Integrative analysis of DNA replication origins and ORC/MCM binding sites in human cells reveals a lack of overlap

Mengxue Tian[1,2], Zhenjia Wang[1], Zhangli Su[2,3], Etsuko Shibata[2,3], Yoshiyuki Shibata[2,3], Anindya Dutta[2,3,*], Chongzhi Zang[1,2,4,*]

[1]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA 22908;

[2]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA 22908;

[3]Department of Genetics, University of Alabama, Birmingham, AL 35233;

[4]Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908

* Correspondence should be addressed to: duttaa@uab.edu (A.D.) or zang@virginia.edu (C.Z.)

Keywords: Chromosomal origins, Origin-recognition-complex, Chromatin, Promoters, G-quadruplex

Abstract

Based on experimentally determined average inter-origin distances of ~100 kb, DNA replication initiates from ~50,000 origins on human chromosomes in each cell cycle. The origins are believed to be specified by binding of factors like the Origin Recognition Complex (ORC) or CTCF or other features like G-quadruplexes. We have performed an integrative analysis of 113 genome-wide human origin profiles (from five different techniques) and 5 ORC-binding profiles to critically evaluate whether the most reproducible origins are specified by these features. Out of ~7.5 million union origins identified by all datasets, only 0.27% were reproducibly obtained in at least 20 independent SNS-seq datasets and contained in initiation zones identified by each of three other techniques (20,250 shared origins), suggesting extensive variability in origin usage and identification. 21% of the shared origins overlap with transcriptional promoters, posing a conundrum. Although the shared origins overlap more than union origins with constitutive CTCF binding sites, G-quadruplex sites and activating histone marks, these overlaps are comparable or less than that of known Transcription Start Sites, so that these features could be enriched in origins because of the overlap of origins with epigenetically open, promoter-like sequences. Only 6.4% of the 20,250 shared origins were within 1 kb from any of the ~13,000 reproducible ORC binding sites in human cancer cells, and only 4.5% were within 1 kb of the ~11,000 union MCM2-7 binding sites in contrast to the nearly 100% overlap in the two comparisons in the yeast, *S. cerevisiae*. Thus, in human cancer cell lines, replication origins appear to be specified by highly variable stochastic events dependent on the high epigenetic accessibility around promoters, without extensive overlap between the most reproducible origins and currently known ORC- or MCM-binding sites.

Introduction

DNA replication is essential for the duplication of a cell and the maintenance of the eukaryotic genome. To replicate the human diploid genome of ~3 billion base pairs, efficient cellular programs are coordinated to ensure genetic information is accurately copied. In each human cell cycle, replication starts from ~50,000 genomic locations called replication origins (Hu and Stillman 2023). At an origin of replication, the double-stranded DNA is unwound, and primase-DNA polymerase alpha lays down the first RNA primer extended as DNA. Once replication is initiated, the helicase involved in unwinding the origin continues to unwind DNA on either side and the replication proteins, including DNA polymerases, help copy the single-stranded DNA. Unlike yeast origins, human and most eukaryotic origins do not seem to have a clear DNA sequence preference(Hu and Stillman 2023), (Leonard and Méchali 2013). Yet it becomes clear that chromatin context and other DNA-based activity are important factors for human and non-yeast eukaryotic origin selection. A great amount of effort has been spent to understand how human origins are specified and sometimes different trends are observed regarding genomic features enriched at human origins. For example, G-quadruplex motifs were found to be associated with human origins(Besnard et al. 2012), but can only explain a small subset of origins even after controlling for technical biases(Foulk et al. 2015). To better resolve these issues, we reasoned that a systematic analysis of all available human origin-mapping datasets will help uncover potential technical and biological details that affect origin detection.

To profile replication origins genome-wide, several sequencing-based methods have been developed, including Short Nascent Strand (SNS)-seq (Foulk et al. 2015), Repli-seq (Hansen et al. 2010), Rerep-seq (Menzel et al. 2020), and Bubble-seq (Mesner et al. 2013). SNS-seq assays across different human cell types identified a subset of origins that can explain ~80% of origins in any tested cell type(Akerman et al. 2020), however no study has systematically examined the consistency within and between different origin-mapping techniques. These

methods are based on different molecular capture strategies and therefore are expected to have technical biases. For instance, SNS-seq utilizes lambda exonuclease (λ-exo) to enrich for RNA-primed newly replicated DNA by removing parental DNA fragments. However, the products may be contaminated with chromosomal DNA fragments or be affected by the cutting bias of λ-exo against GC-rich sequences (Foulk et al. 2015), although it has been proposed that one can experimentally correct and control for such biases(Akerman et al. 2020). SNS-seq is the only method that yields origins at a resolution of a few hundred base-pairs, with all other methods delineating larger initiation zones (IZ) that may have multiple origins. Repli-seq relies on nucleotide pulse-labeling and antibody enrichment of newly replicated DNA superimposed on cells fractionated at different parts of S phase by flow cytometry. Here the results may be contaminated by DNA non-specifically associated with the antibody(Zhao et al. 2020). Rerep-seq only captures the new synthesized sequences that replicate more than once when replication is dysregulated, and these origins tend to be enriched in the early replicating, epigenetically open parts of the genome (Menzel et al. 2020). Bubble-seq (Mesner et al. 2013) generates long reads because it captures the origin-containing replication intermediates by selection of DNA bubbles, and so may enrich for origins that are flanked by pause sites. Okazaki-seq (OK-seq) identifies sites where the direction of the Okazaki fragments changes from left-ward to right-ward on the chromosome revealing sites where the two lagging strands diverge from each other (Petryk et al. 2016). Because the technical biases mentioned above are associated with specific techniques, we suspect that origins captured across different techniques are less likely to be affected by those biases. Given the large number of datasets in the public domain (>100 datasets for human origins), this is an opportune time to study the reproducibility between the studies and determine the most reproducible and consistent origins identified by SNS-seq and confirmed by each of the other methods. This most reproducible group of origins (shared origins) from different research groups using different techniques and different cell lines, will be far fewer than all the origins reported to date and minimize

complications from stochastic (noisy) firing of origins, but they are best expected to fit the current hypotheses regarding origin specification.

Yeast origin recognition complex (ORC) binds to double-stranded DNA with sequence specificity, helps to load minichromosome maintenance protein complex (MCM) and thus prepare the origins for subsequent firing (Bell and Dutta 2002; Remus et al. 2009; Bell and Stillman 1992), (Costa and Diffley 2022). Consistent with this, a role of ORC in loading MCM proteins was also described in Xenopus egg extracts (Rowles et al. 1999; Coleman et al. 1996). All of this leads to the expectation that in human cells there will be significant concordance between ORC binding sites and efficient origins of replication. Efforts to define double-stranded DNA sequences bound specifically by human ORC find very little sequence specificity (Vashee et al. 2003; Hoshina et al. 2013), and this may be responsible for the lack of sequence specificity in human origins. This difference between yeast and human ORC is attributed to sequence features in ORC4, one of the subunits of the origin recognition complex (Lee et al. 2021). When the yeast ORC was humanized by appropriate mutation in the ORC4 subunit, the sequence specificity of ORC binding was lost even in yeast. Despite this loss of sequence specificity, the mutant yeast ORC loaded MCM proteins, and the yeast replicated DNA and survived. Thus, even if human origins of replication do not have specific sequences, one should expect some concordance between ORC binding sites and the most reproducible origins of replication. An additional complexity is that genome-wide analysis found active origins and dormant origins determined by SNS-seq and OK-seq have little difference in ORC or MCM density (Sugimoto et al. 2018; Kirstein et al. 2021). Finally, a few publications have reported significant replication after genetic mutation of Drosophila and human ORC subunit genes that reduce the expressed proteins to near undetectable levels (Park and Asano 2008; Shibata et al. 2016; Okano-Uchida et al. 2018; Shibata and Dutta 2020). Definition of the most reproducible and consistent origins identified by different methods from different research groups in different

cell lines thus provides a unique opportunity to determine how much overlap is seen between the highly reproducible human origins and the ORC-binding sites reported in the literature from ChIP-seq studies.

To understand the genome-wide distribution patterns of replication origins in an unbiased way, we performed an integrative analysis of 113 DNA replication origin datasets. Because SNS-seq is unique in identifying origins of a few hundred bases, while the other methods identify larger initiation zones that contain origins, we first prepared a list of high-resolution SNS-seq origins that have been identified in at least 20 of the 66 SNS-seq datasets in this study. We next determined those reproducible SNS-seq origins that overlap with initiation zones identified by each of the three techniques (Repli-seq, OK-seq and Bubble-seq) to identify 20,250 high-confidence origins that are shared between SNS-seq and every other method identifying initiation zones. Using these shared origins, which overlap significantly with transcription promoters, we tested whether G-quadruplex sites, CTCF binding sites, ORC binding sites or MCM binding sites help specify origins.

Results

A total of 7,459,709 origins from 113 datasets show similar but different genomic features associated with each origin-mapping technique

We collected 113 publicly available replication origin identification studies in different human cell types from five different techniques (Supplementary Figure S1a), including SNS-seq, Repli-seq, Rerep-seq, OK-seq and Bubble-seq. The complete list of datasets used in this analysis can be found in Supplementary Table 1. We processed all the datasets using the first two steps of a pipeline with different parameters considering the various resolution of different techniques, as SNS-seq is unique in identifying high-resolution origins of replication (Fig. 1a). Each of the 113 datasets yielded at least 1000 origins. We merged origins that overlap for at least 1bp from

each other and cut the merged regions into 300bp segments, considering origin lengths were significantly longer for Bubble-seq and OK-seq, methods known to identify initiation zones (Supplementary Fig. S1b). A total of ~7,460,000 union origins were discovered from all techniques (Fig. 1b).

Principal Component Analysis (PCA) of all origin datasets shows that origin locations from the same technique are more similar to each other than from different techniques (Fig. 1c). This is confirmed by pair-wise correlations of the datasets, where it is evident that each method identifies origins that are best correlated with origins identified by that method alone (Supplementary Fig. S3) Although the most popular technique, SNS-seq, shows some variability in the PCA (Fig. 1c), it is also the one that is best correlated with origins identified by other SNS-seq datasets. Re-rep-seq and Bubble-seq seemed to identify initiation zones that were also better correlated with SNS-seq origins (Supplementary Fig. S3). Since Re-rep-seq tends to identify origins that are enriched in early replicating, gene-rich parts of the genome, this result suggests that SNS-seq and Bubble-seq also have some bias for origins in early replicating, gene-rich parts of the genome. However, we did not observe clear trends of clustering of origins identified in similar cell types (Supplementary Fig. S1c), even when we focused only on SNS-seq origins (Supplementary Fig. S1d). The one exception was the SNS-seq origins from T lymphoblasts, which were very closely clustered with each other in six datasets, but these were all done in a the CCRF-CEM cell-line (or a derivative) by one group at the same time(Murai et al. 2018). We only examined SNS-seq data from 2018 onwards, when the lambda exonuclease-based digestion step had been incorporated into SNS-seq protocols, but wondered if there was a steady improvement of reproducibility as one progressed through the years. However, SNS-seq datasets do not become progressively more reproducible as one goes year-by-year from 2018-2022 (Supplementary Fig. S1e). The results suggest that there is significant difference in origins captured by different techniques that cannot be explained by differences in cell choices.

For the most popular technique (SNS-seq) the variability may be least when confined to a single cell line studied at the same time by the same group.

We next checked the genomic characteristics of origins defined by each technique. Regardless of technique used, around half of the detected origins fall in intergenic regions, followed by 30-40% allocated to introns (Fig. 1d). We recently reported 40,110 genomic regions are routinely bound by more than 3000 TF ChIP-seq data, named as TF binding hotspots (Hu et al. 2021). These TF binding hotspots likely represent highly active open chromatin regions. Despite the differences among techniques, the origins in general have lower overlap with TF binding hotspot (<0.5%), compared to the 16% overlap of gene promoter regions (TSS +/-100bp) with such TF binding hotspots (Fig. 1e). We further focused on one specific TF, CCCTC-binding factor (CTCF), which is important in organizing DNA structure and reported to be associated with replication (Su et al. 2020). Origins show significantly lower overlap with constitutive CTCF sites, defined as those that are conserved across cell types (Fang et al. 2020), compared to promoters (Fig. 1f). G-quadruplexes are found to be correlated with both replication and transcription (Lipps and Rhodes 2009). However, compared to promoters, origins defined by these techniques show significantly lower GC content (Fig. 1g). In addition, there is significantly lower overlap of these origins with G-quadruplex sites identified from predicted G-quadruplex motif regions (Bedrat et al. 2016), compared to the overlap of promoters with G-quadruplex sites (Fig. 1h). Thus the reported correlations between CTCF binding sites or G-quadruplexes with origins are not as striking as that of these features with promoters.

We found that all origins, regardless of technique by which identified, are slightly enriched at gene promoters, exons, or introns, compared with the genome background (Fig. 1d). Among the five different techniques, origins defined by Rerep-seq show the highest level of enrichment with promoters, TF hotspots, constitutive CTCF binding sites and G-quadruplexes and the highest

GC content. This is consistent with our knowledge that areas of the genome that are re-replicated when the cell-cycle is disturbed, are enriched in parts of the genome that replicate early in S phase, regions which are enriched in transcriptionally active (and thus epigenetically open) genes and their promoters.

Shared origins are associated with active chromatin and transcription regulatory elements

To address the potential biases caused by each technique, we investigated how many of the ~5 million union SNS-seq origins are reproducible in SNS-seq data and confirmed by other sequencing-based techniques. As we will describe below, these shared origins show significant overlap with promoters. Because the Re-rep-seq origins appear to be slightly different from the origins identified by other techniques, with highest overlap with transcriptionally active genes and promoters, we decided to exclude the Re-rep-seq from the analysis of shared origins and still reached the conclusion summarized above.

SNS-seq origins have the highest resolution. We used the following strategy to determine how many independent confirmations of an SNS-seq origin is sufficient for selecting an SNS-seq origin as a reproducible origin. The occupancy score of each origin defined by SNS-seq (Supplementary Fig. S2a) counts the frequency at which a given origin is detected in the datasets under consideration. Plotting the number of union SNS-seq origins with various occupancy scores with all SNS-seq datasets published after 2018, we sought to determine whether the curve deviates from the random background at a given occupancy score (Fig. 2a). For the random background, we assumed that the number of origins confirmed by increasing occupancy scores decreases exponentially (see Methods and Supplementary Table 2). The threshold occupancy score, to determine whether an origin is a reproducible origin, is the point where the observed number of origins deviates from the expected background number (with an

empirical FDR < 0.1) (Fig. 2a), an occupancy score of 20. Thus the reproducible SNS-seq origins (with a FDR<0.1) were those observed in at least 20 SNS-seq datasets.

We next determined which reproducible SNS-seq origins are confirmed by origins from Repli-seq and replication initiation zones (IZs) from Bubble-seq and OK-seq (Fig. 2b). 20,250 of the reproducible SNS-seq origins were found to overlap with an origin or IZ identified by each of the three other techniques and were called shared origins. These high-confidence shared origins consist of 0.27% of all ~7.5 million union origins. The coordinates of the shared origins are available in the supplementary files.

The shared origins have a greater overlap rate with gene promoters and exons compared to union origins (Fig. 2c). This is in line with the previous observation that replication and transcription are highly coordinated and enrichment of origins at TSS (Ganier et al. 2019; Cook 1999; Karnani et al. 2010). Shared origins also have a substantially higher overlap rate than union origins with TF binding hotspots (Fig. 2d). Moreover, shared origins have a higher overlap rate with constitutive CTCF binding sites, compared to union origins (Fig. 2e), and have a higher GC content and overlap with G-quadruplexes than union origins (Fig. 2f-g).

BART (Wang et al. 2018) analyzes the enrichment of TF binding sites (determined experimentally by ChIP-seq) with areas of interest in the genome. We used BART to perform TF binding site enrichment analysis on the shared origins and identified potential TFs or components of chromatin remodeling factors (e.g., PAF1, EZH2, ZNF282, POLR2A, GTF2I) whose binding sites are associated with the shared origins (Fig. 2h). The high enrichment of activators or repressors of transcription in the factors that have binding sites near shared origins provides more support that the shared origins have properties similar to transcriptional promoters.

To investigate whether the shared origins have a specific chromatin epigenomic signature compared to union origins, we used 5,711 publicly available ChIP-seq datasets for 29 different histone modifications and generated a comprehensive map of histone modification enrichment at shared origins compared to union origins. A substantial enrichment of activating histone marks, including H3K4me3 and H3/H4 acetylation, was observed at shared origins compared to union origins (Fig. 2i). The enrichment of H3K14ac is interesting given the enrichment in the BART analysis of binding sites of BRPF3, a protein involved in this specific acetylation and reported to stimulate DNA replication(Feng et al. 2016), but this modification was not uniquely enriched at the shared origins. These results show that as we move from all origins to a small set of high-confidence shared origins, we see a progressive increase in enrichment of TSS in epigenetically active parts of the genome.

However, transcriptional and epigenetic activators are not the whole story. H3K27me3, a repressive mark, is also enriched at shared origins, although this enrichment is not as high as that of the activating marks. The enrichment of EZH2 binding sites near shared origins is consistent with this observation because EZH2 is part of the Polycomb Repressive Complex 2 known to be a writer of H3K27me3. The paradoxical enrichment of repressive marks is consistent with the shared origins also being near binding sites of MBD (binds methylated DNA and represses transcription), PATZ1 and ZNF282, proteins known to be repressors of transcription.

Having identified a small group of highly reproducible origins (independent of cell-type and technique) we asked whether in a given cell line was one technique superior to the others in identifying such origins. K562 cells have been interrogated for origins by three different techniques: SNS-seq, OK-seq and Repli-seq (Supplementary Fig. S4). Among the 20,250 high-

confidence shared origins, 9,901 (48.9%) overlapped with SNS-seq origins in K562; 3,872 (19.1%) overlapped with OK-seq IZs in K562; 1,163 (5.7%) overlapped with Repli-seq origins in K562. This suggests that in one cell line, even the large initiation zones defined by OK-seq or Repli-seq do not capture most of the high-confidence origins. In the opposite direction, where we estimate what fraction of origins found by a given technique fall in a reproducible origin, the opposite result emerges: 2.7% of SNS-seq origins, 17.2% of OK-seq initiation zones and 7.7% of Repli-seq initiation zones overlapped with the 20,250 shared origins (Supplementary Fig. S4). Thus SNS-seq may be able to identify more of the reproducible origins, but it comes with a high false positive rate.

Human, but not yeast, high-confidence origins have low overlap with known ORC binding sites

Origin recognition complex (ORC) is expected to bind near replication origins during the cell cycle to help define origins (Bell 2002). To investigate the correlation of ORC binding with shared origins detected across different origin-mapping techniques, we analyzed all five publicly available ORC1 and ORC2 ChIP-seq datasets with at least 1,000 peaks (Supplementary Table 1c). We identified a total of 34,894 ORC binding sites in the human genome (Supplementary Table 3). Union (all) of ORC binding sites are enriched at promoters, TF hotspots, constitutive CTCF sites, GC content and G-quadruplexes compared to randomized genome background (Fig. 3a-e), somewhat similar to what we observed for shared origins (Figure 2).

Of the 34,894 ORC binding sites in the human genome 12,712 sites were defined as shared ORC binding sites as they occur in at least two ORC datasets (Occupancy score ≥2). For the majority of genomic features, including enrichment at promoter (Fig. 3a), overlap with constitutive CTCF sites (Fig. 3c), GC content (Fig. 3d) and overlap with G-quadruplex (Fig. 3e), ORC binding sites have similar genomic distribution regardless of how many samples they

appear in. Interestingly, the G-quadruplex enrichment at either union or shared ORC binding sites is a bit lower than in gene promoter regions (Fig. 3e). The only significant difference between the union ORC sites and the shared ORC binding sites is the significantly higher overlap of the latter with TF binding hotspots (Fig. 3b), suggesting that ORC-ChIP seq data also tend to enrich for highly open chromatin regions.

Based on the assumption that ORC binding sites that appeared in more than one dataset are likely to be true-positive ORC binding sites, we analyzed how many of the 12,712 shared ORC binding sites *overlap* with the 20,250 shared origins identified by our analysis above. The overlap was surprisingly low: 1.7% of the shared origins overlapped with shared ORC binding sites (Fig. 3f, Supplementary Fig. S5a). Even when we relaxed the criteria to look at shared origins that are proximal (≤1 kb) to shared ORC binding sites only 1300 (6.4%) of the shared origins were proximal to the shared ORC binding sites (Fig. 3f) (Supplementary Table 4b). The low degree of overlap or proximity of shared origins with shared ORC binding sites indicates that the vast majority of the shared origins are not near experimentally determined ORC binding sites.

In the reverse direction, we asked whether an ORC binding site definitively predicts the presence of an origin nearby. A histogram of the distance between any ORC binding site (union) and the nearest shared origin showed that only 1,086 (3.11%) ORC binding sites are proximal to (≤1 kb) a shared replication origin (Fig. 3g, Supplementary Table 4a). Given there are 34,894 ORC binding site and 20,250 shared origins, if ORC binding was sufficient to determine a high confidence origin, nearly half of the ORC binding sites should have been proximate to the shared origins. This low level of proximity between ORC binding and reproducible origins suggests that the current data on ORC binding sites is unable to predict the presence of a high confidence, actively fired origin.

We examined whether the co-localization is better if the analysis is done with data exclusively from the same human cell type (Supplementary Fig. S5b and S5d). Only 8.8% of the 105,881 union origins identified by SNS-seq, OK-seq, or Repli-seq in K562 cells overlap with ORC2 binding sites mapped in the same cell line, and only 4.9% of the 68,003 union SNS-seq origins mapped in HeLa cells overlap with ORC1 binding sites in the same cell line. The overlaps improve marginally if we focus on shared origins: 12.8% of 9,605 shared origins in K562 cells and 6.1% of 3,390 shared origins in HeLa cells overlap with ORC2 and ORC1 ChIP-seq sites in the concordant cell lines.

Experiments in the yeast S. cerevisiae have demonstrated that ORC binding is critical for defining an origin of replication. Indeed, in contrast to humans, in the more compact genome of the yeast, *S. cerevisiae*, there is higher overlap or proximity (within 1 kb) seen between the Yeast ORC ChIP-seq binding sites (Supplementary Table 1) and the well mapped yeast origins of replication in OriDB (Nieduszynski et al. 2007). In this case, 100% of the shared origins are proximate to (≤1 kb) all (union) or shared ORC binding sites in yeast (Fig. 3f).

Overall, we found that the shared origins are highly co-localized with ORC binding sites in yeast but not in the human cell lines, suggesting that the current methods of ORC binding site determination are failing to identify functional ORC binding.

Properties of the 6.4% shared origins co-localized reproduced ORC binding sites

Because of the poor co-localization of shared origins with ORC binding sites, we next focused on the 6.4% of shared origins (1,300) that are proximate to a shared ORC binding site, a group we will call the “highest confidence origins”. Compared with shared origins or ORC-binding sites alone, these 1,300 origins are even more co-localized with gene promoters (Fig. 4a), TF

binding hotspots (Fig. 4b) and constitutive CTCF binding sites (Fig. 4c). The GC content is not significantly higher for these “highest confidence origins” (Fig. 4d), and neither was their overlap with G-quadruplexes (Fig. 4e).

To understand the correlation between the different classes of origins and replication timing, we calculated the replication timing score for the origins following the ENCODE pipeline (Navarro Gonzalez et al. 2021; Hansen et al. 2010), and found that the mean replication timing score of shared origins suggest that they replicate earlier in S phase compared to the union origins, but there is not much difference between shared origins separated by whether they overlapped with ORC binding sites or not (Fig. 4f).

This suggests that even though shared origins near ORC binding sites are more enriched in TSS, TF binding hotspots and CTCF binding sites they are still similar to shared origins globally in their localization in early replicating, epigenetically active parts of the genome.

To investigate if the “highest confidence origins” (shared origins near ORC binding sites) are associated with higher transcriptional activity, we divided all human genes into four groups based on their expression level in the human K562 cell line and found the genes co-localized with the highest confidence origins exhibit higher expression levels compared to genes co-localized with all shared origins or union origins (Fig. 4g).

BART analysis (Wang et al. 2018) of which protein binding sites are most enriched in 743 origins overlapping with any ORC binding sites shows, as expected, that ORC2 binding sites are enriched near these origins. The other proteins bound near these origins are either transcriptional activators like EPC1, LEF1, ELK, EGR1, PAF1 or transcriptional repressors like KLF13, KRAB, ZNF639 or ZBTB33 (Fig. 4h). These results suggest that the shared origins

overlapping with ORC binding sites tend to be more associated with transcriptional regulation than all shared origins.

Overlap of MCM binding sites with shared origins to define another type of highest confidence origins

The six subunit minichromosome maintenance complex (MCM) is loaded on chromatin in G1 and forms the core of the active helicase that unwinds the DNA to initiate DNA replication (Madine et al. 2000). Since MCM2-7 may be loaded by ORC and move away from ORC to initiate DNA replication, it could be expected that even if the ORC binding sites are not proximate to the 20,250 shared origins, they will be proximate with known MCM binding sites. To test this, we analyzed 18 human MCM ChIP datasets (ENCODE Project Consortium 2012; Ivanov et al. 2018; Utani et al. 2017). We identified 11,394 total MCM3-7 binding sites (union) and 3,209 shared MCM binding sites that are defined by an intersection of MCM3, MCM5 and MCM7 union peaks. Overall MCM binding sites displayed very similar genomic features as ORC binding sites (Fig. 5a-e). We then defined the genomic features common among the shared origins that are close to (≤1kb) MCM binding sites. Like ORC-designated origins (Fig. 4b), the MCM-designated origins showed higher overlap with TF binding hotspots (Fig. 5f). Very interestingly, similar to the high overlap between yeast origins and yeast ORC sites (Fig. 3f), around 95% of yeast origins(Lee et al. 2021; Gros et al. 2015) are close to (≤1kb) the union of all yeast MCM sites. In contrast, only ~4.5% shared human origins are close to (≤1kb) the union of experimentally defined human MCM binding sites (Fig. 5g).

We examined all three types of high confidence origins (ORC designated, MCM3-7 designated and MCM2 designated) in a Venn Diagram (Fig. 5h) and identified 74 that were reproducibly identified by multiple methods (shared origins, near shared ORC binding sites, overlapping with MCM3-7 binding sites and MCM2 binding sites). The coordinates of these origins and their

supporting data (ORC and MCM binding sites) are listed in Supplementary Table 5, and genome browser views for three of them are shown in Fig. 6 to indicate the relationship between the coordinates.

Discussion

Through integrative analysis of publicly available 113 profiles of replication origins from multiple techniques, we identified ~7.5 million union origins in the human genome, of which only 20,250 shared origins are reproducibly identified by at least 20 SNS-seq datasets and confirmed by each of three other techniques. The following conclusions were reached: the shared (highly reproducible) origins are only a small subset of all origins identified and are in epigenetically active chromatin with a strong preference for promoters. However, G-quadruplexes, CTCF binding sites and ORC binding sites are also enriched near promoters, so that although these genomic features are enriched near the shared origins, it is not clear whether these genomic features actually specify origins, or whether they are simply enriched near origins because origins are near promoters. Finally, our results suggest that the co-localization of the shared origins with ORC or MCM2-7 binding sites is very low, much lower than in yeast, and so (a) the currently known binding sites of these proteins should not be used as surrogate markers of origins in human cells and (b) there is a great need to improve our methods of identifying isoforms of these proteins that are strictly functional for origin firing. The current models of origin firing require ORC to bind near origins, to load MCM2-7 nearby, and the latter to be converted to active helicase to fire origins. To rigorously test this model, we need examples of reproducible origins where these conditions are met, and our study helps identify 74 such sites where such experimental verification can be carried out.

Careful analysis shows that individual methods produce origin datasets that are more correlated with each other than across methods. The extensive variation of origins between datasets is

likely because of background noise in all current techniques, and because there is extreme stochasticity of origin selection during DNA replication in individual cells and cell cycles in the same population of cells from a human cell line in culture. There are also differences that are created due to differences in cell lineage. Despite this, we could cut through those differences to focus on the small subset of shared origins that seem to be active in multiple cell lineages.

The shared origins are enriched with active histone marks and enriched in early replicating parts of the genome that are overwhelmingly in an active epigenetic environment. H3K4me3 has been reported to be enriched at replication origins (Picard et al. 2014; Cayrou et al. 2015), and it has also been reported that demethylation of H3K4me3 by KDM5C/JARID1C promotes origin active (Rondinelli et al. 2015). H3ac/H4ac are also reported to be enriched at replication origins (Cadoret et al. 2008; Sequeira-Mendes et al. 2009) and this is regulated by various histone acetyltransferases (HATs) and histone deacetylases (HDACS) (Goren et al. 2008; Wang et al. 2009). Interestingly, H3K27me3 has also been reported to be enriched at replication origins (Picard et al. 2014; Cayrou et al. 2015). The higher enrichment of *all* activating histone modifications relative to the repressive histone modifications (including H3K27me3) (Fig. 2j) rather than individual types of modification, suggest that the shared, highly reproducible origins, are preferentially in epigenetically active euchromatin, this aligns well with the general enrichment of shared origins with TSSs and TF hotspots, which are concentrated in gene-dense, epigenetically open parts of the chromosomes. This also aligns well with the fact that shared origins were not enriched for H4K20me3 and H3K9me3. H4K20me3 has been reported to be enriched near late-firing, heterochromatic initiation zones (Kirstein et al. 2021) and H3K9me3 enriched near late replicating origins (Wang et al. 2017), while the shared origins are biased towards those that are in euchromatin. That origins, like Transcription start sites, prefer areas of the chromatin marked by activating epigenetic marks has been reviewed in 2016 (Prioleau and MacAlpine 2016). We arrive at the same conclusion even though we worked with origins

identified after this review, presumably with significant improvements in method and even though we focused on the most reproducible origins.

We characterized the genomic features of origins and found the shared origins are more co-localized with TSS. Table 1 shows the summary of the overlap of union origins, shared origins, shared origins overlapping ORC binding sites with various genomic features and compares this with the overlap of TSS with the same genomic features. Overall, these results suggest that as we proceed from all origins, to shared origins to highest confidence origins that are proximate to ORC binding sites, we see increasing overlap with promoters and early replicating, transcriptionally active, epigenetically open parts of the genome and that features thought to be important for origin selection (like G-quadruplexes, CTCF binding sites, active chromatin, ORC binding sites) could simply co-occur with origins because of their enrichment near promoters (TSS).

To rigorously test whether the overlaps of shared origins that we observe with various genomic features are significantly above background, we performed a permutation test that evaluates whether the observed overlap is significantly above the mean expected overlap when the experimental dataset is randomized 1,000 times (Table 2). The overlap of *shared origins* with promoters (TSS±4Kb), G-quadruplexes, R-loops, and CTCF binding sites are all significantly above background. However, the overlap of *promoters* with G-quadruplexes, R loops, and CTCF binding sites are also all significantly above background, though the enrichment of G-quadruplexes is minimal. Thus the genomic features characterized are all significantly enriched in both origins and promoters, and may help to specify both, either independently or co-dependently.

G-quadruplexes are of particular interest because they have been experimentally shown to dictate origin-specification (Prorok et al. 2019; Valton et al. 2014). Although there is a higher overlap of shared origins than union origins with G-quadruplexes (43.5% vs. 15%) this overlap is not higher than that between promoters and G-quadruplexes (48.1%) (Fig. 2g). This may suggest that the overlap between origins and G-quadruplexes may be secondary to the overlap between origins and promoters. However, the overlap of shared origins with G-quadruplexes is 3.5-fold above random while that of promoters with G quadruplexes is only 1.3-fold above random, which would be consistent with the idea that G-quadruplexes have a role in specifying origins and the overlap is not simply due to the proximity of origins with promoters.

Similarly, CTCF binding sites have been proposed to contribute to origin-specification (Emerson et al. 2022). Here again, the 2.2% of shared origins that overlap with constitutive CTCF binding sites is less than the 6.4% of promoters that overlap with CTCF binding sites (Fig. 2e). Here, though, the permutation test in Table 2 reveals that the enrichment of origins near CTCF binding sites (4.3-fold above random background) is comparable to the enrichment of promoters near the same sites (4.5-fold). This may, thus, suggest that the overlap of CTCF binding sites with origins is secondary to the overlap of origins with promoters.

Since ORC is an early factor for initiating DNA replication, shared human origins should be proximate to the reproducible ORC binding sites. The vast majority of ORC binding sites are not proximate to the shared origins and conversely, only about 6.4% of the shared origins are proximate (within 1 kb) to the reproducible (shared) ORC binding sites (Fig. 3f). Even with the most relaxed criteria nearly 85% of the most reproducible origins in human cells are not proximate to any ORC binding site (union, Fig. 3f). This low level of proximity contrasts with the ~100% of yeast origins that are within 1 kb of a yeast ORC binding site. Even when we examined cell line-specific origins with the ORC ChIP-seq datasets from the same cell lines, the

co-localization between origins and ORC binding sites remained poor (Supplementary Fig. S5). Another study also noted that only 13% of SNS-seq origins in K562 cells are near ORC2 binding sites (Miotto et al. 2016), but the authors suggested that this could occur if many SNS-seq sites are not real origins but arose from DNA breaks. A vast contamination of SNS-seq origins by sites of DNA breaks is discounted in our analysis because of the reproducibility of the shared origins in multiple labs, multiple lineages and multiple different techniques.

Instead of an unbiased determination of origins mapped by multiple groups to create a small set of reproducible origins, we could also empirically select a few well-curated origin datasets. Towards that end we used the Core origins identified by lambda exonuclease SNS-seq (Akerman et al. 2020) and the Ini-seq2 origins identified by labeling nuclei *in vitro* to identify the earliest replicated parts of the genome (Guilbaud et al. 2022). The shared origins were only a small subset of these origin datasets (Supplementary Fig. S8a and S8c). The percentage of these origins that were near ORC was still far less than what we observed in yeast: 13.7% for core origins and 30.4% for Ini-seq2 origins (Supplementary Fig. S8b and S8d). The higher overlap with Ini-seq2 origins could be because *in vitro* initiation of replication in isolated nuclei in Ini-seq preferentially identifies origins in early replicating, euchromatic parts of the genome, areas that have already been reported to be more enriched for ORC2 binding sites (Miotto et al. 2016).

There could be other explanations for the poor overlap of shared origins with ORC binding sites. We tested whether increasing the stringency of reproduction in SNS-seq data will produce more reproducible origins that are better co-localized with ORC binding sites. However, when we reduced the shared origins based on their reproduction by 30, 40, or 50 (out of 66) SNS-seq datasets, we did not see a marked improvement of their co-localization with currently known ORC binding sites (Supplementary Fig. S9a and S9b).

Finally, the permutation test in Table 2 demonstrates that the co-localization of origins with ORC binding sites was 6.5-fold ($p<0.001$) greater than random expectation but this was less than the co-localization of promoters with ORC binding sites, 10.6-fold greater than random ($p<0.001$). Thus, the co-localization data cannot conclusively say whether ORC binding near an origin is because ORC specifies origins, or whether this is secondary to origins being co-localized near promoters. Note, that the co-localization of ORC1 binding sites and origins in HeLa cells with highly active TSS has been noted in the past (Dellino et al. 2013). In human lymphoblast Raji cells ORC was enriched in promoters, and MCM depleted from gene bodies (Kirstein et al. 2021). ORC1 has also been reported to bind to RNA near promoters and stimulate replication from such sites (Mas et al. 2023). Thus our observation that ORC is highly enriched near promoters is expected, the main advance being that the random permutation test suggests that co-localization of TSS with ORC is more than the co-localization of shared origins with ORC. The question arises, are our results contradicting past studies. To the best of our knowledge, no one has taken such a comprehensive and quantitative approach to assess the proximity of origins with ORC binding sites genome-wide. We determined the percentage of origins that are within 1 kb of ORC-binding sites genome-wide and did a random permutation test to test whether the observed overlap is greater than random expectation. This told us that only a small percentage of the origins are within 1 kb of currently known ORC binding sites on a genome-wide scale, although selected browser shots as in Fig. 6 might suggest otherwise. Furthermore, even though Origins were enriched near ORC binding sites relative to random expectation this enrichment is not more than that of TSS near ORC binding sites.

Taken together, these results suggest three possibilities: (a) the ORC ChIP-seq or origin determination datasets in humans are very noisy and that the ChIP-seq data in particular fail to identify functional ORC binding sites; (b) that the MCM2-7 loaded at the ORC binding sites

move very far to initiate origins, farther than 1 kb from the ORC binding sites; or (c) that there are as yet unexplored mechanisms by which many of the most reproducible origins are specified, mechanisms that might include ORC-independent modes of origin specification, similar to the DnaA independent modes of origin specification in bacteria that use R-loops or DNA breaks to initiate replication (Itoh and Tomizawa 1980; Leela et al. 2021; Goswami and Gowrishankar 2022).

There is significant evidence for possibility (b) above. MCM2-7 is either loaded far from ORC binding sites or move significant distances after loading in Xenopus egg extracts (Harvey and Newport 2003). In yeasts there is evidence that MCM2-7 are pushed far away from the bodies of transcriptionally active genes by the RNA polymerase II and initiate replication at sites distant from where they are loaded on chromatin by ORC (Gros et al. 2015). However, it is worth noting that despite this nearly 100% of the origins in yeast are within 1 kb of ORC binding sites (Fig. 3f). Similar observations have been made in Drosophila, where cyclin E/cdk2 kinase activity promotes the loading of vast excess of MCM2-7 on chromatin relative to ORC, and the MCM2-7 complexes move away from their loading sites due to the activity of the transcriptional apparatus (Powell et al. 2015).

Since most models of replication initiation propose that a stably bound MCM2-7 complex is converted to an active CMG helicase at the time of origin firing, we hoped that even if ORC binding sites are not necessarily close to the shared origins, MCM2-7 binding sites will be more proximate to the shared origins in human cells. However, only 4.5% of shared origins are near any MCM2-7 binding site in human cells and again this is in contrast to the 96.5% of origins in yeast being near any MCM2-7 binding sites (Fig. 5g). Even if we focus on the limited data from the same cell line (K562 or HCT116) only 3.8% (union origins) or 5.5% shared origins overlap with union MCM3-7 binding sites in K562 cells and 19.5% of union origins overlap with any

MCM2 binding sites in HCT116 cells (Supplementary Figure S5b, S5c and S5e). As with ORC, we also did the analysis with two selected origin datasets, the core origins (Akerman et al. 2020) and the Ini-seq2 origins (Guilbaud et al. 2022) and found that still a small percentage of these origins (6.6% and 10.7%, respectively) were near any MCM2-7 binding sites. Finally, a recent paper reported the binding sites of phosphorylated isoforms of MCM2 (Thakur et al. 2022), and since phosphorylation of MCM2 is a prerequisite for MCM2-7 being activated as a helicase, we asked whether any of the phospho-MCM2 binding sites showed better co-localization with the shared origins (Supplementary Fig. S7). Phospho-S108 MCM2 showed the best result among the phospho-isoforms tested, with 22% of the shared origins being proximate to the Phospho-S108 MCM2 binding sites. Thus, phosphorylation on S108 of MCM2 may indeed mark MCM2-7 complexes that become active helicases which fire origins, but even then the co-localization was seen in a disappointingly low percentage of shared origins.

MCM2-7 ChIP-seq data is biased by a lot of noise as human MCM2-7 slides around after initial loading and due to contamination from the actively replicating CMG helicase that moves all over the genome (and pauses at many sites) during S phase, but this should increase the number of MCM2-7 binding sites and not decrease the percentage of origins that are near MCM2-7 sites. The better co-localization of shared origins with phospho-S108 MCM2 suggests that the ChIP methods can be improved to identify a small but critical pool of MCM2-7 that is engaged with the DNA stably in a way that permits origin specification. Alternatively, mapping of the sites bound to the active CMG helicase (MCM2-7, CDC45 and GINS), before the active helicase moves too far from the initiation site, may better enrich for sites that are destined to be origins.

A byproduct of our analysis of ORC ChIP-seq data, was the discovery that ORC1 and ORC2 are not strictly co-located as expected for the subunits of one ORC complex. Only 1,363 (21%) of the 6,501 ORC1 peaks overlapped with the 29,930 ORC2 peaks (Supplementary Fig. S6a),

and the vast majority (68%) of ORC1 peaks are far away (≥2kb) from the closest ORC2 peaks (Supplementary Fig. S6b). As a positive control, we checked other proteins that form well-established complexes like SMARCA4 and ARID1A in SWI/SNF complex, SMC1A and SMS63 in cohesin complex, EZH2 and SUZ12 in PRC2 complex, and found that they show a high proportion of overlap between their peaks as expected (Supplementary Fig. S6a). A possible explanation of the low overlapping rate between ORC1 and ORC2 peaks could be that the datasets are from different cell types (Supplementary Fig. S6c). However, even in ChIP-seq peaks from different cell types, we found continued high overlap between SMC1A and SMC3 peaks as would be expected for complexes that do not have high inter-cell-line variability in binding sites. In contrast, there was a low overlap between EZH2 and SUZ12 peaks taken from different cell-lines suggesting that for other complexes there is significant inter-cell-line variability in binding sites (Supplementary Fig. S6d). Thus, the lack of overlap between the ORC1 and ORC2 association sites on chromatin could be explained by inter-cell-line variation of binding sites for ORC (like the PRC2 complex) or could be explained by the ORC subunits binding to chromatin independent of one another. We prefer the latter explanation because of our experimental result that a substantial fraction of ORC2, 3, 4, 5 and 6 bind to chromatin in the absence of ORC1, and ORC2, 3 and 6 bind to chromatin normally in the absence of ORC5 (Park and Asano 2008; Shibata et al. 2016; Okano-Uchida et al. 2018; Shibata and Dutta 2020). We and others have also noted that human ORC2, 3, 4 and 5 form a stable subcomplex, with much looser association of the subcomplex with ORC1 and ORC6 (Dhar et al. 2001). Note that lineage-specific variation of where ORC may bind to the chromatin or the binding of ORC subcomplexes to chromatin may explain why so few of the ORC binding sites overlap with the shared origins but does not explain why only 6.4% of the shared origins (identified reproducibly in different cell lines) is proximate to any ORC binding sites.

Our analysis provides a characterization of origins in the human genome using multi-source data in the public domain. The sequencing-based profiles from different techniques have different biases in detecting DNA replication origins and show different genomic features. As one adds criteria to identify the most reproducible experimentally determined origins that are close to ORC binding sites, the overlap with TSS, TF-hotspots, constitutive CTCF binding sites and G- quadruplexes increases (Table 1), but in general the overlap with the latter three genomic features does not exceed that of the same features with TSS. Thus, although the overlaps are statistically significant, this correlation analysis cannot determine whether the same genomic features *independently* specify origins of replication and promoters of transcription. This is a question that should be asked experimentally and the 74 experimentally reproduced origin datasets with proximity to ORC and MCM2-7 ChIP-seq sites provides a starting list for such experimental queries. Finally, although our datasets for origins were large, as with all integrative data analyses we are limited by the quality of the data, which can be improved as experimental techniques for both origin identification and ORC or MCM2-7 binding site determination continue to improve.

Methods

Data processing

We collected all publicly available human sequencing datasets of replication origin profiling and ORC ChIP-seq from GEO database (Barrett et al. 2013). Data details of the collected datasets can be found in Supplementary Table 1. Raw sequence data in fastq format were downloaded and processed as follows: FastQC (Andrews S 2010) (v0.11.5) was used for quality control and sequence data were then mapped to human genome (hg38) using bowtie2 (Langdon 2015) (v2.2.9). Sam files were converted into bam files using samtools (Li et al. 2009) (v1.12) and only high-quality reads (q-score ≥30) were retained for subsequent analyses. Peak calling for ORC ChIP-seq data was performed with MACS2 callpeak function (--nomodel --extsize 150 -g hs -B -

-SPMR -q 0.05 --keep-dup 1) (Zhang et al. 2008) (v2.1.4) . Samples with more than 1000 peaks were kept as high-quality samples. SNS-seq, Bubble-seq, and Rerep-seq peak calling was performed with SICER (Zang et al. 2009) with different parameters based on the different resolution of technique: SNS-seq (W200 G600), Bubble-seq (W5000 G15000), Rerep-seq (W200 G600). The OK-seq defined IZs is generated from previous published papers (Petryk et al. 2016), (Wu et al. 2018) and the coordinates information is provided by Drs. Chunlong Chen and Olivier Hyrien. Repli-seq defined origins is generated from UW Encode group and is accessible in GSE34399. To obtain union peaks, peaks from each technique and from all techniques were merged using Bedtools (Quinlan and Hall 2010) merge (v2.29.2). Union origins that are longer than 300bp were cut into separate origins of a maximum length of 300bp. At the end we obtained 7,459,709 union origins.

To compare the locations of origins between samples, we performed a principal component analysis (PCA), where the matrix (to which PCA is applied) is of size 7.5M (number of union origins) x 113 (samples), with a "1" in the (*i*,*j*) position if the *i*th origin was detected in the *j*th dataset. We also performed pairwise pearson correlation of the origins in all the datasets to determine the reproducibility of the results from a given technique and across techniques. The distance metric for heatmap is generated in R using function cor (with parameters: use = “pairwise.complete.obs”, method = “pearson”).

Identification of shared origins

The origins detected by each sample are very different, which makes the identification of common origins difficult. SNS-seq origins have the highest resolution, and so we start with them. We identified 5,069,432 union SNS-seq origins using the same approach. We focus on the union SNS-seq origins that appear above a threshold number of SNS-seq origin datasets (high confidence SNS-seq origins) and then determine those that overlap with any initiation zone

identified by each of the three different techniques to delineate a set of “shared origins”, that initiate replication all the time and can be detected by multiple approaches. Note that this approach does not lose the high resolution of the SNS-seq origins, it merely finds those SNS-seq origins that are most reproduced and detected by each of the other 3 techniques.

To identify the most commonly shared origins between SNS-seq samples, we use an occupancy score for each union SNS-seq origin to show how many samples identify that specific origin. A higher occupancy score means the origin is more commonly present in many samples (Supplementary Fig. S2a). High confidence SNS-seq origins are those that occur in a sufficient number of SNS-seq samples at an FDR of >0.1. We used an exponential model for background (green line) and plotted the distribution of occupancy scores (blue dotted curve) for origins from all SNS-seq samples (Fig. 2a). The exponential model can be expressed as

$$y = A * e^{(K*(x+m))}$$

where x is the occupancy score, and y is the expected number of origins. Origins that are reproduced by multiple samples have higher occupancy scores than the background distribution. An empirical FDR of 0.1 was used to determine the cutoff of occupancy score so that the number of observed origins with occupancy score greater than the cutoff should be 10 times more than expected in the background model (Fang et al. 2020). The high-confidence SNS-seq origins were thus those with occupancy score ≥20 in all 66 SNS-seq samples. Detailed parameters can be found in Supplementary Table 2. Origins mapping to ChrM are removed from SNS-seq shared origins to avoid the interference of mitochondria DNA.

Finally, the high-confidence SNS-seq origins defined by our model that overlap with any IZs from OK-seq, any IZs from Bubble-seq, and any origins defined by Repli-seq are called “shared origins” (Fig. 2b). A total of 20,250 shared origins were identified.

K562 shared origins were identified as K562 SNS-seq origins that overlapped with any of the union OK-seq and Repli-seq peaks. HeLa shared origins were identified as SNS-seq origins present in all 3 SNS-seq samples in HeLa and overlapped with any of the union OK-seq and Repli-seq peaks.

Origins, ORC and MCM ChIP-seq data in yeast

Yeast origins data were obtained from OriDB (Nieduszynski et al. 2007). 829 replication origins mapped to the yeast genome sacCer1 were converted to the sacCer3 genome using UCSC LiftOver (Hinrichs et al. 2006). 289 experimentally confirmed origins, considered as high-confidence origins, were used for co-localization analysis with ORC and MCM binding sites in yeast. ORC and MCM ChIP-seq datasets were collected from the public domain (Supplementary Table 1) and processed using the same procedure as used for human data with reference genome version sacCer3. The shared MCM binding sites were defined as the MCM peaks that occur in all samples.

Enrichment of histone marks at shared origins

Histone modification ChIP-seq peak files were downloaded from CistromeDB (Zheng et al. 2019). A total of 5,711 peak files, each with over 5,000 peaks, covering 29 distinct histone modifications in different human cell types were used to interrogate the association between histone marks and shared origins, using union origins as control. The enrichment analysis was applied for each peak file by comparing the peak number in each histone modification peak file overlapping with shared origins versus the overlapping with union origins. The Odds ratio obtained from two-tailed Fisher's exact test was used as the enrichment score for each peak file (Fig. 2j). Odds ratio and pvalue are calculated using python package scipy. (stats.fisher_exact([[shared_ori_overlap_with_hm, shared_ori_not_overlap_with_hm],

[all_ori_overlap_with_hm - shared_ori_overlap_with_hm, all_ori_not_overlap_with_hm - shared_ori_not_overlap_with_hm]]))

Co-localization analysis

We use co-localization analysis to define shared origins and show the overlapping of CTCF peaks, TF hotspots, TSS regions, G quadraplex peaks, etc. with origins. The co-localization analysis was performed using Bedtools (Quinlan and Hall 2010) intersect -u. At least 1 base pair of intersection is required to be defined as overlapping.

Distance to the closest feature

According to the fact that ORC usually binds to origins but the loaded MCM2-7 can shift before firing an origin, we defined ORC that binds near the origins +/- 1kb as ORC near origins. Bedtools closest -d -t "first" was used to find the closest peak/region and distance for a given region. Origins with a binding site of ORC no further than 1kb are selected as origins near ORC binding site.

G4 binding sites

G4(G-quadraplex) predicted motif sites are obtained from a published G4 motif predictor: G4Hunter(Bedrat et al. 2016). The 1,444,095 G4 motif coordinates is provided by Laurent.

Genomic annotation of promoter, exon, intron, and intergenic region coverage

The coordinates of TSS, exon, intron are from UCSC hg38 version(Navarro Gonzalez et al. 2021). Promoter regions are defined as TSS+/-1kb. Intergenic regions are other genome regions excluding promoter, exon, and intron regions. Genomic annotation for peaks is defined by overlapping with those 4 types of regions by at least 1 bp.

Replication timing score

We used publicly available Repli-seq data in K562 cell line for different cell phase from ENCODE project(Navarro Gonzalez et al. 2021) to measure the replication timing of origins(Supplementary table 1). We used a previously defined weighted average score(Hansen et al. 2010) to combine signal from the six cell phases with the following formula: score=(0.917*G1b)+(0.750*S1)+(0.583*S2)+(0.417*S3)+(0.250*S4)+(0*G2). Higher values correspond to earlier replication.

Permutation Test

The R (version 4.1.3) package named regioneR (version 1.24.0) was used to statistically evaluate the associations between region sets with minor modification. The following parameters were used for running regioneR. Number of iterations: 10,000. Evaluation function: numOverlaps. Randomization function: randomizeRegions. randomizeRegions: hg38. non.overlapping: TRUE. mc.set.seed: FALSE

Acknowledgement. This work was supported by R01 CA60499 to AD and R35 GM133712 to CZ and K99 CA259526 to ZS. We thank all members of both the Zang and Dutta Labs for many helpful suggestions.

Figures and Tables

Figure 1: Total 7,459,709 origins defined by four types of techniques show different genomic features. (a) Data processing pipeline. 113 publicly available profiles of origins are processed following the pipeline. (b) Number of samples collected for each technique. In total 7,459,709 union origins were identified. (c) PCA shows the clustering of origin datasets from different techniques. (d) Genomic annotation (TSS, exon, intron and intergenic regions) of different groups of origins. Background is the percentage of each annotation on the whole genome. (e) Overlap with TF hotspots for different groups of origins and promoters. (f) Overlap with constitutive CTCF binding sites for different groups of origins and promoters. (g) GC content of different groups of origins and promoters. Grey line marks the average GC content of human genome. (h) G-quadruplex overlapping rates of different groups of origins and promoters.

Figure 2: The shared origins are enriched with certain transcription factors and active histone marks. (a) SNS-seq origins fitting distribution to an exponential model shows that an occupancy score $\geq$20 selected for reproducible SNS-seq origins. (b) Conceptual model of how shared origins is calculated. Any SNS-seq shared origin that overlaps with Bubble-seq IZ, OK-seq IZ and Repli-seq origin together is considered as an origin identified by all four techniques (shared origins). (c) Genomic annotation of union origins and shared origins. (d) Overlap with TF hotspots of union origins and shared origins. (e) Overlap with constitutive CTCF binding sites of union origins and shared origins. (f) GC content of union origins and shared origins. (g) G-quadruplex overlapping rates of union origins and shared origins. (h) BART prediction of TFs associated with shared origins. (i) Enrichment of histone marks at shared origins using all origins as control.

Figure 3: Genomic features of shared ORC binding sites and their co-localization with shared origins. (a) Genomic annotation of union ORC and shared ORC binding sites. (b) Overlap with TF hotspot of union ORC and shared ORC binding sites. (c) Overlap with constitutive CTCF binding sites of union ORC and shared ORC binding sites. (d) GC content of union ORC and shared ORC binding sites. (e) Overlap with G-quadruplex of union ORC and shared ORC binding sites. (f) The percentage of high-confidence origins (shared origins in human and confirmed origins in yeast) that overlapped with (left) or are proximate to (≤1kb) (right) to two types of ORC binding sites (union or shared). (g) Distribution of the distance between ORC binding sites and the nearest shared origin.

Figure 4: Shared origins near shared ORC binding sites are more correlated with active transcription. (a) Genomic annotation of shared origins and shared origins near (≤1kb) ORC binding sites. (b) Overlap with TF hotspots of shared origins and shared origins near ORC binding sites. (c) Overlap with constitutive CTCF binding sites of shared origins and shared origins near ORC binding sites. (d) GC content of shared origins and shared origins near ORC binding sites. (e) Overlap with G-quadruplex sites of shared origins and shared origins near ORC binding sites. (f) Y-axis: Replication timing score (from Navarro, 2021) for indicated classes of origins. (g) Annotation of expression level of genes that overlapped with different groups of origins. (h) BART prediction of TFs associated with highest confidence origins.

Figure 5: Genomic features of shared MCM binding sites and their co-localization with shared origins. (a) Genomic annotation of union MCM and shared MCM binding sites. (b) Overlap with TF hotspot of union MCM and shared MCM binding sites. (c) Overlap with constitutive CTCF binding rates of union MCM and shared MCM binding sites. (d) GC content of union MCM and shared MCM binding sites. (e) Overlap with G-quadruplex of union MCM and

shared MCM binding sites. (f) Overlap with TF hotspots of shared origins and shared origins near MCM binding sites. (g) The percentage of high-confidence origins (shared origins in human and confirmed origins in yeast) that overlapped with (left) or are proximate to (≤1kb) (right) to two types of MCM binding sites (union or shared). (h) Venn diagram of shared origins that are near ORC, MCM2 or MCM3-7 binding sites.

Figure 6: Genome browser screenshot for three of the 74 origins from Fig. 5h. The numbers below the SNS-seq shared origins track is the occupancy score of the origins along the length of indicated track.

Table 1: summary of data to show numbers of origins of different types, extent of overlap of each with different genomic features and comparison with Transcription Start Sites (TSS or promoters).

Table 2: Permutation test to ascertain the significance of the overlaps reported in this paper relative to random expectation.

Table 1 Overlap of origins, TSS and ORC binding sites with indicated features.

		% Overlap with				
	Number	Promoter	TF hotspot	CTCF	GC content	G quad
Union Origins	7,459,709	4.1	0.1	0.7	41.3333	15
Shared Origins	20,250	21.1	0.82	2.2	53	43.5
Shared Ori with shared ORC	1,300	46.7	6.54	4.5	54	45.6
TSS+/-100	26,237	100	15.9	6.4	61.5	48.1
Shared ORC	12,712	32.2	20.7	10	53.1	44.1

Table 2 Permutation test of overlap of shared origins or promoters (TSS) with region around promoters, Shared ORC peaks (in > 2 datasets), R-loops, G-quadruplexes and CTCF binding sites. Fold enrichment of observed overlap relative to the mean overlap seen with 1000 randomizations of set A is indicated together with the p-value of the enrichment.

Set_A	Set_B	#Set_A	#Set_B	Observed Overlap	# Random Iterations	Mean Random Overlap	Obs/Random Fold Enrichment	P-Value
origins	4k±promoters	20250	26237	4275	1000	649	6.6	< 0.001
origins	ORC_peaks	20250	12712	347	1000	54	6.4	< 0.001
origins	R-loop_zones	20250	59176	2607	1000	299	8.7	< 0.001
origins	G-quadruplex	20250	1444095	8818	1000	2529	3.5	< 0.001
origins	CTCF_sites	20250	22097	443	1000	104	4.3	<0.001
promoters	ORC_peaks	26237	12712	4832	1000	454	10.6	<0.001
promoters	R-loop_zones	26237	59176	10068	1000	1404	7.2	<0.001
promoters	G-quadruplex	26237	1444095	21809	1000	17242	1.3	<0.001
promoters	CTCF_sites	20250	22097	3644	1000	813	4.5	<0.001

<u>Supplementary data</u>

Supplementary Figures S1-S9: Figure legend on same page as figure.

Supplementary Table 1, collected public data. The original reference for each dataset can be found on the GEO page for each dataset.

a, Metadata of collected public Origin data.

b, Sample number of each cell type.

c, Metadata of collected public ORC ChIP-seq data.

d, Metadata of collected public MCM ChIP-seq data.

e, Metadata of collected public ORC and MCM ChIP-seq data in yeast.

Supplementary Table 2, parameters of model for SNS-seq

Parameters of exponential model for SNS-seq origins.

Supplementary Table 3, union and shared ORC binding sites

a, Union ORC ChIP-seq peaks, with coordinates and occupancy scores.

b, Shared ORC binding sites (defined as union ORC ChIP-seq peaks with an occupancy score >=2).

Supplementary Table 4, highest confidence origins

a, Coordinates of union ORC ChIP-seq binding sites with shared origins in 1kb region.

b, Coordinates of shared origins with shared ORC binding sites in 1kb region.

c, Coordinates of shared origins with union ORC binding sites in 1kb region.

Supplementary Table 5, 74 most confident origins

Coordinates of 74 origins that were reproducibly identified by multiple methods (shared origins, near shared ORC binding sites, overlapping with MCM3-7 binding sites and MCM2 binding sites).

References

Akerman I, Kasaai B, Bazarova A, Sang PB, Peiffer I, Artufel M, Derelle R, Smith G, Rodriguez-Martinez M, Romano M, et al. 2020. A predictable conserved DNA base composition signature defines human core DNA replication origins. *Nat Commun* **11**: 4826.

Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**: D991–D995.

Bedrat A, Lacroix L, Mergny J-L. 2016. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* **44**: 1746–1759.

Bell SP. 2002. The origin recognition complex: from simple origins to complex functions. *Genes Dev* **16**: 659–672.

Bell SP, Dutta A. 2002. DNA replication in eukaryotic cells. *Annu Rev Biochem* **71**: 333–374.

Bell SP, Stillman B. 1992. ATP-dependent recognition of eukaryotic origins of DNA replication by a multiprotein complex. *Nature* **357**: 128–134.

Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, Marin J-M, Lemaitre J-M. 2012. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat Struct Mol Biol* **19**: 837–844.

Cadoret J-C, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, Prioleau M-N. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proceedings of the National Academy of Sciences* **105**: 15837–15842.

Cayrou C, Ballester B, Peiffer I, Fenouil R, Coulombe P, Andrau J-C, van Helden J, Méchali M. 2015. The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res* **25**: 1873–1885.

Coleman TR, Carpenter PB, Dunphy WG. 1996. The Xenopus Cdc6 protein is essential for the initiation of a single round of DNA replication in cell-free extracts. *Cell* **87**: 53–63.

Cook PR. 1999. The Organization of Replication and Transcription. *Science*. https://www.science.org/doi/abs/10.1126/science.284.5421.1790 (Accessed February 4, 2022).

923 Costa A, Diffley JFX. 2022. The Initiation of Eukaryotic DNA Replication. *Annual Review of*
924 *Biochemistry* **91**: 107–131.

925 Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, Segalla S, Cesaroni M, Mendoza-Maldonado R,
926 Giacca M, Pelicci PG. 2013. Genome-wide mapping of human DNA-replication origins:
927 Levels of transcription at ORC1 sites regulate origin selection and replication timing.
928 *Genome Res* **23**: 1–11.

929 Dhar SK, Delmolino L, Dutta A. 2001. Architecture of the human origin recognition complex. *J*
930 *Biol Chem* **276**: 29067–29071.

931 Emerson DJ, Zhao PA, Cook AL, Barnett RJ, Klein KN, Saulebekova D, Ge C, Zhou L, Simandi Z,
932 Minsk MK, et al. 2022. Cohesin-mediated loop anchors confine the locations of human
933 replication origins. *Nature* **606**: 812–819.

934 ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human
935 genome. *Nature* **489**: 57–74.

936 Fang C, Wang Z, Han C, Safgren SL, Helmin KA, Adelman ER, Serafin V, Basso G, Eagen KP,
937 Gaspar-Maia A, et al. 2020. Cancer-specific CTCF binding facilitates oncogenic
938 transcriptional dysregulation. *Genome Biol* **21**: 247.

939 Feng Y, Vlassis A, Roques C, Lalonde M-E, González-Aguilera C, Lambert J-P, Lee S-B, Zhao X,
940 Alabert C, Johansen JV, et al. 2016. BRPF3-HBO1 regulates replication origin activation
941 and histone H3K14 acetylation. *The EMBO Journal* **35**: 176–192.

942 Foulk MS, Urban JM, Casella C, Gerbi SA. 2015. Characterizing and controlling intrinsic biases of
943 lambda exonuclease in nascent strand sequencing reveals phasing between
944 nucleosomes and G-quadruplex motifs around a subset of human replication origins.
945 *Genome Res* **25**: 725–735.

946 Ganier O, Prorok P, Akerman I, Méchali M. 2019. Metazoan DNA replication origins. *Curr Opin*
947 *Cell Biol* **58**: 134–141.

948 Goren A, Tabib A, Hecht M, Cedar H. 2008. DNA replication timing of the human beta-globin
949 domain is controlled by histone modification at the origin. *Genes Dev* **22**: 1319–1324.

950 Goswami S, Gowrishankar J. 2022. Role for DNA double strand end-resection activity of RecBCD
951 in control of aberrant chromosomal replication initiation in Escherichia coli. *Nucleic*
952 *Acids Res* **50**: 8643–8657.

953 Gros J, Kumar C, Lynch G, Yadav T, Whitehouse I, Remus D. 2015. Post-licensing Specification of
954 Eukaryotic Replication Origins by Facilitated Mcm2-7 Sliding along DNA. *Mol Cell* **60**:
955 797–807.

956 Guilbaud G, Murat P, Wilkes HS, Lerner LK, Sale JE, Krude T. 2022. Determination of human DNA
957 replication origin position and efficiency reveals principles of initiation zone organisation.
958 *Nucleic Acids Res* **50**: 7436–7450.

959 Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO,
960 Gartler SM, Stamatoyannopoulos JA. 2010. Sequencing newly replicated DNA reveals
961 widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**: 139–144.

962 Harvey KJ, Newport J. 2003. CpG methylation of DNA restricts prereplication complex assembly
963 in Xenopus egg extracts. *Mol Cell Biol* **23**: 6769–6779.

964 Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS,
965 Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic*
966 *Acids Research* **34**: D590–D598.

967 Hoshina S, Yura K, Teranishi H, Kiyasu N, Tominaga A, Kadoma H, Nakatsuka A, Kunichika T,
968 Obuse C, Waga S. 2013. Human origin recognition complex binds preferentially to G-
969 quadruplex-preferable RNA and single-stranded DNA. *J Biol Chem* **288**: 30161–30171.

970 Hu SS, Liu L, Li Q, Ma W, Guertin MJ, Meyer CA, Deng K, Zhang T, Zang C. 2021. Accurate
971 estimation of intrinsic biases for improved analysis of chromatin accessibility sequencing
972 data using SELMA. *BioRxiv*.

973 Hu Y, Stillman B. 2023. Origins of DNA replication in eukaryotes. *Molecular Cell* **83**: 352–372.

974 Itoh T, Tomizawa J. 1980. Formation of an RNA primer for initiation of replication of ColE1 DNA
975 by ribonuclease H. *Proc Natl Acad Sci U S A* **77**: 2450–2454.

976 Ivanov MP, Ladurner R, Poser I, Beveridge R, Rampler E, Hudecz O, Novatchkova M, Hériché J,
977 Wutz G, van der Lelij P, et al. 2018. The replicative helicase MCM recruits cohesin
978 acetyltransferase ESCO2 to mediate centromeric sister chromatid cohesion. *EMBO J* **37**:
979 e97150.

980 Karnani N, Taylor CM, Malhotra A, Dutta A. 2010. Genomic study of replication initiation in
981 human chromosomes reveals the influence of transcription regulation and chromatin
982 structure on origin selection. *Mol Biol Cell* **21**: 393–404.

983 Kirstein N, Buschle A, Wu X, Krebs S, Blum H, Kremmer E, Vorberg IM, Hammerschmidt W,
984 Lacroix L, Hyrien O, et al. 2021. Human ORC/MCM density is low in active genes and
985 correlates with replication time but does not delimit initiation zones. *eLife* **10**: e62161.

986 Langdon WB. 2015. Performance of genetic programming optimised Bowtie2 on genome
987 comparison and analytic testing (GCAT) benchmarks. *BioData Mining* **8**: 1.

988 Lee CSK, Cheung MF, Li J, Zhao Y, Lam WH, Ho V, Rohs R, Zhai Y, Leung D, Tye B-K. 2021.
989 Humanizing the yeast origin recognition complex. *Nat Commun* **12**: 33.

990 Leela JK, Raghunathan N, Gowrishankar J. 2021. Topoisomerase I Essentiality, DnaA-
991 Independent Chromosomal Replication, and Transcription-Replication Conflict in
992 Escherichia coli. *J Bacteriol* **203**: e0019521.

993 Leonard AC, Méchali M. 2013. DNA replication origins. *Cold Spring Harb Perspect Biol* **5**:
994 a010116.

995 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000
996 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format
997 and SAMtools. *Bioinformatics* **25**: 2078–2079.

998 Lipps HJ, Rhodes D. 2009. G-quadruplex structures: in vivo evidence and function. *Trends in Cell*
999 *Biology* **19**: 414–422.

1000 Madine MA, Swietlik M, Pelizon C, Romanowski P, Mills AD, Laskey RA. 2000. The roles of the
1001 MCM, ORC, and Cdc6 proteins in determining the replication competence of chromatin
1002 in quiescent cells. *J Struct Biol* **129**: 198–210.

1003 Mas AM, Goñi E, Ruiz de Los Mozos I, Arcas A, Statello L, González J, Blázquez L, Lee WTC, Gupta
1004 D, Sejas Á, et al. 2023. ORC1 binds to cis-transcribed RNAs for efficient activation of
1005 replication origins. *Nat Commun* **14**: 4447.

1006 Menzel J, Tatman P, Black JC. 2020. Isolation and analysis of rereplicated DNA by Rerep-Seq.
1007 *Nucleic Acids Research* **48**: e58.

1008 Mesner LD, Valsakumar V, Cieślik M, Pickin R, Hamlin JL, Bekiranov S. 2013. Bubble-seq analysis
1009 of the human genome reveals distinct chromatin-mediated mechanisms for regulating
1010 early- and late-firing origins. *Genome Res* **23**: 1774–1788.

1011 Miotto B, Ji Z, Struhl K. 2016. Selectivity of ORC binding sites and the relation to replication
1012 timing, fragile sites, and deletions in cancers. *Proceedings of the National Academy of*
1013 *Sciences* **113**: E4810–E4819.

1014 Murai J, Tang S-W, Leo E, Baechler SA, Redon CE, Zhang H, Al Abo M, Rajapakse VN, Nakamura E,
1015 Jenkins LMM, et al. 2018. SLFN11 Blocks Stressed Replication Forks Independently of
1016 ATR. *Mol Cell* **69**: 371-384.e6.

1017 Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC,
1018 Nassar LR, Maulding ND, Lee CM, et al. 2021. The UCSC Genome Browser database:
1019 2021 update. *Nucleic Acids Res* **49**: D1046–D1057.

1020 Nieduszynski CA, Hiraga S, Ak P, Benham CJ, Donaldson AD. 2007. OriDB: a DNA replication
1021 origin database. *Nucleic Acids Res* **35**: D40-46.

1022 Okano-Uchida T, Kent LN, Ouseph MM, McCarty B, Frank JJ, Kladney R, Cuitino MC, Thompson
1023 JC, Coppola V, Asano M, et al. 2018. Endoreduplication of the mouse genome in the
1024 absence of ORC1. *Genes Dev* **32**: 978–990.

1025 Park SY, Asano M. 2008. The origin recognition complex is dispensable for endoreplication in
1026 Drosophila. *Proc Natl Acad Sci U S A* **105**: 12343–12348.

1027 Petryk N, Kahli M, d'Aubenton-Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, Thermes C, Chen C-L,
1028 Hyrien O. 2016. Replication landscape of the human genome. *Nat Commun* **7**: 10208.

1029 Picard F, Cadoret J-C, Audit B, Arneodo A, Alberti A, Battail C, Duret L, Prioleau M-N. 2014. The
1030 spatiotemporal program of DNA replication is associated with specific combinations of
1031 chromatin marks in human cells. *PLoS Genet* **10**: e1004282.

1032 Powell SK, MacAlpine HK, Prinz JA, Li Y, Belsky JA, MacAlpine DM. 2015. Dynamic loading and
1033 redistribution of the Mcm2-7 helicase complex through the cell cycle. *EMBO J* **34**: 531–
1034 543.

1035 Prioleau M-N, MacAlpine DM. 2016. DNA replication origins—where do we begin? *Genes Dev*
1036 **30**: 1683–1697.

1037 Prorok P, Artufel M, Aze A, Coulombe P, Peiffer I, Lacroix L, Guédin A, Mergny J-L, Damaschke J,
1038 Schepers A, et al. 2019. Involvement of G-quadruplex regions in mammalian replication
1039 origin activity. *Nat Commun* **10**: 3274.

1040 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
1041 *Bioinformatics* **26**: 841–842.

1042 Remus D, Beuron F, Tolun G, Griffith JD, Morris EP, Diffley JFX. 2009. Concerted loading of
1043 Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139**:
1044 719–730.

1045 Rondinelli B, Schwerer H, Antonini E, Gaviraghi M, Lupi A, Frenquelli M, Cittaro D, Segalla S,
1046 Lemaitre J-M, Tonon G. 2015. H3K4me3 demethylation by the histone demethylase
1047 KDM5C/JARID1C promotes DNA replication origin firing. *Nucleic Acids Res* **43**: 2560–
1048 2574.

1049 Rowles A, Tada S, Blow JJ. 1999. Changes in association of the Xenopus origin recognition
1050 complex with chromatin on licensing of replication origins. *J Cell Sci* **112 (Pt 12)**: 2011–
1051 2018.

1052 Sequeira-Mendes J, Díaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gómez M. 2009.
1053 Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS*
1054 *Genet* **5**: e1000446.

1055 Shibata E, Dutta A. 2020. A human cancer cell line initiates DNA replication normally in the
1056 absence of ORC5 and ORC2 proteins. *J Biol Chem* **295**: 16949–16959.

1057 Shibata E, Kiran M, Shibata Y, Singh S, Kiran S, Dutta A. 2016. Two subunits of human ORC are
1058 dispensable for DNA replication and proliferation. *Elife* **5**: e19084.

1059 Su QP, Zhao ZW, Meng L, Ding M, Zhang W, Li Y, Liu M, Li R, Gao Y-Q, Xie XS, et al. 2020.
1060 Superresolution imaging reveals spatiotemporal propagation of human replication foci
1061 mediated by CTCF-organized chromatin structures. *Proc Natl Acad Sci U S A* **117**: 15036–
1062 15046.

1063 Sugimoto N, Maehara K, Yoshida K, Ohkawa Y, Fujita M. 2018. Genome-wide analysis of the
1064 spatiotemporal regulation of firing and dormant replication origins in human cells.
1065 *Nucleic Acids Research* **46**: 6683–6696.

1066 Thakur BL, Baris AM, Fu H, Redon CE, Pongor LS, Mosavarpour S, Gross JM, Jang S-M, Sebastian
1067 R, Utani K, et al. 2022. Convergence of SIRT1 and ATR signaling to modulate replication
1068 origin dormancy. *Nucleic Acids Res* **50**: 5111–5128.

1069 Utani K, Fu H, Jang S-M, Marks AB, Smith OK, Zhang Y, Redon CE, Shimizu N, Aladjem MI. 2017.
1070 Phosphorylated SIRT1 associates with replication origins to prevent excess replication
1071 initiation and preserve genomic stability. *Nucleic Acids Res* **45**: 7807–7824.

1072 Valton A-L, Hassan-Zadeh V, Lema I, Boggetto N, Alberti P, Saintomé C, Riou J-F, Prioleau M-N.
1073 2014. G4 motifs affect origin positioning and efficiency in two vertebrate replicators.
1074 *EMBO J* **33**: 732–746.

1075 Vashee S, Cvetic C, Lu W, Simancek P, Kelly TJ, Walter JC. 2003. Sequence-independent DNA
1076 binding and replication initiation by the human origin recognition complex. *Genes Dev*
1077 **17**: 1894–1908.

1078 Wang Y, Khan A, Marks AB, Smith OK, Giri S, Lin Y-C, Creager R, MacAlpine DM, Prasanth KV,
1079 Aladjem MI, et al. 2017. Temporal association of ORCA/LRWD1 to late-firing origins
1080 during G1 dictates heterochromatin replication and organization. *Nucleic Acids Res* **45**:
1081 2490–2502.

1082 Wang Z, Civelek M, Miller CL, Sheffield NC, Guertin MJ, Zang C. 2018. BART: a transcription
1083 factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics* **34**:
1084 2867–2869.

1085 Wang Z, Zang C, Cui K, Schones DE, Barski A, Peng W, Zhao K. 2009. Genome-wide mapping of
1086 HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138**: 1019–
1087 1031.

1088 Wu X, Kabalane H, Kahli M, Petryk N, Laperrousaz B, Jaszczyszyn Y, Drillon G, Nicolini F-E, Perot
1089 G, Robert A, et al. 2018. Developmental and cancer-associated plasticity of DNA

1090 replication preferentially targets GC-poor, lowly expressed and late-replicating regions.
1091 *Nucleic Acids Research* **46**: 10157–10172.

1092 Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. 2009. A clustering approach for
1093 identification of enriched domains from histone modification ChIP-Seq data.
1094 *Bioinformatics* **25**: 1952–1958.

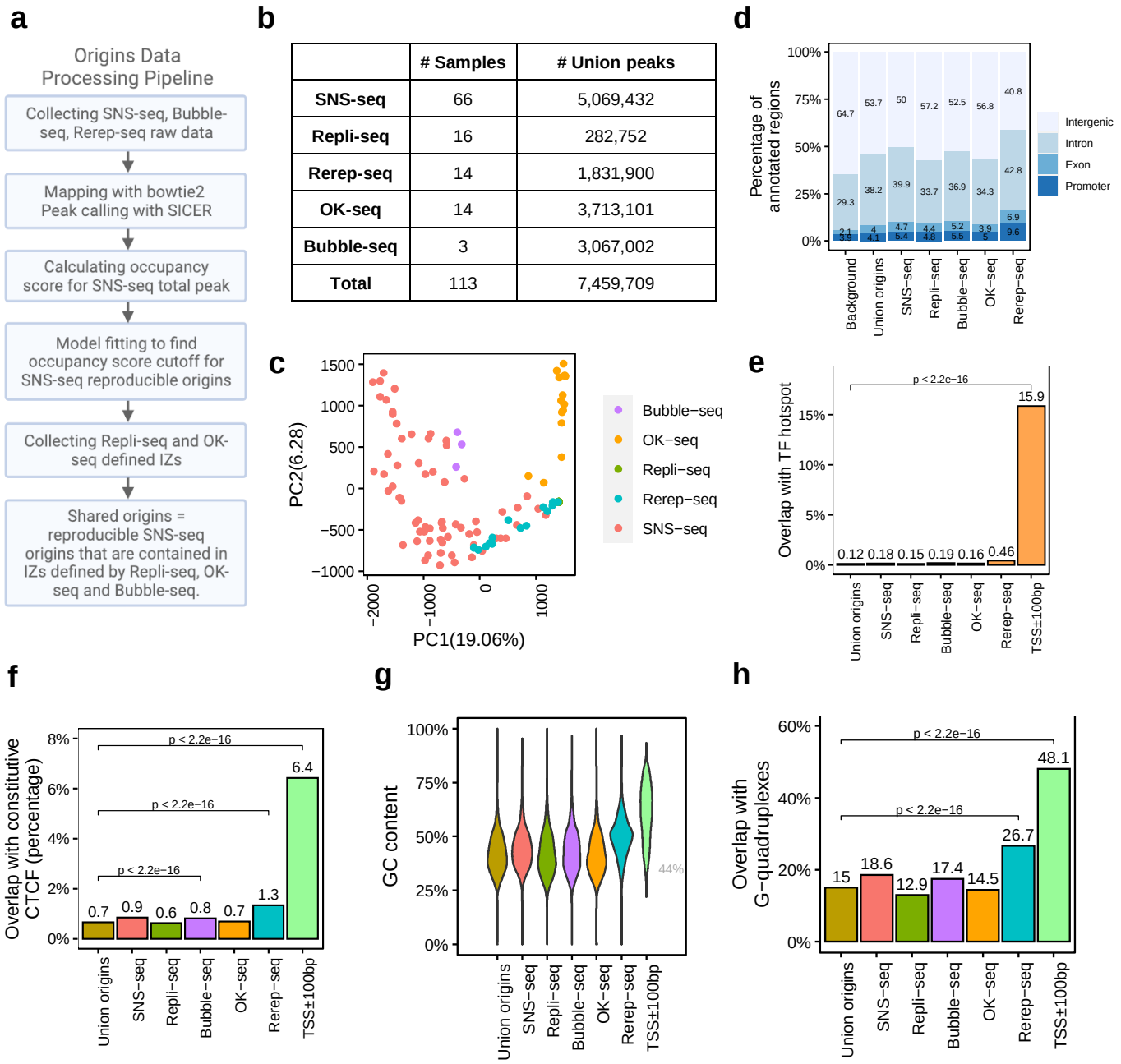
1095 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown
1096 M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.

1097 Zhao PA, Sasaki T, Gilbert DM. 2020. High-resolution Repli-Seq defines the temporal
1098 choreography of initiation, elongation and termination of replication in mammalian cells.
1099 *Genome Biol* **21**: 76.

1100 Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, Chen C-H, Brown M, Zhang X, Meyer CA, et al. 2019.
1101 Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis.
1102 *Nucleic Acids Research* **47**: D729–D735.

1103

Figure 1



	# Samples	# Union peaks
SNS-seq	66	5,069,432
Repli-seq	16	282,752
Rerep-seq	14	1,831,900
OK-seq	14	3,713,101
Bubble-seq	3	3,067,002
Total	113	7,459,709

Figure 1: Total 7,459,709 origins defined by four types of techniques show different genomic features. (a) Data processing pipeline. 113 publicly available profiles of origins are processed following the pipeline. **(b)** Number of samples collected for each technique. In total 7,459,709 union origins were identified. **(c)** PCA shows the clustering of origin datasets from different techniques. **(d)** Genomic annotation (TSS, exon, intron and intergenic regions) of different groups of origins. Background is the percentage of each annotation on whole genome. **(e)** Overlap with TF hotspots for different groups of origins. **(f)** Overlap with constitutive CTCF binding sites for different groups of origins. **(g)** GC content of different groups of origins. Grey line marks the average GC content of human genome. **(h)** G-quadruplex overlapping rates of different group of origins.

Figure 2

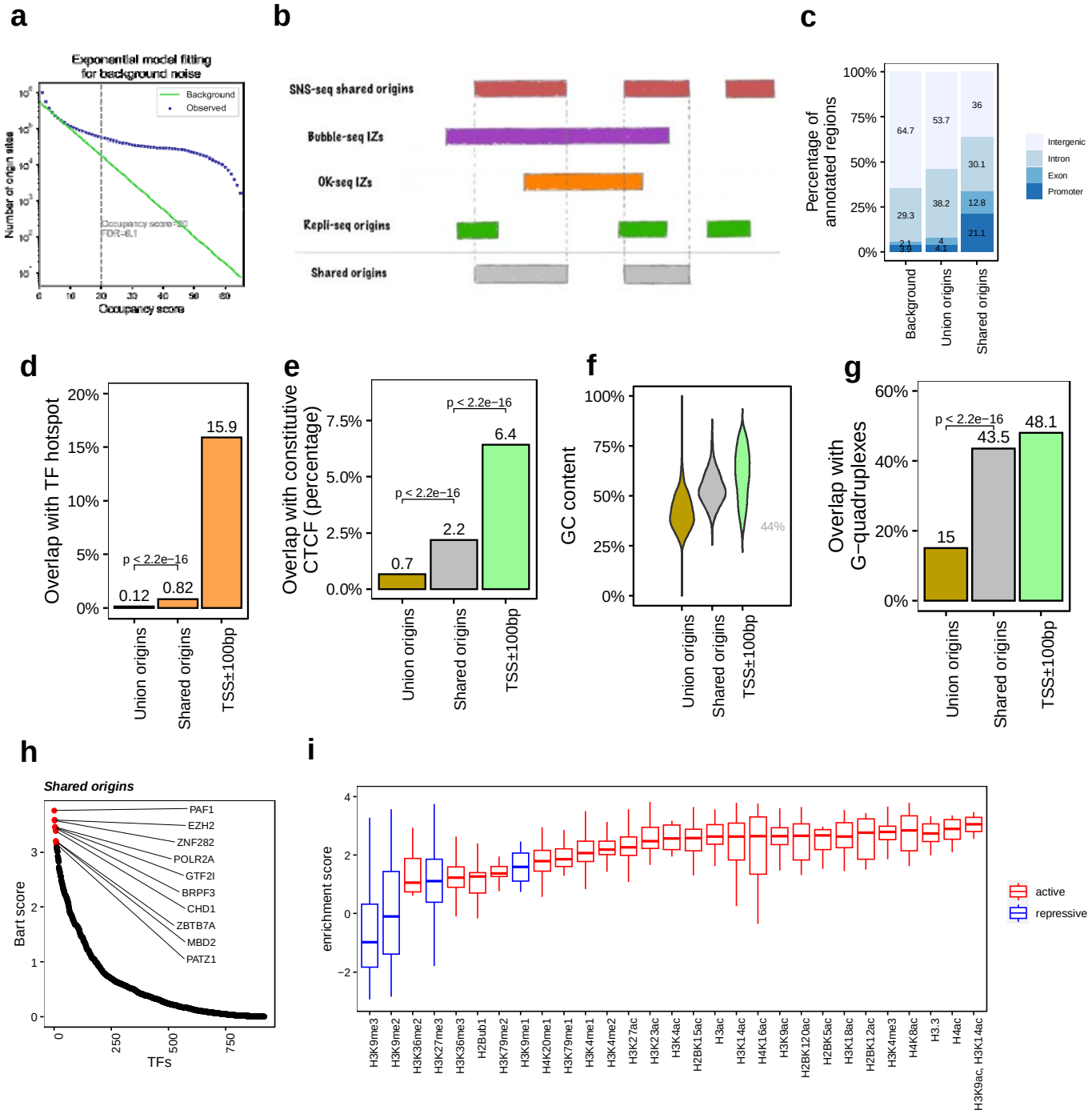


Figure 2: The shared origins are enriched with certain transcription factors and active histone marks. (a) SNS-seq origins fitting distribution to an exponential model. **(b)** Conceptual model of how shared origins is calculated. Every SNS-seq shared origin overlaps with any Bubble-seq IZ, OK-seq IZ and Repli-seq origin together is considered as origin shared by all techniques(shared origins). **(c)** Genomic annotation of union origins and shared origins. **(d)** Overlap with TF hotspots of union origins and shared origins. **(e)** Overlap with constitutive CTCF binding sites of union origins and shared origins. **(f)** GC content of union origins and shared origins. **(g)** G-quadruplex overlapping rates of union origins and shared origins. **(h)** BART prediction of TFs associated with shared origins. **(i)** Enrichment of histone marks at shared origins using all origins as control.

Figure 3

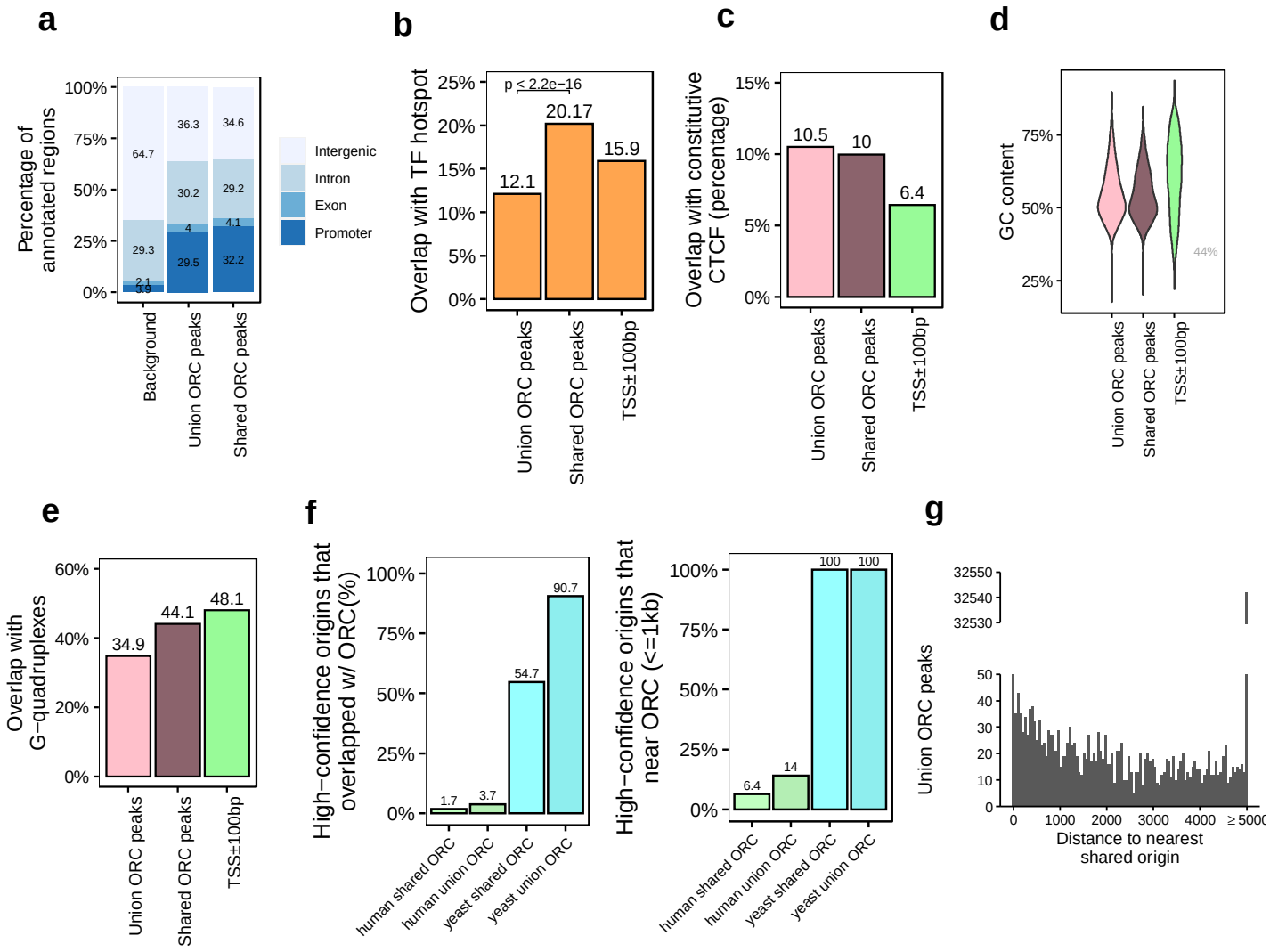


Figure 3: Genomic features of shared ORC binding sites. (a) Genomic annotation of union ORC and shared ORC binding sites. **(b)** Overlap with TF hotspot of union ORC and shared ORC binding sites. **(c)** Overlap with constitutive CTCF binding rates of union ORC and shared ORC binding sites. **(d)** GC content of union ORC and shared ORC binding sites. **(e)** Overlap with G-quadruplex of union ORC and shared ORC binding sites. **(f)** The percentage of high-confidence origins (shared origins in human and confirmed origins in yeast) that overlapped with or near (within1 kb) corresponding ORC binding sites. **(g)** Distribution of the distance between ORC binding sites and the nearest shared origin.

Figure 4

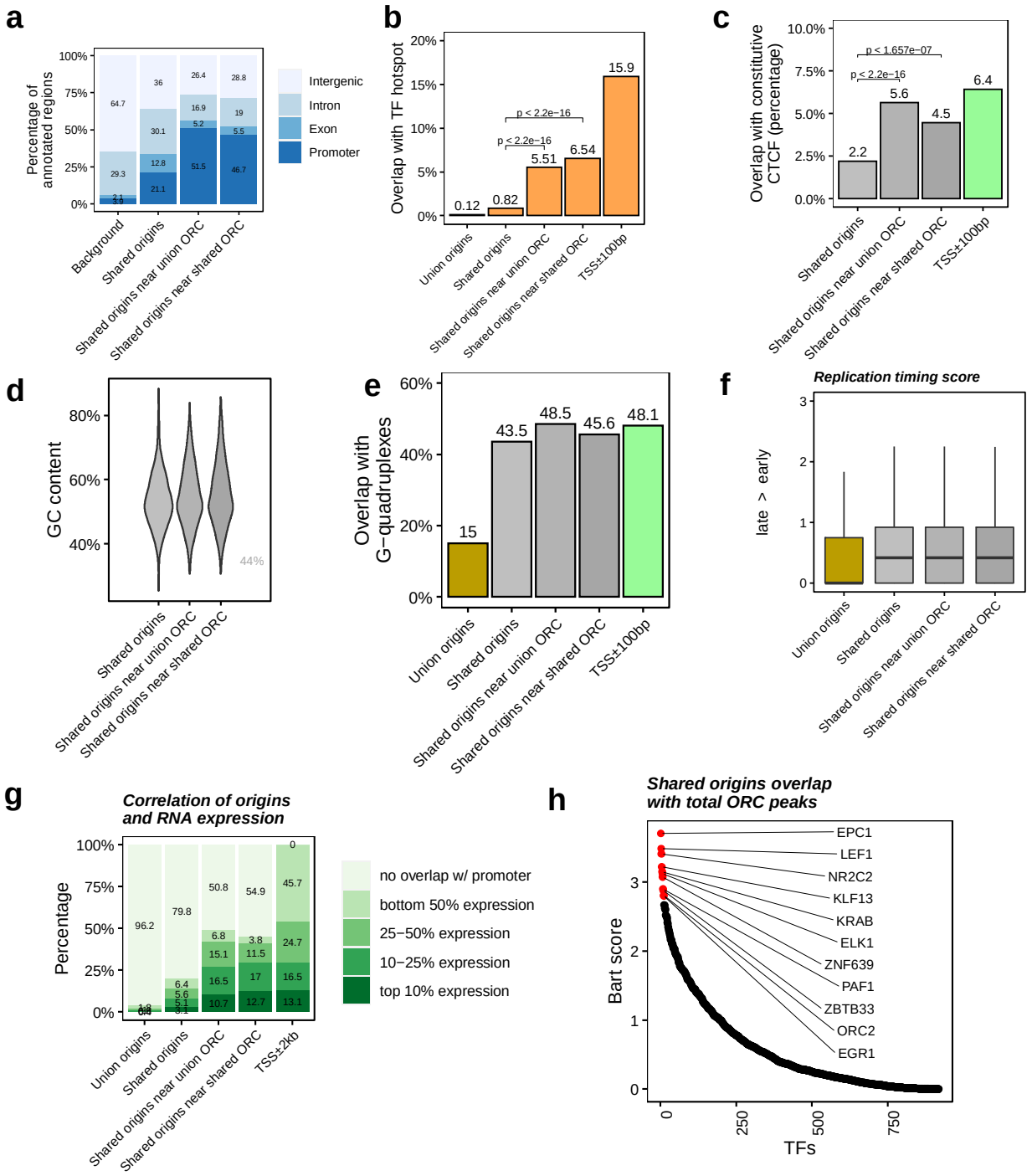


Figure 4: Shared ORC binding sites are more correlated with activate transcription. (a) Genomic annotation of shared origins and shared origins that near (less than 1kb) ORC binding sites. **(b)** Overlap with TF hotspots of shared origins and shared origins that near ORC binding sites. **(c)** Overlap with constitutive CTCF binding sites of shared origins and shared origins that near ORC binding sites. **(d)** GC content of shared origins and shared origins that near ORC binding sites. **(e)** Overlap with G-quadruplex sites of shared origins and shared origins that near ORC binding sites. **(f)** Y-axis: Replication timing score (from Navarro, 2021) for indicated classes of origins. **(g)** Annotation of expression level of genes that overlapped with different groups of origins. **(h)** BART prediction of TFs associated with highest confidence origins.

Figure 5

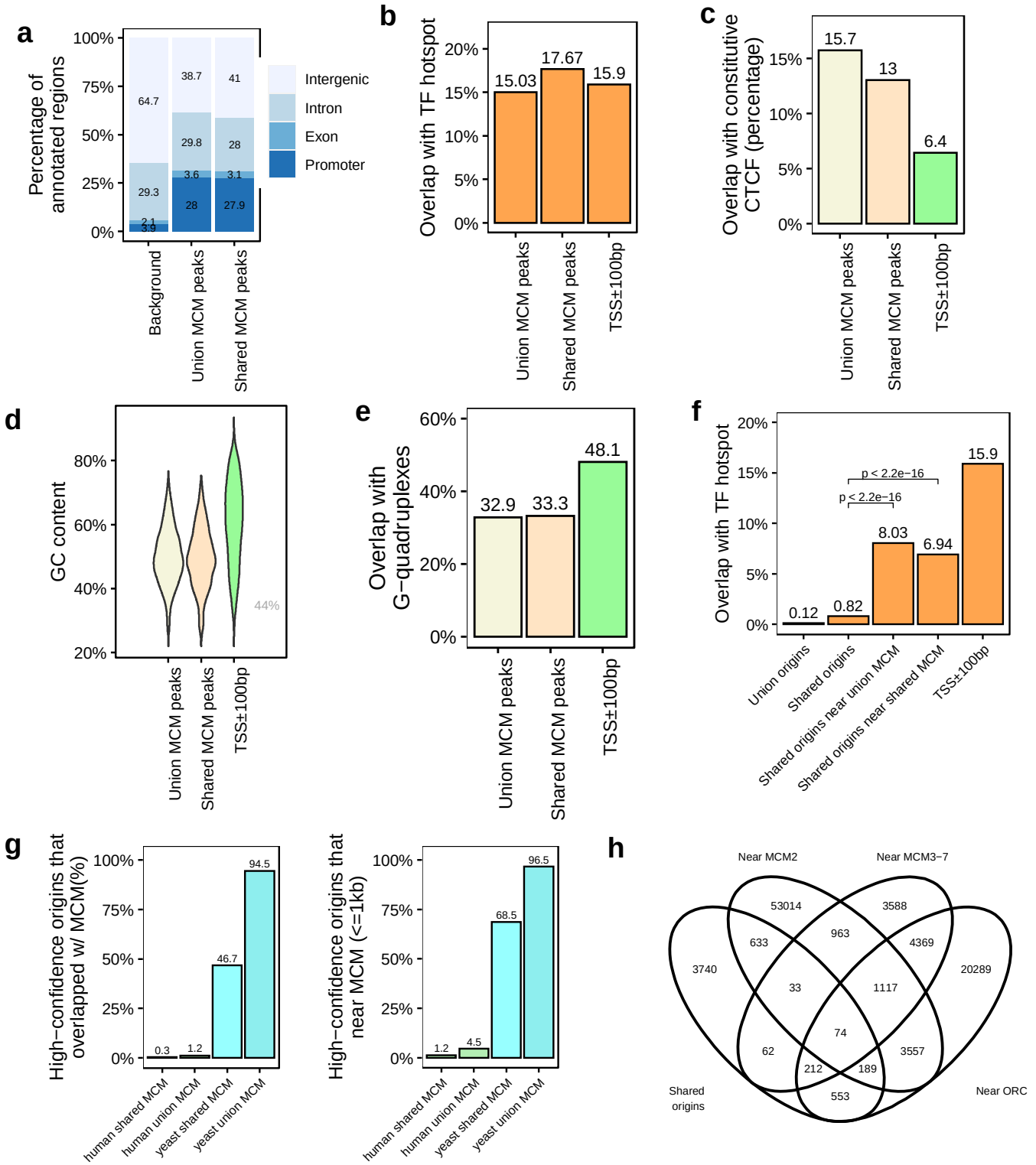


Figure 5: Genomic features of shared MCM binding sites. (a) Genomic annotation of union MCM and shared MCM binding sites. **(b)** Overlap with TF hotspot of union MCM and shared MCM binding sites. **(c)** Overlap with constitutive CTCF binding rates of union MCM and shared MCM binding sites. **(d)** GC content of union MCM and shared MCM binding sites. **(e)** Overlap with G-quadruplex of union MCM and shared MCM binding sites. **(f)** Overlap with TF hotspots of shared origins and shared origins that near MCM binding sites. **(g)** The percentage of high-confidence origins (shared origins in human and confirmed origins in yeast) that overlapped with / near(region boundary less than 1kb) corresponding MCM binding sites. **(h)** Several confident origins defined by different methods. Number is calculated from total peak from each methods overlapping with total peak of all methods.

Figure 6

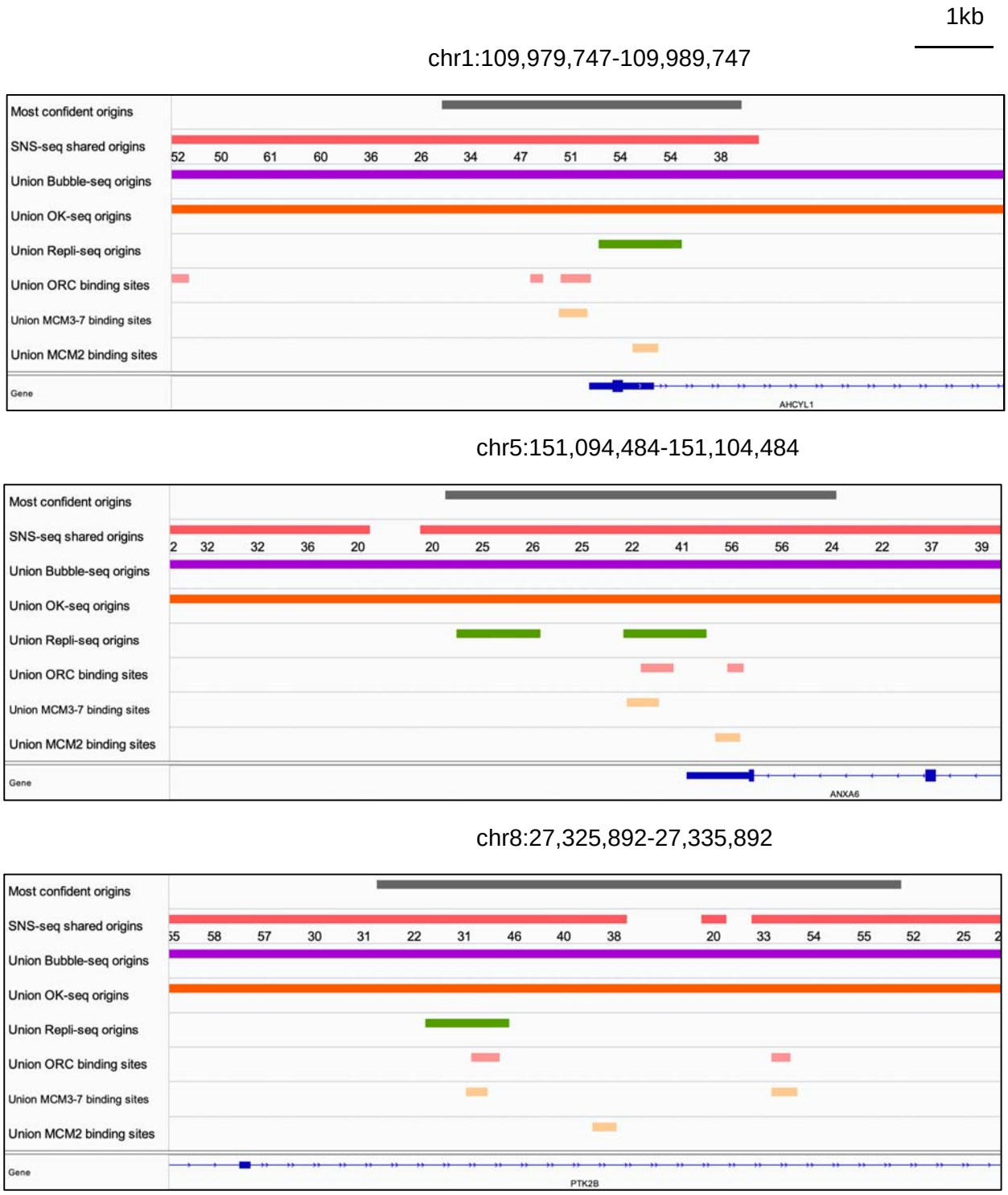


Figure 6: Genome browser screenshot for three of the 74 origins from Fig. 5h
The number on SNS-seq shared origins track is occupancy score of the origin.