High-throughput analysis of DNA replication in single human cells reveals the complex nature of replication timing control

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

* Correspondence to: koren@cornell.edu

Abstract

DNA replication initiates from replication origins firing at different times throughout S phase. Debate remains about whether origins are a fixed set of loci, or a loose agglomeration of potential sites used stochastically in individual cells, and about how consistent their firing time is. We developed an approach to profile DNA replication from whole-genome sequencing of thousands of single cells. We describe "*in silico* flow cytometry", a method for discriminating replicating cells with superior accuracy to FACS and staging them across S phase. Using two microfluidic platforms, we analyzed up to 2,428 individual replicating cells from a single sample. The resolution and scale of the data allow focused analysis of replication initiation sites, demonstrating that the vast majority are in confined genomic regions. While initiation occurs in a remarkably similar order across cells, we unexpectedly identified a subset of initiation regions that constitutively fire in late S phase, and another subset firing randomly throughout S phase. Taken together, high throughput, high resolution sequencing of individual cells reveals previously underappreciated variability in replication initiation and progression.

Introduction

Faithful duplication of the genome is a critical prerequisite to successful cell division. Eukaryotic DNA replication initiates at replication origin loci, which are licensed in the G_1 phase of the cell cycle and fired at different times during the S phase. In many eukaryotes, sequencing of cells at different stages of the cell cycle has been used to profile DNA replication timing, which measures the relative time that different genomic regions are replicated during S phase (reviewed in[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 the reproducible nature of the replication timing program reflects the consistent activity across cells of specific individual replication origins or stochastic firing of different origins in different cells within a given region. Ensemble replication timing measurements have been interpreted to indicate that replication is organized in broad "domains", spanning hundreds of kilobases to several megabases with consistent replication timing governed by the activity of clusters of replication origins[5,6]. Furthermore, some recent replication origin-mapping methods have indicated that replication origins are highly abundant and dispersed throughout the human genome[1,7], suggesting that many sites may function as origins used in a subset of cell cycles. In contrast, high-resolution measurements of hundreds of human replication timing profiles[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 may use different subsets of origins in each cell cycle[13,14].

Recently, replication timing has been analyzed by single-cell sequencing of several hundred mouse or human cells[15-17]. These studies focused on cells in the middle of S phase and analyzed replication at the level of domains, concluding that stochastic variation exists in replication timing and is highest in the middle of S phase. However, single-molecule and single-cell studies have been limited in their throughput and biased toward early S-phase or mid S-phase, respectively. Analyzing many cells is particularly important given that even when the whole genome is captured, a single cell provides only a snapshot of DNA replication at a single moment in time. By assaying many cells at different stages of S phase, it is possible to string these snapshots together to construct a picture of replication states over time. However, the resolution of this picture will be dependent both on capturing cells at many stages of S phase and on assaying a large number of cells.

Here, we report the analysis of whole-genome sequencing of thousands of single replicating cells across ten human cell lines. We developed an *in silico* approach to sort cells by cell cycle

state, allowing us to capture cells throughout the full duration of S phase, and to analyze them in any number of sub-S phase fractions down to single-cell resolution. We found that single cells within a given cell line largely used a consistent set of replication initiation regions, which were discrete genomic loci rather than megabase-scale domains. Furthermore, these initiation regions fired in a predictable, albeit not fixed order. Some initiation regions were consistently fired early in S phase across cells, while others were fired consistently late. However, we also identified a subset of rarely fired initiation regions with a preference for early firing and another subset that fired throughout S phase. We conclude that a consistent set of replication origins explains the vast majority of replication initiation events in single cells, and that existing models of replication timing fall short of explaining the diversity of firing time patterns.

Results

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 two microfluidic systems that isolate and barcode single-cell DNA: the 10x Genomics Single Cell CNV platform, which uses multiple-displacement amplification (shown to be superior to DOP-PCR for copy-number analysis[18]), and direct DNA transposition single-cell library preparation (DLP+)[20], which is an amplification-free method. Both library preparation methods were followed by whole-genome sequencing of single cells.

As an initial proof-of-principle, we analyzed 5,793 cells from the human lymphoblastoid cell line (LCL) GM12878 isolated with the 10x Genomics system, following fluorescence-activated cell sorting (FACS) of G_1-, G_2-, and several fractions of S-phase cells. The resulting sequencing data were sufficient to distinguish replicating cells from non-replicating cells across a five-fold range of sequencing read depths (50-250 reads per Mb). Specifically, local read depth fluctuated more in replicating cells relative to non-replicating cells of similar coverage (Figure 1a). To validate that these fluctuations could be used to computationally distinguish replicating cells from non-replicating cells within an unsorted population, we quantified them using MAPD (median absolute deviation of pairwise differences between adjacent genomic windows), which scales proportionally to read depth (Methods). Indeed, FACS-sorted G_1- and S-phase cells had distinct linear relationships between scaled MAPD and read depth (Figure 1b, left). Therefore, we were able to computationally assign each cell as "G1" or "S" (Figure 1b, right), and compare the resulting fractions to the FACS labels. While *in silico* sorting was highly concordant with FACS, we identified ~2.4% of cells in the G_1-phase FACS fraction that were inferred to be replicating, and reciprocally ~25.1% of cells in the S-phase FACS fraction that were inferred not to be replicating. Thus, *post hoc, in silico* sorting using single cell DNA sequence data provides greater sensitivity and less between-fraction contamination than FACS (Figure S1). Accordingly, we sequenced an additional three GM12878 samples without cell sorting, recovering an additional 3,787 cells in total. We analyzed these cells together with the sorted cell libraries, as described below.

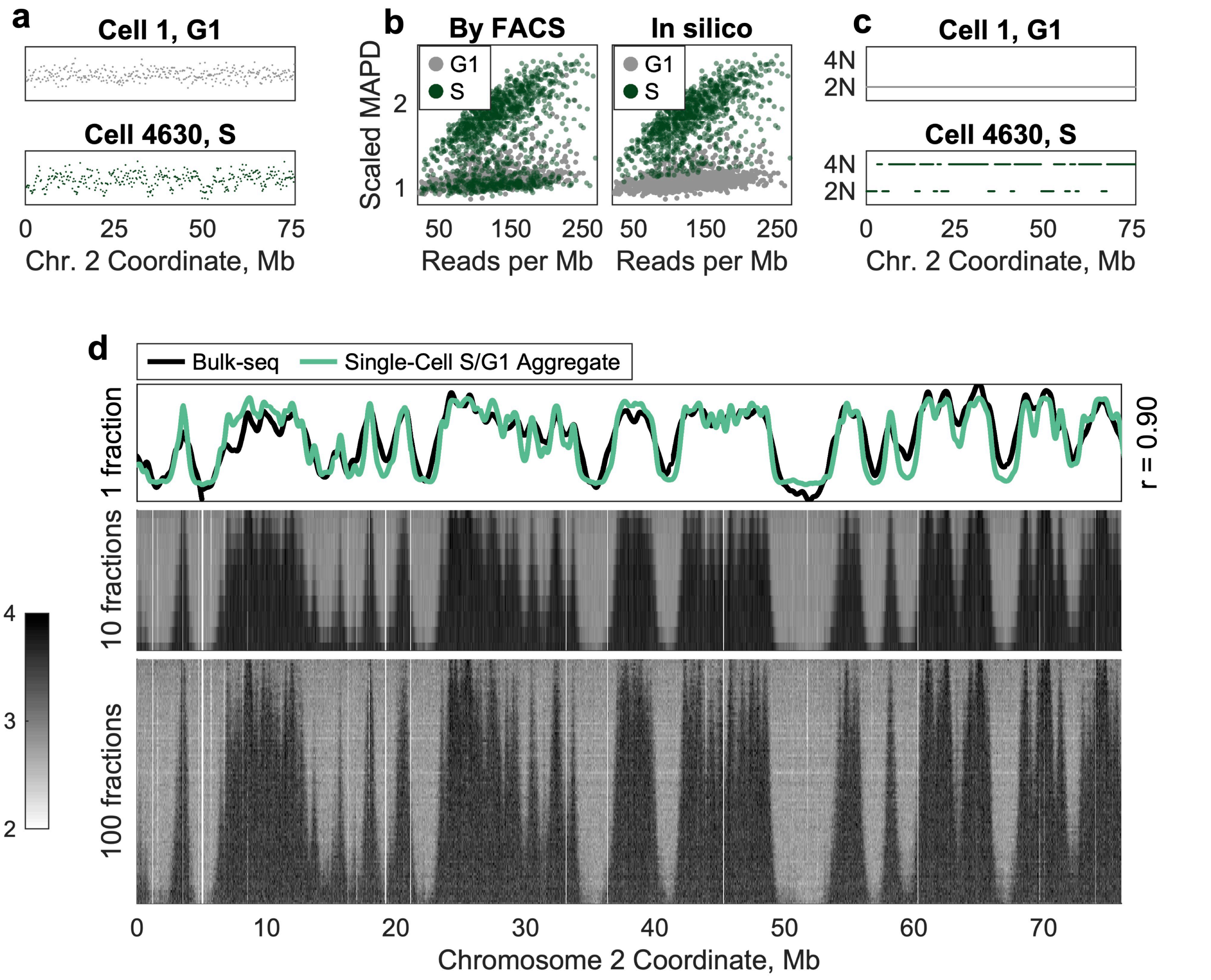

Figure 1. Discrimination of replicating and non-replicating cells by *in silico* flow cytometry.

(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. 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; median absolute pairwise difference between adjacent genomic windows divided by the square root of mean coverage-per-Mb). Left panel: cells are labeled as G_1- (gray) or S-phase (green) based on FACS sorting. Only the G_1- and S-phase fractions are shown. 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) Aggregating data across S-phase cells into one or more fractions reveals a consistent structure of replication progression at different times in S phase. Top panel: an ensemble replication-timing profile inferred from all S-phase cells together (green) was highly correlated with a bulk-sequencing replication-timing profile for the same cell line (black). Middle and bottom panels: single cells were aggregated into 10 or 100 fractions based on S phase progression. Pileups of high read depth (caused by replication in most/all cells in the fraction) are observed in discrete locations across the chromosome. The conical structure of these pileups suggests that replication initiation occurs from fixed loci and proceeds symmetrically in both directions. Each row represents one fraction (containing multiple cells), and each column represents a fixed-size window of 20kb.

In addition to reducing the observed cross-contamination between fractions, *in silico* cell sorting has two major benefits over traditional flow cytometry. First, sequencing biases (particularly, GC-content bias[21]) are known to vary between sequencing libraries, a concern alleviated by using control cells from within the same library as the cells of interest. In contrast, separate G_1- and S-phase libraries would need to be sequenced after FACS. Second, this approach minimizes experimental manipulations in generating the data, as it does not require DNA staining, and inter-experimental variation, for instance in defining FACS gates.

Using G_1 cells identified by this "*in silico* cell sorting" approach, we defined variable-size, uniform-coverage genomic windows that accounted for the effects of mappability and GC-content biases, as well copy number variations, on sequencing read depth[22]. We counted the number of sequencing reads in each window for each cell, and then used a two-state hidden Markov model to infer whether each window contained replicated or unreplicated DNA (Methods). 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).

A discrete benefit of single-cell data is the ability to aggregate similar cells together, effectively increasing the coverage without masking important heterogeneity between subsets of cells. Because the partitioning happens *in silico*, we can consider many different single-cell aggregates of the same data, from a single fraction (spanning all of S phase) down to single cells (wherein each cell is its own fraction). We generated several such aggregates, partitioning cells based on their progression through S phase (% of genome replicated) and summing per-window read counts across cells (Figure 1d). Validating this approach, the single fraction profile – analogous to an ensemble S/G_1 replication timing profile[22] – was highly correlated to a bulk replication timing profile for the same cell line (r = 0.90). Partitioning cells into 10 fractions, a structure emerged similar that seen in high-resolution Repli-seq[10]: conical pileups of high read depth (corresponding to active replication) around peaks observed in bulk sequencing. Many of these regions of high read depth were evident in every fraction, although some (*e.g.*, Figure 1d middle, ~13.8Mb) first appear later in S phase. This same structure was observed – at higher resolution – when cells were partitioned into 100 fractions. Thus, by this approach, we can capture sub-S phase events across all of S-phase without the risk of FACS cross-contamination (Figure S1), at a resolution for which FACS is infeasible (*i.e.*, 100 fractions), and with the ability to compare the same population of cells at multiple levels of resolution.

The logical extension of the partitioning approach is to consider each cell as comprising its own fraction. After filtering out cells that were not replicating or for which a two-fold relationship was not observed between copy-number states, we analyzed 2,428 single GM12878 cells. At this single-cell resolution, we observed consistent pileups of distinct replicated and unreplicated segments across cells (Figure 2a). These pileups were in the same regions observed as peaks in the bulk-sequencing profile and sub-S phase fractions, underscoring that these regions correspond to locations of active replication progression, centered at one or more replication origins. Even at single-cell resolution, these pileups were conical (Figure 2a, insets), consistent with symmetric bidirectional replication fork progression from a common origin locus and appeared visually to be highly localized. Thus, we demonstrate the ability to measure single-cell replication timing in thousands of single cells, in an unbiased manner, and without the need for FACS. This represents roughly ten times more cells than have been reported in previous single-cell replication timing analyses, which have focused primarily on mid-S phase cells[15,16].

We repeated this analysis using single-cell data for 3,040 cells from the LCL GM18507, prepared using amplification-free direct DNA transposition single-cell library preparation (DLP+).[20] We identified 759 replicating cells within this dataset, and again observed pileups in consistent genomic regions, close to peaks in the S/G_1 aggregate replication timing profile (Figure 2b). This dataset enabled us to benchmark our analysis strategy in the absence of amplification bias, ensuring that the observed single-cell pileups were not a persistent technical artifact of the 10x Genomics amplification method and validating the ability to accurately profile single cell replication timing in hundreds to thousands of cells across multiple single-cell sequencing technologies.

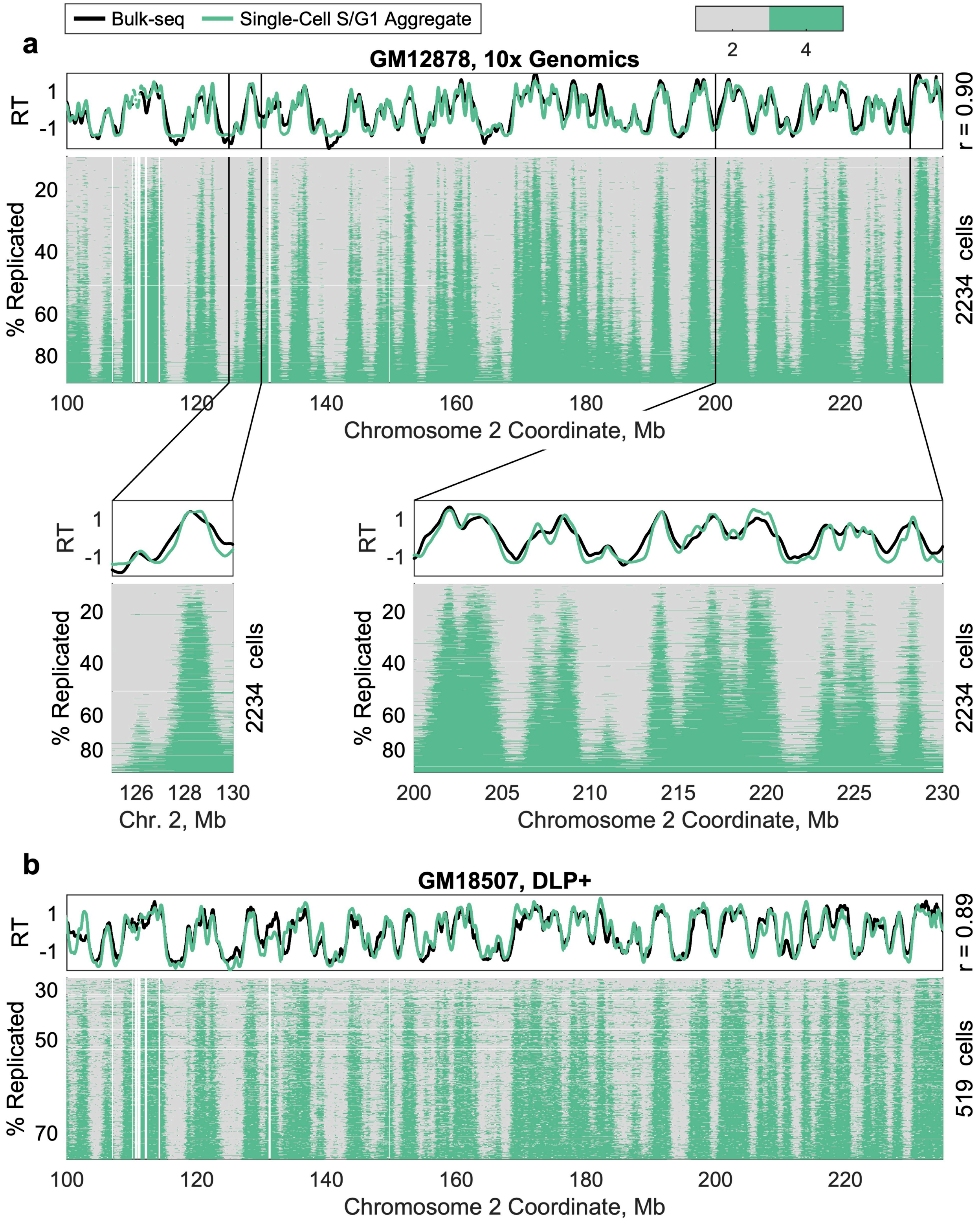

Figure 2. Single-cell replication state data, generated by multiple library preparation protocols.
(a) Single-cell replication profiles for 2,234 GM12878 cells (including both sorted and unsorted cells), following single-cell isolation and library reparation with the 10x Genomics Single-Cell CNV Solution. Consistency of the

replication program is observed across cells at chromosome-scale and at the level of individual peaks (inset). Pileups reflect sharply defined and consistently replication regions, which overlap peaks in the bulk replication timing profile. Variation in activation time during S phase among initiation sites is also observed to mirror the replication timing profile. Each row represents a single cell, sorted by the percent of the genome replicated, and each column represents a fixed-size window of 20kb. 194 cells are not shown due to copy-number aberrations on this chromosome. Low-mappability regions and cell-specific copy-number alterations have been removed (white). Insets show smaller regions.

(b) Single-cell replication profiles for 519 GM18507 cells, following amplification-free direct DNA transposition single-cell library preparation (DLP+). Due to noise, only 480-614 of the 759 S-phase cells were analyzed for any given chromosome. Raw data are from [20].

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[23-26], localized regions[27-29] or broad domains[5,6]) 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, visual inspection of replicated region pileups revealed very little variation across single cells (Figure 2; Figure 3a). To analyze this axis of variation systematically, we began by identifying replicated segments in each single cell. Each replicated segment, which we termed a "track" (by analogy to single-molecule DNA combing tracks), represents the activity of at least one replication origin. Theoretically, if a replication track corresponds to a single replicon, initiating from one origin and expanded by symmetric progression of sister replication forks, the origin of replication should be located at the center of that replication track. Thus, as a first approximation of origin locations, we assigned the center of each replication track as the most likely location of replication initiation for that track. (We excluded tracks longer than 1Mb in this initial analysis to reduce the likelihood of including tracks that reflected the activity of multiple independent origins that have converged.)

Consistent with previous work suggesting that replication initiation potential is diffuse throughout the genome[7], we found that 49.6% of mappable genomic windows were called as a probable initiation site in at least one cell. However, these probable initiation sites were not uniformly distributed across the genome. Rather, highly frequent initiation sites were neighbored by gradually less frequent initiation sites, creating peaks around these local maxima (Figure 3a, bottom). This structure suggests that a more limited group of genomic loci might give rise to replication initiation, as ambiguity in identifying the boundaries of replication tracks would result in slight shifts in the probable initiation site from the true midpoint to a neighboring locus and the observed gradual decrease in initiation frequency with increasing distance from that true midpoint.

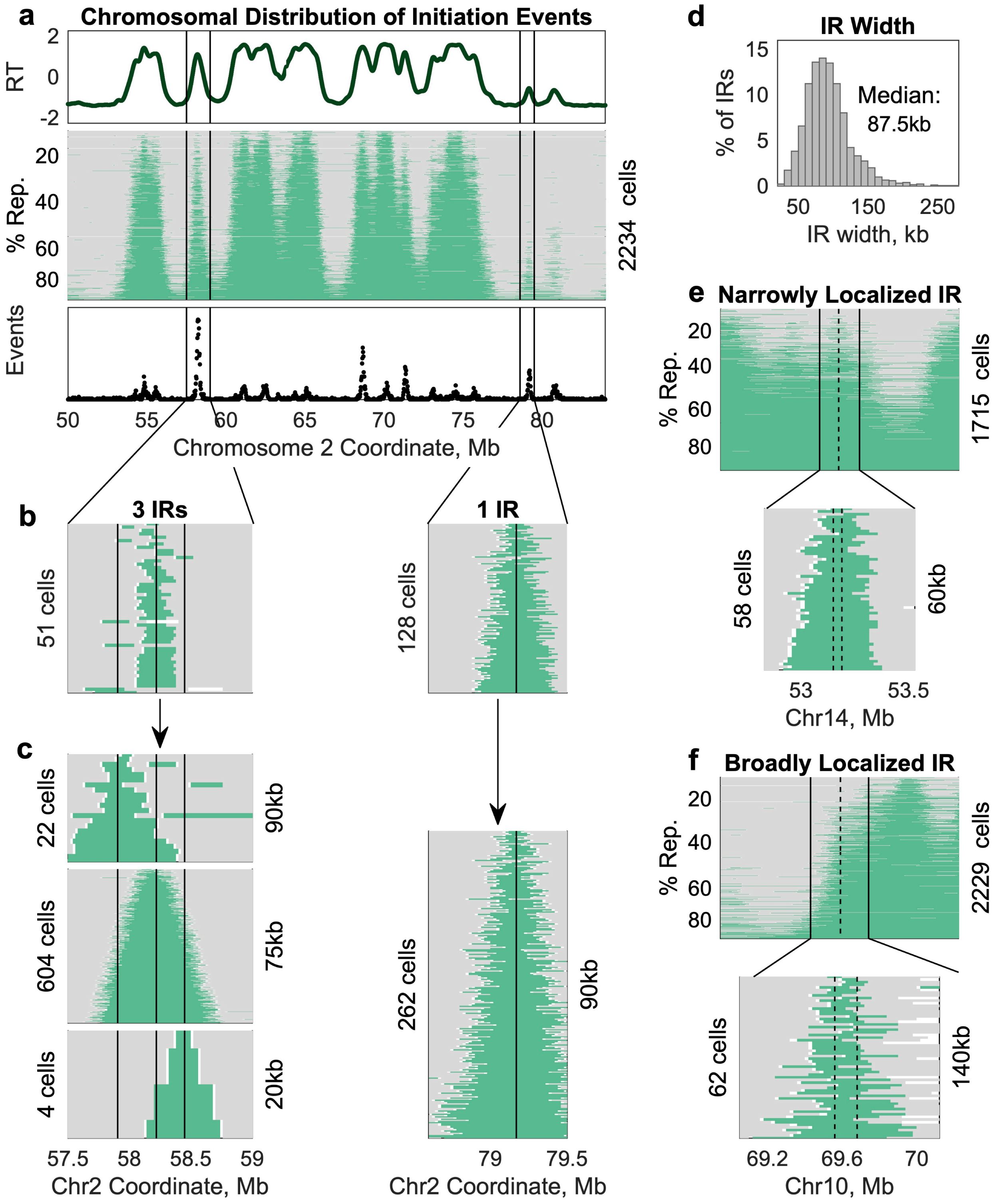

Figure 3. Consistency of single-cell replication initiation sites.

(a) Peaks in the aggregate replication timing profile inferred from all GM12878 S-phase cells (*top*) correspond to segments that are consistently replicated across single cells (*middle*). These aggregate peaks also correspond to regions of dense initiation site calls (*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, termed initiation regions (IRs), which overlap peaks in the aggregate replication timing profile. For each IR (black line), a subset of single cells was identified that contained a replicated region (green track) overlapping the IR center but not extending into either neighboring IR. Some aggregate peaks were found to correspond to multiple neighboring IRs.

(c) Assignment of all replicated tracks shorter than 1Mb to the IR closest to that track's center revealed a cone shape around each IR center, consistent with symmetric replication fork progression. In contrast to **(b)**, some replication tracks centered at the indicated IR extend into a neighboring IR (likely reflecting passive replication of the neighboring origin). This larger set of replication tracks was used to determine the location of the IR: for each track, the center position was assigned as the location of replication initiation in that cell, and the IR was defined as the region between the 25th and 75th percentile of the range of initiation sites across cells. Black lines indicate the center (50th percentile) of the IR.

(d) The location of each IR was identified at kilobase scale (median width: 87.5kb). IRs that were supported by fewer than 5 replication tracks were excluded to avoid skewing the distribution to the left.

(e) 79.2% of IRs could be localized to a region 100kb or narrower. In the example shown, 58 replication tracks were identified that overlapped the IR. The midpoint of these tracks fell within a 60kb range (dotted lines).

(f) Broadly localized IRs may reflect the presence of multiple distinct initiation events that were not disambiguated, technical noise, or mild asymmetry in replication fork progression. In the example shown, 62 cells were identified that overlapped the IR. Visually there appear to be multiple distinct clusters of track midpoints. See Figure S2.

Based on the conclusion that noise in individual cells was likely contributing substantially to variation in initiation site location, we devised a novel approach to cluster these sites into larger initiation regions (IRs) shared across cells, which did not rely solely on a 1Mb length cutoff to determine which replication tracks were informative about individual origins and which represented the activity of multiple independent origins. To accomplish this, replication tracks were sorted from shortest to longest, and sequentially grouped together with other, overlapping replication tracks. This algorithm prioritized information from shorter tracks over longer tracks at each locus; thus, whenever a replication track overlapped two preexisting groups supported by shorter replication tracks, it was excluded from use in defining IR locations. By this process, we identified a total of 7,482 IRs.

As noted above, single-cell pileups corresponded visually to peaks in the S/G1 aggregate replication timing profile (Figure 2; Figure 3a). Indeed, 91.4% of peaks in the aggregate profile coincided with an IR. Of these aggregate peaks that overlapped an IR, 51.1% corresponded to multiple IRs (*e.g.*, Figure 3b, left), while the remaining 49.9% corresponded to a single IR (*e.g.*, Figure 3b, right). This suggests that origins are often clustered in hotspots along the chromosome; the replication-timing peaks corresponding to single IRs could either be regions of lower origin density or, conversely, represent hotspots too dense for individual origins to be detected at this resolution. Thus, single-cell data are concordant with the ensemble replication timing profile, but also caution that smoothing of ensemble profiles likely removes information about distinct initiation sites.

We then assigned all replication tracks shorter than 1Mb to the IR whose center was closest to the midpoint of the track. This includes tracks that potentially overlap multiple fired IRs; however, when all replication tracks assigned to a given IR were sorted by length, a symmetric cone was observed around the IR center (Figure 3c), consistent with sister replication forks progressing away from a single origin or tight cluster of origins at the IR center with similar processivity. For each IR, we calculated how tightly the midpoints of these replication tracks were clustered to assess how precisely the most probable initiation site within the IR was identified. IRs were localized to a median width of 87.5kb (~4 windows; Figure 3d), which corresponds to an inter-IR distance of 120kb to 1.3Mb (median: 260kb). Most IRs (79.2%) were 100kb or narrower (*e.g.,* Figure 3e). Visual inspection of broad IRs (>120kb) suggested that many contain multiple initiation events that were grouped together because of overlap between replication tracks (Figure 3f; Figure S2). Thus, while we cannot determine whether IR width (and variability in IR width) reflects technical noise, inconsistency between cells in the precise location of initiation, or mild asymmetry in sister replication fork progression, we conclude that initiation events are relatively localized, and that at least some of IR widths are likely overestimated. Localized initiation regions are also apparent in the early S fractions of the 10- and 100-fraction profiles (Figure 1d), where the impacts of noise are averaged across many cells.

In our analysis of IRs, we did find evidence of ectopic replication initiation: only 29.7% of IRs contained a peak in the S/G_1 aggregate profile, and 31.2% of IRs were supported by a single replication track. However, these potentially ectopic events comprised a small fraction of all initiation events. Rather, 2,640 IRs (35.3%) accounted for 90% of the replication tracks, indicating that about a third of the IRs are responsible for the vast majority of initiation events genome wide. Thus, contrary to previous studies that analyzed single-cell replication profiles at the level of large chromosomal "domains" [15,16], our data reveal localized initiation regions, which we assume correspond to individual, or tight clusters of, replication origins.

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

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

We first asked whether the single cell data were compatible with strictly determined replication timing, such that every cell initiates replication at every IR in the same order. Strict determinism provides a straightforward prediction to test: the number of IRs replicated in any given cell should predict *which* IRs have been replicated in that cell. For example, a cell that has replicated one IR is predicted to have replicated the IR with the earliest replication timing; a cell that has replicated 100 IRs is predicted to have replicated the 100 IRs with earliest replication timing; and so on. To test how well these predictions matched our data, we counted the number of IRs that were replicated in each cell and used that to assign each IR in that cell an "expected" state – either unreplicated or replicated – assuming that the firing order was fixed (Figure 4a, b). For a given IR, the observed replication state matched the predicted state in the vast majority of cells (Figure 4c), indicating that the firing of IRs in single cells follows a highly predictable order. However, we did observe that, on average, an IR differed from its expected state in 10.9% of cells (Figure 4d). Thus, we can formally rule out the hypothesis that replication timing is strictly determined; IR firing order at the single-cell level is orderly but not entirely predictable.

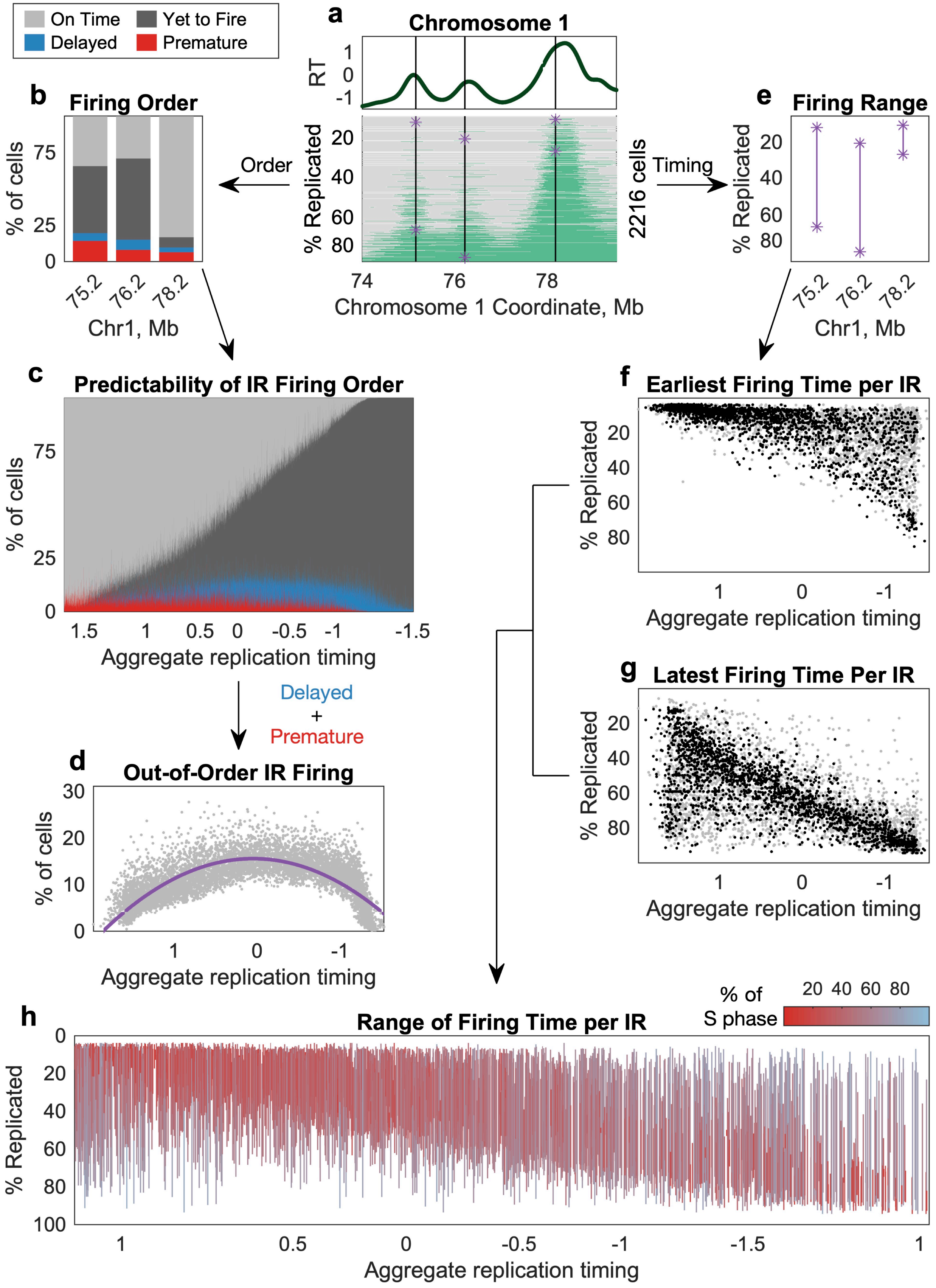
On Time
Yet to Fire
Delayed
Premature
a
Chromosome 1
RT
1
0
-1
% Replicated
20
40
60
80
74
76
78
Chromosome 1 Coordinate, Mb
2216 cells
Order
Timing
b
Firing Order
% of cells
75
25
0
75.2
76.2
78.2
Chr1, Mb
e
Firing Range
% Replicated
20
40
60
80
75.2
76.2
78.2
Chr1, Mb
c
Predictability of IR Firing Order
% of cells
75
25
0
1.5
1
0.5
0
-0.5
-1
-1.5
Aggregate replication timing
Delayed
+
Premature
d
Out-of-Order IR Firing
% of cells
30
20
10
0
1
0
-1
Aggregate replication timing
f
Earliest Firing Time per IR
% Replicated
20
40
60
80
1
0
-1
Aggregate replication timing
g
Latest Firing Time Per IR
% Replicated
20
40
60
80
1
0
-1
Aggregate replication timing
h
Range of Firing Time per IR
% of
S phase
20 40 60 80
% Replicated
0
20
40
60
80
100
1
0.5
0
-0.5
-1
-1.5
1
Aggregate replication timing

Figure 4. Variation in the order and timing of replication initiation in single cells across S phase.
(a) An example region is shown to illustrate the analyses of both IR firing order **(b-d)** and IR firing time (**e-h)**. Black lines indicate the four IRs used as examples in **(b)** and **(e)**. Purple asterisks indicate the earliest cell in which an IR was observed to fire and the latest cell in which it was observed to be unfired.
(b) IRs differ in their degree of consistency across single cells. IRs were ranked from earliest to latest, allowing prediction of which would have fired in each single cell under a strict ordering of IR firing. These predictions were then compared to the single cell data. Cells that have replicated an IR that was not predicted to fire in that cell are considered to have "premature" firing (red), while those that have not replicated an IR predicted to have fired already are considered to have "delayed" firing (blue).
(c) IRs are fired in the expected order in the majority of single cells. However, those 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 IR, sorted from the earliest (left) to the latest (right).
(d) Variability in IR firing is most variable for those that are expected to fire in the middle of S phase. On average, IRs behaved differently than expected (either firing prematurely or delayed) in 10.9% of cells (range: 0.2-27.7%). Each dot represents one IR. Purple line: second-order polynomial fit.
(e) The range of IR firing was defined as spanning from the earliest cell in which an IR was observed to have fired to the latest cell in which it had yet to fire (asterisks in **(a)**). The percent of the genome replicated in each cell was used as the proxy for S phase progression.
(f) For each IR, we identified the least replicated (*i.e.*, earliest) cell containing a replication track assigned to that IR. 95% of IRs were observed to fire in a cell <50% replicated. If the second earliest cell for an IR was within 10% of S phase from the earliest cell, that IR's earliest firing time was considered "corroborated" (black dot); all other IRs are gray.
(g) For each IR, we identified the most replicated (*i.e.*, latest) cell that had not yet replicated the region containing a given IR. If the second latest cell for an IR was within 10% of S phase from the latest cell, that IR's latest firing time was considered "corroborated" (black dot); all other IRs are gray.
(h) IRs with earlier aggregate replication timing tended to have narrower ranges of firing times than those with late aggregate replication timing. Each vertical line represents the range for one IR, color-coded by the % of S phase during which that IR fires (*i.e.*, the length of the line). A small number of constitutively late IRs (short red lines with late aggregate replication timing) can be observed. Only IRs whose earliest and latest values were corroborated by a second cell (*i.e.*, those in black in **(f, g)**) are shown.

Having observed variation across cells, we next asked if that variation was uniform across S phase or concentrated at specific times during S phase. We found a parabolic relationship between replication timing of an IR and the proportion of cells that fired that IR out of the strictly determined order. Thus, variability was lowest at the beginning and end of S phase and highest in the middle of S phase, such that 83.5% of the above-average variability occurred in the 53.0% of IRs with aggregate replication timing between 1 and -1. A similar parabolic trend was previously described by Takahashi *et al.*[15] and was robust in our larger sample size.

We next considered the *extent* of firing time variability, asking when in S phase IRs fire in the instances when they fire out of the predicted order. To answer this question, we identified the least-replicated (*i.e.*, earliest) cell in which an IR was observed to fire and the most-replicated (*i.e.*, latest) cell in which it had yet to fire (Figure 4e). We found that there was an association between the earliest time that an IR fired and its replication timing in the S/G_1 aggregate replication profile (r = -0.64; Figure 4f), indicating that IRs with late aggregate timing tended to start replicating later in S phase than those with early aggregate timing. However, most IRs were observed to have fired in a subset of early S-phase cells: 48% of GM12878 IRs fired at least once in a cell with <10% of its genome replicated, 83% in a cell with <25% replicated, and 95% in a cell <50% replicated. Thus, many IRs with late aggregate timing were not restricted to firing in late S phase. There was also an association between how late into S phase an IR remained unfired and its replication timing in the S/G_1 aggregate replication timing profile (r = -0.67; Figure 4g). Thus, IRs with early aggregate replication timing tended to finish firing across all cells relatively early in S phase, while those with late aggregate timing tended to remain unfired into late S phase.

After determining these earliest and latest cells for each IR, we considered them in a paired manner to determine the *range* of firing times of each IR (Figure 4h). Given that range is sensitive to outliers (*i.e.*, a duplication called "replicated" or a deletion called as "unreplicated"), we focused on IRs for which the minimum and maximum values were "corroborated" by a second cell within 10% of S phase from the extreme. IRs with early aggregate replication timing tended to first fire in early S phase and to complete their replication before the genome was 50% replicated. In contrast, IRs with late aggregate replication timing tended to also first fire in early S phase, but to remain unfired in some cells until the end of S phase. Therefore, the firing time of IRs with early aggregate replication timing was constrained to early S phase, while IRs with late aggregate replication timing appeared to be less constrained. However, we did observe a small number of IRs that fired exclusively in late S phase; these had a more constrained range. We thus proceeded to further analyze these different behaviors in regions with late aggregate replication timing.

Initiation events that appear "late" in ensemble measurements comprise a heterogeneous population of IRs

Our analysis of single-cell replication timing indicated that IRs are fired in a consistent order across most cells, but that IRs with late S/G_1 aggregate replication timing fire across a larger portion of S phase relative to those with early S/G_1 aggregate timing (Figure 4h). We further dissected the nature of these IRs with large firing ranges to better understand whether we were capturing rare occasions of extremely premature firing or perhaps observing a capacity of IRs to fire throughout S phase. In other words: do these IRs fire substantially ahead of schedule in some cells, or do they not have a scheduled time to fire at all?

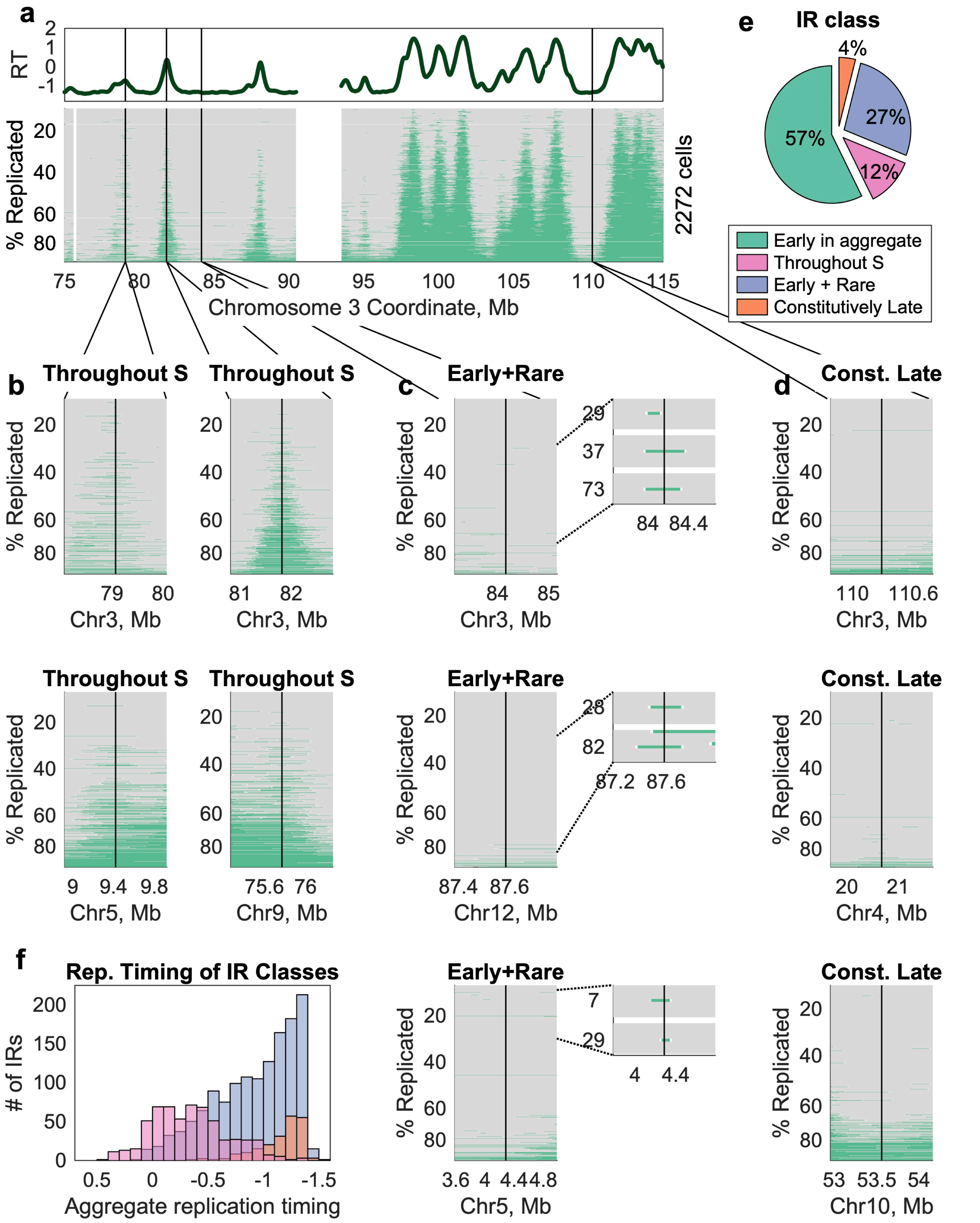
a
RT
2
1
0
-1
% Replicated
20
40
60
80
2272 cells
75
80
85
90
95
100
105
110
115
Chromosome 3 Coordinate, Mb
e
IR class
4%
27%
57%
12%
🟩 Early in aggregate
🟪 Throughout S
🟦 Early + Rare
🟧 Constitutively Late
b
Throughout S
Throughout S
c
Early+Rare
d
Const. Late
% Replicated
20
40
60
80
79
80
Chr3, Mb
81
82
Chr3, Mb
29
37
73
84
84.4
84
85
Chr3, Mb
110
110.6
Chr3, Mb
Throughout S
Throughout S
Early+Rare
Const. Late
9
9.4
9.8
Chr5, Mb
75.6
76
Chr9, Mb
28
82
87.2
87.6
87.4
87.6
Chr12, Mb
20
21
Chr4, Mb
f
Rep. Timing of IR Classes
# of IRs
200
150
100
50
0
0.5
0
-0.5
-1
-1.5
Aggregate replication timing
Early+Rare
7
29
4
4.4
3.6
4
4.4
4.8
Chr5, Mb
Const. Late
53
53.5
54
Chr10, Mb

Figure 5. Three distinct classes of IRs with late aggregate replication timing.
(a-d) "Late" IRs can be classified into three classes based on their behavior across single cells: some fire throughout S phase (**b)**, some fire rarely but often fire early when they do fire (**c**), and some were never observed to fire early **(d)**. The IRs indicated with black lines in (a) are shown in the top row of (b), (c), and (d). Additional examples are shown below.
(e) 27% of IRs with late aggregate replication timing fire infrequently but with a preference for early S phase, while 12% fire throughout S phase. Constitutive late firing is rare (4% of IRs).
(f) IRs that fire throughout S phase (pink) tend to have earlier replication timing than the other two classes of IRs, while those that were constitutively late (orange) had the latest average replication timing.

We found that each of these two explanations for large range of firing times were supported by a substantial fraction of IRs, and that neither behavior was sufficient to explain all cases on its own (Figure 5a). This indicates that some IRs with late aggregate replication timing tend to fire late but sometimes fire very early, while other IRs with late aggregate replication timing fire at many different times in S phase. Specifically, 12% of IRs (27.1%of late IRs) fired inconsistently throughout S phase (Figure 5b, e), with earlier aggregate timing corresponding to more cells firing the IR (compare Figure 5b top left *vs*. top right). On the other hand, 27% of IRs (63.6% of late IRs) fired rarely and almost all the replication tracks associated with these IRs were from cells <50% replicated (Figure 5c, e). Finally, 4% of IRs (9.4% of late IRs) were never observed to fire in a cell <50% replicated (Figure 5d, e). Comparing these three classes, IRs that fired throughout S phase tended to have the earliest aggregate replication timing (median: -0.33 and as early as 0.46), while constitutively late IRs had the latest aggregate timing (median: -1.22; Figure 5f). These unexpected results demonstrate that the late-replicating regions observed in ensemble assays contain origins with heterogeneous firing behavior; these results cannot be fully explained by either a deterministic timing model (which posits these regions contain constitutively late-firing origins) or existing stochastic firing models (which posit that these regions contain low-efficiency origins that become increasingly likely to fire as S phase progresses[11]).

Comprehensive measurement of single-cell DNA replication timing across human cell lines in thousands of cells throughout S phase

Having established a workflow for high-throughput replication analysis of unsorted cells, we performed whole-genome sequencing of 9,658 single cells across eight additional cells lines: two LCLs, three embryonic stem cell lines (ESCs), and three cancer cell lines. As with GM12878, we performed *in silico* cell sorting to distinguish replicating and non-replicating cells within each library (Figure S3a). For each cell line, we generated an aggregate S/G_1 profile that was highly correlated to an S/G_1 bulk replication-timing profile for the same cell line (r = 0.84-0.97; Figure S4). We then generated replication profiles for between 110 and 508 S-phase cells across the different cell lines (Figure 6a, b; Figure S4).

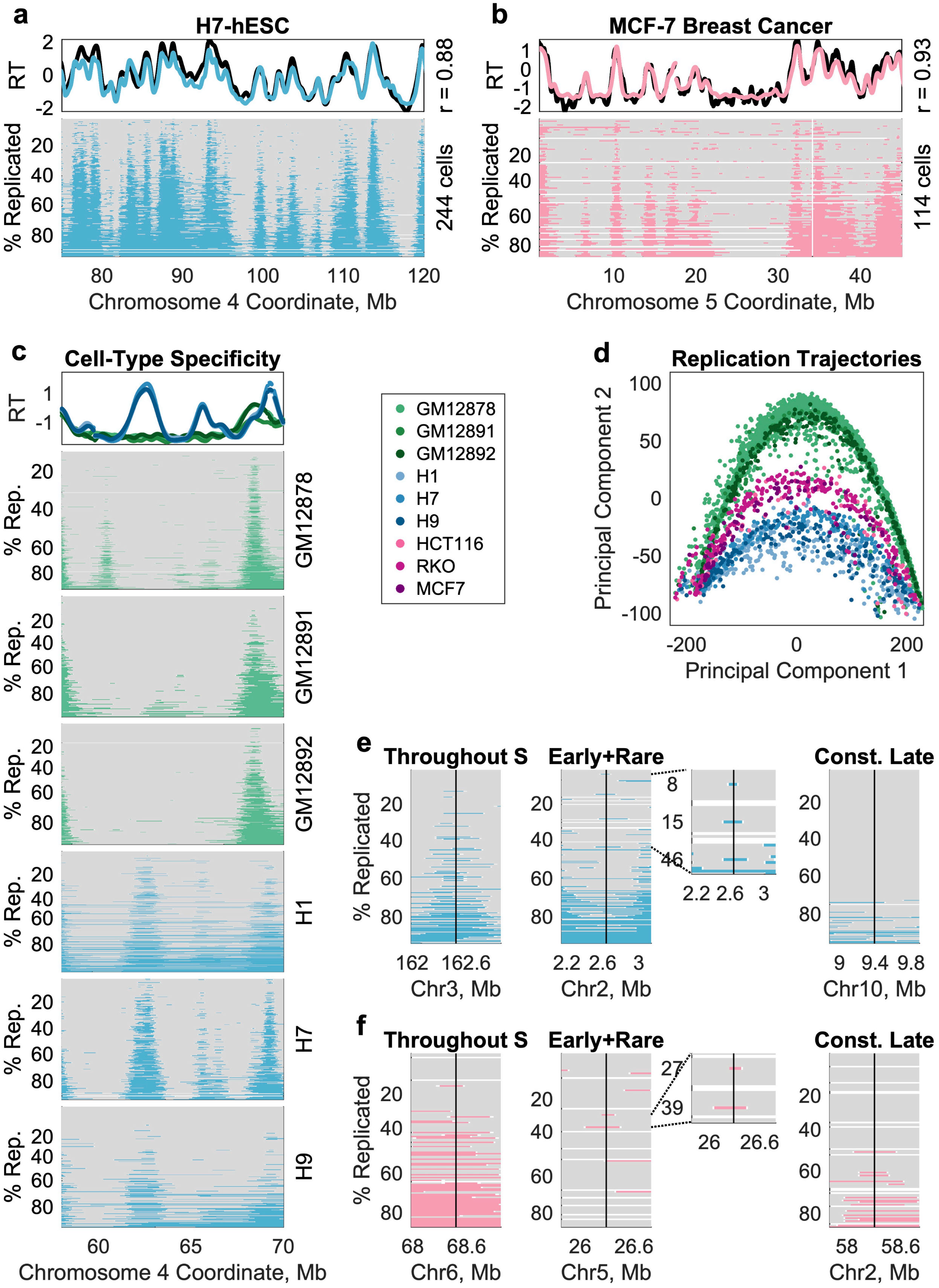
a
H7-hESC
RT
2
0
-2
r = 0.88
% Replicated
20
40
60
80
244 cells
80
90
100
110
120
Chromosome 4 Coordinate, Mb
b
MCF-7 Breast Cancer
RT
1
0
-1
-2
r = 0.93
% Replicated
20
40
60
80
114 cells
10
20
30
40
Chromosome 5 Coordinate, Mb
c
Cell-Type Specificity
RT
1
-1
% Rep.
20
40
60
80
GM12878
GM12891
GM12892
H1
H7
H9
60
65
70
Chromosome 4 Coordinate, Mb
● GM12878
● GM12891
● GM12892
● H1
● H7
● H9
● HCT116
● RKO
● MCF7
d
Replication Trajectories
Principal Component 2
100
50
0
-50
-100
-200
0
200
Principal Component 1
e
Throughout S
Early+Rare
Const. Late
% Replicated
20
40
60
80
8
15
46
2.2 2.6 3
162 162.6
Chr3, Mb
2.2 2.6 3
Chr2, Mb
9 9.4 9.8
Chr10, Mb
f
Throughout S
Early+Rare
Const. Late
% Replicated
20
40
60
80
27
39
26 26.6
68 68.6
Chr6, Mb
26 26.6
Chr5, Mb
58 58.6
Chr2, Mb

Figure 6. Comprehensive measurement of single-cell replication timing across cell types.
(a, b) As in Figure 1e, 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: bulk-sequencing consensus replication timing profiles for LCL (green) and hESC (blue). Lower panels: single cell data reveals that the bulk-sequencing peaks at ~62Mb and ~65.5Mb reflect the presence of hESC-specific initiation sites.
(d) Single cells follow cell-type-specific trajectories of S-phase progression, as determined by principal component analysis (PCA). PCA was performed on replication states in all genomic windows across autosomes. PC1 corresponds to the % of the genome replicated (r = 0.99), with negative values of PC1 reflecting early S phase and positive values reflecting late S phase. Cell types segregate along PC2. Each dot represents a single cell.
(**e, f**) All three categories of IRs with late aggregate replication timing described in Figure 5 were also observed in H7 **(e)** and MCF-7 **(f)**.

The aneuploid breast cancer cell line MCF-7 highlights the broader applicability of *in silico* sorting. While we apply this method to focusing our analysis only on replicating cells, it is also valuable in single-cell analysis of CNAs in cancer. In that context, it is necessary to *remove* replicating cells prior to CNA calling, since both replication and duplications/deletions affect copy number estimation. MAPD has previously been used to filter out "noisy" cells in this type of analysis[30]. However, aneuploidy inflates MAPD values (Figure S3a, compare MCF-7 to other cell lines), making it difficult to effectively set a threshold for filtering. In contrast, explicit modeling of G_1 and S cell populations with distinct linear relationships between read coverage and MAPD efficiently discriminates cells of interest (either for replication analysis or CNA analysis; Figure S3b).

It has been well demonstrated in ensemble experiments that cell types have distinct replication timing programs, which are shared by cell lines of the same cell type[3,31,32]. Thus, we asked whether cell-type differences among these nine cell lines were preserved at the single cell level. Indeed, cell-type differences among the aggregate replication timing profiles were found to be consistent at the single cell level (Figure 6c, Figure S5). These differences in replication state between cell types were sufficient to cluster single cells by cell line and cell type (Figure 6d), suggesting that individual cells of the same cell type follow a similar trajectory through S phase. Two types of replication timing differences can be observed at the ensemble level: differences in peak locations (*i.e.*, in the location of fired origins) and differences in peak amplitude (*i.e.*, in the timing at which a shared origin is fired). We observe both of these classes of variation at the single cell level: cell-type-specific peaks in the S/G_1 aggregate profile that reflect the presence of a cell-type-specific initiation site (*e.g.*, Figure S5a, right) and peaks of different amplitude in the S/G_1 aggregate that correspond to early vs. late firing of a shared initiation site (*e.g.*, Figure S5b, left). Most intriguingly, we also observe a novel type of cell-type difference invisible to ensemble profiling methods: a subset of cell-type differences that appears to be driven by inconsistent usage of an initiation site in one cell type (*e.g.*, Figure S5a, left ~196.1Mb).

We proceeded to call IRs in each cell line and repeated the above analyses of IR order and timing variability. Despite having ~10 times fewer cells relative to GM12878, we were able to identify 1,676-5,077 IRs (compared to 7,482 in GM12878) per cell line in all cell lines except for HCT-116 (discussed below). These IRs were slightly broader than the GM12878 IRs, but still localized (median: 110kb-225kb; Figure S6a). This suggests that increasing the number of cells analyzed will likely yield additional IRs in all cell lines and also further narrow their localization.

Patterns of initiation site localization and timing variability across cell lines were broadly consistent with those observed in GM12878, even though the specific locations of IRs differed between cell types. IRs were fired in a predictable but not fixed order (Figure S6b) that was more disordered for those IRs with mid-S-phase aggregate timing (Figure S6c). With regards to firing time, IRs with late aggregate replication timing fired early in S phase in a subset of cells, although with the smaller sample size, fewer IRs had multiple cells corroborating this behavior (Figure S7a, b). This is consistent with how rarely these events occurred *per IR* in GM12878 and suggests that these events would be observed more frequently in other cell lines when looking across a larger number of cells. However, the fact that so many rare events are observed even in a sample size of ~200 cells suggests that the full scope of variability remains underestimated, including in GM12878.

Finally, the three classes of late IRs were present in each cell line (except HCT-116), and two features were common between GM12878 and other cell lines. First, rarely used IRs with a preference for early firing were more common than IRs that fired throughout S phase; second, a small fraction (4-11% in most cell lines) of IRs were constitutively late (Figure 6e, f; Figure S7c).

The outlier cell line was the colorectal cancer cell line HCT-116, for which we recovered only 110 replicating cells and identified only 758 IRs. In addition to wider IRs, with a median width of 280kb, 78% of IRs identified in HCT-116 had an aggregate replication timing earlier than the genome-wide average replication timing. (In other cell lines, this value was close to 50%, in line with the genome-wide replication timing values.) This bias toward discovering IRs in early-replicating regions creates the impression that variability increases across S phase, particularly when examining HCT-116 alone. These results are presented alongside those of the other cell lines to illustrate how a low cell count can bias IR identification and conclusions drawn from subsequent analyses. However, while not particularly informative about IRs in late-replicating regions, the data from HCT-116 are not incompatible with the trends observed specifically for early IRs across cell lines.

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

Discussion

While ensemble replication profiling methods cannot capture (and may be confounded by) cell-to-cell heterogeneity, previous single-molecule and single-cell methods have been largely limited in their throughput or accuracy. Here, we report a scalable method for analysis of thousands of single replicating cells, across multiple cell lines, and at kilobase resolution. We describe an *in silico* strategy to sort cells across S phase, analogous to and more accurate than traditional flow cytometry, and demonstrate how this method enables simultaneous analysis of replication initiation at population, subpopulation, and single-cell resolutions. In addition, by focusing specifically on replication initiation events called from single cells, we are able to identify which cells are informative about which replication initiation sites, capturing information that is analogous to that collected from lower-throughput single-molecule studies. In a parallel study, Gnan *et al.*[33] developed a similar approach to use single-cell sequencing data to infer DNA replication timing at large scale.

We find that single cells primarily initiate replication at consistent loci, corresponding to peaks in the replication timing profile. Across cells, we are able to pinpoint the locations of 79.2% of these initiation events to regions no larger than 100kb (likely overestimated due to low coverage), challenging the model that there are megabase-long replication domains that are replicated simultaneously[5,6]. Analogously, our data do not support the existence of large constant replication regions (CTRs)[34]. 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[7,10]. Our observation that even tight peaks in ensemble replication-timing profiles often encompass multiple discrete single-cell initiation events lends further credence to the argument that initiation events are even more localized than measured here, with the caveat that we cannot distinguish in our data between single origins and tight clusters of nearby origins. We find evidence for ectopic initiation from regions outside these commonly used initiation regions (as in [7]), although these events comprise a small fraction of overall events. While many previous studies of mammalian replication origins relied on biochemical enrichments of DNA synthesis events and are therefore more prone to false-positive identification of apparent initiation events, single-cell DNA sequencing more reliably represents productive and internally-validated DNA replication[1].

While spatial variability in replication initiation is rare, temporal variation is more common. In general, initiation regions (IRs) expected to fire in the middle of S phase are more variable than those expected to fire earlier or later, consistent with previous reports[15]. At the level of individual IRs, we find that many, particularly those with early aggregate replication timing, have a preferred time of firing that is captured by the aggregate replication timing profile. IRs with early aggregate replication timing tend to be fired in all cells early in S phase, while those with late aggregate replication timing fire across a broader range of S phase. We further find that late replicating IRs can be divided into multiple classes, with only a small subset (<10%) firing constitutively late. Instead, most late IRs can and do fire early – sometimes rarely and sometimes often.

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

While our data are consistent with an important role for origin firing efficiency in determining replication timing, the distinct classes of initiation regions we describe also highlight a shortcoming of considering origin efficiency at the level of individual loci: while low-efficiency origins would be expected to rarely fire in early S phase, their probability of firing should remain constant or even increase as S phase progresses.[11] In other words, origins in late-replicating regions of the genome should fire throughout S phase. Instead, we see that the majority (63.6%) of the inefficient IRs have a low probability of firing in early S phase, and a negligible probability of firing later in S phase, suggesting that the context of replication initiation changes across S phase in a manner that has not been previously characterized.

Our results suggest origin-specific firing efficiencies play a key role in producing the replication timing program; as such, they underscore the value of future work parsing out the contributions of DNA sequence, gene expression, chromatin accessibility, and doubtless other factors to these firing efficiencies. At the same time, a future model for replication timing must also explain why many origins appear to have their highest probability of firing at the beginning of S phase, rather than becoming increasingly likely to fire as S phase progresses – and also why that does not result in large regions of under-replication that persist into G_2 phase, as modeled in [11].

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[18,19,35], this approach promises to become ever more informative for the understanding of the DNA replication timing program.

Methods

Cell Culture

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

Library Preparation and Sequencing

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

For both sorted and unsorted libraries, isolation, barcoding, and amplification of single-cell genomic DNA was performed on the 10x Genomics Chromium Controller instrument, using the 10x Genomics Single Cell CNV Solution kit (10x Genomics, Pleasanton, CA, USA). Paired-end sequencing was performed for 100 cycles with the Illumina NovaSeq 6000 (10x Genomics), 150 cycles with the Illumina HiSeq X Ten (GENEWIZ, Inc., South Plainfield, NJ, USA), or 36 or 75 cycles with the Illumina NextSeq 500 (Cornell University Biotechnology Resource Center, Ithaca, NY, USA). For libraries sequenced multiple times, FASTQ files were merged prior to downstream processing. See Table S1 for details.

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 95th percentile of the high-count abundances.

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

Processing of sequencing reads

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

Computational identification of G_1 cells and definition of G_1 windows

Reads were counted in fixed size windows of 20kb. After removing low-mappability windows (in which fewer <75% of nucleotide positions were uniquely mappable[36]), sets 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 coverage and scaled MAPD. For each sequencing library, an expectation-maximization procedure was used to fit the data as a mixture of two Gaussian functions. The linear fit with the lower y-intercept was assumed to model the G_1 relationship between coverage and scaled MAPD, and cells with a residual ≤ 0.05 from this model 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.*[22]. 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 between libraries and was determined as the number of cells that would define windows of ~20Kb.) This was performed independently for each sequencing library prepared from the same cell line. Per-cell read counts were calculated in these G_1-windows, to account for mappability and GC-content bias, as well as any copy-number variations that were common to many cells within the library.

Finally, we identified and filtered out cell-specific copy-number aberrations (CNA). To do this, we fit a two-component mixed Poisson model to aggregated read counts (15 windows, ~300Kb), and searched for the genomic region with the lowest probability of being observed under either rate coefficient, λ. If the median probability of each window within this region was less than the median probability of all windows genome-wide, we determined it to be a CNA and masked the read counts in that region. This process was performed iteratively until no new regions were discovered. Cells with an autocorrelation in read counts >0.15 after filtering were assumed to have residual undetected CNAs and were excluded from analysis.

Replication state inference

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

Because the HMM does not model the expected two-fold relationship between replicated and unreplicated regions, we assessed the quality of the HMM output using this ratio. Specifically, we calculated the ratio between the average number of reads in windows assigned as replicated to the average number of reads in windows assigned as unreplicated. To be included in further analysis, this ratio was required to be between 1.5 and 2.5.

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

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

Bulk-sequencing replication timing profiles

Replication timing profiles from bulk sequencing assays were used to benchmark single-cell replication profiles. For GM18507, an LCL consensus profile[22] was used. For all other cell lines, a profile for the specific cell line was used. For Illumina Platinum LCLs (GM12878, GM12891, and GM12892)[37] and hESCs (H1, H7, and H9)[8], these data are previously published.

Aggregate replication timing profiles

For each cell line, we generated an aggregate S/G_1 profile, as in [22], 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.

Sub-S phase fraction profiles

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

Identification of initiation regions

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

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

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

Variation in firing order across cells

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

Variation in firing time across cells

To determine the range of firing orders for each IR, we identified the earliest cell containing a replication track attributed to an IR, and the latest cell in which the center of the IR was inferred to be unreplicated (after excluding outlier cells that had not replicated any of the neighboring IRs). The percent of the genome replicated in each of these cells was used as a proxy for time during S phase. Given that range is a metric extremely sensitive to outliers, we considered an IR's range to be "corroborated" if a second cell was observed within 10% of its earliest and latest firing time. We focused on these IRs with corroborated ranges in subsequent analyses.

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

Data availability

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

Code availability

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

Acknowledgements

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

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 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).

6 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).

7 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).

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

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 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).

23 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).

24 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).

25 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).

26 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).

27 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).

28 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).

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

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

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

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

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

34 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).

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

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

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