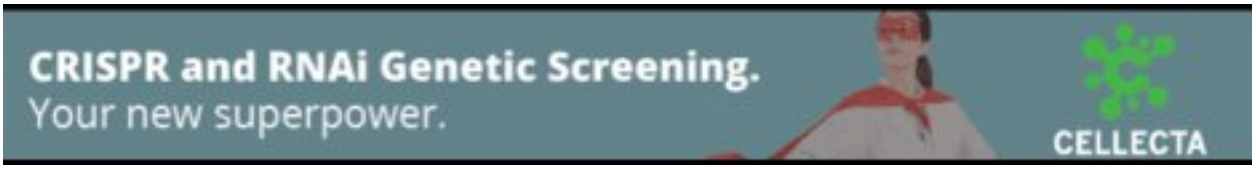
Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription

Larry D. Mesner, Veena Valsakumar, Neerja Karnani, et al.

Genome Res. 2011 21: 377-389 originally published online December 20, 2010
Access the most recent version at doi:10.1101/gr.111328.110



Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription

Larry D. Mesner,[1] Veena Valsakumar,[1] Neerja Karnani,[1] Anindya Dutta,[1,2] Joyce L. Hamlin,[1,3] and Stefan Bekiranov[1,3]

[1]*Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA;* [2]*Department of Pathology, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA*

We have used a novel bubble-trapping procedure to construct nearly pure and comprehensive human origin libraries from early S- and log-phase HeLa cells, and from log-phase GM06990, a karyotypically normal lymphoblastoid cell line. When hybridized to ENCODE tiling arrays, these libraries illuminated 15.3%, 16.4%, and 21.8% of the genome in the ENCODE regions, respectively. Approximately half of the origin fragments cluster into zones, and their signals are generally higher than those of isolated fragments. Interestingly, initiation events are distributed about equally between genic and intergenic template sequences. While only 13.2% and 14.0% of genes within the ENCODE regions are actually transcribed in HeLa and GM06990 cells, 54.5% and 25.6% of zonal origin fragments overlap transcribed genes, most with activating chromatin marks in their promoters. Our data suggest that cell synchronization activates a significant number of inchoate origins. In addition, HeLa and GM06990 cells activate remarkably different origin populations. Finally, there is only moderate concordance between the log-phase HeLa bubble map and published maps of small nascent strands for this cell line.

[Supplemental material is available for this article. The microarray data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE21110.]

The pioneering DNA fiber autoradiographic studies of Huberman and Riggs (1968) published more than 40 yr ago showed that human replication origins are bidirectional, are spaced ~100 kb apart, and adjacent origins often fire simultaneously. These experiments have provided the structural framework for all subsequent studies aimed at identifying the positions and natures of mammalian origins of replication. To understand the critical *cis*- and *trans*-acting elements that control initiation, the biggest challenge has been to localize individual origins among the morass of ~60,000 predicted origins embedded in 6×10^9 bp of DNA in a typical mammalian somatic nucleus. In *S. cerevisiae*, origins largely colocalize with genetic replicators that can be rescued by their ability to support autonomous replication of bacterial plasmids in yeast backgrounds (Stinchcomb et al. 1980; Chan and Tye 1980). Unfortunately, and for unknown reasons, this assay has not led to the global identification of replicators in mammalian genomes.

The alternative strategy has been to determine the chromosomal positions of initiation sites with the expectation that they should colocalize with replicators. Several origin mapping methods have been developed, most of which rely on prior characterization of the locus in question (usually because of prior interest in the local gene), and each is extremely challenging when applied to mammalian cells. In fact, only a few dozen potential origins have been localized so far, and then often by only one approach (for review, see Aladjem et al. 2006; Hamlin et al. 2008). A few of these (e.g., human lamin B2 [Abdurashidova et al. 2000] and DBF4 [Romero and Lee 2008]) appear to correspond to very fixed initiation sites analogous to yeast ARS elements, while the majority represent zones of closely spaced inefficient sites ranging from ~2 kb to >55 kb in length (for review, see Aladjem and Fanning 2004; Gilbert 2005; Hamlin et al. 2008). Thus, it has not been possible to uncover any meaningful shared sequence motifs. Furthermore, because of the small sample size within a given species, it has been difficult to develop useful paradigms for the distributions of origins (whether fixed or zonal) vis-a-vis active genes or local and long-range chromatin characteristics.

With the advent of the genomics era, including microarrays and high-throughput sequencing capability, it is now possible to map active transcription units onto the chromosomes, as well as epigenetic characteristics such as DNA methylation marks, nuclease hypersensitive sites, and the distribution of DNA-binding proteins and covalent histone modifications (for review, see The ENCODE Project Consortium 2007). However, a complete picture of genome activities also must include the locations of replication origins and how they relate to each of these epigenetic features. This will require very pure and comprehensive origin preparations.

Of the several techniques that have been used on an analytical scale to identify individual origins, small nascent strand preparations seemingly held the most promise for identifying all of the active origins in the genome for three reasons: (1) The strands presumably can be purified in an unbiased way from log-phase cultures, (2) one actually ends up with a product that can be hybridized to arrays or sequenced, and (3) origin-centered nascent strands only 1–2 kb in length would potentially afford reasonably

[3]Corresponding authors.
E-mail jlh2d@virginia.edu; fax (434) 924-1789.
E-mail sb3de@virginia.edu; fax (434) 924-1789.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.111328.110.

high resolution and the possibility of meaningful sequence comparisons. Practically speaking, however, purifying small origin-centered nascent strands represents a formidable task: at an average fork rate of ~2 kb/min (Huberman and Riggs 1968), a 2-kb segment of DNA will contain origin-centered nascent strands of this size for a maximum of 0.5 min, which represents less than 1/2400th of a typical 20-h mammalian cell cycle and a similarly small percentage of the corresponding nonreplicating DNA template in a log-phase population. A subsequent modification utilizes lambda-exonuclease to attempt to rid the preparations of the huge background of non-RNA-primed small broken DNA fragments (Bielinsky and Gerbi 1998). In two different studies, nascent strand preparations isolated from human HeLa cells were hybridized to microarrays corresponding to the 44 0.5–1.9-Mb regions under study by the ENCODE consortium (http://genome.ucsc.edu/ENCODE/encode.hg17.html), with the goal of preparing comprehensive maps of origin distributions (Cadoret et al. 2008; Karnani et al. 2010). As we will discuss, the origin maps elaborated by these studies are apparently less than saturating, and are only moderately concordant with one another or with the present study.

In a very different approach to purifying origins, we discovered that restriction fragments containing internal replication bubbles (initiation sites) can be selectively trapped in gelling agarose. In a pilot study, EcoRI fragments from early S-phase Chinese hamster ovary (CHO) cells were subjected to the trapping procedure and a small library of ~5000 clones was prepared (Mesner et al. 2006). The library was shown to be essentially pure by demonstrating the presence of bubbles in 14 of 14 cognate fragments when analyzed in genomic DNA in vivo on two-dimensional (2D) gels.

By this bubble-trapping approach, we have now prepared validated, nearly pure, and comprehensive bubble libraries (>10^6 independent clones each) from early S- and log-phase HeLa cells (an adenocarcinoma), and log-phase GM06990 (an EBV-immortalized human lymphoblastoid cell line). The HeLa and GM06990 cell lines were chosen because they have been extensively studied by the ENCODE Consortium, which has now amassed a large amount of information on several different genetic and epigenetic features of the human genome. In addition, HeLa cells were the source of small nascent strands in the origin mapping studies cited above (Cadoret et al. 2008; Karnani et al. 2010), allowing comparison of the two origin isolation schemes. Finally, although small numbers of potential origins have been identified by the nascent strand technique in karyotypically normal human and murine cell lines (Lucas et al. 2007; Sequeira-Mendes et al. 2009), the GM06990 bubble library represents a nearly saturating origin collection (>10^6 independent clones). This allows comparison of the origin distributions in two very different cell types: one EBV-immortalized cell line with a normal karyotype and of mesodermal lineage (GM06990), and the other highly transformed and probably of ectodermal origin (HeLa).

Importantly, these are the only preparations whose purities have been biologically validated by a method independent of the origin isolation scheme itself (i.e., 2D gels), and in which the large majority of the candidates (85%–90%) correspond to in vivo initiation sites. We have hybridized each origin library to tiling arrays representing the ENCODE pilot regions, which encompass ~1% of the human genome (The ENCODE Project Consortium 2007). Results of our studies paint a fascinating picture of origin characteristics, as well as their distributions relative to underlying DNA sequences, local gene activity, and selected epigenetic features.

Results

Origin libraries prepared from trapped bubble-containing fragments are nearly pure

The starting material for the origin isolation scheme is a very pure preparation of replication intermediates (RIs) isolated by a matrix-enrichment and BND-cellulose chromatographic procedure that we developed (Dijkwel et al. 1991; Mesner et al. 2009), and which does not select for or against any particular kind of RI (Vaughn et al. 1990b). The first origin library was prepared from HeLa cells, which can be reasonably well synchronized by a thymidine/mimosine double-block protocol (Mesner et al. 2009; see Methods). This allowed us to clone potential, anonymous, early firing origin fragments that ultimately could be validated by analyzing their in vivo counterparts in early S-phase genomic DNA on 2D gels (see legend to Fig. 1; Brewer and Fangman 1987). Note that we are not able to detect diploid origins in log-phase genomic DNA owing to the very low signal-to-noise ratios; therefore, both the early S-phase and log-phase trapped DNA samples were initially validated using a probe for the model amplified rDNA origin (Little et al. 1993).

Cell samples were isolated 80 min after release from the mimosine block, which corresponds to early S-phase in the majority

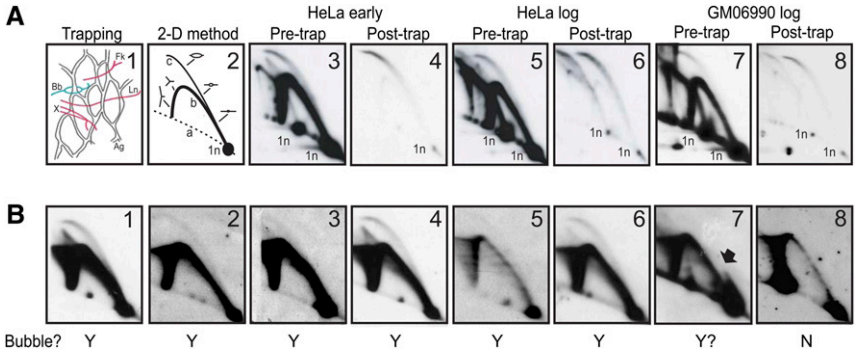


Figure 1. The trapped material, as well as the early S-phase HeLa test library, are extremely pure. (*A*) Frame 1 illustrates the principle of the bubble-trapping procedure, with agarose (Ag) indicated by the network, bubbles (Bb) indicated in red, and single forks (Fk), termination structures (X), and linear fragments (L) indicated in green. Frame 2 diagrams a 2D gel separation of replication intermediates, in which fragments with centered bubbles trace curve c, single forks trace curve b, and linear fragments correspond to the 1n spot. Frames 3, 4, 5, 6, 7, and 8 illustrate the composition of replication intermediates in the rDNA origins before and after trapping in agarose for the early S-phase HeLa sample, one of the log-phase HeLa samples, and the log-phase GM06990 sample, respectively. Note that all samples display two 1n spots, which arise from an RFLP at an EcoRI site in the multiple copies of the rDNA locus; only the smaller 12-kb RFLP is detected in the early S-phase samples, while 12- and 18-kb variants are both detected in the log-phase HeLa and GM06990 samples. (*B*) Eight anonymous cloned fragments from the early S-phase library ranging from 4 kb to 11 kb in length were tested for authenticity by analyzing their cognate genomic restriction fragments with suitable single-copy probes in 2D gels. Fragments 1–6 all display a composite pattern (complete bubble and single fork arc) indicative of initiation zones, while fragment 7 may correspond to a fixed, off-centered origin. The Y and N *below* each panel indicate whether each candidate actually displayed a bubble arc.

of the population. RIs were purified using EcoRI to digest the DNA and were subjected to the agarose bubble-trapping procedure (illustrated in Fig. 1A, frame 1, where Ag, Bb, Fk, X, and L denote agarose, bubbles, single forks, termination structures, and linear fragments, respectively; Mesner et al. 2006; Mesner and Hamlin 2009). After exhaustive electrophoresis to rid the plug of single-forked, X-shaped, and linear fragments, the trapped DNA was recovered from the plug and analyzed.

Figure 1A, frame 2, illustrates the principle of the 2D gel replicon-mapping method (also see legend). Figure 1A, frames 3 and 4, display the patterns obtained for a 12-kb fragment from the amplified rDNA origin in purified RIs from early S-phase HeLa cells before and after trapping in LMP agarose. The starting RI preparation (Fig. 1A, frame 3) displays a complete bubble arc, a very pronounced single-fork arc resulting from passive replication by forks from nearby sites in the rDNA initiation zone, and linear DNA (the 1n spot illustrated on the diagonal curve a in Fig. 1A, frame 2). Note that the second spot to the left and its accompanying faint arc in Figure 1A, frame 3, represent a larger polymorphic rDNA fragment that initiates primarily in mid-S-phase (Larner et al. 1999; see below). Importantly, after trapping, 90%–95% of the material retained in the plug migrates at the position of the bubble arc (Fig. 1A, frame 4). We have shown previously that the material in the 1n spot corresponds to templates for small bubbles destabilized during extraction of DNA from the plug, which is reflected by depletion of the lower end of the bubble arc. Thus, the trapping procedure results in an impressive purification of fragments that contained bubbles in vivo.

The remainder of the trapped material from this early S-phase HeLa DNA was utilized to prepare a recombinant origin library, and more than 10^6 independent clones were isolated. To confirm that the purity of the library is similar to that of the rDNA origin in the starting trapped material, genomic EcoR1 fragments corresponding to eight anonymous clones in the library were analyzed in early S-phase genomic DNA on 2D gels with cognate probes (size range ~4–15 kb). As shown in Figure 1B, frames 1–6, six of these genomic fragments displayed the composite bubble and fork pattern characteristic of initiation zones in early S-phase (Vaughn et al. 1990a; Dijkwel and Hamlin 1992). The fragment shown in Figure 1B, frame 7, may contain a relatively fixed and efficient site situated off-center in the fragment, since the lower end of the bubble arc is visible, but then reverts to the single fork arc at larger bubble sizes (arrow) (Brewer and Fangman 1987). However, an unfortunate distribution of sites for restriction enzymes that efficiently digest matrix-affixed DNA prevents real proof of this proposal. A bubble arc was not detected in the fragment shown in Figure 1B, frame 8, at this exposure. Thus, at least six and possibly seven of eight randomly selected clones arose from fragments that sustained measurable initiation events in early S-phase in vivo.

After this initial success, we utilized the same bubble-trapping procedure to prepare two origin libraries from log-phase HeLa cells and one from log-phase GM06990 cells. The efficacy of the initial trapping step is illustrated for one of the log-phase HeLa replicates and for the GM06990 library in Figure 1A, frames 5 and 6 and frames 7 and 8, respectively. The probe for the amplified rDNA origin demonstrates that >90% of the material retained in the agarose plugs from both preparations corresponds to fragments that contained replication bubbles in vivo, with both the early and midfiring rDNA origin fragments being recovered (Fig. 1A, frames 5–8). Each of the replicate log-phase HeLa preparations displayed similar purities (LD Mesner, unpubl.; also see below). Libraries were constructed from these preparations, and $\geq 1 \times 10^6$ independent clones were recovered from each. Because the relative purity of the trapped early S-phase material and the resulting bubble libraries proved similar, we are confident of the biological validity of the log-phase HeLa and GM06990 libraries as well.

As an independent biological validation, we utilized PCR primers to detect fragments encompassing the beta-globin, lamin B2, and *MYC* origins in the log-phase HeLa libraries (note that the activities of these three origins have not been evaluated in GM06990 cells). All three origins are enriched in the log-phase HeLa libraries relative to genomic DNA, ranging from only approximately fourfold for the beta-globin fragment to ~50-fold and 57-fold for the lamin B2 and *MYC* fragments, respectively (L Wang and JL Hamlin, unpubl.).

Microarray hybridizations and biological replicates are very reproducible

Pooled DNAs from each origin library, as well as log-phase genomic control DNAs from the two parental cell lines, were hybridized to Affymetrix microarrays, which interrogate the unique sequences in the 44 ENCODE regions with 25-bp oligonucleotides spaced ~22 bp apart. The resulting data were normalized and smoothed with Affymetrix tiling array software to estimate $\log_2$ ratios for each library sample versus its genomic control (Cawley et al. 2004).

Figure 2A outlines the general strategy for array analysis, using as an example two independent hybridizations with the early S-phase library in which the samples were labeled either with

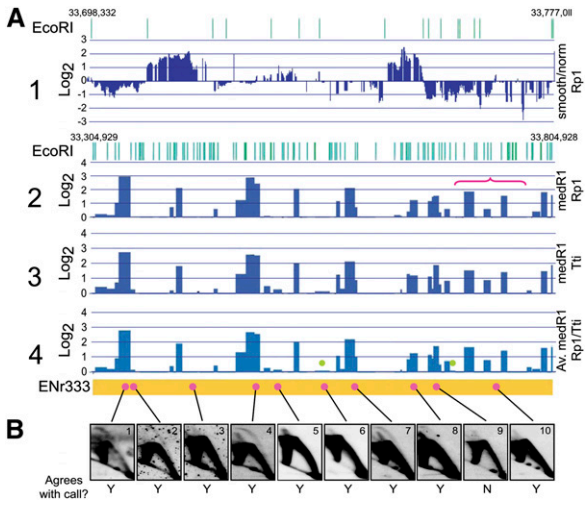


Figure 2. Microarray hybridizations are reproducible and paint a reliable picture of origin distributions in the genome. (*A*) Early S-phase HeLa microarray data at different stages of analysis. (*A1*) Zoomed-in view of normalized, smoothed, $\log_2$ ratios of bubble fragments over genomic control from the Rp1 hybridization for a 78-kb region in Enr333 (33,698,3332–33,777,011), which illustrates details of the array signals and their strict tracking with individual EcoRI fragments (EcoRI sites shown in green at the *top*). (*A2*) Log_2 median signals (zero cut-off) within EcoRI fragments (medRI) from the Rp1 hybridization for the entire 500-kb Enr333 region (33,304,929–33,804,928), showing the EcoRI sites at the *top*; the location of the expanded region in *A1* is indicated with the red bracket. (*A3*) MedRI data for the Tt1 hybridization. (*A4*) Averaged medRI values of Rp1 and Tt1 replicates; the green dots indicate calls that would not survive the 0.2 cut-off value. (*B*) The corresponding genomic fragments for selected positive and negative fragments from panel *A4* were analyzed on 2D gels, with the Y and N *below* the images indicating whether the 2D gel result concords with the microarray call.

biotinylated dATP by random priming (Rp1) or with biotinylated ddATP using terminal transferase (Tt1; Methods). Figure 2A, panel 1, presents smoothed, normalized Rp1 data for a ~79-kb region in ENr333 on chr 20 (residues 33,698,332–33,777,011; bracketed in red in Fig. 2A, panel 2). Signals enriched or depleted in the library relative to total genomic DNA are plotted above and below the baseline (note that regions displaying neither enriched nor depleted signals largely correspond to repetitive regions of the genome excluded from the arrays). As expected, the signals are relatively uniform between any two EcoRI sites (shown as green vertical lines above panels 1 and 2 in Fig. 2A). The remaining panels in Figure 2A represent an EcoRI-centric analysis of the entire 500-kb ENr333 region on chr 20 (33,304,929–33,804,928), in which the median of the smoothed $\log_2$ signals for each fragment (medRI) has been calculated and, as in Figure 2A, panel 1, no cut-off has been applied. Note that the boundary values for the ENCODE region correspond approximately to the outermost EcoRI sites.

The concordance between the 626 Rp1 and 670 Tt1-positive bubble signals is 89.9% and 84.0%, respectively (Fig. 2A, cf. panels 2 and 3; Supplemental Table S1). We generated a scatter plot of medRI values for the Rp1/Tt1 comparison (Supplemental Fig. S3A) and calculated an associated pairwise correlation coefficient of 0.72. Clearly, the two different labeling conditions give very similar microarray patterns. Figure 2A, panel 4, presents the medRI values for the composite signals calculated from the two independent hybridizations treated as replicates (see Supplemental Methods). In subsequent calculations, a relatively liberal medRI cut-off value of 0.2 was used for all of the microarray data based on a number of statistical criteria (see Supplemental Methods; Supplemental Figs. S1, S2). EcoRI fragments whose medRI levels are above the 0.2 cut-off will be referred to henceforward as positive calls, positive fragments, or bubble-containing fragments. The fragments marked with green dots in Figure 2A, panel 4, are examples that did not attain the 0.2 cut-off in ENr333 and would be excluded in all subsequent analyses.

Figure 3A displays the positive calls for ENr333 in the two hybridization replicates, their overlaps, and calls that are unique to each replicate, using the $\log_2$ cut-off of 0.2. As might be expected, the nonoverlapping fragments had lower average signal strengths than those that overlap (0.479 vs. 1.294, respectively) and lower average fragment sizes (5315 vs. 6697; Supplemental Table S1).

Figure 2B is an analysis of the reliability of the positive microarray calls as a measure of library composition itself, and thus, in vivo origin activity at loci other than rDNA. The 10 anonymous fragments from ENr333 indicated with red dots were examined in early S-phase genomic DNA on 2D gels with cognate hybridization probes. With only one exception (fragment 9), positive calls from the array correspond to detectable bubble-containing fragments in the genome (fragments 1, 2, 4, 7, and 8), and the signal strengths are generally consistent with the strengths of the bubble arcs in 2D gels. Furthermore, bubbles were not detected for borderline or negative calls (fragments 3, 5, 6, and 10). Note that the Y and N letters below Fig. 2B indicate whether or not the 2D gel result generally corresponds to the microarray call. Six other positive fragments from four different ENCODE regions also were tested on 2D gels for validity, with 5/6 displaying bubble arcs (LD Mesner, unpubl.).

The two biological replicate log-phase HeLa libraries were independently processed, labeled, and hybridized to the ENCODE arrays. The numbers of positive fragments detected for the two libraries are quite different (594 and 1068 for Rep3 and Rep4, respectively), even though the purities of the individual trapped rDNA bubbles appeared identical on 2D gels (data not shown). This finding suggests the possibility that, although both bubble libraries are quite pure, the Rep3 library may not be as saturating as Rep4, even though both libraries had similar numbers of independent clones (~10^6). More likely is the probability that hybridization to the arrays was not as efficient with Rep3. Nevertheless, an impressive 77.1% of Rep3 fragments were also detected in the larger Rep4 library (17% would be expected by chance). By comparison, 42.9% of Rep4 fragments were present in the Rep3 library, with 9.5% expected by chance (Supplemental Table S1). Therefore, the biological replicates are clearly sampling very similar spectra of initiation sites. This is reinforced by the scatter plot of medRI values shown in Supplemental Figure S3B and a pairwise correlation coefficient of 0.67. Once again, the medRI signals for nonoverlapping fragments were considerably lower and fragment sizes smaller than those for overlapping ones (0.924 vs. 1.788 and 5960 vs. 6981 bp; Supplemental Table S1).

The overlapping and unique calls for each biological replicate are illustrated for ENr333 in Figure 3B. Note that when the medRI cut-off for the larger Rep4 library was raised incrementally from 0.2 to 1.0 (at which point the number of positive calls in each is ~600), only ~59% of clones in each library found matches in the other. Thus, the drop from 77% overlap suggests that even the smaller Rep3 library contains some unique clones that are not present in Rep4. The composite overlapping medRI values surviving the 0.2 cut-off for the two replicate log-phase origin libraries are tabulated in Supplemental Table S1B, and represent a total of 458 shared

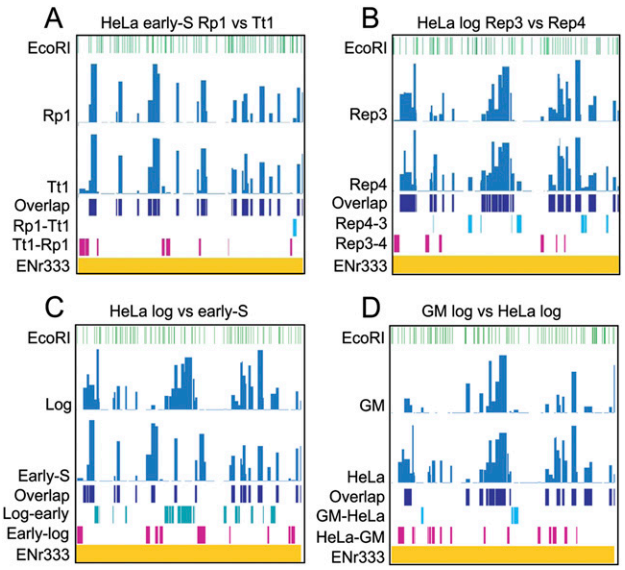


Figure 3. MedRI origin maps reveal bubble fragment clustering, effects of cell synchronization, and major differences in origin distribution between GM06990 and HeLa cells. Four different medRI comparisons are shown for ENr333 in the second and third rows of each panel, with the fourth row indicating overlaps, and fifth and sixth rows representing the nonoverlapping fragments. The notation X–Y indicates fragments in sample X that were not detected in sample Y. Note also that the signals in the last three rows are proportional to the width of each EcoRI fragment, but not the signal strengths. (*A*) Comparison of Rp1 and Tt1 hybridizations with the early S-phase HeLa library. (*B*) Comparison of Rep3 and Rep4 biological replicate log-phase HeLa libraries. (*C*) Comparison of log-phase and early S-phase HeLa libraries. (*D*) Comparison of log-phase HeLa and GM06990 origin libraries.

calls. When positive calls for the two libraries were combined, only 657 survived the 0.2 cut-off because of the many weaker medRI signals from the Rep4 library. This shared 657 value will be used in subsequent comparisons with the single early S-phase HeLa and log-phase GM06990 libraries.

Painting the landscape of origin distributions in the human genome

The early S-phase HeLa library, combined log-phase HeLa libraries, and log-phase GM06990 library illuminated 646, 657, and 988 individual EcoRI fragments within the ENCODE regions, respectively (Supplemental Table S1). Note that this includes end fragments that are interrupted by the ENCODE boundaries. The false discovery rate associated with the 0.2 medRI cutoff is 0.4%, 0.6%, and 1.5% for the early S-phase HeLa, log-phase HeLa, and log-phase GM06990 data, respectively (see Supplemental Methods; Supplemental Fig. S1C). The average insert sizes in the libraries are 6.57, 6.96, and 6.12 kb, respectively, compared with the average of all EcoRI fragments in the 44 ENCODE regions (~4.14 kb; see Supplemental Table S1 for ranges, etc.). The sizes are clearly skewed toward the larger end of the spectrum, as shown in the distributions in Figure 4A (means/modes are indicated in the upper righthand corner of each panel). Thus, the trapping and/or cloning procedure selects for slightly longer fragments. For the three libraries, these values indicate that 10.3, 10.5, and 15.7% of all 6282 EcoRI fragments in the ENCODE regions initiated replication in vivo in early S-phase HeLa, and log-phase HeLa and GM06990 cells, respectively. This corresponds to 15.3, 16.4, and 21.8% of the genome in these regions.

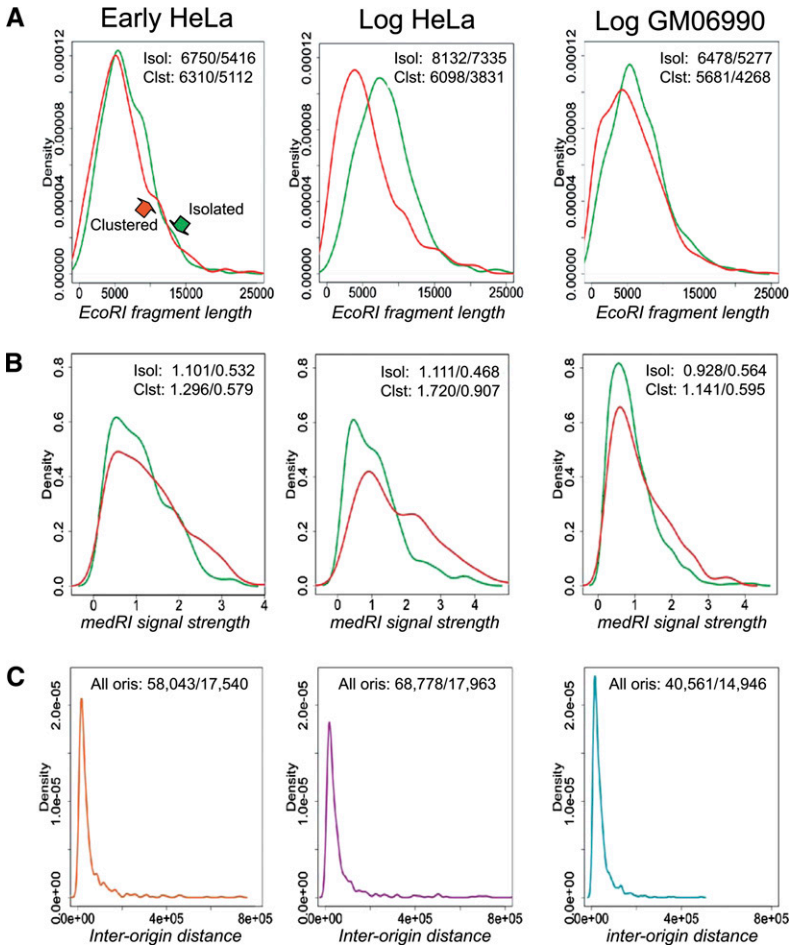


Figure 4. Fragment sizes, signal strengths, and inter-origin distances in the three libraries. (*A*,*B*) The distributions of fragment lengths and signal strengths for isolated (green) and clustered (red) fragments are plotted for the three libraries, in which density corresponds to normalized counts or frequency (where the area under the curve equals 1) in *A*, *B*, and *C*, respectively. The mean and modal values for each distribution are indicated in the *upper right* corner of each panel. (*C*) The interorigin distance distributions for the three libraries, with means and modes shown *within* each panel.

Bubble-containing fragments often cluster into zones

The patterns in Figure 2A, panel 4, for the early S-phase HeLa library, as well as an analysis of the data for the log-phase HeLa and GM06990 samples (see below), indicate that large numbers of positive EcoRI fragments in all three libraries map together into clusters of two or more adjacent fragments (41.2%, 57.8%, and 45.5%, respectively). To test the significance of these values, the positions of all the EcoRI fragments within their respective ENCODE regions were randomly permuted 10,000 times (Supplemental Table S2; Supplemental Methods). Randomization within individual ENCODE regions preserves the number and genomic coverage of positive fragments in a random model, and controls for GC-content biases, local gene density, etc., on the scale of the ENCODE regions (0.5–1.9 Mb). Henceforward, we will refer to any cluster of two or more adjacent fragments as a cluster or an initiation zone.

The resulting values for the number of zones in the real data sets and their enrichments over the random model for the early S-phase HeLa, log-phase HeLa, and log-phase GM06990 libraries are: 111 (1.4-fold), 128 (1.5-fold), and 177 (1.3-fold), respectively, with an associated *P*-value of $<10^{-4}$ for all three comparisons (Supplemental Table S2). In addition, the average numbers of fragments within zones and their fold-enrichments over random for the three libraries are 3.4 (1.25-fold), 4.0 (1.46-fold), and 3.5 (1.34-fold), with *P*-values for the log-phase libraries of $<10^{-4}$ and for the early S-phase library of 3×10^{-4}. Interestingly, the mean lengths of zones are not significantly different from the mean random lengths. This is probably because isolated fragments are longer on average than zonal fragments (see below), and the fact that "zones" in the random model are made up of a mixture of fragments from the experimental sample that were either zonal or isolated. We also observed a wide range of fragment numbers in zones: from two to 11 in log-phase HeLa cells, with zonal lengths from 3.1 to 84.6 kb (average of ~18.1 kb). It is clear from Figure 4A that isolated fragments (green curves) are generally longer than clustered ones (red curves), with the comparative mean lengths being 6750 versus 6310, 8132 versus 6098, and 6478 versus 5681

for early S-phase HeLa, and log-phase HeLa and GM06990 libraries, respectively. On the other hand, the mean and modal medRI signal strengths are higher in clustered fragments for all three libraries, as indicated in Figure 4B and Supplemental Table S2.

Whether isolated or zonal, interorigin distances fall within the expected range for mammalian origins

We used these data to derive the distributions of distances between the centers of individual origins, whether represented by single fragments or fragment clusters that likely correspond to initiation zones (Fig. 4C). For the early S-phase HeLa, log-phase HeLa, and log-phase GM06990 libraries, the distance distributions display modes at ~17.5, 18.0, and 15.0 kb, and have an extremely long right tail extending in some cases to distances of ~1 Mb. For the three libraries, the mean intervals are 58.0, 68.8, and 40.6 kb, respectively. The lower value for GM06990 is, in part, a consequence of the greater coverage of the genome by this origin library.

Comparison of early S- and log-phase HeLa bubble distributions indicates significant overlap, but suggests that cell synchronization alters origin selection

Of the few dozen mammalian origins that have been identified by molecular biological approaches, most fire in early S-phase, in many cases because they were identified as early replicating segments in synchronized cells (for review, see Aladjem et al. 2006; Hamlin et al. 2008). To determine the extent to which these origins are representative, we compared origin distributions between the early S-phase and log-phase HeLa libraries in the ENCODE regions. As an example, medRI values for ENr333 are presented in Figure 3C and Supplemental Table S1. There is clearly considerable overlap between the origin fragments in the early S- and log-phase libraries, with 282 fragments being shared (42.9% and 43.7% of each library, respectively). The scatter plot for the two data sets is shown in Supplemental Figure S3C, and the pairwise correlation coefficient is 0.51.

Obviously, a substantial number of the log-phase library fragments are unique (57.1%), and therefore could correspond to origins that fire at later times in S-phase (but see below). Unexpectedly, however, 56.3% of fragments in the early S-phase library have low or undetectable counterparts in the log-phase library (early log) (Fig. 3C; Supplemental Table S1). At least some of the difference between the two samples is attributable to lack of reproducibility and/or saturation in the biological preparations and hybridization conditions; indeed, fragments unique to each library are generally smaller and their signals weaker than the overlapping fragments, as observed with the replicate comparisons (Supplemental Table S1). However, the much lower R^2 value for the early S-phase versus log-phase comparison (0.51) argues that the two libraries are sampling somewhat different origin populations (Supplemental Fig. S3C). Thus, the relatively standard synchronizing regimen used here has activated a subset of origins that is otherwise inefficient or dormant in undisturbed cultures.

Log-phase HeLa cells and the EBV-immortalized GM06990 lymphoblastoid cell line utilize very different sets of origins

To date, there have been very few studies addressing the differences in origin usage between two different cell types from the same organism. Although the HeLa and GM06990 cell lines are not a matched pair, they provide an opportunity to assess the degree to which cells can alter the spectrum of origin usage in very different genetic and physiological states. The medRI values for ENr333 are graphed in Figure 3D, and summaries of their values across all ENCODE regions are presented in Supplemental Table S1. ENr333 was chosen because, although the GM06990 library illuminates many more bubble fragments than the HeLa library overall (988 vs. 657), many fewer fragments are detected by GM06990 in this particular region. This illustrates that active origin distributions vary considerably from region to region between cell lines. In fact, the two libraries share only 28.3% of total GM06990 fragments and 42.6% of HeLa fragments, with the large remainders being unique to each library. Once again, the sizes and signal strengths for nonoverlapping fragments are lower than overlapping fragments on average (Supplemental Table S1), accounting for some of this discordance. Nevertheless, the pairwise correlation coefficient for GM06990 and HeLa cells is only 0.35 (Supplemental Fig. S3D). In a comparison between the 988 positive GM06990 calls and 1068 calls from the single larger Rep4 HeLa log-phase library, the overlap increased only somewhat to 38.66% and 35.77% for GM06990 and HeLa, respectively.

Initiation zones or clusters are significantly associated with active genes

To understand the complex relationships among DNA sequence, origin activity, and transcription, we compared the distributions of active origins in log-phase HeLa and GM06990 cells with the distributions of known genes (RefSeqs) (Pruitt et al. 2007) and to the cDNA transcriptomes elaborated for these two cell lines by the Affymetrix group (The ENCODE Project Consortium 2007). This analysis was restricted to log-phase data sets, since a transcriptome for early S-phase HeLa cells is not available.

Individual bubble-containing fragments (whether clustered in zones or isolated), as well as zones treated as units, were sorted into nine different categories or classes (illustrated in Fig. 5A and similarly numbered and color-coded in Fig. 5B2 and Table 1; see legends). In Figure 5A, we have represented a gene (RefSeq) as a dark blue rectangle, with the yellow arrow indicating the direction of transcription. Isolated or zonal origin fragments that reside completely within or completely outside of a gene are indicated with dark or light blue rectangles above and below the genes, while fragments that overlap the 5′ or 3′ end of a gene are represented as light or dark green rectangles, respectively. Note that a fragment was considered to overlap a gene if the two shared at least one base pair. The numbers within each overlapping fragment in Figure 5A refer to alternative transcriptional states and correspond to the key below Figure 5B2. In some calculations (see below), a zone is treated as a unit, and the whole zone is color-coded as if it were a single fragment (as in Table 1F,G).

Note that there is ambiguity with fragments or zones that partially overlap a gene or completely contain one or more genes. For example, as shown in the topmost row in Figure 5A for isolated fragment class 3/7, if the actual initiation site (inner pink box) occurs in the right end of the fragment, it would overlap the gene, while initiation in the left end would be entirely intergenic. Likewise, isolated fragment class 2/6 could have initiated at either end and not necessarily within the body of the gene. Note also that the zone below this gene is assigned to category 2/6, but the fragments within it now sort into classes 3/7, 1/5, and 4/8 reading from left to right.

In addition, some fragments or zones contain and/or overlap two or more genes whose transcription status differs (12.8% and 6.8% of fragments for HeLa and GM06990 cells, respectively). There are sixteen possible combinations of this situation, even if only the two-gene case is considered (i.e., pairing each of categories

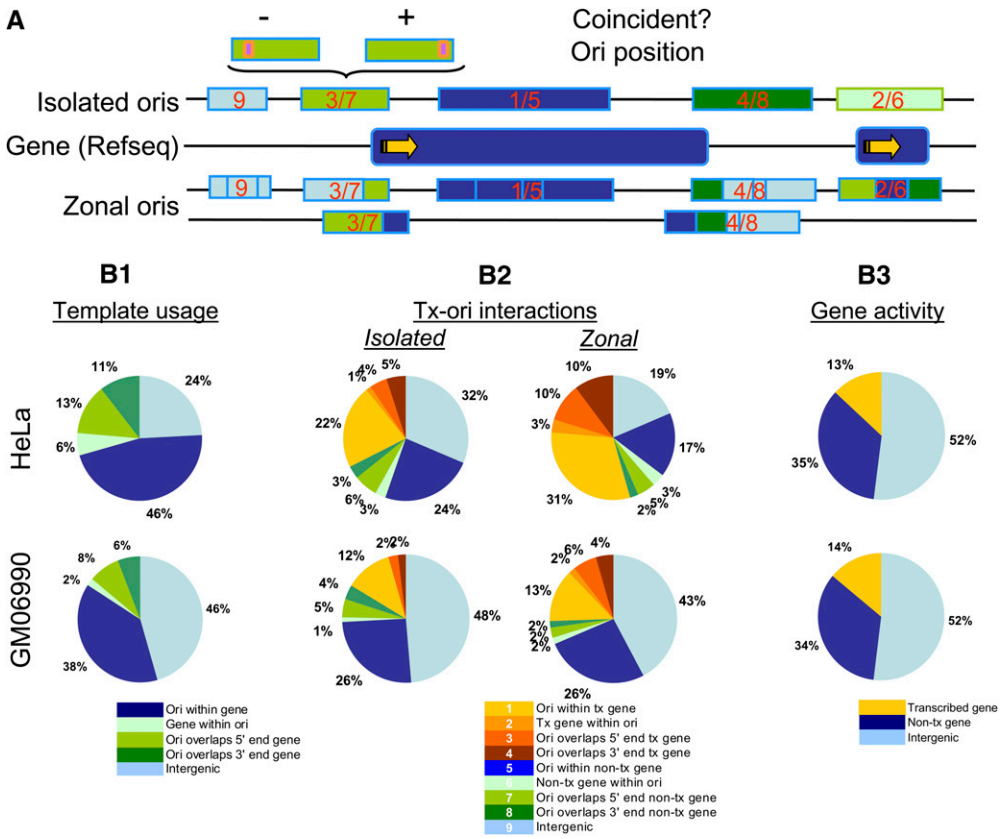


Figure 5. Origin–gene interactions are markedly different between isolated and clustered fragments, and between HeLa and GM06990 cells. (*A*) The black lines represent the DNA template with genes (Refseqs) and bubble-containing fragments, whether isolated or clustered, indicated as rounded or square boxes, respectively, on the axis. Origin fragments are colored either light-blue (intergenic), dark-blue (genic), or light or dark-green (overlapping indeterminate). The yellow arrow indicates the direction of transcription. The box numbers correspond to the nine categories summarized in the text, *below B2*, and in Table 1. The two green boxes at the *top* illustrate the ambiguity associated with a fragment that only partly overlaps a gene, since the transcribed part might not overlap the part that initiates replication. (*B1*) Origin distributions relative to DNA template sequences, where the light-blue sector is intergenic, dark-blue is genic, and the three different green sectors represent categories 6–8, which overlap genes (indicated by key *below*). (*B2*) Distribution of isolated or clustered bubble-containing fragments among intergenic (light-blue), completely nontranscribed genic (dark-blue), partially overlapping nontranscribed genic (three green sectors), or completely or partially overlapping transcribed genomic regions (yellow-to-orange sectors; see key *below*). (*B3*) Percentages of the HeLa and GM06990 genomes that are intergenic (light-blue), nontranscribed genic (dark-blue), or transcribed (yellow; see key *below*).

1–4 with each of categories 5–9; see color key in Fig. 5B2 or Table 1A). To reduce this complexity, we have made assignments according to the hierarchy established in Figure 5B2 and Table 1A. For example, if a bubble fragment overlaps the 5′ end of a nontranscribed gene and the 3′ end of a transcribed gene, it would go into category 4 (not 7). This approach does not bias significance calculations for overlapping transcription and origin activity, since the same assignments were made for randomly permuted fragments (see below). Additionally, a small number of fragments or zones contain or overlap two or more genes that are each transcribed, in which case they were classified according to the same hierarchy. However, if <500 bp of the template was transcription-free in any fragment (i.e., could not accommodate an origin recognition complex [ORC]; Diffley et al. 1994), these fragments were counted as transcribed in the totals shown in each panel, which are therefore sometimes slightly larger than the number of class 1 fragments in Table 1.

The resulting distributions of bubble-containing fragments are summarized in two ways in Table 1 and Figure 5B to aid statistical analysis (Supplemental Methods; Supplemental Table S3). In Table 1, B and C, isolated fragments from the two cell lines were categorized into their respective classes reading from left to right. In Table 1, D and E, zones were sorted into the nine categories reading from top to bottom, while their corresponding fragments were sorted into the nine classes reading from left to right (again we consider a cluster of two or more adjacent fragments to represent a zone). In Table 1, F and G, zones treated as units were assigned to one of the nine categories reading from left to right.

As shown in Figure 5B1, with regard to underlying DNA sequence, there is no clear proclivity to initiate within the 52% of the ENCODE template that is intergenic, since 46.6% and 38.5% of all bubble-containing fragments in HeLa and GM06990 cells (whether isolated or zonal) reside completely within the bodies of active or inactive genes (fragment classes 1 and 5, Table 1B–E). In Figure 5B1 light and dark blue sectors represent the distributions of bubble-containing fragments between intergenic and genic sequences, while the three sectors in shades of green correspond to ambiguous fragments that completely contain a gene or genes or overlap the 5′ or 3′ end of a gene).

Even though ~48% of the template in the ENCODE regions encodes potential transcription units, only 13.2% and 14.0% of the genome is actually transcribed in HeLa and GM06990 cells, respectively (yellow sectors in Fig. 5B3). Yet, there is a striking association between replication initiation zones and active transcription.

Table 1. Origin–gene interactions

A1 Association types

1	Ori within tx gene
2	Tx gene within ori
3	Ori overlaps 5' end tx gene
4	Ori overlaps 3' end tx gene
5	Ori within non-tx gene
6	Non-tx gene within ori
7	Ori overlaps 5' end non-tx gene
8	Ori overlaps 3' end non-tx gene
9	Intergenic

A2 Significance

Significant

Highly significant

B HeLa log library - isolated

	Total # Fgs	Fragment class 1	2	3	4	5	6	7	8	9
	277	60	4	12	14	66	8	17	9	87
Total tx	60									
Total non-tx	188									
Indeterm.	29									

C GM06990 log library - isolated

	Total # Fgs	Fragment class 1	2	3	4	5	6	7	8	9
	538	64	1	13	11	137	6	25	20	261
Total tx	65									
Total non-tx	449									
Indeterm.	24									

D HeLa log library - zonal fragments

Zone type	# Zones	# Fgs in zone	Fragment class 1	2	3	4	5	6	7	8	9
1	21	55	55								
2	24	116	28	13	16	23	5	5	3	4	19
3	22	52	16		22	2	2	2	1	0	7
4	13	41	18			14	1	0	1	1	6
5	17	39					39				
6	8	19					4	6	4	1	4
7	9	25					10		9	0	6
8	2	7					2			2	3
9	12	26									26
Totals	128	380	117	13	38	39	63	13	18	8	71
Total tx		120									
Total non-tx		174									
Indeterm.		86									

E GM06990 log library - zonal fragments

Zone type	# Zones	# Fgs in zone	Fragment class 1	2	3	4	5	6	7	8	9
1	12	28	28								
2	12	36	11	7	5	7	0	0	1	1	4
3	23	63	13		23	2	4	3	2	1	15
4	9	25	8			11	1	1	0	0	4
5	42	105					105				
6	4	12					1	3	1	1	6
7	7	15					4		7	0	4
8	5	10					4			5	1
9	63	156									156
Totals	177	450	60	7	28	20	119	7	11	8	190
Total tx		62									
Total non-tx		337									
Indeterm.		51									

F HeLa log library - zones as units

	Total # zones	Zone category 1	2	3	4	5	6	7	8	9
	128	21	24	22	13	17	8	9	2	12
Total tx	21									
Total non-tx	48									
Indeterm.	59									

G GM06990 log library - zones as units

	Total # zones	Zone category 1	2	3	4	5	6	7	8	9
	177	12	12	23	9	42	4	7	5	63
Total tx	13									
Total non-tx	121									
Indeterm.	43									

The numerical distributions of the nine origin fragment types from log-phase HeLa and GM06990 libraries are color coded as in Figure 5B2; significant associations are colored lavender, while highly significant associations are colored pink. (*A*) Color key. (*B*,*C*) Isolated fragments from the log-phase HeLa or GM06990 libraries are arrayed in a row reading from *left* to *right*, while the numbers of total transcribed, nontranscribed, or indeterminate fragments are shown in the column to the *left*. (*D*,*E*) Zonal fragment distributions are presented as a matrix, with individual zonal fragments arrayed in a row from *left* to *right*, the zones themselves (treated as units) in the first column reading from *top* to *bottom*, and the number of fragments within each zone type in the second column; total transcribed, nontranscribed, or indeterminate categories are summarized *below*. (*F*,*G*) Each zone was treated as a unit and arrayed in a row from *left* to *right*, while the numbers of total transcribed, nontranscribed, or indeterminate fragments are shown in the column.

This is most easily appreciated by comparing the yellow-to-red sectors with the blue-to-green sectors in each of the pie charts in Figure 5B2. In fact, 54.5% and 25.5% of the individual fragments in zones overlap or contain transcribed genes in HeLa and GM06990 cells, respectively (classes 1–4, Fig. 5B2; Table 1D,E). When each cluster or zone is treated as a unit origin, fully 62.5% of HeLa zones overlap or contain transcribed genes, while 31.6% of GM06990 zones do so (categories 1–4, Table 1F,G). By comparison, 32.5% and 16.5% of the isolated fragments in HeLa and GM06990 reside within or overlap active genes (classes 1–4, Fig. 5B2; Table 1B,C).

We assessed the statistical significance of the association between active genes and origin fragments, again using the complete random permutation model, whereby all of the fragments within each ENCODE region were permuted (Table 1B–E). A restricted random permutation model also was generated in which isolated fragments, origin-negative fragments, and zones treated as units were randomly permuted 10,000 times within their ENCODE regions (Table 1F,G; see Supplemental Methods for details of each analysis).

To determine the significance of any origin–gene interactions, we calculated: (1) the probability that any interaction would be obtained by chance (*P*-values), (2) the fold-enrichment of the observed over the mean random interaction values, and (3) the *z*-score, which is the number of standard deviations that separates the observed and mean random values (*z*) (Supplemental Table S3; Supplemental Methods). In Table 1, the highly significant (*P*-value $\leq 10^{-4}$; $F \geq 2$; $z \geq 3$) or significant (*P*-value <0.05; $F < 2$ or $z < 3$) overlaps are highlighted in pink or lavender, respectively. By including *F* and *z* cut-offs, we can identify fragment classes in the complete random permutation model whose enrichment is not merely due to the fact that the number of fragments in zones is, by itself, significant.

As shown by the pink cells in Table 1, for both HeLa and GM06990 cells, zones that completely contain or overlap the 5′ end of a transcribed gene (categories 2 and 3) are the only categories that are highly significant based on either the complete (Table 1D,E) or restricted (Table 1F,G) random model. Table 1, D and E, shows that 11 of the 14 significant or highly significant zonal fragment classes reside within categories 2 and 3 zones in HeLa cells. If we make the two extreme assumptions that origins fire either within or outside of the gene in the 52 indeterminate class 2–4 zonal fragments, then either 69.0% (80/116) or 24.1% (28/116) of all the bubble-containing fragments in category 2 zones overlap

transcribed genes in HeLa cells. The reality is probably somewhere between these extremes.

Surprisingly, a substantial fraction of zonal origin fragments actually reside completely within the bodies of transcribed genes (class 1 fragments; yellow sectors in Fig. 5B2, and Table 1). Consistent with the view that transcription is strongly associated with origin activity, 85.5% of the 117 transcribed genes in the ENCODE regions overlap zonal or isolated bubble-containing fragments in log-phase HeLa cells.

In the GM06990 data set, category 3 zones were found to be the most significant, with only class 3 fragments being highly significant (Table 1E). For the two extreme cases in which all of the indeterminate fragments in category 3 zones (classes 2–4) initiate either completely within or completely outside of the transcription unit, 60.3% or 20.6% of the 63 zonal fragments overlap transcribed genes. As with HeLa cells, transcription is a strong indicator of origin activity, with 73% of the 122 transcribed genes in log-phase GM06990 cells overlapping a zonal or isolated bubble-containing fragment.

Zones are strongly associated with transcribed genes containing activating histone marks in their promoters

We also were able to compare origin distributions in HeLa and GM06990 to the limited data sets available for activating and inactivating histone modifications within the ENCODE pilot regions (H3ac, H3K4me1,2,3, and H4Ac [Sanger Institute]; H4Kac4 and H3K27me3 [Yale University]; and H3K27me3 [Ludwig Institute]; Supplemental Fig. S4; Supplemental Table S4). Due to the strong association between origins and transcribed genes, it was not surprising that activating marks are significantly associated with origin-containing fragments in both cell lines (P-value $< 2.2 \times 10^{-16}$; Supplemental Methods). To determine the percentage of those that are unique to origin activity as opposed to the well-known association with promoters, we annotated isolated fragments and zones according to the following mutually exclusive categories: (1) fragment overlaps by ≥1 bp a transcribed or nontranscribed gene with or without an activating mark in the promoter region (four classes); (2) fragment is intergenic, but its center lies within 7.5 kb of the 5′ or 3′ end of a gene with or without an activating histone mark at those sites (four classes); and (3) fragment is intergenic, but its center lies more than 7.5 kb from any gene with or without activating marks (two classes; see Supplemental Methods and Supplemental Fig. S4 for details).

In HeLa cells, 9.8% and 3.3% of the ENCODE region template corresponds to transcribed genes that either contain or lack activating histone marks at their promoters, respectively, while 20.2% and 15.0% corresponds to nontranscribed genes that either do or do not contain activating marks in their promoters. Results for GM06990 cells were very similar. Remarkably, however, 58.6% and 31.1% of initiation zones in HeLa and GM06990 cells overlap transcribed genes with activating marks. For HeLa cells, associations between isolated fragments or zones and transcribed genes with activating promoter marks were each highly significant ($P < 0.0001$; category 1, Supplemental Table S4A,B), while in GM06990 cells, only the association of zones with such genes was significant (category 1, Supplemental Table S4D). None of the associations with activating marks in intergenic regions were significant. Furthermore, despite the fact that many nontranscribed genes within the ENCODE regions contain activating 5′ marks, their associations are significant only with isolated fragments in HeLa cells (category 3, Supplemental Table S4).

There is only moderate concordance between the HeLa log-phase origin library and small nascent strands isolated from HeLa cells

Two groups have attempted to map the positions of active origins of replication within the ENCODE regions of the HeLa genome by isolating preparations enriched for small nascent strands. In one study (Cadoret et al. 2008; referred to as Study 1), small, single-stranded DNAs were additionally treated with lambda-exonuclease to attempt to remove the large excess of contaminating, broken, non-RNA-primed material (Bielinsky and Gerbi 1998). In fact, 35.0% of 283 nascent strand calls reported in Study 1 overlap the 657 bubble signals by ≥1 bp, while 14.0% of bubble signals overlap nascent strand signals (Table 2; examples of different types of

Table 2. Comparison of log-phase HeLa small nascent strand and bubble distributions

A. Log HeLa bubbles that overlap λ-exo NS (Study 1 - Prioleau)	
Number of medRI bubbles	657
Number of nascent strands	283
Number of nascent strands overlapping medRI bubbles	99
Number of bubbles containing nascent strands	92
% of medRI bubbles overlapping nascent strands	14.0%
% nascent strands overlapping medRI bubbles	35.0%
B. Log HeLa bubbles that overlap λ-exo NS (Study 2 - Dutta)	
Number of med RI bubbles	657
Number of nascent strands	320
Number of nascent strands overlapping medRI bubbles	85
Number of bubbles containing nascent strands	70
% medRI bubbles overlapping nascent strands	10.7%
% nascent strands overlapping medRI bubbles	26.6%
C. Shared log HeLa NS that overlap bubbles	
Number of medRI bubbles	657
Total number of unique NS (Prioleau + Dutta less overlaps)	603
Number of overlapping NS	16
Number of bubbles containing NS (Prioleau)	92
Number of bubbles containing NS (Dutta)	70
Number of bubbles shared by NS (Prioleau and Dutta)	10
% medRI bubbles overlapping shared nascent strands	1.5%
% shared nascent strands overlapping medRI bubbles	1.7%

Data from Cadoret et al. (2008) and Karnani et al. (2010) were analyzed for overlaps as described in the Supplemental Methods. (*A*,*B*) Comparisons of overlaps between bubble signals and nascent strand calls from Study 1 (Cadoret et al. 2008) or Study 2 (Karnani et al. 2010). (*C*) Overlaps between shared nascent strand calls and bubble-containing fragments. (*D*) Cartoon of overlap possibilities: the orange arrow shows an overlap between nascent strand calls from studies 1 and 2; the red arrows illustrate one or two nascent strands contained within a bubble-containing fragment; the green arrow illustrates an overlap between a single nascent strand and two bubbles; the blue arrow shows concordance among the two nascent strand signals and the bubble containing fragments.

overlap are illustrated by the arrows in the diagram in Table 2D). In an independent study from one of our own laboratories using the same method of preparation (Study 2) (Karnani et al. 2010), 26.6% of 320 nascent strand signals overlap 657 bubble signals, while 10.7% of bubble signals overlapped nascent strand calls (Table 2B). When the two nascent strand datasets were compared with each other, only 16 nascent strand signals out of a total of 603 unique calls from the combined studies overlapped by ≥1 nt (indicated with turquoise and orange arrows in Table 2D). Ten of these directly overlapped bubble-containing fragments (turquoise arrow). Of the 10 shared signals, two fall within isolated bubble fragments, and eight fall within zonal fragments; two of the latter each overlap two adjacent bubble-containing fragments (as indicated with the green arrow in Table 2D). The exact locations of these 10 fragments within the ENCODE regions are provided in Supplemental Table S5.

Discussion

Bubble libraries are remarkably pure, reproducible, and nearly saturating

The pilot early S-phase CHO origin library that we constructed previously contained ~ 5000 clones and appeared to be essentially pure (Mesner et al. 2006). We show here that the method can be scaled up in human cell lines more than 200-fold without significant loss of purity. Analysis of the amplified rDNA locus in the trapped material from early S-phase showed that the preparation was >90% pure, and analysis of the genomic counterparts of 20 anonymous clones in early S-phase cells was consistent with this estimate. In total, 30 (possibly 31) of 34 mammalian origin clones isolated from early S-phase cells by trapping correspond to fragments that initiate in vivo (six or seven out of eight in Fig. 1B; five out of six in Fig. 2B, five out of six from other ENCODE regions, and 14 out of 14 hamster origin clones) (Mesner et al. 2006). It is possible that the purity is even higher, since some cloned fragments may be used infrequently enough in vivo to be undetectable on 2D gels, but nevertheless are trapped in agarose. Based on the rDNA analyses, the log-phase HeLa and GM06990 origin libraries appear to be equally pure. Importantly, these are the only origin libraries that have been biologically validated by a method completely independent of the method of isolation itself.

The microarray hybridizations were quite reproducible (~90% concordance). Furthermore, we have recently sequenced the early S-phase library, and a preliminary comparison suggests that the microarray data gives an accurate picture of library composition. An impressive 77% of microarray signals detected in the smaller log-phase HeLa biological replicate (Rep3) were also detected in the larger library (Rep4). In both comparisons, nonoverlapping fragments were smaller than overlapping ones and exhibited significantly lower signals on the arrays. This finding is predicted based on the shorter dwell times of bubbles in small fragments, resulting in less-efficient trapping. Some of the lower signals may also correspond to fragments with lower initiation frequencies in the genome, which do not always get trapped. This might explain why the purities of rDNA fragments in the two log-phase HeLa replicates appear similar, yet Rep3 manifested 594 signals, while Rep4 yielded 1068.

Distribution of initiation sites in the human genome

The microarray data suggests that 10%–15% of all 6282 EcoRI fragments in the ENCODE regions initiate replication in vivo in early S-phase HeLa, log-phase HeLa, and GM06990 cells. These values correspond to 15.3%, 16.4%, and 21.8% of the genome in these regions, but are somewhat overestimated, since the actual initiation site(s) could constitute only a portion of each fragment if the origin is relatively fixed or the fragment contains a small zone. In addition, the larger number of positive calls for the single GM06990 hybridization (988) relative to the HeLa calls (657) undoubtedly accounts for the larger coverage by this library (Supplemental Table S1). It is also likely that some relatively fixed initiation sites will be excluded from the trapped bubble preparations if they are near a local EcoRI site, since bubbles near the end of a fragment would have a very short half-life relative to those near the center. This factor is less important for fragments arising from initiation zones, because starts will occur in different parts of the fragment in different cells, some of which will be near the center and will be trapped. Because the purity and coverage of each origin preparation depends on complete digestion of matrix-affixed DNA, we have been limited to only a few enzymes that yield total digests, and EcoRI is the most reliable.

More than half of all bubble-containing fragments reside in adjacent clusters of two or more fragments. Fragment numbers range from two to 11 in log-phase HeLa cells, with zonal lengths from 3.1 to 84.6 kb (mean of ~18.1 kb). These findings are consistent with results for most of the origins that have been studied on an individual basis (~2–55 kb in length; for review, see Hamlin et al. 2008). Interestingly, the signal strengths of clustered fragments are higher on average than those from isolated fragments. Thus, fixed initiation sites isolated in single EcoRI fragments generally appear to be less efficient than sites residing in zones. Importantly, initiation zones are not the consequence of synchronization protocols, since clustered EcoRI fragments also constitute significant fractions of log-phase HeLa and GM06990 libraries. The fact that most of the clusters were observed in both log-phase HeLa biological replicates argues strongly that these libraries are close to saturation for all but the least-efficient start sites.

It is somewhat surprising that more bubble-containing fragments are not clustered, because if one considers each cluster as a unit origin, then only ~30% of origins represent zones of initiation sites, with the other 70% corresponding to isolated origins. In fact, most origins studied individually represent initiation zones (Aladjem et al. 2006; Hamlin et al. 2008). Furthermore, studies on 14 anonymous early firing bubble-containing fragments identified in the well-synchronized CHO cell line suggested that all probably arose from initiation zones (Mesner et al. 2006). All but one fragment tested on 2D gels in the present study displayed a composite pattern indicative of fragments residing in zones (Vaughn et al. 1990a). The exceptional fragment displayed a pattern that would be expected of an off-center, fixed origin. Notably, this is the only such pattern detected in a total of 34 independent, anonymous origin clones from human and hamster genomes (Figs. 1, 2; Mesner et al. 2006; LD Mesner, unpubl.). One possible factor here is our method of analyzing the microarray data, which counts only contiguous fragments, when we know that small fragments and/or those containing inefficient initiation sites that might reside within a real cluster are under-represented. An additional likely possibility is that many of the isolated fragments are large enough to contain smaller zones (such as *MYC*, which is ~2 kb in length; Waltz et al. 1996).

Interestingly, the average intervals between origins (single fragments and clusters treated as units) are 58.0 kb, 68.8 kb, and 40.6 kb for early S-phase HeLa, and log-phase HeLa and GM06990,

respectively. These averages are underestimates, since, when measured, the majority of mammalian origins appear to be inefficient, probably firing in only 30%–50% of cell cycles (Heintz and Hamlin 1982; Mesner et al. 2006). However, these estimates are well within the range of replicon sizes measured in fiber autoradiographic studies (15–300 kb) (Huberman and Riggs 1968).

Cell culture conditions and cell lineage significantly alter origin distributions

When the array data from early S-phase and log-phase HeLa origin libraries were compared, there was substantial overlap, but the number of fragments unique to the log-phase library was surprisingly large (58%). This was unexpected, since only a few later-firing origins have been identified to date (e.g., Ma et al. 1990; Larner et al. 1999; Lin et al. 2003). Furthermore, many of these fragments appear to be intimately admixed with early firing fragments (log-early; Fig. 3C). This suggests that if they are, indeed, late-firing, they are somehow protected from read-through by forks from nearby early firing origins or that the origins on different alleles fire at different times.

However, a significant fraction of origins identified in the early S-phase library were not obviously present in the log-phase library. This phenomenon is suggestive of older DNA fiber autoradiographic studies and recent single molecule analyses showing that lowering nucleotide pools (as we have done here—first with thymidine and then with mimosine) can activate some inchoate origins and suppress others (Taylor 1977; Anglana et al. 2003; see Gilbert 2007 for in-depth review). These findings also cast doubt on the suggestion that signals found in log-phase but not in early S-phase libraries necessarily correspond to later-firing origins. They also suggest to us that true early-firing origins will somehow have to be identified by methods that do not rely on replication inhibitors, perhaps by centrifugal elutriation or by cell sorting.

We chose HeLa and GM06990 for making origin libraries because, when these studies were initiated, they were among the few model cell lines chosen for the ENCODE pilot project, which would allow us to take advantage of datasets arising from other laboratories. Although HeLa and GM06990 are not a matched pair, they allowed us to assess the degree to which cells can alter the spectrum of origin usage in very different genetic and physiological states. These data indicate that tumorigenesis in HeLa cells and/or the different developmental lineages of the two cell lines resulted in a surprisingly different spectrum of origin activation. We are now preparing libraries from matched normal and tumorigenic lymphoblastoid cell lines to attempt to uncover the contribution of tumorigenesis to origin selection.

Origin activity is significantly associated with transcriptional activity and permissive chromatin environments

We were surprised to find that there is no clear proclivity to initiate replication in intergenic sequences: 46% and 38% of bubble-containing fragments in HeLa and GM06990 reside fully within the bodies of gene sequences. This is a conservative estimate of genic template usage, as it does not include genic regions in ambiguous fragments in which initiation may have occurred. Nevertheless, the data suggest that any required genetic replicators in genes must have coevolved with gene structure.

Even more remarkable is the very pronounced association between replication initiation and transcription per se: 58% and 25% of zonal bubble-containing fragments lie within or overlap active genes in HeLa and GM06990 cells, respectively, while 32% and 16% of isolated fragments do so. The most significant categories by far were zones that encompassed or overlapped the 5′ ends of transcribed genes. While individual origins have been shown to often reside near the ends of genes (for review, see Aladjem et al. 2006), these studies are biased somewhat, as origin-finding schemes often relied on restriction maps and probes available because of prior interest in the gene. The beauty of the whole-genome approach is its lack of bias. In this regard, our data are entirely consistent with previous studies showing a strong association between purified small nascent DNA strands and transcription start sites in human and murine cells (Lucas et al. 2007; Cadoret et al. 2008; Sequeira-Mendes et al. 2009; Karnani et al. 2010). In addition, recent studies on murine cells undergoing differentiation also implicate the changing transcriptional program in determining, or being entrained with, replication timing, which ultimately depends on origin activation (Hiratani et al. 2004).

Fully 31% and 22% of zonal and isolated fragments lie completely within the bodies of active genes in HeLa cells, while 25% and 16% do so in GM06990. These findings were unexpected, given the many experiments suggesting that the two processes are mutually exclusive on the same template (e.g., Snyder et al. 1988; Haase et al. 1994; Hyrien et al. 1995; Mesner and Hamlin 2005). Note that we have effectively ruled out the possibility that the trapping procedure might also select for D-loops displaced during transcription by showing that the very active *dhfr* gene in CHO cells is not detectably retained in the agarose plug (Mesner et al. 2006). Furthermore, fragments containing other novel looped structures that might have been trapped undoubtedly would migrate differently than canonical bubbles or single forks when the genomic counterpart was tested on 2D gels, since several studies have shown that these gels are remarkably sensitive to only slight alterations in overall structure (Brewer and Fangman 1987; Martin et al. 1991; Schvartzman et al 1993; Kalejta and Hamlin 1996; Kalejta et al 1996).

One possible explanation for intragenic initiation may be that replication initiation and transcription do not occur on the same allele. Alternatively, transcription and initiation of replication may occur at different times in the cell cycle. Indeed, there is ample evidence for both allelic and/or timing differences vis-a-vis initiation of replication and transcription (e.g., Kitsberg et al. 1993; Goren and Cedar 2003; Karnani et al. 2007). An important question for which we presently have no answer is why interactions between genes and zonal origins are so highly significant when those with isolated fragments are only modestly so. An additional question is why there are such significant differences in origin–gene interactions between HeLa and GM06990 cells. Even though a larger percentage of the GM06990 genome is represented by the bubble library (21.8% vs. 16.4%), partial or complete overlap between bubble-containing fragments and actively transcribed genes is less than half that of HeLa cells (20.6% vs. 45.2%, respectively). Perhaps the shear aneuploidy and resulting gene copy number imbalances in the highly tumorigenic HeLa cell line result in noncanonical allelic differences.

Despite the highly significant associations between transcribed genes and origins, transcription per se may not be the sole driver of origin activation, since many origins (whether isolated or zonal) appear to be free of transcription, at least by the strict criterion that RNAs defined by the Affymetrix group had to coincide with annotated genes in our analysis (Table 1). However, low levels of novel transcripts have been detected whose 5′ and 3′ ends have

not been accurately mapped and curated (The ENCODE Project Consortium 2007), and these could contribute to activation of some origins.

Given the very strong association between many origins and actively transcribed genes, it was not surprising to learn that activating marks are significantly associated with bubble-containing fragments in both cell lines. Indeed, only 9.8% and 3.3% of the ENCODE template in HeLa and GM06990 cells represent transcribed genes with activating marks at their promoters, yet 58.6% and 31.1% of initiation zones overlap these genes. Like the relationships with transcription itself, these mechanisms apparently are not the only drivers of origin activation, since many bubble-containing fragments are distributed among all of the other categories, whose associations with origins are not statistically significant (Supplemental Table S4).

Concordance between origin distributions measured with bubble libraries and small nascent strand preparations

When compared with origin maps constructed for log-phase HeLa cells with small nascent strands in Study 1 (Cadoret et al. 2008) and Study 2 (Karnani et al. 2010), only 35.1% and 26.6% of nascent strands fall within bubble-containing fragments, respectively, and only 1.7% of the shared nascent strand signals actually overlapped a bubble (Table 2). The 10 bubble signals shared with the two nascent strand studies do not apparently correspond to "super" origins, since the average $\log_2$ signal is 1.48, whereas the average for all 657 positive bubble fragments in the log-phase HeLa library is 1.46.

We assume that the two datasets from Studies 1 and 2 are both valid. Indeed, ~90% of the overlapping fragments from Study 2 have been shown to be occupied by the ORC complex (Karnani et al. 2010). One likely explanation for the minimal overlap with the bubbles or with each other is that neither preparation is saturating. Furthermore, a given zonal cluster should give rise to many small nascent strands, but only a small percentage of bubble fragments overlapped more than one small nascent strand (Table 2). One factor could be that only ~10^8 cells were utilized for the small nascent strand preparations, whereas the HeLa bubble libraries arose from ~0.5×10^9–1×10^9 cells. However, the preparations in studies 1 and 2 both contained the *MYC* origin, suggesting that each sampled the most active origins in log-phase HeLa cells. In fact, all three origins that have been used as gold standards in nascent strand preparations are enriched in the log-phase HeLa bubble libraries relative to genomic DNA, ranging from approximately fourfold for a beta-globin fragment to an average of ~57-fold and ~50-fold for *MYC* and lamin B2 fragments, respectively (L Wang and JL Hamlin, unpubl.). These values fall within the expected range based on earlier studies (Giacca et al. 1994; Waltz et al. 1996; Wang et al. 2004).

Methods

Cell culture, propagation, and synchronization

HeLa cells were obtained from the American Type Culture Collection and GM06990 cells from the Coriell Institute for Medical Research. Cell culture conditions and the thymidine/mimosine double-block synchronization regimen were exactly as described previously (Mesner et al. 2009). The early S-phase HeLa cultures were harvested 80 min after release from mimosine for preparation of replication intermediates (RIs).

Bubble trapping and origin library preparation

RIs were purified ~200-fold exactly as described (Mesner et al. 2009) using EcoRI to digest the DNA. Bubble-containing EcoRI fragments were isolated as previously described (Mesner et al. 2006; Mesner and Hamlin 2009). After trapping in and recovery from the agarose, ~20% of the sample was used for 2D gel analysis and the remainder was cloned into the EcoRI site of pGem7 (Promega). Library clones were pooled and aliquoted after minimal growth on liquid medium and plates to avoid the propagation of siblings and to reduce the differential effects of clone size and composition on bacterial growth rate. Approximately 10^6 individual clones with inserts were obtained for each of the four libraries (one early S-phase HeLa library, two log-phase HeLa libraries, and one log-phase GM06990 library).

Validation of trapped and cloned bubble-containing fragments on 2D gels

The trapped material, genomic counterparts of several anonymous clones from the early S-phase library, as well as cognate genomic EcoRI fragments for several selected positive microarray origin candidates were validated by analysis on 2D gels (Brewer and Fangman 1987) exactly as previously described (Mesner et al. 2009). For the trapped material, a probe from the amplified rDNA origin was used for detection (Little et al. 1993); nonrepetitive probes for individual anonymous clones were generated by PCR amplification of library or genomic DNA using homologous primers developed for the cognate fragments with the BLAST program (http://www.ncbi.nlm.nih.gov).

Microarray hybridization

DNA from the pooled libraries, as well as a total log-phase genomic control DNA sample for the relevant cell lines, were purified by standard methods. For the early S-phase HeLa library, the DNA was labeled either with biotinylated dATP by random priming (Rp1) or with ddATP using terminal transferase (Tt1), and each labeled preparation was hybridized once to Affymetrix genomic tiling arrays as previously described (Jeon et al. 2005). In subsequent experiments with the log-phase HeLa (Rep3 and Rep4) and GM06990 libraries and their respective genomic controls, DNA samples were labeled with ddATP and each preparation was hybridized once to the microarrays.

Statistical analyses

See Supplemental Methods.

Acknowledgments

The corresponding principal investigators for the wet-bench and computational aspects of this study are J.L.H. and S.B., respectively. We thank Pieter Dijkwel and Rebecca Pickin for critical reading of the manuscript and the other members of our laboratories for helpful discussions. This work was supported by RO1-HG002937 (J.L.H.), RO1-GM26108 (J.L.H.), U01-HG03157 (A.D.), and R01-CA60499 (A.D.) from the NIH.

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:** 2023–2026.

Aladjem MI, Fanning E. 2004. The replicon revisited: An old model learns new tricks in metazoan chromosomes. *EMBO Rep* **5:** 686–691.

Aladjem MI, Falaschi A, Kowalski D. 2006. Eukaryotic DNA replication origins. (ed. M. DePamphilis), pp. 31–62. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Anglana M, Apiou F, Bensimon A, Debatisse M. 2003. Dynamics of DNA replication in mammalian somatic cells: Nucleotide pool modulates origin choice and interorigin spacing. *Cell* **114:** 385–394.
Bielinsky AK, Gerbi SA. 1998. Discrete start sites for DNA synthesis in the yeast ARS1 origin. *Science* **279:** 95–98.
Brewer BJ, Fangman WL. 1987. The localization of replication origins on ARS plasmids in *S. cerevisiae*. *Cell* **51:** 463–471.
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. *Proc Natl Acad Sci* **105:** 15837–15842.
Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116:** 499–509.
Chan CS, Tye BK. 1980. Autonomously replicating sequences in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci* **77:** 6329–6333.
Diffley JF, Cocker JH, Dowell SJ, Rowley A. 1994. Two steps in the assembly of complexes at yeast replication origins *in vivo*. *Cell* **78:** 303–316.
Dijkwel PA, Hamlin JL. 1992. Initiation of DNA replication in the dihydrofolate reductase locus is confined to the early S period in CHO cells synchronized with the plant amino acid mimosine. *Mol Cell Biol* **12:** 3715–3722.
Dijkwel PA, Vaughn JP, Hamlin JL. 1991. Mapping of replication initiation sites in mammalian genomes by two-dimensional gel analysis: Stabilization and enrichment of replication intermediates by isolation on the nuclear matrix. *Mol Cell Biol* **11:** 3850–3859.
The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.
Giacca M, Zentilin L, Norio P, Diviacco S, Dimitrova D, Contreas G, Biamonti G, Perini G, Weighardt F, Riva S. 1994. Fine mapping of a replication origin of human DNA. *Proc Natl Acad Sci* **91:** 7119–7123.
Gilbert DM. 2005. Origins go plastic. *Mol Cell* **20:** 657–658.
Gilbert DM. 2007. Replication plasticity, Taylor-made: Inhibition vs recruitment of origins under conditions of replication stress. *Chromosoma* **116:** 341–347.
Goren A, Cedar H. 2003. Replicating by the clock. *Nat Rev Mol Cell Biol* **4:** 25–32.
Haase SB, Heinzel SS, Calos MP. 1994. Transcription inhibits the replication of autonomously replicating plasmids in human cells. *Mol Cell Biol* **14:** 2516–2524.
Hamlin JL, Mesner LD, Lar O, Torres R, Chodaparambil SV, Wang L. 2008. A revisionist replicon model for higher eukaryotic genomes. *J Cell Biochem* **105:** 321–329.
Heintz NH, Hamlin JL. 1982. An amplified chromosomal sequence that includes the gene for dihydrofolate reductase initiates replication within specific restriction fragments. *Proc Natl Acad Sci* **79:** 4083–4087.
Hiratani I, Leskovar A, Gilbert DM. 2004. Differentiation-induced replication-timing changes are restricted to AT-rich/long interspersed nuclear element (LINE)-rich isochores. *Proc Natl Acad Sci* **101:** 16861–16866.
Huberman JA, Riggs AD. 1968. On the mechanism of DNA replication in mammalian chromosomes. *J Mol Biol* **32:** 327–341.
Hyrien O, Maric C, Mechali M. 1995. Transition in specification of embryonic metazoan DNA replication origins. *Science* **270:** 994–997.
Jeon Y, Bekiranov S, Karnani N, Kapranov P, Ghosh S, MacAlpine D, Lee C, Hwang DS, Gingeras TR, Dutta A. 2005. Temporal profile of replication of human chromosomes. *Proc Natl Acad Sci* **102:** 6419–6424.
Kalejta RF, Hamlin JL. 1996. Composite patterns in neutral/neutral two-dimensional gels demonstrate inefficient replication origin usage. *Mol Cell Biol* **16:** 4915–4922.
Kalejta RF, Lin HB, Dijkwel PA, Hamlin JL. 1996. Characterizing replication intermediates in amplified CHO dihydrofolate reductase domain by two novel gel electrophoretic techniques. *Mol Cell Biol* **16:** 4923–4931.
Karnani N, Taylor C, Malhotra A, Dutta A. 2007. Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res* **17:** 865–876.
Karnani N, Taylor CM, Malhotra A, Dutta A. 2010. Genomic study of replication initiation in human chromosomes reveals the influence of transcription regulation and chromatin structure on origin selection. *Mol Biol Cell* **21:** 393–404.
Kitsberg D, Selig S, Brandeis M, Simon I, Keshet I, Driscoll DJ, Nicholls RD, Cedar H. 1993. Allele-specific replication timing of imprinted gene regions. *Nature* **364:** 459–463.
Larner JM, Lee H, Little RD, Dijkwel PA, Schildkraut CL, Hamlin JL. 1999. Radiation down-regulates replication origin activity throughout the S phase in mammalian cells. *Nucleic Acids Res* **27:** 803–809.
Lin CM, Fu H, Martinovsky M, Bouhassira E, Aladjem MI. 2003. Dynamic alterations of replication timing in mammalian cells. *Curr Biol* **13:** 1019–1028.
Little RD, Platt TH, Schildkraut CL. 1993. Initiation and termination of DNA replication in human rRNA genes. *Mol Cell Biol* **13:** 6600–6613.
Lucas I, Palakodeti A, Jiang Y, Young DJ, Jiang N, Fernald AA, Le Beau MM. 2007. High-throughput mapping of origins of replication in human cells. *EMBO Rep* **8:** 770–777.
Ma C, Leu TH, Hamlin JL. 1990. Multiple origins of replication in the dihydrofolate reductase amplicons of a methotrexate-resistant chinese hamster cell line. *Mol Cell Biol* **10:** 1338–1346.
Martin PL, Hernandez P, Martinez-Robles MI, Schvartzman JB. 1991. Unidirectional replication as visualized by two-dimensional agarose gel electrophoresis. *J Mol Biol* **220:** 843–853.
Mesner LD, Hamlin JL. 2005. Specific signals at the 3′ end of the *DHFR* gene define one boundary of the downstream origin of replication. *Genes Dev* **19:** 1053–1066.
Mesner LD, Hamlin JL. 2009. Isolation of restriction fragments containing origins of replication from complex genomes. *Methods Mol Biol* **521:** 315–328.
Mesner LD, Crawford EL, Hamlin JL. 2006. Isolating apparently pure libraries of replication origins from complex genomes. *Mol Cell* **21:** 719–726.
Mesner LD, Dijkwel PA, Hamlin JL. 2009. Purification of restriction fragments containing replication intermediates from complex genomes for 2-D gel analysis. *Methods Mol Biol* **521:** 121–137.
Mitsis PG, Kwagh JG. 1999. Characterization of the interaction of lambda exonuclease with the ends of DNA. *Nucleic Acids Res* **27:** 3057–3063.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35:** D61–D65.
Romero J, Lee H. 2008. Asymmetric bidirectional replication at the human DBF4 origin. *Nat Struct Mol Biol* **15:** 722–729.
Schvartzman JB, Martinez-Robles MI, Hernandez P. 1993. The migration behaviour of DNA replicative intermediates containing an internal bubble analyzed by two-dimensional agarose gel electrophoresis. *Nucleic Acids Res* **21:** 5474–5479.
Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gomez M. 2009. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genetics* **5:** e1000446. doi: 10.1371/journal.pgen.1000446.
Snyder M, Sapolsky RJ, Davis RW. 1988. Transcription interferes with elements important for chromosome maintenance in *Saccharomyces cerevisiae*. *Mol Cell Biol* **8:** 2184–2194.
Stinchcomb DT, Thomas M, Kelly J, Selker E, Davis RW. 1980. Eukaryotic DNA segments capable of autonomous replication in yeast. *Proc Natl Acad Sci* **77:** 4559–4563.
Taylor JH. 1977. Increase in DNA replication sites in cells held at the beginning of S phase. *Chromosoma* **62:** 291–300.
Vaughn JP, Dijkwel PA, Hamlin JL. 1990a. Replication initiates in a broad zone in the amplified CHO dihydrofolate reductase domain. *Cell* **61:** 1075–1087.
Vaughn JP, Dijkwel PA, Mullenders LH, Hamlin JL. 1990b. Replication forks are associated with the nuclear matrix. *Nucleic Acids Res* **18:** 1965–1969.
Waltz SE, Trivedi AA, Leffak M. 1996. DNA replication initiates non-randomly at multiple sites near the c-myc gene in HeLa cells. *Nucleic Acids Res* **24:** 1887–1894.
Wang L, Lin CM, Brooks S, Cimbora D, Groudine M, Aladjem MI. 2004. The human beta-globin replication initiation region consists of two modular independent replicators. *Mol Cell Biol* **24:** 3373–3386.

Received June 4, 2010; accepted in revised form November 29, 2010.