Supporting Information

Hansen et al. 10.1073/pnas.0912402107

Methods

Cell Culture and RNA Analysis. The University of Washington Institutional Review Board/Human Subjects Review Board approved these studies for human subjects. Anonymous normal fibroblast cultures (BJ), normal lymphoblastoid cell lines (GM06990 and H0287), and a normal lymphoblastoid cell line from an ICF carrier (GM08729) were purchased from the Coriell Cell Repositories (Camden, NJ), or furnished previously by W. Raskind (University of Washington). TL010 is a male lymphoblastoid line described previously (1). Standard growth conditions for lymphoblastoid cells and fibroblasts were as described (2, 3). K562 is a female chronic myelogenous leukemia cell line with erythroid properties that was obtained from the American Tissue Culture Collection (ATCC, Manassas, VA) and grown under ATCC-recommended conditions. NIH-approved human embryonic stem cells (BG02; male) were obtained from BresaGen (Athens, GA) and were cultivated at the University of Washington embryonic stem cell core facility under feeder-free conditions (4). All cells were grown in exponential phase for replication timing and gene expression studies.

At cell harvest, a subset of cells was stored at −20°C in RNALater (Ambion, Austin, TX). Total RNA was purified using Ribopure (Ambion) according to vendor-recommended protocols. Total RNA quality was assessed on RNA 6000 nano chips using a bioanalyzer (Agilent, Santa Clara, CA). Approximately 3 μg of total RNA for each cell type was sent to the University of Washington Center for Array Technology for labeling and hybridization to Affymetrix Human Exon 1.0 ST arrays (Affymetrix, Santa Clara, CA). A Whole Transcript Sense Target Labeling Assay (Affymetrix) was used to reduce rRNA, perform in vitro transcription, and create labeled sense strand DNA for hybridization to the arrays. Hybridizations were carried out according to the manufacturer's protocol. Intensity files from the scanned exon arrays were provided by the array facility and exon expression data were analyzed using Affymetrix ExACT 1.2.1 software. Samples were quantile normalized with PM-GCBG background correction and PLIER (probe logarithmic intensity error) summarized. The data are available from the NCBI GEO database under accession numbers GSM472898, GSM472903, GSM472910, GSM472944, and GSM472945.

Repli-Seq Procedure. Repli-Seq is based on a previously described STS replication assay (1, 5) and has been modified for sequence analysis of newly replicated DNA. The procedure is similar for attached and unattached cell types. Briefly, 5–20 million cultured cells in exponential growth phase were incubated with 50 μM BrdU in growth medium so as to label the newly replicating DNA in vivo (6). Cells were harvested 1–1.5 h later either by direct centrifugation for unattached cells or by trypsinization and centrifugation for attached cells. After washing the cells, their DNA was stained with DAPI (4,6-diamidino-2-phenylindole) in nonionic detergent (Nonidet P-40), and this material was syringed several times and filtered in preparation for flow cytometry. The staining solution also contains 10% dimethyl sulfoxide for freezing extra cells at −70°C for later flow sorting, if needed.

Flow cytometry was performed either on an EPICS Elite (1) or an Influx (Cytopeia/BD Biosciences, San Jose, CA) flow cytometer and sorting was performed according to the contiguous DAPI fluorescence channels described in Fig. 1 (G1a, G1b, S1, S2, S3, S4, and G2/M; the G1a fractions generally had high backgrounds and were not processed for sequencing). Because different lineages have different cell-cycle distributions and cytometry patterns, the standard method for DAPI channel definitions is to divide the G1 peak-to-G2/M peak value by 6 for the G1b- and S-phase increments in DAPI content (starting from the G1 peak); G1a is one half of this increment and G2 is twice this increment.

The number of cells sorted per channel varied among the different cell lines but was typically 50,000. The cells were sorted two channels at a time into tubes containing cell lysis buffer over a period of about 90 min for all channels to be sorted (1). DNA was then purified from the lysates by phenol extraction and the BrdU-DNA was isolated following heat denaturation by immunoprecipitation using an anti-BrdU monoclonal antibody (BD Biosciences). Following purification from the immunoprecipitate, the newly replicated single-stranded BrdU-DNA was then made double-stranded by random-primed extension with a random hexamer labeling kit (Invitrogen, Carlsbad, CA). The purified material was quantified with PicoGreen staining (Invitrogen) using a NanoDrop 3300 fluorospectrometer (Thermo Scientific, Waltham, MA).

Illumina sequencing libraries were constructed from the random-primed BrdU-DNA material according to standard procedures. Illumina's genomic prep kit protocol was used with minor modifications (Illumina, Hayward, CA): 10–50 ng of purified random-primed products was subjected to end repair by combining fractionated DNA, 1× T4 DNA ligase buffer (containing 0.1 mM ATP), 0.4 mM dNTP mix, 15 U T4 DNA polymerase, 5 U DNA polymerase (Klenow, large fragment), and 50 U T4 polynucleotide kinase (New England Biolabs, Ipswich, MA). The reaction mixture was incubated at 20°C for 30 min and purified using a MinElute micro spin column (Qiagen, Valencia, CA). Following repair, nontemplated adenines were added to the 3′ ends of purified fragments with 1 mM dATP and 5 U Klenow fragment (3′-5′ exo-minus) (NEB) and incubated at 37°C for 30 min.

After column purification, adapters (Illumina) were ligated to the ends of DNA fragments in a 50-μL reaction by combining 17.5 μL "A-tailed" fragments, 25 μL 2× Quick DNA ligase buffer (NEB), 2.5 μL adapter oligo mix (Illumina), and 5 μL Quick DNA ligase (NEB) and incubated for 15 min at room temperature. The adapter-ligated material was then purified with MinElute columns (Qiagen) and enriched by a limited amplification: 50 μL reactions consisting of 5 μL adapter-ligated DNA (equivalent to about one tenth of the ligated material), 25 μL 2× Phusion High-Fidelity PCR Master Mix (NEB), 0.5 μL each of primers 1.1 and 2.1 (Illumina), and 19 μL nuclease-free water were assembled and cycled with the following thermal profile: 98°C for 30 min, followed by 16 cycles of 98°C for 10 min, 65°C for 30 min, and 72°C for 30 min, with a final extension at 72°C for 5 min.

Amplified libraries were purified with a MinElute micro spin column (Qiagen) and size-fractionated using agarose electrophoresis. Fragments in the 250–500 bp range were purified from the gel using the Qiagen gel-extraction kit according to the manufacturer's recommendations and quantified with PicoGreen staining. One sixth of each enriched library was sequenced in a single lane on a Genome Analyzer (Illumina) using standard sequencing-by-synthesis technology.

Sequence Analysis and Sequence Tag Densities. Raw sequencing data were analyzed using the Illumina data analysis pipeline. Standard parameterization of the pipeline was used for image analysis, base calling, and alignment. Twenty-seven-base reads are evaluated and aligned to the human reference genome with and without chromosome Y removed (hg18) using version 0.3.0 of ELAND (Efficient Large-Scale Alignment of Nucleotide Databases) (Illumina). The removal of chromosome Y facilitated ascertainment of replication patterns in the p- and q-terminal pseudoautosomal regions of chromosome X in both male and female cells. Unique reads containing up to two mismatches were mapped to the genome. Per lane, uniquely mapping sequences were typically about 4 million, but additional lanes were sequenced and combined if they were much below 2 million (some G1 and G2 fractions). The total sequence counts for each cell line's fractions are given in Table S1 for the "no-Y" genome maps. Mapped sequence tags containing simple repeats and other low-complexity sequences appeared to be nonspecific background regions ("bad spots") and were removed by calculating sequence tag densities in 150 bp windows and removing tags within windows containing five or more tags.

After filtering bad spots, the density of BrdU-DNA-derived sequence tags along the genome was calculated for each cell-cycle fraction using 50 kb sliding windows at 1 kb intervals. These tag densities were then normalized to a global density of 4 million tags per genome for each fraction [because flow cytometry windows suggest that the G1b- and S-phase fractions represent equivalent increases in DNA content, and the number of S-phase cells in the G2/M fraction is similar to the others by cell-cycle analysis (MultiCycle AV software, Phoenix Flow Systems, San Diego, CA)]. To avoid potential variability in signal and background related to tag mapability variation, sequence bias, or copy-number differences, we further normalized the 50 kb densities of each cell line to a percentage of total replication for that line at each genomic coordinate [as was originally done for our STS-based replication time analysis (5)]. These percent-normalized sequence tag density values (PNDV) were then used for visualization of replication time patterns on a mirror of the UCSC Genome Browser and for computational analysis as described below. Mitochondrial sequence tag counts were not used for normalization because non-S-phase nuclei contribute a major, yet variable, portion of the G1 and G2 signals, and also the retention of mitochondrial replication signal

in the sorted DAPI-stained nuclei was more variable than we previously found with propidium iodide-stained cells (1).

Validation of Repli-Seq Using an STS-Based Replication Assay. BrdU-DNA isolated from the cell-cycle fractions was amplified with locus-specific primers and analyzed by Southern blot autoradiography according to a standard STS-based replication timing assay (1, 5). Primer sequences and PCR parameters for STSs are available on request. Allele-specific *SNRPN* replication timing expression was determined as previously described (7).

Analysis of Lineage-Specific Differences in Replication Timing. To simplify the computational search for variations in replication timing in different cell lines, we combined the PNDV for G1 and S1 for each 1000 bp window of the genome to yield a cumulative "early" replication signal for each cell type. Similarly, a "late" signal was calculated by adding the PNDV for S4 and G2 for each position and for each cell type. To avoid spurious signals, regions were removed from analysis if they contained low replication signals in any cell line using a cutoff threshold of 50 sequence tags per 50 kb (from the tag densities that were normalized to 4 million tags per genome). The masked regions also included gaps (with 10 kb pads), segmental duplications, and the entire Y chromosome (not evaluated for replication time as described above). The filtered region for this analysis is provided in Table S2.

The absolute replication timing (ART) was calculated by taking the natural log of the ratio of early to late replication timing for each cell type, yielding a negative number with positions that replicate late and a positive number for early-replication portions of the genome. Replication time variability between cell lines was examined by calculating pairwise ART differences for all combinations; absolute differences greater than 1.8 were considered significant based on background calculations (Table S2).

Relative replication timing (RRT) was developed as a way to compare the replication of one cell type to another at significantly different sites. The RRT for any cell type at a given position was calculated by adding the number of cell types that have significantly later replication timing and subtracting the number of cell types that have significantly earlier replication timing. A positive RRT at any position indicates a cell type has relatively earlier replication than most other cell types. Regions with no significant differences between cell types were classified as "constant" in their replication timing. These constant sites were further divided into pan-lineage-early/constant-early, pan-lineage-late/constant-late, and pan-lineage-mid/constant-mid categories depending on whether the average ART score at each site was greater than 1, less than −1, or in between, respectively.

Correlation of Lineage-Specific Replication Timing with Lineage-Specific Transcriptional Differences. The positions of probes from the Affymetrix Human Exon 1.0 ST array were mapped onto the 1-kb genomic windows that were found to show significant replication timing differences. If a window contained at least five probes and had valid expression data for all cell types examined, it was included in analysis; for each cell line, a single expression score was calculated for each included window that represents the average expression of all probes mapped to that window. For each cell line the distribution of expression was examined, and sites displaying significantly high or low expression relative to the median expression for all cell types at each position (based on a normal distribution) were retained for each cell type. Box plots correlating these significant differences in expression to RRT were then generated using an R software script.

Replication Initiation Zones/Very Early Replication. To estimate the minimum number of early-replication initiation zones present in the genome for each cell line, the percent-normalized G1b regions were searched for regions with 24% or more of the total replication signal. These regions were padded by 10 kb and merged. Merged segments less than 30 kb or those within 25 kb of gaps were filtered out to reduce spurious signals. Also filtered out were G1 segments having ≥40% overlap with segmental duplications whose fraction of matching bases elsewhere in the genome is ≥98%. The G1 regions were analyzed and compared between different cell lineages using the UCSC Table Browser "summary/statistics" function on the custom browser tracks (8). G1 segments in K562 were not included for analysis (inclusion significantly limits the regions analyzed because of the many chromosomal deletions and other abnormalities that make the percent-normalized G1 background higher than in the other lines).

Biphasic Replication Domains. To determine regions where replication time is biphasic, cell-cycle PNDV signals were added in pairs to generate expanded replication time categories (G1+S1, S1+S2, S2+S3, S3+S4, S4+G2). A biphasic region was defined as a 1 kb bin that contained at least 40% of the total replication signal in each of two nonadjacent expanded replication time categories. For example, this condition is met for a specific cell type at a specific position if G1+S1 > 40% and S2+S3 > 40%, but S1+S2 < 40%. This measure was calculated for all possible combinations of expanded replication time tracks.

Chromatin Accessibility/DNaseI Sensitivity. To compare replication timing with chromatin sensitivity to DNaseI digestion, we digested BJ fibroblast and K562 erythroid nuclei according to procedures previously described (9). Size-fractionated double-cut fragments were processed into Illumina sequencing libraries, sequenced, and mapped to the genome (10), with mapping for the human genome as described above for Repli-Seq. Approximately 20 million uniquely mapping 27 bp genomic reads were obtained for each cell type, and the density of DNaseI cleavages in a 150 bp sliding window (step 20) was calculated. Density values were used for comparisons with Repli-Seq data. The data are available as released from the ENCODE Project through the UCSC Genome Browser at http://genome.ucsc.edu (BJ fibroblasts: subId = 295, narrow peak data; K562 erythroid cells: subId = 106, narrow peak data).

Genomic Feature Analysis of Different Replication Classes. Enrichments for particular genomic features within different replication classes (Table S3) were calculated as follows. For any given set of features (e.g., RefSeq transcription starts) of total number, N, the observed number, n, is counted in the locations of interest (e.g., biphasic regions). The expected number in the same locations is computed under a null assumption approximated by assuming a uniform distribution of the elements across the genome (minus the masked regions that have low or problematic mapability). The expected number is thus $m = N \times (L/G)$, where L is the cumulative length of the regions of interest and G is the size of the adjusted genome. The enrichment or depletion over what is expected is then $(n - m)/m$. Most features were obtained from current hg18 tables available from UCSC; recombination hot spots (HapMap release 21, combined Phase I and Phase II) were downloaded from http://www.hapmap.org and converted to hg18 using the LiftOver utility (UCSC).

1. Hansen RS, et al. (1997) A variable domain of delayed replication in FRAXA fragile X chromosomes: X inactivation-like spread of late replication. *Proc Natl Acad Sci USA* 94: 4587–4592.
2. Hansen RS, Canfield TK, Stanek AM, Keitges EA, Gartler SM (1998) Reactivation of XIST in normal fibroblasts and a somatic cell hybrid: Abnormal localization of XIST RNA in hybrid cells. *Proc Natl Acad Sci USA* 95:5133–5138.
3. Hansen RS, Ellis NA, Gartler SM (1988) Demethylation of specific sites in the 5′ region of the inactive X-linked human phosphoglycerate kinase gene correlates with the appearance of nuclease sensitivity and gene expression. *Mol Cell Biol* 8: 4692–4699.
4. Wang L, et al. (2007) Self-renewal of human embryonic stem cells requires insulin-like growth factor-1 receptor and ERBB2 receptor signaling. *Blood* 110:4111–4119.
5. Hansen RS, Canfield TK, Lamb MM, Gartler SM, Laird CD (1993) Association of fragile X syndrome with delayed replication of the FMR1 gene. *Cell* 73:1403–1409.
6. Vassilev LT, Burhans WC, DePamphilis ML (1990) Mapping an origin of DNA replication at a single-copy locus in exponentially proliferating mammalian cells. *Mol Cell Biol* 10:4685–4689.
7. Kawame H, Gartler SM, Hansen RS (1995) Allele-specific replication timing in imprinted domains: absence of asynchrony at several loci. *Hum Mol Genet* 4:2287–2293.
8. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32 (Database issue):D493–D496.
9. Sabo PJ, et al. (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* 3:511–518.
10. Hesselberth JR, et al. (2009) Global mapping of protien-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 6:283–289.

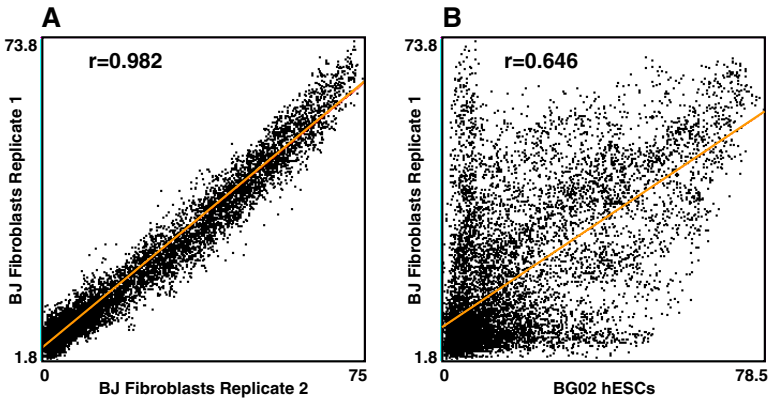

Fig. S1. Repli-Seq reproducibility. (*A*) BJ1-BJ2 biological replicate replication timing comparison. Repli-Seq was applied to two biological replicates of BJ fibroblasts and comparable results were obtained. Shown is the scatter plot of BJ1-BJ2 comparisons of G1B data along chromosome 11 showing a high degree of correlation ($r^2 = 0.965$). The plot was generated with the correlation function of the UCSC Table Browser using 10-kb windows (8). (*B*) BJ1-BG02 replication timing comparison. A similar plot for BJ1-BG02 comparison reveals much more divergence ($r^2 = 0.417$) than in the BJ1-BJ2 comparison.

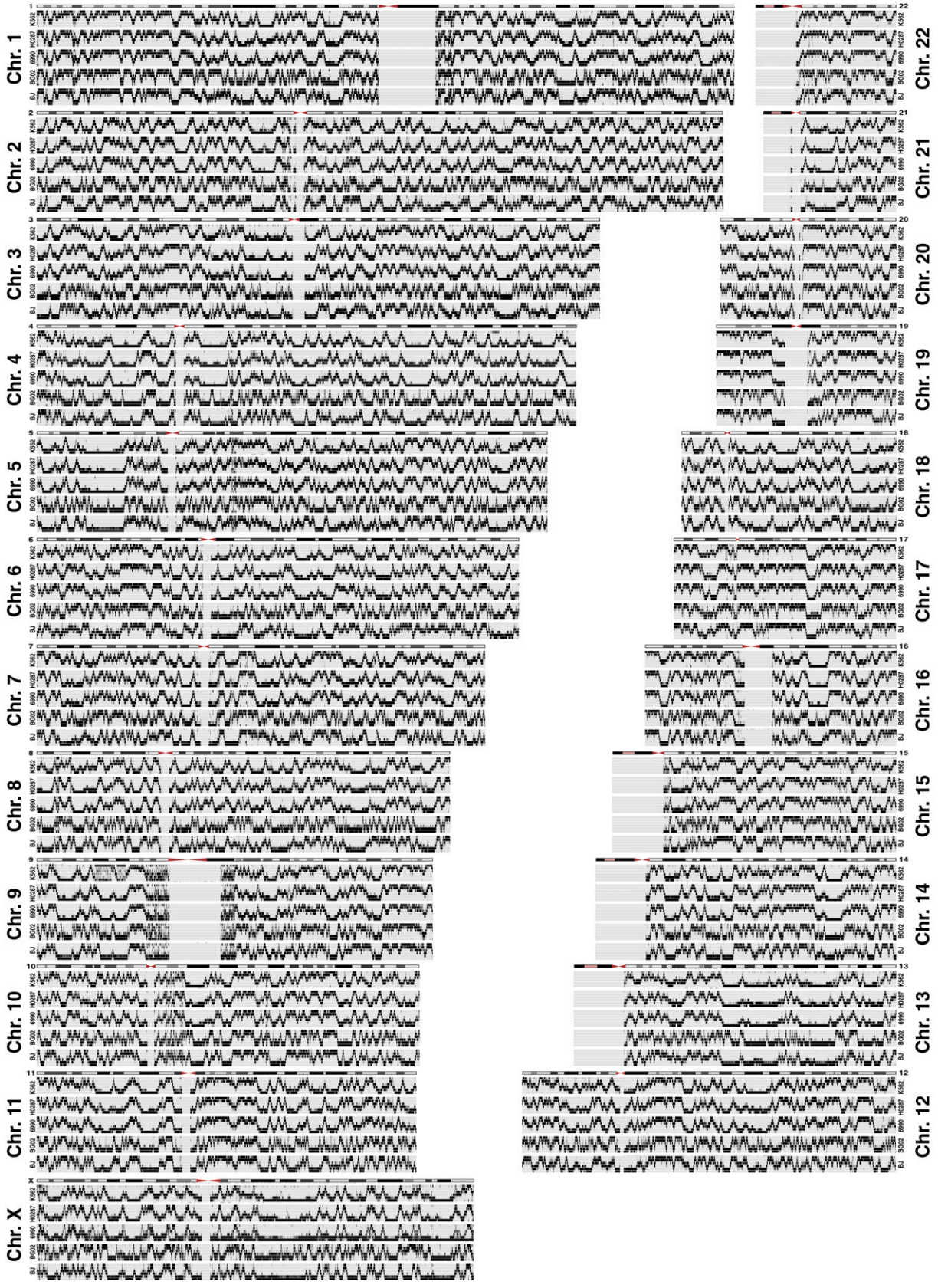

Fig. S2. Whole genome replication patterns in five cell lines. Repli-Seq patterns for all chromosomes are shown as compressed views of percentage-normalized data (without using the Table S2 filter).

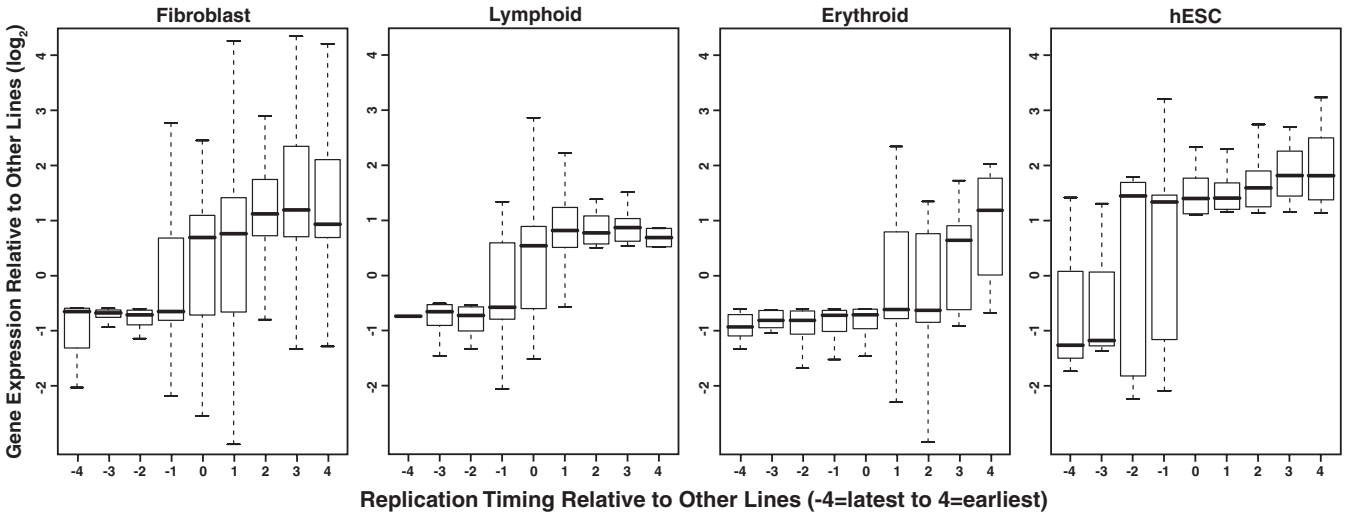

Fig. S3. Lineage-specific expression differences in regions with plastic replication timing. Shown are box plots correlating significant differences in expression level to relative replication timing for each of four cell types. RNA expression data (Affymetrix exon arrays) were ascertained in regions of plastic replication timing for each cell type and exons displaying significantly high or low expression levels relative to the median expression for all cell types were identified. Note that both lineage-specific early and late replication are highly correlated with lineage-specific gene expression and repression (respectively).

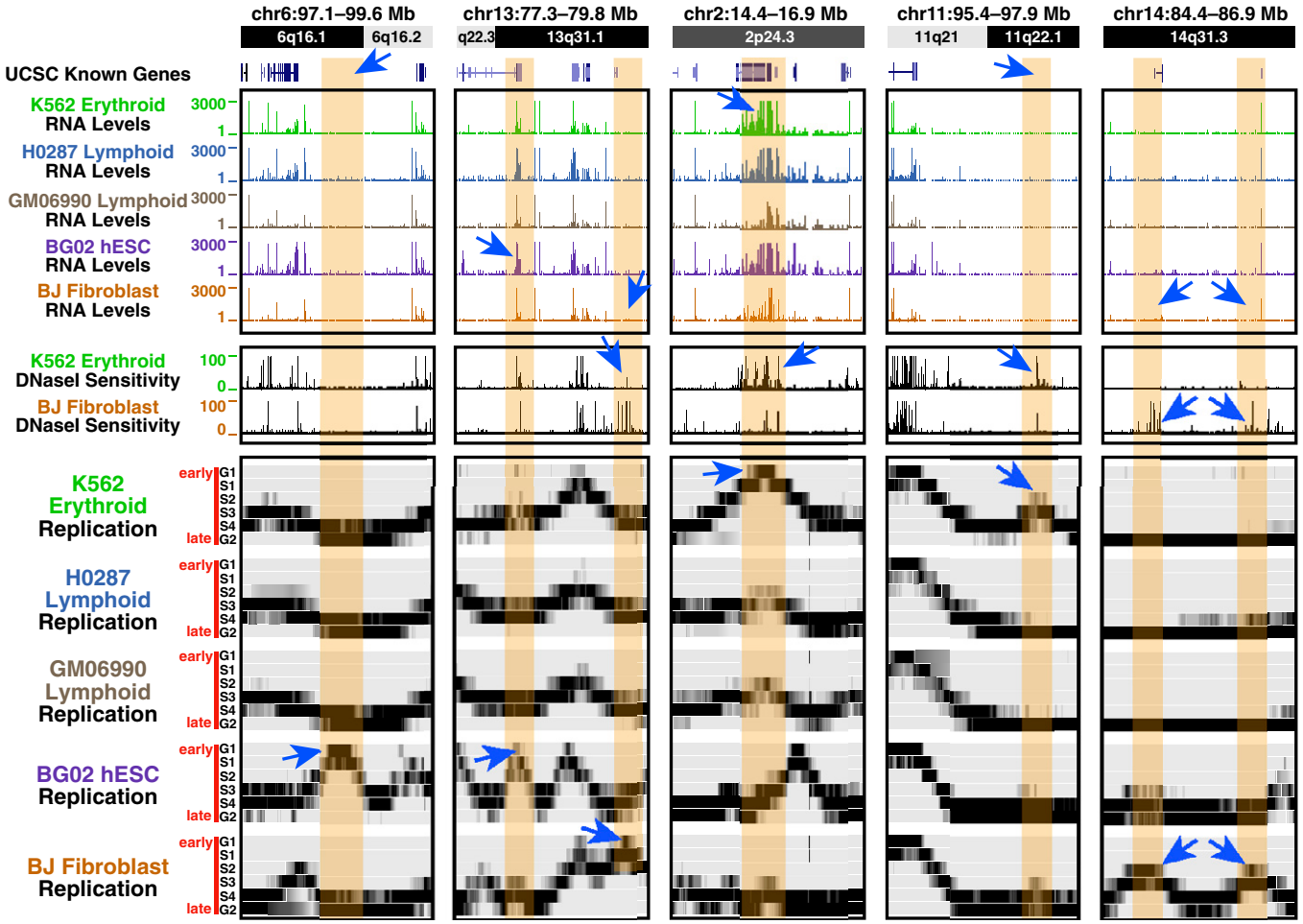

Fig. S4. Tissue-specific replication timing: DNaseI sensitivity concordance and gene expression discordance. Shown are examples where lineage-specific patterns of early or late replication are not correlated with exon RNA levels but are generally concordant with DNaseI sensitivity patterns (arrows highlight relevant features). In the first panel, the BG02-specific early replication peak is completely within a large gene desert. In the second panel, the first early peak in BG02 is not associated with increased expression relative to other lineages that also have high expression even though they replicate later than BG02. The last early peak in BJ is lineage-specific and associated with increased DNaseI sensitivity, but not with changes in exon RNA in this gene-poor region. In the third panel, the K562-specific early peak is associated with increased DNaseI sensitivity compared with the late replicating BJ fibroblasts, but RNA levels are high in both. In the fourth panel, a peak of mid-S replication specific to K562 in a gene desert identifies a later-replicating initiation zone and it associated with increased DNaseI sensitivity. In panel 5, two mid-S peaks specific to BJ fibroblasts identify later-activated initiation zones and are associated with increased DNaseI sensitivity without RNA expression.

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (DOC)
Table S4 (DOC)
Table S5 (DOC)