Published online 22 June 2013 *Nucleic Acids Research, 2013, Vol. 41, No. 16* ***7615–7624***
doi:10.1093/nar/gkt552

Loss of heterozygosity preferentially occurs in early replicating regions in cancer genomes

Brent S. Pedersen[1] and Subhajyoti De[1,2,*]

[1]Department of Medicine, University of Colorado School of Medicine, Aurora, CO, USA and [2]Molecular Oncology Program, University of Colorado Cancer Center, Aurora, CO, USA

Received April 22, 2013; Revised May 20, 2013; Accepted May 25, 2013

ABSTRACT

Erroneous repair of DNA double-strand breaks by homologous recombination (HR) leads to loss of heterozygosity (LOH). Analysing 22 392 and 74 415 LOH events in 363 glioblastoma and 513 ovarian cancer samples, respectively, and using three different metrics, we report that LOH selectively occurs in early replicating regions; this pattern differs from the trends for point mutations and somatic deletions, which are biased toward late replicating regions. Our results are independent of *BRCA1* and *BRCA2* mutation status. The LOH events are significantly clustered near RNA polII-bound transcription start sites, consistent with the reports that slow replication near paused RNA polII might initiate HR-mediated repair. The frequency of LOH events is higher in the chromosomes with shorter inter-homolog distance inside the nucleus. We propose that during early replication, HR-mediated rescue of replication near paused RNA polII using homologous chromosomes as template leads to LOH. The difference in the preference for replication timing between different classes of genomic alterations in cancer genomes also provokes a testable hypothesis that replicating cells show changing preference between various DNA repair pathways, which have different levels of efficiency and fidelity, as the replication progresses.

INTRODUCTION

Loss of heterozygosity (LOH) is a common class of genomic alterations observed in cancer genomes, which occurs due to heterozygous deletion of one allele, or duplication of a maternal or paternal chromosome or chromosomal region and concurrent loss of the other allele; the latter is known as copy neutral LOH or uniparental disomy. Copy neutral LOH events arise via homologous recombination (HR)—a DNA double-strand break repair pathway (1). HR is active during and shortly after DNA replication—when sister chromatids and homologous chromosomes are easily available (2). DNA replication is spatially segregated such that some genomic regions are replicated early and others later during S phase (3). It was recently demonstrated that local DNA replication timing (RT) affects the patterns of point mutations (4–6), somatic copy number alterations (4,7,8) and rearrangements (9) in cancer and normal genomes—late replicating regions accumulate more mutations than early replicating regions (10). These findings prompt the question of whether LOH events, which are primarily replication-dependent phenomena, also show distinct patterns in the context of DNA RT.

Here, integrating genomic alteration data for 597 glioblastoma (GBM) (11) and 591 ovarian cystadenoma (12) samples from the cancer genome atlas (TCGA), and DNA RT data for multiple cell types (3), we survey the RT pattern of the genomic regions affected by LOH events, and discuss the findings in the context of the temporal expression pattern of the genes involved in the HR- and non-homologous end-joining (NHEJ)-mediated repair. We then compare and contrast the RT preference for LOH events with that for point mutations and somatic copy number alterations in cancer genomes. We further analyse the findings in the context of factors that are known to contribute to replication stress during early replication, and also the nuclear localization of homologous chromosome pairs. Finally, we conclude by discussing our findings in light of erroneous HR-mediated repair during early replication.

MATERIALS AND METHODS

We mapped all data sets to human reference genome version hg18. Various genomic and epigenomic features were downloaded from the UCSC genome browser (13) as appropriate.

DNA RT data set

We obtained RT data measured using a massively parallel sequencing-based technique across multiple human cell

*To whom correspondence should be addressed. Tel: +1 303 724 6461; Fax: +1 303 724 1799; Email: subhajyoti.de@ucdenver.edu

types from Hansen *et al.* (3). In this study, the RT of different genomic regions was categorized as 'constant early', 'constant mid', 'constant late' and 'variable across cell types'. Some regions had no RT assigned because of coverage, mappability and other technical issues. We focused on genomic regions that had constant early and constant late RT across several human cell types throughout this article. Constant early and constant late RT regions covered 585.13 and 521.14 Mb of the genome, respectively. The remaining regions are termed as 'other_RT' regions.

LOH and other genomic alterations data sets

We have obtained genomic data for 597 GBM (11) and 591 ovarian cystadenoma samples (12) from TCGA. LOH status for the GBM and ovarian cancer samples was analysed using Illumina HumanHapMap550K and Human1MDuo microarrays, respectively, and processed by the Hudson Alpha Institute for Biotechnology using published protocols (11,12). The somatic copy number alteration data for the same samples were obtained from TCGA (11,12). We excluded the samples with potential systematic biases, and also the LOH events that were likely to occur via heterozygous deletion (Supplementary Module SM1), using our previously published approach (14). Our final data set had 22 392 and 74 415 LOH events in 363 GBM and 513 ovarian cancer samples, respectively.

Analytical approach and estimation of statistical significance

We used Bedtools (15) for calculating overlap between two genomic features (e.g. LOH and early replicating regions; getOverlap function) and for estimating intersection between multiple features (multiIntersectBed function). Some genomic regions did not have any RT assigned because of mappability, coverage and other technical issues. Hence, often some LOH end points did not have any RT assigned, but the genomic regions in their proximity did. To maximize biologically relevant overlap between the data sets, we considered a window of 1 kb centering each LOH end point and assigned the RT of that window as the RT of these end points.

We calculated (i) the observed (or expected) proportion of LOH end points in early RT regions as:

$$\frac{\textit{number of LOH end points in RT regions}}{(\textit{number of LOH end points in early RT regions} + \textit{number of LOH end points in late RT regions})}$$

(ii) the observed (or expected) proportion of LOH events with both end points in early RT regions as:

$$\frac{\textit{the number of LOH events with both end points in early RT}}{\left(\begin{array}{l}\textit{the number of LOH with both end points in early RT} \\ \textit{+ both end points in late RT+ one end point in early} \\ \textit{RT and the other in late RT regions}\end{array}\right)}$$

and (iii) the observed (or expected) proportion of the length of LOH events in early RT regions as:

$$\frac{\textit{number of base pairs of LOH overlapping with early RT regions}}{\left(\begin{array}{l}\textit{number of base pairs of LOH overlapping with early} \\ \textit{RT regions + number of base pairs of} \\ \textit{LOH overlapping with early RT regions}\end{array}\right)}$$

We found excluding the LOH end points and stretches of genomic regions affected by LOH events that reside in other_RT regions provides a more meaningful interpretation of the observed preference for early (or late) RT regions and its statistical significance, compared with the cases where other_RT regions were included in the analysis.

We estimated statistical significance of the observed overlaps between LOH and RT patterns using permutation analysis with 10 000 iterations. It was shown that permutation allows preservation of higher-order genomic structures, and hence provides a more realistic *P*-value compared with other statistical tests. During the permutation analysis, we performed genome-wide shuffling using the `shuffleBed` function of the Bedtools (15) with default seed and other parameters, and also keeping the length of the LOH events unchanged. We also used two alternative permutation strategies: `shuffleBed` with the `-chrom` option to permute the LOH events within respective chromosomes, and shuffleBed with the `-chrom` and `-excl` options to permute the LOH events within respective chromosomes, after excluding selected (e.g. centrometric) regions.

Cell cycle-related gene expression

We obtained data on dynamic expression patterns of the genes during the cell cycle from multiple independent experiments in baker's yeast (16–19) and human cell lines (20) as deposited in Cyclebase 2.0 (21). Peak time, periodicity and regulation of these genes were calculated using methods proposed by Gauthier *et al.* (22), and archived in the database. In brief, the P(per) was defined (22) as the chance of observing as great a periodicity by random shuffling of the individual time-point values of the expression profile. First, a Fourier score was obtained for each gene profile. Next, simulated profiles were generated from random shuffling of the data within the original profile 1 million times. The relative proportion of simulated profiles whose Fourier scores were greater than or equal to the gene's true Fourier score was reported as the P(per). Due to the normalization techniques used by Gauthier *et al.* (22), P(per) can take values >1. A small P(per) indicated a highly periodic pattern of expression. If the expression data for a given gene were available from multiple experiments, the P(per) from individual experiments were multiplied to generate the final P(per).

$$F_i = \sqrt{\left(\sum \sin(\omega t)\cdot x_i(t)\right)^2 + \left(\sum \cos(\omega t)\cdot x_i(t)\right)^2}$$
where $\omega = 2\pi/(\text{interdivision time})$

The P(reg) was defined (22) as an estimate that the magnitude of variance between experiments. First, for a given gene the standard deviation was obtained for the log-ratio profile. Then simulated profiles were created from the global distribution for 1 million iterations. The proportion of shuffled profiles whose standard deviations were greater than or equal to the gene's standard deviation was calculated, and normalized to create the final P(reg). Due tothe normalization techniques used by Gauthier *et al*. (22), P(reg) can take values >1. A small *P*-value for regulation indicated low variance and a strongly regulated gene.

Peak time was calculated as a percentage, with both 0 and 100 representing the M/G1 transition phase during the cell cycle. To compute a peak time for a single gene across all available experiments, a sine wave was fitted to the combined expression profile, and the time scale was 'shifted' such that time was represented as a fraction of the cell cycle. In those cases where the expression pattern lacked periodicity at the cell cycle time scale, or the expression pattern between experiments was inconsistent, the peak time was reported as 'Uncertain'.

Genomic and epigenomic features associated with replication stress

We analysed 76 common fragile sites (23), early replicating fragile sites (24), human genes obtained from Ensembl v54 (25), transcription start sites as in Ensembl v 54 (25), and the sites of RNA polII occupancy in GM12878, HUVEC, HeLa and K562 cell lines (13,26,27). Because transcription start sites are a single base-pair wide, we considered a window of ±5 kb while testing for overlap in both observed and expected cases. The regions marked as 'standard peaks' (StdPk.NarrowPk track from the ENCODE/Stanford/Yale/USC/Harvard group) were chosen as the sites of RNA polII occupancy in the four ENCODE cell lines (26,27).

Distance between homologous chromosomes

We obtained the data on the distance between homologous chromosomes in the EJ-30 human epithelial cancer cell line from Heride *et al*. (28). In brief, the authors used fluorescence *in situ* hybridization using advanced microscopy and image analysis tools to analyse in 3D the radial positions of 10 chromosomes (chr1, chr4, chr8 chr10, chr14, chr16, chr17, chr18, chr19 and chr21). Most of the chromosomes occupied specific nuclear positions in the genome and had small variance in inter-homolog distance (28). The nuclear localization and inter-homologous distance estimated in this study were comparable with that estimated in other human cell types (28,29).

RESULTS

Data sets analysed

We used RT data measured using a massively parallel sequencing-based technique across multiple human cell types (3). Some genomic regions replicated early (or late) irrespective of cell types (noted as constant early or constant late RT regions, respectively), whereas others had variable patterns. Throughout this article we focused on the genomic regions that were classified as constant early RT (total length 585.13 Mb) and constant late RT (total length 521.14 Mb).

We obtained the LOH data as available for 597 GBM (11) and 591 ovarian cancer (12) samples from TCGA. We performed extensive quality control steps, excluding the samples with potential systematic biases (e.g. batch effects, low signal to noise ratio), and also the LOH events that were likely to occur via heterozygous deletion (see Methods and Supplementary Module SM1). Our final data set had 22 392 and 74 415 copy neutral LOH events in 363 GBM and 513 ovarian cancer samples, respectively.

Genomic regions affected by LOH events are replicated predominantly early

HR-mediated repair can initiate near one end point of LOH events and proceed unidirectionally to the other end point, or start somewhere between and proceed bidirectionally up to the two end points of the LOH events. To investigate DNA RT patterns of the LOH events after considering these possibilities, we adopted three metrics, analysing DNA RT patterns—(i) at the LOH end points, (ii) over the length of the LOH events and (iii) focusing on only the small (<10 kb) LOH events, which are likely to have the same RT throughout the length.

First metric

To study DNA RT patterns at the LOH end points, we overlaid RT data and the LOH end points from TCGA ovarian cancer samples (12) on the human reference genome (Figure 1A), and found that 40 189 and 21 621 LOH end points occurred in constant early and constant late RT regions, respectively. There were, on average, 0.134 LOH end points per megabase (Mb) per sample in the constant early RT regions, and 0.081 LOH end points per Mb per sample in constant late RT regions in the filtered ovarian cancer data set. We compared the proportion of LOH end points in early (or late) RT regions with that expected by chance using permutation analysis (see Methods for details), and found that the observed preference for LOH end points to occur in the early RT regions was significantly higher compared with that expected by chance (permutation test; *P*-value $<1 \times 10^{-3}$; Figure 1B).

We then repeated the analyses for TCGA GBM samples (11) and found that there were, on average, 0.055 LOH end points per Mb per sample in the constant early RT regions and 0.040 LOH end points per Mb per sample in constant late RT regions in the filtered data set. Once again, a permutation analysis revealed that in GBM samples, LOH end points also preferentially occurred in early RT regions (permutation test; *P*-value $<1 \times 10^{-3}$; Figure 1C).

To examine whether the aggregated patterns are biased by a small number of outlier samples, we repeated the analyses for individual samples. Although small number of LOH events in individual samples made the trends

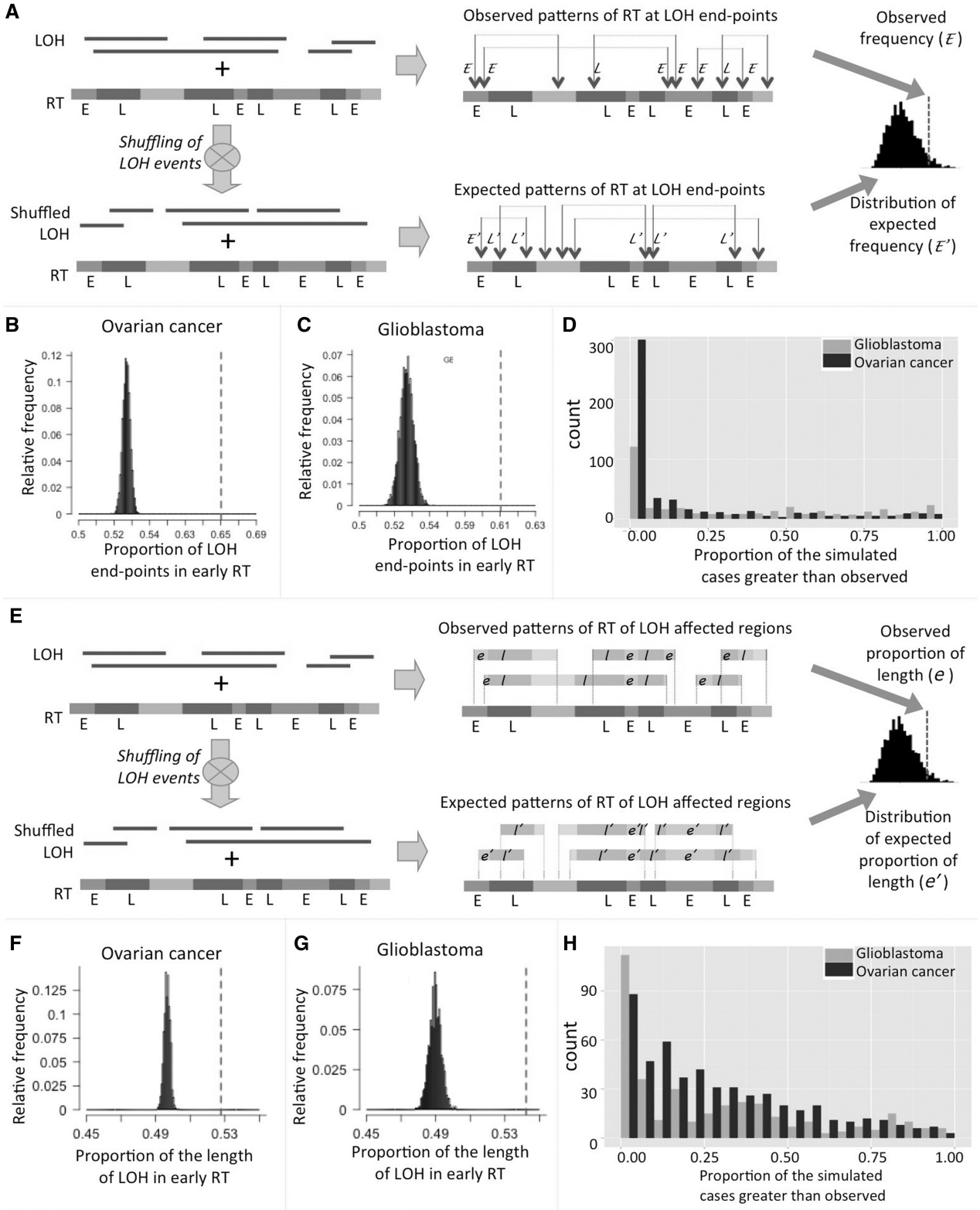

Figure 1. (**A**) A schematic representation showing patterns of DNA RT at the LOH end points. Comparisons between the observed proportions of the (**B**) ovarian cancer and (**C**) GBM LOH end points in early RT regions (dashed vertical bar) with that expected when the LOHs are shuffled across the genome. (**D**) Proportion of the individual ovarian cancer and GBM samples, where observed proportion of LOH end points in early RT regions is higher than that in the shuffled distribution. (**E**) A schematic representation showing overlap between genomic regions affected by LOH events and early RT regions. Comparisons between the observed proportions of the length of the (**F**) ovarian cancer and (**G**) GBM LOH events covered by early RT regions (dashed vertical bar) with that expected when the LOHs are shuffled throughout the genome. (**H**) Proportion of the individual ovarian cancer and GBM samples, where observed proportion of the length of LOH events in early RT regions is higher than that in the shuffled distribution. We also obtained consistent results using alternative permutation approaches, as described in the Supplementary Module SM3.

noisier, we found similar patterns for a majority of the GBM and ovarian cancer samples (Figure 1D)—highlighting that our aggregated results were not due to certain outlier samples.

Next, we calculated how often both the end points of LOH events in TCGA ovarian cancer (12) and GBM (11) samples resided in similar (i.e. both end points in constant early or constant late) or different (i.e. one end point in constant early and the other end point in constant late) RT regions (Methods). We found that the observed proportion of LOH events with early RT at both end points was significantly higher compared with that expected by chance (permutation test; P-value $<1 \times 10^{-3}$) for both the ovarian cancer and GBM data set, and that our aggregated results were not biased by outlier samples (Supplementary Module SM2). Taken together, our findings suggest that LOH end points preferentially occurred in early RT regions.

Second metric

To study RT patterns over the length of the LOH events, we calculated the proportion of the length of the genomic region affected by LOH events that replicated early and those that replicated late during the S phase (see Methods, Figure 1E). We found that the proportion of genomic regions affected by LOH events replicated predominantly early was higher compared with those replicated late, and the trend was statistically significant compared with that expected by chance (permutation test; P-value $<1 \times 10^{-2}$, Figure 1F–G), and that our aggregated results were not due to certain outlier samples (Figure 1H).

Third metric

We then focused on the small (<10 kb) LOH events, the majority of which are likely to have same RT across their length. For both ovarian cancer and GBM samples, using analytical approaches similar to that described previously, we found that the small LOH events were also significantly likely to have early RT at their end points and also over their length, compared with that expected by chance (permutation test; P-value $<1 \times 10^{-3}$).

Finally, we carried out extensive control calculations to account for potential caveats. We performed additional permutation analysis by: (i) randomizing the LOH events only within the same chromosomes, (ii) after excluding centromere regions and (iii) grouping the LOH events as those <1, 1–5 and >5 Mb in size, and in each case found consistent results for both the cancer data sets (Supplementary Module SM3). We found similar results irrespective of the germ line and somatic mutation status at the *BRCA1* and *BRCA2* loci (Supplementary Module SM3). DNA RT is correlated with many genomic and epigenomic features. Integrating chromatin (26), cytogenetic banding patterns (30) and GC content (13) data, we found that our results are consistent even after controlling for these potential covariates (Supplementary Module SM3). Integrating long-range interaction and repeat element data, we found that the two end points of LOH events frequently harbor similar repeat classes, and also are in proximity of each other in the 3D nucleus (Supplementary Module SM3); these attributes might facilitate co-operative HR-mediated repair within the same replication factory, but further studies are warranted. Taken together, our findings suggest that LOH events preferentially occur in early RT regions, and the results are similar across different cancer types, and robust toward the choice of data sets and statistical approaches.

LOH end points have different RT preferences compared with other types of genomic alteration

Different classes of genomic alterations, e.g. point mutations, somatic copy number alterations and LOH arise because of erroneous repair of DNA lesions by various DNA repair pathways. Recently, it was reported that local DNA RT also affects the patterns of point mutations (4–6) and copy number alterations (4,7,8)—point mutations are enriched in late replicating regions, and end points of somatic copy number alterations, especially deletions, occur at a high frequency in late replicating regions (10). Here we reported that, in contrast, the LOH end points selectively occur in early replicating regions in multiple cancer types (Figure 2A). The difference in RT patterns between these distinct classes of genomic alterations led us to ask whether the DNA repair pathways, especially the HR pathway that mediates LOH events (1), also show systematic changes in expression during different phases of the cell cycle.

HR pathway genes are active during early replication

We surveyed the temporal pattern of expression of the genes involved in the canonical DNA double-strand repair pathways during the cell cycle in yeast and humans. We obtained data on the dynamic expression pattern of the genes in the HR (i.e. *RAD50*, *RAD51*, *RAD52*, *RAD54*, *BRCA2*, *XRCC2 XRCC3*, *NBN*, *MRE11*, *MUS81*, *GEN1*, *SHFM1*, *RBBP8*) and NHEJ pathway (i.e. *KU70*, *KU80*, *LIG4*, *HYRC*, *XRCC4*) from multiple independent experiments (16–20) as deposited in the Cyclebase 2.0 (21). We found that mRNA expression of *RAD51* and *RAD54*, which are important for initiation of HR-mediated repair, was high during early replication (G1-S phase) and decreased rapidly afterward (S-G2 phase); the pattern was consistent across independent experiments in both humans and yeast, and showed significant periodicity and low variance (Figure 2B–F; P(per) $<1 \times 10^{-5}$; see Methods for periodicity and variance calculation). Expression of other genes in the HR pathway, or those involved in the NHEJ pathway, did not show distinct cell cycle specific pattern (Supplementary Module SM4). Although we could not examine protein-level expression and post-transcriptional modifications on these genes, the observed findings are consistent with the model that HR-mediated repair is active even during early stages of DNA replication. This is in agreement with the report by Kadyk and Hartwell (1992) that HR-mediated DNA repair using homologous chromosomes leading to LOH can occur during G1 stage of the cell cycle (31). It prompted us to investigate whether certain types of replication stress might trigger HR-mediated repair during

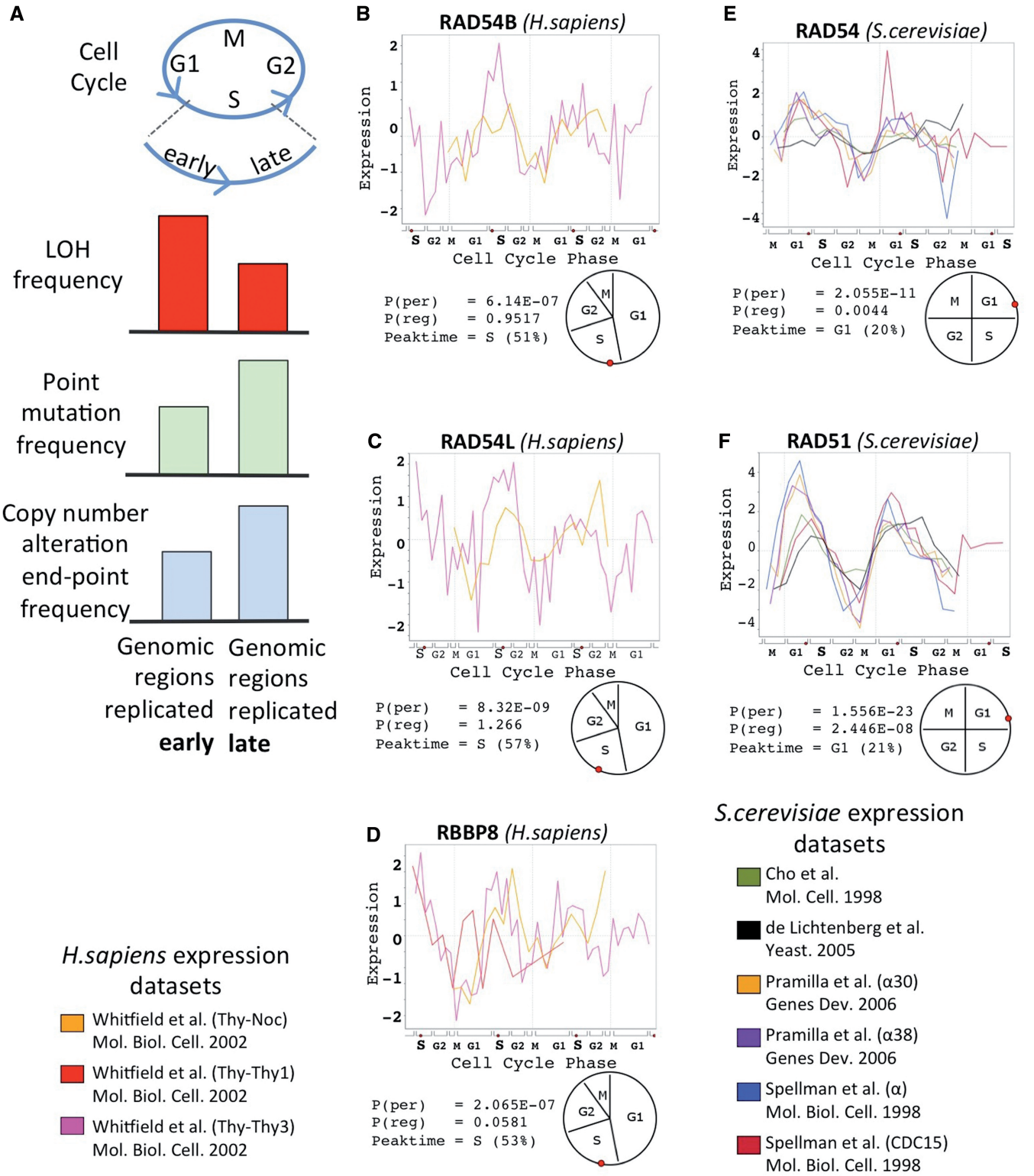

Figure 2. (**A**) Point mutations and copy number alteration (especially deletion) end points are prevalent in late RT regions, but LOH end points are more common in early replicating regions. Temporal patterns of expression of (**B**) *RAD54B*, (**C**) *RAD54L*, (**D**) *RBBP8*, (**E**) *RAD54* and (**F**) *RAD51* during cell cycle in yeast and human cell lines, derived from multiple independent experiments. Measure of periodicity P(per), variance between experiments P(reg) and peak time of expression are listed for each gene, as obtained from CycleBase 2.0 (21). High periodicity and tight regulation are indicated by small values of P(per) and P(reg), respectively.

early replication using homologous chromosomes in the nucleus.

LOH events overlap with sites of high RNA polII occupancy

We next investigated whether certain genomic features, which are commonly associated with replicative stress, also significantly overlap with LOH events (Figure 3A). Common fragile sites are frequent sources of genomic instability (23). We first analysed whether the LOH events significantly overlap with the 76 well-characterized common fragile sites (23) and the newly reported early replicating fragile sites (ERFS) (24), which are implicated in somatic copy number alterations and translocations in lymphoma subtypes, respectively. Interestingly, we did not observe any significant overlap between short (<10 kb; the *third metric*) LOH events and both classes of fragile sites (CFS or ERFS; permutation test; *P*-value $>5 \times 10^{-2}$; Figure 3B) described above. The results were similar using the other two metrics as well.

Even though transcription and replication are meant to be spatially segregated, collision of replication fork with paused RNA polII is another key cause of replicative stress (32). Combining transcription start sites and RNA polII occupancy data from multiple ENCODE cell lines, we found that the sites of high RNA polII occupancy significantly overlapped with transcription start sites (permutation test; *P*-value $<1 \times 10^{-4}$), which is consistent with the reports that promoter-proximal RNA polII pausing is common (33). Integrating LOH data from GBM and ovarian cancer samples, and using the third metric (LOH of size <10 kb), we found that both the transcription start sites of genes and the sites of RNA polII occupancy significantly overlapped with the short (<10 kb) LOH events (permutation test; *P*-value $<1 \times 10^{-3}$; Figure 3B). We did not have spatial resolution in the data sets to test whether RNA polII paused at pre-initiation complex or in early elongation contributed to this pattern. Nevertheless, a vast majority of these genes were expressed in the tumor samples and also in matched normal controls (12). Although the sites of RNA polII occupancy, present in two or more ENCODE cell lines, accounted for <1.5% of the early replicating regions, they overlapped with more than >10% of the short (<10 kb)

A

Feature	Description	Reference
Common fragile sites	Regions showing site-specific gaps and breaks on metaphase chromosomes after partial inhibition of DNA synthesis.	Durkin SG, Glover TW. Annu Rev Genet. 2007;41:169-92.
Early replicating fragile sites	Recurrent, early replicating, and AID-independent DNA lesions arising from replication stress	Barlow JH et al. Cell. 2013 Jan 31;152(3):620-32.
Sites of origin of replication	Peak regions in Repli-seq analysis across multiple ENCODE cell lines	Hansen RS et al. Proc Natl Acad Sci U S A. 2010 Jan 5;107(1):139-44.
Sites of replication termination	Valley regions in Repli-seq analysis across multiple ENCODE cell lines	Hansen RS et al. Proc Natl Acad Sci U S A. 2010 Jan 5;107(1):139-44.
Human genes	Genes as defined by Ensembl v54 (human reference genome hg18)	Flicek P et al. Nucleic Acids Res. 2013 Jan;41(Database issue):D48-55.
Transcription start site	Transcription start sites defined by Ensembl v54 (human reference genome hg18)	Flicek P et al. Nucleic Acids Res. 2013 Jan;41(Database issue):D48-55.
RNA polII bound sites	RNA pol II peaks analysis across multiple ENCODE cell lines	ENCODE project consortium. Nature. 2012 Sep 6;489(7414):57-74.

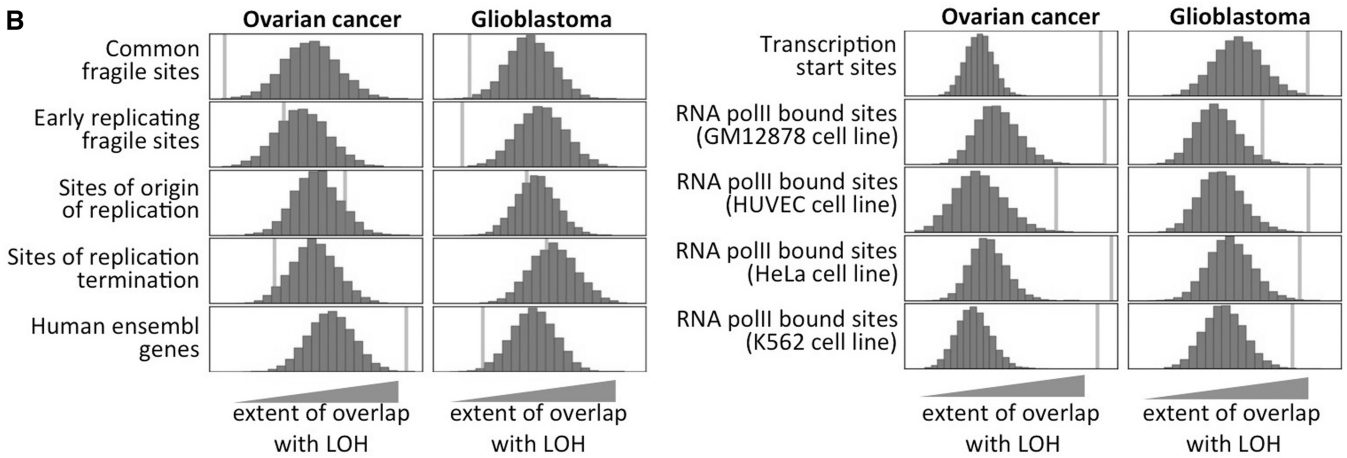

Figure 3. (**A**) Summary of different genomic features analysed in the context of LOH. (**B**) Comparisons of the observed (gray vertical line) extent of overlap between these features with short (<10 kb) LOH events in GBM and ovarian cancer samples, with that expected (light gray bars) when the LOHs are shuffled across the genome.

LOH events in ovarian cancer samples. We found similar evidence for preferential overlap between the sites of RNA polII occupancy and LOH events using the other two metrics as well (Supplementary Module SM5), and even after adjusting for potential covariates such as GC content, gene density and size of the LOH events (permutation test; P-value $<5 \times 10^{-2}$). Even though we could not possibly test every possible covariate or analyse S-phase RNA polII occupancy data from the same samples, consistency of our findings across multiple data sets and analytical approaches hint that these issues are perhaps unlikely to bias our conclusions. Interpreting our findings in the context of recent reports (24,33–36), it is tempting to suggest that during early S phase, replicative stress in the vicinity of RNA polII (37), which are trapped as per-initiation complex or paused in early elongation (38), can invoke HR-mediated repair.

LOH frequency inversely correlates with inter-homologous chromosome distance

Before and during early replication (G1-S phase), homologous chromosomes frequently contribute templates for HR-mediated rescue of replication (31). Eukaryotic chromosomes occupy distinct nuclear territories, such that some pairs of homologous chromosomes (e.g. chr19) are closer to each other than other pairs (e.g. chr4; Figure 4A). Despite cell type-specific variation in nuclear organization, some chromosome pairs have shorter distance than others in the nucleus across different cell types (28,29). We investigated whether the relative frequency of LOH events differed between human chromosomes and whether relative proximity of homologous chromosomes correlated with this pattern. Indeed, overlaying inter-chromosomal distance data (28), we found that the relative frequency of LOH events per chromosome had significantly (Pearson correlation test; P-value $<5 \times 10^{-2}$) inverse correlation with inter-homolog distance in the nucleus (Figure 4B–C) for both GBM and ovarian cancer. Our findings were generally robust toward variation in inter-homologous chromosome distance (Supplementary Module SM6). We also obtained similar results using 3D fluorescence *in situ* hybridization-based inter-chromosome distance data for human fibroblast (29) (Supplementary Module SM6). It is likely that during early replication, when sister chromatids are forming, proximity of homologous chromosome copies is a key factor affecting HR-mediated repair leading to gene conversion and LOH.

DISCUSSION

Taken together, we have demonstrated that (i) LOH events preferentially occur in early RT regions, which is

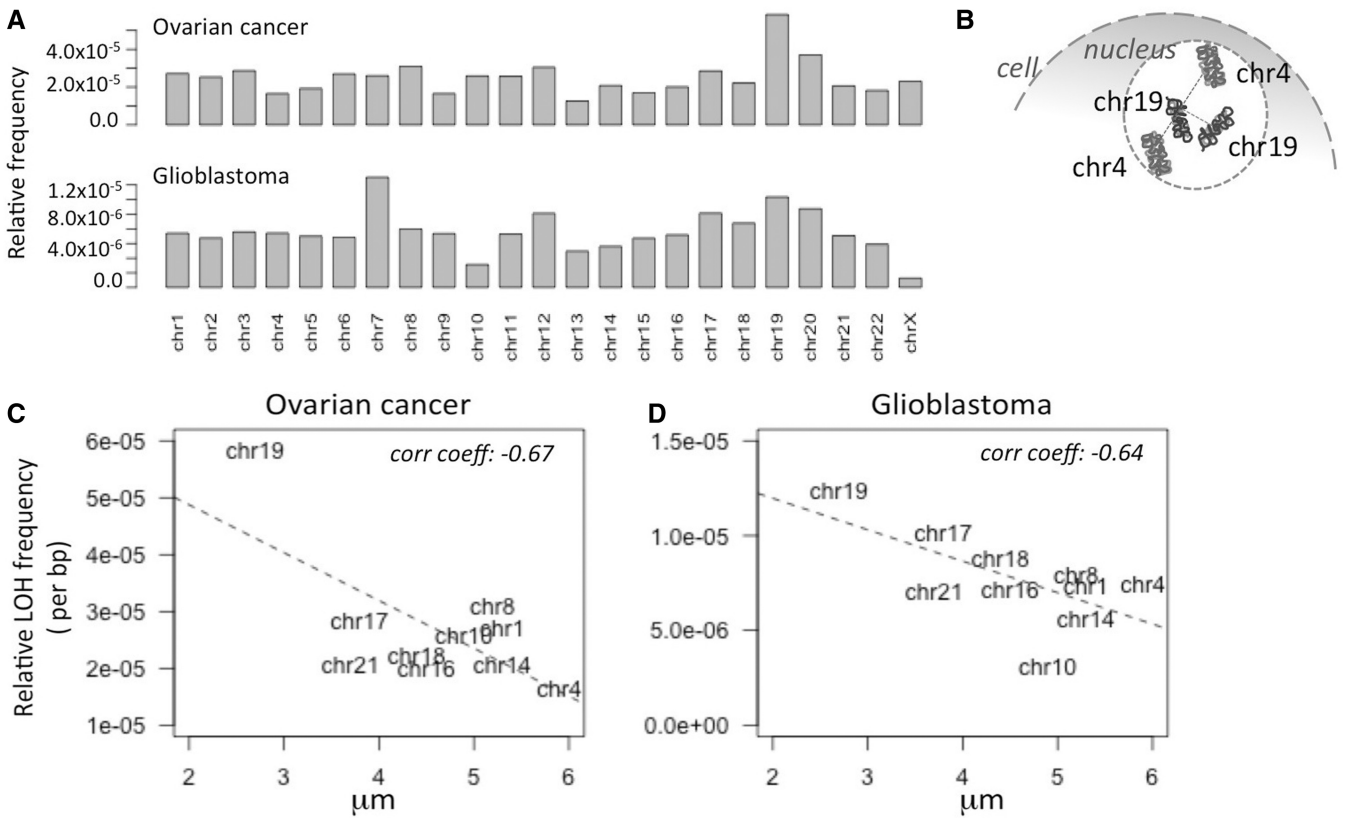

Figure 4. (**A**) The relative frequency of LOH events (per sample, per bp) in different chromosomes for the TCGA ovarian cancer and GBM samples. (**B**) Different chromosomes have different inter-homolog distance inside the nucleus. Scatterplot showing distribution of inter-homolog distance of human chromosomes against the relative frequency of LOH events per chromosome in the (**C**) ovarian cancer and (**D**) GBM samples adjusted by chromosome lengths.

consistent with the temporal patterns of expression of HR pathway genes, (ii) RT preference for LOH events contrasts that for point mutations and somatic copy number alterations, (iii) LOH events significantly overlap with sites of high RNA polII occupancy near transcription start sites and (iv) the relative frequency of LOH events in human chromosomes correlates with the distance between homologous chromosomes in the nucleus. The preference for early RT was observed irrespective of the size of the LOH events, and mutation status of *BRCA1* and *BRCA2*.

RNA polII pausing at pre-initiation complex or in early elongation is widespread in metazoans including humans (33,38). Paused RNA polII is known to interfere with advancing replisome contributing to replication stress (37), R-loop formation (39) and induce HR-mediated rescue (35). Although such repair is predominantly done using the sister chromatid as a template, the homologous chromosome copy may also be used, although at a lower frequency (31). The relative location of the homologous sequence, derived from sister chromatid or homologous chromosome, is suspected to influence the choice of template and the efficiency of HR-mediated repair (31). In light of these observations, our findings are consistent with a model that during early replication when sister chromatids are forming, HR-mediated rescue of replication forks near paused RNA polII using homologous chromosomes leads to LOH events in cancer genomes.

We also note potential caveats of the analysis. First, we prefer to take a conservative stance while inferring causality from correlation. Because data on chromatin, long-range interactions and temporal expression of the HR and NHEJ pathway genes were not derived from the same samples, we cautiously interpreted the findings. Second, we acknowledge that RNA polII pausing could be one of the factors that contribute to replicative stress leading to HR-mediated rescue (35,36), and many of these factors can be inter-related; thus, a more comprehensive survey is required to estimate their effects during early replication. Third, we were unable to consider intra-tumor heterogeneity, tissue-specific variation in RT and post-transcriptional modifications on the DNA repair pathway genes during cell cycle in our analysis. Nevertheless, our results are consistent across different tumor types, robust against the choice of data sets, size classes of LOH events, statistical approaches and potential covariates. Moreover, they are in agreement with current literature regarding the sources of replicative stress and HR-mediated repair. So, we anticipate that these issues are unlikely to bias our conclusions. Nevertheless, independent validation of our findings would establish the conclusions firmly.

Our findings highlight an important distinction between LOH and other classes of genomic alterations such as point mutations and somatic copy number alterations. Point mutations (4–6,40,41) and somatic copy number alterations (particularly deletions) (4,7,8) frequently occur in late RT regions. In contrast, we found that LOH events preferentially occur in early RT regions. In the early RT regions, which are also enriched in protein-coding genes (3), LOH-mediated gene conversion can potentially replace wild-type alleles with recessive deleterious alleles, leading to increased risk of manifestation of recessive deleterious traits, complicating the resulting phenotype in the affected individuals. Damage-induced hypermutability and error-prone repair of such regions could lead to further genetic changes (42,43). Furthermore, the difference in RT preference between different classes of genomic alterations also provokes a testable hypothesis whether replicating cells show any changing preference between various DNA repair pathways, which have different levels of efficiency and fidelity (1), as the replication progresses.

SUPPLMENTARY DATA

Supplementary Data are available at NAR Online. Supplementary Modules M1–M6.

ACKNOLWLEDGEMENTS

The authors thank Robert Sclafani, David Schwartz, Nancy Maizels, James DeGregori, Kornelia Polyak and the anonymous reviewers for insightful discussions and critical comments.

FUNDING

University of Colorado School of Medicine; American Cancer Society [ACS-IRG 57-001-53]; National Cancer Institute Physical Sciences Oncology Center initiative [U54-CA143798]. Funding for open access charge: University of Colorado School of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Chapman,J.R., Taylor,M.R. and Boulton,S.J. (2012) Playing the end game: DNA double-strand break repair pathway choice. *Mol. Cell*, **47**, 497–510.
2. Mao,Z., Bozzella,M., Seluanov,A. and Gorbunova,V. (2008) DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle*, **7**, 2902–2906.
3. Hansen,R.S., Thomas,S., Sandstrom,R., Canfield,T.K., Thurman,R.E., Weaver,M., Dorschner,M.O., Gartler,S.M. and Stamatoyannopoulos,J.A. (2010) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA*, **107**, 139–144.
4. Koren,A., Polak,P., Nemesh,J., Michaelson,J.J., Sebat,J., Sunyaev,S.R. and McCarroll,S.A. (2012) Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.*, **91**, 1033–1040.
5. Liu,L., De,S. and Michor,F. (2013) DNA replication timing and higher-order nuclear organization determine patterns of single nucleotide substitutions in cancer genomes. *Nat. Commun.*, **4**, 1502.
6. Woo,Y.H. and Li,W.H. (2012) DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat. Commun*, **3**, 1004.
7. Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.
8. De,S. and Michor,F. (2011) DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotechnol.*, **29**, 1103–1108.

9. Drier,Y., Lawrence,M.S., Carter,S.L., Stewart,C., Gabriel,S.B., Lander,E.S., Meyerson,M., Beroukhim,R. and Getz,G. (2013) Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.*, **23**, 228–235.
10. Donley,N. and Thayer,M.J. (2013) DNA replication timing, genome stability and cancer: late and/or delayed DNA replication timing is associated with increased genomic instability. *Semin. Cancer Biol.*, **23**, 80–89.
11. Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
12. Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.
13. Karolchik,D., Hinrichs,A.S. and Kent,W.J. (2012) The UCSC Genome Browser. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis... [et al.]*, **Chapter 1**, Unit 1 4.
14. Pedersen,B.S., Konstantinopoulos,P.A., Spillman,M.A. and De,S. (2013) Copy neutral loss of heterozygosity is more frequent in older ovarian cancer patients. *Genes Chromosomes Cancer*, **52**, 740–746.
15. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
16. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
17. de Lichtenberg,U., Wernersson,R., Jensen,T.S., Nielsen,H.B., Fausboll,A., Schmidt,P., Hansen,F.B., Knudsen,S. and Brunak,S. (2005) New weakly expressed cell cycle-regulated genes in yeast. *Yeast*, **22**, 1191–1201.
18. Pramila,T., Wu,W., Miles,S., Noble,W.S. and Breeden,L.L. (2006) The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes Dev.*, **20**, 2266–2278.
19. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
20. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.
21. Gauthier,N.P., Jensen,L.J., Wernersson,R., Brunak,S. and Jensen,T.S. (2010) Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Res.*, **38**, D699–D702.
22. Gauthier,N.P., Larsen,M.E., Wernersson,R., de Lichtenberg,U., Jensen,L.J., Brunak,S. and Jensen,T.S. (2008) Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Res.*, **36**, D854–D859.
23. Durkin,S.G. and Glover,T.W. (2007) Chromosome fragile sites. *Annu. Rev. Genet.*, **41**, 169–192.
24. Barlow,J.H., Faryabi,R.B., Callen,E., Wong,N., Malhowski,A., Chen,H.T., Gutierrez-Cruz,G., Sun,H.W., McKinnon,P., Wright,G. *et al.* (2013) Identification of early replicating fragile sites that contribute to genome instability. *Cell*, **152**, 620–632.
25. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
26. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J. *et al.* (2012). ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
27. Rosenbloom,K.R., Sloan,C.A., Malladi,V.S., Dreszer,T.R., Learned,K., Kirkup,V.M., Wong,M.C., Maddren,M., Fang,R., Heitner,S.G. *et al.* (2013) ENCODE Data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.*, **41**, D56–D63.
28. Heride,C., Ricoul,M., Kieu,K., von Hase,J., Guillemot,V., Cremer,C., Dubrana,K. and Sabatier,L. (2010) Distance between homologous chromosomes results from chromosome positioning constraints. *J. Cell Sci.*, **123**, 4063–4075.
29. Bolzer,A., Kreth,G., Solovei,I., Koehler,D., Saracoglu,K., Fauth,C., Muller,S., Eils,R., Cremer,C., Speicher,M.R. *et al.* (2005) Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.*, **3**, e157.
30. Furey,T.S. and Haussler,D. (2003) Integration of the cytogenetic map with the draft human genome sequence. *Hum. Mol. Genet.*, **12**, 1037–1044.
31. Kadyk,L.C. and Hartwell,L.H. (1992) Sister chromatids are preferred over homologs as substrates for recombinational repair in *Saccharomyces cerevisiae*. *Genetics*, **132**, 387–402.
32. Mortusewicz,O., Herr,P. and Helleday,T. (2013) Early replication fragile sites: where replication-transcription collisions cause genetic instability. *EMBO J.*, **32**, 493–495.
33. Wu,J.Q. and Snyder,M. (2008) RNA polymerase II stalling: loading at the start prepares genes for a sprint. *Genome Biol.*, **9**, 220.
34. Saleh-Gohari,N., Bryant,H.E., Schultz,N., Parker,K.M., Cassel,T.N. and Helleday,T. (2005) Spontaneous homologous recombination is induced by collapsed replication forks that are caused by endogenous DNA single-strand breaks. *Mol. Cell. Biol.*, **25**, 7158–7169.
35. Michel,B., Flores,M.J., Viguera,E., Grompone,G., Seigneur,M. and Bidnenko,V. (2001) Rescue of arrested replication forks by homologous recombination. *Proc. Natl Acad. Sci. USA*, **98**, 8181–8188.
36. Iraqui,I., Chekkal,Y., Jmari,N., Pietrobon,V., Freon,K., Costes,A. and Lambert,S.A. (2012) Recovery of arrested replication forks by homologous recombination is error-prone. *PLoS Genet.*, **8**, e1002976.
37. Bermejo,R., Lai,M.S. and Foiani,M. (2012) Preventing replication stress to maintain genome stability: resolving conflicts between replication and transcription. *Mol. Cell*, **45**, 710–718.
38. Adelman,K. and Lis,J.T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.*, **13**, 720–731.
39. Aguilera,A. and Garcia-Muse,T. (2012) R loops: from transcription byproducts to threats to genome stability. *Mol. Cell*, **46**, 115–124.
40. Schuster-Bockler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, **488**, 504–507.
41. Stamatoyannopoulos,J.A., Adzhubei,I., Thurman,R.E., Kryukov,G.V., Mirkin,S.M. and Sunyaev,S.R. (2009) Human mutation rate associated with DNA replication timing. *Nat. Genet.*, **41**, 393–395.
42. Burch,L.H., Yang,Y., Sterling,J.F., Roberts,S.A., Chao,F.G., Xu,H., Zhang,L., Walsh,J., Resnick,M.A., Mieczkowski,P.A. *et al.* (2011) Damage-induced localized hypermutability. *Cell Cycle*, **10**, 1073–1085.
43. De,S. and Babu,M.M. (2010) A time-invariant principle of genome evolution. *Proc. Natl Acad. Sci. USA*, **107**, 13004–13009.