available data in iPSCs (Supplementary Fig. 9). Three chromatin remodelers, EP300 (P300), SP1, and RBBP5, were also enriched at rtQTLs. EP300 is a histone acetyltransferase that catalyzes at least six acetylation marks in the replication initiation histone modification combinations, including H3K56ac29.

Transcription factors (TFs) bind DNA in a sequence-specific manner at characteristic motifs. This offers an opportunity to test, at base-pair resolution, whether TF binding affects replication timing at rtQTLs (Methods). Strikingly, OCT4 and NANOG had significantly higher binding affinity for early- compared to late-replicating alleles in both hESCs and iPSCs, while EP300 and ATF3 (Activating Transcription Factor 3, which is enriched at EP300 sites43), were linked to early replication at least in hESCs (Fig. 5a, b). These associations appeared to be independent from gene expression, as they were reproduced for rtQTLs > 250 kb away from expressed genes. For these early-replication-associated TFs, the rtQTLs fell within the TF binding motifs such that a single base-pair change disrupted or even abolished binding; this was associated with delayed replication of the rtQTL-affected initiation site (Fig. 5c). An unexpected finding was rtQTL alleles with the opposite effect, i.e., higher binding affinity associated with later-replication. We infer that in these cases protein binding suppresses replication initiation (Fig. 5). These included CTCF, an insulator of topologically associated domains (TADs); REST (NRSF), a repressor of transcription52; ZNF143, which associates with the CTCF-cohesin cluster53, and at least in hESCs also RAD21 (part of the cohesin complex) and YY1, which co-localize with CTCF at TAD boundaries54–57. These associations were yet stronger when considering only motifs with biochemically confirmed TF binding when data were available (Methods). Since some TF motifs were inferred using ChIP-seq data, they may be motifs for co-factors, and in these cases, rtQTLs may affect the binding of the co-factor instead of the TF per se. Furthermore, TFs and the histone modifications appeared to be independent determinants of replication timing. The majority (82%) of the me³acH3K56ac histone modification regions did not overlap any of the TFs that were positively associated with replication timing, and the histone modification predicted replication timing peak locations even better when it was not associated with activator TFs (median distance: 38.8 kb) than at regions overlapping TF binding (53.7 kb; random was 161.71). Taken together, we conclude that some rtQTLs alter DNA binding protein motifs, abolish a DNA binding site or generate a new one, and consequently alter DNA replication timing through specific TF (or co-factor) binding. This analysis uncovers several factors that can thus regulate DNA replication timing. In addition, different factors influence subsets of replication initiation sites, further illuminating the complex combinatorial landscape that controls human DNA replication timing. Finally, these results demonstrate one molecular mechanism whereby a single base-pair alteration could affect the replication timing of an extended genomic region.

Discussion

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

Studying chromatin structure at rtQTL sites revealed histone modification combinations that predicts a subset of replication initiation sites across cell types. These combinations represent noncanonical functions of histone H3 lysine methylations that form a previously undescribed bivalent chromatin state27 present at localized sites throughout the genome. Prior biochemical evidence supports an involvement of these histone marks in DNA
**Fig. 4 Histone marks associate with DNA replication timing.**

a Association of rtQTL genotypes with individual (left panel) or combinations (right panel) of histone marks. Positive (negative) values indicate that individuals with early (late)-replicating genotypes are more likely to carry a histone mark at those rtQTL sites. Right panel: while individual H3 methylation marks associate with late replication, the H3K4me3-H3K9me3-H3K36me3 combination is strongly associated with early replication, and even more so when combined with H3K56ac. Note the different Y scale.

b Examples of rtQTLs associated with histone mark variations. Replication timing and corresponding histone ChIP-seq tracks for individual cell lines homozygous for the early- or late-replicating alleles. Early replication correlates with the presence of the specified histone marks.
We propose that these histone modifications promote local replication activity, although we do not necessarily imply that they mark the locations of replication origins per se. We suspect that there could be various such combinations of histone modifications, spanning localized chromosomal regions, that influence the propensity of replication initiation. Further studies with yet more comprehensive replication timing and chromatin data could refine and expand the associations between the epigenome and replication timing regulation.

rtQTLs further associated with inter-individual variation in histone marks and TF binding affinity. In many cases, several cis-
Fig. 5 rtQTLs affect replication timing by altering TF binding motifs. a, b Binding of TFs such as OCT4 and NANOG promotes earlier replication, while binding of CTCF, REST and other factors is associated with later replication in hESCs (a) and iPSCs (b). Chi-squared test, FDR <10%. c Examples of rtQTLs altering binding affinity of TFs that function as replication activators or repressors. Heterozygous profiles are not plotted. Center panels: Chip-seq tracks. Lower panels: sequence logos of the motifs containing the rtQTL SNPs, motif binding names, and changes in binding affinity (calculated based on motif scores). Asterisk indicates that the motif was on the negative strand and the sequence shown is the reverse complement. Red arrows: locations of the rtQTL SNPs. For activators, the rtQTL allele associated with early replication encodes an intact binding motif, while the allele associated with late replication abolishes the motif. Repressors have the opposite pattern: the early allele abolishes the motif.

Whole-genome sequence data. Whole-genome sequence data and genotype calls for 121 hESC lines were obtained from Merkle et al.19. We denote these as “Merkle_batch1”. Nine additional hESC lines were sequenced in another batch from Merkle et al.39, denoted as “Merkle_batch2”. We further used whole-genome sequence data from 326 iPSC lines from the HipSci Project36 (ENA accession number: PRJEB15299), denoted as “HipSci”. We sequenced an additional 15 hESC lines and 17 iPSCs (dbGaP accession number: phs001957), denoted as “in_house_hESC” and “in_house_iPSC”, respectively.

For the in-house datasets, DNA was extracted using the MasterPure Complete DNA and RNA Purification Kit (Lucigen). Sequencing libraries were prepared using the Illumina TruSeq PCR-free kit and sequenced on an Illumina HiSeq X Ten to ~16-fold coverage with 150 × 2 paired-end reads. Sequencing was performed at GeneWiz (South Plainfield, NJ). Reads were aligned to GRCh37 using BWA, and genetic variants were called following the GATK Best Practices. Variants were filtered using GATK’s variant quality score recalibration, such that SNPs had a 99.9% sensitivity to true variants (HapMap 3.3 and Omni 2.5M)38 and a 99.0% sensitivity to true indels (Mills / 1000 Genomes indels)39,40.

Cell culturing

iPS derivation. Two-milliliter Skin biopsies were made using the AccuPunch Biopsy kit (Acuderm Inc.) after local anesthesia using lidocaine. Biopsies were dissected in 4-8 pieces and placed under a glass coverslip held to the bottom of a six-well dish using silicon grease in DMEM medium containing 10% FBS (Atlanta Biologicals #10828010) and 10ng/ml bFGF (Thermo Fisher #13256029), Rock inhibitor Y-27632 (Selleckchem #S1049) and embryonic stem cell grade FBS (HyClone). After ESC outgrowth became visible, Rock inhibitor was phased out, and passaged manually for one to two passages in the same media. Thereafter, ESCs were passaged enzymatically using TrypLE Express (Thermo Scientific #12605036) and were cultured on Geltrex (Thermo Fisher #A1413302) in StemFlex medium (Thermo Fisher #A3349401).

ESC derivation. Human embryonic stem cells (ESC) were derived as previously described31,42. Briefly, the trophoderm of blastocysts was ablated using 20–40 laser pulses of 400ms at intensity set to 100% (Hamilton Thorne). The inner cell mass was plated on irradiated mouse embryonic fibroblast feeders (Globalstem), DMEM HG (Thermo Fisher #10569010), Glutamax (Thermo Fisher #35000709) and Pen/Strp (Thermo Fisher #15141063) and 18% KOSR (Thermo Fisher #10828010) and 10ng/ml BFGF (Thermo Fisher #13256029), Rock inhibitor Y-27632 (Selleckchem #S1049) and embryonic stem cell grade FBS (HyClone). After ESC outgrowth became visible, Rock inhibitor was phased out, and passaged manually for one to two passages in the same media. Thereafter, ESCs were passaged enzymatically using TrypLE Express (Thermo Scientific #12605036) and were cultured on Geltrex (Thermo Fisher #A1413302) in StemFlex medium (Thermo Fisher #A3349401). Blastocysts were obtained after ICSI of donated oocytes and sperm.

Inference of DNA replication timing. DNA replication timing was inferred by analyzing sequence read depth (corrected for GC content bias) in non-overlapping windows of 10 kb of uniquely alignable sequence using GenomeSTRiP18,63. Among the 121 hESC lines from Merkle_batch1, five did not optimally thrive in culture, resulting in read depth profiles with low correlations to other samples; these cell lines were excluded from further analysis. We excluded 26 of the 326 iPSC lines from the HipSci dataset for the same reasons. As described below, further filtering were performed for Merkle_batch1 and HipSci datasets. Replication timing inference for the in-house datasets is described separately (see the “validation of rtQTLs” section below).

Whole-genome sequencing and gene expression analysis. From the hipSci dataset for the same reasons. As described below, further filtering were performed for Merkle_batch1 and HipSci datasets. Replication timing inference for the in-house datasets is described separately (see the “validation of rtQTLs” section below). We first filtered out genomic windows for which the DNA copy number is influenced by mapping inaccuracies, technical outliers or regions likely to be CNVs and/or SVs. Specifically, we kept windows with a >1% MAF in the 1000 Genomes European individuals; (2) overlapped short CNVs (median size: 3.51 kb) identified in the directly analyzed cell lines (applicable for Merkle_batch1 only, Merkle et al., submitted39). Filters (2) and (3) remove polymorphic copy number regions that could both confound the inference of replication timing as well as potentially lead to the false association of SNPs with CNVs/SVs instead of true replication timing
variants. (4) There were still genomic regions remaining with consistently extreme copy number values even after exclusion of mapped CNVs. These regions typically correspond to genomic regions containing structural variations, segmental duplications, or other genomic regions with poor reference sequence continuity. To identify these windows, we calculated the median value (across samples) of each genomic window. We then used the median of these values to represent the “common” copy number of the genome. Any window with a median of more than 0.4 copies above or below the 25%/75% percentile of peak windows was considered to be a subclonally duplicated or deleted (sub-integer change in copy number over a large region). These criteria were implemented to further remove outlier data points. Specific parameters for criteria 4 and 5 (as well as the filtering steps described below) were chosen based on extensive testing and are optimized to be the least disruptive to the data as possible, learning towards stringency. Altogether, 28,769 data windows (11.0% of all windows) were removed, leaving 232,027 windows after filtering for Merkle_batch1. For the HipSci dataset, 239,516 windows remained.

We further removed, in individual cell lines, genomic windows that were copy number outliers in specific samples (rather than across all samples). We removed data points that were at least 0.6 copies above or below the common copy number (see above), or at least 0.25 copies above or below the median copy number of that replication timing window, in any particular sample. Together, these two parameters ensured the efficient filtering of absolute or relative outliers, respectively. We also removed data points that were in a large CNV (median size: 3.02 Mb) identified in that individual (applicable for Merkle_batch1 only). These CNVs are predominantly of cell culture and not shared with other samples, thus the reason to remove them from the applicable samples rather than from the data in general. In addition, we used correlations and autocorrelations to identify samples in which an entire chromosome or chromosome arm appeared to be subclonally duplicated or deleted (sub-integer change in copy number over a large region). These criteria were implemented to further remove outlier data points. Data after the above filtering steps are referred to as “filtered raw data”. We provide a visual demonstration to clarify and justify the parameter choices (Supplementary Fig. 1a). Our filtering criteria are effective in removing CNVs and outlier data points, resulting in DNA copy number profiles optimal for replication timing inference.

Processing of the X chromosome data was performed separately for males and females. For males, because they only carry one X chromosome, all the thresholds above were divided by two. The filtered raw data was further normalized to Z-score (i.e., autosomal mean of zero and standard deviation of one) by subtracting the mean then dividing by the standard deviation of all data points, and smoothed using a penalized smoothing spline using the {R}pdeni package with smoothing parameter $10^{-16}$ (chosen after testing multiple parameters to be the most representative of the underlying raw data; this parameter is also consistent with previous papers). For each chromosome, we smoothed across gaps only if the gaps were shorter than 300 kb. Continuous genomic segments (between gaps) that were smaller than 300 kb were removed from further analysis, since smoothing of these short segments was sensitive to individual outlier data points and to edge effects. Data after the above normalization and smoothing is referred to as “smoothed data” (Fig. 1a) and was used in further analyses. The total length of replication timing windows in the smoothed data was 2330.66 Mb for autosomes (referred to as the “analyzable genome”), 121.15 Mb for the X chromosome in females, and 121.19 Mb for the X chromosome in males. For analyses involving the analyzable genome, only autosomal rtQTLs were counted.

For correlation calculations involving sib pairs vs. non-sib pairs (Merkle_batch1) and cell lines derived from the same donor vs. different donors (HipSci), we used replication timing data from all autosomes. The Wilcoxon rank-sum test was used to assess significance. For the analysis regarding IBID segments in sib pairs, we first filtered pairwise IBID using TRUFFLE, then binned the IBID segments into 2.5 Mb regions. The purpose was to minimize bias in correlation estimation because of variable IBID segment sizes. We calculated pairwise correlation in these regions, and assigned the estimate to one of three groups (IBD 0/1/2). ANOVA was used to assess the significance of the difference in average correlation among IBD 0/1/2 groups. For all box plots in this study, the center line represents median, box limits represent the first and third quartile, and the whiskers represent the maximum and minimum. Outliers as determined by the R boxplot function were not plotted. The correlation was calculated using the corr function in the Statistics::Basic::Perl module. Pearson correlation of replication timing among samples and with previous Repli-Seq measurements were calculated in R. The R package was used for variance component analyses. Linear mixed-effects model was fit for each replication timing window, with the following variables: donor, sex, age, and technical factors (Pluritest pluripotency score and novelty scores, culturing media, date of derivation, and method of derivation). When assessing the effect of rtQTLs, we also included rtQTL genotype in the model, and partitioned the donor effect into two parts: those explained and those not explained by rtQTLs. Of note, all statistical tests used in this study were two-sided unless otherwise noted. Details of statistical methods used in this paper are listed in Supplementary Table 2.

Identification of replication timing peaks. We identified peaks in the Merkle_batch1 dataset. For each sample, peaks were identified in the replication timing profile as local maxima. Peaks across all samples were then clustered using agglomerative hierarchical clustering in MATLAB (functions linkage and cluster). All resulting peaks contained a distance of 4 Mb, which yields a list of peak clusters, each containing one or more peak locations. When a cluster contained multiple peaks from the same sample, the peak closest to the cluster center was retained and all other peaks from the sample were dropped. For each peak cluster, the boundary was defined as the full range of peak locations in this cluster for hESCs, and the 5th to 95th percentile of peak windows for iPSCs (to account for the more common presence of outliers in the iPSC dataset). We only used peak clusters that contained peaks from more than 10% of the samples.

Identification of replication timing variants. We searched for replication timing variants using the Merkle_batch1 and HipSci datasets. We expect genomic regions with significant replication timing variation among samples to have greater standard deviation (SD) in their replication timing across samples. We calculated the SD of replication timing across the genome, excluding windows spanning genome gaps, and identified local peaks in the smoothed SD data. Peaks with SD values greater than the genome-wide SD mean were called as replication timing variants. To evaluate the significance of these tentative variant regions, we performed pairwise t-tests on raw replication timing data for all pairs of samples in 500 kb windows centered at the called variant peaks. We required variants to be significant at the 0.05 levels after Bonferroni correction for the number of sample pairs and the number of variant regions tested.

In a second step, we extended each replication timing variant until cross-individual replication timing variation fell below significance. Specifically, we tested adjacent 200 kb windows, sliding 100 kb at a time, extending until there were less than 0.1% significantly different cell line pairs. In cases where replication timing variants overlapped each other after extension, they were merged if the correlation of replication timing across samples at the SD peaks was greater than 0.9, and otherwise separated at the SD valleys (local minima) between peaks. Finally, we removed replication timing variants on chromosome 19 due to a high false-positive rate related to the extreme GC content and therefore noise in the data on this chromosome. We further removed variants driven by no more than 1% of the samples.

Data processing prior to rtQTL mapping. Sample selection. We performed principal component analysis (PCA) on the genotypes of the hESCs lines, using the 1000 Genomes Phase 3 European, East Asian, and African samples as references. Eight samples appearing to have non-European ancestry (admixed or East Asian) were removed from rtQTL mapping, leaving 108 individuals for further analysis. PCA was performed using the SNPRelate package in {R}. We also performed PCA with the HipSci dataset, and confirmed that all samples were of European ancestry. A total of 192 unrelated samples in the HipSci dataset were used for rtQTL mapping. While we kept sib pairs in the ESC dataset, all rtQTLs in ESC were reproducible (at nominal $p < 0.05$) when using only unrelated samples.

Genotype imputation. Imputation was performed with IMPUTE267 using the 1000 Genomes Project Phase 3 reference panel and default parameters. Variants with minor allele frequency (MAF) ≤ 1% in Europeans or Americans were not used for imputation. Imputed variants with average genotype probability ≥ 80% were used in subsequent analyses.

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

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

rtQTL mapping. Selection of phenotype PCs in rtQTL mapping. We followed the eQTL mapping framework used in the GTEx Project (https://gtexportal.org/home/documentationPage) to map rtQTLs. We included the genotype (first three, similar to GTEx) and phenotype (first k) PCs in rtQTL mapping to account for genetic and non-genetic confounding factors, respectively. To find the optimal k, we tested each integer from 1 to 40. We consider the optimal k as the one leading to the highest number of windows harboring rtQTLs identified in rtQTL mapping. In this analysis, permutation parameter “permut 50 500” was used in fastQTL. Windows level $p$ values were calculated, and the R package “jenner” was used to identify windows harboring rtQTLs at the 200 kbp for ESC and iPSCs. This resulted in 24 and 22 selected as the optimal k for ESC and iPSC rtQTL mapping analysis, respectively, which was used in all subsequent rtQTL mapping analyses.
Cis-rtQTL mapping using fastQTL. We implemented a two-step approach to map rtQTLs using fastQTL. We generally restricted our analysis to cis-rtQTLs, defined as a 1 Mb upstream or downstream of the center of each tested replication timing window. The 1 Mb distance threshold is standard in eQTL mapping studies and allows capturing more distal cis effects (Fig. 1m). The first three genotype PCs and first 24 or 22 (for ESC and iPSC, respectively) phenotype PCs were included as covariates.

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

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

Evaluation of inference of rtQTL mapping. To ensure that the computed variant-level p-values were not inflated, we calculated inflation factor with the Genomic Control method. We selected 200 windows (100 selected from windows carrying putative rtQTLs, and the other 100 randomly selected from the rest of the genome) and computed their association with genome-wide variants. We obtained variant-level statistics (which follows \( \chi^2 \) distribution under the null hypothesis) and computed the ratio of their median to the median of \( \chi^2 \) (0.456) as the genetic inflation factor. We calculated a genomic inflation factor (\( \lambda \)) as 1.03 and 1.00 for the ESC and iPSC dataset, respectively, thus the nominal \( p \)-values were not inflated; this was also supported by quantile-quantile plots.

Identification of rtQTLs. The following procedure was used to identify discrete rtQTLs, i.e., independent (not in LD) association signals, based on the significant SNPs mapped using the aforementioned two-step approach. For clarity, we denote independent association signals as rtQTLs, each of which contains multiple SNPs that are part of the association signal.

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

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

Since nearby replication timing windows are highly correlated, the same rtQTL can be detected across multiple windows. We consolidated association signals detected in different windows if they satisfy all of the following three criteria: (1) the tag variants are in LD (\( r^2 \geq 0.2 \)), (2) the replication timing windows are correlated (\( R^2 \geq 0.1 \)), and (3) the distance between the windows is less than 2 Mb. In addition to separating rtQTLs by LD, we subsequently performed conditional association for each identified rtQTL. For a given rtQTL, we tested the association between each gene X chromosome variant in this rtQTL and replication timing using fastQTL. We included as covariates the top genetic variant of this rtQTL as well as genotype and replication timing PCs. If a genetic variant had \( p < 0.05 \) after Bonferroni correction (based on the number of rtQTL genetic variants in the given rtQTL), this rtQTL was divided into multiple rtQTLs, each representing an independent association signal. rtQTLs identified using both approaches (LD and conditional analysis) were reported.

We attempted wide ranges of values for the thresholds used in this section, and found that the rtQTL results were highly robust to the choice of thresholds. We used three parameters to define independent regions and genetic variants. We considered two SNPs in LD if their \( r^2 \) was greater than \( a \) (set to 0.2), and we considered a rtQTL-associated replication timing windows if their correlation \( R^2 \) was greater than \( b \) (set to 0.1) and they were no more than \( c \) (set to 2 Mb) apart. We first fixed parameter \( c = 2 \) Mb and tested various combinations of \( a \) and \( b \) (0.1, 0.2, 0.5, and 0.8). As shown in Supplementary Table 3, the proportion of the genome associated with ESC-rtQTLs, the average size of rtQTL associated regions, and the mean number of rtQTLs at multi-rtQTL regions were highly robust to the choices of \( a \) and \( b \). We then fixed \( a = 0.2 \) and \( b = 0.1 \), and tested various choices of \( c \) (0.5, 1, 2, 3, 4, 5 Mb). We did not test \( c < 0.5 \) Mb because replication timing windows show high correlation (\( r > 0.6 \)) when they are less than 0.5 Mb apart. As shown in Supplementary Table 4, the rtQTL results were also highly robust to the choice of \( c \).

In rtQTL mapping, we also discarded putative rtQTLs that were supported by less than two genetic variants or had the minimum \( p \)-value was greater than 10–3. These two criteria were designed to filter out very weak statistical associations (median number of variants supporting an rtQTL was 133, and median \( p \)-value for both ESC and iPSC rtQTLs was on the order of 10–5). Therefore, the use of these two filters does not result in overestimation of genomic regions associated with rtQTLs or the number of independent rtQTLs per region.

Filtering of rtQTLs. The putative rtQTLs identified were subjected to further filtering. First, we determined the boundaries of regions that significantly associated with each putative rtQTL. Starting at the window that most strongly associated with the tag variant (i.e., with the lowest \( p \)-value) of an rtQTL, we extended the associated regions to include five windows (50 kb or less). We refer to this region as the “associated region”.

Next, we filtered false positives suspected to be potentially caused by short CNVs. During data processing (described above), we removed windows in which copy number measurement are potentially influenced by CNVs. However, short CNVs, specifically spanning only one or two windows, may still be filtered and could lead to false-positive rtQTLs (if they are in LD with SNPs). This type of false positive was identified by utilizing the raw unsmoothed data as follows: if a putative rtQTL is a false positive caused by a CNV, the association signal would be (1) observed in a small number of unsmoothed raw windows (with CNV), and (2) will be more strongly associated with the raw data than with the smoothed data (in which the CNV will be smoothed within a broader region, thus decreasing association). We computed the association \( p \)-values of the tag variant of each rtQTL with the (1) smoothed data within the associated region, (2) filtered raw data within the associated region, and (3) the raw unsmoothed data from the raw dataset, the current dataset was used for LD calculation. For calculations involving LD from 1000 Genomes Phase 3 Europeans was used whenever available. For SNPs not called in the 1000 Genomes dataset, the current dataset was used for LD calculation.

In total, we identified 608 ESC rtQTLs, among which 603 were on autosomes and five were on the X chromosome in males. No rtQTLs were found on the X chromosome in females. This was not due to the reduced number of individuals tested, but likely resulted from (1) data that less structured during data timing processing. To mitigate this, we further required (2) that the minimum \( p \)-value of such removed data must be higher, or no more than one order of magnitude lower, than that of the smoothed data. This further removes CNVs that are minimized by smoothing but can still be present and lead to false associations (which would be stronger for unsmoothed data). We found that a cutoff of one order of magnitude of significance performs well without removing non-CNv data.

Finally, we sought to remove very weak rtQTL associations that are less likely to be true. We therefore required that the minimum \( p \)-value from the raw data must be less than 0.01 and that the associated region must be larger than one replication timing window.
respectively). The early- and late-replicating alleles were equally likely to be the reference allele (binomial p = 0.53), thus rtQTL mapping was not influenced by reference allele selection. We also used three additional datasets to validate ESC rtQTLs. The first dataset contains 9 hESCs in Merkle_batch2 and the 8 hESCs in Merkle_batch1 that were excluded in rtQTL mapping due to ancestry. The second and third datasets are the in-house hESC and iPSC dataset, respectively.

For the first dataset, validation was performed in fastQTL. Validation using the second and third datasets were performed in MATLAB by calculating the Pearson correlation p-value between the strongest rtQTL genetic variant and the replication timing locus with the strongest association in the discovery set. We tested rtQTLs of which the top genetic variant was polymorphic and had all three genotypes in the discovery dataset. rtQTLs excluded if the alternative allele of the top genetic variant in the validation dataset was not consistent with that of in the discovery set. This left 427 regions that could be tested in the third dataset, and 396 regions in the fourth dataset. Replication timing of these two datasets were inferred using GenomeSTRiP (as described above) in 2.5 kb windows of uniquely alignable sequence29. For each sample, windows with copy number > 3 or < 1 were removed. We used a segmentation algorithm (segment in MATLAB) to further remove outlier data points (segments with mean > 2.45 or < 1.55 were removed). The data were then smoothed using a cubic smoothing spline with parameter 10−17.

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

SMARD validation analysis was carried out as previously described23. Briefly, cells were pulse-labeled sequentially with 25 μM IdU and CldU. The cells were then embedded in 1% IncCert agarose and lysed. The remaining embedded genomic DNA was digested with restriction endonucleases. Pulsed-field gel electrophoresis (PFGE) was used to separate DNA according to size. The segment containing the locus-of-interest was identified with Southern blot and the gel slice was excised. Agarose was then melted, and individual DNA strands were stretched on silanized glass slides. Immunostaining was employed to detect the halogenated nucleotides in the replicated DNA. Biotinylated FISH probes were used to identify DNA molecules containing the locus-of-interest.

Multi-rtQTLs. To identify multi-rtQTL regions, we considered separate rtQTLs to be associated with the same region if the replication timing locus most strongly associated with them were correlated (R2 ≥ 0.2) across individuals, were in physical proximity (<2 Mb apart), and each provided additional explanatory power for replicate timing. Secondary rtQTLs were either not in the primary ones (130 and 265 multi-rtQTL regions in the ESC and iPSC dataset, respectively), or provided additional explanatory power despite being in LD (5 cases in ESC and 10 cases in iPSC).

We found that the distance threshold (2 Mb) used in multi-rtQTL identification was robust. As described above, we discovered 135 multi-rtQTL regions in the hESC dataset. We tested various values for d (0.25, 0.5, 1.2, 3, 5, 10 Mb), and found that the number of multi-rtQTL regions identified was stable to change in d (Supplementary Table 5). Specifically, when we set d = 10 Mb, though we identified multi-rtQTL regions with up to eight rtQTLs, there was only a modest increase in the average number of rtQTLs per multi-rtQTL region (12%) and total number of multi-rtQTL regions (11%).

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

We examined the possible interaction between primary and secondary rtQTLs in regions with two, three, and four rtQTLs. If there was no interaction, we expect that the replication timing in these regions will be positively linearly correlated with the sum of replicate timing across the alternative (additive) model. As described above, we found 1564 rtQTLs associated with 31.4% (731.1 Mb) of the genome. These estimates were very similar to the estimates using the original approach (1617 rtQTLs, 31.8%/741.8 Mb), which suggests that the approach described above was appropriate and did not overestimate the effects of rtQTLs.

Factors driving the increased rtQTL calling ability. We identified 1,617 rtQTLs in hESCs and iPSCs, which was two orders of magnitudes more than our previous study. Several factors contribute to the much larger number of rtQTLs identified in this study.

First and foremost, the quality of the data is much greater in the current study. This is mostly driven by the deeper sequence data (~30x sequence coverage, compared to 2–4x coverage in our 2014 study). In addition, ESCs and iPSCs are more proliferative than LCLs, leading to stronger replication signal in the sequence data.

To demonstrate the effect of data quality on rtQTL mapping, we artificially added noise to the iPSC replication timing data from this study to match the level of noise in our 2014 study. We did this for chromosomes 2, 3, and 4 as a representative data subset and evaluated noise using autocorrelation of raw replication timing data along chromosomes. After adding noise (Supplementary Figure 1c), the number of rtQTLs identified on the three chromosomes dropped substantially, from 282 to 17 (~6%).

A second critical factor enabling the enhanced rtQTL identification in the current study is our improved analytical framework. In our previous study, we were very strict in identifying only the strongest replication timing variants, and we only performed rtQTL testing on those. In contrast, we aimed here to identify as many rtQTLs as possible under the statistical power constraints. Limiting our current search to the same criteria we used in the previous study led to a 3.8-fold decrease in the number of rtQTLs identified (from 603 to 158 for hESC rtQTLs). Other factors contributing to the increased number of identified rtQTLs are an improved two-stage rtQTL mapping approach, and more optimal calibrations of principal component (PC) regression for replication timing and genotype.

Validation of rtQTLs. To validate the iPSC rtQTLs, we examined their reproducibility in the Merkle_batch ESC dataset (108 European ancestry samples only). Validation was performed using fastQTL20 by testing the association between the strongest rtQTL SNP and the replication timing locus closest to the locus with the strongest association in the discovery set. In Supplementary Table 3, the alternative allele of the top rtQTL SNPs was not available in the validation dataset (Merkle_batch ESCs), an rtQTL SNP from the same rtQTL that has p value less than two orders of magnitude higher than that of the strongest rtQTL SNP was used instead. We found that the -log10(p-values) of rtQTL mapping in the discovery and validation datasets (Pearson r = 0.75, p = 1.28×10−176). We then repeated this analysis in the opposite direction (validate ESC rtQTLs using HipSci iPSCs) and obtained similar results (Pearson r = 0.76, p = 7.81×10−112). These observations support that the rtQTLs identified in this study are highly reproducible.
individuals carrying zero or four early-replicating alleles were estimated from actual data. We then compared –2x(log likelihood ratio) with the chi-squared distribution with freedom of 2 to obtain a p-value. We also tested for interactions between pairs of genetic variants in affecting replication timing using a linear model, explicitly modeling main and interaction terms. Genotypes were coded as -1, 0, and 1 for this model. P-values for the interaction term were extracted, followed by Bonferroni correction. We then applied this test to the discovered rQTLs. We then applied this test to replication timing windows and genetic variants (in cis within 1 Mb of a given replication timing window) genome-wide. To reduce the number of tests, we pruned replication timing windows and genetic variants (in applied this test to the discovered rtQTLs. We then applied this test to replication timing using a linear model, explicitly modeling main and interaction terms. Genotypes were coded as -1, 0, and 1 for this model.

Data sources
Epigenetic enrichment analyses
Enrichment calculations were excluded from this analysis. Enrichment calculations were performed on overlapping with the feature, and whether this is more or less likely (i.e., enriched or down-regulated) than the actual feature. Overlapping enrichment calculations necessarily appear in the exact same location, unless the enrichment calculation and the actual feature overlap. Statistical significance was assessed using the one-tailed binomial test. The binomial parameter \( p \) was estimated from 100 random permutations, from which we estimated the probability of random SNPs (matched with the rQTLs, see details below) overlapping with the feature. Correction for multiple testing was applied when multiple features from the same category (e.g., histone marks) were tested. For each rQTL, we searched for random SNPs that match the characteristics of the tag variant of the rQTL (denoted as “actual tag variant”) and used the matched variants (“matched tag variants”) to tag the random sets of SNPs used in permutations. We required that the matched tag variants must be at least 2 Mb away from the actual tag variant. The matched tag variants must have satisfied all three following criteria: (1) have similar minor allele frequency (<5% difference), (2) have similar distances to the nearest replication initiation site and terminus (<50 kb difference), and (3) have similar replication timing (<0.5 units difference) with the actual tag variant. We require the matched tag variants to have the same number of nearby peaks (LD \( r^2 \geq 0.2 \)) than that of the other epigenetic features. With this method, we identify me3achypers regions that overlap with an initiation site (i.e., the interval between boundaries of the initiation sites). If a me3achyper region does not overlap with any initiation site, its physical distance was calculated as the distance to the nearest initiation site boundary.

Using epigenetic features to predict replication initiation site locations
Identification of epigenetic feature combinations. To identify combinations of chromatin marks enriched at rQTLs, we used a stepwise, iterative approach. The hESC rQTLs and epigenetic data were used. We considered 29 histone marks (Supplementary Fig. 4c) and also included H2A.Z, DNase I hypersensitivity, and binding sites of 51 TFs and other DNA binding factors (referred to as TFs for simplicity). First, we tested each individual epigenetic feature (histone mark or TF) to identify features that are enriched at rQTL SNPs. Enrichment was examined using the same permutation-based approach described above. The only difference was that each rQTL individual SNP was considered independently (as opposed to being considered together with other SNPs assigned to the same rQTL). As our goal was to identify co-occurrence of epigenetic features at the same exact genomic locations. Statistical significance was assessed using Fisher’s exact test. We corrected for multiple testing at 5% FDR using the Benjamini-Hochberg correction.

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

To identify me3achyper regions, we identified regions that carry one of the 13 histone marks and regions that overlap with peaks at least 11 variable acetylation marks. We merged me3achyper regions that co-occurred within 10 kb. In Supplementary Fig. 5e–g, the position of initiation sites found in >10% of the samples were determined based on local maxima in the averaged replication timing profile. When calculating distances (fractional and physical), the distance was set to zero for me3achyper regions that overlap with an initiation site (i.e., the interval between boundaries of the initiation sites). If a me3achyper region does not overlap with any initiation site, its physical distance was calculated as the distance to the nearest initiation site boundary.

Receiver operating characteristic (ROC) curves. To obtain ROC curves, we randomly selected 1000 replication timing windows at replication timing peaks as positives, and another 1000 windows far (>750 kb) from the nearest peak as negatives. The distance threshold was in place to ensure that the selected negative windows would not be too proximal to peaks. We then performed prediction on these 2000 windows. If a window was within cutoff distance of a me3achyper region (in LCL, we considered the top three combination marks [H3K4me3-H3K9me3-H3K36me3]), we predict it as “peak”. Otherwise, we predict it as “not peak”. Then we compared the predictions with the truth and counted true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) windows. We calculated the true positive rate as TP/(TP+FN) and the false positive rate as FP/(TN+FP) and constructed the ROC curve. For permutations, we randomly shifted the location of the me3achyper regions between 1 Mb and 2Mb and obtained ROC curves and AUCROC based on these random intervals.

Replication initiation site prediction in LCLs and iPSCs. Following best practices, we used the me3achyper histone modifications, which were discovered using hESC rQTLs, to predict replication initiation sites in LCLs and iPSCs. For predictions, we randomly shifted the location of the me3achyper regions between 1 Mb and 2Mb and obtained ROC curves and AUCROC based on these random intervals.

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-27115-9 | www.nature.com/naturecommunications
discovered based on rtQTLs but then shown to be predictive of peak locations in general, peaks with rtQTLs were excluded in the analysis of prediction performance (Fig. 3). We assessed the generalizability of the replication initiation histone modifications in LCLs and iPSCs. LCL is a cell type that has a distinct epigenetic and replication landscape from hESC lines⁷⁴–⁷⁶, and iPSCs have similar but not identical to replication timing profiles in hESC lines (r = 0.90). Replication timing profile for the GM12878 LCL and 192 unrelated iPSCs were inferred from whole-genome sequencing data⁷⁰,⁷⁸. In ESC and iPSC, when calculating the physical distance of predicted initiation sites to actual initiation sites, we defined peak boundaries using a clustering-based approach (see above). In LCL, because of the limited number of peaks we defined initiation site boundaries as 100 kb upstream and downstream of the local maxima in the replication timing profiles. For all distance calculations (ESC, iPSC, and LCL), we excluded replication timing peaks that overlap rtQTLs in any cell type, such that these calculations were not biased by the locations used for the discovery of the histone modifications. Data for HSK4m1e3, H3K4m3e2, and H3K36me3 for 18 LCLs were obtained from⁷⁷, and merged with the ENCODE data. Imputed acetylation data for the GM12878 was from the Roadmap Epigenomics Project⁷². Histone mark data for five iPSCs was from the Roadmap Epigenomics Project⁷². If a histone modification combination location was found in one cell type (either hESC or LCL), but no combination location was found within 100 kb in the other cell type, we denoted this region as cell-type-specific. Otherwise, this region was denoted as “shared” between the two cell types.

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

We aggregated information from all rtQTL SNPs, except those that are monomorphic in the seven cell lines. We assigned each cell line by genotype to one of two categories, i.e., early-replicating, heterozygous, and late-replicating, at each rtQTL SNP. For each epigenetic feature, we tested whether the cell lines carrying the early-replicating genotypes are more (or less) likely to harbor it than the cell lines carrying the late-replicating genotypes, using the two-tailed binomial test. The binomial parameter \( p \) was calculated as \( \frac{\text{Perm}_\text{early} - \text{Perm}_\text{late}}{\sqrt{\text{Perm}_\text{early} \times (1 - \text{Perm}_\text{early})}} \), where \( \text{Perm}_\text{early} \) is the proportion of late-replicating genotypes carrying this feature, and \( \text{Perm}_\text{early} \) and \( \text{Perm}_\text{late} \) are estimated from ten permutations (described below). Bonferroni correction was used to correct for multiple testing at the 0.05 level. In each permutation, we used random SNPs matched for rtQTLs (for details see the “enrichment analyses” section), and randomly designated one genotype as the early-replicating genotype. We obtained genotype and epigenetic information from the seven cell lines at these random SNPs and calculated the proportion of early- and late-replicating genotypes carrying the feature in ten permutations (\( \text{Perm}_\text{early} \) and \( \text{Perm}_\text{late} \)).

Transcription factors. To identify TFs that regulate replication timing, we tested whether the early-replicating alleles (the CAVIAR 90% causal set) influence chromatin structure and replication across the entire beta-globin locus. We used this principle to identify TFs associated with replication timing. We tested the motifs of all TFs studied in Supplementary Table 16. We used the motif discovery pipeline in CAVIAR, Version 2.2 (available at https://www.thekorenlab.org/CAVIAR) to identify TFs that are enriched in the full or partial gene body of the beta-globin locus. These TFs were tested for enrichment in the beta-globin locus using the RankProd package (www.r-project.org). The results were confirmed by searching the factor binding predictions using the Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/; ENCODE data were from https://www.encodeproject.org/). LCL histone modification data were from Kassowski et al. (ref. 47; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GE850883).

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

Received: 17 August 2021; Accepted: 29 October 2021; Online published: 19 November 2021

References
1. Prioleau, M.-N. & MacAlpine, D. M. DNA replication origins—where do we begin? Genes Dev. 30, 1637–1649 (2016).
2. Hamlin, J. L., Mesner, L. D. & Dijkwel, P. A. A winding road to origin discovery. Chromosoma 18, 45–61 (2010).
3. Aladjem, M. I. & Redon, C. E. Order from clutter: selective interactions at mammalian replication origins. Nat. Rev. Genet. 18, 101 (2016).
4. Forrester, W. C. et al. A deletion of the human beta-globin locus activation region (CAVIAR 90% causal set) influences altered chromatin structure and replication across the entire beta-globin locus. Genes Dev. 4, 1637–1649 (1990).
5. Liu, G., Malott, M. & Leffak, M. Multiple Functional Elements Comprise a Mammalian Chromosomal Replicator. Mol. Cell. Biol. 23, 1832 (2003).
6. Aladjem, M. I., Rodewald, L. W., Kolman, J. L. & Wahl, G. M. Genetic Dissection of a Mammalian Replicator in the Human β-Globin Locus. Science 291, 1005 (1998).
7. Altman, A. L. & Fanning, E. The Chinese Hamster Dihydrofolate Reductase Replication Origin Beta is Active at Multiple Ectopic Chromosomal Locations and Requires Specific DNA Sequence Elements for Activity. Mol. Cell. Biol. 21, 1098 (2001).
8. Aladjem, M. I. et al. Participation of the Human β-Globin Locus Control Region in Initiation of DNA Replication. Science 270, 815 (1995).
9. Kajita, R. F. et al. Distal Sequences, but Not ori-B/DBR-1, Are Essential for Initiation of DNA Replication in the Chinese Hamster DHFR Origin. Mol. Cell. Biol. 2, 797–806 (1998).
10. Wang, L. et al. The Human β-Globin Replication Initiation Region Consists of Two Modular Independent Replicators. Mol. Cell. Biol. 24, 3373 (2004).
11. Sima, J. et al. Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. Cell 176, 1–15 (2019).
12. Conner, A. L. & Aladjem, M. I. The chromatin background of DNA replication: Lessons from genetics and genome-scale analyses. Biochem. Biophys. Acta 1819, 794–811 (2012).
13. Sahu, C. G. et al. A mechanism for epigenetic control of DNA replication. Genes Dev. 32, 224–229 (2018).
29. Cayrou, C. et al. New insights into replication origin characteristics in Drosophila follicle cells. *Nature 430*, 372–376 (2004).
30. Vogelauer, M., Rubbi, L., Lucas, I., Brewer, B. J. & Grunstein, M. Histone Acetylation Regulates the Time of Replication Origin Firing. *Mol. Cell 10*, 1239–1249 (2002).
31. Rizzardi, L. F., Dorn, E. S., Strahl, B. D. & Cook, J. G. DNA Replication Origin Function Is Promoted by H3K4 Di-methylation in Saccharomyces cerevisiae. *Genetics 192*, 371 (2021).
32. Koren, A. et al. Genetic variation in human DNA replication timing. *Cell 159*, 1015–1026 (2014).
33. Leeke, F. T. & Struhl, K. Biological insights from the whole genome analysis of human embryonic stem cells. *bioRxiv*, 2020.2020.2026.337352, https://doi.org/10.1101/2020.11.10.2026.337352 (2020).
34. Mishra, S. et al. Cross-talk between Lysine-Modifying Enzymes Controls Site-Specific CTCF Binding and Replication Origin Activation During Early Human Embryogenesis. *Genome Biol. 21*, 76 (2020).
35. Hormozdari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics 198*, 497–508 (2014).
36. Gehrhardt, U. et al. Co-acting DNA sequence at a replication origin promotes repeat expansion to fragile X full mutation. *J. Cell Biol. 206*, 599–607 (2014).
37. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature 550*, 204–213 (2017).
38. Long, H. et al. H3A.Z facilitates licensing and activation of early replication origins. *Nature 577*, 576–581 (2020).
39. Jakob, C. R., Muñoz, S. A., Olcenhock, B. A. & Blobel, G. A. Histone H3 Lysine 9 Methylation and HP1gamma Are Associated with Transcription Elongation through Mammalian Chromatin. *Mol. Cell 19*, 381–391 (2005).
40. Bernstein, B. E. et al. A Bivalental Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell 125*, 315–326 (2006).
41. Cayrou, C. et al. Genome-scale analysis of meiotic recombination reveals. *Mol. Reprod. Dev. 81*, 1324–1328 (2014).
42. Mishra, S. et al. Cross-talk between Lysine-Modifying Enzymes Controls Site-Specific DNA Amplifications. *Cell 174*, 803–817 (2018).
43. Liang, Z., Diamond, M., Smith, J. A., Schnell, M. & Daniel, R. Proliferating cell nuclear antigen is required for loading of the SMCC/KMD3C histone dimethylase onto chromatin. *Epigenetics Chromatin 4*, 18 (2011).
44. Kasowski, M. et al. Extensive Variation in Chromatin States Across Humans. *Science 342*, 750–752 (2013).
45. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA 100*, 9440–9445 (2003).
46. Churchill, G. A. (2004). *Statistics and data analysis for the life sciences* (Springer, New York) 131–134 (2014).
47. Eberle, M. A. et al. A reference data set of 5.4 million phased human variants revealed by sequencing a three-generation 17-member pedigree. *Genome Res. 27*, 157–164 (2017).
48. Bailey, S. D. et al. ZNF143 provides sequence specificity to ensure chromatin state transitions at gene promoters. *Nat. Commun. 6*, 6186 (2015).
49. Liang, Z., Diamond, M., Smith, J. A., Schnell, M. & Daniel, R. Proliferating cell nuclear antigen is required for loading of the SMCC/KMD3C histone dimethylase onto chromatin. *Epigenetics Chromatin 4*, 18 (2011).
50. Khare, S. P. et al. Histone–a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res. 40*, D337–D342 (2012).
51. Zhao, J., Li, X., Guo, M., Yu, J. & Yan, C. The common stress responsive transcription factor AFX3 binds genomic sites enriched with p300 and H3K27ac for transcriptional regulation. *BMC Genomics 17*, 335 (2016).
52. Hwang, J.-Y. & Zukin, R. S. REST, a master transcriptional regulator in neurodegenerative disease. *Curr. Opin. Neuropathol. 48*, 193–200 (2018).
53. Bailey, S. D. et al. ZNF143 provides sequence specificity to ensure chromatin state transitions at gene promoters. *Nat. Commun. 6*, 6186 (2015).
54. Nora, E. P. et al. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell 169*, 930–944 e922 (2017).
55. Mills, R. E. et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res. 16*, 1182–1190 (2006).
56. Yamada, M. et al. An integrated map of structural variation in 2,504 human genomes. *Nature 526*, 75 (2015).
57. Yang, M. et al. Human oocytes reprogram adult somatic nuclei of a type 1 diabetic to diploid pluripotent stem cells. *Nature 510*, 533–536 (2014).
58. Pauli, D. et al. Nuclear genome transfer in human oocytes eliminates mitochondrial DNA variants. *Nature 493*, 632–637 (2013).
59. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarthy, S. A. Discovery and genotyping of genome-wide variants by sequencing on a population scale. *Nat. Genet. 43*, 269–276 (2011).
60. Koren, A. et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet. 91*, 1033–1040 (2012).
61. Yamada, M. et al. Human oocytes reprogram adult somatic nuclei of a type 1 diabetic to diploid pluripotent stem cells. *Nature 510*, 533–536 (2014).
62. Pauli, D. et al. Nuclear genome transfer in human oocytes eliminates mitochondrial DNA variants. *Nature 493*, 632–637 (2013).
63. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarthy, S. A. Discovery and genotyping of genome-wide variants by sequencing on a population scale. *Nat. Genet. 43*, 269–276 (2011).
64. Koren, A. et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet. 91*, 1033–1040 (2012).
65. Yamada, M. et al. Human oocytes reprogram adult somatic nuclei of a type 1 diabetic to diploid pluripotent stem cells. *Nature 510*, 533–536 (2014).
66. Pauli, D. et al. Nuclear genome transfer in human oocytes eliminates mitochondrial DNA variants. *Nature 493*, 632–637 (2013).
67. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarthy, S. A. Discovery and genotyping of genome-wide variants by sequencing on a population scale. *Nat. Genet. 43*, 269–276 (2011).
68. Koren, A. et al. Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet. 91*, 1033–1040 (2012).
69. Yamada, M. et al. Human oocytes reprogram adult somatic nuclei of a type 1 diabetic to diploid pluripotent stem cells. *Nature 510*, 533–536 (2014).
70. Pauli, D. et al. Nuclear genome transfer in human oocytes eliminates mitochondrial DNA variants. *Nature 493*, 632–637 (2013).
71. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarthy, S. A. Discovery and genotyping of genome-wide variants by sequencing on a population scale. *Nat. Genet. 43*, 269–276 (2011).
Acknowledgements
This work was supported by the National Institutes of Health (DP2-GM123495 to A.K.) and the National Science Foundation (MCB-1921341 to A.K.). X.Z. was supported by a Stein Fellowship. F.T.M. is a New York Stem Cell Foundation Robertson Investigator (NYSCF-R-156), and is supported by the Wellcome Trust and Royal Society (211221/Z/18/Z), and by the Chan Zuckerberg Initiative (191942).

Author contributions
Q.D., M.M.E., A.N.B., M.L.H., Y.H., Y.T., C.J.C., S.G., R.E.H., and F.T.M. analyzed data. N.W. performed experiments. X.Z. and J.H. provided statistical expertise. K.E., J.G., D.E., A.G.C., and A.K. provided supervision. Q.D. and A.K. wrote the paper with feedback from all co-authors.

Competing interests
K.E. is a co-founder of Q-State Biosciences, Quralis and Enclear Therapies, and currently employed at BioMarin Pharmaceutical. No other authors declare competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-27115-9.

Correspondence and requests for materials should be addressed to Amnon Koren.

Peer review information Nature Communications thanks the anonymous reviewer for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2021