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

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 at Birmingham, 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

DNA replication initiates from ~50,000 origins on human chromosomes in each cell-cycle and the origins are hypothesized 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 site datasets to critically evaluate whether the most reproducible origins are specified by these features. Out of ~7.5 million 300 bp chromosomal fragments reported to harbor origins by all the datasets, only 0.27% were reproducibly detected by four techniques (20,250 shared origins), suggesting extensive variability in origin usage and identification in different circumstances. 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, in contrast to the nearly 100% overlap between the two 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 ORC-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). 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). Notably any technical biases are uniquely associated with each assay, suggesting a false positive origin is unlikely to be captured across different techniques. 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, determine the most reproducible and consistent origins identified by different methods from different research groups and in different cell types and evaluate the hypotheses regarding origin specification at least on these most reproducible origins.

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 Stillman 1992)(Bell and Dutta 2002; Remus et al. 2009), (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 at least 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 profiles. 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 split the zones into 300 base fragments and identified SNS-seq defined origins that overlap with origins identified by the various methods. We thus identified ~7.5 million 300 base fragments that harbor origins in total, of which 20,250 high-confidence origins are shared across four techniques. Using these high-confidence shared origins, which overlap significantly with transcription promoters, we could test whether G-quadruplex sites, ORC binding sites, G-quadruplexes or CTCF binding sites help specify origins and compare the enrichment of these features on origins versus promoters.

Results

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

To compare the replication origin profiles generated by various methods, we collected 113 publicly available replication origin profiling datasets in different human cell types from five different techniques (Supplementary Figure 1a), 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 analyzed all the datasets using pipeline considering the difference of resolution of each technique to avoid data processing biases (Fig. 1a). All the 113 datasets contain at least 1000 origins (Supplementary Fig. 1b) with origin lengths being significantly longer for Bubble-seq and OK-seq, methods known to identify initiation zones (Supplementary Fig. 1b). A total of ~7,460,000 union origins were discovered from all techniques (Fig. 1b).

Principal Component Analysis (PCA) of all origin profiles shows that profiles from the same technique are more similar to each other than to different techniques, although the most popular

technique, SNS-seq, shows a very large variability in the PCA (Fig. 1c). However, we did not observe clear trends of clustering of origins identified in similar cell types (Supplementary Fig. 1c), or separated by peak count or year of study (Supplementary Fig. 1d). The results suggest that there is significant difference in origins captured by different techniques that cannot be explained by differences in cell choices or numbers of peaks identified and that the most popular technique (SNS-seq) has some internal variability in origins identified that cannot be explained by refinement of technique over the years.

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 different 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 ~7.5 million union origins are shared among four sequencing-based techniques. As we will describe below, these shared origins show significant overlap with promoters. Because the re-rep origins appear to be slightly different from the origins identified by other techniques, with highest overlap with transcriptionally active genes and promoters, we excluded the origins identified by this technique from this analysis 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 origin as a reproducible origin. The occupancy score of each origin defined by SNS-seq (Supplementary Fig. 2a) counts the frequency at which a given origin is detected in the datasets under consideration. Plotting the number of origins with various occupancy scores when all SNS-seq datasets published after 2018 are considered together (the union origins) shows that 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). This analysis was done to identify the origins that are reproducibly called (exceed the threshold occupancy score) by multiple datasets in SNS-seq.

Once these reproducible origins are identified, we determined which origins identified by SNS-seq (the technique with the highest sequencing resolution) are confirmed by origins from Repli-seq and replication initiation zones (IZs) from Bubble-seq and OK-seq (Fig. 2b). Because the initiation zones can be very large, we divided the zones into 300 bp segments for calculation of overlap. These 20,250 high-confidence shared origins consist of 0.27% of all ~7.5 million union origins. The co-ordinates of the origins identified by each of the techniques and 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 (Cook 1999; Karnani et al. 2010)(Ganier et al. 2019). Shared origins also have a substantially higher overlap rate than union origins with TF binding hotspots (Fig. 2d), suggesting that they are more likely to interact with transcription factors. 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). These results show that the shared origins are not a random selection from all origins but represent a set of consensus or high-confidence origins that are less likely to be biased by different techniques and are enriched in epigenetically active chromatin.

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.

Human, but not yeast, high-confidence origins have low overlap with 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 1000 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 four independent origin-determination techniques. The overlap was surprisingly low: 1.7% of the shared origins overlapped with shared ORC binding sites (Fig. 3g, Supplementary Fig. 3a). 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 overlapped with the shared ORC binding sites (Fig. 3g) (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 a 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 also showed that only 1,086 (3.11%) ORC binding sites are proximate to (≤1 kb) a shared replication origins (Fig. 3f, 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 ORC binding is not sufficient to predict the presence of a high confidence, actively fired origin.

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. 3g).

Overall, we found that the shared origins are more co-localized with ORC binding sites in yeast compared to human cell lines.

Properties of shared origins with reproduced ORC binding sites

We next determined the genomic features of the 1300 "highest confidence" origins where a shared origin (defined by four different techniques) is proximate to a shared ORC binding site. Compared with shared origins or ORC-binding sites alone, the 1300 "highest confidence" origins that included both 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 the “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 shared origins including the highest confidence origins are more correlated with earlier replication activities compared to the union origins, but there is not much difference between origins separated by whether they overlapped with ORC binding sites or not (Fig. 4f). This suggests that the most reproducible origins (shared origins) are 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.

Overall, these results suggest that the most reproducible origins (including the highest confidence origins that are proximate to ORC binding sites) overlap with early replicating, transcriptionally active, epigenetically open parts of the genome.

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 is expected that the 20,250 shared replication origins will be proximate with known MCM binding sites. To this end, 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. 3g), 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 commonly identified by four techniques. This low yield of highly reproducible origins may reflect variations in the techniques, variations among cell lineages, noise in the methods used or the general stochastic nature of origins selection in different cells in different cell-cycles. We find that the variation persists even if we focus on one cell lineage (Supplementary Fig. 1c). The extensive variation of origins between datasets is likely to be both both because the background noise correction has to be improved for all current techniques,

and because there is extreme stochasticity of origin selection during DNA replication in human cell lines.

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. The shared origins have higher GC content than union origins, but this is not as high as that noted for TSS. G-quadruplexes have been suggested as factors that dictate origin-specification. Although there is a higher overlap of shared origins than union origins with G-quadruplexes (43.5% vs. 15%) this overlap is comparable to that between shared ORC sites and G-quadruplexes (44.1%) and neither is higher than that between TSS and G-quadruplexes (48.6%). This may suggest that the overlap between origins and G-quadruplexes may be secondary to the overlap between origins and TSS.

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 TSS that overlap with CTCF binding sites. This may also be construed to suggest that the overlap of CTCF binding sites with origins is secondary to the overlap of origins with TSS.

Since ORC is an early factor for initiating DNA replication, we expected that shared human origins will 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. 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. 3g). This low level of proximity contrasts with the ~100% of yeast origins that are within 1 kb of a yeast ORC binding site. Taken together, these results suggest either that the ChIP-seq and origin determination datasets in humans are much noisier than in yeast, or that the MCM2-7 loaded at the ORC binding sites move much further away to initiate origins far from the ORC binding sites, or that there are as yet unexplored mechanisms of origin specification in human cancer cells.

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 median expected overlap when the experimental dataset is randomized 10,000 times (Table 2). The overlap of *shared origins* with promoters (TSS±4Kb), shared ORC binding sites, G-quadruplexes and R-loops are all significantly above background. However, the overlap of *promoters* with shared ORC binding sites, G-quadruplexes and R loops 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.

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 MCM2-7 binding sites will be more proximate to the reproducible origins in human cells. However, taking all the datasets into consideration, only 4.5% of the shared origins are proximate to any (union) MCM2-7 binding sites (Fig. 5g). Even if we focus on the limited data from the same cell line (HCT116), at most <25% of all origins are proximate to any of the MCM2 binding sites (Supplementary Figure 3c). Again, the contrast with yeast is striking where ~95% of the experimentally defined origins are near any MCM2-7 binding sites. One obvious caveat is that the MCM2-7 ChIP-seq

data may be 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. We suggest that the ChIP methods in human cells fail to define a small but critical pool of MCM2-7 that is engaged with the DNA stably in a way that permits origin specification.

Interestingly, through the analysis of ORC ChIP-seq data, we found 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 were overlapped with the 29,930 ORC2 peaks (Supplementary Fig. 4a), and the vast majority (68%) of ORC1 peaks are far away (≥2kb) from the closest ORC2 peaks (Supplementary Fig. 4b). As a positive control, we checked other proteins that form well-established complexes like SMARCA4 and ARID1A in SWI/SNF complex, SMC1A and SMC3 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. 4a). 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. 4c). Using 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. 4d). 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). Such lineage-specific variation of where ORC may bind to the 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.

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 activation (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). 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 epigenetically active chromatin favors origins specification in general. 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.

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 start 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 paper (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).

Identification of shared origins

The origins detected by each sample are very different, which makes the identification of common origins difficult. So, we focus on the origins that appear in most samples, this group of origins initiate replication all the time and can be detected by all approaches, they are more likely to have important roles. SNS-seq origins have the highest resolution, and so we start with them. To identify the most commonly shared origins between SNS-seq samples, we use occupancy score for each union peak to show how many samples are occupied by a specific peak. Higher occupancy score means the peak are more commonly shared by many datasets (Supplementary Fig. 2a). We consider the origins detected by more samples as confidence origins and define the shared origins as those occur in sufficient number of samples from all four techniques and each technique. We used an exponential model (blue curve in corresponding figure) to fit the distribution of occupancy scores (blue dotted curve in corresponding figure) for origins merged from all SNS-seq samples (Fig. 2a). Origins shared by each group of datasets have higher occupancy scores than the background distribution. 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. An empirical FDR of 0.1 was used to determine the cutoff of occupancy score (say x') so that the number of observed origins with occupancy score greater than x' should be 10 times more than expected in the background model (Fang et al. 2020). The high-confidence shared origins were finally determined as 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.

The SNS-seq shared 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 is shared origins (Fig. 2b).

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 lease 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.

Tables

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).

		%	%	%	%	%
	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 to ascertain the significance of the overlaps reported in this paper relative to random expectation.

Region_Set_A	Region_Set_B	#Region_Set_A	#Region_Set_B	Observed Overlap	# Random Iterations	Random Overlap	Observed/ Random Fold_Enrichment	P-Value
shared_origins	ORC_Peaks shared>2	20250	12712	349	10000	54	6.5	< 0.0001
shared_origins	R-loop_zones	20250	59176	2945	10000	325	9.1	< 0.0001
shared_origins	4k±promoters	20250	26237	4960	10000	738	6.7	< 0.0001
shared_origins	G-quadruplexes	20250	1444095	13300	10000	3139	4.2	< 0.0001
promoters	ORC_Peaks shared>2	26237	12712	4832	10000	454	10.6	<0.001
promoters	R-loop_zones	26237	59176	10068	10000	1402	7.2	<0.001
promoters	G-quadruplexes	26237	1444095	21809	10000	17220	1.3	<0.001

Figures

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. (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) Distribution of the distance between ORC binding sites and the nearest shared origin. (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 ORC binding sites (union or shared).

Figure 4: 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) Replication timing score of shared origins and shared origins near ORC binding sites. (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. (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.

Supplementary data

Supplementary Figures S1-S4: 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.

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).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Wu X, Kabalane H, Kahli M, Petryk N, Laperrousaz B, Jaszczyszyn Y, Drillon G, Nicolini F-E, Perot G, Robert A, et al. 2018. Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions. *Nucleic Acids Research* **46**: 10157–10172.

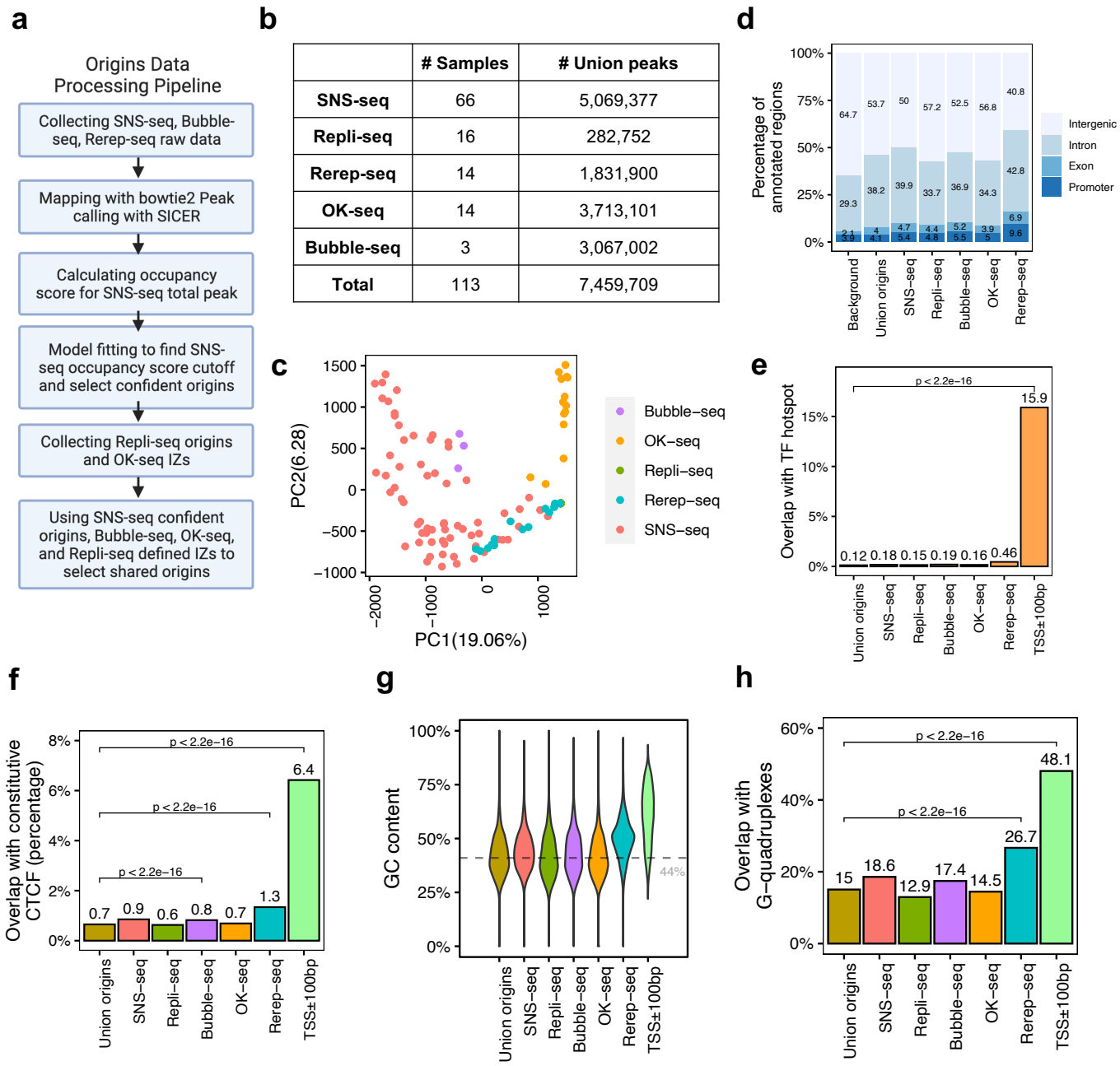
Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**: 1952–1958.

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

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

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

Figure 1



	# Samples	# Union peaks
SNS-seq	66	5,069,377
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

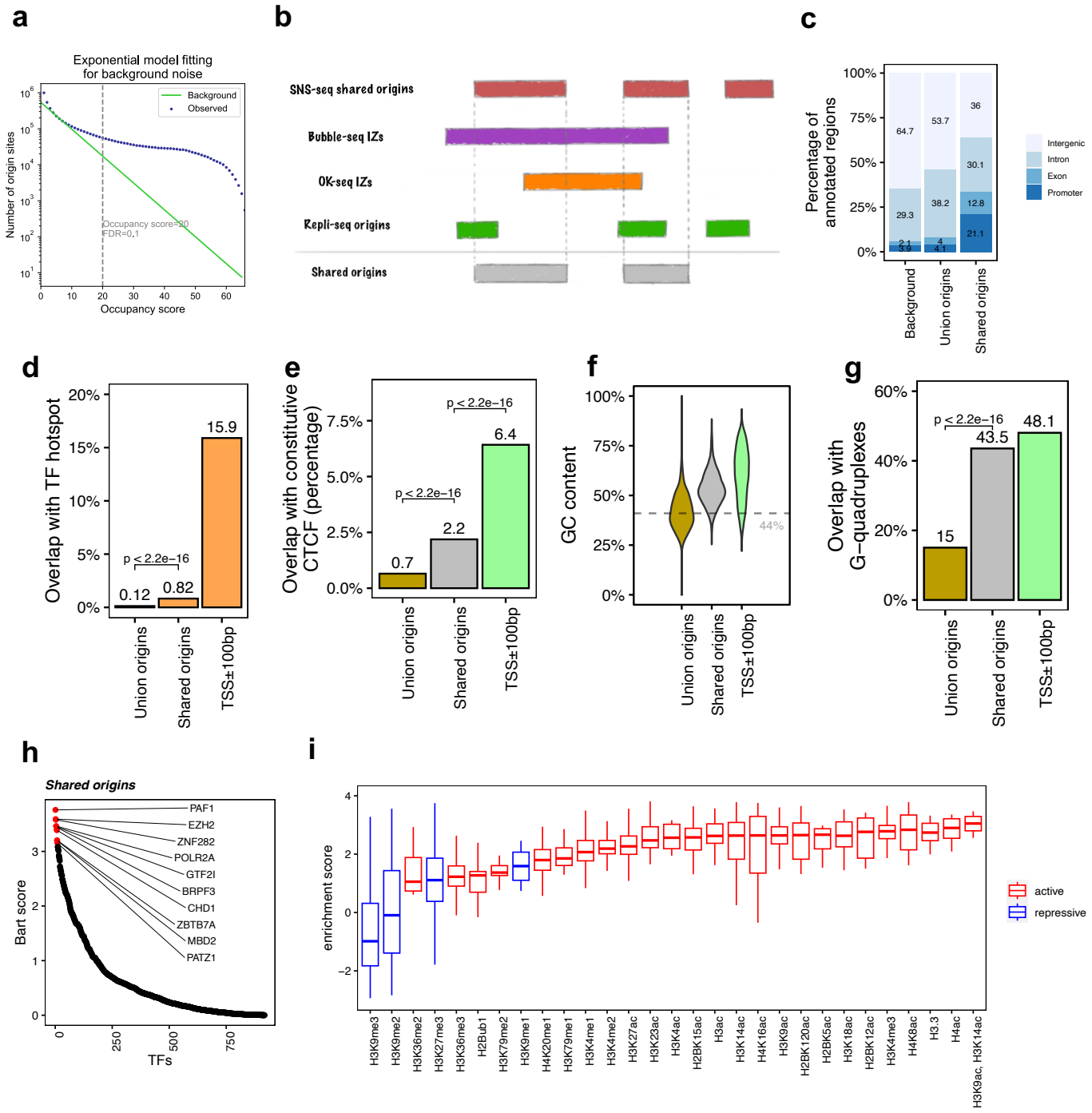


Figure 2: The shared origins are enriched with certain transcription factors and active histone marks. (a) SNS-seq origins fitting distribution to an exponential model. **(b)** Conceptual model of how shared origins is calculated. Every SNS-seq shared origin overlaps with any Bubble-seq IZ, OK-seq IZ 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

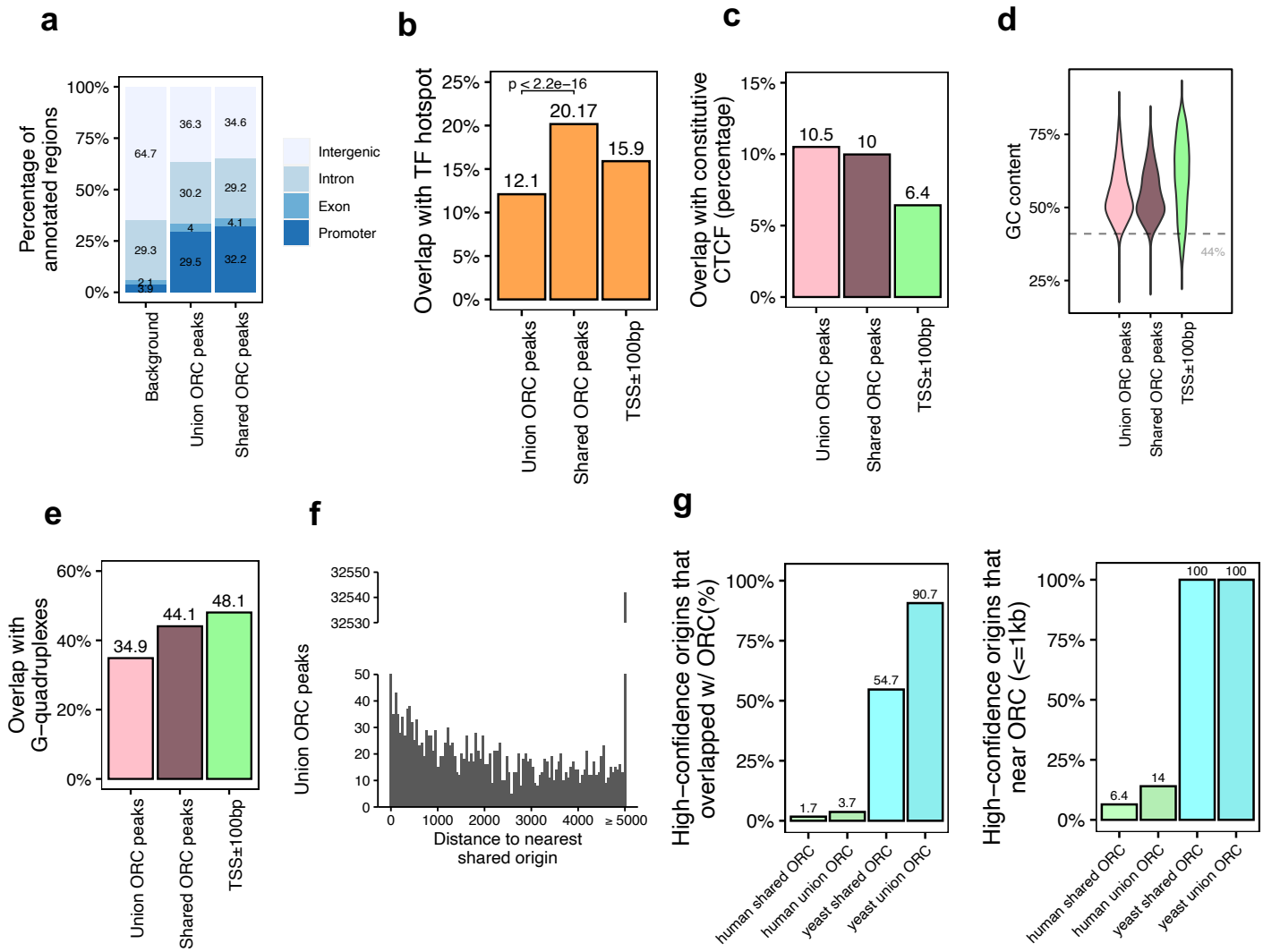


Figure 3: Genomic features of shared ORC binding sites. (a) Genomic annotation of union ORC and shared ORC binding sites. **(b)** Overlap with TF hotspot of union ORC and shared ORC binding sites. **(c)** Overlap with constitutive CTCF binding rates of union ORC and shared ORC binding sites. **(d)** GC content of union ORC and shared ORC binding sites. **(e)** Overlap with G-quadruplex of union ORC and shared ORC binding sites. **(f)** Distribution of the distance between ORC binding sites and the nearest shared origin. **(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 ORC binding sites.

Figure 4

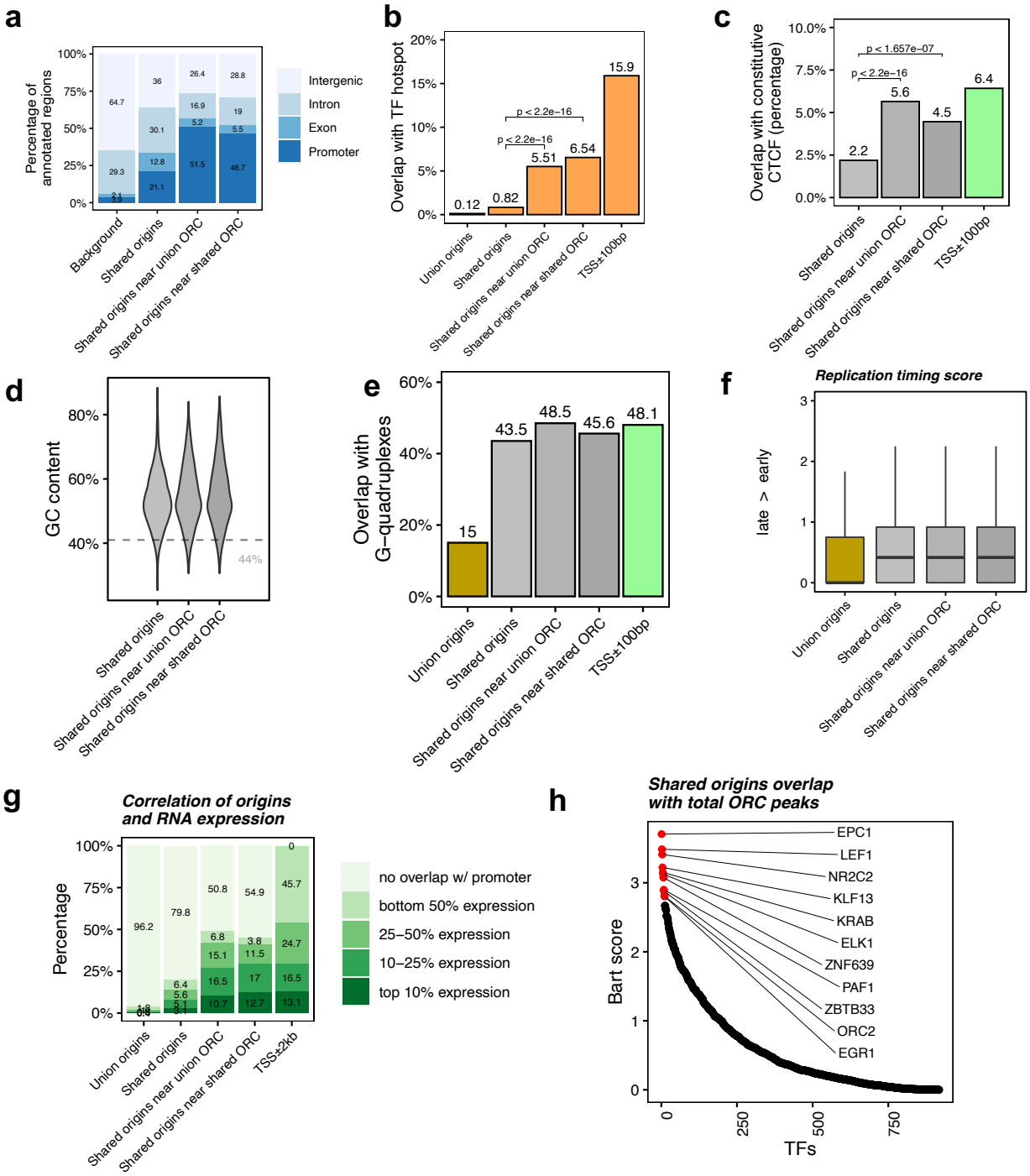


Figure 4: Shared ORC binding sites are more correlated with activate transcription. (a) Genomic annotation of shared origins and shared origins that near (less than 1kb) ORC binding sites. **(b)** Overlap with TF hotspots of shared origins and shared origins that near ORC binding sites. **(c)** Overlap 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)** Replication timing score of shared origins and shared origins that near ORC binding sites. **(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

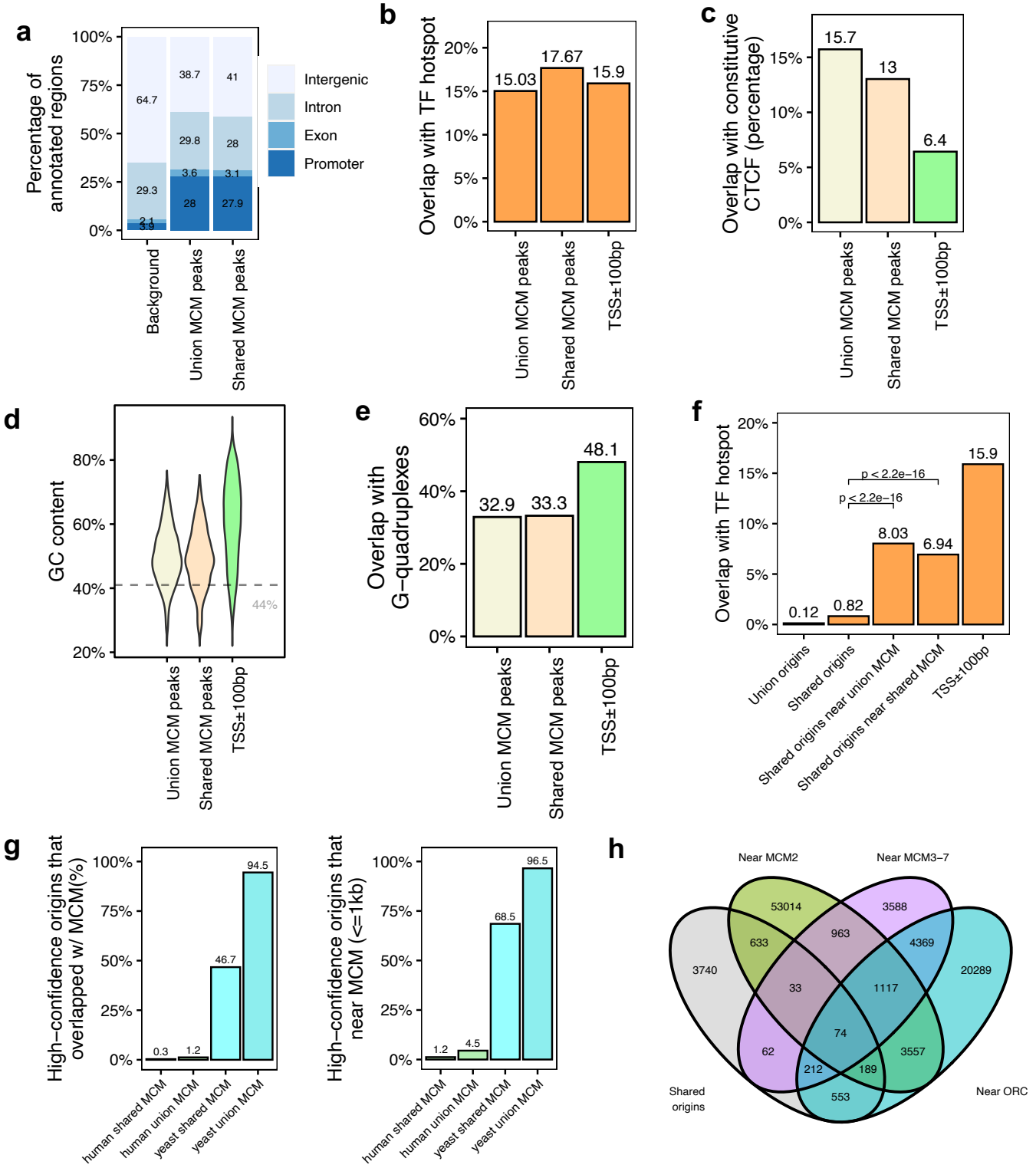


Figure 5: Genomic features of shared MCM binding sites. (a) Genomic annotation of union MCM and shared MCM binding sites. **(b)** Overlap with TF hotspot of union MCM and shared MCM binding sites. **(c)** Overlap with constitutive CTCF binding rates of union MCM and shared MCM binding sites. **(d)** 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

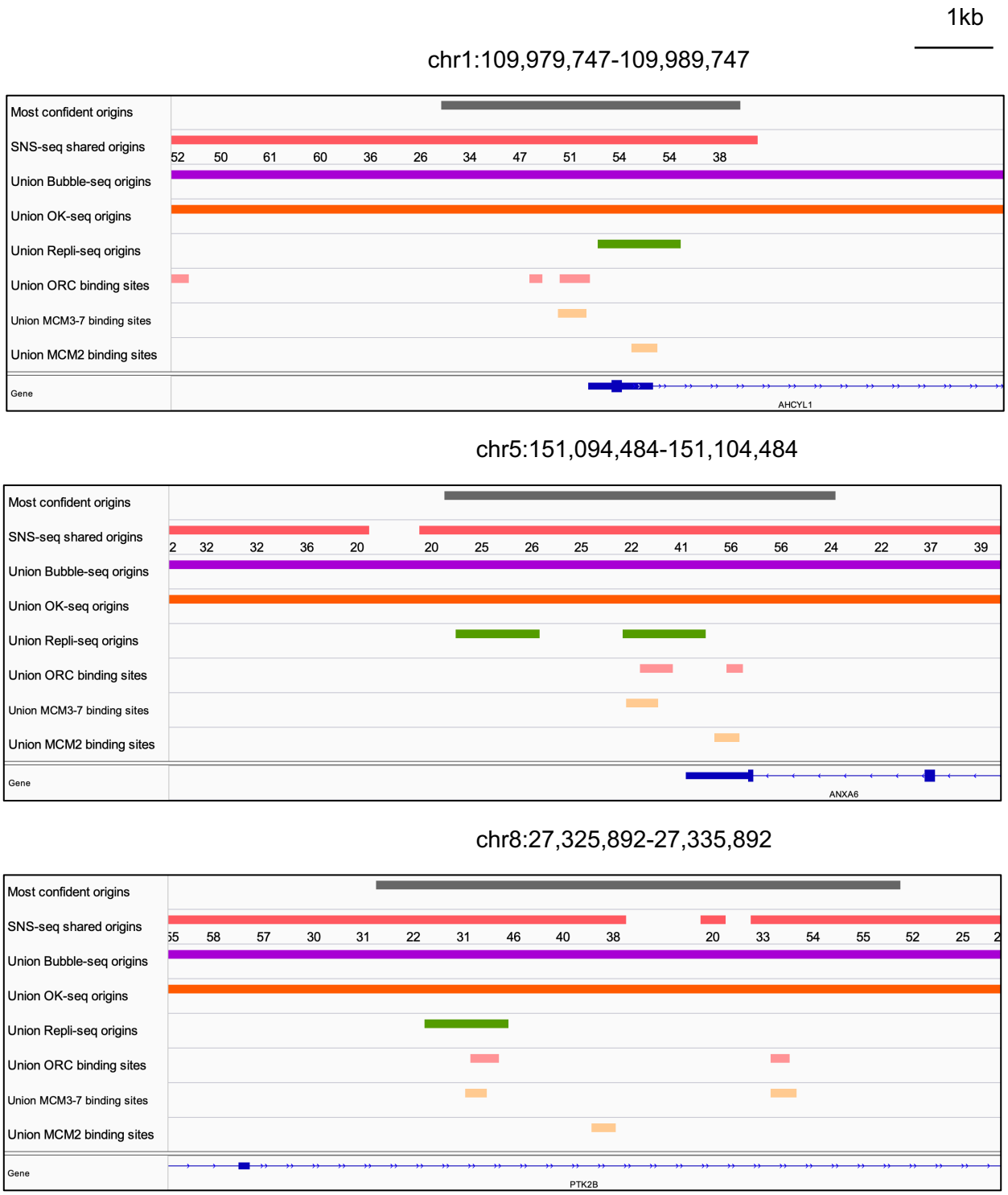


Figure 6: Genome browser screenshot for three of the 74 origins from Fig. 5h
The number on SNS-seq shared origins track is occupancy score of the origin.