Detection and characterization of replication origins defined by DNA polymerase epsilon

Roman Jaksik[1], David A. Wheeler[2*], Marek Kimmel[1,3]

[1] *Department of Systems Biology and Engineering and Biotechnology Centre, Silesian University of Technology, Gliwice, Poland*

[2] *Human Genome Sequencing Centre, Baylor College of Medicine, Houston, Texas, USA*

[3] *Departments of Statistics and Bioengineering, Rice University, Houston, Texas, USA*

Corresponding author:

Roman Jaksik

roman.jaksik@polsl.pl

Keywords

DNA replication, replication origins, mutator phenotype, polymerase ε, chromatin structure

[*] Current address: Clinical Genomics Group, Department of Computational Biology, St Jude Children's Research Hospital, Memphis, TN 38103

Abstract

Background: Although the process of DNA replication is highly conserved the location of origins of replication (ORI) may vary from one tissue to the next or one round of replication to the next in eukaryotes, suggesting flexibility in the choice of locations to initiate replication. Lists of human ORI therefore vary widely in number and location and there are no methods available to compare them.

Results: Here we report the genome-wide localization of ORI in POLE-mutated human tumors using whole genome sequencing data. Mutations accumulated after many rounds of replication of unsynchronized dividing cell populations in tumors allow to identify constitutive origins, which we show are shared with high-fidelity between individuals and tumor types. Using a Smith–Waterman-like dynamic programming approach, we compared replication origin positions obtained from multiple different methods. The comparison allowed us to define a consensus set of replication origins, identified consistently by multiple ORI detection methods.

Conclusions: Many DNA features co-localized with the consensus set of ORI, including chromatin loop anchors, G-quadruplexes, S/MARs and CpGs. Among all features, the H2A.Z histone exhibited the most significant association. Our results show that mutation-based detection of replication origins is a viable approach to determining their location and associated sequence features.

Background

DNA replication origins are crucial for initiation of the DNA synthesis, guiding the recruitment of proteins that form the pre-replication complex (pre-RC), including Mcm2-7 helicase. Helicase leads to the creation of a replication bubble, making the DNA accessible to polymerases, which replicate DNA in a bidirectional manner (3). The first step in pre-RC formation is recruitment of the origin recognition complex (ORC) that binds to specific regions in the DNA. These regions, referred to as the DNA replication origins (ORI) are selected based on sequence specificity in yeast (4, 5); however, in humans the recognition mechanism utilizes various DNA characteristics (7), and only a limited number of origins are active at each cell cycle (8). Efficiency of activation of the origins is used to classify them into three categories: constitutive, flexible and dormant (7). Constitutive origins

are used by all cells, independent of the cell type in each cell cycle, whereas activation of the flexible origins may vary in position or from one cell cycle to the next, or one cell type to the next (7). Dormant origins become active in stress conditions that affect the S phase, including serum starvation and DNA damage (9). Variation in origin activation by Flexible or Dormant origins may be one reason for large differences in the number of ORI between different cell types (10-12) and could also help explain differences in results obtained using various methods. Large variation in numbers and positioning of ORI in higher organisms constitutes one of the major confounding factors in the study of human ORI.

The target recognition mechanism of ORC requires DNA characteristics similar to those required by transcription factors at transcriptions start sites (TSS). They include: nucleotide composition, chromatin state, DNA methylation, and secondary structure of DNA (7, 13). Since ORI must be accessible to protein binding, their locations were shown to coincide with the nucleosome free regions, histone acetylation and DNAse sensitive sites (14). Additionally, a low DNA methylation level is an important factor, making some promoter regions suitable targets for ORC binding (15). As a result. several of the best studied ORI are in the vicinity of TSS of known genes such as *MYC*, *TOP1* and *LMNB2* (16-18).

Despite the similarities to TSS there exists no definite evidence of the existence of specific sequence motifs required to initiate ORC assembly in humans. However, in some unicellular eukaryotic genomes *Cis*-acting sequences determine the location of replication origins (4) and in yeast two sequence elements are necessary: a 17-bp autonomously replicating consensus sequence (ACS) that binds Origin Recognition Complex (ORC): WWWWTTTAYRTTTWGTT (19) and a broader sequence context encompassing 200 to 300 bp that appears to be important for depleting nucleosomes from the origin (20-23). In human cells, nucleotide composition (2, 24, 25) and G-quadruplexes (26) have been shown to increase the replication origin activity also affecting its location. DNA structure is also believed to play an important role in replication origin location and activity. Chromatin loops were shown to be associated with origins (27), and also it is believed that DNA replication is initiated in regions attached to the nuclear matrix (MARs) (28). In a more recent

work DNA replication origins were also shown to be overrepresented at the borders of topologically associating domains (TADs) (25).

Previously it was proposed (29) that mutational patterns emanated by the replicative DNA polymerases might effectively map the origins of replication, using mutations identified in whole genome sequencing experiments. Using this information, we developed a method for the detection of constitutive origins of replication, mORI (mutationally-defined ORI), based solely on mutation data from mutator phenotype tumor genomes. We also developed a novel method for the comparison of genomic positions which we used to compare multiple replication origin detection methods. Finally, we used the identified replication origins to characterize DNA structure in the vicinity of constitutive replication origins, determining the factors that are associated with their location.

Results

Detection of replication origins based on mutation patterns

DNA Polymerase ε is assumed to be responsible for the leading strand synthesis (30), a result first discovered in yeast, and subsequently reinforced by strong strand biases of context-specific mutations observed in cancers with mutation in the proof-reading domain of the enzyme – Fig. 1A (29, 31). To capture this feature and to identify potential replication origins we analyzed the distribution of mutations in POLE exonuclease damage tumors by developing a POLE-exo associated Mutation Asymmetry score (PMA). We obtained whole genome sequencing data (WGS) of tumors harboring POLE-exo mutation generated by the TCGA and ICGC projects (Supplementary Table 1) and focused on 20 cases out of 43 in which at least 20% of all mutations are either TCT→TAT and TCG→TTG — the two mutations most commonly found in these patients ((32) see mutation signatures 10a and 10b).

Although TCT→TAT and TCG→TTG variants were reported to make up the majority of all mutations in POLE-exo mutant tumors, we noticed that other mutation types might be of relevance since they also occurred in similar spatial patterns around putative ORIs. Since our replication origin detection method (mORI) benefited from higher mutation numbers we conducted it in two stages. In

the first stage we identified replication origins based on the PMA score calculated using only the reported TCT→TAT and TCG→TTG variants – Fig.1B.

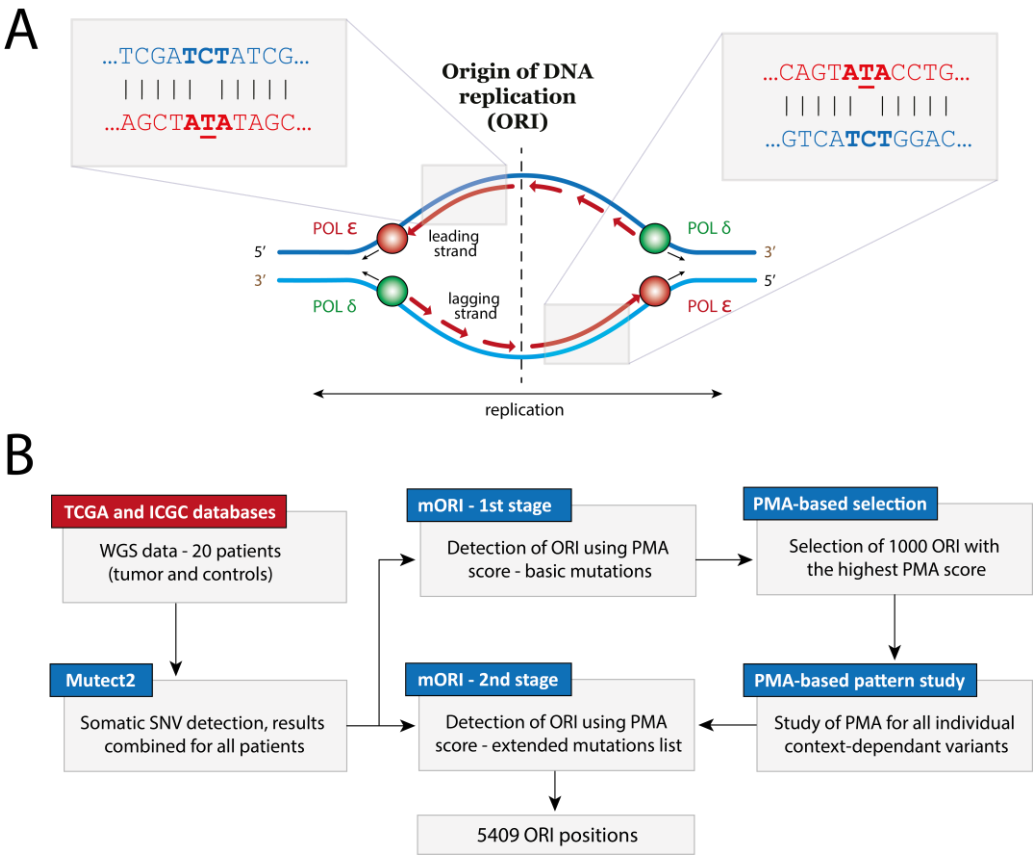

Fig. 1: A) DNA replication in human cells. Polymerase ε synthesizes the leading strand while polymerase δ is responsible for the synthesis of the lagging strand (30). Both polymerases start the synthesis from the replication origin (ORI) after replication is initiated by the origin recognition complex (ORC). Damage in the exonuclease domain of polymerase ε lead to an increased mutation rate with a preference for C→A mutation in TxT context, observed as TCT->TAT the newly replicated strand. This creates an asymmetric pattern of TCT>TAT and AGA>ATA in the vicinity of replication origins. **B)** Diagram of the two-stage ORI detection method (mORI), based on the PMA score

We then used 1000 identified ORI positions with the highest PMA score to study the distribution around these ORI of all 96 possible triple context-dependent variants. Those that showed similar patterns, belonging to clusters A and B shown on Fig. 2, were used in the second stage of the detection, which was based this time on 58 mutation types. Mutations in cluster A show a positive PMA score while those in cluster B a negative one, in order to account for this difference, in the second stage of the detection we combined mutations with C/A reference allele from cluster A with G/T from cluster B and G/T from cluster A with C/A from cluster B. This increased the total number of

mutations used from on average 422k to 606k per sample, also increasing the number of identified replication origins from 5132 to 5409.

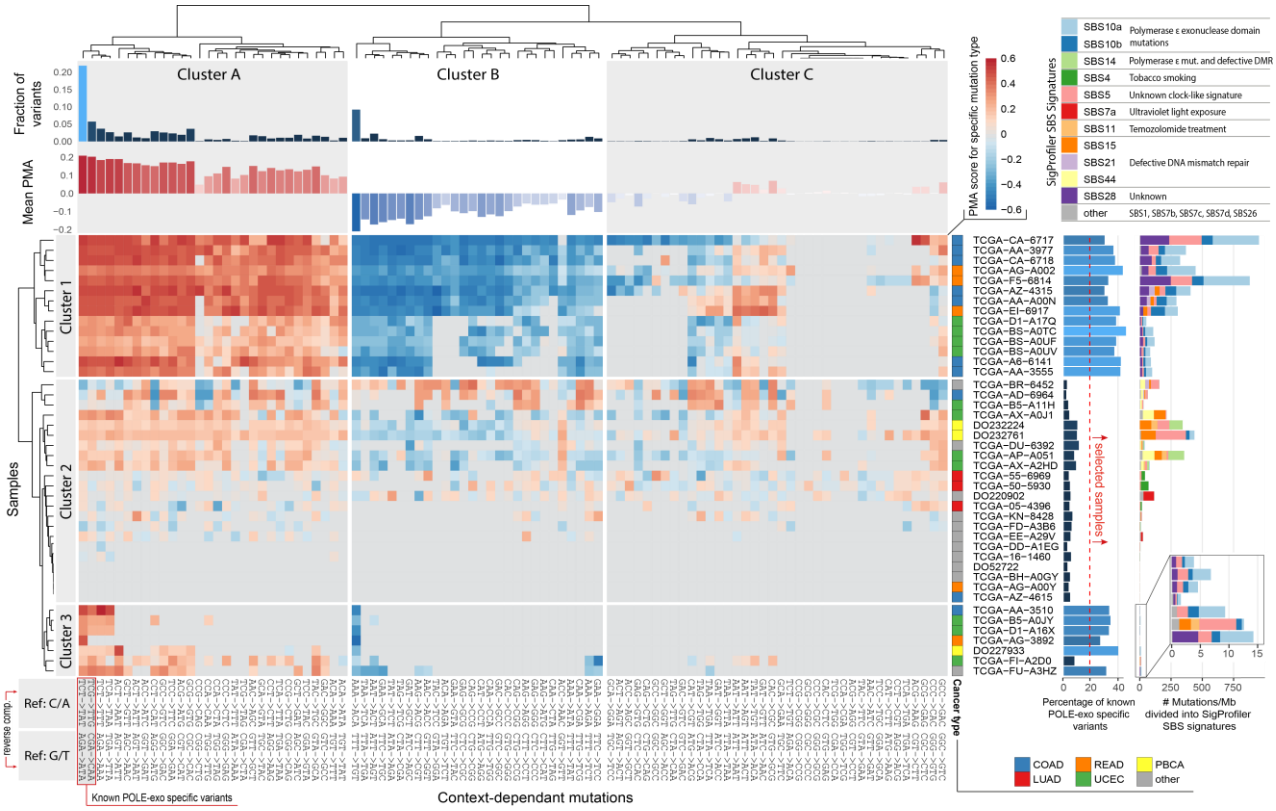

Fig.2: Heatmap of the PMA score calculated at ORI positions identified using known POLE-exo specific variants (marked on the plot). Samples (rows) and mutation types (columns), both clustered into 3 groups. Columns were clustered using only samples in which percentage of POLE-exo specific variants exceeded 20% (bar plot located on the right side). Mutations from clusters A and B were used to calculate PMA score in the final detection. Bar plots on the right show the percentage of POLE-exo specific variants (marked in the lower left corner) and total number of mutations per Mb in each of the samples divided into SigProfiler signatures listed in the upper right corner of the plot. Color-based annotation next to the samples IDs determines the cancer type (COAD - colon adenocarcinoma; LUAD - lung adenocarcinoma; READ - rectum adenocarcinoma; UCEC - uterine corpus endometrial carcinoma; PBCA - pediatric brain cancer).

Fig. 2 shows the PMA score obtained for the 1000 replication origins identified using TCT→TAT and TCG→TTG (referred to as POLE-exo specific variants) for individual mutation types and samples. The majority of 20 selected samples, which show at least 20% POLE-exo specific variants also show a similar pattern in other mutation types, including AAA→ACA / TTT→TGT, which make up around 9% of all mutations in all of the studied samples. Clusters A and B marked on Fig.2 contain a total of 58 mutation types which show different frequency in relation to the replication origin location. Cluster A contains, among others, all C→T mutations no matter the context while cluster B all A→C

mutations. The context however seems to affect the frequency of the mutation occurrence with TxT context being associated with the highest number of variants.

The samples differ significantly in terms of mutation profiles. Samples from cluster 1 (Fig.2) show a high number of mutations per Mb, from 52 in TCGA-D1-A17Q up to 952 in TCGA-CA-6717. This cluster represents colon, lung and rectum adenocarcinomas. SigProfiler decomposition of those variants shows a combination of SBS10a, SBS10b, SBS5 and SBS28 signatures the first two of which are characteristic of POLE-exo mutants, the second two are often observed in cancers, especially the SBS5 clocklike signature. Cluster 3 shows a similar SigProfiler decomposition, however the variant frequency is much lower, from an average 2 per Mb in TCGA−AG−3892 up to 14 in TCGA−FU−A3HZ. Cluster 3 contains similar cancer types as cluster 1, with an addition of one cervical squamous cell carcinoma and endocervical adenocarcinoma (TCGA-FU-A3HZ) and one pediatric brain cancer (DO227933). Sample TCGA−FI−A2D0 is an outlier in cluster 3 in terms of mutational signature, showing a majority of variants associated with SBS5 (unknown clock-like signature) and some from SBS15 (defective DNA mismatch repair) unseen in other samples from this group. This sample was not selected for the ORI detection as it doesn't pass the 20% cutoff of POLE-exo specific variants. Cluster 2 contains the remaining samples which also have less than 20% of POLE-exo specific variants, representing 12 different cancer types. While some of the cases exceed 100 mutations/Mb they show a significantly different pattern associated with defective DNA mismatch repair (signatures SBS15 and SBS44) but also signature associated with polymerase ε mutation (SBS14).

The ORI detection algorithm described is expected to identify only the positions which are conserved among individual cell division cycles and among individuals, except for cell type-specific sites and sites that can change very often depending on the number of cell divisions. We believe these sites are highly dependent on the specificity of the DNA structure in their vicinity, guiding the ORC to those locations.

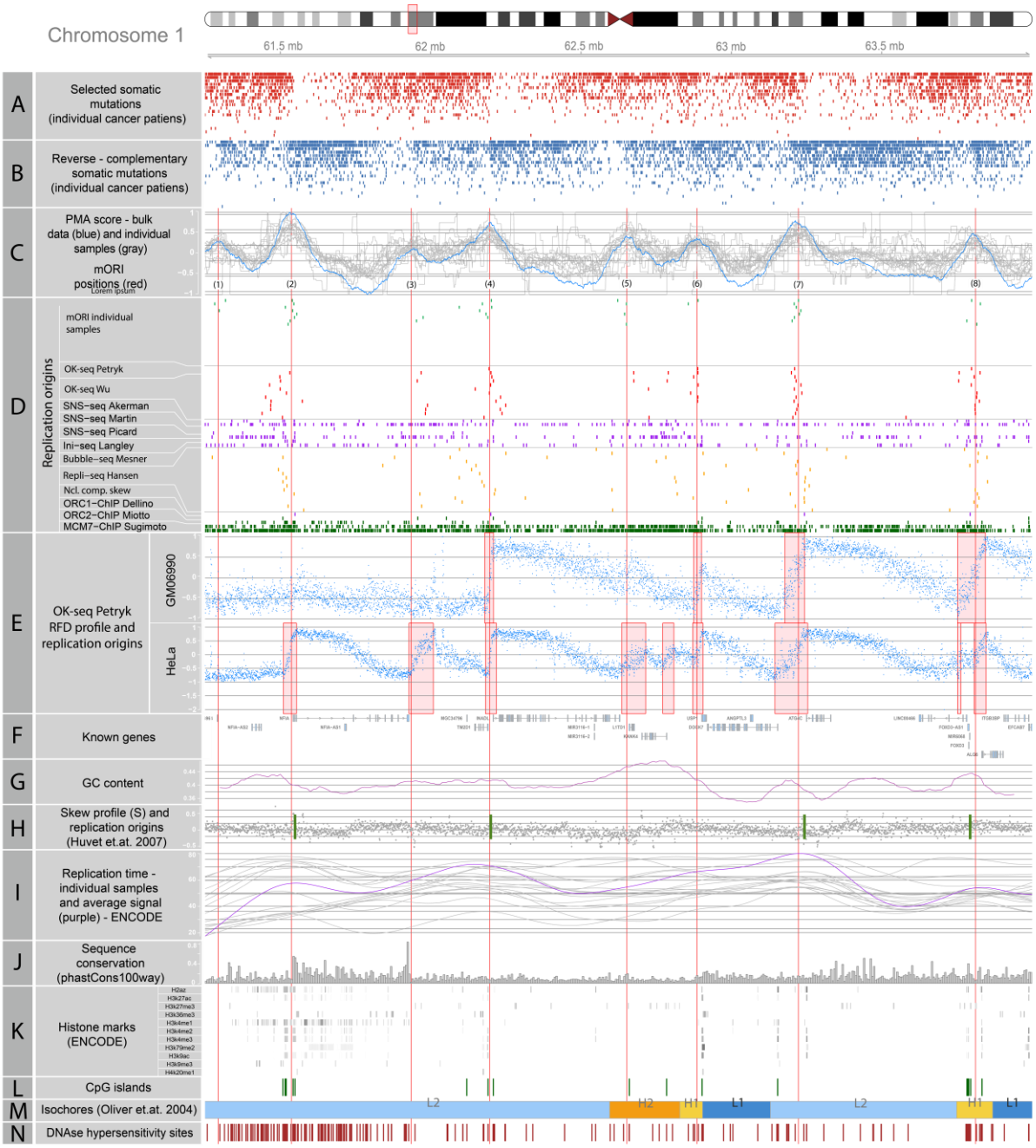

Fig.3: Replication origins identified in a fragment of chromosome 1. **A)** Individual context dependent somatic mutations used in the ORI detection algorithm (C/A reference allele from cluster A and G/T from cluster B, as shown on Fig.2), each row represents one individual patient; **B)** mutations of reverse complementary type to those shown on panel A (G/T reference allele from cluster A and C/A from cluster B as shown on Fig.2); **C)** The consensus PMA score calculated for the combined samples (blue line) and individual sample PMA score (gray), red lines mark the peak positions which represent eight replication origins identified by mORI numbered 1-8; **D)** Replication origins identified by other NGS-based methods, (see Supp Table 1) in other samples from various tissues; each row represents one sample; except the Akerman SNS-seq track where the top one corresponds to core and bottom to stochastic origins **E)** RFD profile (blue dots) for two OK-seq samples from (1), along with identified ORI positions (red rectangles); **F)** exons of known genes; **G)** GC content calculated in 1kb windows; **H)** nucleotide compositional skew profile (gray dots) and replication origin positions from (2); **I)** replication time obtained for individual ENCODE samples (gray line) and an average for all samples (purple line); **J)** sequence conservation score (phastCons100way); **K)** ENCODE histone marks; **L)** CpG islands (UCSC data); **M)** isochore positions from (6), lowest GC - L1,L2,H1,H2,H3 – highest GC; **N)** DNAse hypersensitivity sites obtained for K562 cell line (ENCODE data)

Distribution of identified replication origins

Fig.3 shows ORI sites identified in a 3 Mb segment on the p-arm of Chromosome 1, demonstrating the extreme strand bias of selected mutations in their vicinity (the entire list of identified replication origins is available in the Supplementary Table S2). The coefficient also shows local minima, which we assume represent sites where polymerase ε meets polymerase δ, i.e., the endpoints of the replication process. The minima are not as evident as maxima, showing a different distribution and in many cases are not placed halfway between the two neighboring replication origins. Possibly, this could be caused by variation in the firing time of individual replication origins (33), or the polymerase efficiency which, among other things, is known to be GC content-dependent (34).

Locations of the replication origins are reported to be associated with DNA features that are non-randomly distributed over the entire genome (7). For this reason, the identified replication origins might also be non-randomly distributed. To test this, we computed the average distance between identified ORI and compared it to the average distance between random positions in the genome. The average distances between the locations in the two sets were then compared, showing a statistically significant difference (p-value $< 2.2 \times 10^{-16}$), indicating that the variance of the calculated distances is significantly smaller for the identified ORI compared to the random positions, consistent with the former being more regularly spaced.

Comparison of ORI detection methods

Replication origin detection can be carried out using various experimental approaches, summarized in Table 1, which provide from 1 thousand to over 0.5 million positions in the human genome. The detection is expected to vary significantly in terms of sensitivity and specificity also targeting various classes of replication origins (constitutive, flexible and dormant), resulting in significantly different numbers of detected sites. Our replication origin detection method is based on an asymmetric mutational pattern between the leading and lagging strand. In that respect it is similar to replication fork directionality (RFD) profiles, shown on Fig.3, which originate from the Okazaki fragment purification and sequencing (OK-seq) (1, 35). We compared the mutational pattern in the present work to the experimental RFD profiles available for GM06990, a “normal” cell line, and for HeLa, a

cell line of cancerous origin, we also calculated the RFD profiles and carried out origin detection for 10 additional cell lines based on data provided by Wu (35). The number of ORIs detected using our method (5409) is similar to the number described by Petryk et al. for GM06990 (5684) and samples from Wu et al. (on average 5374).

Comparison of two sets of genomic coordinates is usually carried out by calculating the distance between elements located on a specific chromosome or finding the overlaps after converting the coordinates into ranges (36, 37). This however requires a definition of maximum allowed distance, used to find the overlaps, or pairing the coordinates between both sets for which the distance will be calculated. The latter can be very difficult especially if the overall number of features in both compared sets is different. The problem of comparing two sets of ORI is very similar to the one solved by Needleman-Wunsch global sequence alignment algorithm, where it is possible to introduce gaps into one of the compared sequences in order to account for insertions and deletions. We adopted this dynamic programming approach to work with ORI positions represented as numeric vectors on each chromosome instead of nucleotide sequences and used it to optimally pair to sets of origin positions, which led to an estimation of the pairwise distance between the two sets. Instead of rewards for match and mismatch we used the absolute distance between locations and used a gap penalty to control the number of gaps introduced to optimize the comparison -- gaps represent unmatched positions, from one method or the other (see Materials and Methods). We extended the algorithm to allow comparison of multiple ORI sets, which works similarly to ClustalW multiple sequence alignment (38).

We used both algorithms, termed numeric vector alignment (NVA) and multiple numeric vector alignment (MNVA), to compare the positions of replication origins between methods that provide similar number of replication origins (up to 10 000 positions in the genome). Panels A and B on Fig.4 show the results of MNVA conducted on replication origins we identified, based on mutation patterns in individual samples, with the highest number of mutations, as well as positions detected using the union of all samples. We also used experimentally identified ORI sites that originate from the works of Mesner, Picard, Langley, Petryk and Wu and from the computational method developed by Huvet (see Table 1). Each row represents one origin and the color scale corresponds to the

fraction of methods in which it was detected. By applying the MNVA algorithm we were able to collapse nearly 150 thousand genomic coordinates, across all chromosomes, from 31 distinct sets, into 10161 locations. Out of those 4666 were identified in at least 50% and 518 by more than 90% of the samples, from all ORI detection methods. Positions of all compared origins combined using the MNVA algorithm are available as the supplementary table S3. Fig. 4C shows the results of NVA pairwise comparisons between all origins identified. Fig.4 not only shows which positions are identified by the majority or minority of methods (panels A and B) but also the global similarities between them (panel C). Origins detected by mORI were most similar to those detected by the method of Huvet et al. and to the OK-seq. As expected, the three sets of randomly generated genomic positions were poor matches to all other sets.

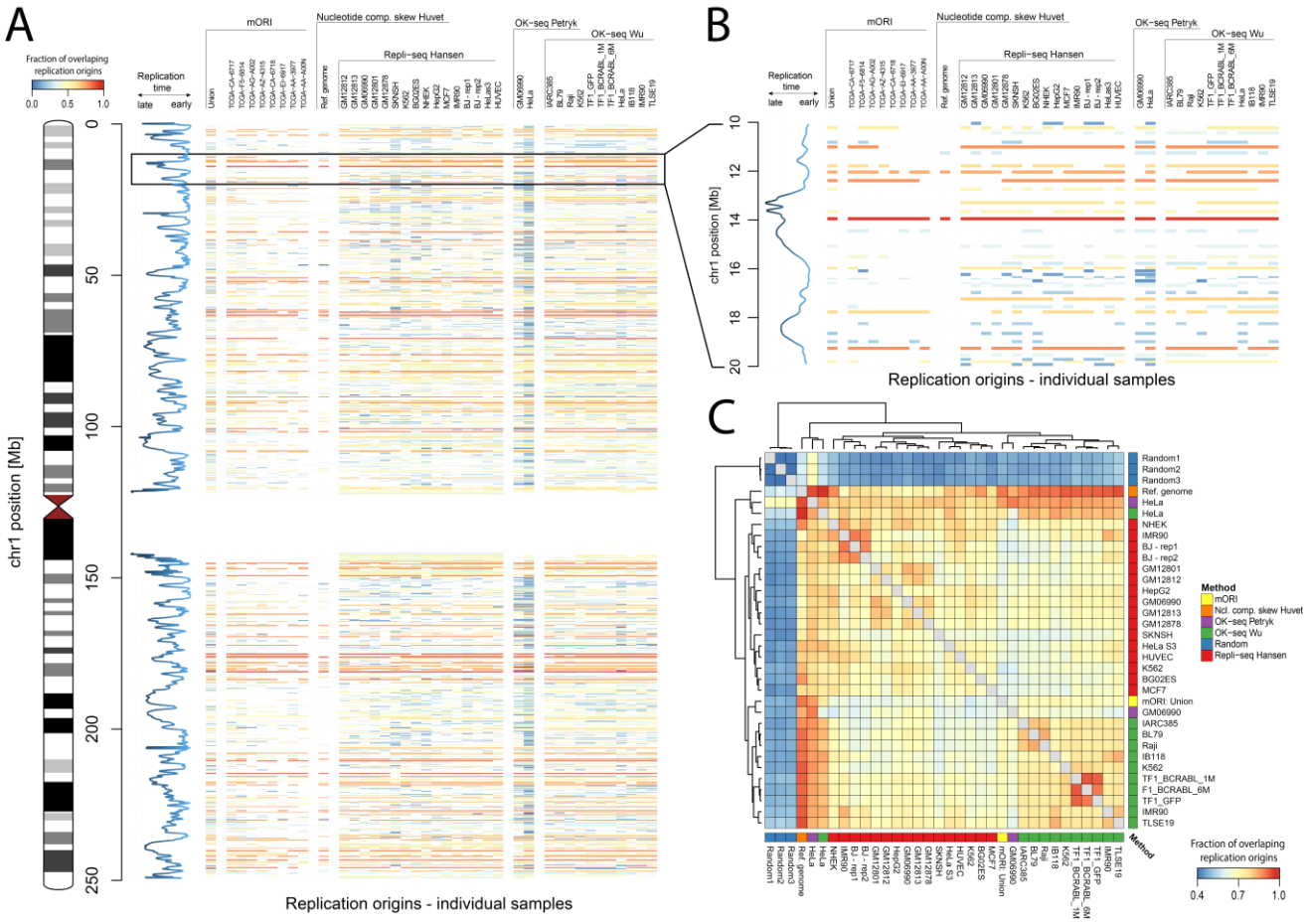

Fig.4: Similarity between various replication origin detection methods. **A**) results obtained using the multiple vector alignment algorithm for chromosome 1, each row represents one origin and the color scale corresponds to the fraction of methods in which it was detected; blue line represents the replication time data. **B**) enlarged fragment of panel A for 10 – 20Mb region of chr1. **C**) heatmap showing overall correlation between ORI detection methods, including a set of 5000 random genomic positions for comparison. The correlations are obtained using a dynamic programming-based vector alignment algorithm (see Methods).

DNA features associated with ORI

The mORI detection method is based on the PMA score, which additionally can be used to rank the origins. We assume higher PMA scores corresponds not only to the detection confidence but also the frequency at which the origin is utilized in a group of multiple analyzed samples. This provides the possibility to identify a set of consistently utilized replication origin positions that can be used to study the DNA features associated with their location. We conducted such analysis using all 5409 ORI positions, identified based on mutation patterns, and only using a subset of 1000 origins, with the highest values of the PMA score. In Fig. 5A we compared the number of repeat sequences, G-quadruplexes, histone marks, CpG islands, transcription start sites, S/MARs and chromatin loops at various distances from the replication origins, measured by 20kb windows from the putative ORI. The occurrences were summed over all studied origins and divided by the minimum occurrence levels, showing only the fold change for each interval.

As expected, the location of replication origins is associated with DNA features that affect the accessibility of DNA to protein binding. Many replication origins are located within promoter and gene-rich regions (7), especially those which are active in a given cell (39). Fig. 5A shows an increase in the number of transcription start sites (TSS) and DNase hypersensitive sites in the vicinity of replication origins (+/-50kb). Stronger association can be observed for the CpG island locations which are also overrepresented in the vicinity or replication origins.

Simple tandem repeats and repeated sequences derived from the Repbase repository (40) show only a minor increase in the vicinity or ORI. The highest increase among repeated sequences is observable for the Alu SINEs, which were reported to be associated with ORI location (10, 41).

G-quadruplexes (G4) were previously reported as being overrepresented in the vicinity of replication origins (26). To test this association we used G4 positions, identified in the (42) study based on G4-seq, which form under physiological K+ conditions and which are stabilized by pyridostatin (PDS). Results provided by both approaches are shown separately for each strand on Fig. 5A. Compared to other approaches, the increase in the G4 number in the vicinity of ORI is small, but clearly observable and offset from center in opposite direction on plus and minus strands.

Promoters of transcriptionally active genes are located inside nucleosome-free regions, which are reported to be associated with replication origins (20, 21). Chromatin structure limits DNA availability, and therefore it is one of the major factors that affect replication origin activity. Among all of the features that we compared H2A.Z histone exhibited the highest occurrence at the replication origins detected by mORI. Long has recently shown H2A.Z histone to epigenetically regulate the licensing and activation of early replication origins (43). Interestingly, our data show that H3k36me3 and H3k79me2 exhibit depletion at the mORI sites, while showing an increased occurrence frequency at the +/- 50-100kb distance.

Topological organization of the DNA inside the nucleus is another important feature that affects DNA replication. Origins located near the anchor points of chromatin loops were shown to have a higher activity (27). To validate this feature we compared the locations of chromatin loops (44), with the locations of the highest scoring 1000 mORI replication origins (Fig. 5B, see Materials and Methods for details). The number of replication origins shows a clear enrichment in the vicinity of loop anchors marked with a vertical dashed line; however, the number of origins drops significantly, approaching the loop center. We also found week association between scaffold/nuclear matrix attached regions (S/MARs), gathered in the MARome database (45), and replication origins which were previously reported to have possible association (28).

The number of mutations associated with the POLE-exo damage is assumed to be correlated with the number of tumor cell divisions (approximately 600 mutations per cell cycle (46)), implying tumor age contributes to the inter-patient differences that were observed (Fig. 3). However, epigenetic modifications to the DNA may also have a significant impact on the location of the replication origins (e.g., DNA methylation (47)) affecting the patient specificity of mutation patterns. We compared the methylation levels, determined in the TCGA project, in the vicinity of mORI replication origins with those for the distant parts of the genome (Fig. 5C). We observed a small drop in the methylation levels in the vicinity of replication origins (~10%) suggesting that the relation between both is not direct and might result from correlation with other structural features of the DNA.

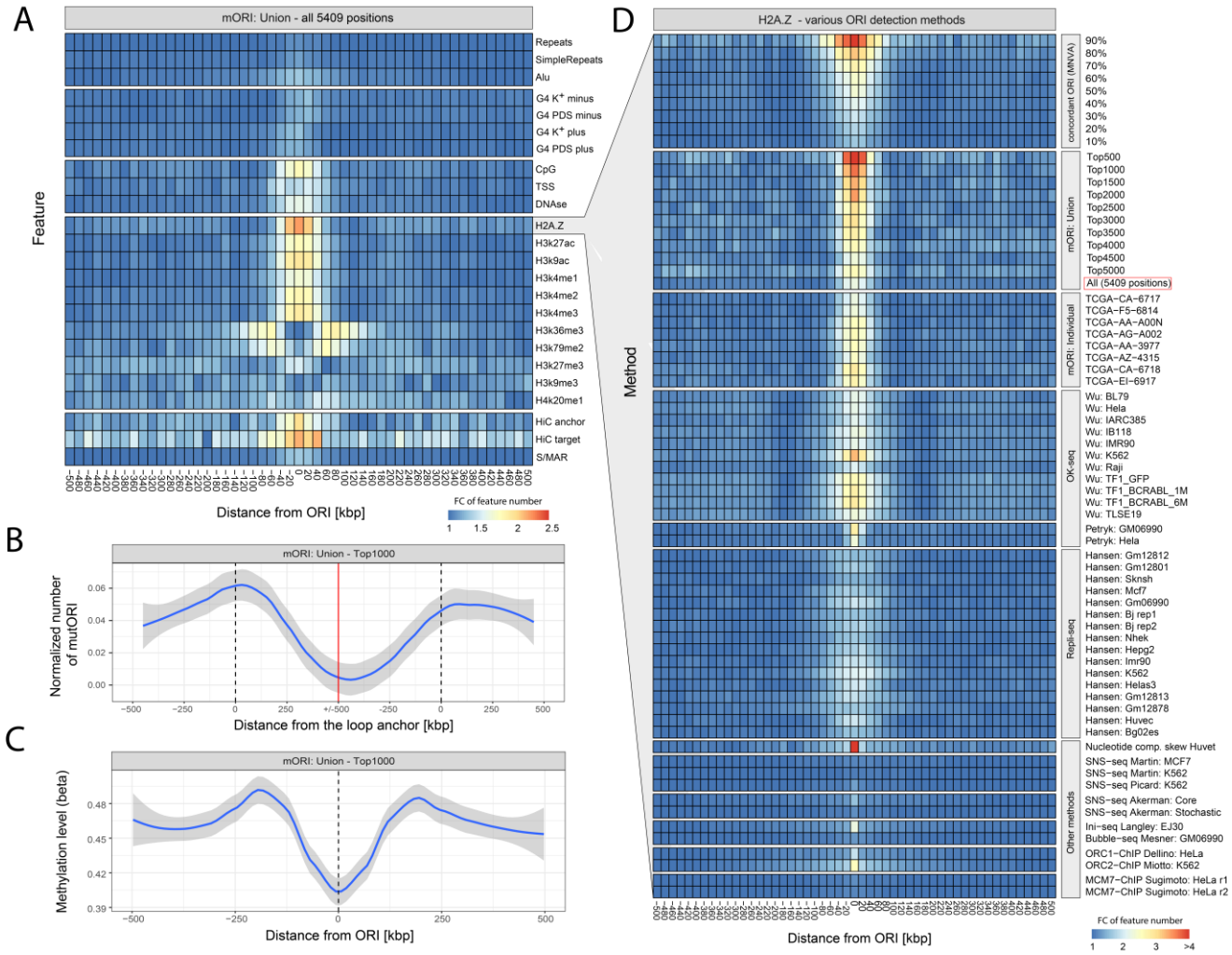

Fig.5: Correlation of mORI sites with epigenetic features of the genome. **A)** Number of specific features at a given distance from the replication origins, expressed as the fold change with respect to the minimum value from each row. **B)** Number of identified ORI at a specific distance from the chromatin loop anchor. Dashed vertical line shows the loop anchor points, red line marks the loop center. Since loops have different size but the plot shows a fixed interval of 1MB, the number of ORI was divided by the number of loops which reach this specific length. **C)** Average methylation level of the TCGA samples used for the detection of replication origins at a various distance from the Top 1000 origin positions. **D)** Statistics for H2A.Z histone obtained for various replication origin detection methods and results of the method concordance obtained using the MNVA algorithm.

The PMA score allowed us to select ORI conserved across multiple samples. Using the sample comparison based on MNVA, we also selected a subset of origin positions detected using other approaches. Fig. 5C shows the frequency of H2A.Z histone in the vicinity of replication origins identified using various approaches. The first panel was created using a subset of positions obtained using the MNVA algorithm, based on a various cutoff level, that defines the minimal agreement between all 31 analyzed samples shown below. We excluded the results of our mORI method obtained for individual samples (shown on the “mORI: individual” panel) since the data used to derive

them were also used in the combined sample study (mORI: union). The MNVA panel shows that the higher the agreement between the methods, the stronger the association between the replication origin position and frequency of H2A.Z occurrence. A similar trend can be observed in the second panel obtained for the results of mORI selection based on various cutoff levels associated with the PMA score. The higher the PMA score of the identified replication origins, the more frequently H2A.Z is associated with an ORI. While both approaches, based on method concordance and PMA score, can be successfully used to identify a set of conserved origins, the H2A.Z frequency is increased in a narrower interval in the vicinity of ORI selected based on the PMA score and the same situation can be observed for other studied DNA features. This suggests that the origin position is more precisely estimated using the mutation based approach, compared to the agreement between multiple methods, which is why we selected it for the study shown on Fig. 5A-C. Subsequent panels of Fig. 5D show the H2A.Z frequency in the vicinity of origins identified using mORI method applied to individual samples with high mutation counts and other replication origin detection approaches also shown on Fig. 4. Similarly to the ORI identified based on mutation patterns the positions obtained in the OK-seq methods also show high association with H2A.Z, however out of all approaches the most clear association was obtained for the computational method developed by Huvet et al.

Discussion

We employed on genomic scale the mutator phenotype associated with damaged polymerase ε, found in cancer cases (mainly colon adenocarcinoma and endometrial carcinoma), to identify genomic positions of ORI and developed an approach to compare positions obtained by other means, such as Repli-Seq or OK-seq. Mutations in the exonuclease domain of Polymerase ε affect the proofreading mechanism, leading to characteristic context-specific mutations grouped asymmetrically on the leading strand near ORI. Although exonuclease mutations in POLD1 are much less frequent than in POLE, several cases have been reported (48) and it would be interesting to determine whether POLD1 can also effectively mark ORI.

There are important biological differences in ORI determined by POLE mutational processes compared to approaches depending on isolation and sequencing of newly synthesized DNA, which

may lead to deeper insights into the function of mammalian ORI. Mutationally-defined ORI (mORI) are inscribed directly onto the DNA through the action of the replicative enzyme over many rounds of cell division in vivo; Repli/OK-Seq obtain data from a single round of replication in cell culture. One outcome of our mutational approach is that only the most consistently firing origins will be clearly observed, which may reduce the overall ORI count relative to other methods. Repli/OK-Seq are dependent on mapping small fragments of DNA, which may prove limiting in understanding how highly repetitive regions of the genome are replicated. mORI in principle could use long read technologies to overcome mapping difficulties.

We formalized the detection of mutationally-derived ORI by constructing, the POLE-exo mutation asymmetry (PMA) score, which assumes maximum value at the inferred position of ORI. ORI predicted by PMA scores are consistent overall with those obtained by other sequence-based methods. Using a subset of methods that provide a comparable number of replication origins we conducted a two-stage comparison of replication origin positions. In the first stage we conducted a pairwise comparison between each two methods using a numerical vector alignment (NVA) algorithm based on dynamic programming similar to Needleman-Wunsch for global nucleotide sequence alignment. mORI locations were most similar to the OK-seq-derived ORI among sequencing-based method. Subsequently, we compared multiple methods by extending the NVA algorithm to a multiple vector version (MNVA), using an algorithm similar to ClustalW multiple sequence alignment (38). Based on the results we assessed the overlap among all ORI sets and selected origins which are identified in a specific fraction of samples by various methods. We found that 4666 ORI were identified in at least 50% and 518 by more than 90% of the samples, across all methods.

Given the large concordance observed between origins detected using cell types from different patients, we assume that despite cell type and tissue of origin, the majority of ORI positions that we identified can be also observed in normal human cells. We believe that the replication origin positions identified in our study, both based on the mutation patterns and multi-method overlap provide a good basis to study the DNA features that characterize replication origin positions.

Therefore we sought associations of mORI locations with various DNA properties previously shown to co-occur with ORI location. Repeated sequences and G-quadruplexes were weakly associated with mORI origins, while DNA methylation levels and DNase hypersensitive sites were moderately associated. Replication origins from mORI were demonstrated to occur in the vicinity of chromatin loop anchors as was suggested previously (27). Weak association with S/MARs (28) as a potential factor influencing ORI location was also observed. Replication origins identified using mORI were also associated with epigenetic modifications, including methylation level and specific chromatin histones, most importantly H2A.Z. We further used H2A.Z as a benchmark to compare other ORI detection methods as well as various criteria used to select a subset of mORI and ORI identified in multiple samples/methods. Association between ORI and H2A.Z is stronger the higher is our prediction score (PMA) and also stronger as more methods identify a particular origin. Additionally, among the sequencing-based ORI detection methods we showed that Ok-seq, Ini-seq and ORC2-based ChIP-seq exhibit the highest association with H2A.Z. However an even stronger association with H2A.Z was observed for the computational method based on nucleotide composition skew (2). The association with H2A.Z is much weaker for methods that provide high number of origins, especially those based on SNS-seq (25, 49, 50), Ini-seq (51) and Bubble-seq (52). This suggests that they may represent a more variable class of origins—as mentioned, mORI is biased toward constitutive ORI.

The mORI detection method is applicable to cells characterized by the POLE mutator phenotype, which currently is only associated with cancer. This study used those POLE-exo mutated tumors found in ca. 1% of TCGA patients, primarily colorectal and endometrial cancers, but larger studies have been reported with representative tumors from many organ systems including brain, ovaries, prostate pancreas and lung (48). Moreover, it should be possible to introduce exonuclease-mutated POLE into the nuclear genome of any cell type using gene editing technologies.

Conclusions

We developed a novel method of replication origin detection, MutORI, based on the mutation patterns of POLE-exo tumors. MutORI identifies replication origins using whole genome sequencing data without any modifications, and then classifies origin utilization based on the value of the PMA score, revealing a set of constitutive replication origins in a single step. We applied this methodology to create the first ORI dataset generated from living tissues, rather than cell culture, and used it to characterize DNA structural features in the vicinity of identified ORI positions. The highest association of ORI was with histone H2A.Z, and the association increases with PMA score. We additionally proposed a new replication origin comparison methodologies, for pairwise ORI comparison, based on Needleman-Wunsch global nucleotide alignment, and multiple ORI sets by adapting a ClustalW multiple sequence alignment approach. These methods enabled a comparative analysis of replication origin positions from different cell types and methods. We report the relevant DNA characteristics associated with the locations of the commonly observed ORI. Selection of ORI positions based on our comparison method allowed us to identify 518 highly conserved ORI, which are detected in over 90% of samples from multiple cell lines and using various detection methods.

Methods

Detection of replication origins based on somatic mutations (mORI)

For the purpose of the mutation pattern detection we define a POLE-exo mutation asymmetry score PMA which is calculated for each n-th position of the genome:

$$PMA_n = \frac{4}{W_n^2}\left(\sum_{i=n-d}^{n-1} k_i^{lead} \sum_{i=n}^{n+d} k_i^{lag} - \sum_{i=n-d}^{n-1} k_i^{lag} \sum_{i=n}^{n+d} k_i^{lead}\right) \quad (1)$$

$$W_n = \sum_{i=-d}^{d} k_i^N \quad (2)$$

where:

d – window size

k_i^{lead} – number of mutations specific to the upstream of replication origin region e.g. T**C**T → T**A**T, identified at position *i* of the genome for all patients

k_i^{lag} – number of reverse complementary mutations to those from k_i^{lag} e.g. A**G**A → A**T**A, identified at position *i* of the genome

k_i^N – total number of all mutation types, at position *i* for all patients

$$k_i^{lag} + k_i^{lead} \leq k_j^N \in \{0,1, \dots m\}$$

m – number of patients

k_i^{lag} and k_i^{lead} represent mutations specific to the upstream and downstream region from a particular genome position *i* at which the coefficient is calculated. The maximum value of the coefficient is 1 which corresponds to the highest possible difference in mutation occurrence between regions upstream and downstream from a given position. The minimum is -1 which corresponds to a reverse pattern that can be observed in the vicinity of regions where two polymerases meet from opposite directions. The coefficient was calculated every 1kb over the entire genome, using 200kb window (100kb upstream and downstream from the selected position). We then smoothed the coefficient (moving average n=100) and identified local maxima using the peakPick R package (neighlim=100, peak.npos=50). Maxima's with PMA lower than 0.1 were omitted. The peaks were then filtered using Fisher's exact test with Benjamini-Hochberg multiple testing correction, assuming 0.01 significance level. This removed regions identified using a small number of mutations which are likely false positives. We tested other approaches to peak filtering including methods utilizing strict criteria based on the PMA score, the number of mutations in the sliding window, or other tests and multiple testing correction methods. All methods used showed a high concordance affecting only the stringency of the tests which is dependent on the parameters and significance levels utilized.

Replication origin position obtained using other methods

The positions of identified replication origins were compared to positions obtained using other approaches, both experimental and one additional computational. Table 1 summarizes the results

showing the number of available samples and identified replication origins. It also highlights which methods were used in the comparison.

Table 1. Summary of large scale replication origin detection methods in human cells.

Type	Subtype	Method	Cells	# Samples [1]	# Orgins/sample	Reference
Based on next generation sequencing	Synchronized cells	SNS-seq	hESC H9, HC, HMEC, HMEC-derivatives	6	40 000 - 110 000	(25)
			K562	1	239 107	(50)
			Primary basophilic erythroblasts	1	[3] 266 378	(53) [6]
			HCT116	1	100 301	(54) [6]
			HeLa; IMR-90; hESC H9; iPSC	4	[3] 236 480	(26) [6]
			K562; MCF7	2	62 971; 94 195	(49)
		Repli-seq	HCT116; WA01; WA09	3	n/a	(55)
			K562; HeLa-S3; …	16	[3] 4 812	(10)
		Ini-seq	EJ30	1	25 053	(51)
		Bubble-seq	GM06990	1	124 646	(52)
	Asynchronous cells	OK-seq	hTERT RPE-1	1 (two cond.)	not determined	(56) [6]
			K562; HeLa; …	11	[2,3] 5 374	(35)
			HeLa; GM06990	2	9 836; 5 684	(1)
		ORC1-ChIP-seq	HeLa	1	13 626	(57)
		ORC2-ChIP-seq	K562	1	52 251	(58)
		MCM7-ChIP-seq	HeLa	1	[5] 434 207	(59)
	NA	Nucleotide comp. skew	Reference genome	1	1 060	(2)
Based on microarrays	Synchronized cells	Repli-chip	HeLa-S3; GM06990; …	9	n/a	(60) [6]
		Bubble-chip [4]	HeLa; GM06990	2	128; 177	(61) [6]
		NS-BrIP [4]	HeLa	1	815	(62) [6]
		NS-LExo [4]	MCF7; BT-474; H520	3	8 281; 4 432; 3 201	(63) [6]
			HeLa	1	320	(62) [6]
			HeLa	1	283	(64) [6]

[1] excluding replicates
[2] replication origins were obtained by processing the data using our algorithm
[3] average for all samples/replicates
[4] based on tiling microarrays that do not cover the entire genome
[5] average for two replicates (includes dormant origins)
[6] not used in the method comparison described in the results section

Genomic positions of experimentally identified ORI positions were downloaded from 11 various sources, we only used data that originate from NGS based, genome wide studies, listed in Table 1. In some cases the data required some pre-processing, as described in the table below:

Table 2: Sources of alternative ORI positions used in this study

Method	Reference	Comments
OK-seq	(35)	BAM files for hg19 were downloaded from the SRA database, under accession number PRJEB25180. Replication origins were identified using our algorithm described below.
	(1)	BED files with ORI positions for hg19 were provided by the authors of the publication.
SNS-seq	(25)	Genomic coordinates of ORI positions for hg38 were downloaded from the manuscript supplement - Table S1. The coordinates were converted to hg19 using liftOver function from the *rtracklayer* Bioconductor library (65). ORI from Q1 and Q2 were combined into the "core" group, while the remaining Q3-10 into "stochastic".
	(50)	BED files with ORI positions for hg19, for Hela, IMR90 and K562 cell lines, were downloaded from http://pbil.univ-lyon1.fr/members/fpicard/oriseq/
	(49)	hg19 ORI positions were downloaded from the DeOri database (11) for both HCT116 and K562 cell lines

Repli-seq	(10)	ENCODE Repli-seq data were downloaded from GEO dataset GSE34399. We used the peak locations of the wavelet-smoothed signal available as the BED files with Pk suffix.
Ini-seq	(51)	Genomic coordinates of ORI positions for hg19 were downloaded from the manuscript supplement - Table S3 (Determination by SICER at E = 10e-5).
Bubble-seq	(52)	ORI positions for hg18 were downloaded in a bedGraph format from GEO dataset GSE38809 (GM_combined_RD_bubbles file). The coordinates were converted to hg19 using liftOver function from the *rtracklayer* Bioconductor library (65).
ORC1-ChIP-seq	(57)	ORI positions for hg18 were downloaded in a BED format from GEO dataset GSM922790 (ChIPseq_Orc1_GradientHela_enrichRegions file). The coordinates were converted to hg19 using liftOver function from the *rtracklayer* Bioconductor library (65).
ORC2-ChIP-seq	(58)	Genomic coordinates for hg19 were downloaded from the manuscript supplement (Table S1).
MCM7-ChIP-seq	(59)	Genomic coordinates for hg19 were downloaded from GEO dataset GSE107248 (MCM7ChIP_exp1/2.txt files).
Nucleotide comp. skew	(2)	ORI positions for hg17 were downloaded from the manuscript supplement (N-domain_positions.xls file). The coordinates were converted to hg19 using liftOver function from the *rtracklayer* Bioconductor library (65).

Detection of origins based on OK-seq data

The ORI detection algorithm which we developed for OK-seq data is similar to the approach used in (1). OK-seq is based on replicative incorporation of the EdU followed by size fractionation of EdU labeled fragments and Illumina sequencing. The orgins are identified using RFD profiles, computed for 1kb window using the following formula:

$$RFD = (C-W)/(C+W)$$

where C and W correspond to the number of reads mapped on Crick and Watson strands respectively. The RFD values range from -1 to 1 which correspond to the highest proportion of leftward and rightward moving forks respectively. In the vicinity of the replication origin the RFD profile should show a significant change in the RFD value from -1 to 1.

To calculate the RFD profile we first used samtools view (66), ver. 1.11, to separate reads from both strands (-b -h -F 16 parameters to get the forward strand reads and -b -h -f 16 for the reverse strand). For this purpose we used aligned reads stored in a BAM file. We than used bedtools coverage (67) ver. 2.25.0 to obtain the number of reads from both strands in 1kb windows defined in a BED file, generated using bedtools makewindows, for hg19 reference genome. Based on those results we calculated the RFD profile using the formula defined above, however we rescaled the number of Crick reads (C), so that the total number of W and C reads is identical for each BAM file.

To detect the positions of replication origins based on the RFD profile we first fitted a linear model to the RFD values from a 200kb windows, moving by 1kb across the entire genome. The slope parameter of the linear model has the highest value at positions where the RFD profile shows a shift from -1 to 1, which is assumed to represent replication origins (1). We smoothed (moving average n=200) the slope values of the linear models, calculated for the entire genome and later identified the local maxima, using the peakPick R package (neighlim=100, peak.npos= 100, deriv.lim = 1).

We applied this algorithm to the entire (35) dataset (11 samples) for which positions of replication origins were not provided by the authors. The analysis was based on BAM files downloaded from SRA (PRJEB25180), which were aligned to the hg19 reference genome.

Comparison of replication origin positions using NVA and MNVA

Genomic coordinates of DNA replication origins were compared using a modified version of the Needleman-Wunch global sequence alignment algorithm, named numeric vector alignment (NVA). The modification allows to use it for comparison of two sets of numeric vectors, instead of nucleotide or amino acid sequences (character vectors). The alignment is performed for each chromosome individually. The modified algorithm doesn't require a substitution matrix, which defines scores given for matches and miss-matches between specific nucleotides. Instead it uses the absolute distance between both locations and the only parameter required is the gap penalty that controls the number of gaps introduced to the compared vectors.

Individual elements of the scoring matrix *S* are defined as:

$$S_{i,j} = \min(S_{i-1,j-1} + D_{i,j},\ S_{i,j-1} + d,\ S_{i-1,j} + d)$$

where *D* is the absolute distance between locations *i* and *j* from the A and B vectors:

$$D_{i,j} = |A_i - B_j|$$

and *d* is the gap penalty parameter. The scoring system is reversed compared to the classical Needleman-Wunch algorithm since the higher is the distance between the elements the less similar are the vectors. The alignment score (value associated with the overall quality of the alignment) is defined as the sum of absolute distances between all paired vector elements. For each unpaired

element we add the gap penalty. In our study we used *d*=1000000. The method also returns a consensus vector, which is the average of paired elements from both individual vectors.

To compare a set of multiple replication origin positions, used to create Fig. 4A-B, we created an algorithm similar to the ClustalW multiple sequence alignment (38), named multiple numeric vector alignment (MNVA). The algorithm comprises of the following steps:

1) Calculate all possible pairwise alignments using NVA, record the alignment score for each pair of vectors
2) Create a guide tree based on the pairwise alignment scores matrix obtained in previous step, using the Neighbor Joining algorithm.
3) Align the sequences with NVA by progressive method based on the tree obtained in previous step. In each following iteration the selected vector is aligned to a consensus obtained in the previous iteration.
4) Create the final alignment matrix by aligning each individual vector to the vector obtained in the final iteration of previous step (without introducing new gaps).

The function returns a *n* by *m* matrix where *n* is the number of provided sequences and m is the length of the vector with introduced gaps, obtained in final iteration of step 3. The matrix elements include the original values of each vector separated by added gaps, marked with the dash character. Implementations of both algorithms, NVA and MNVA can be downloaded from our GitHub repository.

Generation of random genomic positions

Random genomic positions used in the ORI comparisons were selected using the createRandomRegions function from the regioneR package (68), which allows to exclude regions with gaps in the reference genome. For the study of distances between ORI identified using PMA score which was compared to random genomic positions we additionally selected the positions only from regions were mutations were identified. This approach allows omitting the regions with unknown sequence and low complexity, which due to lack of mutation data would bias the results.

Analysis of genomic features in the vicinity or replication origins

The following table lists all genomic features visualized on Fig. 4 and used to determine the number of features at a specific distance from replication origins used to create Fig. 5A,D:

Table 3: Sources of coordinate locations of various DNA features used in this study.

Feature	Source
RFD profiles	RFD profile data for HeLa and GM06990 cells was downloaded from the database described in (1). Positions of replication origins marked with red rectangles are based on BED files with ORI positions provided by the authors.
Known genes	Exon locations for hg19 used to make gene visualizations, were obtained from the UCSC FTP server: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/refGene.txt.gz. Alternative splice variants for a particular gene were combined into a single entry that contained all possible exons.
GC content	GC content was calculated in 1kb intervals for the hg19 reference genome based on the BSgenome.Hsapiens.UCSC.hg19 and seqinr (69) R libraries.
Skew profile (S) and replication origins	Compositional skew was calculated according to the specifications from (2), in 1kb intervals across the entire hg19 reference genome.
Replication time	Replication time data for hg19, from 15 cell lines, obtained in the ENCODE project, were downloaded from GEO (ID: GSE34399) in a bigWig file format. The data represent smoothed wave signals for 1kb windows, obtained in the Repli-seq experiment.
Sequence conservation	Sequence conservation data was obtained from the phastCons100way UCSC track using phastCons100way.UCSC.hg19 Bioconductor library (70)
Histone marks	Histone marks, including H2az, H3k27ac, H3k27me3, H3k36me3, H3k4me1, H3k4me2, H3k4me3, H3k79me2, H3k9ac, H3k9me3, H4k20me1 were downloaded from the UCSC table browser, for K562 cell line.
CpG islands	CpG island locations for hg19 were downloaded from the AnnotationHub (AH5086 track) using AnnotationHub Bioconductor library (71)
Isochores	Isochore locations for hg19 were downloaded from https://bioinfo2.ugr.es/isochores database (6) and divided into one of 5 groups: L1, L2, H1, H2, H3, based on their average GC content according to the following thresholds: L1 ∈ [0, 37); L2 ∈ [37,41); H1 ∈ [41,46); H2 ∈ [46,53); H3 ∈ [53,100)
DNAse hypersensitivity sites	DNAse hypersensitivity peaks originating from the ENCODE project, were downloaded from the UCSC table browser, for K562 cell line (72).
Repeats	Repeat sequence locations for hg19 were downloaded from the AnnotationHub (AH5122 track) using AnnotationHub Bioconductor library (71)
Simple repeats	Locations of simple repeats for hg19 were downloaded from the AnnotationHub (AH5124 track) using AnnotationHub Bioconductor library (71)
Alu sequences	Alu sequences are a subset of the Repeats track which contains all repeats from the Alu family (37 types).
G-quadruplexes	G-quadruplex locations for hg19 reference genome were downloaded from GEO (ID: GSE110582). The G4 locations originate the from a study based on G4-seq (42)
Transcription start sites (TSS)	Locations of transcription start sites (TSS) were determined based on the UCSC gene annotation file downloaded from the FTP server: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/refGene.txt.gz
Chromatin loops	Chromatin loop data were obtained from the GEO database (ID: GSE63525) for K562 cells (44).
S/MARs	Locations of scaffold/nuclear matrix attached regions (S/MARs) for hg19 were downloaded from the MARome database in a BED file format (45)

Study of methylation levels at ORI locations

Processed methylation data (beta values) for hg19 were downloaded from the GDC Data Portal (TCGA project) for 19 out of 20 samples, which were used for replication origin detection based on mutation pattern. The only omitted sample was DO227933, for which methylation data was not available in ICGC database at the time we wrote the manuscript. The downloaded data originated from both Illumina Infinium HumanMethylation450 and HumanMethylation27 BeadChips. While both platforms differ significantly in the number of tested CpGs we were only interested in the variability of methylation levels at a certain distance of identified replication origins without a direct comparison between the samples. For this reason all samples were combined into one set which was used to calculate average methylation levels at a particular genomic location. The differences between both platforms affected the weight of individual sample in the average methylation levels, however we find that to be of small relevance to the conclusions we have drown based on those statistics.

Comparison of ORI locations and chromatin loops

Chromatin loop data were obtained from the Gene Expression Omnibus database (ID: GSE63525) for K562 cells (44). The locations were used to estimate the distribution of replication origins in the vicinity of loop anchor by calculating the number of replication origins at a specific distance from the closest loop. We considered the loop itself and the +/- 500kb surrounding region; however, since the loops have different lengths we divided the number of origins by the total number of loops that reach a specific length, similarly as we proposed in our previous work (73) for transcription factor binding sites.

Abbreviations

G4 - G-quadruplexes

ICGC - International Cancer Genome Consortium

MNVA - Multiple numeric vector alignment

mORI - mutationally-defined ORI

NVA - Numeric vector alignment

ORI – origins of replication

PMA - POLE-exo associated Mutation Asymmetry score

RFD - replication fork directionality

S/MARs - scaffold/nuclear matrix attached regions

TADs - topologically associating domains

TCGA – The Cancer Genome Atlas

TSS - transcriptions start sites

WGS - whole genome sequencing data

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Details concerning POLE-exo mutants, positions of identified replication origins and comparison of replication origin positions, based on MNVA algorithm, identified using multiple approaches are available as supplementary tables.

Implementations of the main R functions used in this study are available in the replicationOrigins R package, which is available through GitHub: https://github.com/rjaksik/replicationOrigins. The package includes implementations of the following algorithms:

- Replication origin detection based on mutation patterns
- OK-seq origin detection based on RFD profiles
- Numeric vector alignment (NVA)
- Multiple numeric vector alignment (MNVA)

The remaining code used for additional analysis and visualizations is available per reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the Polish National Science Centre grant No. 2016/23/D/ST7/03665.

Authors' contributions

RJ implemented the data processing methods, conducted the analysis and drafted the paper. DW contributed to the study design, interpretation of the results and writing of the manuscript. MK supervised the project and contributed to the study design, aided the data analysis and contributed to the writing of the manuscript.

Acknowledgements

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. Calculations were carried out using infrastructure of the Ziemowit computer cluster (www.ziemowit.hpc.polsl.pl) in the Laboratory of Bioinformatics and Computational Biology, The Biotechnology, Bioengineering and Bioinformatics Centre Silesian BIO-FARMA, created in the POIG.02.01.00-00-166/08 and expanded in the POIG.02.03.01-00-040/13 projects.

References

1. Petryk N, Kahli M, d'Aubenton-Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, et al. Replication landscape of the human genome. Nature communications. 2016;7:10208.
2. Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, et al. Human gene organization driven by the coordination of replication and transcription. Genome research. 2007;17(9):1278-85.
3. Cooper G, Hausman R. The Cell: A Molecular Approach, Sixth Edition. Sunderland: Sinauer Associates; 2013.
4. Masai H, Matsumoto S, You Z, Yoshizawa-Sugata N, Oda M. Eukaryotic chromosome DNA replication: where, when, and how? Annual review of biochemistry. 2010;79:89-130.

587 5. Marahrens Y, Stillman B. A yeast chromosomal origin of DNA replication defined by multiple functional
588 elements. Science. 1992;255(5046):817-23.
589 6. Oliver JL, Carpena P, Hackenberg M, Bernaola-Galvan P. IsoFinder: computational prediction of isochores
590 in genome sequences. Nucleic acids research. 2004;32(Web Server issue):W287-92.
591 7. Fragkos M, Ganier O, Coulombe P, Mechali M. DNA replication origin activation in space and time.
592 Nature reviews Molecular cell biology. 2015;16(6):360-74.
593 8. DePamphilis ML. Origins of DNA replication in metazoan chromosomes. J Biol Chem. 1993;268(1):1-4.
594 9. Gilbert DM. Replication origin plasticity, Taylor-made: inhibition vs recruitment of origins under
595 conditions of replication stress. Chromosoma. 2007;116(4):341-7.
596 10. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, et al. Sequencing newly
597 replicated DNA reveals widespread plasticity in human replication timing. Proceedings of the National
598 Academy of Sciences of the United States of America. 2010;107(1):139-44.
599 11. Gao F, Luo H, Zhang CT. DeOri: a database of eukaryotic DNA replication origins. Bioinformatics.
600 2012;28(11):1551-2.
601 12. Hyrien O. Peaks cloaked in the mist: the landscape of mammalian replication origins. The Journal of cell
602 biology. 2015;208(2):147-60.
603 13. Cvetic C, Walter JC. Eukaryotic origins of DNA replication: could you please be more specific? Seminars in
604 cell & developmental biology. 2005;16(3):343-53.
605 14. Mechali M. Eukaryotic DNA replication origins: many choices for appropriate answers. Nature reviews
606 Molecular cell biology. 2010;11(10):728-38.
607 15. Delgado S, Gomez M, Bird A, Antequera F. Initiation of DNA replication at CpG islands in mammalian
608 chromosomes. Embo Journal. 1998;17(8):2426-35.
609 16. Vassilev L, Johnson EM. An initiation zone of chromosomal DNA replication located upstream of the c-
610 myc gene in proliferating HeLa cells. Mol Cell Biol. 1990;10(9):4899-904.
611 17. Keller C, Ladenburger EM, Kremer M, Knippers R. The origin recognition complex marks a replication
612 origin in the human TOP1 gene promoter. J Biol Chem. 2002;277(35):31430-40.

613 18. Abdurashidova G, Deganuto M, Klima R, Riva S, Biamonti G, Giacca M, et al. Start sites of bidirectional
614 DNA synthesis at the human lamin B2 origin. Science. 2000;287(5460):2023-6.
615 19. Theis JF, Newlon CS. The ARS309 chromosomal replicator of Saccharomyces cerevisiae depends on an
616 exceptional ARS consensus sequence. Proceedings of the National Academy of Sciences of the United
617 States of America. 1997;94(20):10786-91.
618 20. Berbenetz NM, Nislow C, Brown GW. Diversity of eukaryotic DNA replication origins revealed by genome-
619 wide analysis of chromatin structure. PLoS genetics. 2010;6(9):e1001092.
620 21. Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM. Conserved nucleosome positioning defines
621 replication origins. Genes Dev. 2010;24(8):748-53.
622 22. Breier AM, Chatterji S, Cozzarelli NR. Prediction of Saccharomyces cerevisiae replication origins. Genome
623 biology. 2004;5(4):R22.
624 23. Nieduszynski CA, Knox Y, Donaldson AD. Genome-wide identification of replication origins in yeast by
625 comparative genomics. Genes Dev. 2006;20(14):1874-9.
626 24. Wang L, Lin CM, Lopreiato JO, Aladjem MI. Cooperative sequence modules determine replication
627 initiation sites at the human beta-globin locus. Human molecular genetics. 2006;15(17):2613-22.
628 25. Akerman I, Kasaai B, Bazarova A, Sang PB, Peiffer I, Artufel M, et al. A predictable conserved DNA base
629 composition signature defines human core DNA replication origins. Nature communications.
630 2020;11(1):4826.
631 26. Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, et al. Unraveling cell type-specific
632 and reprogrammable human replication origin signatures associated with G-quadruplex consensus
633 motifs. Nature structural & molecular biology. 2012;19(8):837-44.
634 27. Courbet S, Gay S, Arnoult N, Wronka G, Anglana M, Brison O, et al. Replication fork movement sets
635 chromatin loop size and origin choice in mammalian cells. Nature. 2008;455(7212):557-60.
636 28. Wilson RHC, Coverley D. Relationship between DNA replication and the nuclear matrix. Genes Cells.
637 2013;18(1):17-31.

638 29. Shinbrot E, Henninger EE, Weinhold N, Covington KR, Goksenin AY, Schultz N, et al. Exonuclease
639 mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human
640 origins of replication. Genome research. 2014;24(11):1740-50.
641 30. Lujan SA, Williams JS, Kunkel TA. DNA Polymerases Divide the Labor of Genome Replication. Trends in
642 cell biology. 2016;26(9):640-54.
643 31. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, et al. Mutational Strand
644 Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell. 2016;164(3):538-
645 49.
646 32. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational
647 processes in human cancer. Nature. 2013;500(7463):415-21.
648 33. Koren A, Handsaker RE, Kamitaki N, Karlic R, Ghosh S, Polak P, et al. Genetic Variation in Human DNA
649 Replication Timing. Cell. 2014;159(5):1015-26.
650 34. Arezi B, Xing W, Sorge JA, Hogrefe HH. Amplification efficiency of thermostable DNA polymerases.
651 Analytical biochemistry. 2003;321(2):226-35.
652 35. Wu X, Kabalane H, Kahli M, Petryk N, Laperrousaz B, Jaszczyszyn Y, et al. Developmental and cancer-
653 associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-
654 replicating regions. Nucleic acids research. 2018;46(19):10532.
655 36. Standage DS, Brendel VP. ParsEval: parallel comparison and analysis of gene structure annotations. BMC
656 bioinformatics. 2012;13:187.
657 37. Zytnicki M, Luo Y, Quesneville H. Efficient comparison of sets of intervals with NC-lists. Bioinformatics.
658 2013;29(7):933-9.
659 38. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple
660 sequence alignment through sequence weighting, position-specific gap penalties and weight matrix
661 choice. Nucleic acids research. 1994;22(22):4673-80.

662 39. Zehnbauer B, Vogelstein B. Supercoiled loops and the organization of replication and transcription in
663 eukaryotes. BioEssays : news and reviews in molecular, cellular and developmental biology.
664 1985;2(2):52–4.
665 40. Kojima KK. Human transposable elements in Repbase: genomic footprints from fish to humans. Mobile
666 DNA. 2018;9:2.
667 41. Saegusa Y, Sato M, Galli I, Nakagawa T, Ono N, Iguchi-Ariga SM, et al. Stimulation of SV40 DNA
668 replication and transcription by Alu family sequence. Biochimica et biophysica acta. 1993;1172(3):274-
669 82.
670 42. Marsico G, Chambers VS, Sahakyan AB, McCauley P, Boutell JM, Antonio MD, et al. Whole genome
671 experimental maps of DNA G-quadruplexes in multiple species. Nucleic acids research. 2019;47(8):3862-
672 74.
673 43. Long H, Zhang L, Lv M, Wen Z, Zhang W, Chen X, et al. H2A.Z facilitates licensing and activation of early
674 replication origins. Nature. 2020;577(7791):576-81.
675 44. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human
676 genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80.
677 45. Narwade N, Patel S, Alam A, Chattopadhyay S, Mittal S, Kulkarni A. Mapping of scaffold/matrix
678 attachment regions in human genome: a data mining exercise. Nucleic acids research. 2019;47(14):7247-
679 61.
680 46. Shlien A, Campbell BB, de Borja R, Alexandrov LB, Merico D, Wedge D, et al. Combined hereditary and
681 somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers.
682 Nature genetics. 2015;47(3):257-62.
683 47. Rein T, Kobayashi T, Malott M, Leffak M, DePamphilis ML. DNA methylation at mammalian replication
684 origins. J Biol Chem. 1999;274(36):25792-800.
685 48. Campbell BB, Light N, Fabrizio D, Zatzman M, Fuligni F, de Borja R, et al. Comprehensive Analysis of
686 Hypermutation in Human Cancer. Cell. 2017;171(5):1042-56 e10.

687 49. Martin MM, Ryan M, Kim R, Zakas AL, Fu H, Lin CM, et al. Genome-wide depletion of replication initiation
688 events in highly transcribed regions. Genome research. 2011;21(11):1822-32.
689 50. Picard F, Cadoret JC, Audit B, Arneodo A, Alberti A, Battail C, et al. The spatiotemporal program of DNA
690 replication is associated with specific combinations of chromatin marks in human cells. PLoS genetics.
691 2014;10(5):e1004282.
692 51. Langley AR, Graf S, Smith JC, Krude T. Genome-wide identification and characterisation of human DNA
693 replication origins by initiation site sequencing (ini-seq). Nucleic acids research. 2016;44(21):10230-47.
694 52. Mesner LD, Valsakumar V, Cieslik M, Pickin R, Hamlin JL, Bekiranov S. Bubble-seq analysis of the human
695 genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins.
696 Genome research. 2013;23(11):1774-88.
697 53. Mukhopadhyay R, Lajugie J, Fourel N, Selzer A, Schizas M, Bartholdy B, et al. Allele-specific genome-wide
698 profiling in human primary erythroblasts reveal replication program organization. PLoS genetics.
699 2014;10(5):e1004319.
700 54. Fu H, Besnard E, Desprat R, Ryan M, Kahli M, Lemaitre JM, et al. Mapping replication origin sequences in
701 eukaryotic chromosomes. Current protocols in cell biology. 2014;65:22 0 1-17.
702 55. Zhao PA, Sasaki T, Gilbert DM. High-resolution Repli-Seq defines the temporal choreography of initiation,
703 elongation and termination of replication in mammalian cells. Genome biology. 2020;21(1):76.
704 56. Chen YH, Keegan S, Kahli M, Tonzi P, Fenyo D, Huang TT, et al. Transcription shapes DNA replication
705 initiation and termination in human cells. Nature structural & molecular biology. 2019;26(1):67-77.
706 57. Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, Segalla S, et al. Genome-wide mapping of human DNA-
707 replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing.
708 Genome research. 2013;23(1):1-11.
709 58. Miotto B, Ji Z, Struhl K. Selectivity of ORC binding sites and the relation to replication timing, fragile sites,
710 and deletions in cancers. Proceedings of the National Academy of Sciences of the United States of
711 America. 2016;113(33):E4810-9.

712 59. Sugimoto N, Maehara K, Yoshida K, Ohkawa Y, Fujita M. Genome-wide analysis of the spatiotemporal
713 regulation of firing and dormant replication origins in human cells. Nucleic acids research.
714 2018;46(13):6683-96.
715 60. Ryba T, Hiratani I, Sasaki T, Battaglia D, Kulik M, Zhang J, et al. Replication timing: a fingerprint for cell
716 identity and pluripotency. Plos Comput Biol. 2011;7(10):e1002225.
717 61. Mesner LD, Valsakumar V, Karnani N, Dutta A, Hamlin JL, Bekiranov S. Bubble-chip analysis of human
718 origin distributions demonstrates on a genomic scale significant clustering into zones and significant
719 association with transcription. Genome research. 2011;21(3):377-89.
720 62. Karnani N, Taylor CM, Malhotra A, Dutta A. Genomic study of replication initiation in human
721 chromosomes reveals the influence of transcription regulation and chromatin structure on origin
722 selection. Mol Biol Cell. 2010;21(3):393-404.
723 63. Valenzuela MS, Chen Y, Davis S, Yang F, Walker RL, Bilke S, et al. Preferential localization of human
724 origins of DNA replication at the 5'-ends of expressed genes and at evolutionarily conserved DNA
725 sequences. PloS one. 2011;6(5):e17308.
726 64. Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, et al. Genome-wide studies highlight
727 indirect links between human replication origins and gene regulation. Proceedings of the National
728 Academy of Sciences of the United States of America. 2008;105(41):15837-42.
729 65. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers.
730 Bioinformatics. 2009;25(14):1841-2.
731 66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format
732 and SAMtools. Bioinformatics. 2009;25(16):2078-9.
733 67. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics.
734 2010;26(6):841-2.
735 68. Gel B, Diez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor
736 package for the association analysis of genomic regions based on permutation tests. Bioinformatics.
737 2016;32(2):289-91.

69. Charif D, Lobry JR. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural approaches to sequence evolution: Molecules, networks, populations. New York: Springer Verlag; 2007.
70. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research. 2005;15(8):1034-50.
71. Morgan M, Shepherd L. AnnotationHub: Client to access AnnotationHub resources. R package version 2220. 2020.
72. Ho L, Crabtree GR. Chromatin remodelling during development. Nature. 2010;463(7280):474-84.
73. Jaksik R, Rzeszowska-Wolny J. The distribution of GC nucleotides and regulatory sequence motifs in genes and their adjacent sequences. Gene. 2012;492(2):375-81.