# High-throughput analysis of DNA replication in single human cells reveals constrained variability in the location and timing of replication initiation

Dashiell J. Massey[1], Amnon Koren[1]
[1] Department of Molecular Biology and Genetics, Cornell University, Ithaca NY 14853, USA

## Abstract

DNA replication occurs throughout the S phase of the cell cycle, initiating from replication origin loci that fire at different times. Debate remains about whether origins are a fixed set of loci used across all cells or a loose agglomeration of potential origins used stochastically in individual cells, and about how consistent their firing time during S phase is across cells. Here, we develop an approach for profiling DNA replication in single human cells and apply it to 2,305 replicating cells spanning the entire S phase. The resolution and scale of the data enabled us to specifically analyze initiation sites and show that these sites have confined locations that are consistently used among individual cells. Further, we find that initiation sites are activated in a similar, albeit not fixed, order across cells. Taken together, our results suggest that replication timing variability is constrained both spatially and temporally, and that the degree of variation is consistent across human cell lines.

## Introduction

Faithful duplication of the genome is a critical prerequisite to successful cell division. Eukaryotic DNA replication initiates at replication origin loci, which are licensed in the $G_1$ phase of the cell cycle and fired at different times throughout the S phase. In many eukaryotes, sequencing of cells at different stages of the cell cycle has been used to profile DNA replication timing, which measures the relative time that different genomic regions are replicated during S phase (reviewed in[1]). This replication timing program is highly reproducible across experiments[2], suggesting strict regulatory control; and conserved across phylogeny[3,4], suggesting selection under evolutionary constraint. However, the molecular mechanisms that determine the locations and preferred activation times of replication origins in mammalian genomes remain unclear. Furthermore, there is debate over whether replication origins are located at fixed discrete genomic sites, dense clusters of sites that fire stochastically in different cells, or even throughout the genome at sites of diffuse origin potential. In particular, some recent replication origin-mapping methods have led to speculation that human replication origins may be highly abundant and dispersed throughout the genome[1,5]. Ensemble replication timing measurements have been interpreted to indicate that replication follows broad "domains", spanning hundreds of kilobases to several megabases with consistent replication timing governed by the activity of clusters of replication origins[6,7]. In contrast, high-resolution measurements of hundreds of human replication timing profiles[8,9], or replication timing across multiple S-phase fractions[10], support initiation of replication from more localized genomic regions. While these replication timing methods reveal genomic regions that reproducibly replicate at characteristic times during S phase, it remains contested whether these represent conserved pattern across cells or reflect the average behavior of single cells. Previous work has modeled how the stochastic firing of replication origins could be sufficient to explain the replication timing profile[11,12], and single-molecule experiments (e.g. with DNA combing) have suggested that cells use different subsets of origins in each cell cycle[13,14].

Recently, replication timing has been analyzed by single-cell sequencing using up to several hundred mouse or human cells[15-17]. These studies focused on cells from the middle of S phase and analyzed replication at the level of domains, concluding that stochastic variation exists in replication timing and is highest in the middle of S phase.

However, single-molecule and single-cell studies have been limited in their throughput, and biased toward early S-phase or mid S-phase, respectively. Analyzing a large number of cells is particularly important given that even when the whole genome is captured, a single cell provides only a snapshot of DNA replication at a single moment in time. By assaying many cells at different stages of S phase, it is possible to string these snapshots together to construct a picture of replication states over time. However, the resolution of this picture will be dependent both on capturing cells at many stages of S phase and on assaying a large number of cells.

Here, we report the analysis of whole-genome sequencing of thousands of single replicating cells across nine cell lines. We successfully distinguished replicating and non-replicating cells,

allowing us to capture cells throughout the full duration of S phase without cell sorting. We find that single cells within a given cell line use a consistent set of replication initiation sites. The majority of these initiation sites are discrete genomic loci rather than megabase-scale domains. Furthermore, initiation sites are constrained in the order in which they are activated, with each initiation site firing within a limited time window during S phase. We conclude that a consistent set of replication origins firing at relatively fixed times in S phase explain the vast majority of replication initiation events in single cells, constraining the degree of spatial and temporal variation observed in replication timing.

## Results

### A high-throughput, high-resolution approach for single cell replication timing measurement

Previous sequencing-based studies measured DNA replication timing in a relatively small number of cells, mostly limited to mid-S phase cells[15,16]. To analyze single cells, these studies performed DNA amplification using DOP-PCR, which is known to yield suboptimal DNA copy number measurements[18,19]. Consequently, these studies were limited to analyzing replication timing at the level of large chromosomal domains (typically on the order of megabases). As an alternative approach, we devised a method to study DNA replication timing across the entire span of S phase, in hundreds to thousands of cells, and with higher spatial resolution than previous methods. Specifically, we used the 10x Genomics Single Cell CNV platform, a microfluidics platform for isolation, barcoding, and multiple-displacement amplification (MDA; which has been shown to be superior to DOP-PCR[18]) of single cells, which was then followed by whole-genome sequencing. As an initial proof-of-principle, we analyzed 5,793 cells isolated and sequenced from the human lymphoblastoid cell line (LCL) GM12878 following fluorescence-activated cell sorting (FACS) of $G_1$- and S-phase cells. Even at a sequence coverage of ~0.01X, replicating cells could be distinguished from non-replicating cells. Specifically, local read depth fluctuated more in replicating cell relative to non-replicating cells of similar coverage (Figure 1a). We quantified these fluctuations using the median absolute deviation of pairwise differences between adjacent genomic windows (MAPD), which scales proportionally to read depth (Methods). FACS-sorted $G_1$- and S-phase cells had distinct linear relationships between scaled MAPD and read depth (Figure 1b, left), which allowed us to computationally assign each cell as "G1" or "S" on the basis of a two-component Gaussian mixture model (Figure 1b, right). While *in silico* sorting was highly concordant with the FACS fractions, we identified ~14% of cells in the $G_1$-phase FACS fraction that were actually replicating, and reciprocally ~25% of cells in the S-phase FACS fraction that were not replicating. Thus, *post hoc*, *in silico* sorting using single cell DNA sequence data provides greater sensitivity and less between-fraction contamination than FACS.

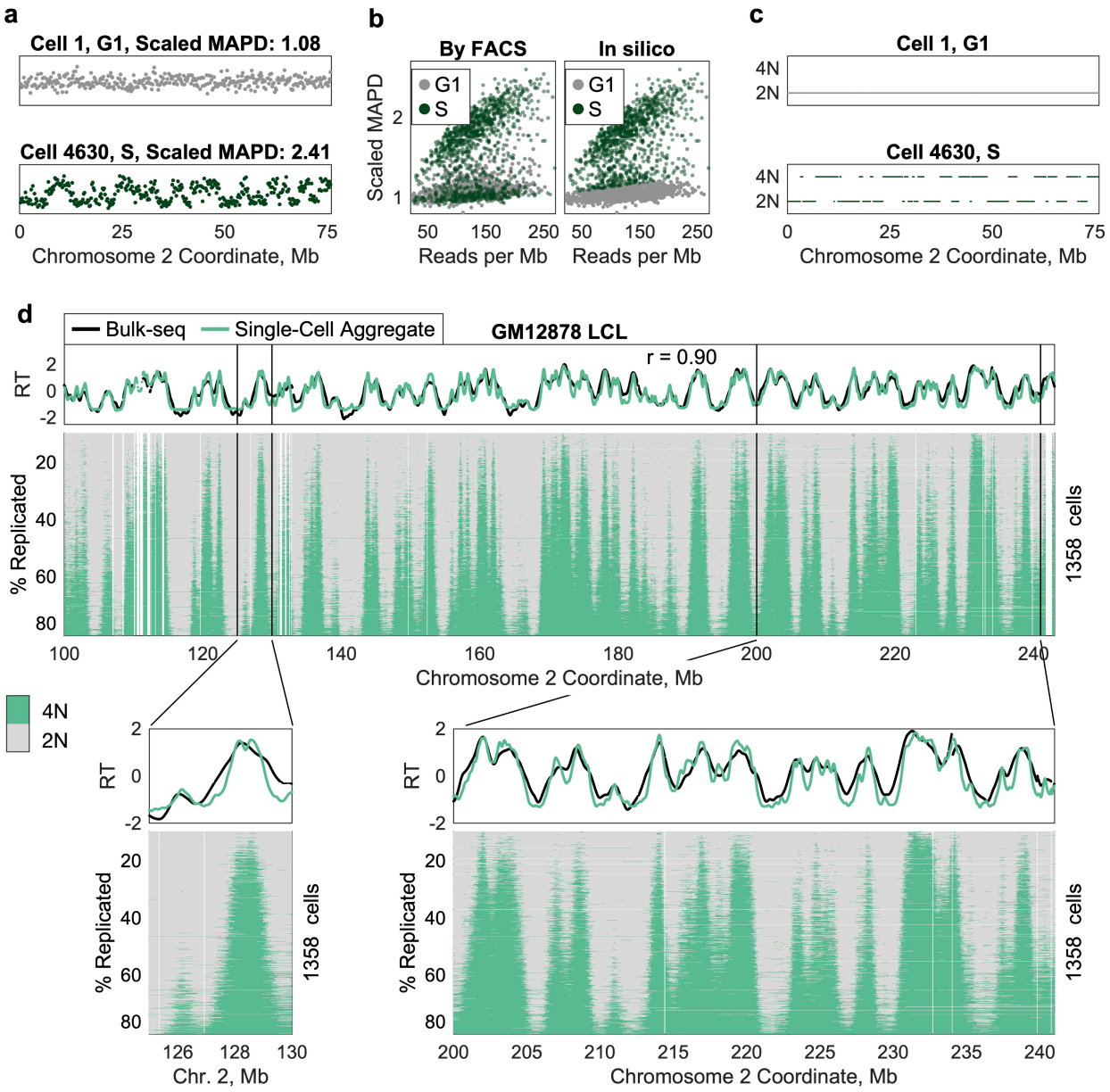


Figure 1. A method for high-throughput measurement of single-cell DNA replication timing.

**(a)** Non-replicating $G_1$ cells (*e.g.* Cell 1, *top*) have a relatively uniform sequencing read depth across the genome, whereas S-phase cells (*e.g.* Cell 4630, *bottom*) display fluctuations in read depth, consistent with the presence of two underlying copy number states. Scaled MAPD (median absolute pairwise difference between adjacent genomic windows divided by the square root of mean coverage-per-Mb) measurements affirm the classification of cells as replicating or non-replicating. Each dot represents raw read count in a 200kb window.

**(b)** Flow-sorted single cells (*left*) can be accurately sorted *in silico* (*right*). Replicating S-phase cells display a higher degree of read-depth fluctuation relative to non-replicating $G_1$-phase cells sequenced to equivalent coverage (quantified by scaled MAPD). Left panel: cells are labeled as $G_1$- (gray) or S-phase (green) based on FACS sorting. Right panel: the same cells are labeled as $G_1$- or S-phase based on scaled MAPD, revealing widespread S-phase contamination in the $G_1$ FACS sample.

**(c)** Replication profiles were inferred for each single cell, using a two-state hidden Markov model. Non-replicating cells (*e.g.* Cell 1, *top*) display a single copy number (2N), while replicating cells (*e.g.* Cell 4630, *bottom*) display two

distinct copy number states (2N and 4N). Each dot represents the inferred replication state in a 20kb window. The same region is shown from **(a)**.

**(d)** Single-cell replication profiles for 1,358 GM12878 cells (including FACS-sorted cells and unsorted cells) reveal continuous replication progression, with sharply defined and consistently replicated regions overlapping peaks in the bulk replication timing profile. Variation in activation time along S-phase progression among initiation sites also mirrors the replication timing profile. Top panel: *In silico* sorting shown in **(b)** was used to generate an S/$G_1$ aggregate replication timing profile (green) from the single-cell library, which was highly correlated to the bulk-sequencing profile (black). Bottom panel: each row represents a single cell, sorted by the percent of the genome replicated, and each column represents a fixed-size window of 20kb. Low-mappability regions and cell-specific copy-number alterations have been removed (white). Insets show smaller regions.

After assigning replicating and non-replicating cells, we used the non-replicating cells as an internal "$G_1$ control" to define variable-size, uniform coverage genomic windows that account for mappability and GC-content biases, as well as the effects of copy number variations, on sequencing read depth[20]. The ability to identify these control cells *post hoc* has two major benefits. First, sequencing biases (particularly, GC-content bias[21]) are known to vary between sequencing libraries, a concern that is alleviated by using control cells from within the same library as the cells of interest. Second, this approach enables us to capture the full span of S-phase without risking contamination of FACS fractions and without the need to define subjective sorting gates. For each genomic window in each cell, we then inferred the replication state (replicated or unreplicated) using a two-state hidden Markov model. This confirmed the uniform DNA copy number across the genome in $G_1$ cells, and fluctuating regions of replicated and unreplicated DNA in S-phase cells (Figure 1c; Methods).

We also aggregated all of the sequencing reads across single cells inferred to be in $G_1$- or S-phase, and generated an aggregate S/$G_1$ replication timing profile[20]. This aggregate profile was highly correlated to concomitantly or independently sequenced bulk S/$G_1$ or unsorted replication timing profiles generated from the same cell line ($r = 0.9$; Figure 1d, top). This provides a further validation of the *in silico* sorting assignments, as supported by the failure to reproduce the bulk replication profile when randomly assigning cells to replication states.

## Comprehensive measurement of single-cell DNA replication timing across thousands of cells, throughout S phase, and across cell lines

Having established a workflow for high-throughput replication analysis of unsorted cells, we performed whole-genome sequencing of 13,445 additional single cells, from GM12878 as well as eight other cells lines including embryonic stem cell lines and cancer cell lines. As with the flow-sorted GM12878 cells, we were able to distinguish replicating and non-replicating cells, and to generate an aggregate S/$G_1$ profile that was highly correlated to an independently generated replication-timing profile for that cell line ($r = 0.83$-$0.97$). We then generated replication profiles for between 76 and 1,358 S-phase cells across the different cell lines. By sorting the cells by the percent of the genome replicated, we observed a consistent pattern of single-cell replication and progression of DNA replication along S phase that corresponded well to the aggregate and bulk replication-timing profiles. In particular, there were distinct peaks in the single cell data that

aligned with the locations of peaks in the ensemble replication timing profiles (Figure 1d, Figure 2a,b). Thus, our approach enables accurate determination of replication state across hundreds or thousands of cells across the entire S phase and in different human cell types. In contrast to previous studies that analyzed single-cell replication profiles at the level of large chromosomal "domains" [15,16], our data reveals discrete initiation from localized initiation "sites" (peaks in the replication profiles), which we assume correspond to individual, or tight clusters of, replication origins. These initiation sites can be observed and analyzed at the single cell level, as further dissected below.

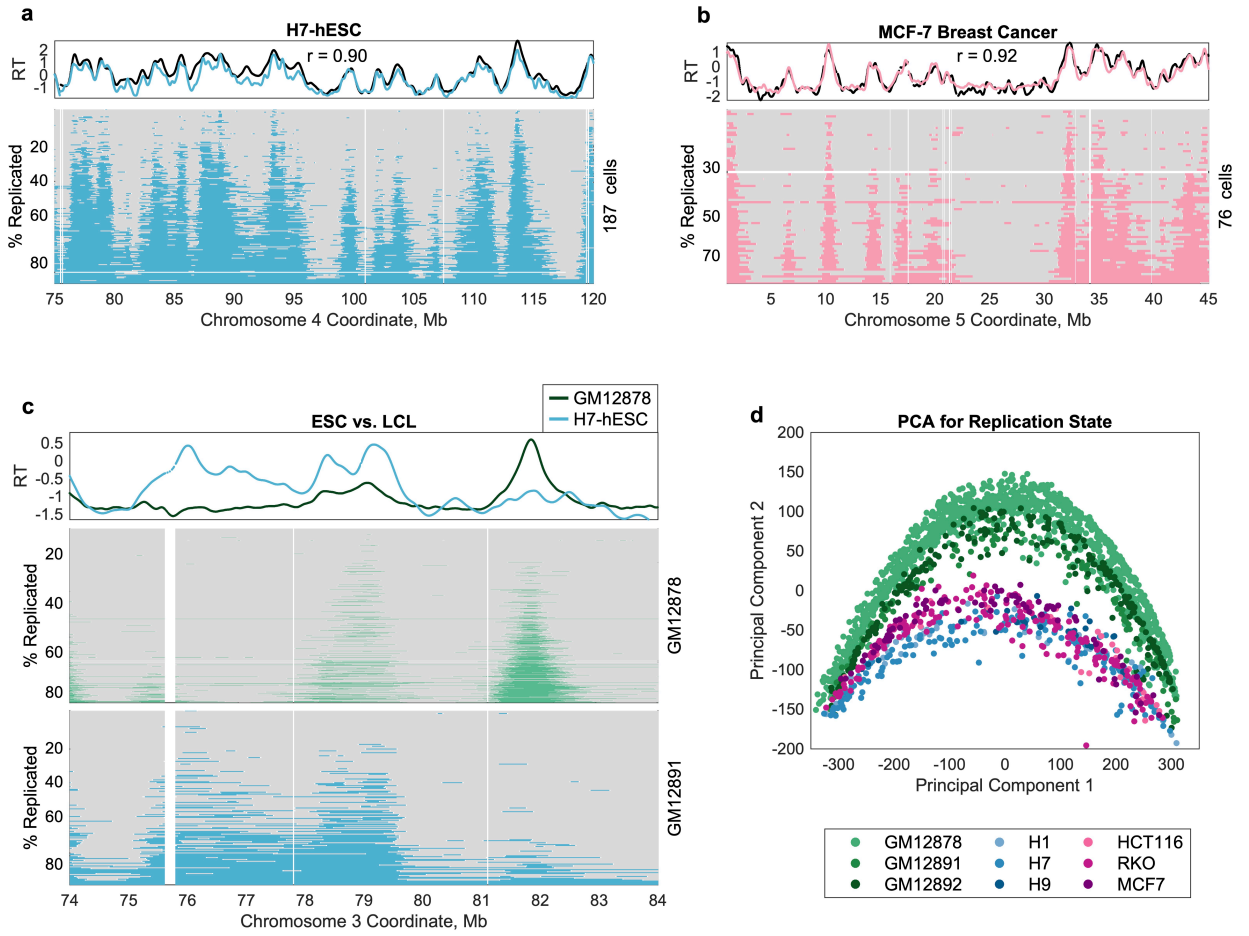


Figure 2. Comprehensive measurement of single-cell replication timing across cell types.

**(a, b)** As in Figure 1d, for the embryonic stem cell line H7 **(a)** and for the breast cancer cell line MCF-7 **(b)**.

**(c)** Replication timing variation between cell types is observed at the single-cell level. Top panel: aggregate profiles for GM12878 (black) and H7-hESC (blue) as in **(d)** and **(e)**. Middle and bottom panels: single cell data reveals that the peaks at ~76Mb and ~82Mb reflect cell-type specific initiation sites in ESC and LCL, respectively. In contrast, the peak at ~79Mb that appears to be early-replicating in ESC and late-replicating in LCL in the aggregate profiles is replicated at a relatively consistent time across single H7-hESC cells, but replicates throughout S phase in a subset of single GM12878 cells.

**(d)** Single cells follow cell-type-specific trajectories of S-phase progression, as determined by principal component analysis (PCA). PCA was performed on all genomic windows across autosomes. PC1 corresponds to the % of the genome replicated (r = 0.99; from left to right). Each dot represents a single cell.

When comparing single-cell replication profiles across cell types, differences between cell types observed in the aggregate replication timing profiles were also observed at the single cell level (Figure 2c). While some of these differences can be attributed to the presence of a cell-type-specific initiation site, others reflect a shared initiation site that tends to be fired early in one cell type and late in another. Most intriguingly, some cell-type differences appear to result from shared initiation sites that are consistently used in one cell type but only in a subset of cells in the other (e.g. Figure 2c). These differences in replication state between cell types are sufficient to cluster cells by principal component analysis (Figure 2d).

**Sites of replication initiation are consistent in single cells**

The nature of DNA replication initiation events is among the most debated aspects of mammalian DNA replication, both regarding its spatial scale (specific loci[22-25], localized regions[26-28] or broad domains[6,7]) and the degree of spatial and temporal stochasticity across cells[1,11,12]. Our comprehensive single-cell DNA replication data enables us to rigorously address these subjects.

We focused first on the spatial dimension of variability among cells. As noted above, very little variation in initiation site locations was visually observed in the single-cell data (Figure 1d; Figure 2a, b; Figure 3a). To analyze this axis of variation systematically, we first called peaks in the aggregate replication timing profiles. In GM12878, the cell line for which we have the largest sample size, we identified 3,792 autosomal peaks for which we could delineate the boundaries with high confidence (the peak was ≥100kb and ≥0.05 replication timing standard deviations from at least one of its neighboring valleys). Next, we compiled a list of replicated segments across all cells that overlapped a single aggregate peak, thus excluding regions that we knew to contain multiple potential replication origins or that could be attributed to passive replication by a neighboring initiation site. This produced a set of 228,759 replicated segments, which we term replication "tracks" (mean ± standard deviation: 648 ± 331 kb). For each peak in the aggregate profile, we identified all informative replication tracks. We then assigned the center of each track as the most likely location of replication initiation in that cell, and asked how closely these single-cell initiation locations were clustered (Figure 3b). The median width of replication initiation regions was 120kb, with some localized to regions of just 40kb – merely two data windows (Figure 3c, d, e). As noted above based on visual inspection of the data, these results are at odds with the notion that there are megabase-scale replication domains that are simultaneously replicated[6,7,15,16].

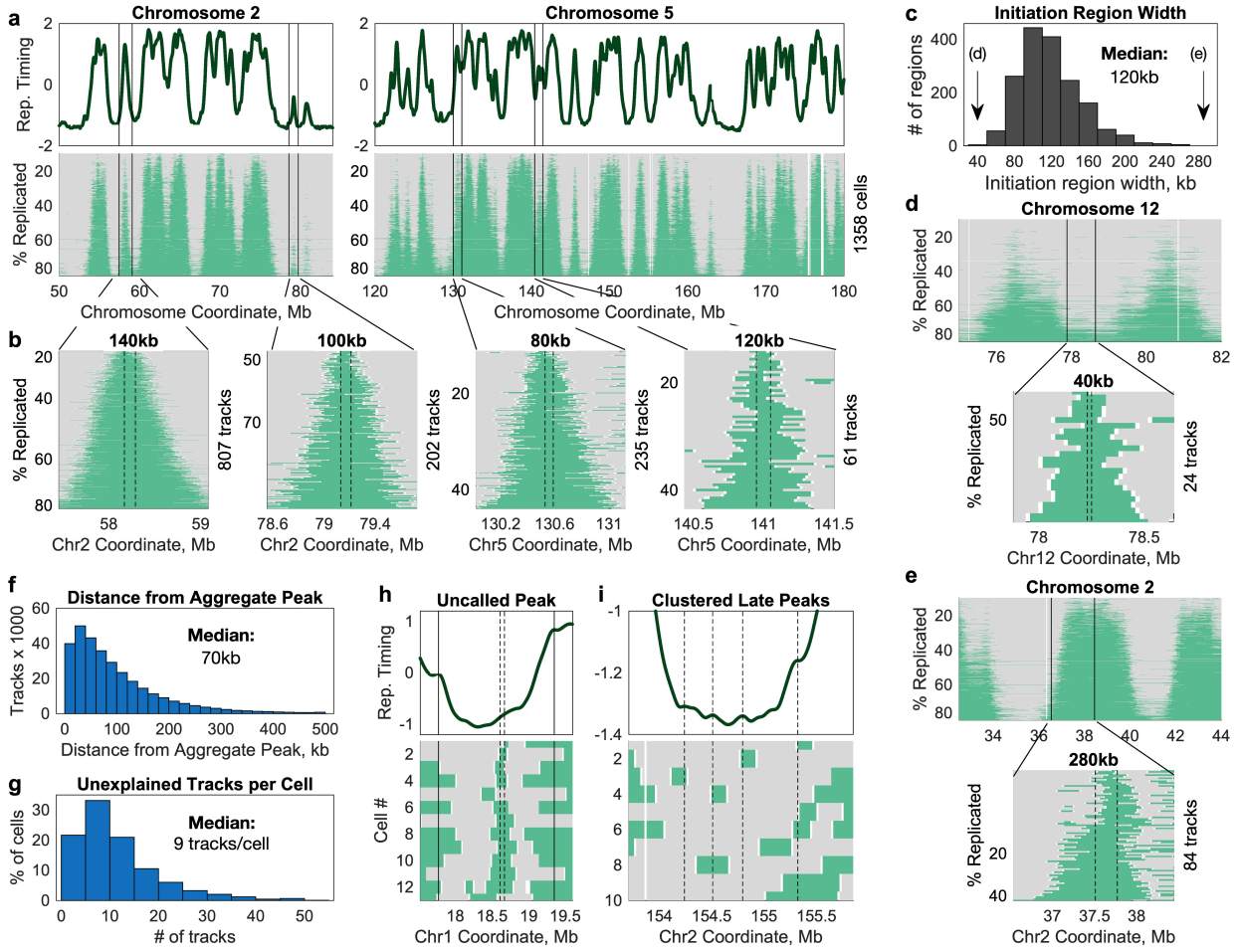


Figure 3. Consistency of single-cell replication initiation sites.

**(a)** Peaks in the aggregate replication timing profile inferred from all 1,358 GM12878 S-phase cells (*top*) correspond to segments that are consistently replicated across single cells (*bottom*). Two example regions are shown. The indicated regions correspond to the full width of the insets in **(b).**

**(b)** Replicated regions in single cells are centered at consistent locations, which overlap peaks in the aggregate replication timing profile. For each peak in the aggregate profile, a subset of single cells was identified that contained a replicated region (green track) overlapping that peak but do not extend into either neighboring peak. For each replicated region, the center position was assigned as the location of replication initiation in that cell, and the common initiation site (dotted black lines) was defined as the region between the 25$^{th}$ and 75$^{th}$ percentile of the range of initiation locations across tracks.

**(c)** The distribution of replication initiation region interquartile widths, as in the distance between dotted lines in **(b)**, for all initiation sites in the genome.

**(d)** Four initiation sites were localized to a 40kb region. In the example shown, analysis of 24 replication tracks permitted the identification of a narrowly defined replication initiation site (*bottom*), despite the lack of a clearly defined region across all single cells (*top*).

**(e)** One initiation site was localized to a 280kb region. This initiation site was located in an early-replicating region. Analysis of 84 replication tracks identified a wide replication initiation site (*bottom*). Interestingly, there appear to be two clusters of replication tracks, with distinct center positions. This suggests the possibility that "wider" initiation regions identified in our analysis actually represent multiple discrete initiation sites that cannot be discriminated in the aggregate profile.

**(f)** 95% of single-cell replication tracks are within 280kb of the nearest peak in the aggregate profile. This is consistent with the maximum initiation region width calculated in (**c**), suggesting that at most 5% of replication initiation events are initiated from locations other than the peaks identified in the aggregate replication timing profile.

**(g)** 58% of cells contain 10 or fewer replication tracks genome-wide that are > 280kb from a peak called in the aggregate replication timing profile, and 99% contain fewer than 50 such tracks.
**(h)** An example of single-cell replication tracks not identified in the aggregate profile (and thus counted towards the 5% of uncalled initiation events). The locally early replication in the aggregate replication timing profile (*top*) is insufficient to call this as a peak, and this region is >280kb from the nearest neighboring initiation sites (solid lines), nominating it as a uncalled peak. Nonetheless, single-cell replication tracks (*bottom*) identify an 80kb replication initiation region shared by 13 single cells. As in (**b**), dotted lines represent the 25$^{th}$ and 75$^{th}$ percentile estimates for the location of the initiation site.
**(i)** Single-cell replication tracks support ambiguous peaks observed in the aggregate replication timing profile. In the region shown, four local maxima are identified (black dotted lines) in the aggregate profile in a late-replicating region flanked by two much earlier-replicating regions. Each of these local maxima is supported as a bona fide replication initiation site by at least one replication track that cannot be attributed to any of the others (*bottom*). As in (**h**), each of these single-cell replication tracks was counted toward the number of unexplained initiation events.

We next considered the reciprocal question of how often replication initiates in single cells from initiation sites other than the peaks called in the aggregate replication timing profile. For this analysis, we also included replication tracks that did not overlap any of the aggregate peaks (total n = 312,741; mean ± standard deviation: 515 ± 334 kb). Again, we assigned the center of each track as the most likely location of replication initiation for that track. We found that 94.88% of single-cell initiation sites were within 280kb of an aggregate peak (median distance: 70kb), consistent with replication initiating at that peak (Figure 3f). This suggests that at most 5% of the genome is replicated from initiation sites other than the peaks called in the aggregate profile. Furthermore, 99% of cells contained no more than 50 of these ectopic initiation events, further underlining how rarely they occur in any given cell (Figure 3g).

Several compelling examples suggest that at least some of the remaining initiation events are also common across cells (and missed among the 3,792 prominent peaks we operationally defined). For instance, in one example, a narrowly defined (80kb) initiation site was shared by 13 cells, and contained a "shoulder" in the aggregate profile (Figure 3h). In a second example, single cells supported the presence of four low-confidence peaks in a late-replicating region (Figure 3i). While the latter example may represent cases of rare, ectopic initiation outside of common initiation sites, both instances establish only a lower bound for the number of cells activating a particular initiation site, as additional initiation events at these sites may have been missed in cells at other stages of replication progression. In both cases, the single-cell data is consistent with the majority of initiation events captured by the aggregate replication profiles, either as prominent peaks or as "lower-confidence" peaks or shoulders that likely depend on data smoothing. Thus, ensemble replication timing profiling captures the vast majority of replication initiation events, and ectopic, "off-site" initiation is rare even in single cells.

## Initiation sites fire in a consistent, but not strictly deterministic order, across cells

Given that single cells appear to initiate replication from the same genomic locations at least 95% of the time, we turned our focus to the temporal axis of variation: how consistent is the order in which single cells initiate replication at these loci?

To answer this question, we assumed as our null a strictly deterministic model in which every cell fires every initiation site in the same order. Under this model, a cell that has fired a single initiation site will have fired the earliest-replicating initiation site, and a cell that has fired five initiation sites will have fired the five earliest-replicating initiation sites. More broadly, under the null model, the number of initiation sites that have been replicated in a given cell also determines *which* sites have been initiated. To look for divergence from this model in the single-cell data, we first ranked each of the peaks identified in the aggregate profile from the earliest to the latest. This allowed us to predict which initiation sites should have fired in each cell, based on the number of initiation sites observed to have been replicated (Figure 4a,b). These predictions were supported in the majority of cells, indicating that the order of replication initiation is relatively similar between single cells (Figure 4c). However, this order is not fixed, and the predicted replication state of a given initiation site was different than expected, on average, in 9.1% of cells (Figure 4d).

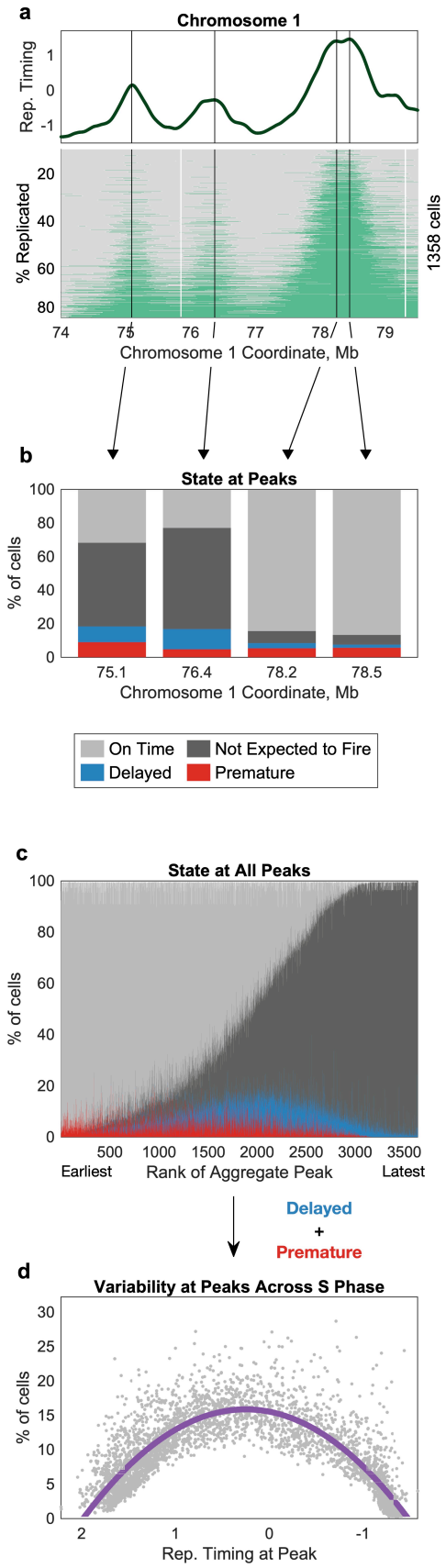


Figure 4. Variation in the timing of replication initiation in single cells changes across S phase.

**(a)** Initiation sites differ in their degree of consistency across single cells. Peaks in the aggregate replication timing profile (*top*) were ranked from earliest to latest, allowing prediction of which initiation sites would have fired in each single cell under a strict ordering of origin firing. These predictions were then compared to the single cell data (*bottom*).

**(b)** The two early-replicating peaks shown in this example (78.2Mb and 78.5Mb) were fired out of order in 7.5% and 8.4% of cells, respectively; while the two later-replicating peaks (75.1Mb and 76.4Mb) were fired out of order in 18.3% and 16.8% of cells, respectively.

**(c)** Initiation sites are fired in the expected order in the majority of single cells. However, initiation sites expected to fire in the middle of S-phase vary more than those at the beginning or end of S phase. Each column represents an initiation site, sorted from the earliest (left) to the latest (right). Cells that have replicated an initiation site that was not predicted to fire in that cell are considered to have "premature" firing (red), while those that have not replicated an initiation site predicted to have fired already are considered to have "delayed" firing (blue), as in (**b**).

**(d)** Variability in initiation site firing is most variable for those sites that are expected to fire in the middle of S phase. On average, initiation sites behaved differently than expected (either firing prematurely or delayed) in 9.1% of cells (range: 0.3%-40%). Notably, mid-S-phase initiation sites (aggregate replication timing between 0 and 1) were more variable than earlier or later initiation sites. Each dot represents one initiation site. Red line: second-order polynomial fit.

Deviation from the expected order of replication initiation was not uniformly distributed across S phase. Specifically, variability was low for early-replicating initiation sites, increased in mid-replicating initiation sites, and then decreased again for late-replicating sites. Mid-early replicating initiation sites (replication timing value between 0 and 1) comprised 20.29% of all initiation sites but 44.67% of the initiation sites that were variable in more than 10% of cells. Indeed, 725 of the 738 (98.24%) mid-early replicating initiation sites were variable in more than 10% of cells.

Next, we analyzed the replication timing of instances in which initiation occurred out of the expected order. While we cannot determine the precise order in which initiation sites fired in a given cell (since many are already well replicated at the time of measurement), individual cells do provide information about the bounds of an initiation site's rank. For each peak in the aggregate profile, we identified the earliest cell to have replicated that region (i.e. the minimum rank) and the latest cell that had not yet replicated that region (i.e. the maximum rank) (after excluding outliers; Methods). This produces a range of firing order for each initiation site (Figure 5a,b). For each initiation site, this range of firing got progressively later across S phase, but throughout was constrained to a confined time window surrounding the expected initiation order (Figure 5c). This suggests that the aggregate replication timing profile measures the preferred time during S phase at which each initiation site fires, and that initiation sites have limited potential to fire throughout S phase. Additionally, the continuous trend in the range of firing order suggests that early- and late-replicating initiation sites do not differ categorically, e.g. early-replicating initiation sites are not more efficient at firing at their preferred time than late-replicating initiation sites. Rather, replication timing reflects a continuum of firing potential that exists across S phase.

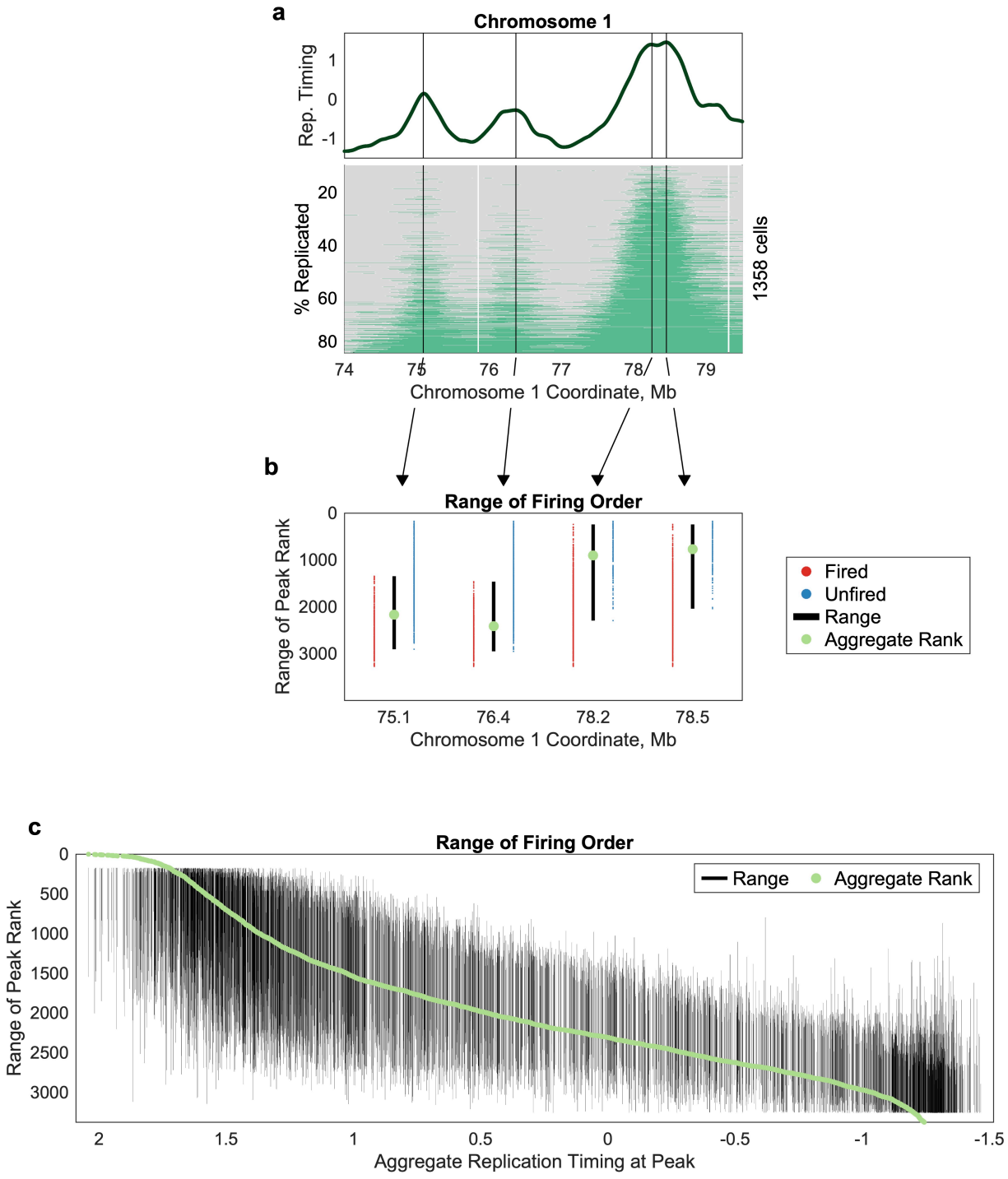


Figure 5. The order of initiation events is constrained in single cells.

**(a)** Peaks in the aggregate profile were ranked and used to make predictions about single cell data, as in Figure 4a. The same example region is reproduced here.

**(b)** The order of firing for each initiation site is constrained to a narrow range within S phase. For each initiation site, a range of firing orders (black line) was constructed, spanning from the earliest cell to fire that initiation site (red) to the latest cell that had not yet fired that site (blue). As a null, each peak is expected to have the same rank it holds in the aggregate (green).

**(c)** Across S phase, the average range of firing order is of a relatively consistent magnitude. The center of the range of firing order becomes progressively and continuously later throughout S phase. Range of firing order (black) and expected rank (green) were determined for each peak in the aggregate profile, as in (**c**).

## Discussion

While ensemble replication profiling methods cannot capture (and may be confounded by) cell-to-cell heterogeneity, previous single-molecule and single-cell methods have been largely limited in their throughput or accuracy. Here, we report a scalable method for analysis of thousands of single replicating cells, across multiple cell lines, and at kilobase resolution. At the same time, our replicon-centered analytical approach enables us to identify which cells are informative about which replication initiation sites, capturing information that is analogous to that collected from lower-throughput single-molecule studies.

We find that single cells initiate replication at consistent loci, corresponding to peaks in the replication timing profile. Furthermore, we are able to pinpoint the locations of these replication initiation sites to regions ranging from 40-280kb (which themselves may be over-estimates due to low-coverage), challenging the model that there are megabase-long replication domains that are replicated simultaneously[6,7]. Analogously, our data do not support the existence of large constant replication regions (CTRs)[29]. While it is conceivably straightforward to envision how measurements with limited resolution would give the impression of domains or CTRs where none exist, it appears more difficult to reconcile the sharp and discrete initiation peaks in our single cell data with the idea of large regions with constant replication timing. In contrast, our data is consistent with recent high-resolution studies that suggest that replication initiation is confined to regions of several tens of kilobases[5,10], although we do not find compelling evidence for widespread ectopic initiation from region outside commonly used initiation sites[5]. While many previous studies of mammalian replication origins relied on biochemical enrichments of DNA synthesis events[1], single-cell DNA sequencing more reliably represents productive and internally-validated DNA replication.

While spatial variability in replication initiation is rare, temporal variation is more common. In general, initiation sites expected to fire in the middle of S phase are more variable than those expected to fire earlier or later. At the level of individual initiation sites, we find that each site has a preferred time of firing that is captured by the ensemble replication timing profile and that out-of-order firing is constrained to a narrow range of S phase. In contrast to previous work[5,16], we do not find compelling evidence of initiation sites that normally fire at the end of S phase firing very early, or vice versa. Interestingly, although mid-S phase initiation sites fire in a different order in a greater number of cells than early- or late-S phase initiation sites, the corresponding increase in the width of the range of firing times is modest. Thus, replication timing heterogeneity might be mechanistically similar across S phase. Finally, the range of potential firing times gets continuously and progressively later throughout S-phase progression, suggesting that if there are distinct classes of replication origins that differ in their efficiency or variability, those classes do not correspond neatly to early and late replicating origins. Overall, while our data are consistent with clusters of origins organized in a limited

number of broader initiation zones, they most strongly support the notion that the most physiologically relevant scale at which to study replication timing is neither basepair-resolved origins nor megabase-scale CTRs, but rather initiation sites spanning several tens of kilobases. At this scale, the locations and timing of replication follow a near-deterministic program without much need to invoke stochastic processes.

Single-cell DNA sequencing of proliferating cell samples, without experimental manipulation (e.g. cell synchronization or sorting), can reveal the dynamics of DNA replication in exquisite detail. Applying this approach across cell types, genetic backgrounds, and experimental conditions will reveal how replication is altered at the spatiotemporal levels in different physiological contexts. With constantly improving methods for high-throughput single cell isolation and accurate whole-genome amplification[18,19,30], this approach promises to become ever more informative for the understanding of the DNA replication timing program.

## Methods

### Cell Culture

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

### Library Preparation and Sequencing

Isolation, barcoding, and amplification of single cell genomic sequencing libraries was performed on the 10x Genomics Chromium Controller instrument, using the 10x Genomics Single Cell CNV kit (10x Genomics, Pleasanton, CA, USA). Paired-end sequencing was performed for 150 cycles with the Illumina HiSeq X Ten (GENEWIZ, Inc., South Plainfield, NJ, USA) or for 100 cycles with the Illumina NextSeq 500 (Cornell University Biotechnology Resource Center, Ithaca, NY, USA).

Sorted GM12878 cells were cultured, isolated, and sequenced by 10x Genomics.

### Processing of Single-Cell Barcodes

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

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

**Processing of sequencing reads**

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

**Computational identification of $G_1$ cells**

Reads were initially counted in fixed-size 20Kb windows. After removing windows with < 75% mappability[31], set of 50 windows were aggregated together to calculate the median absolute deviation of pairwise differences between adjacent windows (MAPD). MAPD was then scaled by the square root of the mean number of reads per aggregated window (mean coverage/Mb), to produce a linear relationship between mean coverage/Mb and MAPD. Using an expectation-maximization procedure, we fit two linear models between these variables as a Gaussian mixture. A predicted MAPD was calculated for each cell using the $G_1$ linear model ($y = 0.001x + 0.95$), and cells with a residual ≤ 0.05 were assigned as $G_1$.

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

Finally, we identified and filtered out cell-specific copy-number aberrations (CNA). To do this, we used the MATLAB function `findchangepts` to iteratively search the whole-genome profile for the two most likely locations of abrupt changes in mean and slope. The region between these two breakpoints was then tested as a potential CNA, and was filtered out if the region median was < 0.9 or > 1.1 times the genome-wide median ***and*** there was a significant difference between the putative CNA and the rest of the genome by two-sample t-test.

**Replication state inference**

For each cell, we assigned each $G_1$-defined window as "replicated" or "unreplicated" using a two-component hidden Markov model (HMM). To initialize the model, we first fit a two-component mixed Poisson model to aggregated read counts (15 windows, ~300Kb) and assigned each window to the mean it was closer to. If this initial model had higher dispersion than a single-component Poisson model, we assigned the cell as $G_1$. Otherwise, we refined the initial window assignments using the HMM, which modeled read counts as the mixture of two Poisson processes. Cells for which the HMM failed to converge after 40 iterations were filtered out.

Because the HMM does not model the expected two-fold relationship between replicated and unreplicated regions, we assessed the quality of the HMM output using this relationship. Specifically, we calculated the ratio between the average number of reads in windows assigned as replicated to the average number of reads in windows assigned as unreplicated. To be included in further analysis, this

ratio was required to be between 1.7 and 2.2. Additionally, to find any cells that contained uncorrected CNAs, we calculated the average copy-number assigned to each chromosome, and removed cells for which the standard deviation between chromosomes was greater than 0.4.

Finally, for the ease of analysis, we interpolated the data back onto fixed-size 20Kb windows. Windows for which a value other than 2 or 4 was interpolated were masked, as were windows with low mappability (< 75%).

**Aggregate replication timing profiles**

For each cell line, we generated an aggregate S/$G_1$ profile, as in Koren *et al.*[20], except that we generated the $G_1$ and S fractions *in silico* by aggregating reads across all cells assigned to that fraction. Briefly, the $G_1$ fraction was used to generated variable-size windows with a fixed number of reads (n = 200), and the number of S-phase reads was then counted in the same windows. This profile was smoothed in a gap-aware fashion with a cubic smoothing spline (MATLAB function `csaps`), with a smoothing parameter of $10^{-16}$, and normalized to a mean of 0 and standard deviation of 1.

**Localizing initiation sites**

To identify high confidence peaks in the aggregate replication timing profile, we defined segments that contained one local maximum and its two flanking local minima. We then filtered this initial list by removing any local maximum that was <100Kb from ***both*** flanking local minima or was <0.05 replication standard deviations greater than ***both*** flanking local minima.

To identify initiation events in single cells, we first identified all segments of the genome that were inferred to be in a replicated state. However, because this included both actively and passively replicated regions, we then focused only on "informative" segments, those that overlapped a single aggregate peak. For each informative segment, we assigned the center position as the most likely site of replication initiation. Then, we analyzed the distributions of these initiation sites for each peak with at least 20 informative segments. We defined the region of replication initiation as spanning from the 25$^{th}$ percentile to the 75$^{th}$ percentile of the center positions of informative segments for that peak.

To assess how often ectopic replication occurs at the single-cell level, we identified all replicated segments that overlapped *no more than one* aggregate peak as "informative". We again assigned the center of each segment as the most likely location of replication initiation, and for each location, we calculated how far it was from the nearest aggregate peak.

**Variation in timing across cells**

To assess variation in timing across cells, we compared the data to a null model under which every cell replicates the same initiation sites in the same order. Under this model, the number of initiation sites fired also dictates which sites are fired. Thus, we counted the number of replicated regions overlapping initiation sites in each cell, and then predicted which regions those would be under the null model. For each peak in the aggregate, we then calculated the variability of that initiation site by counting how many cells did not follow our prediction.

To determine the range of firing orders, we identified all of the cells that had already replicated a given segment and all of the cells that had not yet replicated that segment. After removing outliers from each list, we calculated the range of firing orders as spanning from the minimum number of replicated initiation sites in the already-replicated population to the maximum number of replicated initiation sites in the not-yet-replicated population.

# References


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

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

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

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

5 Wang, W. *et al.* Genome-Wide Mapping of Human DNA Replication by Optical Replication Mapping Supports a Stochastic Model of Eukaryotic Replication. *Biorxiv*, doi:10.1101/2020.08.24.263459 (2021).

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

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

8 Ding, Q. *et al.* The Genetic Architecture of DNA Replication Timing in Human Pluripotent Stem Cells. *BioRxiv* **doi.org/10.1101/2020.05.08.085324** (2020).

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

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

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

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

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

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

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

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

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

18 Gonzalez, V. *et al.* Accurate Genomic Variant Detection in Single Cells with Primary Template-Directed Amplification. *BioRxiv* (2020).

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

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

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

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

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

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

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

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

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

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

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

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

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