RESEARCH

Modeling single cell DNA replication dynamics and aneuploidy in genomically unstable cancers

Adam C. Weiner[1,2], Marc J. Williams[1], Hongyu Shi[1,3], Sohrab P. Shah[1*] and Andrew McPherson[1*]

*Correspondence:
shahs3@mskcc.org;
mcphera1@mskcc.org
[1]Computational Oncolocy, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA
Full list of author information is available at the end of the article

Abstract

DNA replication is a highly coordinated cell cycle process that can become dysregulated in cancer, increasing both proliferation and mutation rates. Single-cell whole genome sequencing holds potential for studying replication dynamics of cancer cells; however, computational methods for identifying S-phase cells and inferring single-cell replication timing profiles remain immature for samples with heterogeneous copy number. Here we report a new method, PERT, which jointly infers replication and somatic copy number states of S-phase cells. This method enabled us to analyze the replication dynamics of >10,000 S-phase single-cell genomes across various triple negative breast cancers and cell lines with subclonal copy number heterogeneity. We show that PERT robustly predicts cell cycle phase, quantifies replication timing variability, and approximates relative proliferation rates between tumor subclones. Our results illuminate how aberrant DNA replication processes can both drive and result from evolution of human tumors.

Keywords: DNA replication; genomic instability; single-cell whole genome sequencing; probabilistic modeling

Background

DNA replication occurs during the synthesis (S)-phase of the cell cycle in preparation for cell division. Replication origin firing and fork progression is highly coordinated in healthy cells but can become dysregulated in diseases of uncontrolled cell proliferation such as cancer [1, 2, 3]. Cancer cells often bypass cell cycle checkpoints during mitosis (M)- and S-phases to achieve high proliferation and mutation rates. In genomically unstable cancers, these mutations often come in the form of somatic copy number alterations (CNAs) [4, 5]. Scaled single-cell whole genome sequencing (scWGS) has enabled analysis of 1000s of genomes from a single tumor sample [6, 7, 8, 9, 10, 11] but computational identification of S-phase cells and accurate quantification of downstream DNA replication dynamics remain challenging for genomically unstable samples. Earlier scWGS studies of intratumor evolution performed fluorescence activated cell sorting (FACS) on DAPI intensity to isolate polyploid G1-phase cells prior to sequencing to prevent transient replication copy number changes from confounding study of somatic copy number changes [10, 11]; however, understanding the impact of DNA replication on tumor evolution requires sequencing all cells to computationally discern the two signals.

Two key DNA replication dynamics theorized to be dysregulated in cancer are replication timing (RT) and distribution of cell cycle phases. RT is the relative

time within S-phase that regions of the genome replicate at submegabase resolution [12, 13, 14]. Early RT regions correspond to active chromatin found in high-throughput chromosome conformation capture (Hi-C) A compartments while late RT regions correspond to repressive chromatin found in Hi-C B compartments [15, 16]. RT is heterogeneous across different tissue types and is correlated with mutational patterns as early and late RT loci have different DNA damage and repair rates [17, 18]. Copy-neutral structural variation in cancers have been shown to alter RT [19, 20]; however, computational methods to study bulk and single-cell RT (scRT) on genomically unstable samples remain immature. For this reason, very little is known about how RT is conserved after CNAs emerge and how RT itself contributes to the CNA landscape. Separately, distribution of cell cycle phases is important for studying cancer as it relates to cell proliferation. G1-, G2- and the senescent G0-phases of the cell cycle are much more variable in duration than S- and M-phases, causing highly proliferative samples to have more S- and M-phase cells. Pathologists have leveraged this phenomenon as visual quantification of mitotic cells in histological sections remains the gold-standard method to evaluate tumor proliferative activity and grade [21, 22]. Additionally, transcriptional signatures associated with S-phase activity are upregulated in highly proliferative tumors [23]. Platinum chemotherapies such as cisplatin also leverage cell cycle dynamics to target S-phase enriched hyperproliferative cancer cells as they generate DNA adducts which block replication and trigger downstream cell death [24]. These phenomena suggest that investigating the cell cycle distributions of individual tumor subclones would provide a missing link to connect a clone's genotype to treatment-naive fitness and chemosensitivity.

Replication timing can be assayed through bulk and single-cell whole genome sequencing of S-phase cells [25, 26, 27, 28, 29], enabling potential joint study on the cell-to-cell variability of CNAs and scRT programs. Methods developed to infer copy number (CN) states from scWGS [6, 30, 31, 32, 33] succeed for G1/2-phase cells but fail for S-phase cells due the enrichment of replication origins in high GC genomic loci which confounds GC bias correction. Similarly, computational methods developed to investigate replication dynamics from scWGS data struggle to appropriately deconvolute GC and RT signals or account for subclonal CNAs. For cell cycle phase, the classifier reported in Laks *et al* [7] uses a random forest classifier and manually constructed features trained on two scWGS direct library preparation (DLP+) cell cycle sorted cell lines. This classification approach and limited training data inhibits performance in libraries that have different GC biases or ploidies from the training data, making it challenging to classify out-of-distribution libraries. For scRT estimation, Kronos [29] uses CN calls to threshold each bin in S-phase cells as replicated or unreplicated; however, it assumes that the replication signal remains unaltered by GC bias correction during CN calling and is only designed for cases in which all cells have the same somatic CN profile.

We present a new method, Probabilistic Estimation of single-cell Replication Timing (PERT), to jointly infer single-cell copy number and replication states from scWGS data. PERT uses a Bayesian framework that models observed read depth as a combination of somatic CN, replication, GC bias, and sequencing error. As input, PERT requires binned read depth and single-cell CN results; the user can

specify both the input CN caller and bin size. PERT uses a multi-step learning process with informative priors to favor parsimonious solutions in which S- and G1/2-phase cells have similar somatic CN profiles, S-phase cells follow a shared RT profile, and all cells in both phases having the same GC bias. Such modeling choices make PERT robust to clone- and cell-level CNAs and fluctuations in GC bias which can be platform dependent. Through extensive benchmarking, we demonstrate that PERT outperforms alternative approaches for scRT inference and S-phase classification. Applying PERT to both genomically unstable cell lines and patient derived xenografts of human cancers, we investigated the bi-directional relationship between CNAs and RT shifts, used RT shifts to phase X-inactivated alleles, and used clone cell cycle distributions to assess proliferation rates of individual tumor subclones on- and off-cisplatin.

Results

PERT: probabilistic estimation of single-cell replication timing

PERT jointly models the impacts of CN, replication, GC bias, and overdispersion on the observed read depth (Z) and uses a multi-step learning scheme to accurately infer somatic CN (X) and replication states (Y) for all bins ($1, ..., N$ cells, $1, ..., M$ loci) in a scWGS sample. PERT takes binned read depth and CN states from existing single-cell CN callers (i.e. HMMcopy [31], 10x CellRanger-DNA [9]) as input (Fig. 1a). PERT identifies an initial set of high-confidence G1/2-phase cells in which input CN states are assumed to accurately represent somatic CN. All remaining cells are assumed to have "unknown" CN states and phases as they are either in S-phase, have poor sequencing quality (i.e. chemistry failure), or have CN calling failures due to incomplete GC bias correction and/or ploidy overestimation. Cells are then passed into a probabilistic model which is trained in three steps. First, PERT leverages the observed read depth and somatic CN states from high-confidence G1/2 cells to learn library-level GC bias (β_μ, β_σ) and read depth overdispersion (λ) terms (Fig. S1a). Second, PERT uses CN prior concentrations (η) and conditions on terms learned in Step 1 to learn per-bin somatic CN and replication states in unknown cells (Fig. S1b). Third, PERT conditions on latent variables learned in Steps 1-2 to see if any S-phase cells were included in the set of high-confidence G1/2 cells (Fig. S1c). CN prior concentrations are set for each cell by using the consensus CN profile of the most similar G1/2 clone (clone method) or a composite scoring of the most similar G1/2 clone and cell CN profiles (composite method) (Fig. S1d). The probability that a given bin is replicated (ϕ) is modeled as a sigmoidal function through latent variables corresponding to each cell's time within S-phase (τ), each locus's replication timing (ρ), and a stochasticity term (α). After learning per-bin replication states in all cells, PERT thresholds cells into predicted S, G1/2, and low quality (LQ) phases by examining the fraction of replicated loci per-cell as G1/2-phase cells appear as fully replicated or unreplicated (Methods).

PERT is implemented in Python using Pyro [34] and is publicly available on github. Our implementation includes user tutorials and allows for manual tuning of priors and hyperparameters. See Fig. 1b for a description of all terms in the graphical model and Methods for additional mathematical, inference, and implementation details.

Evaluation of performance on simulated data

We benchmarked PERT's accuracy at inferring somatic CN, replication states, and cell cycle phase through quantitative simulation experiments. We simulated parameter sweeps over the number of clones, the cell-specific CNA rate, replication stochasticity, read depth overdispersion, GC bias coefficients, and RT profiles (Table 1). A representative example of true and PERT inferred CN and scRT profiles can be seen for dataset P5.8 which has 3 clones and a cell-specific CNA rate of 0.02, resulting in a replication state accuracy (bins correctly assigned as replicated or unreplicated) of 96.5% (Fig. 2a-d). We benchmarked PERT against the DLP+ cell cycle classifier presented in Laks *et al* for per-cell phase accuracy [7] and Kronos for per-bin scRT accuracy [29] as they are the only available softwares for predicting cell cycle and scRT states, respectively. The Laks classifier achieved similar performance to the initial version (AUC=0.92) on a train-test split after removing low importance sequence-level features not available in our simulated data, with its most important features remaining high correlations and slopes between per-bin GC content and read depth, indicating that it performs similarly on simulated and real scWGS data (Fig. S2a,b).

PERT significantly outperformed the Laks classifier and Kronos in all simulated datasets. The performance gap between PERT and Kronos was significant ($p_{adj} < 10^{-4}$) for all parameter combinations and increased as a function of cell CNA rate, number of clones, and overdispersion (Fig. 2e,f, S3). PERT's replication and CN accuracies decayed at high cell CNA and overdispersion (λ) rates but remained robust to extreme replication stochasticity (α), GC bias (β_{mu}), and clone number (Figs. S3,4). Kronos errors came at bins with subclonal or cell-specific CNAs whereas PERT errors arose from cell-specific CNAs not shared between S- and G1/2-phase cells and swapping of very early and late S-phase cells (Fig. S2c-g). PERT accurately estimated the fraction of replicated loci in all cells and thus achieved 93% phase accuracy across all simulated datasets with $\lambda = 0.75$, performing significantly better ($p_{adj} < 10^{-4}$) than the Laks classifier in all parameter sweeps (Fig. 2g,h, S5). Importantly, PERT's cell cycle phase accuracy was robust to fluctuations in GC bias slope ($\beta_{\mu,0}$), indicating that PERT's performance will not fluctuate between DLP+, 10x, and other scWGS modalities with unique GC bias relationships (Fig. S5i). Together this analysis demonstrates that PERT provides significant improvement upon existing methods to infer scRT states, and thus cell cycle phase, from scWGS data – particularly in cases where CNAs arise with subclonal structure.

PERT identifies distinct replication timing profiles in cell cycle sorted samples

We investigated PERT's ability to identify unique RT profiles within the same sample using cell cycle sorted DLP+ samples from two unrelated cell lines: lymphoblastoid GM18507 cells with diploid genomes and breast cancer T47D cells with aneuploid genomes [7]. Cells were sorted into G1-, S-, and G2-phase FACS fractions according to DAPI staining intensity prior to sequencing and we merged both lines into one sample for analysis with PERT (Fig. S6a,b). PERT found distinct CN number profiles for S-phase cells of each line and its predicted phases were highly concordant with FACS (Fig. 3a-f). These samples were enriched for mid-S-phase

cells due to conservative DAPI thresholds (Fig. S6c,d). We computed cell line pseudobulk RT profiles and found that 15% (794/5258) of loci had an absolute RT difference > 0.25 between GM18507 and T47D, consistent with each cell line having a unique RT program. RT has been shown to be influenced by nuclear organization, with genomic loci in inactive chromatin being late replicating and active chromatin being early replicating [12, 15, 16, 27]. Consistent with this we found cell line specific correlation between the inferred RT profiles and Hi-C A/B (active/inactive) compartments of T47D and other lymphoblastoid cell lines [35] (Fig. 3h). To ensure that the latent variable for global RT (ρ) did not prevent accurate inference of cell line RT profiles, we split the cell lines and ran PERT independently. We found that both cell lines had RT profile Pearson correlations of 0.99 between their merged and split PERT runs (Fig. 3i, S6e,f). Similarly, PERT inferred accurate clone-specific RT profiles for simulated data in which each clone had a unique ENCODE cell line RT profile [35] (Fig. S7). This analysis reveals that PERT can accurately infer replication states in samples where multiple distinct RT profiles are present.

PERT is robust to poor initialization and identifies potential FACS errors

We next assessed PERT's accuracy when many true G1/2-phase cells were excluded from the initial set of high-confidence G1/2-phase cells. This scenario is common for libraries with many low quality cells (i.e. chemistry failure) or G1/2-phase cells with CN calling errors (i.e. wrong ploidy selected). We thus devised a permutation experiment in which a subset of GM18507 and T47D FACS G1/2-phase cells were mislabeled as S-phase during initialization to examine whether PERT successfully recovered them as G1/2-phase. We found that >90% of all mislabeled cells were accurately recovered (PERT=G1/2) across all permutation datasets without compromising identification of S-phase cells and cell-line specific RT profiles (Fig. 4a,b, S8a). Mislabeled cells which were predicted to be PERT=S were disproportionately FACS=G2 and 80-95% replicated with orthogonal per-cell features concordant with S-phase (Fig. 4a, S8b-g). Additionally, we found many cell-specific CNAs in the set of PERT=G1/2, FACS=S cells (Fig. 4c). We hypothesize that many discrepancies between FACS and PERT phases were FACS errors since cells with unique CNAs possess higher or lower DAPI intensities than other cells in the same phase. This analysis demonstrates that PERT's performance is analogous, if not superior, to experimental cell cycle sorting and is capable of analyzing unsorted scWGS samples where many G1/2-phase cells are excluded from the initial high-confidence G1/2 set.

PERT as a data-driven S-phase classifier in unsorted libraries

We next investigated PERT's performance on previously published unsorted DLP+ data of genetically engineered mammary epithelial 184-hTERT and high grade serous ovarian cancer OV2295 cell lines which contain many clonal, subclonal, and cell-specific CNAs [7, 36] (Additional File 1). Previous analysis of this data used G1/2-phase cells to perform phylogenetic, allelic, and CN analysis but did not use S-phase cells to study DNA replication. We found that all samples had more late- than early-S-phase cells which is consistent with literature that most loci replicate during early S-phase while very late RT loci take much longer to replicate [37, 38] (Fig.

4d). When comparing PERT vs Laks predicted phases, the majority of Laks=G1/2 (14859/16165, 91.9%) and Laks=S (3383/5353, 63.2%) cells were concordant with PERT phase but 34.8% (1863/5353) of Laks=S and 55.7% (1898/3407) of Laks=LQ cells were called as PERT=G1/2 (Fig. 4e). PERT detected ∼4x fewer LQ cells than Laks (3407 to 829) with many of the recovered cells having erroneously high ploidy and thus higher breakpoint counts according to the input CN states (Fig. 4f,g). Reducing the number of LQ cells enables PERT to increase the yield of scWGS libraries, such as SA1292, which had many G1/2-phase cells excluded from previous analysis (Figs. S9,10, Additional Files 1,2). This evidence suggests that PERT's direct modeling of how DNA replication, somatic CN, and library-level biases combine to produce observed read count provides high-fidelity cell cycle prediction across diverse data.

Linking copy number alterations and replication timing

We further analyzed the Funnell *et al.* hTERT samples [36] to study the link between CNAs and RT as they accrued CNAs during serial passaging due to successive deletion of TP53, BRCA1, and BRCA2 (Fig. 5a). For each genotype, DLP+ sequencing of ancestral and intermediate populations enabled us to interrogate whether RT influences CNA acquisition and whether CNAs cause shifts in RT.

We first sought to find how much samples and clones diverged from one another in their CN vs RT distances. hTERT samples and clones had highly correlated RT profiles despite their divergence in CN distance while OV2295 had the lowest pairwise RT correlation for all hTERTs (Fig. S11a-d). One of the two TP53-/- BRCA2-/- samples (SA1055) had clonal whole genome doubling (WGD) but its RT remained similar to the diploid TP53-/- BRCA2-/- sample (SA1056) and other diploid hTERTs (Fig. S11a,b). OV2295, SA906a, and SA1054 all had subclonal WGD events; however, the WGD clones had higher RT correlation with diploid clones from the same sample than WGD clones from different samples (Fig. S11c,d). These findings support previous reports showing RT is heritable [16, 39] and provide novel evidence additional copies gained through WGD maintain the same RT as their diploid ancestors.

We next examined if CNAs preferentially emerged in early or late RT loci. We computed a reference RT profile from SA039 clone A since this clone represents the ancestral hTERT WT population with no CNAs (Fig. S11e-g). Counting sample gain, loss and unaltered bins in pseudobulk CN profiles, we found that gains preferentially emerged from early RT loci and losses from late RT loci ($p_{adj} < 10^{-4}$) (Fig. 5b-d). We did not find an association between CNA breakpoint location and ancestral RT (Fig. S11h). These findings suggest that the RT of a genomic locus influences its gain and loss rates [37, 40].

We last sought to determine whether RT is altered by the presence of CNAs. We tested this by examining subclonal CNAs, labeling each bin as having a subclonal gain, loss, or being unaltered compared to other clones of the same sample (Fig. 5e). Using unaltered loci as a control distribution for the expected RT deviation between any two clones, we found a shift to earlier RT in bins with losses and a shift to later RT in bins with gains (Fig. 5f,g, S11i). This finding that CNAs produce genomes with unique RT programs was also supported by an increase in cell-to-cell scRT variability statistic, T-width [26], as a function of the number of clones

in a sample (Fig. S11j). These results provide evidence that gains and losses can perturb the RT much like copy-neutral structural variation [19, 20] as they alter the structural organization of the genome; however, the effect size of CNA-induced RT shifts are smaller than RT shifts between different cell types. Thus, cell-of-origin likely dictates the position of active replication origins along the genome whereas CNAs can alter the relative time within S-phase at which these origins fire.

Chromosome X replication timing shifts measure the ratio of active to inactive alleles

Given that X-inactivation produces an allele that replicates much later than the active allele [27, 39], we hypothesized that the greatest RT shifts would occur from CNAs which disrupted the 1:1 balance of chrX active (Xa) to inactive (Xi) alleles. To study the relationship between RT and X-inactivation we ran PERT and SIGNALS, a single-cell allele-specific copy number caller, on the previously described hTERT and OV2295 cell lines [36]. Across all samples, 5 were diploid and allelically balanced on chrX and 4 had chrX events (Additional File 1). We used the 5 diploid balanced samples to compute a reference pseudobulk RT profile and computed RT shifts between samples with chrX events and this reference. Using RT shifts on autosomes as a background distribution, we found that all samples with chrX events had significantly earlier (3/4) or later (1/4) chrX RT ($p_{adj} < 10^{-4}$, Fig. 5h). OV2295 and SA1056 had fully clonal chrX loss of heterozygosity (LOH) and earlier chrX RT, indicating that no Xi alleles were present in these samples (Additional File 1). OV2295's clonal chrX LOH and subclonal WGD suggests an evolutionary history in which Xi was lost prior to WGD, producing XaXa genotypes in the WGD clones. In contrast to OV2295 and SA1056, SA1054 and SA1055 had multiple subclonal chrX gains on opposite parental alleles. For SA1055, clone A had a gain of the chrX SIGNALS A allele and significantly earlier RT whereas clones D and E had a gain of the chrX SIGNALS B allele and significantly later RT (Fig. 5i,j). These results show us that the SA1055 SIGNALS A and B alleles were Xa and Xi, respectively. We repeated our analysis for SA1054 and found that the SIGNALS B allele was Xi in this sample too (Figs. 12,13). These results demonstrate that PERT can be used in tandem with allele specific copy number information to identify which chrX parental allele is epigenetically inactive using only scWGS data as input.

Clone cell cycle distributions reflect proliferation rate and cisplatin sensitivity

PERT's ability to predict cell cycle phase and infer somatic CN in S-phase cells enables novel interrogation into the cell cycle distribution of individual clones. Proliferation rates are integral to understanding intratumor clonal evolution because highly proliferative cells are known to have a relative fitness advantage in the treatment-naive setting and greater sensitivity to platinum-based chemotherapies [24]. We confirmed that high PERT G1/2-phase fractions correlate with low proliferation and high scRNA G1-phase fraction by running PERT on three gastric cancer cell lines with co-registered doubling times, 10x scWGS, and 10x scRNA measurements [8] (Fig. S14a, Additional File 2). We then analyzed time-series DLP+ data generated from triple negative breast cancer (TNBC) patient derived xenografts (PDXs) with and without cisplatin treatment to investigate whether PERT can predict proliferative fitness of tumor clones [41] (Fig. 6a). Previous analysis of this

data revealed an inversion of the clonal fitness landscape upon cisplatin exposure but did not identify any genotypic or phenotypic features to explain why such an inversion occurs. We ran PERT on all TNBC PDX libraries and used the fraction of G1/2-phase vs S-phase cells assigned to each clone at each timepoint to determine when a given clone is enriched or depleted for S-phase cells (Fig. 6b-e, Fig. S15i-iv, Methods). Each clone's continuous S-phase enrichment (SPE) score was compared to its subsequent expansion rate within G1/2-phase, revealing that SPE correlated with expansion in untreated samples and contraction in treated samples (Fig. 6f). This result is in line with reports that cisplatin's genotoxic insults preferentially kill S-phase cells [24]. We also observed a positive correlation between clone SPE and expansion in untreated time-series hTERT samples (Fig. S14b-f). We find a relationship between clone SPE and expansion despite confounders such as serial passaging between DLP+ timepoints and destructive sequencing which introduce artificial evolutionary bottlenecks that distort the relationship between true proliferation rate and observed clone expansion rate. We found that RT profiles were unique between PDXs (Pearson r 0.82-0.87) and highly conserved between on- and off-treatment groups of the same PDX (Pearson r 0.94-0.99) (Fig. S16a), suggesting that cisplatin does not disrupt genome-wide RT coordination. Additionally, we found on-treatment SA1035 S-phase cells to be earlier in S-phase than off-treatment SA1035 S-phase cells ($p_{adj} < 10^{-4}$); however, the on- vs off-treatment distributions had no significant difference for the other two PDXs ($p_{adj} > 0.05$) (Fig. S16b). Finally, on-treatment samples had >2x higher S-phase fraction (SPF) than off-treatment samples in SA1035 and SA535 (Fig. S16c,d). We hypothesize that the relationship between SPF and cisplatin is a consequence of DNA platinum adducts stalling replication fork progression [42], producing a longer S-phase and thus higher SPF. In total, these results show that PERT is capable of measuring the relative proliferation rate between competing subclones of a tumor – an important feature for predicting their on- and off-treatment evolutionary fitness.

Discussion

PERT introduces a probabilistic framework for jointly inferring somatic copy number and replication states from single-cell whole genome sequencing data of genomically unstable samples. PERT is superior to previous scRT methods which assume all cells have identical somatic copy number profiles and attribute all read depth fluctuations to replication [26, 27, 28, 29]. Additionally, using PERT's inferred fraction of replicated loci per cell provides robust cell cycle phase prediction which outperforms pre-trained classifiers [7] that perform poorly when GC bias of a given library does not match those of the training libraries. PERT can be run on any high-throughput scWGS data modality containing S-phase cells (DLP+, 10x, etc) at multiple bin sizes (500kb, 20kb, etc) since it models the data at the read depth level. Using PERT to identify S-phase cells in unsorted scWGS libraries is superior to FACS approaches used in acoustic cell tagmentation libraries [11, 10] as PERT enables sequencing of all cells in a given tumor, not just those within a certain DAPI range. PERT's recovery of many G1/2-phase cells with erroneous initial CN states allows for more accurate estimation of clone sizes and increases the number of cells that can be used for clone-pseudobulk DNA sequence analysis (such as SNV calling or detection of rearrangement junctions).

PERT's methodological improvements allowed for quantification of clonal and subclonal RT shifts and estimation of the relative proliferation rate of tumor subclones. Our finding that losses preferentially emerged from late RT loci and gains from early RT loci suggests that certain gains and losses might result from mechanisms of over- and under-replication or reflect differential fitness of gains vs losses in gene-rich early vs gene-poor late RT loci [37, 38, 40]. Similarly, our finding that bins with subclonal losses shift to earlier RT and bins with gains shift to later RT could be explained by transcriptional buffering of large CNAs as early RT is linked with high transcription rate [12, 43]. However, it is unclear whether this effect is due to selection – where losses in low transcription late RT alleles are tolerated better than losses in high transcription early RT alleles – or plasticity – where the remaining allele undergoes epigenetic modulation after the CNA occurs. We also found that clonal and subclonal chrX allelic imbalances resulted in significant RT shifts, allowing us to phase X-inactivation to each sample's A or B allele. Finally, we leveraged cell lines of various doubling times and time-series sampling of TNBC PDXs to demonstrate that PERT S-phase fraction measures proliferation rates, with S-phase enriched clones expanding in the untreated context and contracting in the cisplatin-treated context.

The continued improvement and generation of scWGS data across various cancer types will enable future studies into the relationship between DNA replication and genomic instability. As sequencing technologies improve, it will be possible to use PERT to measure replication dynamics at sub-kilobase scales, allowing researchers to quantify genome-wide statistics such as origin firing efficiency, fork speed, and fork collapse rate using scWGS data alone. These advances will be critical in attempts to define various replication stress etiologies and, ultimately, match patient tumors with appropriate targeted therapies [2, 44]. We believe that PERT's accurate prediction of subclonal cell cycle phase distributions will be helpful for identifying senescent or hyperproliferative clones [45, 46, 47]; however, we caution that batch effects such as storage technique and reagent concentrations could alter the number of S-phase cells sequenced, complicating future inter-sample analyses. Nevertheless, we believe that using cell cycle distributions to prospectively predict which clones will respond to chemotherapy can provide complementary information to other genomic features such as gain of oncogenes, loss of tumor suppressors, and WGD which can have variable phenotypic impacts depending on unmeasurable factors of the tumor microenvironment [48, 49, 50].

In conclusion, we anticipate that studying DNA replication programs in genomically unstable cancer should provide insights into each tumor subclone's etiology, evolutionary fitness, and drug sensitivities. PERT offers a novel probabilistic framework for detecting S-phase cells and explaining their observed read depth as a combination of copy number and replication signals. Our approach provides a new tool to disentangle the relative contribution of fixed genomic alterations and other dynamic processes on DNA replication programs in cancer.

Conclusions

PERT allows for single-cell DNA replication dynamics to be measured in genomically unstable cancer cells with subclonal and cell-specific copy number aberrations,

leading to better understanding on how DNA replication influences and is modulated throughout tumor evolution.

Methods

PERT model

The input for PERT is binned read depth (Z) and CN states for all scWGS cells. Input CN states are obtained through single-cell CN callers such as HMMcopy [6, 31] or 10x CellRanger-DNA [8, 9]. PERT first identifies a set of high-confidence G1/2-phase cells where the input CN states reflect accurate somatic CN. All remaining cells have their input CN states dropped as they are initially considered to have unknown CN states and cell cycle phase. Most S-phase cells should be present in the unknown initial set. High-confidence G1/2-phase cells are phylogenetically clustered into clones based on CN using methods such as sitka [51] or MEDICC2 [52]. Optionally, users can provide their own sets of clustered high-confidence G1/2-phase and unknown cells. These sets of cells are passed into a probabilistic model which infers somatic CN (X) and replication states (Y) through three distinct learning steps. In Step 1, PERT learns parameters associated with library-level GC bias (β_μ, β_σ) and sequencing overdispersion (λ) by training on high-confidence G1/2 cells (Fig. S1a). Step 1 conditions on CN (X), replication (Y), and coverage/ploidy scaling terms (μ) because input CN states are assumed to accurately reflect somatic CN states and all bins are unreplicated ($Y = 0$) in high-confidence G1/2 cells. Once β_μ, β_σ and μ have been learned in Step 1, we can condition on them in Step 2 (Fig. S1b). Step 2 learns latent parameters representing each cell's time in S-phase (τ_n), each locus's replication timing (ρ_m), and global replication stochasticity (α) to compute the probability that a given bin is replicated ($Y_{n,m} = 1$) or unreplicated ($Y_{n,m} = 0$). Only unknown cells are included in Step 2. Prior belief on each unknown cell's CN state is encapsulated using a prior distribution (π) which has concentration parameters (η) conditioned on the input CN of the most similar high-confidence G1/2 cells. A full list of model parameters, domains, and distributions can be found in Table 1. Step 3 is an optional final step which learns CN and replication states for high-confidence G1/2 cells (Fig. S1c). This step is necessary to determine if any S-phase cells are present in the initial set of high-confidence G1/2 cells. Step 3 conditions upon replication timing (ρ) and stochasticity (α) values learned in Step 2 to ensure that such properties are conserved between both sets of cells.

PERT is designed for scWGS data with coverage depths on the order of 0.01-0.1x and thus 500kb bin sizes are used by default in this manuscript; however, the model can be run on count data of any bin size as long as sufficient memory and runtime are allocated.

Equations for Step 1

Given that we have accurate CN caller results for high-confidence G1/2 cells, we can solve for each cell's coverage/ploidy scaling term μ_n and condition upon it

$$\mu_n = \frac{\sum_{m=0}^{M} Z_{n,m}}{\sum_{m=0}^{M} X_{n,m}}. \tag{1}$$

The latent variables are arranged together in function block f through the following equations to produce the bin-specific negative binomial event counts $\delta_{n,m}$. The GC bias rate of each individual bin ($\omega_{n,m}$) depends on the GC content of the locus (γ_m) and the GC bias coefficients ($\beta_{n,k}$) for the cell

$$\omega_{n,m} = e^{\sum_{k=0}^{K} \beta_{n,k} * \gamma_m^k}. \tag{2}$$

The expected read count per bin is computed as follows:

$$\theta_{n,m} = X_{n,m} * \omega_{n,m} * \mu_n. \tag{3}$$

The expected read count per bin is then used in conjunction with the negative binomial event success probability term (λ) to produce a number of negative binomial event count for each bin

$$\delta_{n,m} = f(X_{n,m}, \gamma_m, \lambda, \mu_n, \beta_{n,k}) = \frac{\theta_{n,m} * (1-\lambda)}{\lambda}. \tag{4}$$

Where we place the constraint $\delta_{n,m} \geq 1$ to avoid sampling errors in bins with $\theta_{n,m} \approx 0$. Finally, the read count at a bin is sampled from an overdispersed negative binomial distribution $Z_{n,m} \sim \mathrm{NB}(\delta_{n,m}, \lambda)$ where the expected read count for $Z_{n,m}$ is $\theta_{n,m}$ and the variance is $\frac{\theta_{n,m}}{(1-\lambda)}$.

Equations for Steps 2-3

Steps 2-3 have equations which differ from Step 1 since it must account for replicated bins and cannot solve for μ_n analytically. The probability of each bin being replicated ($\phi_{n,m}$) is a function of the cell's time in S-phase (τ_n), the locus's replication timing (ρ_m), and the replication stochasticity term (α). Replication stochasticity (α) controls how closely cells follow the global RT profile by adjusting the steepness of a sigmoid curve. The following equation corresponds to function block g:

$$\phi_{n,m} = g(\alpha, \tau_n, \rho_m) = \frac{1}{1 + e^{-\alpha(\tau_n - \rho_m)}}. \tag{5}$$

Equations corresponding to function block f differ from those in Step 1. The total CN ($\chi_{n,m}$) is double the somatic CN ($X_{n,m}$) when a bin is replicated ($Y_{n,m} = 1$)

$$\chi_{n,m} = X_{n,m} * (1 + Y_{n,m}). \tag{6}$$

The GC rates ($\omega_{n,m}$) and negative binomial event counts ($\delta_{n,m}$) are computed the same as in Step 1 (Eq 2, Eq 4). However, the expected read count uses total instead of somatic CN

$$\theta_{n,m} = \chi_{n,m} * \omega_{n,m} * \mu_n. \tag{7}$$

Since CN is learned in Steps 2-3, the coverage/ploidy scaling term (μ_n) must also be learned. We use a normal prior $\mu_n \sim \mathrm{N}(\mu_\mu^n, \mu_\sigma^n)$ where the approximate total ploidy and total read counts are used to estimate the mean hyperparameters (μ_μ^n). Total ploidies for each cell are approximated using the CN prior concentrations (η) and times within S-phase (τ) to account for both somatic and replicated copies of DNA that are present. We fixed the standard deviation hyperparameters (μ_σ^n) to always be 10x smaller than the means to ensure that $\mu_n \geq 0$ despite use of a normal distribution (used for computational expediency).

$$\mu_\mu^n = \frac{\sum_{m=0}^{M} Z_{n,m}}{(1+\tau_n)\sum_{m=0}^{M} \mathrm{argmax}_p(\eta_{n,m,p})} \tag{8}$$

$$\mu_\sigma^n = \frac{\mu_\mu^n}{10} \tag{9}$$

Constructing the CN prior concentrations

There are two ways to construct the CN prior concentrations within PERT. The first is to use the most similar high-confidence G1/2 clone to define the concentrations for each unknown cell (clone method). We assign each unknown cell its clone (c_n) via Pearson correlation between the cell read depth profile (Z_n) and the clone pseudobulk read depth profile (Z_c)

$$c_n = \mathrm{argmax}_c(\mathrm{corr}(Z_n, Z_c)). \tag{10}$$

Clone pseudobulk CN and read depth profiles represent the median profile across all high-confidence G1/2 cells in a given clone c. Once we have clone assignments for each unknown cell, the CN concentration of all possible states P at each genomic bin ($\eta_{n,m,p}$) is constructed to be w times larger for the state p that matches the clone pseudobulk CN state ($X_{c_n,m}$) for that same bin compared to all other states. The default setting is $w = 10^6$.

$$\eta_{n,m,p} = \begin{cases} w & \text{if } p = X_{c_n,m} \\ 1 & \text{else} \end{cases} \tag{11}$$

The second way to construct the prior is to leverage additional information from the most similar high-confidence G1/2 cells when constructing $\eta_{n,m,p}$ (composite method). The rationale for the composite method is that there might be rare CNAs within a clone which only appear in a handful of cells but do not appear in the clone pseudobulk CN profile X_c. To find the most similar high-confidence G1/2

cells, we compute the read depth correlation between the unknown cell (Z_{n_s}) and the high-confidence G1/2 cells from the best matching clone (Z_{n_g})

$$\psi = \text{corr}(Z_{n_s}, Z_{n_g}). \tag{12}$$

The consensus clone CN profile and top J matches for each unknown cell are then used to construct the CN prior ($\eta_{n,m,p}$). Each row of ψ is sorted to obtain the top J high-confidence G1/2 matches $n_{g(0)}, ..., n_{g(J-1)}$. All entries are initialized to 1 ($\eta_{n,m,k} = 1$) before adding varying levels of weight (w) to states where the CN matches a G1/2-phase cell or clone pseudobulk CN profile. The default settings are $w = 10^5$ and $J = 5$.

$$\eta_{n,m,p} = \begin{cases} +1 & \text{everywhere} \\ +w * 2 * J & \text{if } p = X_{c_n,m} \\ +w * (J-0) & \text{if } p = X_{n_{g0},m} \\ +w * (J-1) & \text{if } p = X_{n_{g1},m} \\ ... & \\ +w & \text{if } p = X_{n_{g(J-1)},m} \end{cases} \tag{13}$$

By default, the composite method is used during Step 2 and the clone method is used during Step 3; however, the user may select between both methods during Step 2. Using the clone method during Step 2 should be seen as a 'vanilla' version of PERT which should be used when very few cell-specific CNAs are present. The clone method is used for Step 3 since the composite method would produce many self-matching cells. A comparison of the two methods can be seen when benchmarking PERT on simulated data.

Model initialization and hyperparameters

Splitting cells into initial sets of high-confidence G1/2-phase and unknown cells is performed by thresholding heuristic per-cell features known to correlate with cell cycle phase. PERT uses clone-normalized number of input CN breakpoints between neighboring genomic bins (BKnorm) and clone-normalized median absolute deviation in read depth between neighboring genomic bins (MADNnorm). Note that breakpoints between chromosome boundaries are not counted.

$$\text{BK}_n = \sum_{m=0}^{M-1} \begin{cases} 1 & \text{if } X_{n,m} \neq X_{n,m+1} \\ 0 & \text{else} \end{cases} \tag{14}$$

$$\text{BKnorm}_n = \text{BK}_n - \frac{1}{C} \sum_{c=0}^{C} \text{BK}_c \tag{15}$$

$$\text{MADN}_n = \text{Med}(\sum_{m=0}^{M-1} Z_{n,m} - Z_{n,m+1}) \tag{16}$$

$$\text{MADNnorm}_n = \text{MADN}_n - \frac{1}{C}\sum_{c=0}^{C} \text{MADN}_c \tag{17}$$

Under default settings, PERT initializes cells with MADNnorm<0 and BKnorm<0 as high-confidence G1/2-phase with all other cells as unknown phase. Initial cell phases can also be input by users based on experimental measurements or alternative metrics such as 10x CellRanger-DNA's 'dimapd' score (used in [8, 29, 28]), the Laks *et al* classifiers' S-phase probability and quality scores [7], or read depth correlation with a reference RT profile [32].

To discourage the model from converging at local minima instead of global minima, each cell's time in S-phase (τ_n) is initialized using scRT results from a clone-aware adaptation of Dileep *et al* [26] which thresholds the clone-normalized read depth profiles into replicated and unreplicated bins. Each unknown cell n is assigned to clone c with the highest correlation between cell and clone pseudobulk read depth profiles (Eq 10). The read depth of each cell is then normalized by the CN state with highest probability within the CN prior ($\eta_{n,m,p}$)

$$y_{n,m} = \frac{Z_{n,m}}{\text{argmax}_p(\eta_{n,m,p})}. \tag{18}$$

The clone-normalized read depth profiles (y_n) are then binarized into replication state profiles (Y_n) using a per-cell threshold ($t_n \in [0, 1]$) that minimizes the Manhattan distance between the real data and its binarized counterpart.

$$t_n = \text{argmin}_t \left| y_{n,m} - \begin{cases} 1 & \text{if } y_{n,m} \geq t_n \\ 0 & \text{else} \end{cases} \right| \tag{19}$$

$$Y_{n,m} = \begin{cases} 1 & \text{if } y_{n,m} \geq t_n \\ 0 & \text{else} \end{cases} \tag{20}$$

The fraction of replicated bins per cell from the deterministic replication states $Y_{n,m}$ are then used to initialize the parameter representing each cell's time in S-phase (τ_n) within PERT's probabilistic model.

$$\tau_n = \frac{1}{M}\sum_{m=0}^{M} Y_{n,m}. \tag{21}$$

Initialization of τ_n is particularly important because the model might think an early S-phase cell (<20% replicated) is a late S-phase cell (>80% replicated), or vice versa, as both have relatively 'flat' read depth profiles compared to mid-S-phase cells. Thus τ_n will rarely traverse mid-S-phase values during inference when its initial and true values lie far apart. Additional parameter initializations include $\lambda = 0.5$ for negative binomial overdispersion and $\beta_{\sigma,k} = 10^{-k}$ for the standard deviation of each GC bias polynomial coefficient k. Unlike τ_n, the model is unlikely to get stuck at local minima with these parameters so they are initialized to the same values globally.

The latent variables β_μ, ρ, and α are sampled from prior distributions with fixed hyperparameters. The mean of all GC bias polynomial coefficients (β_μ) are drawn from the prior $\mathrm{N}(0, 1)$. Each locus's replication timing (ρ) is drawn from the prior $\mathrm{Beta}(1, 1)$ to create a uniform distribution on the domain $[0, 1]$. The replication stochasticity parameter (α) is drawn from the prior distribution $\Gamma(\mathrm{shape} = 2, \mathrm{rate} = 0.2)$ which has a mean of $\frac{\mathrm{shape}}{\mathrm{rate}} = 10$ and penalizes extreme values on a positive real domain.

PERT phase predictions

We used the PERT model output to predict 'G1/2', 'S', and 'low quality' (LQ) phases for each cell. G1/2-phase cells were defined by having <5% or >95% replicated bins. Of the remaining cells with 5-95% replicated bins, those with high read depth autocorrelation (>0.5), replication state autocorrelation (>0.2), or fraction of homozygous deletions ($X = 0$, >0.05) were deemed to be low quality. All other cells were deemed to be in S-phase. Using 500kb bins, autocorrelation scores were the average of all autocorrelations ranging from 10 to 50 bin lag size. Thresholds used for splitting S and LQ phases can be adjusted by users should the default settings produce unexpected output.

Model construction and inference

PERT is written using Pyro which is a probabilistic programming language written in Python and supported by PyTorch backend [34]. PERT uses Pyro's implementation of Black Box Variational Inference which enables use biologically-informed priors instead of being limited to conjugate priors [53].

Simulated datasets

To benchmark PERT's ability to accurately infer single-cell replication and somatic CN states, we simulated datasets with varying clonal structures and cell-specific CNA rates. Somatic CN states are simulated by first drawing clone CN profiles and then drawing cell-specific CNAs that deviated from said clone CN profile. All CNAs are drawn at the chromosome-arm level. 400 S- and 400 G1/2-phase cells are simulated in each dataset.

Once CN states have been simulated, we simulate the read depth using PERT as a generative model. We condition the model on the provided β_μ, β_σ, λ, α, ρ, γ, and X parameters when generating cell read depth profiles. All read depth values (Z) are in units of reads per million. RepliSeq data for various ENCODE cell lines are used to set ρ values for each clone [35]. We did not need to condition the model

on τ since the Beta$(1, 1)$ prior hyperparameters forced values to be sampled from a uniform distribution ranging from 0 to 1. A table of all the parameters used in each simulated dataset can be found in Table 2.

We called CN on simulated binned read count data using HMMcopy. Given that Kronos was designed as an end-to-end pipeline that takes in raw BAM files, we forked off the Kronos repository and edited their 'Kronos RT' module to accept binned read count and CN states as input. Cells were split into S- and G1/2-phase Kronos input populations according to their true phase. Code to our forked repository can be found at `https://github.com/adamcweiner/Kronos_scRT`. Similarly, we removed features from the Laks *et al* cell cycle classifier that used alignment information such as the percentage of overalapping reads per cell. The Laks classifier was retrained with said features removed prior to deployment on simulated data (Fig. S2a,b).

Experimental methods

Detailed descriptions of the data generation methods are described in Laks *et al*, Funnel *et al*, and Salehi *et al* [7, 36, 41]. Such descriptions include generation of the cell cycle FACS datasets, generation of engineered hTERT cell lines, xenografting, time series passaging, and DLP+ sequencing.

DLP+ data processing

All DLP+ data was passed through `https://github.com/shahcompbio/single_cell_pipeline` before downstream analysis. This pipeline aligned reads to the hg19 reference genome using BWA-MEM. Each cell was then passed through HMMcopy using default arguments for single-cell sequencing. HMMcopy's output provided read count and gc-corrected integer CN states for each 500kb genomic bin across all cells and loci. Loci with low mappability (<0.95) and cells with low read count (<500,000 reads) were removed. Cells were also filtered for contamination using the FastQ Screen which tags reads as matching human, mouse, or salmon reference genomes. If >5% of reads in a cell are tagged as non-human the cell is flagged as contaminated and subsequently removed.

Cells were only passed into phylogenetic trees if they were called as G1/2-phase and high quality by classifiers described in Laks *et al* [7]. In certain cases, cells might be manually excluded from the phylogenetic tree if they pass the cell cycle and quality filters but have an abnormally high number of HMMcopy breakpoints. All cells included in the phylogenetic tree are passed into PERT as the set of high-confidence G1/2 cells; all cells outside the tree are passed as unknown cells.

Phylogenetic clustering based on CN profiles

We used the clone IDs from Funnel *et al* for high-confidence G1/2 cells [36]. These single-cell phylogenetic trees were generated using sitka [51]. Sitka uses CN breakpoints (also referred to as changepoints) across the genome as binary input characters to construct the evolutionary relationships between cells. Sitka was run for 3,000 chains and a consensus tree was computed for downstream analysis. The consensus tree was then cut at an optimized height to assign all cells into clones (clusters). For datasets with no sitka trees provided or select datasets, cells were clustered into

clones using K-means where the number of clones was selected through Akaike information criterion. We performed a K-means reclustering for the Salehi *et al* TNBC PDX data [41] as sitka produced small clusters which inhibited robust tracking of S-phase clone fractions across multiple timepoints.

Pseudobulk profiles

Many times in the text we describe "pseudobulk" replication timing, copy number, or read depth profiles within a subset of interest (i.e. cells belonging to the same clone or sample). To compute pseudobulk profiles, we group all the cells of interest and then take the median values for all loci in the genome. When computing pseudobulk CN profiles, we only include the cells of the modal (most common) ploidy state before computing median values for all loci.

Comparison of RT profiles to Hi-C A/B compartments

Hi-C compartment data were downloaded from ENCODE for T47D and B-lymphoblast (GM- prefix) cell lines using the accession codes ENCFF713FCA, ENCFF220LEI, ENCFF733ZUJ, ENCFF907MWF, ENCFF522SPQ, and ENCFF4-11JKH [35]. Genomic coordinates were lifted to human reference hg19 for comparison. Due to varying quality and sequencing platforms of each Hi-C library, we used Spearman instead of Pearson correlation.

Gastric cancer cell line data

10x Chromium single-cell DNA (10x scWGS) data of gastric cancer cell lines NCI-N87, HGC-27, and SNU-668 were downloaded from SRA (PRJNA498809). Copy number calling was performed using the CellRanger-DNA pipeline using default parameters. Data was aggregated from 20kb to 500kb bins for analysis with PERT. Each cell line's doubling time and fraction of G1-phase scRNA cells were extracted from Andor *et al* [8].

Clone S-phase enrichment scores

To test whether a clone (c) is significantly enriched or depleted for S-phase cells at a given timepoint (t), we must compare that clone's fraction in both S- and G1/2-phases. We first define the following variables as such:

$N_{s,c,t}$: Number of S-phase cells belonging to clone c at time t

$N_{g,c,t}$: Number of G1/2-phase cells belonging to clone c at time t

$N_{s,t}$: Total number of S-phase cells across all clones at time t

$N_{g,t}$: Total number of G1/2-phase cells across all clones at time t

N_t: Total number of cells in a population at time t (all clones, all phases)

We can then define the fractions of S- and G1/2-phase cells assigned to clone c at time t ($f_{s,c,t}$, $f_{g,c,t}$):

$$f_{s,c,t} = \frac{N_{s,c,t}}{N_{s,t}} \tag{22}$$

$$f_{g,c,t} = \frac{N_{g,c,t}}{N_{g,t}}. \tag{23}$$

Each clone's continuous S-phase enrichment score ($\xi_{c,t}$) is the difference between the S- and G1/2-phase fractions. Positive values indicate the clone is enriched for S-phase cells.

$$\xi_{c,t} = f_{s,c,t} - f_{g,c,t}. \tag{24}$$

Using the fraction of G1/2-cells belonging to clone c, we can compute the expected total number of cells in clone c and time t across all cell cycle phases

$$E(N_{c,t}) = f_{g,c,t} * N_t. \tag{25}$$

We produce a p-value for enrichment of S-phase cells using a hypergeometric test scipy.stats.hypergeom(M=N_t, n=$N_{s,t}$, N=$E(N_{c,t})$).sf($N_{s,c,t}$). To produce a p-value for S-phase depletion we subtract this enrichment p-value from 1. All p-values are Bonferroni-corrected by dividing by the total number of statistical tests. p-adjusted thresholds of 1e-2 are used for saying a clone is significantly enriched or depleted for S-phase cells within a given library.

Clone expansion scores

For time-series DLP+ experiments, we computed clone expansion scores for each clone c at time t ($S_{c,t}$) by examining the fraction of G1/2-phase cells belonging to clone c at timepoint t ($f_{g,c,t}$) and the subsequent timepoint ($f_{g,c,t+1}$). Positive values indicate the clone expands by the next timepoint.

$$S_{c,t} = f_{g,c,t+1} - f_{g,c,t} \tag{26}$$

Statistical tests

When boxplots are presented in the figures, hinges represent the 25% and 75% quantiles and whiskers represent the ±1.5x interquartile range. Statistical significance is tested using independent t-tests from scipy.stats unless otherwise noted. Bonferroni correction is implemented for all statistical tests to limit false discovery. The number of stars is a shorthand for the adjusted p-value of a given statistical test ([$< 10^{-4}$, ****], [$< 10^{-3}$, ***], [$< 10^{-2}$, **], [< 0.05, *], [≥ 0.05, ns]). Shaded areas surrounding linear regression lines of best fit represent 95% confidence intervals obtained via boostrapping (n=1000 boostrap resamples).

Quantifying cell-to-cell scRT variability

To quantify cell-to-cell scRT variability, we computed T_{width} as defined in Dileep *et al* and Kronos-scRT [26, 29]. T_{width} is the time needed for genomic regions to be replicated in 25% to 75% of cells in a S-phase that is assumed to last 10 hours. Large T_{width} values indicate that the ordering of DNA replication is highly variable between cells whereas small T_{width} values indicated that it is deterministic.

Funding

This project was funded in part by Cycle for Survival supporting Memorial Sloan Kettering Cancer Center. S.P.S. holds the Nicholls Biondi Chair in Computational Oncology and is a Susan G. Komen Scholar. A.C.W. is supported by National Cancer Instititue (NCI) Ruth L. Kirschstein National Research Service Award for Predoctoral Fellows F31-CA271673. M.J.W. is supported by NCI Pathway to Independence award K99-CA256508.

Availability of data and materials

Data availability

Data from Funnel *et al* [36] can be found on zenodo https://zenodo.org/record/6998936#.Y0h3luzMLzc. Raw DLP+ sequencing data from Laks *et al* and Salehi *et al* [7, 41] are available from the European Genome-Phenome under study IDs EGAS00001004448 and EGAS00001003190, respectively. All other data will be made available for controlled access upon final publication.

Code availability

The following code repositories are publicly available and contain tutorials for installation and use.

- Package containing PERT model and other tools for scRT analysis: https://github.com/shahcompbio/scdna_replication_tools
- DLP+ single-cell whole genome sequencing pipeline: https://github.com/shahcompbio/single_cell_pipeline

The following repositories will be made available upon final publication.

- Analysis scripts and figure generation: https://github.com/shahcompbio/scdna_replication_paper
- LaTeX files and figures for manuscript generation: https://github.com/adamcweiner/pert_manuscript

Ethics approval and consent to participate

Not applicable

Competing interests

S.P.S. is a shareholder of Imagia Canexia Health Inc. and is a consultant to AstraZeneca Inc., outside the scope of this work.

Authors' contributions

A.C.W., S.P.S., and A.M. designed the methodology and wrote the manuscript. All authors helped to design experiments and/or analyze the data.

Authors' information

S.P.S. and A.M. jointly supervised this work.

Author details

[1]Computational Oncolocy, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA. [2]Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medicine, New York, USA. [3]Gerstner Sloan Kettering Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer Center, New York, USA.

References

1. Guilbaud, G., Rappailles, A., Baker, A., Chen, C.L., Arneodo, A., Goldar, A., d'Aubenton-Carafa, Y., Thermes, C., Audit, B., Hyrien, O.: Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. PLoS Comput Biol **7**(12), 1002322 (2011). doi:10.1371/journal.pcbi.1002322
2. Gaillard, H., García-Muse, T., Aguilera, A.: Replication stress and cancer. Nature Reviews Cancer **15**(5), 276–289 (2015). doi:10.1038/nrc3916
3. Rivera-Mulia, J.C., Dimond, A., Vera, D., Trevilla-Garcia, C., Sasaki, T., Zimmerman, J., Dupont, C., Gribnau, J., Fraser, P., Gilbert, D.M.: Allele-specific control of replication timing and genome organization during development. Genome Res **28**(6), 800–811 (2018). doi:10.1101/gr.232561.117
4. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.L., Brenton, J.D., Tavaré, S., Caldas, C., Aparicio, S.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature **486**(7403), 346–52 (2012). doi:10.1038/nature10983
5. Burrell, R.A., McClelland, S.E., Endesfelder, D., Groth, P., Weller, M.-C., Shaikh, N., Domingo, E., Kanu, N., Dewhurst, S.M., Gronroos, E., Chew, S.K., Rowan, A.J., Schenk, A., Sheffer, M., Howell, M., Kschischo, M., Behrens, A., Helleday, T., Bartek, J., Tomlinson, I.P., Swanton, C.: Replication stress links structural and numerical cancer chromosomal instability. Nature **494**(7438), 492–496 (2013). doi:10.1038/nature11935
6. Zahn, H., Steif, A., Laks, E., Eirew, P., VanInsberghe, M., Shah, S.P., Aparicio, S., Hansen, C.L.: Scalable whole-genome single-cell library preparation without preamplification. Nat Methods **14**(2), 167–173 (2017). doi:10.1038/nmeth.4140
7. Laks, E., McPherson, A., Zahn, H., Lai, D., Steif, A., Brimhall, J., Biele, J., Wang, B., Masud, T., Ting, J., Grewal, D., Nielsen, C., Leung, S., Bojilova, V., Smith, M., Golovko, O., Poon, S., Eirew, P., Kabeer, F., Ruiz de Algara, T., Lee, S.R., Taghiyar, M.J., Huebner, C., Ngo, J., Chan, T., Vatrt-Watts, S., Walters, P., Abrar, N., Chan, S., Wiens, M., Martin, L., Scott, R.W., Underhill, T.M., Chavez, E., Steidl, C., Da Costa, D., Ma, Y., Coope, R.J.N., Corbett, R., Pleasance, S., Moore, R., Mungall, A.J., Mar, C., Cafferty, F., Gelmon, K., Chia, S., Marra, M.A., Hansen, C., Shah, S.P., Aparicio, S.: Clonal decomposition and dna replication states defined by scaled single-cell genome sequencing. Cell **179**(5), 1207–122122 (2019). doi:10.1016/j.cell.2019.10.026
8. Andor, N., Lau, B.T., Catalanotti, C., Sathe, A., Kubit, M., Chen, J., Blaj, C., Cherry, A., Bangs, C.D., Grimes, S.M., Suarez, C.J., Ji, H.P.: Joint single cell dna-seq and rna-seq of gastric cancer cell lines reveals rules of in vitro evolution. NAR Genom Bioinform **2**(2), 016 (2020). doi:10.1093/nargab/lqaa016

9. Velazquez-Villarreal, E.I., Maheshwari, S., Sorenson, J., Fiddes, I.T., Kumar, V., Yin, Y., Webb, M.G., Catalanotti, C., Grigorova, M., Edwards, P.A., Carpten, J.D., Craig, D.W.: Single-cell sequencing of genomic dna resolves sub-clonal heterogeneity in a melanoma cell line. Communications Biology **3**(1), 318 (2020). doi:10.1038/s42003-020-1044-8
10. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W.R., Hicks, J., Wigler, M.: Tumour evolution inferred by single-cell sequencing. Nature **472**(7341), 90–4 (2011). doi:10.1038/nature09807
11. Minussi, D.C., Nicholson, M.D., Ye, H., Davis, A., Wang, K., Baker, T., Tarabichi, M., Sei, E., Du, H., Rabbani, M., Peng, C., Hu, M., Bai, S., Lin, Y.-w., Schalck, A., Multani, A., Ma, J., McDonald, T.O., Casasent, A., Barrera, A., Chen, H., Lim, B., Arun, B., Meric-Bernstam, F., Van Loo, P., Michor, F., Navin, N.E.: Breast tumours maintain a reservoir of subclonal diversity during expansion. Nature **592**(7853), 302–308 (2021). doi:10.1038/s41586-021-03357-x
12. Hulke, M.L., Massey, D.J., Koren, A.: Genomic methods for measuring dna replication dynamics. Chromosome Research **28**(1), 49–67 (2020). doi:10.1007/s10577-019-09624-y
13. Zhang, Q., Bassetti, F., Gherardi, M., Lagomarsino, M.C.: Cell-to-cell variability and robustness in s-phase duration from genome replication kinetics. Nucleic Acids Res **45**(14), 8190–8198 (2017). doi:10.1093/nar/gkx556
14. Vouzas, A.E., Gilbert, D.M.: Mammalian dna replication timing. Cold Spring Harbor Perspectives in Biology **13**(7) (2021). doi:10.1101/cshperspect.a040162
15. Rivera-Mulia, J.C., Buckley, Q., Sasaki, T., Zimmerman, J., Didier, R.A., Nazor, K., Loring, J.F., Lian, Z., Weissman, S., Robins, A.J., Schulz, T.C., Menendez, L., Kulik, M.J., Dalton, S., Gabr, H., Kahveci, T., Gilbert, D.M.: Dynamic changes in replication timing and gene expression during lineage specification of human pluripotent stem cells. Genome Research **25**(8), 1091–1103 (2015). doi:10.1101/gr.187989.114. http://genome.cshlp.org/content/25/8/1091.full.pdf+html
16. Klein, K.N., Zhao, P.A., Lyu, X., Sasaki, T., Bartlett, D.A., Singh, A.M., Tasan, I., Zhang, M., Watts, L.P., Hiraga, S.-i., Natsume, T., Zhou, X., Baslan, T., Leung, D., Kanemaki, M.T., Donaldson, A.D., Zhao, H., Dalton, S., Corces, V.G., Gilbert, D.M.: Replication timing maintains the global epigenetic state in human cells. Science **372**(6540), 371–378 (2021). doi:10.1126/science.aba5545
17. Tomkova, M., Tomek, J., Kriaucionis, S., Schuster-Böckler, B.: Mutational signature distribution varies with dna replication timing and strand asymmetry. Genome Biology **19**(1), 129 (2018). doi:10.1186/s13059-018-1509-y
18. Yaacov, A., Vardi, O., Blumenfeld, B., Greenberg, A., Massey, D.J., Koren, A., Adar, S., Simon, I., Rosenberg, S.: Cancer mutational processes vary in their association with replication timing and chromatin accessibility. Cancer Research **81**(24), 6106–6116 (2021). doi:10.1158/0008-5472.CAN-21-2039
19. Ryba, T., Battaglia, D., Chang, B.H., Shirley, J.W., Buckley, Q., Pope, B.D., Devidas, M., Druker, B.J., Gilbert, D.M.: Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. Genome Res **22**(10), 1833–44 (2012). doi:10.1101/gr.138511.112
20. Donley, N., Thayer, M.J.: Dna replication timing, genome stability and cancer: late and/or delayed dna replication timing is associated with increased genomic instability. Semin Cancer Biol **23**(2), 80–9 (2013). doi:10.1016/j.semcancer.2013.01.001
21. Ibrahim, A., Lashen, A., Toss, M., Mihai, R., Rakha, E.: Assessment of mitotic activity in breast cancer: revisited in the digital pathology era. Journal of Clinical Pathology **75**(6), 365 (2022). doi:10.1136/jclinpath-2021-207742
22. Bogunovic, D., Neill, D.W., Belitskaya-Levy, I., Vacic, V., Yu, Y.-L., Adams, S., Darvishian, F., Berman, R., Shapiro, R., Pavlick, A.C., Lonardi, S., Zavadil, J., Osman, I., Bhardwaj, N.: Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. Proceedings of the National Academy of Sciences **106**(48), 20429 (2009). doi:10.1073/pnas.0905139106
23. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., Louis, D.N., Rozenblatt-Rosen, O., Suvà, M.L., Regev, A., Bernstein, B.E.: Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. Science **344**(6190), 1396–1401 (2014). doi:10.1126/science.1254257
24. Donaldson, K.L., Goolsby, G.L., Wahl, A.F.: Cytotoxicity of the anticancer agents cisplatin and taxol during cell proliferation and the cell cycle. International Journal of Cancer **57**(6), 847–855 (1994). doi:10.1002/ijc.2910570614
25. Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., Stamatoyannopoulos, J.A.: Sequencing newly replicated dna reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A **107**(1), 139–44 (2010). doi:10.1073/pnas.0912402107
26. Dileep, V., Gilbert, D.M.: Single-cell replication profiling to measure stochastic variation in mammalian replication timing. Nature Communications **9**(1), 427 (2018). doi:10.1038/s41467-017-02800-w
27. Takahashi, S., Miura, H., Shibata, T., Nagao, K., Okumura, K., Ogata, M., Obuse, C., Takebayashi, S.-i., Hiratani, I.: Genome-wide stability of the dna replication program in single mammalian cells. Nature Genetics **51**(3), 529–540 (2019). doi:10.1038/s41588-019-0347-5
28. Massey, D.J., Koren, A.: High-throughput analysis of single human cells reveals the complex nature of dna replication timing control. Nature Communications **13**(1), 2402 (2022). doi:10.1038/s41467-022-30212-y
29. Gnan, S., Josephides, J.M., Wu, X., Spagnuolo, M., Saulebekova, D., Bohec, M., Dumont, M., Baudrin, L.G., Fachinetti, D., Baulande, S., Chen, C.-L.: Kronos scrt: a uniform framework for single-cell replication timing analysis. Nature Communications **13**(1), 2329 (2022). doi:10.1038/s41467-022-30043-x
30. Wang, R., Lin, D.-Y., Jiang, Y.: Scope: A normalization and copy-number estimation method for single-cell dna sequencing. Cell Systems **10**(5), 445–4526 (2020). doi:10.1016/j.cels.2020.03.005
31. Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A., Shumansky, K., Chin, S.F., Turashvili, G., Hirst, M., Caldas, C., Marra, M.A., Aparicio, S., Shah, S.P.: Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted

pathways in triple-negative breast cancer. Genome Res **22**(10), 1995–2007 (2012). doi:10.1101/gr.137570.112

32. Michael, P.S., Geoff, M., Florian, M.: Scabsolute: Measuring single-cell ploidy and replication status. bioRxiv, 2022–1114516440 (2022). doi:10.1101/2022.11.14.516440
33. Zaccaria, S., Raphael, B.J.: Characterizing allele- and haplotype-specific copy numbers in single cells with chisel. Nature Biotechnology **39**(2), 207–214 (2021). doi:10.1038/s41587-020-0661-6
34. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., Goodman, N.D.: Pyro: deep universal probabilistic programming. J. Mach. Learn. Res. **20**(1), 973–978 (2019)
35. An integrated encyclopedia of dna elements in the human genome. Nature **489**(7414), 57–74 (2012). doi:10.1038/nature11247
36. Funnell, T., O'Flanagan, C.H., Williams, M.J., McPherson, A., McKinney, S., Kabeer, F., Lee, H., Salehi, S., Vázquez-García, I., Shi, H., Leventhal, E., Masud, T., Eirew, P., Yap, D., Zhang, A.W., Lim, J.L.P., Wang, B., Brimhall, J., Biele, J., Ting, J., Au, V., Van Vliet, M., Liu, Y.F., Beatty, S., Lai, D., Pham, J., Grewal, D., Abrams, D., Havasov, E., Leung, S., Bojilova, V., Moore, R.A., Rusk, N., Uhlitz, F., Ceglia, N., Weiner, A.C., Zaikova, E., Douglas, J.M., Zamarin, D., Weigelt, B., Kim, S.H., Da Cruz Paula, A., Reis-Filho, J.S., Martin, S.D., Li, Y., Xu, H., de Algara, T.R., Lee, S.R., Llanos, V.C., Huntsman, D.G., McAlpine, J.N., Shah, S.P., Aparicio, S.: Single-cell genomic variation induced by mutational processes in cancer. Nature **612**(7938), 106–115 (2022). doi:10.1038/s41586-022-05249-0
37. Bertolin, A.P., Hoffmann, J.S., Gottifredi, V.: Under-replicated dna: The byproduct of large genomes? Cancers (Basel) **12**(10) (2020). doi:10.3390/cancers12102764
38. Moreno, A., Carrington, J.T., Albergante, L., Al Mamun, M., Haagensen, E.J., Komseli, E.S., Gorgoulis, V.G., Newman, T.J., Blow, J.J.: Unreplicated dna remaining from unperturbed s phases passes through mitosis for resolution in daughter cells. Proc Natl Acad Sci U S A **113**(39), 5757–64 (2016). doi:10.1073/pnas.1603252113
39. Edwards, M.M., Zuccaro, M.V., Sagi, I., Ding, Q., Vershkov, D., Benvenisty, N., Egli, D., Koren, A.: Delayed dna replication in haploid human embryonic stem cells. Genome Research (2021). doi:10.1101/gr.275953.121
40. Shaikh, N., Mazzagatti, A., De Angelis, S., Johnson, S.C., Bakker, B., Spierings, D.C.J., Wardenaar, R., Maniati, E., Wang, J., Boemo, M.A., Foijer, F., McClelland, S.E.: Replication stress generates distinctive landscapes of dna copy number alterations and chromosome scale losses. Genome Biology **23**(1), 223 (2022). doi:10.1186/s13059-022-02781-0
41. Salehi, S., Kabeer, F., Ceglia, N., Andronescu, M., Williams, M.J., Campbell, K.R., Masud, T., Wang, B., Biele, J., Brimhall, J., Gee, D., Lee, H., Ting, J., Zhang, A.W., Tran, H., O'Flanagan, C., Dorri, F., Rusk, N., de Algara, T.R., Lee, S.R., Cheng, B.Y.C., Eirew, P., Kono, T., Pham, J., Grewal, D., Lai, D., Moore, R., Mungall, A.J., Marra, M.A., Hannon, G.J., Battistoni, G., Bressan, D., Cannell, I.G., Casbolt, H., Fatemi, A., Jauset, C., Kovačević, T., Mulvey, C.M., Nugent, F., Ribes, M.P., Pearsall, I., Qosaj, F., Sawicka, K., Wild, S.A., Williams, E., Laks, E., Li, Y., O'Flanagan, C.H., Smith, A., Ruiz, T., Lai, D., Roth, A., Balasubramanian, S., Lee, M., Bodenmiller, B., Burger, M., Kuett, L., Tietscher, S., Windhager, J., Boyden, E.S., Alon, S., Cui, Y., Emenari, A., Goodwin, D., Karagiannis, E.D., Sinha, A., Wassie, A.T., Caldas, C., Bruna, A., Callari, M., Greenwood, W., Lerda, G., Eyal-Lubling, Y., Rueda, O.M., Shea, A., Harris, O., Becker, R., Grimaldi, F., Harris, S., Vogl, S.L., Weselak, J., Joyce, J.A., Watson, S.S., Vázquez-Garćıa, I., Tavaré, S., Dinh, K.N., Fisher, E., Kunes, R., Walton, N.A., Sa'd, M.A., Chornay, N., Dariush, A., González-Solares, E.A., González-Fernández, C., Yoldas, A.K., Millar, N., Whitmarsh, T., Zhuang, X., Fan, J., Lee, H., *et al.*: Clonal fitness inferred from time-series modelling of single-cell cancer genomes. Nature (2021). doi:10.1038/s41586-021-03648-3
42. Wagner, J.M., Karnitz, L.M.: Cisplatin-induced dna damage activates replication checkpoint signaling components that differentially affect tumor cell survival. Mol Pharmacol **76**(1), 208–14 (2009). doi:10.1124/mol.109.055178
43. Schukken, K.M., Sheltzer, J.M.: Extensive protein dosage compensation in aneuploid human cancers. Genome Research (2022). doi:10.1101/gr.276378.121
44. da Costa, A.A.B.A., Chowdhury, D., Shapiro, G.I., D'Andrea, A.D., Konstantinopoulos, P.A.: Targeting replication stress in cancer therapy. Nature Reviews Drug Discovery **22**(1), 38–58 (2023). doi:10.1038/s41573-022-00558-5
45. Schmitt, C.A., Wang, B., Demaria, M.: Senescence and cancer — role and therapeutic opportunities. Nature Reviews Clinical Oncology **19**(10), 619–636 (2022). doi:10.1038/s41571-022-00668-4
46. Bartkova, J., Rezaei, N., Liontos, M., Karakaidos, P., Kletsas, D., Issaeva, N., Vassiliou, L.-V.F., Kolettas, E., Niforou, K., Zoumpourlis, V.C., Takaoka, M., Nakagawa, H., Tort, F., Fugger, K., Johansson, F., Sehested, M., Andersen, C.L., Dyrskjot, L., Ørntoft, T., Lukas, J., Kittas, C., Helleday, T., Halazonetis, T.D., Bartek, J., Gorgoulis, V.G.: Oncogene-induced senescence is part of the tumorigenesis barrier imposed by dna damage checkpoints. Nature **444**(7119), 633–637 (2006). doi:10.1038/nature05268
47. Sheltzer, J.M., Torres, E.M., Dunham, M.J., Amon, A.: Transcriptional consequences of aneuploidy. Proc Natl Acad Sci U S A **109**(31), 12644–9 (2012). doi:10.1073/pnas.1209227109
48. Vázquez-García, I., Uhlitz, F., Ceglia, N., Lim, J.L.P., Wu, M., Mohibullah, N., Niyazov, J., Ruiz, A.E.B., Boehm, K.M., Bojilova, V., Fong, C.J., Funnell, T., Grewal, D., Havasov, E., Leung, S., Pasha, A., Patel, D.M., Pourmaleki, M., Rusk, N., Shi, H., Vanguri, R., Williams, M.J., Zhang, A.W., Broach, V., Chi, D.S., Da Cruz Paula, A., Gardner, G.J., Kim, S.H., Lennon, M., Long Roche, K., Sonoda, Y., Zivanovic, O., Kundra, R., Viale, A., Derakhshan, F.N., Geneslaw, L., Issa Bhaloo, S., Maroldi, A., Nunez, R., Pareja, F., Stylianou, A., Vahdatinia, M., Bykov, Y., Grisham, R.N., Liu, Y.L., Lakhman, Y., Nikolovski, I., Kelly, D., Gao, J., Schietinger, A., Hollmann, T.J., Bakhoum, S.F., Soslow, R.A., Ellenson, L.H., Abu-Rustum, N.R., Aghajanian, C., Friedman, C.F., McPherson, A., Weigelt, B., Zamarin, D., Shah, S.P.: Ovarian cancer mutational processes drive site-specific immune evasion. Nature **612**(7941), 778–786 (2022). doi:10.1038/s41586-022-05496-1
49. Drews, R.M., Hernando, B., Tarabichi, M., Haase, K., Lesluyes, T., Smith, P.S., Morrill Gavarró, L., Couturier, D.-L., Liu, L., Schneider, M., Brenton, J.D., Van Loo, P., Macintyre, G., Markowetz, F.: A pan-cancer compendium of chromosomal instability. Nature **606**(7916), 976–983 (2022). doi:10.1038/s41586-022-04789-9
50. Vasudevan, A., Baruah, P.S., Smith, J.C., Wang, Z., Sayles, N.M., Andrews, P., Kendall, J., Leu, J., Chunduri,

N.K., Levy, D., Wigler, M., Storchová, Z., Sheltzer, J.M.: Single-chromosomal gains can function as metastasis suppressors and promoters in colon cancer. Developmental Cell **52**(4), 413–4286 (2020). doi:10.1016/j.devcel.2020.01.034

51. Sohrab, S., Fatemeh, D., Kevin, C., Farhia, K., Nicole, R., Tyler, F., Marc, J.W., Daniel, L., Mirela, A., Kieran, R.C., Andrew, M., Samuel, A., Andrew, R., Sohrab, S., Alexandre, B.-C.: Cancer phylogenetic tree inference at scale from 1000s of single cell genomes. bioRxiv, 2020–0506058180 (2023). doi:10.1101/2020.05.06.058180
52. Kaufmann, T.L., Petkovic, M., Watkins, T.B.K., Colliver, E.C., Laskina, S., Thapa, N., Minussi, D.C., Navin, N., Swanton, C., Van Loo, P., Haase, K., Tarabichi, M., Schwarz, R.F.: Medicc2: whole-genome doubling aware copy-number phylogenies for cancer evolution. Genome Biology **23**(1), 241 (2022). doi:10.1186/s13059-022-02794-9
53. Ranganath, R., Gerrish, S., Blei, D.: Black Box Variational Inference. In: Kaski, S., Corander, J. (eds.) Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 33, pp. 814–822. PMLR, Reykjavik, Iceland (2014). https://proceedings.mlr.press/v33/ranganath14.html

Figures

Figure 1 a) PERT takes scWGS binned read count and CN calls as input and learns somatic copy number, replication states, and cell cycle phase predictions for all cells. Representative FACS isolated G1-phase (left) and S-phase (right) cells from GM18507 (top) and T47D (bottom) cell lines; points colored by HMMcopy state. b) PERT graphical model and table of all parameters. Gray nodes represent terms that are conditioned during Step 2 of the 3-step learning procedure.

Figure 2 a-d) True vs inferred somatic copy number and replication states for 400 S-phase cells from representative simulated dataset P5.8 which has 3 clones and a cell-specific CNA rate of 0.02. Composite CN prior method was used during inference. e-f) Per-bin replication accuracy in S-phase cells for Kronos and PERT with clone and composite CN prior concentrations across varying numbers of clones and cell-specific CNA rates. g) True vs inferred fraction of replicated bins per cell. h) Confusion matrix of true vs PERT cell cycle phase for simulated datasets with $\lambda = 0.75$.

Figure 3 a-c) HMMcopy, PERT copy number, and PERT replication states for GM18507 and T47D cells predicted as S-phase by PERT. d-f) UMAP embeddings of read depth where cells are colored by cell line, FACS cell cycle phase, and PERT predicted phase g) Difference between inferred T47D and GM18507 RT profiles. h) Spearman correlation between inferred T47D and GM18507 RT profiles and Hi-C A/B compartment scores from ENCODE. All cell lines with the 'GM' prefix are lymphoblastoids. i) Pearson correlation between inferred T47D and GM18507 RT profiles when the cell lines are merged into one sample or split when running PERT.

Figure 4 a) Inferred fraction of replicated bins per cell for all cells in permuted datasets, split by FACS phase. b) Fraction of FACS=G1/2 cells mislabeled as S-phase predicted as PERT=G1/2 across all permutation rates. c) Example of a FACS=S PERT=G1/2 GM18507 cell. Bins are colored by HMMcopy (top) and PERT replication (bottom) state. Arrows point to cell-specific CNAs. Title includes quality score and S-phase probability from Laks *et al* classifiers. d-g) Analysis of unsorted DLP+ samples from hTERT and OV2295 cell lines. d) Inferred fraction of replicated bins per S-phase cell. e) Confusion matrix of Laks *et al* vs PERT phase. f-g) Distribution of each cell's HMMcopy ploidy and breakpoint count where each violin corresponds to a unique position in the confusion matrix.

Tables

Additional Files

Supplementary Figures

Supplementary Fig. 1 a) PERT first learns overdispersion (λ) and GC parameters (β_μ, β_σ) from high-confidence G1/2 cells where we condition all bins as unreplicated ($Y = 0$) and CN states (X) according to CN caller results. b) PERT conditions the parameters learned in Step 1 to learn latent replication and somatic CN states in unknown cells. c) Replication timing (ρ) and stochasticity (α) terms learned in Step 2 are conditioned as Step 3 learns latent replication and somatic CN states in high-confidence G1/2-phase cells to search for any missing S-phase cells. d) Overview of clone and composite methods to set copy number prior concentrations (η). Pearson correlation is used to determine similarity.

Figure 5 a) hTERT genotype lineage diagram showing progression from WT to mutant alleles. b) Distribution of ancestral hTERT WT RT values split by clonal CNA type. c) RT profile of ancestral hTERT WT SA039 clone A. d) CN profiles for all hTERT samples, normalized by SA039 clone A and ploidy. Values > 0 are gains, < 0 are losses, and $= 0$ are unaltered. e) SA1292 clone CN profiles. Loci are split into subclonal gain, loss, and unaltered sites. f) SA1292 clone minus reference RT across different CNA types. Reference RT is the weighted average of all unaltered clones at gain and loss loci or a random clone at unaltered loci. g) Same as f but computed across all hTERT samples. h) Difference between sample and reference RT for samples with chrX events, split by chromosome. Samples with no chrX events were used as the reference. i) SIGNALS total and allele specific CN from SA1054 high-confidence G1/2-phase cells. j) Difference between SA1054 clone and reference RT for clones with chrX allelic imbalance. Allelically balanced clones B-C were used as the reference. Clones D-E were combined as they have identical chrX CN and allelic ratios.

Figure 6 a) Schematic of time-series DLP+ sampling for untreated and cisplatin-treated TNBC PDXs. b-e) Representative SA1035 untreated sample. b-c) Relative fraction of each clone within G1/2- and S-phase cells. d-e) Comparison of each clone's fraction in S- vs G1/2-phase populations at each timepoint. Dashed gray line represents equal prevalence in both phases. Triangles denote clone and timepoint combinations significantly ($p < 0.01$) enriched or depleted for S-phase cells via hypergeometric test. Distance from the dashed gray line represents each point's continuous SPE score. f) Relationship between SPE score and clone expansion for all TNBC PDX clone and timepoint combinations with >10 G1/2-phase cells, split by cisplatin status. Lines represent linear regression fits with shaded areas representing 95% confidence intervals.

Table 1 Parameters used in the PERT model

Name	Domain/Distribution	Description
λ	$(0, 1)$	NB event success probability (overdispersion)
δ	$[1, \infty)$	# of NB events (Eq 4)
γ	$[0, 1]$	GC content of locus
β_μ	$\sim \mathrm{N}(0, 1)$	GC bias polynomial coefficient means
β_σ	$(0, \infty)$	GC bias polynomial coefficient stdevs
β	$\sim \mathrm{N}(\beta_\mu, \beta_\sigma)$	GC bias polynomial coefficients
μ_μ	$(0, \infty)$	coverage/ploidy mean (Eq 8)
μ_σ	$(0, \infty)$	coverage/ploidy stdev (Eq 9)
μ	$\sim \mathrm{N}(\mu_\mu, \mu_\sigma)$	coverage/ploidy scaling term
τ	$[0, 1]$	time in S-phase
ρ	$\sim \mathrm{Beta}(1, 1)$	replication timing
α	$\sim \Gamma(2, 0.2)$	replication stochasticity term
ϕ	$(0, 1)$	replication probabilities (Eq 5)
Y	$\sim \mathrm{Bernoulli}(\phi)$	replication status
η	$[0, 1]$	copy number prior concentration
π	$\sim \mathrm{Dir}(\eta)$	copy number probabilities
X	$\sim \mathrm{Cat}(\pi)$	copy number state
Z	$\sim \mathrm{NB}(\delta, \lambda)$	read depth
K		GC bias polynomial degree
L		# libraries
M		# genomic loci
N		# cells
P		# copy number states

Table 2 Parameters values for simulated datasets

Datatag	# clones	CNA rate	λ	α	ENCODE RT	β_μ
D1	1	0	0.75	10	MCF7	[1.2, 0]
D2	1	0	0.75	5	MCF7	[1.2, 0]
D3	1	0	0.75	15	MCF7	[1.2, 0]
D4	1	0	0.75	10	MCF7	[-1.2, 0]
D5	1	0	0.5	10	MCF7	[1.2, 0]
D6	1	0	0.9	10	MCF7	[1.2, 0]
D7	1	0.02	0.75	10	MCF7	[1.2, 0]
D8	1	0.05	0.75	10	MCF7	[1.2, 0]
D9	1	0	0.99	10	MCF7	[1.2, 0]
D10	1	0	0.6	10	MCF7	[1.2, 0]
P1	3	0	0.75	10	MCF7	[1.2, 0]
P1	3	0	0.75	10	MCF7	[1.2, 0]
P2	3	0.02	0.75	10	MCF7	[1.2, 0]
P3	3	0.05	0.75	10	MCF7	[1.2, 0]
P4	3	0	0.75	5	MCF7	[1.2, 0]
P5	3	0.02	0.75	5	MCF7	[1.2, 0]
P6	3	0.05	0.75	5	MCF7	[1.2, 0]
P7	3	0	0.75	15	MCF7	[1.2, 0]
P8	3	0.02	0.75	15	MCF7	[1.2, 0]
P9	3	0.05	0.75	15	MCF7	[1.2, 0]
P10	4	0.02	0.75	10	GM-06990 -12801 -12812 -12813	[1.2, 0]
P11	4	0.02	0.75	10	BJ MCF7 HEPG2 GM12813	[1.2, 0]

Supplementary Fig. 2 a-b) Feature importance scores and train-test ROC curve for the DLP+ cell cycle classifier described in Laks *et al.* Features beginning with 'slope' or 'correlation' measure the relationship between GC and various forms of normalized read depth. c-g) Kronos and PERT scRT errors for true S-phase cells in dataset P5.8. c-d) True vs inferred S-phase times for each method. e) HMMcopy states for all S-phase cells. f-g) Kronos and PERT per-bin replication accuracies. Rows in e-g are sorted by clone and true S-phase time.

Supplementary Fig. 3 Comparison of Kronos vs PERT with clone and composite CN prior methods in S-phase cells. a-b) Parameter sweep across cell-specific CNA rate. c-e) Parameter sweep across number of clones. f) Accuracy for all unique combinations of simulation parameters (datatag). There are 10 unique datasets for each datatag. g) Parameter sweep across the replication stochasticity term α. h) Parameter sweep across the read depth overdispersion term λ. i) Parameter sweep across the GC bias slope term $\beta_{\mu,0}$.

Supplementary Fig. 4 Same as Supplementary Fig. 3 but for copy number performance between two PERT variants.

Supplementary Fig. 5 Same as Supplementary Fig. 3 but for cell cycle phase performance between PERT (composite) and the Laks classifier.

Supplementary Fig. 6 a-b) HMMcopy states for FACS G1/2- and S-phase cells. c-d) Histogram of inferred fraction of replicated bins for FACS S-phase cells where colors represent cell line and clone ID. e-f) Inferred cell line RT profiles for chromosome 1 when each cell line is split or merged before passing into PERT.

Supplementary Fig. 7 a) Pairwise Pearson correlation between true and inferred clone RT profiles in simulated datasets where each clone follows a unique ENCODE cell line RT profile. Cell lines with the 'GM' prefix are derived from B-lymphocytes, BJ from human foreskin fibroblasts, MCF7 from breast cancer, and HepG2 from liver cancer. b-e) True and inferred clone RT profiles across the whole genome and chr1.

Supplementary Fig. 8 a) Pairwise Pearson correlation between inferred cell line RT profiles across all permuted datasets. Datasets A-C have the lowest permutation rate (0.01); U-W have the highest permutation rate (0.75). b-g) Pairwise scatterplots of orthogonal cell cycle phase features for FACS=G1/2 cells mislabeled as S-phase. Cells are colored by their predicted PERT phase.

Supplementary Fig. 9 a) Reads per million, b) HMMcopy states, c) PERT copy number states, and d) PERT replication states for cells predicted as S-phase and e-h) G1/2-phase by PERT. Rows (cells) are sorted first by clone ID and then fraction of replicated bins. i) Confusion matrix of Laks *et al.* vs PERT cell cycle phases. j-k) PCA embedding of reads per million where cells are colored by clone ID and PERT phase.

Supplementary Fig. 10 SA1292 cell features correlated with quality and cell cycle phase where each violin corresponds to a unique position of the Laks vs PERT phase confusion matrix. a) Corrected number of HMMcopy

copy number breakpoints. b) HMMcopy ploidies. c) Corrected median absolute deviation of reads per million between neighboring bins. d) Normalized autocorrelation of reads per million profiles. e) Intercept and f) slope coefficients from a linear regression fit between GC and read count.

Supplementary Fig. 11 a-d) Pairwise CN similarity and RT Pearson correlation between hTERT and OV2295 samples and clones with >10 S-phase cells. e-g) Clone CN and RT profiles for hTERT WT sample SA039. h) Distribution of ancestral hTERT WT RT values split by whether a locus contains a clonal CNA breakpoint across all hTERT samples. i) SA1292 RT profiles of clones B and C relative to clone A. Arrows indicate subclonal CNAs noted in Fig. 5e. j) Each hTERT sample's per-cell time from scheduled replication (T-width) statistic vs number of clones.

Supplementary Fig. 12 ChrX RT profiles across all hTERT and OV2295 cell lines. a) Sample RT profiles. b-j) Clone RT profiles in each sample. The number of S-phase cells in each clone is denoted by n in the legend.

Supplementary Fig. 13 a-c) Difference between clone and reference RT for clones with chrX allelic imbalance. Allelically balanced clones E-J were used as the reference. Clones A-B were combined as they have identical copy number and allelic ratios on chrX. d) SIGNALS total and allele specific CN from high-confidence G1/2-phase cells.

Supplementary Fig. 14 a) PERT percent of G1/2-phase cells in each clone vs measured doubling time and percent of scRNA cells in G1-phase for gastric cancer cell lines sequenced with 10x scWGS [8]. Dashed lines represent weighted least squares fits where weight is the number of cells per clone. b) Schematic of time-series DLP+ sampling for hTERT WT and p53-/- cell lines. b) Relationship between SPE and clone expansion for each clone and timepoint combination with >10 G1/2-phase cells across all three samples and c-e) split by sample. Lines represent linear regression fits with shaded areas representing 95% confidence intervals.

Supplementary Fig. 15 i) Relative fraction of each clone within G1/2- and ii) S-phase populations. iii-iv) Comparison of each clone's fraction in S- vs G1/2-phase populations at each timepoint. Dashed gray line represents equal prevalence in both phases. Triangles denote clone and timepoint combinations significantly ($p_{adj} < 10^{-2}$) enriched or depleted for S-phase cells via hypergeometric test. Distance from the dashed gray line represents each point's continuous SPE score. v) Relationship between SPE and expansion for each clone and timepoint combination with >10 G1/2-phase cells. Lines represent linear regression fits with shaded areas representing 95% confidence intervals. a-h) Each row corresponds to a unique sample.

Supplementary Fig. 16 a) Pairwise Pearson correlation between treated and untreated sample RT profiles. b) Distribution of cell S-phase times between treated and untreated samples. c) Total fraction of S-phase cells between treated and untreated samples. d) Total fraction of S-phase cells in each clone across all samples, split by treatment status.

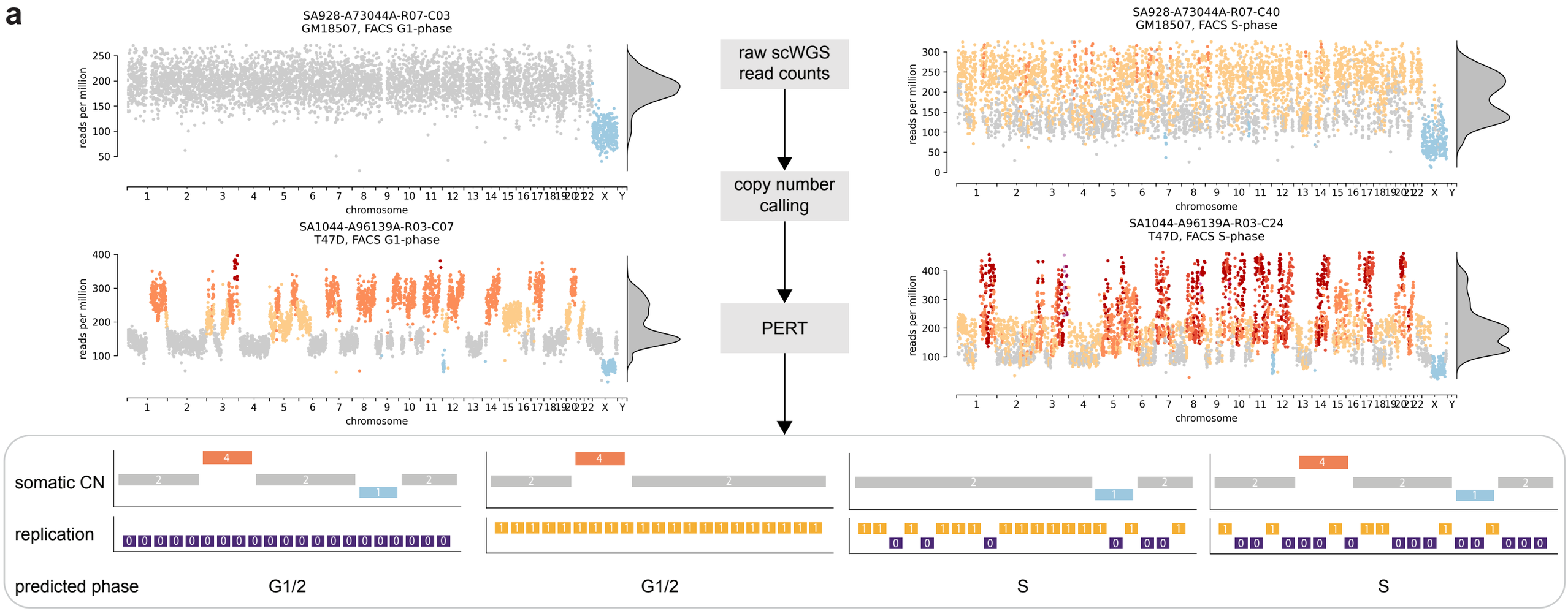
Additional file 1 — HMMcopy and SIGNALS heatmaps for all DLP+ samples

Heatmaps of HMMcopy and SIGNALS results for all high-confidence G1/2-phase cells in each sample initially reported in Funnell *et al* [36].

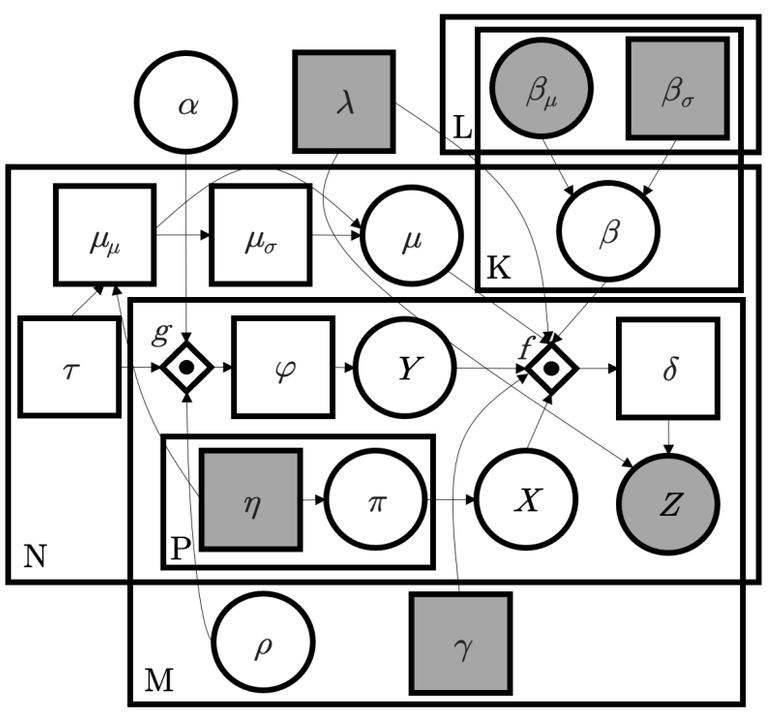
Additional file 2 — PERT heatmaps

Heatmaps of PERT input (read depth and called CN states) and output (somatic CN and replication states) for all scWGS samples analyzed. Top rows correspond to cells predicted as S-phase by PERT, bottom rows correspond to cells predicted as G1/2-phase by PERT. Each figure represents a unique sample. Input CN states were from HMMcopy for all DLP+ samples and CellRanger-DNA for all 10x scWGS samples (NCI-N87, HGC-27, and SNU-668). PERT was run at 500kb resolution unless otherwise noted as 20kb for select samples.

a



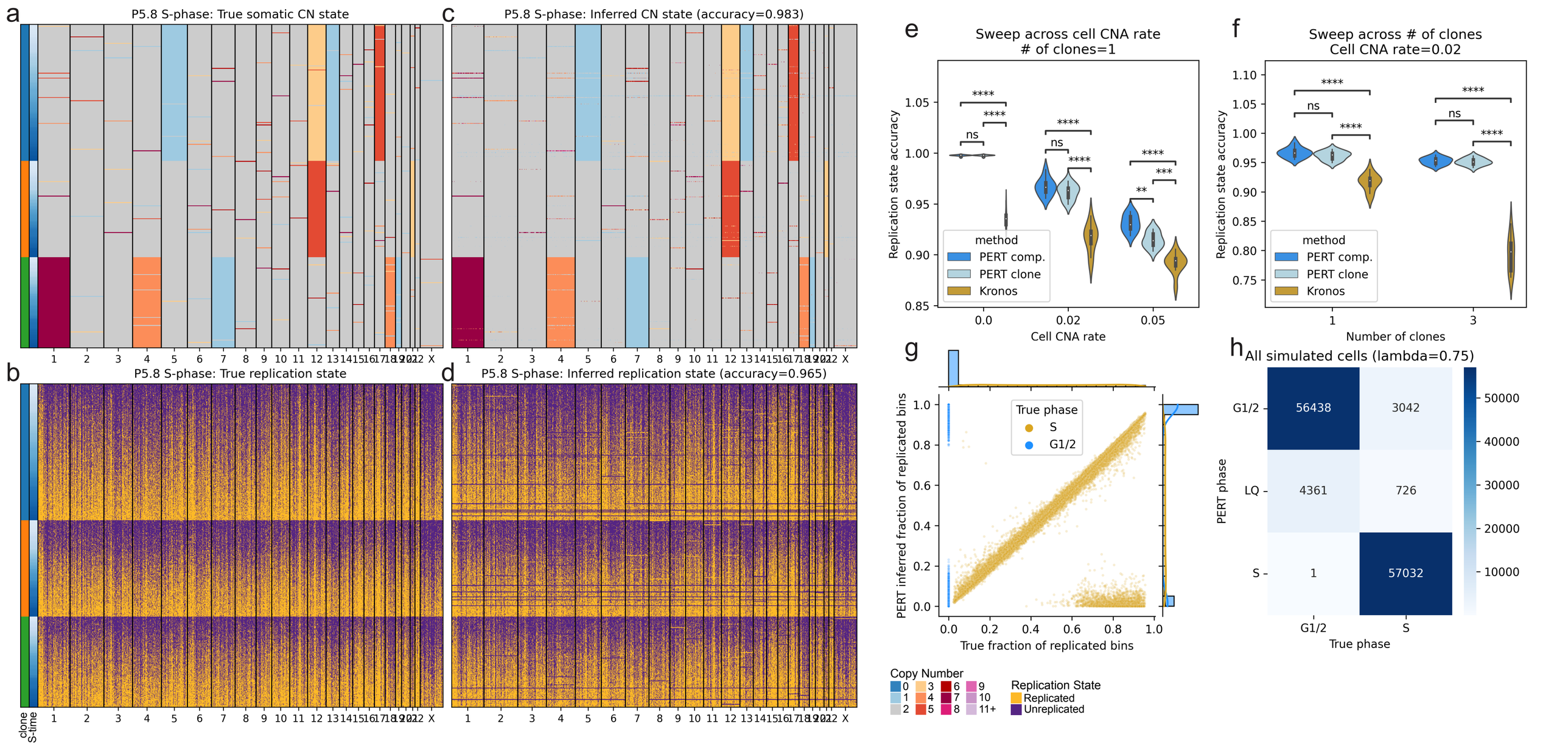
b

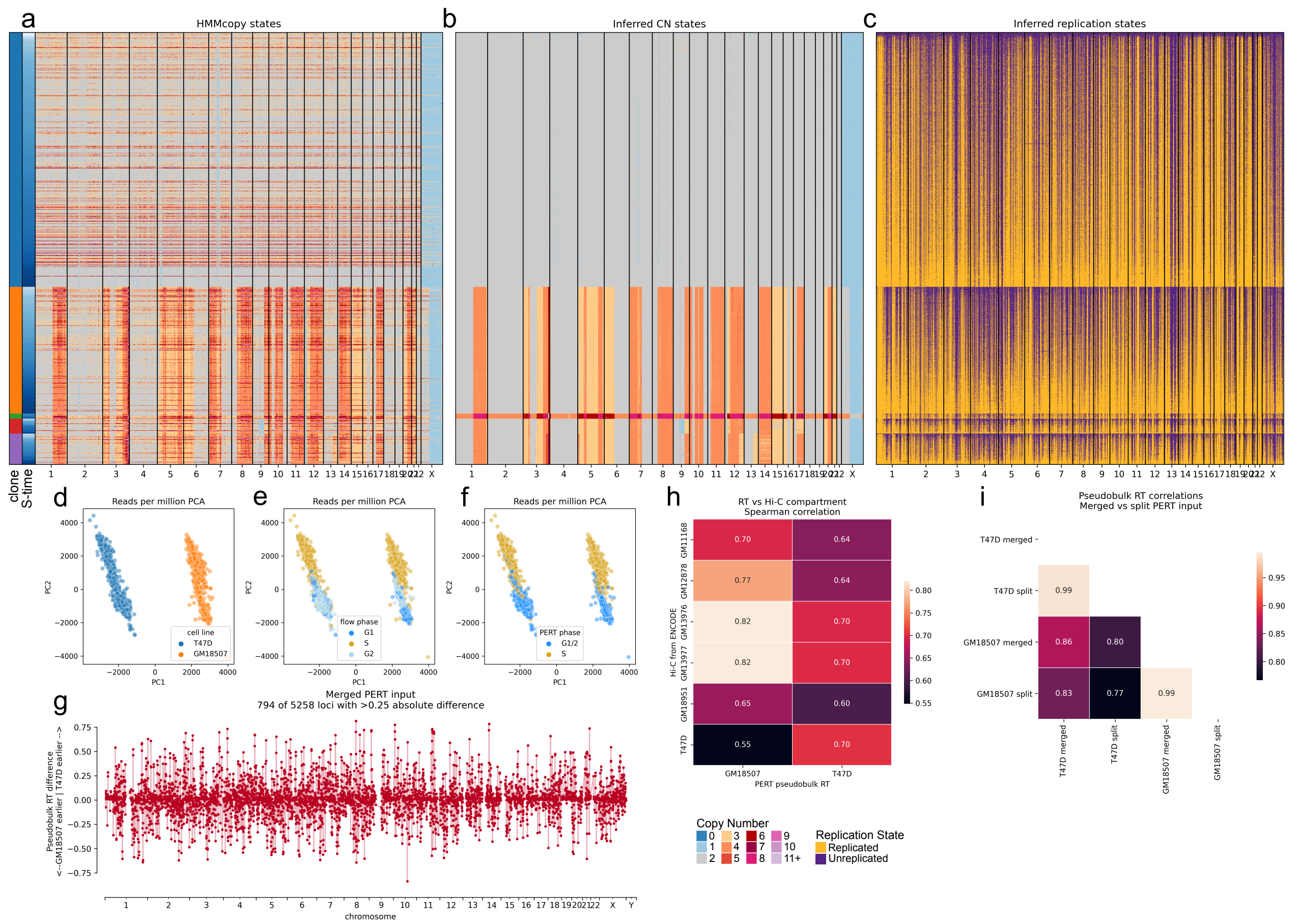


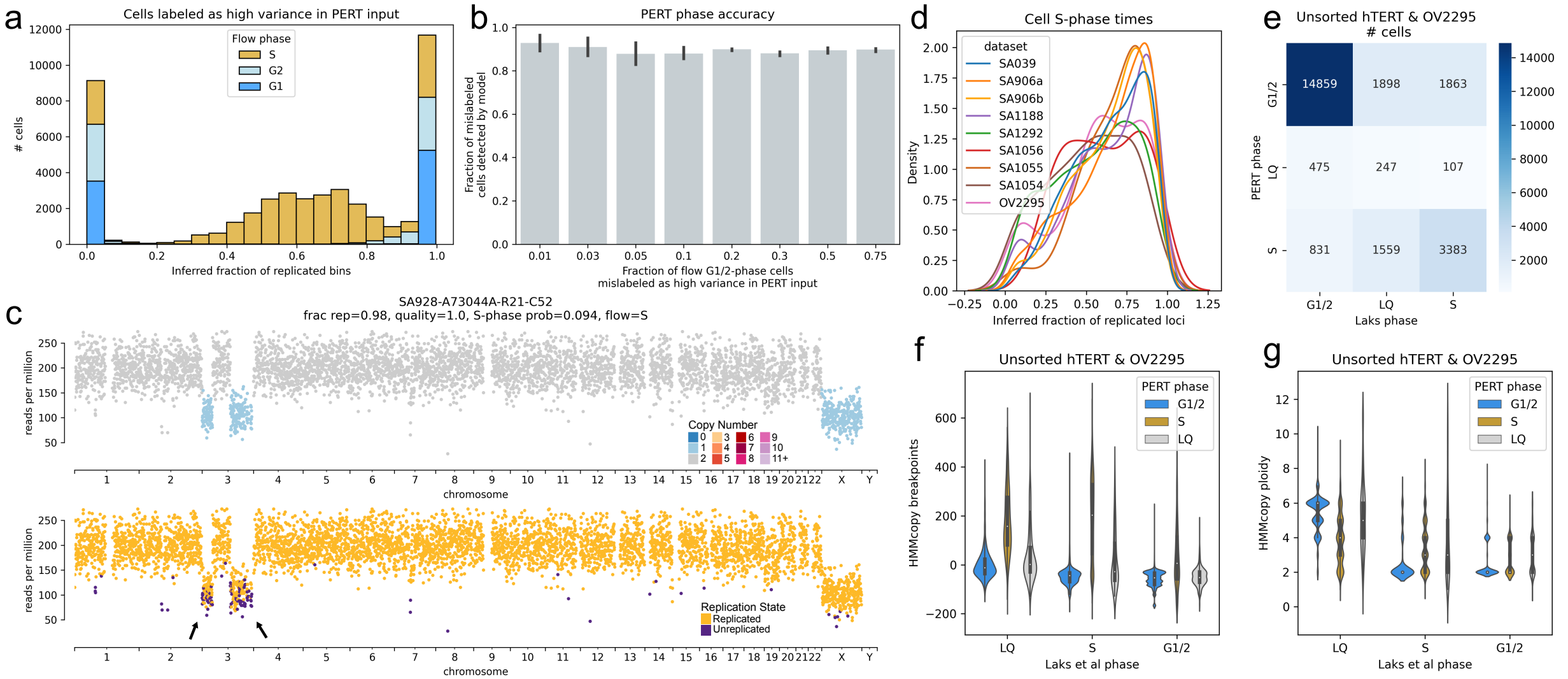
Name	Domain/Distribution	Description
λ	$(0, 1)$	NB event success probability (overdispersion)
δ	$[1, \infty)$	# of NB events (Eq 4.4)
γ	$[0, 1]$	GC content of locus
β_μ	$\sim \mathrm{N}(0, 1)$	GC bias polynomial coefficient means
β_σ	$(0, \infty)$	GC bias polynomial coefficient stdevs
β	$\sim \mathrm{N}(\beta_\mu, \beta_\sigma)$	GC bias polynomial coefficients
μ_μ	$(0, \infty)$	coverage/ploidy mean (Eq 4.8)
μ_σ	$(0, \infty)$	coverage/ploidy stdev (Eq 4.9)
μ	$\sim \mathrm{N}(\mu_\mu, \mu_\sigma)$	coverage/ploidy scaling term
τ	$[0, 1]$	time in S-phase
ρ	$\sim \mathrm{Beta}(1, 1)$	replication timing
α	$\sim \Gamma(2, 0.2)$	replication stochasticity term
ϕ	$(0, 1)$	replication probabilities (Eq 4.5)
Y	$\sim \mathrm{Bernoulli}(\phi)$	replication status
η	$[0, 1]$	copy number prior concentration
π	$\sim \mathrm{Dir}(\eta)$	copy number probabilities
X	$\sim \mathrm{Cat}(\pi)$	copy number state
Z	$\sim \mathrm{NB}(\delta, \lambda)$	read depth
K		GC bias polynomial degree
L		# libraries
M		# genomic loci
N		# cells
P		# copy number states

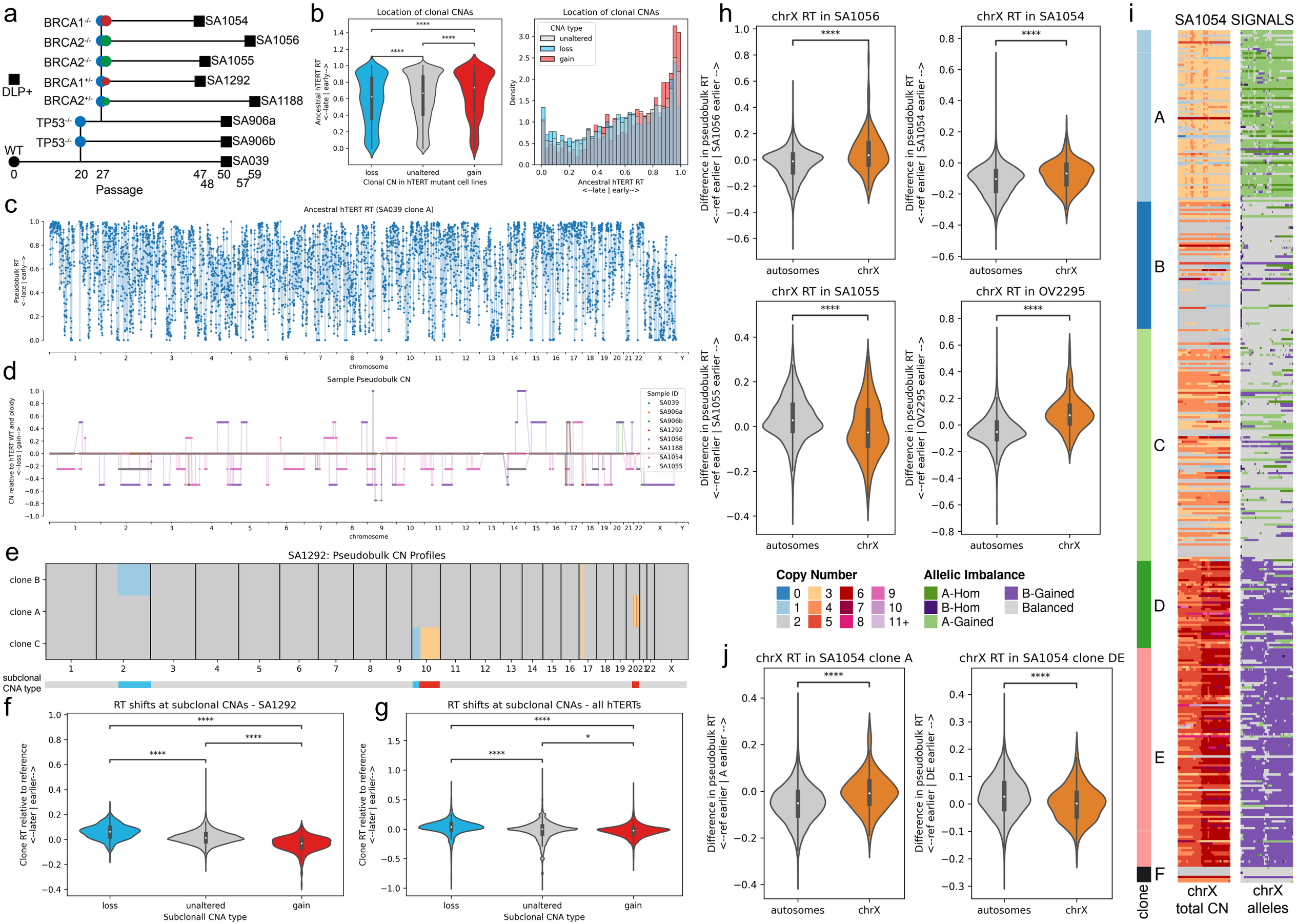
Replication State
- Replicated
- Unreplicated

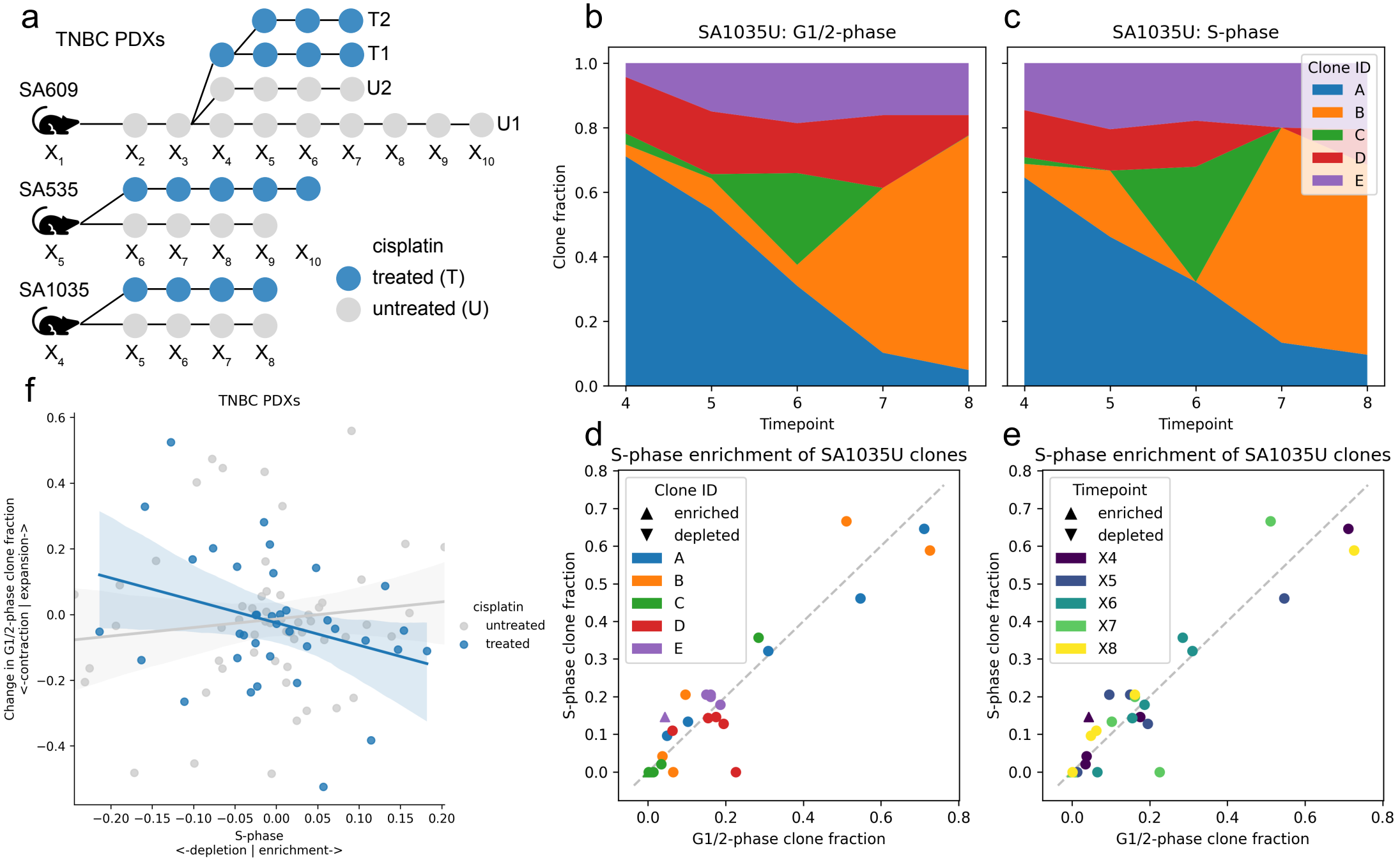
Copy Number
0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11+


a
P5.8 S-phase: True somatic CN state
c
P5.8 S-phase: Inferred CN state (accuracy=0.983)
b
P5.8 S-phase: True replication state
d
P5.8 S-phase: Inferred replication state (accuracy=0.965)
clone
S-time
e
Sweep across cell CNA rate
# of clones=1
Replication state accuracy
Cell CNA rate
method
PERT comp.
PERT clone
Kronos
f
Sweep across # of clones
Cell CNA rate=0.02
Number of clones
g
PERT inferred fraction of replicated bins
True fraction of replicated bins
True phase
S
G1/2
h
All simulated cells (lambda=0.75)
PERT phase
True phase
56438
3042
4361
726
1
57032
Copy Number
Replication State
Replicated
Unreplicated


a
HMMcopy states
b
Inferred CN states
c
Inferred replication states
clone
S-time
d
Reads per million PCA
cell line
T47D
GM18507
e
Reads per million PCA
flow phase
G1
S
G2
f
Reads per million PCA
PERT phase
G1/2
S
PC1
PC2
g
Merged PERT input
794 of 5258 loci with >0.25 absolute difference
Pseudobulk RT difference
<--GM18507 earlier | T47D earlier -->
chromosome
h
RT vs Hi-C compartment
Spearman correlation
Hi-C from ENCODE
GM11168
GM12878
GM13976
GM13977
GM18951
T47D
GM18507
PERT pseudobulk RT
0.70
0.64
0.77
0.64
0.82
0.70
0.82
0.70
0.65
0.60
0.55
0.70
i
Pseudobulk RT correlations
Merged vs split PERT input
T47D merged
T47D split
GM18507 merged
GM18507 split
0.99
0.86
0.80
0.83
0.77
0.99
Copy Number
0 1 2 3 4 5 6 7 8 9 10 11+
Replication State
Replicated
Unreplicated


a
Cells labeled as high variance in PERT input
Flow phase
S
G2
G1
cells
Inferred fraction of replicated bins
b
PERT phase accuracy
Fraction of mislabeled cells detected by model
Fraction of flow G1/2-phase cells mislabeled as high variance in PERT input
d
Cell S-phase times
dataset
SA039
SA906a
SA906b
SA1188
SA1292
SA1056
SA1055
SA1054
OV2295
Density
Inferred fraction of replicated loci
e
Unsorted hTERT & OV2295
cells
PERT phase
Laks phase
G1/2
LQ
S
14859
1898
1863
475
247
107
831
1559
3383
c
SA928-A73044A-R21-C52
frac rep=0.98, quality=1.0, S-phase prob=0.094, flow=S
reads per million
Copy Number
chromosome
Replication State
Replicated
Unreplicated
f
Unsorted hTERT & OV2295
PERT phase
HMMcopy breakpoints
Laks et al phase
g
Unsorted hTERT & OV2295
HMMcopy ploidy


a
BRCA1-/-
BRCA2-/-
BRCA2-/-
BRCA1+/-
BRCA2+/-
TP53-/-
TP53-/-
WT
DLP+
SA1054
SA1056
SA1055
SA1292
SA1188
SA906a
SA906b
SA039
0
20
27
47
48
50
57
59
Passage
b
Location of clonal CNAs
Ancestral hTERT RT <--late | early-->
loss
unaltered
gain
Clonal CN in hTERT mutant cell lines
Location of clonal CNAs
CNA type
unaltered
loss
gain
Density
Ancestral hTERT RT <--late | early-->
c
Ancestral hTERT RT (SA039 clone A)
Pseudobulk RT <--late | early-->
chromosome
d
Sample Pseudobulk CN
CN relative to hTERT WT and ploidy <--loss | gain-->
Sample ID
SA039
SA906a
SA906b
SA1292
SA1056
SA1188
SA1054
SA1055
chromosome
e
SA1292: Pseudobulk CN Profiles
clone B
clone A
clone C
subclonal CNA type
f
RT shifts at subclonal CNAs - SA1292
Clone RT relative to reference <--later | earlier-->
loss
unaltered
gain
Subclonall CNA type
g
RT shifts at subclonal CNAs - all hTERTs
Clone RT relative to reference <--later | earlier-->
loss
unaltered
gain
Subclonal CNA type
h
chrX RT in SA1056
chrX RT in SA1054
chrX RT in SA1055
chrX RT in OV2295
Difference in pseudobulk RT <--ref earlier | SA1056 earlier -->
Difference in pseudobulk RT <--ref earlier | SA1054 earlier -->
Difference in pseudobulk RT <--ref earlier | SA1055 earlier -->
Difference in pseudobulk RT <--ref earlier | OV2295 earlier -->
autosomes
chrX
Copy Number
0
1
2
3
4
5
6
7
8
9
10
11+
Allelic Imbalance
A-Hom
B-Hom
A-Gained
B-Gained
Balanced
i
SA1054 SIGNALS
A
B
C
D
E
F
clone
chrX total CN
chrX alleles
j
chrX RT in SA1054 clone A
chrX RT in SA1054 clone DE
Difference in pseudobulk RT <--ref earlier | A earlier -->
Difference in pseudobulk RT <--ref earlier | DE earlier -->
autosomes
chrX


a
TNBC PDXs
SA609
T2
T1
U2
U1
X1 X2 X3 X4 X5 X6 X7 X8 X9 X10
SA535
X5 X6 X7 X8 X9 X10
SA1035
X4 X5 X6 X7 X8
cisplatin
treated (T)
untreated (U)
b
SA1035U: G1/2-phase
Clone fraction
Timepoint
c
SA1035U: S-phase
Clone ID
A
B
C
D
E
Timepoint
d
S-phase enrichment of SA1035U clones
Clone ID
enriched
depleted
A
B
C
D
E
S-phase clone fraction
G1/2-phase clone fraction
e
S-phase enrichment of SA1035U clones
Timepoint
enriched
depleted
X4
X5
X6
X7
X8
S-phase clone fraction
G1/2-phase clone fraction
f
TNBC PDXs
Change in G1/2-phase clone fraction
<-contraction | expansion->
S-phase
<-depletion | enrichment->
cisplatin
untreated
treated