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

Human origins of replication (ORI) are recognized by the cellular machinery through mechanisms which are still poorly understood. The process of DNA replication is highly conserved, but the location of ORI may vary considerably from one round of replication to the next, and depends on many factors, making them difficult to map in the genome. We investigated the possibility that at genomic scale, the mutator phenotype associated with DNA polymerase ε exonuclease domain mutation could identify genomic positions of replication origins. Here we report the genome-wide localization of DNA replication origins in POLE-mutated tumors using whole genome sequencing data from The Cancer Genome Atlas project. We show that mutation-based detection of replication origins allows to identify constitutive origins shared between various individuals.

We developed a novel method to compare two or more sets of genomic coordinates based on Smith–Waterman-like dynamic programming, which we used to compare replication origin positions obtained using multiple different methods. The comparison allowed us to define a consensus set of replication origins, identified consistently by multiple ORI detection methods. Many DNA features co-localized with 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.

Introduction

DNA replication origins are crucial for initiation of the DNA synthesis, guiding the recruitment of proteins that form the pre-replication complex (pre-RC). Subsequent recruitment of helicases leads to creation of a replication bubble, making the DNA accessible to polymerases which replicate DNA in a bidirectional manner (Cooper and Hausman 2013). 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), which in yeast are selected based on sequence specificity, (Marahrens and Stillman 1992; Masai et al. 2010); however in humans the

recognition mechanism utilizes various DNA characteristics (Fragkos et al. 2015), and only a limited number of origins are active at each cell cycle (DePamphilis 1993). Efficiency of activation of the origins is used to classify them into three categories: constitutive, flexible and dormant (Fragkos et al. 2015). 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 in from one cell cycle to the next, or one cell type to the next (Fragkos et al. 2015). Dormant origins become active in stress conditions that affect the S phase, including serum starvation and DNA damage (Gilbert 2007). 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 (Hansen et al. 2010; Gao et al. 2012; Hyrien 2015) 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 (Cvetic and Walter 2005; Fragkos et al. 2015). 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 (Mechali 2010). Additionally, a low methylation level is an important factor, making some promoter regions suitable targets for ORC binding (Delgado et al. 1998). As a result. several of the best studied ORI are in the vicinity of TSS of known genes such as *MYC*, *TOP1* and *LMNB2* (Vassilev and Johnson 1990; Abdurashidova et al. 2000; Keller et al. 2002).

Despite the similarities to TSS there exists no definite evidence of the existence of specific sequence motifs required for ORC binding. However, in some unicellular eukaryotic genomes *Cis*-acting sequences determine the location of replication origins (Masai et al. 2010) and in yeast two sequence elements are necessary: a 17-bp autonomously replicating consensus sequence (ACS) that binds Origin Recognition Complex (ORC): WWWWTTTAYRTTTWGTT (Theis and Newlon 1997) and a broader sequence context encompassing 200 to 300 bp that appears to be important for depleting nucleosomes from the origin (Breier et al. 2004; Nieduszynski et al. 2006; Berbenetz et al.

2010; Eaton et al. 2010). In human cells, high AT content (Wang et al. 2006) and G-quadruplexes (Besnard et al. 2012; Hoshina et al. 2013) 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 (Courbet et al. 2008), and also it is believed that DNA replication is initiated in regions attached to the nuclear matrix (MARs) (Wilson and Coverley 2013).

In the present work we employed, at the genomic scale, Shinbrot et al. (Shinbrot et al. 2014) finding to map the DNA origins of replication, using mutations identified in whole genome sequencing experiments. Additionally we have shown that the mutation pattern also covers other context-dependent mutation types. Using this information, we developed a method for the detection of constitutive origins of replication - MutORI, 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 (Lujan et al. 2016), 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 (Shinbrot et al. 2014; Haradhvala et al. 2016). 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 (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 ((Alexandrov et al. 2013) 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 (MutORI) 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.

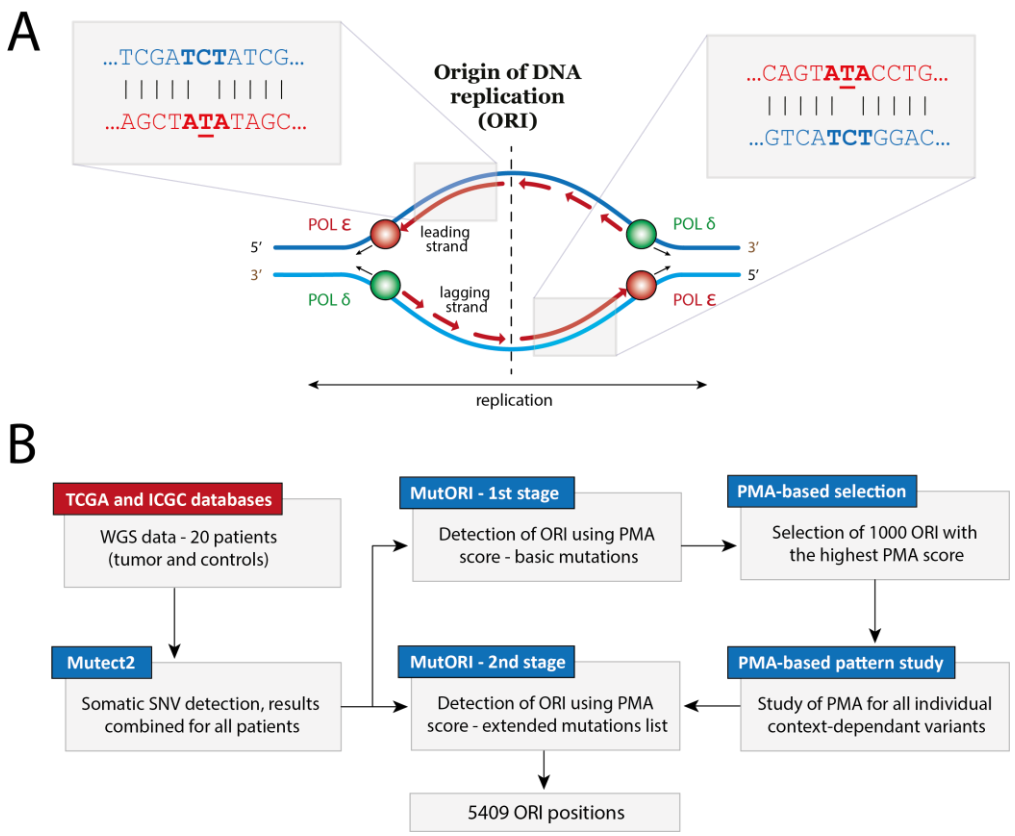

Fig. 1: A) DNA replication in human cells. Polymerase ε synthesizes the leading strand while polymerase δ is responsible for the synthesis of the lagging strand (Lujan et al. 2016). Both polymerases start the synthesis from the replication origin (ORI) after replication is initiated by the origin recognition 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 – MutORI, 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 422 to 606 thousand per sample, also increasing the number of identified replication origins from 5132 to 5409.

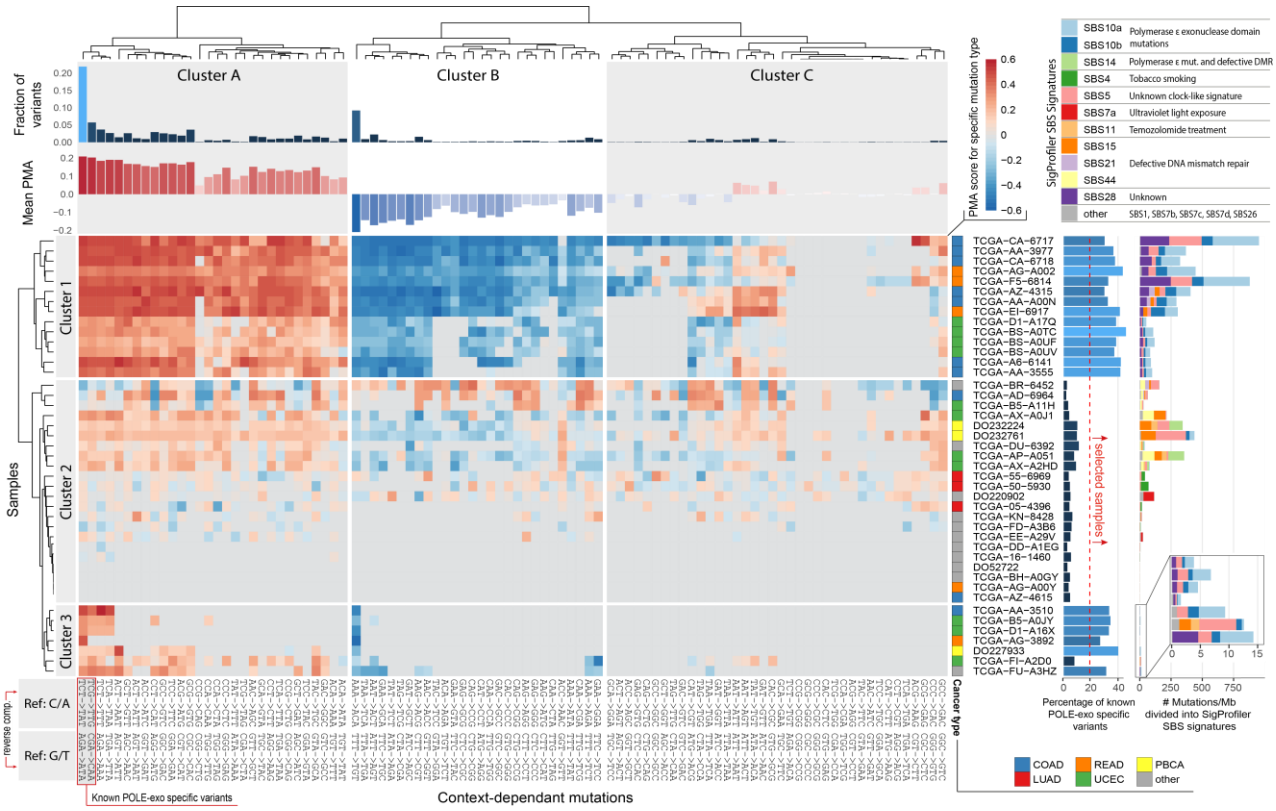

Fig.2: Heatmap of the PMA score calculated at ORI positions identified using known POLE-exo specific variants (marked on the plot). Samples (rows) and mutation types (columns), both clustered into 3 groups. Columns were clustered using only samples in which percentage of POLE-exo specific variants exceeded 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.

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 (Koren et al. 2014), or the polymerase efficiency which, among other things, is known to be GC content-dependent (Arezi et al. 2003).

Locations of the replication origins are reported to be associated with DNA features that are non-randomly distributed over the entire genome (Fragkos et al. 2015). 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.

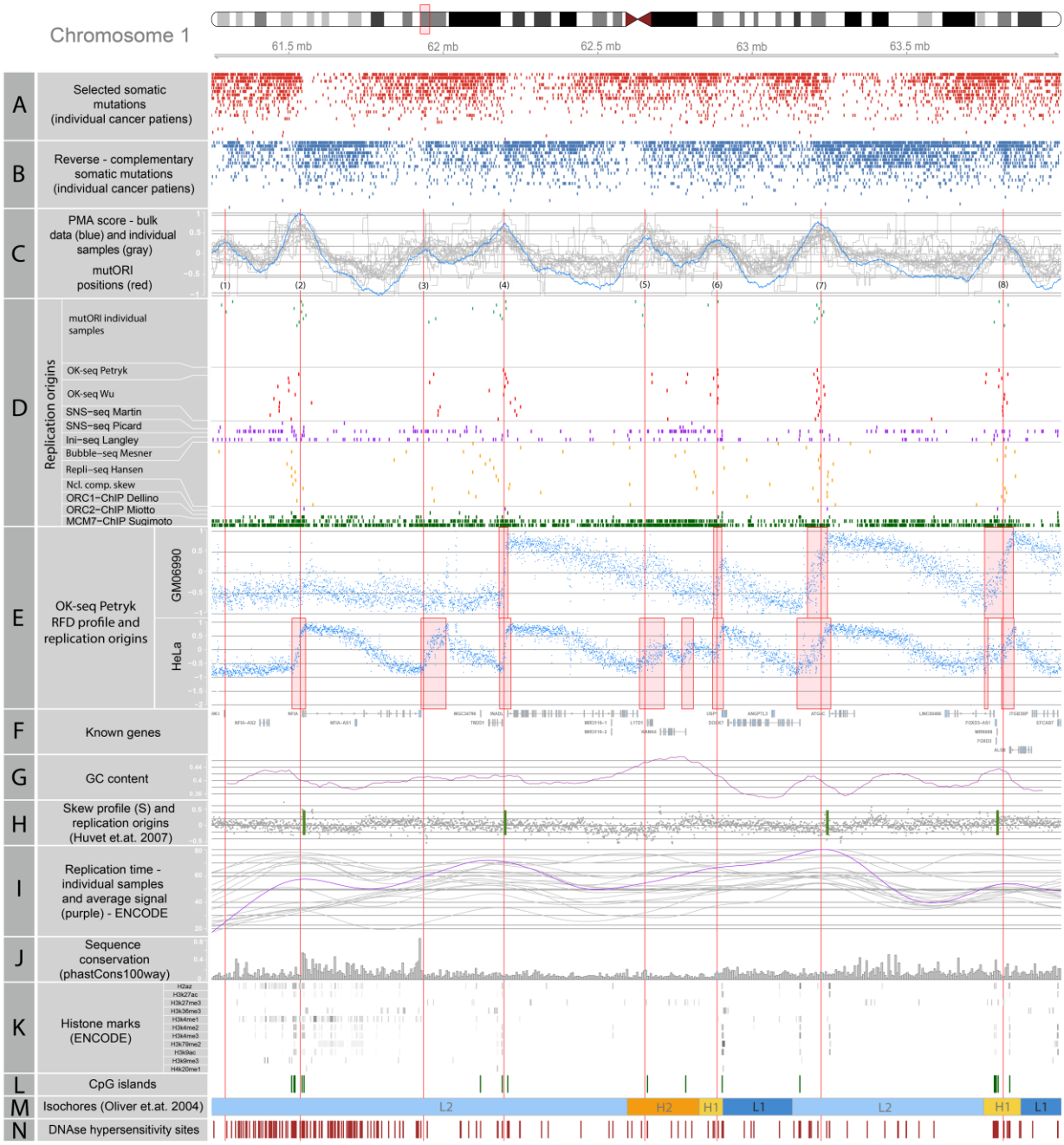

Fig.3: Replication origins identified in a fragment of chromosome 1. **A)** Individual context dependent somatic mutations used in the ORI detection algorithm (C/A reference allele from cluster A and G/T from cluster B, as shown on Fig.2), each row represents one individual patient; **B)** mutations 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 MutOri numbered (1)-(8); **D)** Replication origins identified by other NGS-based methods, (see Supp Table 1) in other samples from other tissues; each row represents one sample; **E)** RFD profile (blue dots) for two OK-seq samples from (Petryk et al. 2016), 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 (Huvet et al. 2007); **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 (Oliver et al. 2004), lowest GC - L1,L2,H1,H2,H3 – highest GC; **N)** DNAse hypersensitivity sites obtained for K562 cell line (ENCODE data)

184

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) (Petryk et al. 2016; Wu et al. 2018). 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 (Wu et al. 2018). 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 (Standage and Brendel 2012; Zytnicki et al. 2013). 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 two sets of numeric vectors, instead of nucleotide sequences, and used it to pair replication origin positions and compare the overall distances between the two sets. Instead of rewards for match and mismatch we used the absolute distance between locations requiring only a gap penalty to control the number of gaps introduced to the compared vectors (see Materials and Methods), which represent unmatched positions, specific to one method only. We also extended the

algorithm to a version that allows to compare multiple sets of positions, which works similarly to ClustalW multiple sequence alignment (Thompson et al. 1994).

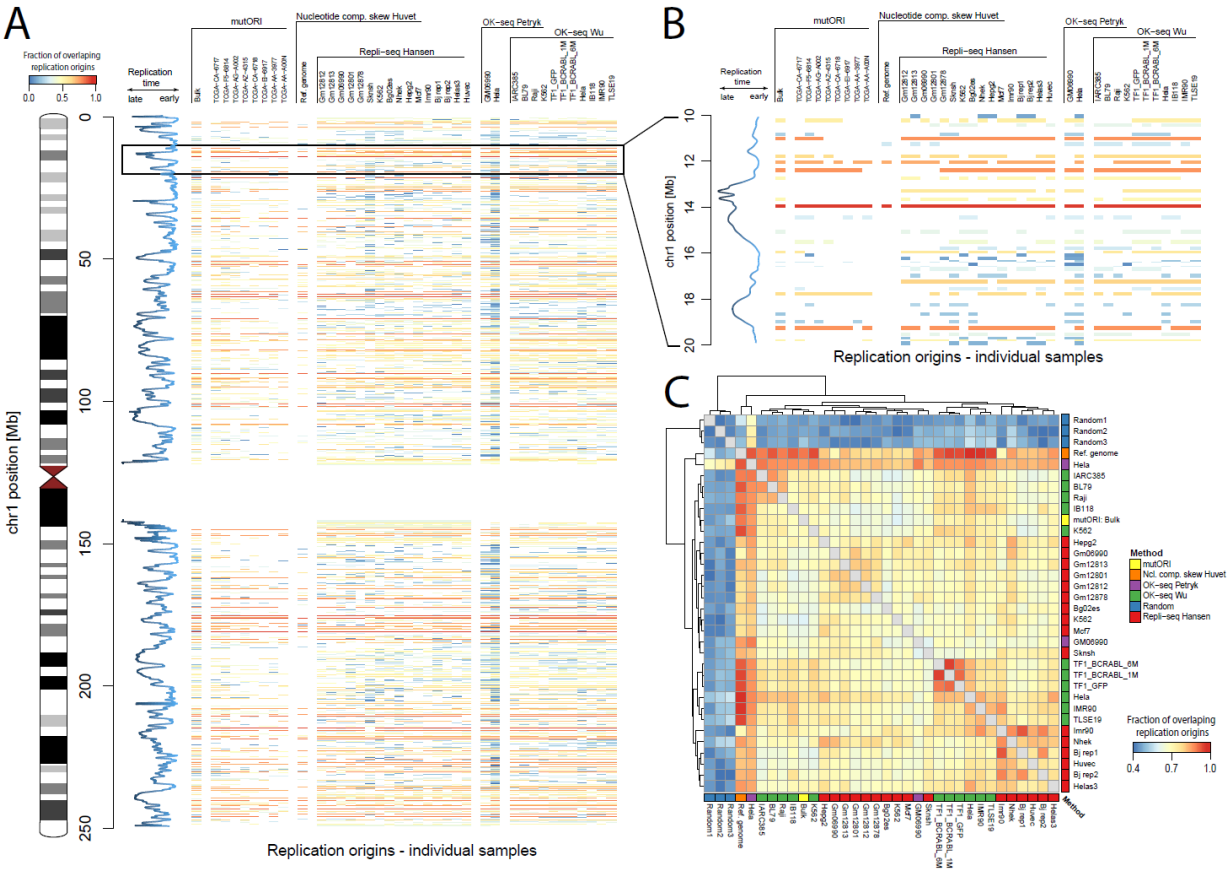

Fig.4: Similarity between various replication origin detection methods. **A**) results obtained using the multiple vector alignment algorithm for chromosome 1, each row represents one origin and the color scale corresponds to the fraction of methods in which it was detected; blue line represents the 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 random genomic positions for comparison. The correlations are obtained using a dynamic programming-based vector alignment algorithm (see Methods).

We used both algorithms, named 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 (Fig.4). Panel A and B show the results of MNVA conducted on replication origins we identified, based on mutation patterns in individual samples with a large number of mutations, as well as consensus positions using 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. Positions of all

compared origins combined using the MNVA algorithm are available as the supplementary table S3. Fig. 4C shows the results of 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). Our approach is similar to the method developed by Huvet et al. and to the OK-seq approaches, while the lowest, as expected, to the three sets of randomly generated genomic positions.

DNA features associated with ORI

The MutORI 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 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 (Fragkos et al. 2015), especially those which are active in a given cell (Zehnbauer and Vogelstein 1985). 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 (Kojima 2018) 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 (Saegusa et al. 1993; Hansen et al. 2010).

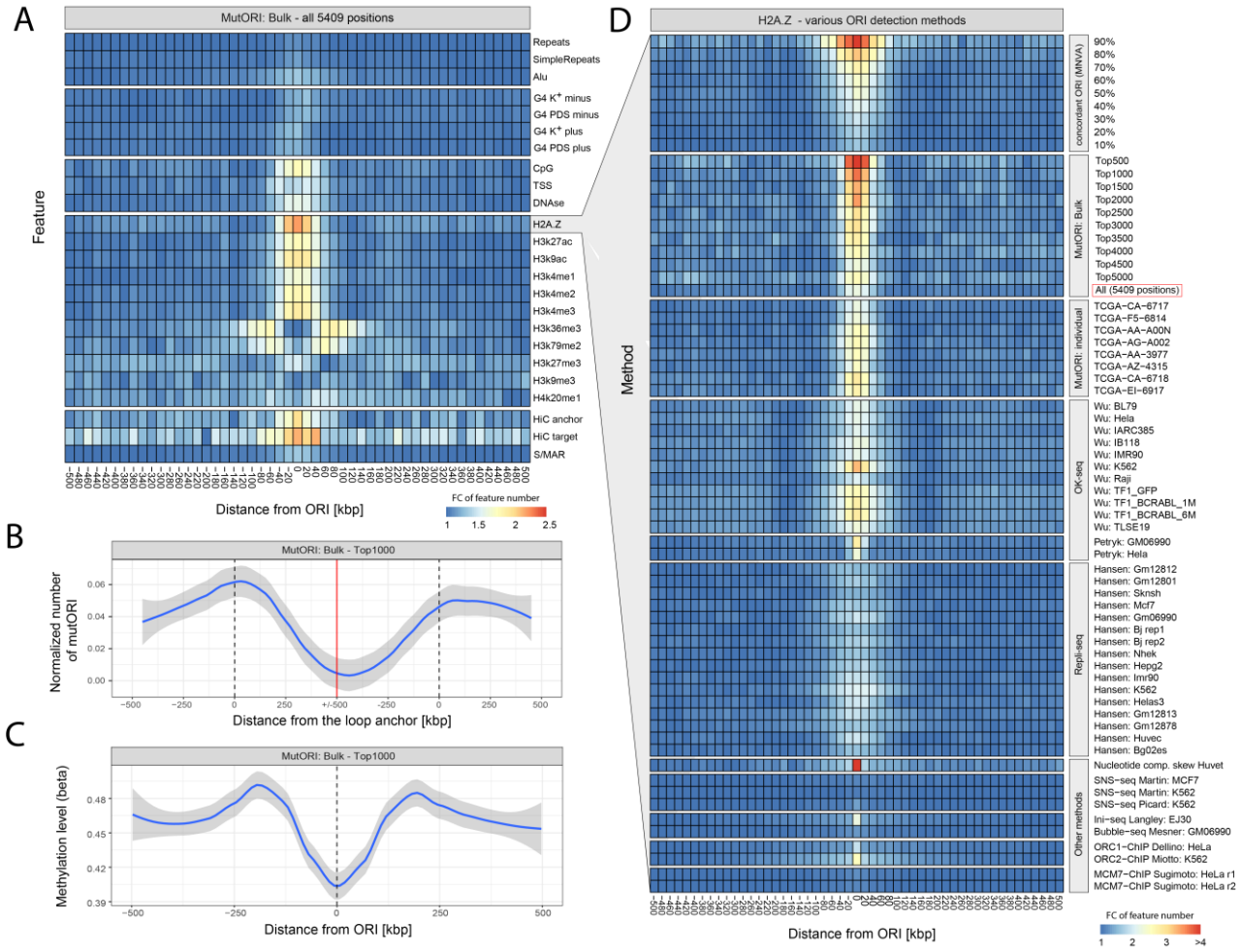

Fig.5: Correlation of MutORI sites with epigenetic features of the genome. **A)** Number of specific features at a given distance from the replication origins, expressed as the fold change with respect to the minimum value from each row. **B)** Number of identified ORI at a specific distance from the 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.

G-quadruplexes (G4) were previously reported as being overrepresented in the vicinity of replication origins (Besnard et al. 2012; Hoshina et al. 2013). To test this association we used G4 positions, identified in the (Marsico et al. 2019) 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 (Berbenetz et al. 2010; Eaton et al. 2010). 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 MutORI. Long has recently shown H2A.Z histone to epigenetically regulate the licensing and activation of early replication origins (Long et al. 2020). Interestingly, our data show that H3k36me3 and H3k79me2 exhibit depletion at the MutORI 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 (Courbet et al. 2008). To validate this feature we compared the locations of chromatin loops (Rao et al. 2014), with the locations of the highest scoring 1000 MutORI 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 (Narwade et al. 2019), and replication origins which were previously reported to have possible association (Wilson and Coverley 2013).

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 (Shlien et al. 2015)), 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 (Rein et al. 1999)) affecting the patient specificity of

mutation patterns. We compared the methylation levels, determined in the TCGA project, in the vicinity of MutORI 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.

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. Figure 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 MutORI method obtained for individual samples (shown on the “MutORI: individual” panel) since the data used to derive them were also used in the combined sample study (MutORI: consensus). 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 MutORI 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 MutORI 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 the 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 on leading strand near ORI. We formalized this idea by constructing, the POLE-exo mutation asymmetry (PMA) score, which assumes maximum value at the inferred position of ORI.

ORI predicted by PMA score are consistent with some of those obtained by different methods. In many cases, PMA predicted fewer ORI predicted than those obtained using SNS-seq or Ini-seq methods. One reason may be that mutator phenotype is associated with a subset of most commonly activated ORI. 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 the Needleman-Wunsch global nucleotide sequence alignment. Using those findings we determined that our ORI set is most similar to the OK-seq approaches and to the method based on nucleotide composition skew (Huvet et al. 2007). Following this initial check, we compared multiple methods by extending the NVA algorithm to a multiple vector version (MNVA) allowing us to compare multiple sets of genomic positions, using an algorithm similar to ClustalW multiple sequence alignment (Thompson et al. 1994). 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.

Based on replication origins identified using the MutORI approach, and a set of origins which were detected consistently by other methods, we sought statistically significant associations with various DNA properties previously associated with ORI location. We found weak association with repeated sequences and G-quadruplexes, and moderate association with methylation levels and

DNase hypersensitive sites. Replication origins from MutORI were statistically demonstrated to occur in the vicinity of chromatin loop anchors as was suggested previously (Courbet et al. 2008). Week association with S/MARs (Wilson and Coverley 2013) as a potential factor influencing ORI location was also observed. Replication origins identified using MutORI 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 MutORI 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 (Huvet et al. 2007).

The most significant limitation of the MutORI detection method is that it is applicable only to cells characterized by the POLE mutator phenotype, which occurred in ca. 1% of TCGA patients. Most of the mutations in the exonuclease domain of POLE, identified using human cancer cells, were found at the amino acid positions 286 and 411 with other positions being significantly less frequent and in many cases specific to only one patient (see Supplementary Table S1). Not all of the POLE-exo mutants showed an elevated frequency of TCT→TAT and TCG→TTG mutations. This might result from differences in amino acid position of the variant, differences in the tumor age, allelic frequencies (tumor heterogeneity) and the time during tumor development at which the mutation occurred. Additionally, not all of the identified POLE-exo variants were validated, therefore is possible that in some cases the mutations might be false positives. We excluded such potential cases by focusing only on those where the TCT→TAT and TCG→TTG mutation frequency exceeds 20% of all mutations.

Some of the differences between the genomic distribution of replication origins in our study compared to other methods may result from the fact that we analyzed data from TCGA tumors, while other studies rely on immortalized cell lines, some of which are of non-cancerous origin. However we

observed a large concordance between origins detected using cell types from different patients and for this reason 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.

To summarize, the most significant advantage of our replication origin detection method is that not only it can identify replication origins using whole genome sequencing experiments without any modifications, but also classify origin activity based on the value of the PMA score. This allows to narrow down the analysis only to a set of constitutive replication origins. Additionally the replication origin comparison methodology that we propose, based on Needleman-Wunsch global nucleotide sequence alignment algorithm, allows to identify conserved replication origin positions, aiding the process of finding relevant DNA characteristics associated with their location.

Materials and methods

Detection of replication origins based on somatic mutations (MutORI)

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**<u>C</u>**T → T**<u>A</u>**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. Detailed information on the comparison and data sources are described in the supplementary methods.

Table 1. Summary of genome wide replication origin detection methods in human cells

Type	Method	Cells	# Samples [1]	# Orgins/sample	Reference
Based on next generation sequencing	OK-seq	hTERT RPE-1	1 (two cond.)	not determined	(Chen et al. 2019) [6]
		K562; HeLa; …	11	[2,3] 5 374	(Wu et al. 2018)
		HeLa; GM06990	2	9 836; 5 684	(Petryk et al. 2016)
	SNS-seq	K562	1	239 107	(Picard et al. 2014)
		Primary basophilic erythroblasts	1	[3] 266 378	(Mukhopadhyay et al. 2014) [6]
		HCT116	1	100 301	(Fu et al. 2014) [6]
		HeLa; IMR-90; hESC H9; iPSC	4	[3] 236 480	(Besnard et al. 2012) [6]
		K562; MCF7	2	62 971; 94 195	(Martin et al. 2011)
	Repli-seq	HCT116; WA01; WA09	3	n/a	(Zhao et al. 2020)
		K562; HeLa-S3; …	16	[3] 4 812	(Hansen et al. 2010)
	Ini-seq	EJ30	1	25 053	(Langley et al. 2016)
	Bubble-seq	GM06990	1	124 646	(Mesner et al. 2013)
	ORC1-ChIP-seq	HeLa	1	13 626	(Dellino et al. 2013)
	ORC2-ChIP-seq	K562	1	52 251	(Miotto et al. 2016)
	MCM7-ChIP-seq	HeLa	1	[5] 434 207	(Sugimoto et al. 2018)
	Nucleotide comp. skew	Reference genome	1	1 060	(Huvet et al. 2007)
Based on microarrays	Repli-chip	HeLa-S3; GM06990; …	9	n/a	(Ryba et al. 2011) [6]
	Bubble-chip [4]	HeLa; GM06990	2	128; 177	(Mesner et al. 2011) [6]
	NS-BrIP [4]	HeLa	1	815	(Karnani et al. 2010) [6]
	NS-LExo [4]	MCF7; BT-474; H520	3	8 281; 4 432; 3 201	(Valenzuela et al. 2011) [6]
		HeLa	1	320	(Karnani et al. 2010) [6]
		HeLa	1	283	(Cadoret et al. 2008) [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

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 (Thompson et al. 1994), 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.

Software availability

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.

Acknowledgements

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This work was supported by the Polish National Science Centre grant No. 2016/23/D/ST7/03665. 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.

Disclosure Declaration

None.

References

Abdurashidova G, Deganuto M, Klima R, Riva S, Biamonti G, Giacca M, Falaschi A. 2000. Start sites of bidirectional DNA synthesis at the human lamin B2 origin. *Science* **287**(5460): 2023-2026.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500**(7463): 415-421.

Arezi B, Xing W, Sorge JA, Hogrefe HH. 2003. Amplification efficiency of thermostable DNA polymerases. *Analytical biochemistry* **321**(2): 226-235.

Berbenetz NM, Nislow C, Brown GW. 2010. Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS genetics* **6**(9): e1001092.

Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, Marin JM, Lemaitre JM. 2012. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature structural & molecular biology* **19**(8): 837-844.

Breier AM, Chatterji S, Cozzarelli NR. 2004. Prediction of Saccharomyces cerevisiae replication origins. *Genome biology* **5**(4): R22.

Cadoret JC, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, Prioleau MN. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proceedings of the National Academy of Sciences of the United States of America* **105**(41): 15837-15842.

Chen YH, Keegan S, Kahli M, Tonzi P, Fenyo D, Huang TT, Smith DJ. 2019. Transcription shapes DNA replication initiation and termination in human cells. *Nature structural & molecular biology* **26**(1): 67-77.

Cooper G, Hausman R. 2013. *The Cell: A Molecular Approach, Sixth Edition*. Sinauer Associates, Sunderland.

502 Courbet S, Gay S, Arnoult N, Wronka G, Anglana M, Brison O, Debatisse M. 2008. Replication fork movement
503 sets chromatin loop size and origin choice in mammalian cells. *Nature* **455**(7212): 557-560.
504 Cvetic C, Walter JC. 2005. Eukaryotic origins of DNA replication: could you please be more specific? *Seminars*
505 *in cell & developmental biology* **16**(3): 343-353.
506 Delgado S, Gomez M, Bird A, Antequera F. 1998. Initiation of DNA replication at CpG islands in mammalian
507 chromosomes. *Embo Journal* **17**(8): 2426-2435.
508 Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, Segalla S, Cesaroni M, Mendoza-Maldonado R, Giacca M,
509 Pelicci PG. 2013. Genome-wide mapping of human DNA-replication origins: levels of transcription at
510 ORC1 sites regulate origin selection and replication timing. *Genome research* **23**(1): 1-11.
511 DePamphilis ML. 1993. Origins of DNA replication in metazoan chromosomes. *J Biol Chem* **268**(1): 1-4.
512 Eaton ML, Galani K, Kang S, Bell SP, MacAlpine DM. 2010. Conserved nucleosome positioning defines
513 replication origins. *Genes Dev* **24**(8): 748-753.
514 Fragkos M, Ganier O, Coulombe P, Mechali M. 2015. DNA replication origin activation in space and time.
515 *Nature reviews Molecular cell biology* **16**(6): 360-374.
516 Fu H, Besnard E, Desprat R, Ryan M, Kahli M, Lemaitre JM, Aladjem MI. 2014. Mapping replication origin
517 sequences in eukaryotic chromosomes. *Current protocols in cell biology* **65**: 22 20 21-17.
518 Gao F, Luo H, Zhang CT. 2012. DeOri: a database of eukaryotic DNA replication origins. *Bioinformatics* **28**(11):
519 1551-1552.
520 Gilbert DM. 2007. Replication origin plasticity, Taylor-made: inhibition vs recruitment of origins under
521 conditions of replication stress. *Chromosoma* **116**(4): 341-347.
522 Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM,
523 Stamatoyannopoulos JA. 2010. Sequencing newly replicated DNA reveals widespread plasticity in
524 human replication timing. *Proceedings of the National Academy of Sciences of the United States of*
525 *America* **107**(1): 139-144.
526 Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE,
527 Braunstein LZ et al. 2016. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of
528 DNA Damage and Repair. *Cell* **164**(3): 538-549.
529 Hoshina S, Yura K, Teranishi H, Kiyasu N, Tominaga A, Kadoma H, Nakatsuka A, Kunichika T, Obuse C, Waga S.
530 2013. Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and
531 single-stranded DNA. *J Biol Chem* **288**(42): 30161-30171.
532 Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, Thermes C. 2007. Human gene
533 organization driven by the coordination of replication and transcription. *Genome research* **17**(9):
534 1278-1285.
535 Hyrien O. 2015. Peaks cloaked in the mist: the landscape of mammalian replication origins. *The Journal of cell*
536 *biology* **208**(2): 147-160.
537 Karnani N, Taylor CM, Malhotra A, Dutta A. 2010. Genomic study of replication initiation in human
538 chromosomes reveals the influence of transcription regulation and chromatin structure on origin
539 selection. *Mol Biol Cell* **21**(3): 393-404.
540 Keller C, Ladenburger EM, Kremer M, Knippers R. 2002. The origin recognition complex marks a replication
541 origin in the human TOP1 gene promoter. *J Biol Chem* **277**(35): 31430-31440.
542 Kojima KK. 2018. Human transposable elements in Repbase: genomic footprints from fish to humans. *Mobile*
543 *DNA* **9**: 2.
544 Koren A, Handsaker RE, Kamitaki N, Karlic R, Ghosh S, Polak P, Eggan K, McCarroll SA. 2014. Genetic Variation
545 in Human DNA Replication Timing. *Cell* **159**(5): 1015-1026.
546 Langley AR, Graf S, Smith JC, Krude T. 2016. Genome-wide identification and characterisation of human DNA
547 replication origins by initiation site sequencing (ini-seq). *Nucleic acids research* **44**(21): 10230-10247.
548 Long H, Zhang L, Lv M, Wen Z, Zhang W, Chen X, Zhang P, Li T, Chang L, Jin C et al. 2020. H2A.Z facilitates
549 licensing and activation of early replication origins. *Nature* **577**(7791): 576-581.
550 Lujan SA, Williams JS, Kunkel TA. 2016. DNA Polymerases Divide the Labor of Genome Replication. *Trends in*
551 *cell biology* **26**(9): 640-654.
552 Marahrens Y, Stillman B. 1992. A yeast chromosomal origin of DNA replication defined by multiple functional
553 elements. *Science* **255**(5046): 817-823.

554 Marsico G, Chambers VS, Sahakyan AB, McCauley P, Boutell JM, Antonio MD, Balasubramanian S. 2019.
555 Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic acids research*
556 **47**(8): 3862-3874.
557 Martin MM, Ryan M, Kim R, Zakas AL, Fu H, Lin CM, Reinhold WC, Davis SR, Bilke S, Liu H et al. 2011. Genome-
558 wide depletion of replication initiation events in highly transcribed regions. *Genome research* **21**(11):
559 1822-1832.
560 Masai H, Matsumoto S, You Z, Yoshizawa-Sugata N, Oda M. 2010. Eukaryotic chromosome DNA replication:
561 where, when, and how? *Annual review of biochemistry* **79**: 89-130.
562 Mechali M. 2010. Eukaryotic DNA replication origins: many choices for appropriate answers. *Nature reviews*
563 *Molecular cell biology* **11**(10): 728-738.
564 Mesner LD, Valsakumar V, Cieslik M, Pickin R, Hamlin JL, Bekiranov S. 2013. Bubble-seq analysis of the human
565 genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins.
566 *Genome research* **23**(11): 1774-1788.
567 Mesner LD, Valsakumar V, Karnani N, Dutta A, Hamlin JL, Bekiranov S. 2011. Bubble-chip analysis of human
568 origin distributions demonstrates on a genomic scale significant clustering into zones and significant
569 association with transcription. *Genome research* **21**(3): 377-389.
570 Miotto B, Ji Z, Struhl K. 2016. Selectivity of ORC binding sites and the relation to replication timing, fragile
571 sites, and deletions in cancers. *Proceedings of the National Academy of Sciences of the United States*
572 *of America* **113**(33): E4810-4819.
573 Mukhopadhyay R, Lajugie J, Fourel N, Selzer A, Schizas M, Bartholdy B, Mar J, Lin CM, Martin MM, Ryan M et
574 al. 2014. Allele-specific genome-wide profiling in human primary erythroblasts reveal replication
575 program organization. *PLoS genetics* **10**(5): e1004319.
576 Narwade N, Patel S, Alam A, Chattopadhyay S, Mittal S, Kulkarni A. 2019. Mapping of scaffold/matrix
577 attachment regions in human genome: a data mining exercise. *Nucleic acids research* **47**(14): 7247-
578 7261.
579 Nieduszynski CA, Knox Y, Donaldson AD. 2006. Genome-wide identification of replication origins in yeast by
580 comparative genomics. *Genes Dev* **20**(14): 1874-1879.
581 Oliver JL, Carpena P, Hackenberg M, Bernaola-Galvan P. 2004. IsoFinder: computational prediction of
582 isochores in genome sequences. *Nucleic acids research* **32**(Web Server issue): W287-292.
583 Petryk N, Kahli M, d'Aubenton-Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, Thermes C, Chen CL, Hyrien O. 2016.
584 Replication landscape of the human genome. *Nature communications* **7**: 10208.
585 Picard F, Cadoret JC, Audit B, Arneodo A, Alberti A, Battail C, Duret L, Prioleau MN. 2014. The spatiotemporal
586 program of DNA replication is associated with specific combinations of chromatin marks in human
587 cells. *PLoS genetics* **10**(5): e1004282.
588 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD,
589 Lander ES et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of
590 chromatin looping. *Cell* **159**(7): 1665-1680.
591 Rein T, Kobayashi T, Malott M, Leffak M, DePamphilis ML. 1999. DNA methylation at mammalian replication
592 origins. *J Biol Chem* **274**(36): 25792-25800.
593 Ryba T, Hiratani I, Sasaki T, Battaglia D, Kulik M, Zhang J, Dalton S, Gilbert DM. 2011. Replication timing: a
594 fingerprint for cell identity and pluripotency. *Plos Comput Biol* **7**(10): e1002225.
595 Saegusa Y, Sato M, Galli I, Nakagawa T, Ono N, Iguchi-Ariga SM, Ariga H. 1993. Stimulation of SV40 DNA
596 replication and transcription by Alu family sequence. *Biochimica et biophysica acta* **1172**(3): 274-282.
597 Shinbrot E, Henninger EE, Weinhold N, Covington KR, Goksenin AY, Schultz N, Chao H, Doddapaneni H, Muzny
598 DM, Gibbs RA et al. 2014. Exonuclease mutations in DNA polymerase epsilon reveal replication strand
599 specific mutation patterns and human origins of replication. *Genome research* **24**(11): 1740-1750.
600 Shlien A, Campbell BB, de Borja R, Alexandrov LB, Merico D, Wedge D, Van Loo P, Tarpey PS, Coupland P,
601 Behjati S et al. 2015. Combined hereditary and somatic mutations of replication error repair genes
602 result in rapid onset of ultra-hypermutated cancers. *Nature genetics* **47**(3): 257-262.
603 Standage DS, Brendel VP. 2012. ParsEval: parallel comparison and analysis of gene structure annotations.
604 *BMC bioinformatics* **13**: 187.

605 Sugimoto N, Maehara K, Yoshida K, Ohkawa Y, Fujita M. 2018. Genome-wide analysis of the spatiotemporal
606 regulation of firing and dormant replication origins in human cells. *Nucleic acids research* **46**(13):
607 6683-6696.
608 Theis JF, Newlon CS. 1997. The ARS309 chromosomal replicator of Saccharomyces cerevisiae depends on an
609 exceptional ARS consensus sequence. *Proceedings of the National Academy of Sciences of the United*
610 *States of America* **94**(20): 10786-10791.
611 Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple
612 sequence alignment through sequence weighting, position-specific gap penalties and weight matrix
613 choice. *Nucleic acids research* **22**(22): 4673-4680.
614 Valenzuela MS, Chen Y, Davis S, Yang F, Walker RL, Bilke S, Lueders J, Martin MM, Aladjem MI, Massion PP et
615 al. 2011. Preferential localization of human origins of DNA replication at the 5'-ends of expressed
616 genes and at evolutionarily conserved DNA sequences. *PloS one* **6**(5): e17308.
617 Vassilev L, Johnson EM. 1990. An initiation zone of chromosomal DNA replication located upstream of the c-
618 myc gene in proliferating HeLa cells. *Mol Cell Biol* **10**(9): 4899-4904.
619 Wang L, Lin CM, Lopreiato JO, Aladjem MI. 2006. Cooperative sequence modules determine replication
620 initiation sites at the human beta-globin locus. *Human molecular genetics* **15**(17): 2613-2622.
621 Wilson RHC, Coverley D. 2013. Relationship between DNA replication and the nuclear matrix. *Genes Cells*
622 **18**(1): 17-31.
623 Wu X, Kabalane H, Kahli M, Petryk N, Laperrousaz B, Jaszczyszyn Y, Drillon G, Nicolini FE, Perot G, Robert A et
624 al. 2018. Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-
625 poor, lowly expressed and late-replicating regions. *Nucleic acids research* **46**(19): 10532.
626 Zehnbauer B, Vogelstein B. 1985. Supercoiled loops and the organization of replication and transcription in
627 eukaryotes. *BioEssays : news and reviews in molecular, cellular and developmental biology* **2**(2): 52–
628 54.
629 Zhao PA, Sasaki T, Gilbert DM. 2020. High-resolution Repli-Seq defines the temporal choreography of
630 initiation, elongation and termination of replication in mammalian cells. *Genome biology* **21**(1): 76.
631 Zytnicki M, Luo Y, Quesneville H. 2013. Efficient comparison of sets of intervals with NC-lists. *Bioinformatics*
632 **29**(7): 933-939.

633