Genome-scale analysis of replication timing: from bench to bioinformatics

Tyrone Ryba[1], Dana Battaglia[1], Benjamin D Pope[1], Ichiro Hiratani[2] & David M Gilbert[1]

[1]Department of Biological Science, Florida State University, Tallahassee, Florida, USA. [2]Biological Macromolecules Laboratory, National Institute of Genetics, Mishima, Japan. Correspondence should be addressed to D.M.G. (gilbert@bio.fsu.edu).

Published online 2 June 2011; doi:10.1038/nprot.2011.328

Replication timing profiles are cell type–specific and reflect genome organization changes during differentiation. In this protocol, we describe how to analyze genome-wide replication timing (RT) in mammalian cells. Asynchronously cycling cells are pulse labeled with the nucleotide analog 5-bromo-2-deoxyuridine (BrdU) and sorted into S-phase fractions on the basis of DNA content using flow cytometry. BrdU-labeled DNA from each fraction is immunoprecipitated, amplified, differentially labeled and co-hybridized to a whole-genome comparative genomic hybridization microarray, which is currently more cost effective than high-throughput sequencing and equally capable of resolving features at the biologically relevant level of tens to hundreds of kilobases. We also present a guide to analyzing the resulting data sets based on methods we use routinely. Subjects include normalization, scaling and data quality measures, LOESS (local polynomial) smoothing of RT values, segmentation of data into domains and assignment of timing values to gene promoters. Finally, we cover clustering methods and means to relate changes in the replication program to gene expression and other genetic and epigenetic data sets. Some experience with R or similar programming languages is assumed. All together, the protocol takes ~3 weeks per batch of samples.

INTRODUCTION

Although the mechanisms that specify the timing and placement of origin firing in higher eukaryotes remain a mystery, all eukaryotes have a defined RT program that is largely conserved between closely related species[1], including humans and mice[2,3]. Analyses of RT in various cell types have yielded insights into genome organization and repackaging events during development, suggesting an important role for the timing program itself or for 3D genome organization in regulating developmental gene expression[1,3,4]. In this protocol, we describe approaches for measuring genome-wide RT. As data processing and analysis often cause a bottleneck in these studies, the protocol also covers methods used routinely in our laboratory for downstream analysis[3,5,6]. Although this protocol emphasizes mammalian cells, as applied to analyze RT changes in various mouse and human cell types[3,5,6], it can be adapted to any proliferating cell type; such variations have been used to analyze RT in *Drosophila*[7–9], *Arabidopsis*[10] and budding yeast[11].

Overview of the procedure: generating experimental data (Steps 1–61)

The first portion of the protocol describes how to derive raw data for genome-wide RT analysis. Given that the protocol measures the timing of events during the cell cycle, some form of synchronization is required. Synchronization can be achieved either prospectively, before cell collection, or retroactively, after the cells have been collected. In yeasts, prospective synchrony methods are well established, and in many cases, the same method can be used to compare different strains[12,13]. However, most synchronization schemes for multicellular organisms are cumbersome and optimized for specific cell lines[14–16], and most require the use of metabolic inhibitors that can interfere with normal regulation of replication[17,18]. By contrast, retroactive synchronization using fluorescence-activated cell sorting (FACS) to select cells based on the increase in DNA content during S phase can be applied to any proliferating cell population without the need for any previous manipulation beyond dissociation of cells into a single-cell suspension[19]. Moreover, most prospective synchronization regimes for studying RT verify the quality of synchronization by FACS analysis of DNA content; as DNA content defines S-phase interval, selection of cells for DNA content is the most direct means to the desired end. The resolution of S-phase intervals is determined by the fineness of DNA content windows selected. The only situations in which prospective synchronization alternatives may need to be considered are for cells that are very difficult to dissociate or those that are severely aneuploid, such that DNA content does not reflect the time during S phase.

In the original method[20,21], cells were labeled with BrdU for a fraction of S-phase and sorted into several different time points during S-phase. BrdU-substituted DNA could then be isolated either on the basis of its increased density or by using BrdU-specific antibodies, and specific loci could be examined by hybridization or PCR[20–22]. With microarray analysis, replication of the entire genome can be queried in a single-array hybridization by limiting the analysis to two differentially labeled samples, allowing all probes to be assigned one internally normalized relative RT value and rapid comparison of many samples[3,5,6,23,24]. One limitation of assigning one RT value per map position is that it cannot distinguish cases in which homologous loci replicate asynchronously, a situation that is estimated to occur in a small percentage of the genome[19]. However, the protocol can be adapted for analysis of these genomic segments by dividing and sorting S-phase into finer fractions[19].

The two most popular variations of retroactive synchronization by FACS are described in PROCEDURE section below. In the first method, BrdU-labeled cells are divided into early and late S-phase fractions, and BrdU-labeled DNA synthesized either early or late can then be labeled and hybridized onto a microarray. This method produces a high signal-to-noise ratio, as immunoprecipitation (BrdU-IP) substantially enriches DNA synthesized in each half of S-phase. However, BrdU-IP efficacy can fluctuate and must

be closely monitored. In the second method, unlabeled cells are sorted into total S-phase versus G1-phase populations and DNA from these stages is differentially labeled and used as the target. This obviates BrdU-IP, but the dynamic range is limited to the twofold copy number increase during S-phase. Both methods yield similar results, evidenced by a direct comparison in the same cell line in one study[6]. In both methods, DNA from each fraction is differentially labeled with Cy3 and Cy5 dyes and then co-hybridized to a whole-genome oligonucleotide microarray. The ratio of the abundance of each probe in each fraction is then used to generate a RT profile.

Overview of the procedure: normalization and computational analysis of RT data sets (Steps 62–88)

In this section of the protocol, we focus on methods specifically useful for RT analysis using whole-genome comparative genomic hybridization (CGH) microarrays[25], which we have used to investigate the type, degree and mechanism of RT changes in mouse and human cell lines[3,5,6,23,24]. General methods for normalizing and analyzing microarray experiments for chromatin modifications or transcription at gene promoters have been described in detail in other works[26–29]. Similar to two-color microarray designs comparing an experimental sample with a reference, our RT experiments use a two-channel design comparing early versus late fraction enrichment for each target. Typically, we include two dye-swap replicates per sample to address bias due to dye-specific effects, such as more rapid photobleaching of Cy5 dye than Cy3. Our philosophy is to minimize the number of transformations applied to the data and apply only minimally invasive global methods for removing bias and scaling data sets to allow comparisons between them.

All the analysis described here uses the R framework for statistical computing[30–32]. Through user-submitted packages that facilitate a wide variety of methods, R has become an indispensible tool for many common computational tasks. Although R has an initially steep learning curve due to its command line interface, help is available in many locations and forms, including books[33–35], online manuals (http://cran.r-project.org/) and mailing lists aggregated in the R mailing lists archive (http://tolstoy.newcastle.edu.au/R/). Help can also be found within R itself; the command str() is often helpful for viewing the structure of variables and data sets and the ? operator (e.g., ?data.frame()) provides a help page for the corresponding function. We use the R package LIMMA (linear models for microarray data), also available with a user interface through the limmaGUI package, for normalization and scaling[27,36]. The steps for this process are straightforward and are illustrated using two biological replicate data sets of mouse L1210 lymphoblast cells; these data sets are available in raw form in **Supplementary Data 1–4**, and after normalization and smoothing at http://www.ReplicationDomain.org/.

We provide this section as a verified route for extracting information from the microarray experiments described in the PROCEDURE; however, users with sufficient experience with R or having different requirements for their data are free to modify the analysis as needed, and a wide array of alternative and additional methods are available through Bioconductor[31]. Although our methods for downstream analyses were tested primarily with NimbleGen CGH microarrays, most are applicable to any data format containing chromosome, genomic position and RT information for each probe.

Experimental design

BrdU incorporation. The nucleotide analog BrdU can be used to pulse-label newly synthesized DNA during the S-phase. For mammalian cell types that have 8- to 12-h S-phases, incubation with BrdU for 2 h has been empirically determined to provide sufficient incorporation to ensure successful BrdU-IP in subsequent steps, yet the incubation time is long enough to identify even subtle differences in RT, such as between female cells with one versus two early replicating X chromosomes[5]. Success has also been achieved with BrdU labeling times as short as 1 h, but subsequent BrdU-IP can be problematic, as there is very little substituted DNA relative to the background of unsubstituted DNA that will contribute to noise in the BrdU-IP[6]. The BrdU-labeling times for cells with S-phase lengths substantially different from mammalian cells, such as amphibian[20] or fly[8] cells, should be adjusted appropriately.

FACS sorting fractions of S-phase. For first-time users, we recommend that at least 5×10^6 cells be used; however, with experience and a sufficient fraction of S-phase cells, fewer than 0.5×10^6 starting cells can be successfully profiled. The important parameter is to obtain 20,000–30,000 cells in each of the early and late S-phase fractions. As described in PROCEDURE Step 1A, ethanol-fixed cells can be stained with propidium iodide (PI) and sorted on the basis of DNA content. Alternative fluorochromes that do not require RNase digestion, such as chromomycin A3, can also be used with ethanol-fixed cells[20,21]. Some cell types tend to clump or produce a lot of cellular debris when fixed in ethanol. For these cell types, the fixation step can be skipped and DNA can be stained with DAPI (4,6-diamidino-2-phenylindole) in permeabilized cells, as described in PROCEDURE Step 1B. The advantage of the method described in Step 1A is that cells fixed in ethanol can be stored at −20 °C (empirically determined to be the optimal temperature) or shipped to collaborators. The cells should be placed on dry ice during shipping, with a partition between the tube and the dry ice to prevent cell freezing. All steps, particularly storage, should be done in the dark, as BrdU-substituted DNA is light sensitive.

During FACS analysis, forward and side scatter analyses should be used to select an appropriate population of cells free of doublets or cell debris, both of which can hinder accurate sorting of desired populations. Lasers used in this protocol include 488-nm blue to detect PI or 407-nm violet to detect DAPI in cells that have been stained for DNA content. Two separate fractions of S-phase, early and late, are typically chosen for collection, but more can be collected if desired[20,21].

Immunoprecipitation of BrdU-labeled DNA. DNA from BrdU-labeled cells should be sonicated into fragments ranging from 250 bp to 2 kb and then immunoprecipitated using a BrdU-specific antibody. Sonication into fragments of this size helps eliminate IP of DNA that has not been labeled with BrdU. If samples have been stored at −20 °C before beginning the IP, thaw them in a 56 °C water bath to completely dissolve SDS, and then add 200 µl of SDS-PK buffer, prewarmed to 56 °C with 0.05 mg ml^{-1} glycogen, to each sample before performing the phenol-chloroform extraction in PROCEDURE Step 13.

Quality control check of S-phase DNA. Because of the sensitivity and large number of steps involved, BrdU-IP is one of the trickiest parts of the protocol. To ensure quality, screen BrdU-IPs by PCR

TABLE 1 | Primers used for human and mouse BrdU IP screening.

Name	Sequence (5′–3′)	Base pairs (bp)
Human test regions		
Mitochondrial DNA	Forward, CCTAGGAATCACCTCCCATTCC	168
	Reverse, GTGTTTAAGGGGTTGGCTAGGG	
HBA1	Forward, GACCCTCTTCTCTGCACAGCTC	257
	Reverse, GCTACCGAGGCTCCAGCTTAAC	
HBB	Forward, CCTGAGGAGAAGTCTGCCGTTA	241
	Reverse, GAACCTCTGGGTCCAAGGGTAG	
MMP15	Forward, CAGGCCTCTGGTCTCTGTCATT	249
	Reverse, AGAGCTGAGAAACCACCACCAG	
BMP1	Forward, GATGAAGCCTCGACCCCTAGAT	177
	Reverse, ACCCGTCAGAGACGAACTTGAG	
PTGS2	Forward, GTTCTAGGCTGGTGTCCCATTG	230
	Reverse, CTTTCTGTACTGCGGGTGGAAC	
NETO1	Forward, GGAGGTGGAATGCTAGGGACTT	286
	Reverse, GCTGAGTGTGGCCTTAAGAGGA	
SLITRK6	Forward, GGAGAACATGCCTCCACAGTCT	281
	Reverse, GTCCTGGAAGTTGAGTGGATGG	
ZFP42	Forward, CTTGTGGGGACACCCAGATAAG	233
	Reverse, AACCACCTCCAGGCAGTAGTGA	
DPPA2	Forward, AGGTGGACAGCGAAGACAGAAC	168
	Reverse, GGCCATCAGCAGTGTCCTAAAC	
Mouse test regions		
Mitochondrial DNA	Forward, GACATCTGGTTCTTACTTCA	346
	Reverse, GTTTTTGGGGTTTGGCATTA	
Hba-a1	Forward, AAGGGGAGCAGAGGCATCA	439
	Reverse, AGGGCTTGGGAGGGACTG	
Hbb-b1	Forward, CAGTAAGCCACAGATCCTATTG	369
	Reverse, CCCATAGTGACTATTGACTGTG	
Pou5f1	Forward, CCCTCCCTAAGTGCCAGTTTCT	194
	Reverse, GTAATCGCCCTCAGCAGTGTCT	
Mmp15	Forward, AACAGAAGGCCTGCCTTGAC	360
	Reverse, TGCATAGCACGACAGCATTG	
Zfp42	Forward, TGAGATTAGCCCCGAGACTGAG	211
	Reverse, CGTCCCCTTTGTCATGTACTCC	
Dppa2	Forward, CCACAGGAAGACAGGAAGCAGT	199
	Reverse, AGCCAGACAGGAGCCCTAGAGT	
Ptn	Forward, CTGGAATGAGTTACTGACGGGG	230
	Reverse, CTGGCCCCACTGTGTAATAAGC	
Mash1	Forward, GAAGATGAGCAAGGTGGAGACG	182
	Reverse, AGTAGGACGAGACCGGAGAACC	
Akt3	Forward, GAAGTGTGGGTTGAACCTCTGG	173 bp
	Reverse, GCACCCTCTCCACTGTTCTGAT	

amplification, using primers specific to DNA markers of known relative replication time (i.e., early or late phase). Although real-time PCR can be performed, we find gel electrophoresis to be sufficient to evaluate enrichment of DNA in each IP sample. Importantly, as PCR results can vary between aliquots of the same sample and RT can vary between cell types[3,5], consistency across multiple samples from the same cell type is the best way to verify quality. Use the primer sets listed in **Table 1** for mouse or human cell types, or substitute suitable alternatives to screen several IPs from both early and late S-phase fractions.

Amplification methods for immunoprecipitated single-stranded DNA. Once purified by IP and screened for sample quality, BrdU-incorporated DNA must be amplified to obtain sufficient amounts for array hybridization. If multiple samples pass PCR screening, pool DNA from parallel IPs to use as the starting material for whole-genome amplification (WGA); otherwise, use a single-screen IP. Perform WGA as desired (we use the GenomePlex Complete Whole Genome Amplification and Reamplification Kits from Sigma), load amplified samples onto a gel to determine size range and screen once more by PCR to ensure that no bias was introduced during amplification.

Labeling and hybridization of amplified samples. The specific steps required in this section will largely depend on the chosen array platform. Although we focus on NimbleGen products to avoid the ambiguity inherent to generalized methods, the products can be applied successfully to other platforms[8,9], including deep sequencing of BrdU-IP DNA[37]. Currently, mammalian RT data generated from microarray hybridization and deep sequencing are of equal quality[3,6], whereas the microarray method remains more cost effective and the bioinformatics are considerably less demanding for the typical laboratory. Future advances reducing BrdU-labeling times

and sequencing limitations may make this method more cost effective and accessible[38]. Once a platform is chosen, the labeling and hybridization steps are fairly straightforward. Briefly, 1 µg of early or late replicating DNA may be labeled with either Cy5 or Cy3 random 9-mer dyes by Klenow reaction, precipitated with isopropanol and resuspended and quantified in nuclease-free water. Finally, equal quantities of labeled early- and late-fraction DNA should be combined (specific quantity will depend on array design).

Array design. Array design is also an important consideration, and the nature of your study should be a guide in selecting between the variety of available standard and custom designs. For our genome-wide studies in both mouse and human cell lines, 385K- and 3 × 720K-feature CGH tiling arrays have sufficient probe densities, showing no disadvantage compared with high-density 2.1M CGH tiling arrays[5,6], but they have considerable cost and convenience advantages. Tiling designs with roughly evenly spaced probes also facilitate the interpretation and analysis of genetic features.

Array scanning. Carry out scanning according to the manufacturer's recommendations, avoiding unnecessary laser exposure. Take care to align channels with respect to signal intensity frequencies, although minor differences between channels usually do not impact smoothed timing profiles after normalization.

Quality control of microarray data. Before analysis, the overall quality of a microarray experiment should be examined from several angles. In general, there are six qualities that are important for reliable results of RT analyses on CGH arrays that should be verified at the corresponding PROCEDURE steps:

(1) Comparable signal intensity distributions for red and green channels (Step 74).
(2) Unbiased signal ratios with respect to signal intensity (Step 75).
(3) Comparable timing value distributions between experiments (Step 76).
(4) A high overall signal-to-noise ratio of the experiment (Step 84).
(5) Lack of artifacts in raw and false-color microarray images (Step 85).
(6) High correlations between replicate experiments (Step 86B).

Downstream analysis. When comparing the timing program with other genetic and epigenetic properties, you should note that differences in formats between chromatin immunoprecipitation (ChIP)-on-chip, ChIP-seq and other approaches will require some care in processing, and even data sets from similar platforms often have idiosyncrasies that must be accounted for. In particular, take care to ensure that RT and other data types are compared in compatible genomic builds and equivalent cell types; use a method of quantification consistent with the methods and goals of the studies involved.

MATERIALS

REAGENTS

▲ **CRITICAL** All solutions are prepared with ddH_2O and stored at room temperature (22 °C) unless otherwise indicated.

- *Cells of interest*: Cultures can be grown in a cell culture dish of any size, but must be in an actively dividing state for use in this protocol. A minimum of 50,000 S-phase cells is required for the protocol. However, we recommend that cultures with at least 120,000 S-phase cells be used.
- BrdU (5-bromo-2′-deoxyuridine; Sigma Aldrich, cat. no. B5002). Prepare stock solutions of 10 mg ml^{-1} and 1 mg ml^{-1} in ddH_2O and store at −20 °C.
- Cell culture medium appropriate for cell type
- Trypsin-EDTA (1×; Mediatech, cat. no. 25-053-Cl)
- Accutase (Innovative Cell Technologies, cat. no. AT104). For long-term storage, store at −20 °C. Thaw overnight at 4 °C before use. Once thawed, store at 4 °C for up to 2 months. Warm to room temperature before each use.
- PBS (see REAGENT SETUP)
- FBS (GIBCO, cat. no. 16000). Prepare 1% (vol/vol) in PBS and store at 4 °C.
- DAPI (BioChemika, cat. no. 32670). Dissolve stock in ddH_2O to a final concentration of 10 mg ml^{-1}. Store at −20 °C protected from light.
- PI (Sigma, cat. no. P4179-100MG; see REAGENT SETUP)
- RNase A (10 mg ml^{-1}; Sigma, cat. no. R6513; store at −20 °C)
- Proteinase K (20 mg ml^{-1}; Amresco, cat. no. E195; store at −20 °C)
- Glycogen (20 mg ml^{-1}; Fermentas, cat. no. RO561; store at −20 °C)
- Isopropanol (Sigma, cat. no. 59304)
- Ethanol (100%; Sigma, cat. no. E7023)
- Ethanol (70% (vol/vol); Sigma, cat. no. E7023)
- Tris base (Fisher Scientific, cat. no. BP152-5)
- HCl (EMD, cat. no. HX0603P-5)
- NaCl (Fisher Scientific, cat. no. BP358-1)
- KCl (Sigma, cat. no. P3911-500G)
- Na_2HPO_4 (Sigma, cat no. S3264-500G)
- KH_2PO_4 (Fisher Scientific, cat no. P380-500)
- SDS (Invitrogen, cat. no. 15525-017)
- EDTA (Invitrogen, cat. no. 15576-028)
- SDS-PK buffer (see REAGENT SETUP)
- Tris-saturated Phenol (Fisher, cat. no. BP226-500; store at −20 °C)
  ! **CAUTION** It is caustic and harmful if inhaled or ingested. Wear gloves and other appropriate protective equipment. Use adequate ventilation. Store at −20 °C.
- Chloroform (Sigma, cat. no. 34854) **!** **CAUTION** It is a probable carcinogen and is harmful if inhaled or ingested. Wear gloves and other appropriate protective equipment. Use adequate ventilation.
- BrdU-specific antibody (BD Biosciences Pharmingen, cat. no. 555627). Store at 4 °C.
- Ammonium acetate (Fisher Scientific, cat. no. A639-500; see REAGENT SETUP) **!** **CAUTION** It is an irritant and is harmful if swallowed. Wear gloves and other appropriate protective equipment. Use adequate ventilation. Store at room temperature.
- Rabbit anti-mouse IgG (Sigma, cat. no. M-7023). Store at 4 °C
- Anti-Mouse IgG-AlexaFluor488 (Invitrogen/Molecular Probes, cat. no. A-11029). Store at 4 °C.
- Taq DNA Polymerase with ThermoPol Buffer (New England BioLabs, cat. no. M0267)
- dNTPs (10 µM; Bioline, cat. no. BIO-39025)
- Ethidium bromide (10 mg ml^{-1}; Fisher, cat. no. BP102-5)
- Agarose (OmniPur, cat. no. 2125)
- PCR primers for BrdU-IP quality verification (Steps 45–49), see **Table 1**
- HCl/0.5% (0.1 M (vol/vol)) Triton X-100 (Sigma, cat. no. T9284) in ddH_2O. Store at room temperature
- Sodium tetraborate ($Na_2B_4O_7 \cdot 10H_2O$; 0.1 M, Sigma, cat. no. S-9640, pH 8.5 in ddH_2O)
- Tween-20 (0.5% (vol/vol); Sigma, cat. no. P-1379) / 1% (wt/vol) BSA (Fisher Scientific, cat. no. BP1600-1) in PBS
- Triton X-100 (0.1% (vol/vol); Sigma, cat. no. T9284) in PBS
- Triton X-100 (0.5% (vol/vol); Sigma, cat. no. T9284) in PBS

- BSA (Fisher Scientific, cat. no. BP1600-1)
- GenomePlex Complete Whole Genome Amplification Kit (Sigma, cat. no. WGA2)
- GenomePlex WGA Reamplification Kit (Sigma, cat. no. WGA3)
- QIAquick PCR Purification Kit (QIAGEN, cat. no. 28106)
- NimbleGen Dual-Color DNA Labeling Kit (NimbleGen, cat. no. 05223547001)
- NimbleGen Hybridization Kit (NimbleGen, cat. no. 05583683001)
- NimbleGen Wash Buffer Kit (NimbleGen, cat. no. 05584507001)

EQUIPMENT

- Nylon mesh (37 µm; Small Parts, cat. no. CMN-0040-D)
- Filter syringe (BD Biosciences, cat. no. H8293-005663)
- Round-bottom polystyrene tube (5 ml; Falcon, cat. no. 352054)
- Round-bottom tube (15 ml; Falcon, cat. no. 2059)
- FACSAria cell sorter (BD Biosciences, or a comparable sorter)
- Hemocytometer
- Vortexer
- Sonicator (Heat Systems-Ultrasonics W-380, Heat Systems Ultrasonics)
- Thermocycler
- Spectrophotometer (NanoDrop, Thermo Scientific)
- Electrophoresis apparatus
- Appropriate NimbleGen Arrays and Mixers
- Appropriate NimbleGen Hybridization System
- Appropriate NimbleGen microarray scanner
- Parafilm

REAGENT SETUP

PBS (1×) To prepare 1 liter, dissolve 8 g NaCl, 0.2 g KCl, 1.44 g Na_2HPO_4 and 0.24 g KH_2PO_4 in 800 ml of ddH_20. Adjust pH to 7.4 with HCl and adjust the final volume to 1 liter. Sterilize by autoclaving. Store at room temperature.

Trypsin-EDTA (0.2×) To prepare 50 ml, combine 10 ml of 1× Trypsin-EDTA with 40 ml of 1× PBS. Store at 4 °C for up to 1 month. Warm to room temperature before each use.

Propidium iodide (1 mg ml^{-1}) To prepare 20 ml, dissolve 20 mg PI powder in autoclaved ddH_2O to obtain a final volume of 20 ml and filter by syringe. Store protected from light for up to 1 year at 4 °C.

DAPI staining solution To prepare ~1 ml, add 10 µl of 10% (vol/vol) Triton X-100 and 2 µl of 1 mg ml^{-1} DAPI to 1 ml of PBS. Prepare the solution fresh before each use.

BrdU-specific antibody (12.5 µg ml^{-1}) Dilute antibody in 1× PBS from the stock concentration of 0.5 mg ml^{-1} to a final concentration of 12.5 µg ml^{-1}. Freshly prepare 40 µl of diluted antibody for each sample and discard unused diluted antibody.

Ammonium acetate (10 M) To prepare 100 ml, dissolve 77.08 g ammonium acetate in 50 ml of ddH_2O. Add ddH_2O to adjust the final volume to 100 ml. Syringe-filter and store at room temperature.

Tris-HCl (1 M; pH 8.0) To prepare 500 ml, dissolve 60.57 g Tris base in 400 ml of ddH_2O. Add HCl to adjust pH to 8.0. Add additional ddH_2O to adjust the final volume to 500 ml. Sterilize by autoclaving. Store at room temperature.

EDTA (0.5 M) To prepare 1 liter, dissolve 186.1 g disodium EDTA in 800 ml of ddH_2O. Stir vigorously on a magnetic stirrer. Adjust the pH to 8.0 by addition of NaOH. Add ddH_2O to a final volume of 1 liter. Sterilize by autoclaving. Store at room temperature.

TE buffer (1×) To prepare 1 liter, add 10 ml of 1 M Tris-HCl (pH 8.0) to 2 ml of 0.5 M EDTA (pH 8.0) and adjust the final volume to 1 liter with autoclaved ddH_2O. Store at room temperature.

Phenol-chloroform solution To prepare 50 ml, combine 25 ml of Tris-saturated phenol with 25 ml of chloroform. Allow separation of layers before use. We recommend that the solution be stored overnight before use to allow adequate separation or centrifuged at maximum speed for 10 min before use to achieve separation. Store at 4 °C protected from light.

SDS-PK buffer To prepare 50 ml, combine 34 ml autoclaved ddH_2O, 2.5 ml of 1 M Tris-HCl (pH 8.0), 1 ml of 0.5 M EDTA, 10 ml of 5 M NaCl and 2.5 ml of 10% (wt/vol) SDS. Store at room temperature. Warm to 56 °C before use to completely dissolve SDS.

IP buffer (10×) To prepare 50 ml, combine 28.5 ml of ddH_2O, 5 ml of 1 M sodium phosphate (pH 7.0), 14 ml of 5 M NaCl and 2.5 ml of 10% (wt/vol) Triton X-100. Store at room temperature.

IP buffer (1×) To prepare 50 ml, add 5 ml of 10× IP buffer to 45 ml of autoclaved ddH_2O. Store at room temperature.

Digestion buffer To prepare 50 ml, combine 44 ml of autoclaved ddH_2O, 2.5 ml of 1 M Tris-HCl (pH 8.0), 1 ml of 0.5 M EDTA and 2.5 ml of 10% (wt/vol) SDS. Store at room temperature.

EQUIPMENT SETUP

Sonicator Adjust sonicator settings as needed to achieve a 250 bp to 2 kb distribution of DNA fragment sizes. We use a water bath-type sonicator (Heat Systems-Ultrasonics W-380) with a 2-s, 50% duty cycle and an output setting of 7 for 4 min.

PROCEDURE

BrdU labeling and staining of cells for FACS

1| To perform PI staining and ethanol fixation before sorting, follow option A; this method is most commonly used, as it allows for shipping or long-term storage, and it has worked well for most mouse cell lines[5,6]. For cells that break or clump in ethanol, follow option B; note that a drawback of option B is that cells need to be sorted immediately following BrdU labeling. Alternatively, carry out the procedure for S/G1 sorting described in **Box 1** instead of Steps 1–57 (see also **Fig. 1**). This method obviates the need for BrdU-IP and WGA and can alleviate concerns that sorting early and late fractions of S-phase or WGA introduce a temporal bias; however, in our experiments, E/L (early/late) fractionation has produced results equivalent to the S/G1 method as well as sorting of additional S-phase fractions[3,37].

(A) Labeling and PI staining of cells for FACS after ethanol fixation ● TIMING 3.5 h

(i) Add BrdU to cells in culture medium at a final concentration of 50 µM.

(ii) Incubate cells for 2 h in a CO_2 incubator at 37 °C with 5% CO_2.

(iii) For adherent cells, rinse gently with ice-cold PBS twice. For suspension cells, collect cells in a 15-ml tube and proceed directly to Step 1A(vi).

(iv) Detach adherent cells using 0.2× Trypsin-EDTA for 2–3 min or Accutase for 3–6 min.
▲ **CRITICAL STEP** Incubate cells at 37 °C with the enzyme treatment and/or use gentle trituration if necessary to achieve a single-cell suspension, as this is essential for accurate FACS sorting.

(v) Add 5 ml of cell culture medium (containing FBS if trypsin has been used) to the cell culture dish or flask, pipette gently and transfer contents to a 15-ml round-bottom tube.

BOX 1 | METHOD FOR SORTING ACCORDING TO S/G1-PHASE ● TIMING 1 D

In this method, cells are sorted into two fractions, G1- and S-phase, based on DNA content, and RT is derived from the twofold copy number increase for early versus late replicating sequences in pure S-phase populations. DNA analysis using flow cytometry can be performed simply by the use of a single DNA-binding fluorescent dye, such as PI or DAPI, as originally described[59]. Although this method is adequate, simultaneous measurement of BrdU-labeled DNA by performing BrdU/PI double staining for cell cycle analysis can discriminate G1- and early S-phase cells much more efficiently than by PI-only staining. In addition, some cell types, particularly those derived from differentiated stem cells or primary tissues, can produce debris that interferes with good S-phase sorting, and a short BrdU label described here can eliminate debris that is not labeled with BrdU. The advantage of this method is that it eliminates the need for BrdU-IP and WGA steps (described in Steps 13–44 and 51–57), which need to be carefully controlled. However, direct comparisons have shown that this method produces a lower signal-to-noise ratio than the method described in the main PROCEDURE[6]. A much shorter BrdU pulse label is used in this protocol at lower concentration, because we are only trying to identify the cells in S-phase. With longer BrdU-labeling time periods, G2/M cells become labeled. It should be noted that we originally used the standard protocol for BrdU/PI analysis provided by Becton-Dickinson, which is fine for analysis. However, we found that the high concentration of HCl in this method sheared genomic DNA to very small fragments that precluded subsequent steps of the protocol. By titrating HCl, we found that 0.1 M HCl produced the optimal compromise between good S- versus G1-phase separation and minimal DNA shearing. For BrdU/PI double-staining, correction of spectral overlap is critical for successful experiments. Spectral overlap exists between emission spectra of PI and fluorescein isothiocyanate (FITC)/Alexa Fluor 488 (for BrdU). Without correction, the BrdU/PI plots typically look similar to **Figure 1a**. For this correction, the adjustment of the ratio between PI and Alexa Fluor 488 (or FITC) gains can significantly reduce the skewing shown in **Figure 1a**. Subtraction of the FITC signal from the PI signal (i.e., compensation) may also be required. To perform these corrections, a 'BrdU-only' control is required, prepared by staining BrdU-labeled cells without the addition of PI. A 'PI-only' control also helps, prepared by staining non-BrdU-labeled cells for BrdU and PI. (*Note*: BrdU-labeled specimen stained for PI only does not reflect background signals derived from the anti-BrdU antibody and thus is not as good as unlabeled cells.) This step can be time-consuming, but is critical for successful sorting. We suggest that you first adjust the gains of forward scatter and side scatter, and then adjust the PI and AlexaFluor488 gains by trial and error to obtain the best possible BrdU/PI plot. You may be able to obtain a reasonable BrdU/PI plot without compensation; otherwise, compensate by subtracting FITC signal from PI signal. The lower the percentage subtracted, the better.

S/G1 FACS sorting ● TIMING 1 d

1. For adherent cells, remove cell culture medium from exponentially growing cells and replace with cell culture medium containing BrdU at a final concentration of 10 μM. For suspension cells, add BrdU to the cell culture medium at a final concentration of 10 μM. In order to obviate the amplification step before labeling and array hybridization, start with 6 million cells. One should also prepare a small sample of non-BrdU-labeled, ethanol-fixed cells for PI-only control and set aside a small number of BrdU-labeled cells for BrdU-only control.
2. Incubate cells for 15 min in a CO_2 incubator at 37 °C and 5% CO_2.
3. Fix as described in Steps 1A(iii–x) of the main PROCEDURE.

■ **PAUSE POINT** Cells can be stored as in Step 1A.

4. Aliquot (multiples of) 3 × 10^6 cells in 1.5-ml tube(s), centrifuge for 5 min at 200*g* at room temperature. Removal of supernatant is much easier with 1.5-ml tubes as the pellets are very loose.
5. Aspirate the supernatant completely with a P200 pipette. Here and elsewhere, an additional pulse spin (~3 s) will help with discarding residual supernatant.
6. Loosen the pellet by brief vortexing.
7. Add 1 ml of 0.1 M HCl/0.5% (vol/vol) Triton X-100; resuspend by tapping.
8. Incubate for 15 min at room temperature in the dark.
9. Centrifuge for 5 min at 200*g* at room temperature, then aspirate the supernatant completely.
10. Add 1ml of 0.1 M sodium tetraborate and resuspend by tapping.
11. Centrifuge for 5 min at 200*g* at room temperature, then aspirate the supernatant completely.
12. Add 0.15-μg anti-BrdU antibody in 0.5 ml of 0.5% (vol/vol) Tween-20/1% (vol/vol) BSA/PBS and resuspend by tapping.
13. Incubate for 30 min at room temperature in the dark.
14. Centrifuge for 5 min at 200*g* at room temperature, then aspirate the supernatant completely.
15. Add 0.5 ml of 0.5% (vol/vol) Tween-20/1% (vol/vol) BSA/PBS.
16. Centrifuge for 5 min at 200*g* at room temperature, then aspirate the supernatant completely.
17. Add 1 μg of anti-mouse IgG-Alexa Fluor 488 in 100 μl of 0.5% (vol/vol) Tween-20/1% (vol/vol) BSA/PBS and resuspend by tapping (or when 1–2 × 10^6 cells are used, add 0.5 μg of mouse-specific IgG–Alexa Fluor 488 in 50 μl).
18. Incubate for 30 min at room temperature in the dark.
19. Centrifuge for 5 min at 200*g* at room temperature, then aspirate the supernatant completely.
20. Add 0.5 ml of 0.5% (vol/vol) Tween-20/1% (vol/vol) BSA/PBS.
21. Centrifuge for 5 min at 200*g* at room temperature, then aspirate supernatant completely.
22. Resuspend the pellet in 1 ml of 5 μg ml^{-1} PI in PBS (for 'BrdU-only' control, just add PBS).
23. Transfer to a round-bottom 5-ml tube (i.e., Falcon 2054).
24. Adjust the concentration to 2 × 10^6 ml^{-1} by adding 5 μg ml^{-1} PI in PBS. For BrdU-only sample, adjust to the same concentration by adding PBS without PI.
25. Filter with a 37-μm mesh filter.
26. Bring to flow lab for sorting (on ice, in the dark). Resume the main PROCEDURE at Step 58.

(vi) Count the number of cells collected using a hemocytometer. Collect enough cells to obtain at least 20,000–30,000 (preferably >150,000) cells in each fraction after sorting (Step 2); this will generally require 0.5–1 × 10^6 cells, with more cells required if few cells are in S-phase. For first-time users, we recommend starting with 4 × 10^6 to 8 × 10^6 cells.
(vii) Centrifuge at ~200*g* for 5 min at room temperature.
(viii) Aspirate the supernatant carefully and resuspend the cells in 2.5 ml of ice-cold PBS containing 1% (vol/vol) FBS.
(ix) Add 7.5 ml of ice-cold 100% ethanol dropwise while gently vortexing.
▲ **CRITICAL STEP** Note that vortexing should be performed gently to avoid cell damage.
(x) Seal the cap of the 15-ml tube with Parafilm and mix gently but thoroughly.
■ **PAUSE POINT** If necessary, cells can be stored in the dark at −20 °C indefinitely.

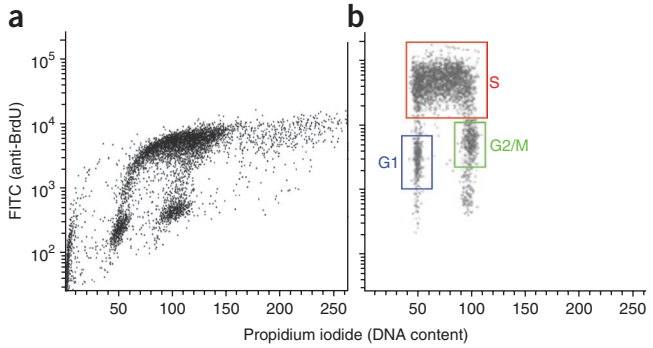

Figure 1 | Two-dimensional cell-cycle sorting for S- and G1-phases. Cells labeled with BrdU and stained as described in **Box 1**, and then analyzed on a FACS instrument. (**a**) A typical noncorrected BrdU/PI plot. Note how the plot is skewed to the right because of spectral overlap. (**b**) A corrected BrdU/PI plot. Sorting windows for nicely separated G1, S and G2/M phases of the cell cycle are indicated.

(xi) Resuspend the BrdU-labeled, ethanol-fixed cells by tapping and inverting the tube.
(xii) Transfer 4 × 10^6 to 8 × 10^6 cells to a 5-ml polystyrene round-bottom tube.
(xiii) Centrifuge at ~200*g* for 5 min at room temperature.
(xiv) Decant supernatant carefully.
(xv) Resuspend the cell pellet in 2 ml of PBS with 1% (vol/vol) FBS. Mix well by tapping the tube.
(xvi) Centrifuge at ~200*g* for 5 min at room temperature.
(xvii) Decant supernatant carefully.
(xviii) Resuspend the cell pellet in PBS with 1% (vol/vol) FBS to achieve a solution of 3 × 10^6 cells per ml.
(xix) Add 1 mg ml^{-1} of PI to a final concentration of 50 µg ml^{-1}.
(xx) Add 10 mg ml^{-1} of RNase A to a final concentration of 250 µg ml^{-1}.
(xxi) Tap the tube to mix and incubate for 20–30 min at room temperature (22 °C) in the dark.
(xxii) Filter cells by pipetting them through a 37-µm nylon mesh into a 5-ml polystyrene round-bottom tube.
(xxiii) Place samples on ice in the dark and proceed directly to FACS sorting.

(B) BrdU labeling and DAPI staining of cells for FACS ● TIMING 3 h
(i) Follow Steps 1A(i–vii).
(ii) Aspirate the supernatant carefully.
(iii) Add 5 ml of ice-cold PBS and pipette gently but thoroughly.
(iv) Centrifuge at ~200*g* for 5 min at room temperature.
(v) Aspirate the supernatant carefully.
(vi) Resuspend the cell pellet in DAPI staining solution to achieve a solution of 5 × 10^6 to 10 × 10^6 cells per ml.
(vii) Filter the cells by pipetting them through a 37-µm nylon mesh into a 5-ml polystyrene round-bottom tube.
(viii) Place the samples on ice in the dark and proceed directly to FACS sorting.

2| Run the sample on FACSAria cell sorter (alternatively, any comparable cell sorter can be used).
▲ **CRITICAL STEP** It is very important to place live samples chilled on ice or at 4 °C during FACS analysis to avoid cell-cycle progression in the absence of BrdU. Protect samples from light.

3| Use forward and side scatter information to select the desired population of cells to be included in the sort, and exclude doublets or cell debris.

4| Create a histogram that plots cell count on the *y* axis and DNA content (fluorochrome intensity) on the *x* axis (see **Fig. 2**).

5| Select two distinct S-phase populations to be sorted into separate fractions, as indicated in **Figure 2**.

6| Sort cells into fresh 5-ml round-bottom tubes and place at 4 °C during the sort.
■ **PAUSE POINT** Store cells immediately on ice in the dark until all samples have been sorted.
? TROUBLESHOOTING

7| Centrifuge at 400*g* for 10 min at 4 °C. Alternatively, if fewer than 150,000 cells have been collected for each fraction, proceed directly to Step 9.

8| Decant supernatant gently, only once.
▲ **CRITICAL STEP** Some residual sheath fluid can be left in the tube to prevent losing the cell pellet, which can easily detach from the tube during this step.

9| Add 1 ml of SDS-PK buffer containing 0.2 mg ml^{-1} of proteinase K and 0.05 mg ml^{-1} of glycogen for every 100,000 cells collected and mix vigorously by tapping the tube.

10| Incubate samples in a 56 °C water bath for 2 h.

11| Mix the sample thoroughly and aliquot 200 µl, equivalent to ~20,000 cells, per 1.5-ml tube.
■ **PAUSE POINT** Samples can be stored for at least 6 months at −20 °C in the dark before use.

12| Add 200 µl of SDS-PK buffer with 0.05 mg ml^{-1} of glycogen to each aliquot and proceed directly to BrdU-IP.

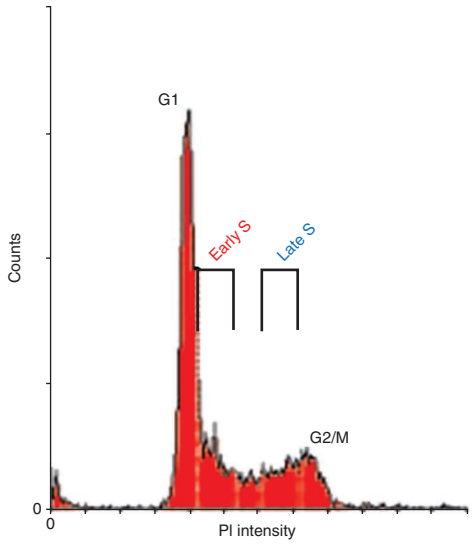

Figure 2 | A typical cell-cycle profile for a mammalian fibroblast population obtained during FACS analysis by plotting cell count versus DNA content. In this example, cellular DNA was stained with PI; accordingly, DNA content is represented by PI intensity. A G1 peak, representing cells with 2*N* DNA content, and a G2/M peak, representing cells that have undergone replication and therefore possess a 4*N* DNA content, are labeled. The area between these two peaks is representative of cells in S phase and can be sorted into two fractions, as indicated here, to obtain early and late S-phase samples.

BrdU immunoprecipitation ● **TIMING 2–3 d**

13| Extract once with phenol-chloroform, collecting the upper phase in a 1.5-ml tube.

14| Extract once with chloroform, collecting the upper phase in a 1.5-ml tube.

15| Add 1 volume of isopropanol and mix well.

16| Store at −20 °C for 20 min.
■ **PAUSE POINT** Samples can be stored in the dark at −20 °C overnight.

17| Centrifuge at 16,000*g* for 30 min at 4 °C.

18| Discard the supernatant and add 750 µl of 70% ethanol to the pellet.

19| Centrifuge at 16,000*g* for 5 min at 4 °C, then remove all ethanol and let the pellet dry.

20| Resuspend the pellet in 500 µl of 1× TE.
■ **PAUSE POINT** If necessary, the pellet can be stored overnight at 4 °C.

21| Sonicate DNA to an average size of approximately 0.7–0.8 kb. Settings required for a 250-bp to 2-kb range should be determined empirically for each sonicator type. See EQUIPMENT SETUP.

22| Incubate the sample at 95 °C for 5 min to heat-denature the DNA.

23| Cool the sample on ice for 2 min.

24| Add 60 µl of 10× IP buffer to a clean 1.5-ml tube.

25| Add the denatured DNA from Step 22 to the tube from Step 24.

26| Add 40 µl of 12.5 µg ml^{-1} anti-BrdU antibody.

27| Incubate for 20 min at room temperature with constant rocking.
▲ **CRITICAL STEP** Cover tubes with foil and keep samples in the dark.

28| Add 20 µg of rabbit anti-mouse IgG.
▲ **CRITICAL STEP** Cover tubes with foil and keep samples in the dark.

29| Incubate for 20 min at room temperature with constant rocking.

30| Centrifuge at 16,000*g* for 5 min at 4 °C.

31| Remove the supernatant completely.
▲ **CRITICAL STEP** If the pellet becomes loose, then briefly centrifuge the sample again in order to completely remove the supernatant without disturbing the pellet. Several centrifugations may be necessary to completely remove the supernatant.

32| Add 750 µl of 1× IP buffer that has been chilled on ice.

33| Centrifuge at 16,000*g* for 5 min at 4 °C.

34| Remove supernatant completely, as in Step 31.

35| Resuspend the pellet in 200 µl of digestion buffer with freshly added 0.25 mg ml^{-1} proteinase K. Incubate the samples overnight at 37 °C.

36| Add 100 µl of fresh digestion buffer with freshly added 0.25 mg ml^{-1} proteinase K.

37| Incubate the samples for 60 min at 56 °C.

38| Extract once with phenol-chloroform, collecting the upper phase in a 1.5-ml tube.

39| Extract once with chloroform, collecting the upper phase in a 1.5-ml tube.

40| Add 0.625 µl of 20 mg ml^{-1} glycogen, 100 µl of 10 M ammonium acetate and 750 µl of 100% ethanol and mix well.

41| Store at −20 °C for 20 min.
■ **PAUSE POINT** Samples can be stored in the dark at −20 °C indefinitely.

42| Centrifuge at 16,000*g* for 30 min at 4 °C.

43| Remove supernatant, rinse pellet with 70% (vol/vol) ethanol and dry.

44| Resuspend the pellet in 80 µl of 1× TE (for a final concentration of 250 cell equivalents per µl).
■ **PAUSE POINT** Store DNA at 4 °C for up to 1 month or at –20°C for longer storage.

PCR method for quality control of BrdU-IP ● **TIMING 4–6 h**

45| Prepare enough PCR master mix to screen all IP samples with each primer set listed in **Table 1**. An example PCR mix is listed in the table below. Mitochondrial primer sets should be used at 1.0 µM concentration instead of 0.5 µM; add 0.63 µl of forward and reverse 20 µM combined primers and adjust with nuclease-free water accordingly.

Component	Amount per reaction (µl)	Final
Taq buffer (10×)	1.25	1×
dNTPs (10 mM)	0.25	0.2 mM
Taq polymerase (20 U $µl^{-1}$)	0.06	1.2 U
F/R 20 µM combined primers	0.31	0.5 µM
Nuclease-free water	Up to 12.5	

46| Aliquot 11.5 µl of master mix per tube and add 1 µl of IP sample.

47| Run the samples in thermocycler with the following conditions:

Cycle number	Denature	Anneal	Extend
1	94 °C, 2 min		
2–39	94 °C, 45 s	60 °C, 45 s	72 °C, 2 min
40			72 °C, 5 min

48| Add 2.5 µl of 6× loading dye to every 12.5-µl reaction and load 6 µl onto 1.5% (wt/vol) agarose gel. Run the gel at 125 V for 16 min.

49| Score each IP based on anticipated enrichment of amplicon DNA (see Experimental design). Multiple samples from the same cell type should amplify consistently, with enrichment consistent with genes of known RT for the given cell type.
▲ **CRITICAL STEP** Before proceeding, verify sample quality with corresponding primer sets listed in **Table 1**.
? TROUBLESHOOTING

50| If several IPs of the same sample and S-phase fraction pass the screening, pool equal amounts of each IP to a final volume of 50 µl (e.g., if two IPs pass, combine 25 µl of each in the pool).

Whole-genome amplification ● TIMING 8 h
51| Precipitate DNA fractions by adding 1.25 µl of 2 mg ml^{-1} glycogen, 20 µl of 10 M ammonium acetate and 150 µl of ethanol to each 50-µl IP sample (if pooling multiple samples, a total volume of 50 µl should still be used). Mix well, let it stand at −20 °C for 20 min and then centrifuge for 30 min at maximum speed at 4 °C.

52| Rinse the pellets with 70% (vol/vol) ethanol, air-dry them, and then resuspend them in 10 µl of nuclease-free water.

53| Transfer the 10-µl samples to 0.2-ml PCR tubes and carry out WGA using an appropriate method or kit. In our experiments, the GenomePlex Complete Whole Genome Amplification Kit has worked well, starting from the library preparation step (i.e., skipping fragmentation)[39].

54| Purify entire WGA products using an appropriate PCR purification kit, such as QIAquick. Elute in 30 µl nuclease-free water prewarmed to 65 °C and determine the concentration using Nanodrop.

55| Dilute WGA samples to appropriate concentration (we use 1 µl DNA of 20 ng $µl^{-1}$) and, if necessary to obtain sufficient material for the chosen array platform, perform a second round of WGA. We follow the GenomePlex WGA Reamplification Kit, Reamplification Procedure A.

56| Purify entire reamplified WGA products as in Step 54.

57| Screen purified final products using the PCR method described in Steps 46–49.
■ **PAUSE POINT** Samples can be stored in the dark at −20 °C for up to 1 month.
? TROUBLESHOOTING

Labeling and hybridizing ● TIMING 1–3 d
58| Differentially label reamplified early and late WGA DNA fractions with Cy3 and Cy5 dyes from Step 57 (or nonamplified DNA prepared as in Box 1) according to the method most appropriate for the chosen array platform. We follow the sample labeling instructions for the NimbleGen Dual-Color DNA Labeling Kit.

59| Hybridize the samples to array(s) using the corresponding method or kit. We use the NimbleGen Hybridization Kit.

60| After hybridization, wash array(s) as needed. We perform this step, according to the manufacturer's instructions, using the NimbleGen Wash Buffer Kit.

61| Scan array(s) with an appropriate microarray scanner and software package. We use the NimbleGen scanner GenePix 4000B and the accompanying NimbleGen arrays user's guide, CGH analysis v5.1. Newer equipment is accompanied with a newer version of the user's guide and operated slightly differently. For NimbleGen arrays, raw images should be saved as .tif files, and two .pair files (one each for Cy3 and Cy5 channels) will be created per experiment.

Normalization of raw data sets ● TIMING 1 d

62| If necessary, install R from http://www.r-project.org/. Create RGL (Red Green List) files from the original NimbleGen .pair files, as described in Steps 63–68. These files contain columns for both red (Cy5) and green (Cy3) channel signal intensities; example pair files used in Step 65 and throughout are available in **Supplementary Data 1–4**.

63| Set the working directory using the command 'setwd' in the R console to specify the appropriate file path. Here and in later steps, the '>' symbol denotes the R prompt at the beginning of a line and should be omitted when typing the command.

```
> setwd("D:\RT project\Raw datasets")
```

64| Read the first five rows of data from the raw data files and determine the data type of each column using the sapply() function:

```
> tab5rows <- read.delim("318990_4L1210LymphoblastP1_532.pair", header = T, nrows =
5, skip=1)
> classes <- sapply(tab5rows, class)
```

▲ **CRITICAL STEP** When reading large tables in R, such as .pair files, explicitly noting the number of rows and data type of each column as illustrated here and in Step 65 will save a substantial amount of memory and calculation time. Occasionally, the sapply() function will set the genomic position columns of large data sets as an integer type, which lacks the memory space to store large numbers. If so, set the type manually with > classes[x] = 'numeric' (where x is the column number containing position information) after creating the classes variable.

65| Read the raw data sets into memory. Note that variable names and file names may be substituted here and elsewhere, as appropriate. The 'nrows' parameter can be a modest overestimate; the correct number of rows will be present in the final table, but an estimate allows the system to allocate the correct amount of memory.

```
> mLymph1Cy3 <- read.delim("L1210LymphoblastR1_532.pair", header=T, nrows=390000,
comment.char = "", colClasses=classes, skip=1)    # Supplementary Data 1

> mLymph1Cy5 <- read.delim("L1210LymphoblastR1_635.pair", header=T, nrows=390000,
comment.char = "", colClasses=classes, skip=1)    # Supplementary Data 2

> mLymph2Cy3 <- read.delim("L1210LymphoblastR2_532.pair", header=T, nrows=390000,
comment.char = "", colClasses=classes, skip=1)    # Supplememtary Data 3

> mLymph2Cy5 <- read.delim("L1210LymphoblastR2_635.pair", header=T, nrows=390000,
comment.char = "", colClasses=classes, skip=1)    # Supplementary Data 4
```

66| Extract the Cy3 and Cy5 channel signal intensities from the raw data sets, for example,

```
> mLymph1 <- data.frame(S_Cy5=mLymph1Cy5[,10],S_Cy3=mLymph1Cy3[,10])
> mLymph2 <- data.frame(S_Cy5=mLymph2Cy5[,10],S_Cy3=mLymph2Cy3[,10])
```

67| Write the columns extracted in Step 66 to separate RGL files for normalization

```
> write.table(mLymph1, file="L1210LymphoblastR1.rgl.txt", row.names=F,
quote=F, sep="\t", eol="\r\n") write.table(mLymph2,
file="L1210LymphoblastR2.rgl.txt", row.names=F, quote=F, sep="\t",
eol="\r\n")
```

68| Create a 'targets' text file that describes the target files for normalization**.** We will name this file 'T.txt' (see **Supplementary Data 5** for an example targets file). Note that, to be read correctly, the file should be tab delimited and should contain only one carriage return at the end of the final line. Place this file in the same directory as the raw .pair files and .rgl files generated above.

69| Install a current version of the LIMMA package according to the instructions at http://bioinf.wehi.edu.au/limma/ or by using the command line interface:

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("limma")
> biocLite("statmod")
```

70| Perform LOESS and scale normalization using LIMMA as described in Steps 71–73 and verify the results as described in Steps 74–85. This process is more straightforward than many two-color normalization methods, as NimbleGen arrays do not have print tip groups, spot background areas or mismatch spots that must be accounted for. LOESS normalization (normalize within arrays) corrects the internal dependence of red-green ratios on their intensity independently for each array and is examined further in Steps 74 and 75. Scale normalization (normalize between arrays) equalizes the distribution of timing values between multiple samples for comparisons and can be verified in Step 76.

71| Load the LIMMA package and read the raw data sets listed in the file created in Step 68. This will generate a MAlist-type data object, r, which stores the ratios (M-values) and average intensity values (A-values) of raw samples before normalization:

```
> library(limma)
> t = readTargets("T.txt", row.names="Name")
> r = read.maimages(t, source="generic",columns=list(R="S_Cy5", G="S_Cy3"))
```

72| Perform LOESS normalization. This will generate a second MAlist-type data object, MA.l, which stores the samples after within-array normalization.

```
> MA.l = normalizeWithinArrays(r, method="loess")
```

73| Perform scale normalization. This will generate a third MAList object, MA.q, which stores the samples after between-array normalization. As with ChIP-chip methods[13,14], this type of scale normalization may not be appropriate for examining subsets of the genome in which large unbalanced timing changes are expected (e.g., timing of the X chromosome before and after inactivation), but is ideal for whole-genome analyses.

```
> MA.q = normalizeBetweenArrays(MA.l, method="scale")
```

74| Check the distribution of spot intensities for red and green channels after each stage of normalization (**Fig. 3**). These distributions should be fairly well aligned and should have tails with high signal values. Experiments in which signal intensity drops off more sharply will often show higher levels of noise in the final data set. (Here and in subsequent steps, text following the '#' symbol represents non-executed comments.)

```
> plotDensities(r)       # Raw data
> plotDensities(MA.l)    # After within-array normalization
> plotDensities(MA.q)    # After between-array normalization
```

75| Create MA plots to check for a relationship between the ratio of dye intensities (M) and their average (A)(**Fig. 4**). Points will often be skewed to low Cy5/Cy3 ratios at low intensities due to photobleaching of Cy5, but should be corrected after within-array loess normalization in LIMMA. This bias is the most common artifact for NimbleGen arrays but other types can also be diagnosed with MA plots[40].

```
> plotMA(r, array=1)      # Raw data, replicate 1
> plotMA(MA.l, array=1)   # After within-array normalization
```

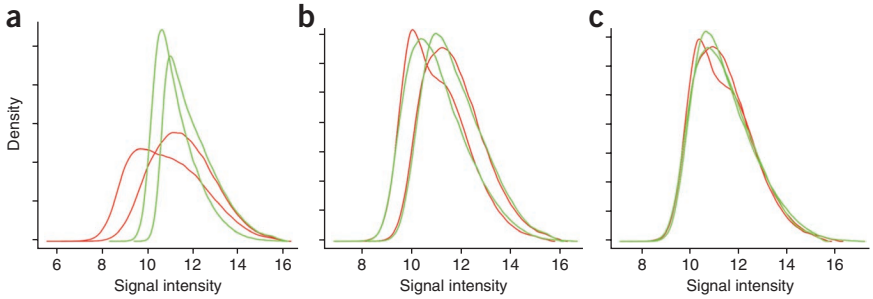

Figure 3 | Distribution of signal intensities before and after normalization. (**a**–**c**) Panels depict the distribution of Cy5 (red) and Cy3 (green) signal values before normalization (**a**), after within-array normalization (**b**), and after between-array normalization (**c**) in LIMMA. (**b**,**c**) As RT is a relative property, equivalent amounts of DNA are transcribed before and after the middle of S-phase, allowing distributions to be transformed to a common scale for each channel (**b**) and array (**c**).

76| Verify that the distribution RT values are equivalent across experiments after normalization by creating boxplots of Cy5/Cy3 ratios for each experiment (**Fig. 5**). These distributions may be slightly different before normalization (and after within-array normalization), but first and third quartiles (the box boundaries) of all experiments should be equal after between-array normalization.

```
> boxplot(MA.l$M~col(MA.l$M), names=colnames(MA.l$M))
> boxplot(MA.q$M~col(MA.q$M), names=colnames(MA.q$M))
```

77| Create an intermediate file containing the normalized data sets by typing, for example:

```
> write.table(MA.q$M, file="Loess_mLymph_112909.txt", quote=F, row.names=F, sep="\t")
```

This tab-delimited text file will be further processed in Steps 79–85 to sort and average the normalized data sets and check other quality control measures.

78| Remove the other objects from memory.

```
> rm(r, MA.l, MA.q, mLymph1Cy3, mLymph1Cy5, mLymph2Cy3, mLymph2Cy5, mLymph1,
mLymph2); gc(reset=T)
```

Or, remove all objects.

```
> rm(list=ls())
```

79| Assign position and chromosome information to the normalized data sets. This can be accomplished using the original .pair files, which typically contain this information in columns 'POSITION' and 'SEQ_ID', respectively (option A). Some data formats, such as HD2 triplex arrays, contain a different format of SEQ_ID column with chromosome and chromosome end points combined (e.g., 'chr11:1–134452384') or no SEQ_ID column. In these cases, extract chromosome labels from the PROBE_ID column (option B)

(A) Copy position and chromosome columns from original .pair files

(i) Read the intermediate file created in Step 77:

```
> tab5rows = read.table("Loess_mLymph_112909.txt", header = T, nrows = 5)
> classes = sapply(tab5rows, class)
> RT = read.table("Loess_mLymph_
112909.txt", header=T, nrows=389306,
comment.char = "", colClasses=classes)
```

(ii) Next, read the original .pair file containing POSITION and SEQ_ID columns:

```
> tab5rows = read.delim("L1210Lym
phoblastP1_635.pair", header = T,
nrows = 5, skip=1)
> classes = sapply(tab5rows, class)
> a = read.delim("L1210LymphoblastP
1_635.pair", header=T, nrows=389306,
comment.char = "", colClasses=classes)
```

(iii) Finally, remove unmapped probes from the files loaded in Steps 79A(i) and (ii) and assign position and

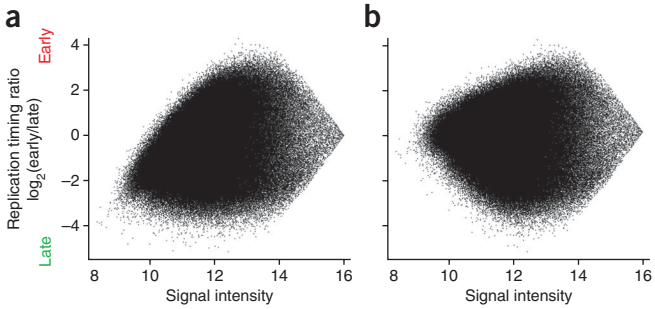

Figure 4 | Dependence of timing ratios on signal intensity. (**a**,**b**) MA plots from LIMMA illustrate the relationship between red/green ratios (*y* axis) and signal intensity (*x* axis) before (**a**) and after (**b**) within-array normalization. The skew of low-intensity data pointing toward Cy3 (here, late) values is a common characteristic of two-color arrays, and is corrected after normalization.

chromosome information to the normalized data sets:

```
> RT = subset(RT, a$POSITION != 0)
> a = subset(a, a$POSITION != 0)
> RT$CHR = a$SEQ_ID; RT$POSITION = a$POSITION
```

(B) Parse position and chromosome information from PROBE_ID column

(i) Load the normalized and .pair files as outlined in Steps 79A(i) and 79A(ii).

(ii) Split the PROBE_ID column into the elements preceding and following 'FS'; for example, 'CHR12FS006244334' will become 'CHR12' and '006244334'.

```
> x = strsplit(as.character(a$PROBE_ID), "FS")
> y = unlist(x)      # chr [1:770156] "CHR01""003001832" ...
```

(iii) Separate the odd- and even-numbered indices of this object into separate columns and convert the position strings to numeric values.

```
> y1 = y[c(TRUE, FALSE)]    # chr [1:385078] "CHR01""CHR01" ...
> y2 = y[c(FALSE,TRUE)]    # chr [1:385078] "003001832""003018759" ...
> y2 = as.numeric(y2)    # num >[1:385078] 3001832 3018759 ...
```

(iv) Finally, assign the position and chromosome information to the normalized data set:

```
> RT = data.frame(CHR=y1, POSITION=y2, RT, stringsAsFactors=F)
```

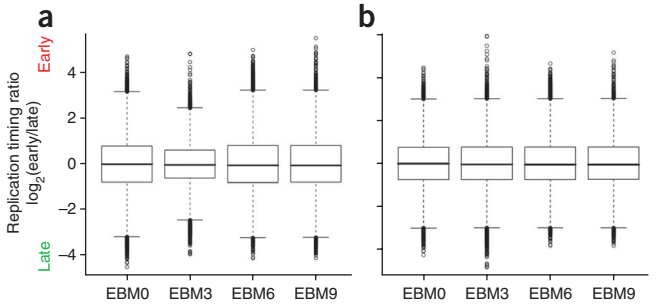

Figure 5 | Verification of scale normalization between data sets. (**a**,**b**) Exemplary boxplots of timing values before (**a**) and after (**b**) normalization between arrays, for a 9-d differentiation from embryonic stem cells to neural precursor cells with 3-d intermediates: (EBM0 (embryonic stem cells); EBM3, EBM6 and EBM9 (neural precursor cells))[5]. Modest differences in the distribution of timing values (with box boundaries representing the first and third quartiles) are equalized after scaling.

80| Sort data sets by chromosome and position. This will ensure that the plotting and autocorrelation checks in Steps 81 and 84 are accurate and that they are required for most downstream analysis. By the default sorting method, the order of mouse chromosomes will be 1, 10–19, 2–9, X and then Y. This order itself is unimportant but should be consistent across experiments to prevent errors in downstream analysis.

```
> RT = RT[order(RT$CHR, RT$POSITION),]
```

81| Plot timing values across a chromosome (**Fig. 6**). This serves to verify the orientation for early/late domains, as well as the overall technical quality of the experiments. Check the data set structure using 'str(RT)' for the correct column numbers to plot and adjust the *y* axis span ('ylim') as needed.

? TROUBLESHOOTING

```
> RTb = subset(RT, RT$CHR == "chr1")  # Create a subset of timing values in chromosome 1
> par(mar=c(3.1,4.1,1,1),mfrow=c(2,1))  # Set plot margins; include two rows in layout
> plot(RTb[,1]~RTb$POSITION,pch=19,cex=0.2,col="grey",ylim=c(-3,3))  # Plot replicate 1
> plot(RTb[,2]~RTb$POSITION,pch=19,cex=0.2,col="grey",ylim=c(-3,3))  # Plot replicate 2
```

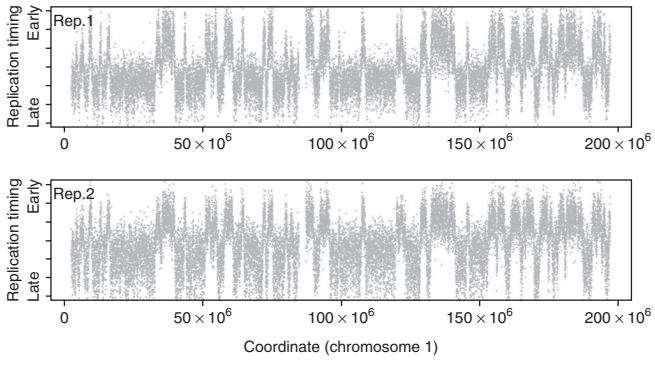

Figure 6 | Replication timing values across chromosome 1. For each replicate, individual $\log_2$(Cy5/Cy3) probe intensities are plotted in gray (*y* axis) against their position on chromosome 1 (*x* axis). Because of photobleaching of Cy5 diagnosed in Step 75, timing is skewed toward early values in replicate 1 (top, Rep. 1) and late values in replicate 2 (bottom, Rep. 2), illustrating the practical advantages of dye-swap replicates.

82| Using known regions of early or late replication, verify that the timing values are properly oriented. If not, reverse them by multiplying the appropriate data columns by −1

```
> RT[,1] = RT[,1] * -1
```

83| Rename data sets and average replicates as desired, then write a finalized file containing normalized data to the current working directory (see Step 63), for example,

```
> names(RT)[1:2] = c("mLymphR1", "mLymphR2")
> RT$mLymphAve = (RT[,1] + RT[,2]) / 2
```

```
> write.table(RT, "
LoessScale+CHRPOS_mLymph_
112909.txt", row.names=F,
quote=F, sep="\t")
```

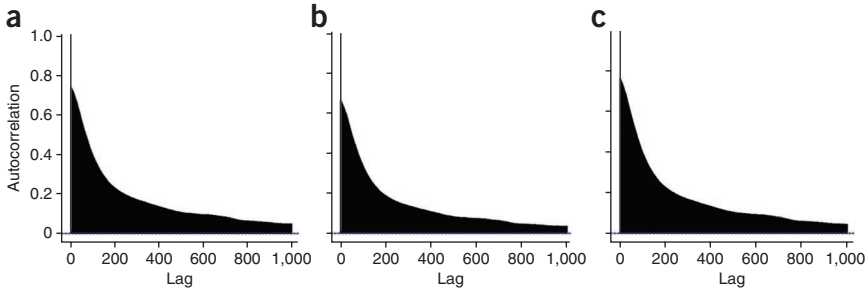

Figure 7 | Autocorrelation functions for two RT experiments and their average. Correlation values (*y* axis) decline as a function of genomic distance between data points (with 'lag' on the *x* axis representing the separation between probes), but should start above 0.6 for high-quality data sets and improve upon averaging replicates.

84| For each data set, determine the autocorrelation function (ACF), which describes the correlation between neighboring data points as a function of their genomic distance (**Fig. 7**). As nearby loci should replicate almost synchronously, the ACF is a useful measure of overall data quality. High-quality data sets will have a correlation between nearest neighbor timing values of R = 0.60–0.80. This measure of signal-to-noise ratio will improve as more replicates with equivalent states are averaged.

```
> acf(RT[,1],lag=1000)$acf[2]    # Replicate 1: R = 0.742

> acf(RT[,2],lag=1000)$acf[2]    # Replicate 2: R = 0.665

> acf(RT$mLymphAve, lag=1000)$acf[2]  # Averaged 1 and 2: R = 0.762
```

? TROUBLESHOOTING

85| To check for spatial artifacts, examine the original .tif images (**Fig. 8**) for common characteristics of regional bias, such as streaks, blank regions or overabundance of either channel in any region of the array[41]. Note that the 'rtiff' package may first need to be installed as in Step 72. As most probes on tiling microarray designs are randomly distributed with respect to genomic location, spatial artifacts in the scanned images should not affect timing values to a large extent in any particular location in the genome, but may reduce the overall signal-to-noise ratio of the experiment if they cover a substantial portion of the array.

```
> library(rtiff)
> Cy5 = readTiff("318990_3MEFfemale_532.tif")
> plot(Cy5)
```

Static properties of the timing program in a given cell type ● TIMING 3 h

86| After normalization, choose among several common options to analyze the characteristics of timing data sets. Although optional, each method is complementary and useful for a wide range of downstream analysis. To derive an overall timing profile from noisier raw data points, apply a loess smoothing function (option A). Use a correlation metric, generally after LOESS smoothing, to determine the overall levels of similarity among two or more data sets (option B). Perform segmentation (option C) to define the boundaries of replication domains and determine their average timing.

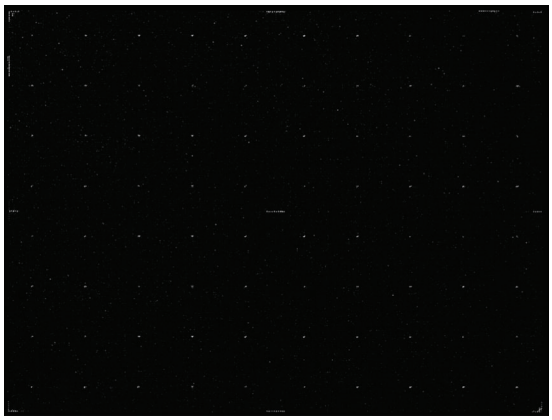

Figure 8 | A typical NimbleGen microarray image after a successful experiment. The lighter points in a grid pattern are control features that aid with spot alignment.

(A) LOESS smoothing

(i) Apply LOESS smoothing to each chromosome as outlined below (**Fig. 9**). For human and mouse data sets, we perform smoothing with a bandwidth of 300 kb; other systems may have different optimal smoothing spans that should be determined empirically using the smallest span that reproduces most of the features between replicate profiles.

```
> chrs = levels(RT$CHR);
str(chrs)    # Create a list of all
chromosomes
```

```
> AllLoess = NULL        # Initialize a variable to store all loess-smoothed data

> for (chr in chrs) {        # For each chromosome,

> RTl = NULL        # Create a variable to store loess-smoothed values

> RTb = subset(RT, RT$CHR == chr)  # Subset the data set to a single chromosome

> lspan = 300000/(max(RTb$POSITION)-min(RTb$POSITION))  # Set smoothing span

> cat("Current chromosome: ", chr, "\n")  # Output current chromosome to console

> RTla = loess(RTb$ mLymphR1~ RTb$POSITION, span = lspan)  # Smooth data set 1

> RTlb = loess(RTb$mLymphR2~ RTb$POSITION, span = lspan)  # Smooth data set 2

> RTlc = loess(RTb$mLymphAve ~ RTb$POSITION, span = lspan)# Smooth data set 3

> RTl = data.frame(CHR=RTb$CHR, POSITION=RTb$POSITION, RTla$fitted, RTlb$fitted, RTlc$fitted)        # Combine the data sets for the current chromosome

> AllLoess = rbind(AllLoess, RTl)    # Combine current chromosome with overall data set

> }          # End for loop

> x = as.data.frame(AllLoess)    # Reformat the smoothed data sets as a data frame
```

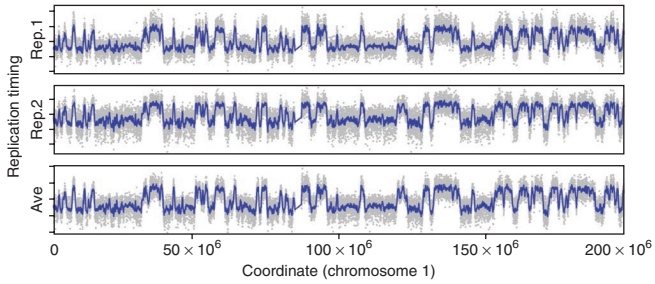

Figure 9 | Raw (gray) and LOESS-smoothed (blue) RT values along chromosome 1.

(ii) Rename the LOESS-smoothed data sets as desired and save these to a tab-delimited text file. Note that column names within a data frame cannot begin with a number.

```
> names(x)[3:5] = c("x300smo_mLymphR1", "x300smo_mLymphR2", "x300smo_mLymphAve)

> write.table(x, "300kb_LoessSmo_mLymph_112909.txt", row.names=F, quote=F, sep="\t")
```

(iii) Plot the results of LOESS smoothing as follows (**Fig. 9**). The "mfrow" parameter may be adjusted for different numbers of data sets.

```
> RTc = subset(RT, CHR == "chr1")        # Subset of raw timing data in chr1

> LSc = subset(LS, CHR == "chr1")        # Subset of smoothed data in chr1

> par(mar=c(2.2,5.1,1,1), mfrow=c(3,1), col="grey", pch=19, cex=0.5, cex.lab=1.8, xaxs="i")

> plot(RTc$mLymphR1~RTc$POSITION, ylab="mLymph R1", xaxt="n")  # Plot raw data points

> lines(LSc$x300smo_mLymphR1~LSc$POSITION, col="blue3", lwd=3)  # Overlay loess line

> plot(RTc$mLymphR2~RTc$POSITION, ylab="mLymph R2", xaxt="n")
```

```
> lines(LSc$x300smo_ mLymphR2~LSc$POSITION, col="blue3", lwd=3)

> plot(RTc$mLymphAve~RTc$POSITION, xlab="Coordinate (bp)", ylab="mLymph ave")

> lines(LSc$x300smo_ mLymphAve~LSc$POSITION, col="blue3", lwd=3)
```

(B) Correlations between data sets

(i) Once the technical quality of the array data is established, compare biological replicate experiments to determine the relative level of biological similarity between samples. When comparing different cell types, to isolate biological rather than array quality differences, we typically use LOESS-smoothed averaged replicate data, rather than individual, raw or normalized data:

```
> cor(x[,c(4:6)]
```

	Rep1	Rep2	Ave
Lymphoblast Rep1	1.000	0.978	0.995
Lymphoblast Rep2	0.978	1.000	0.994
Lymphoblast Ave	0.995	0.994	1.000

The cor() function defaults to Pearson correlation, but other methods are available (see ?cor in R). If missing values are present, add 'na.rm=T' to remove them.

(C) Segmentation

(i) Perform circular binary segmentation as outlined in Steps 86C(ii–iv) (**Fig. 10**). Biologically, these segments (or 'replication domains') appear to correspond to domains of coordinately regulated, synchronously firing origins that may be part of replication factories. We perform segmentation as follows using the DNACopy algorithm designed by Venkatraman *et al.*[42], which performs favorably compared with alternatives for CGH copy number analysis[43–45].

(ii) First, load the DNAcopy package and prepare a CNA (copy number array) object for segmentation

```
> library(DNAcopy)

> mLymph = CNA(RT$mLymphAve, RT$CHR, RT$POSITION, data.type="logratio",
sampleid = "mLymph")
```

(iii) Next, segment the CNA object with the desired parameters. The parameters shown are those that we have used for analysis of mouse and human timing data sets, with autocorrelations near 0.8[3,6]; data of different quality or in different formats may require these to be determined empirically.

```
> Seg.mLymph = segment(mLymph, nperm=10000, alpha=1e-15, undo.splits="sdundo",
undo.SD=1.5, verbose=2)
```

(iv) Examine the resulting segmentation object 'Seg.mLymph', which contains the raw data and segmentation breakpoints assigned by circular binary segmentation[46]. The number of segments assigned can be determined using str(Seg.mLymph$output) and visualized using various functions built into DNAcopy (**Fig. 10**).

```
> par(ask=T,mar=c(3.1,4.1,1,1))     # Set figure margins; ask before replotting

> plot(Seg.mLymph, plot.type="c")     # Plot each chromosome separately

> plot(Seg.mLymph, plot.type="s")     # Plot overview of all chromosomes
```

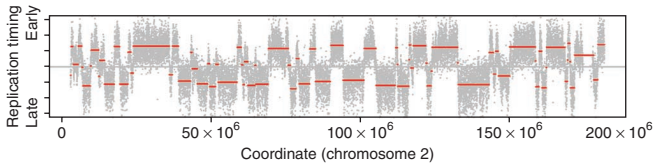

Figure 10 | Raw data (gray) overlaid with segmented timing domains (red) along chromosome 2, as defined by circular binary segmentation[42].

```
> plot(subset(Seg.mLymph,chromlist="chr2"), pch=19, pt.cols=c("gray","gray"),
xmaploc=T, ylim=c(-3,3))          # Plot a single chromosome with alternate format
```

(v) Create a tab-delimited text file containing segment end points and average RT values for each segment. The file will be written to the current working directory (see Step 63).

```
> write.table(Seg.mLymph$output, row.names=F, quote=F, sep="\t")
```

(vi) After segmentation, calculate the sizes of replication domains from the segmented data set to examine average sizes for domains with early or late timing.

```
> Lymph = Seg.mLymphR1$output      # Extract domain information

> Lymph$size = Lymph$loc.end - Lymph$loc.start    # Calculate domain sizes

> LymphEarly = subset(Lymph, Lymph$seg.mean > 0)  # Create subset of early domains

> LymphLate = subset(Lymph, Lymph$seg.mean < 0)  # Create subset of late domains

> boxplot(LymphEarly$size, LymphLate$size)    # Distribution of early/late do-
main sizes
```

Dynamic changes in the timing program ● TIMING 3 h

87|To examine changes in the replication program during differentiation, use one or several of the methods in this step to leverage the segmentation and loess smoothing methods introduced in Step 86. As no single method is sufficient to fully describe the type, degree and distribution of timing changes during development, we cover several complementary ways to measure these properties and explore the relationships between cell types. These include the following: (A) The amount of the genome changing RT (percentage change analysis); (B) The degree and relationships of RT changes between cell types (clustering approaches); and (C) The properties of domains that change timing on differentiation (switching domain analysis).

(A) Percentage change analysis

(i) Determine the amount of the genome with differential timing between two or more cell types using an arbitrary, percentile or significance-based cutoff for RT changes. We recommend scaling data sets to equivalent ranges and applying an empirical cutoff for changes verifiable by PCR to quantify these genome wide, as shown here. As most methods for quantifying timing changes are sensitive to scale differences, data sets should be first scaled and normalized together in LIMMA (see Steps 62–76).

```
> RTd1 = RT$mLymphR1 - RT$mLymphR2     # Calculate timing differences between
data sets

> mLength = length(RTd1)       # Determine total number of probes

> s = 0.67          # Set cutoff for significant changes

> sum(abs(RTd1)>s)/mLength     # Percentage changing, R1 vs. R2

> sum(RTd1 < -s)/mLength       # Early to Late changes: 1.6% of all probes

> sum(RTd1 > s)/mLength       # Late to Early changes: 1.3% of all probes
```

(B) Clustering approaches

(i) Perform clustering to aggregate experiments with similar timing patterns. For *k*-means clustering we have used the programs Cluster[47] and TreeView (http://rana.lbl.gov/EisenSoftware.htm) and refer readers to their corresponding guides. For hierarchical clustering, we use the 'pvclust' package in R[48] to compute clusters based on the stability of connections between cell types and ascribe *P* values to each node.

(ii) To improve the precision of individual RT measurements and lessen the considerable computational expense of most

clustering algorithms, average individual timing values into larger windows before clustering. We typically average data sets in windows of ~200 kb.

```
> mLymph.R1 = NULL; mLymph.R2 = NULL    # Initialize variables to store averaged data

> nWin = 35           # 5.8 kb median probe spacing * 35 = 203 kb

> mLength = nrows(RT)/nWin      # Calculate number of windows

> for (x in 1:mLength) {       # For each potential window,

>   z1 = x * nWin         # Determine probe number at window start

>   z2 = (x+1) * nWin       # Determine probe number at window end

>   mLymph.R1[x] = mean(RT$mLymphR1[z1:z2]) # Average replicate 1 across window

>   mLymph.R2[x]= mean(RT$mLymphR2[z1:z2])   # Average replicate 2 across window

>   cat("Current window: ", x, "/", mLength, "\n")  # Write the current window to
the console

> }            # End for loop

> RTWind = data.frame(mLymph.R1, mLymph.R2)  # Write the results to a new data frame
```

(iii) Load the pvclust[48] package and use its corresponding function to cluster data sets using multiscale bootstrap re-sampling, which will assign *P* values to each node in the hierarchical clustering dendrogram. See ?pvclust after loading the package for additional options and settings.

```
> library(pvclust)

> cluster.bootstrap <- pvclust(RTWind, nboot=1000, method.dist="abscor")
```

(iv) Plot the cluster dendrogram as performed below.

```
> plot(cluster.bootstrap)    # Plot overall dendrogram

> pvrect(cluster.bootstrap)    # Outline data sets that cluster at a significant level
```

▲ **CRITICAL STEP** Take care when interpreting the results of hierarchical clustering, as a wide variety of topologies are possible for a single dendrogram, as any node can be flipped horizontally without changing the connections between clusters; agglomerative clusters can change substantially when new experiments are added; and the exact connections produced (although usually not the overall structure of the dendrogram) often change for different clustering algorithms or distance metrics.

(C) Properties of RT switching domains

(i) Perform segmentation on the differences between timing profiles to define the boundaries of domains that switch to earlier or later replication (switching domains) and analyze the properties of genetic and epigenetic elements within them. To compute these domains, first subtract the normalized (not LOESS-smoothed) values of the two experiments to be compared and create a CNA object in a manner similar to Step 86C(ii).

```
> dRT = CNA(RT$NPCave-RT$ESCave, RT$CHR, RT$POSITION, data.type="logratio",
sampleid="NPC-ESC dRT")
```

(ii) Next, segment the resulting object, calculate domain sizes and write the segments to a tab-delimited text file.

```
> Seg.dRT = segment(dRT, nperm=10000, alpha=1e-15, undo.splits = "sdundo", undo.
SD=1.5, verbose=2); dRTdom = Seg.dRT$output
```

```
> dRTdom$size = dRTdom$loc.end - dRTdom$loc.start

> write.table(dRTdom, "Switching segments, mNPC vs. mESC.txt", row.names=F,
quote=F, sep="\t")
```

(iii) Identify domains with the largest timing changes in either direction, as well as domains with stable timing between data sets, using cutoffs from the quantile() function.

```
> quantile(dRTdom$seg.mean, probs = c(0.05, 0.95))   # Top 5% of changes to
early/late

> quantile(dRTdom$seg.mean, probs = c(0.40, 0.60))   # Middle 20% of smallest changes

> LtoEdom = subset(dRTdom, dRTdom$seg.mean > 1.28552)  # Isolate late-to-early domains

> EtoLdom = subset(dRTdom, dRTdom$seg.mean < -1.32328)  # Isolate early-to-late domains

> middleDom = subset(dRTdom, dRTdom$seg.mean > -0.14808  # Isolate non-switching domains

& dRTdom$seg.mean < 0.23698)

> boxplot(middleDom$size, LtoEdom$size, EtoLdom$size)  # Plot distributions of
domain sizes
```

Comparison and alignment to outside data sets ● TIMING 6 h

88| Choose among several alternative approaches to compare the timing program with the vast array of genome-wide or gene-centric data made available through initiatives such as ENCyclopedia Of DNA Elements[49–51] and public repositories such as Gene Expression Omnibus[52,53]. To study gene-level regulation, assign RT values (option A) and epigenetic marks (option B) to lists of RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) or other gene locations. For domain-level analysis, average values within the boundaries of replication domains (option C).

(A) Assignment of RT values to gene promoters

(i) Assign LOESS-smoothed timing values to gene promoters as outlined below. Although the main purpose of LOESS is to derive an overall smoothed RT profile, the smoothed data object produced can be interrogated at any set of genomic coordinates, making it especially valuable for comparing data sets from different array platforms and coordinates. Using this approach, we assign timing data to the RefSeq gene promoter locations of NCBI as follows:

(ii) Begin by loading the required data sets; these include a table of RefSeq gene locations for the desired species (found at http://www.ncbi.nlm.nih.gov/RefSeq/) and a list of smoothed RT values created in Step 86 and loaded as in Step 65.

(iii) Next, create a list of chromosomes to be analyzed and variables to store the data in each chromosome.

```
> chrs = levels(RefSeq$CHR)    # Create a list of chromosomes to be analyzed

> AllSm = NULL       # Variable to store smoothed data for all chromosomes

> ChrSm = NULL       # Variable to store smoothed data for one chromosome
```

(iv) Run the following loop to calculate RT values at transcription start sites of RefSeq genes. Advanced R users may substitute an appropriately reformatted function if desired, and the approach below may be used generically to apply values from any data type regulated on large scales (relative to array probe density) to any list of genomic coordinates.

```
> for(chr in chrs) {       # For each chromosome,

>  RTc = subset(RT, CHR == chr)     # Create subset of timing values in the chromosome

>   RSc = subset(RefSeq, CHR == chr)   # Create subset of RefSeq genes in the chromosome

>  cat("Current chromosome: ", chr, "\n")  # Output current chromosome to console
```

```
> lspan = 300000/(max(RTc$POSITION)-min(RTc$POSITION))    # Set smoothing span

> smLym1 = loess(RT$mLymphR1 ~ RT$POSITION, span = lspan)  # Smooth data set 1

> smLym2 = loess(RT$mLymphR2 ~ RT$POSITION, span = lspan)  # Smooth data set 2

> smLym3 = loess(RT$mLymphAve ~ RT$POSITION, span = lspan)  # Smooth data set 3

> Lym1 = predict(smLym1, RSc$TSS)  # Predict (interpolate) values at transcrip-
tion start sites

> Lym2 = predict(smLym2, RSc$TSS)  # Predict values for data set 2

> Lym3 = predict(smLym3, RSc$TSS)  # Predict values for data set 3

> ChrSm = data.frame(CHR=chr,POSITION= RSc$TSS, Lym1, Lym2, Lym3)

> AllSm = rbind(AllSm, ChrSm)    # Combine information for all experiments/chromosomes

>    }       # End for loop
```

(v) As in Steps 78 and 79, write the results of analysis to an external file before unloading the data from memory:

```
> write.table(AllSm, "Mouse lymphoblast RT at RefSeq gene positions.txt",
quote=F, sep="\t")
```

(B) Assignment of histone and other epigenetic marks to gene promoters

(i) Assign epigenetic and other data sets to gene promoters using Steps 88B(ii–v). Unlike RT, values from epigenetic data sets are often too sparse, relative to their unit of regulation, to apply the method in option 88A. For this example, we assign values from a generic genome-wide ChIP-seq experiment to windows +500 to −2,500 bases from RefSeq gene promoters.

(ii) As in option A, first load the required data sets as described in Steps 65 and 88A(ii). Two files are required: one with columns describing the genomic coordinate, orientation (+/−) and chromosome of each gene (read into a variable named "RefSeq") and another with the coordinate, chromosome and data value for each mark (read into variable 'Marks').

(iii) Create a list of chromosomes to be analyzed and variables to store the assigned values:

```
> chrs = levels(Marks$CHR); AllGenes = NULL; AllHist = NULL
```

(iv) Run the following loop to assign values near transcription start sites to RefSeq genes. We generally set the apply function to assign the highest value within the promoter window to the gene; other approaches include averaging the number of reads within the body of genes[54], individually analyzing equally spaced bins across open reading frames[55] and assessing promoters with significant binding above background[56]. Bear in mind that the transcription start site may not be the best target for all modifications; indeed, for trimethylated lysine 36 of histone H3 (H3K36me3) marking transcription elongation, values at the transcription end point or exon 5′ ends may better represent overall enrichment[57].

```
> for (chr in chrs) {       # For each chromosome,

>  RSc = subset(RefSeq, CHR == chr)  # Create subset of RefSeq genes in the chromosome

>  MKc = subset(Marks, CHR == chr)  # Create subset of mark values in the chromosome

> for(m in 1:nrow(RSc)) {     # For each gene in the chromosome,

>   if(RSc[m,]$Strand == "+") {    # If the gene is in the forward orientation,

>     RTcSub = subset(RTc, (RTc$Start < RSc[m,]$txStart +500) & (RTc$Start
> RSc[m,]$txStart - 2500))    # Collect values from txStart +500 to -2500bp
```

```
>     AllHist = rbind(AllHist, apply(RTcSub, 2, max)[3:12]) # Assign max value to gene

>     AllGenes = rbind(AllGenes, RSc[m,]$Gene)   # Combine with overall list

>       }     # End if

>    if(RSc[m,]$Strand == "-") {    # If the gene is in the reverse orientation,

>     RTcSub = subset(RTc, (RTc$Start < RSc[m,]$txEnd +2500) &
(RTc$Start >        RSc[m,]$txEnd - 500))    # Collect values from txEnd +2500
to -500 bp

>     AllHist = rbind(AllHist, apply(RTcSub, 2, max)[3:12]) # Assign max value to gene

>     AllGenes = rbind(AllGenes, RSc[m,]$Gene)   # Combine with overall list

>      }     # End if

>   cat("Chromosome:", chr, " Gene:", m, "/", nrow(RSc), "\n") # Output current gene

>     }         # End gene loop

>     }       # End chromosome loop
```

(v) Finally, similarly to previous steps, combine the gene and epigenetic mark information into a single table and output as tab-delimited text.

```
> OutFile = data.frame(cbind(AllGenes, AllHist), stringsAsFactors=F)

> write.table(OutFile, file="Histone modifications at RefSeq gene positions.txt",
row.names=F, quote=F, sep="\t")
```

(C) Integration of epigenetic mark values over replication domains

(i) Use the method below to correlate domain-wide RT values and the average level of epigenetic marks within timing domains segmented in Step 86C (for static timing domains) or 87C (for domains that switch timing). Given that the magnitude of correlations between genetic properties generally increases when measured in larger windows, it is important to quantify these relationships in windows consistent with biologically regulated unit sizes.

(ii) Read the replication domains created in Step 86C or 87C into variable 'Seg.RT'.

```
> Seg.RT = read.table("Lymph-1 segments.txt",header=T)
```

(iii) Create a list of chromosomes and variables in which to store average epigenetic mark and timing values.

```
> chrs = levels(Seg.RT$chrom)

> MarksData = NULL; RTData = NULL
```

(iv) Run the loop below to assign the average values of one or multiple epigenetic data sets to each replication domain or modify as needed.

```
> dom = 0            # Initialize domain number to 0

> for(chr in chrs) {          # For each chromosome,

>  Seg.RTb = subset(Seg.RT, Seg.RT$chrom == chr)  # Get timing domains in chromosome

>  MarksB = subset(Marks, Marks$CHR == chr)    # Get mark data in chromosome

>  for (i in 1:dim(Seg.RTb)[1]) {       # For each domain,
```

```
>    cat("Current chr:", chr, " Domain:", dom, "\n")  # Output current domain

>    MarksD = subset(MarksB, MarksB$Start > Seg.RTb[i,]$loc.start &
MarksB$Start <   Seg.RTb[i,]$loc.end)       # Find subset of marks in domain

>    MarksD = MarksD[,3:12]     # Exclude chr/pos from mark data

>    MarksD[,1:10] = MarksD[,1:10] - MarksD[,1]  # Subtract control values, if needed

>    MarksData = rbind(MarksData, apply(MarksD,2, "mean")) # Average mark data in domain

>     dom = dom + 1         # Increment domain number

>     }           # End domain loop

>     }           # End chromosome loop
```

(v) Finally, find the correlations between domain-wide RT and each type of epigenetic mark and create scatter plots to visualize these relationships.

```
> cor(Seg.RT$seg.mean, data.frame(MarksData))

> plot(Seg.RT$seg.mean, data.frame(MarksData)[1])
```

? TROUBLESHOOTING

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 2**.

TABLE 2 | Troubleshooting table.

Step	Problem	Possible reason	Solution
6	Cell aggregation or debris accumulation prevents accurate cell sorting	Failure to achieve single-cell suspension with certain problematic cell types	Incubate with enzyme treatment, such as Trypsin-EDTA or Accutase, for a longer period of time. Use gentle trituration to ensure that cell aggregates are broken apart before fixation and/or sorting. Occasional pausing and filtering of cell samples during FACS may help
		Vortexing during ethanol fixation was too harsh	Use the lowest vortex setting available while adding ethanol dropwise
49	Inconsistent PCR bands between aliquots of the same sample	Contamination between fractions during FACS, probably due to problems in cell fixation	Switch from PI staining (with fixation) to DAPI staining (without fixation)
		Inconsistent number of cells aliquotted to each tube	Mix contents thoroughly before aliquotting and freezing for storage. Aliquot 20,000 cells while the samples are hot to avoid pipetting errors as a result of SDS formation in the solution
		Insufficient BrdU labeling time	Incubate growing cells with BrdU for a longer period of time. Cells with longer S-phase require longer BrdU incubation times
		Varying efficiency of BrdU-IP between samples caused by loss of DNA-protein pellet	Use caution when removing supernatant from the loose DNA-protein pellet. Centrifuge the sample multiple times, as needed, to remove supernatant without disturbing the pellet

(continued)

TABLE 2 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
57	Samples do not pass screening	Bias created during WGA	Increase the amount of starting material for WGA. For instance, start with 100 µl of IP sample pool instead of 50 µl at Step 52
81	Skew toward early or late values	Bias created during WGA or labeling, or excessive photobleaching during scanning	Check early versus late WGA yields, and avoid multiple scans of the array
84	Low autocorrelation (high noise level)	Values are not properly sorted by chromosomal location	Ensure that chromosome and position columns are properly assigned to experimental values and sorted as in Step 80
		Low signal intensity (in Step 74)	Check yield after labeling and amplification steps, as well as scanner settings
88C(v)	Large difference in domain numbers between similar data sets	Sensitivity of segmentation algorithms to differences in data quality	Either adjust the parameter undo.SD (using similar autocorrelation-level data sets as a guide) or add Gaussian noise to higher-quality data sets to equalize their ACF (Step 84) before segmentation

● TIMING

Steps 1–12, BrdU labeling and FACS sorting: 5–6 h
Steps 13–44, BrdU immunoprecipitation: 2–3 d
Steps 45–50, PCR assay: 4–6 h
Steps 51–57, Whole-genome amplification: ~5 h
Box 1, S/G1 FACS sorting: ~1 d
Step 58, Dye labeling: 3–4 h
Step 59, Hybridization: ~1 h plus hybridization time
Steps 60 and 61, Washing and scanning: 1–2 h
Steps 62–85, Normalization: ~1 d
Step 86, Static properties: ~3 h
Step 87, Dynamic properties: ~3 h
Step 88, Outside data sets: ~6 h

ANTICIPATED RESULTS

Our research has shown that the described method is a powerful tool for genome-scale analysis of RT. However, meaningful data analysis is dependent on the quality of available data. Therefore, measures should be taken throughout the protocol to ensure that each phase of the procedure produces quality starting material for subsequent phases. Anticipated results for various steps of the protocol are described here. Typical FACS plots showing successful DNA content analysis and indicating appropriate S-phase fractions to be collected are shown in **Figure 1**.

Following cell sorting and BrdU-IP, marker genes with known relative RT (**Table 1**) should be amplified by PCR for multiple IP samples and detected by electrophoresis on an agarose gel. Among the mouse sequences listed in **Table 1**, mitochondrial DNA replicates throughout the cell cycle[58] and will typically be equally represented in early and late S-phase fractions. *Hba-a1*, *Pou5f1* and *Mmp15* are generally early replicating markers, whereas *Hbb-b1*, *Zfp42*, *Dppa2*, *Ptn*, *Mash1* and *Akt3* are generally late replicating markers. Note that some genes switch RT at some point during development; for instance, *Zfp42* and *Dppa2* are early replicating in ESCs, but late replicating in all somatic cell types examined to date. Therefore, consistency across multiple samples from the same cell type is usually the most reliable way to assess the quality of IP samples. Among the human sequences listed in **Table 1**, mitochondrial DNA is equally represented in early and late S-phase fractions, whereas *HBA1*, *MMP15* and *BMP1* are generally early replicating markers. *PTGS2*, *NET01*, *SLITRK6*, *ZFP42* and *DPPA2* are generally late replicating. High-quality IP reactions show consistency in the relative amount of BrdU-labeled DNA in respective S-phase fractions between samples of the same cell type. This PCR analysis should be performed again directly following WGA in order to ensure that no bias has been introduced during this step of the procedure. If no bias is detected, 4–8 µl of purified WGA3 DNA should be run on a 1.5% agarose gel in order to determine its quality. Quality DNA will range in size from 100 to 1,000 bp, with an average size of ~400bp. In addition, WGA3 DNA should have an A_{260}/A_{280} value ≥1.8 and an A_{260}/A_{230} value ≥1.9

in order to function as high-quality starting material for the labeling reaction. NimbleGen arrays user's guide for CGH analysis should be consulted for anticipated results of the hybridization and scanning procedures.

After a successful experiment, domains of coordinate RT (replication domains) will be clearly visible in the raw data after plotting these across a chromosome (**Fig. 6**). Less-successful experiments will have autocorrelation values below 0.6 (**Fig. 7**), and visibly higher levels of noise, thereby limiting the resolution of smaller replication domains. Further, low signal in MA plots (**Fig. 4**) and signal intensity distributions (**Fig. 3**) will also often present with low autocorrelation, and may indicate a low volume of Cy-labeled DNA or problems with scanning. If several replicate experiments were done, they should have high (>0.90) correlations between LOESS-smoothed timing values (Step 86B).

Note: Supplementary information is available in the HTML version of this article.

ACKNOWLEDGMENTS We thank J. C. Rivera Mulia and A. Rycyk for helpful comments on the manuscript. Research in the Gilbert lab is funded by NIH Grants GM083337 and GM085354.

AUTHOR CONTRIBUTIONS D.M.G. and I.H. conceived the study and designed the experiments. T.R. and I.H. devised the computational methods. T.R., D.B., B.D.P. and D.M.G. wrote the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Published online at http://www.natureprotocols.com/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Pope, B.D., Hiratani, I. & Gilbert, D.M. Domain-wide regulation of DNA replication timing during mammalian development. *Chromosome Res.* **18**, 127–136 (2010).
2. Yaffe, E. *et al.* Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. *PLoS Genet.* **6**, e1001011 (2010).
3. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–770 (2010).
4. Gilbert, D.M. *et al.* Space and time in the nucleus: developmental control of replication timing and chromosome architecture. *Cold Spring Harb. Symp. Quant. Biol.* doi:10.1101/sqb.2010.75.011 (2010).
5. Hiratani, I. *et al.* Genome-wide dynamics of replication timing revealed by *in vitro* models of mouse embryogenesis. *Genome Res.* **20**, 155–169 (2010).
6. Hiratani, I. *et al.* Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.* **6**, e245 (2008).
7. Schwaiger, M. *et al.* Heterochromatin protein 1 (HP1) modulates replication timing of the Drosophila genome. *Genome Res.* **20**, 771–780 (2010).
8. Schwaiger, M. *et al.* Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. *Genes Dev.* **23**, 589–601 (2009).
9. Schübeler, D. *et al.* Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. *Nat. Genet.* **32**, 438–442 (2002).
10. Lee, T.-J. *et al. Arabidopsis thaliana* chromosome 4 replicates in two phases that correlate with chromatin state. *PLoS Genet.* **6**, e1000982 (2010).
11. Koren, A., Soifer, I. & Barkai, N. MRC1-dependent scaling of the budding yeast DNA replication timing program. *Genome Res.* **20**, 781–790 (2010).
12. Raghuraman, M.K. & Brewer, B.J. Molecular analysis of the replication program in unicellular model organisms. *Chromosome Res.* **18**, 19–34 (2010).
13. Hayashi, M. *et al.* Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast. *EMBO J.* **26**, 1327–1339 (2007).
14. Karnani, N., Taylor, C.M. & Dutta, A. Microarray analysis of DNA replication timing. *Methods Mol. Biol.* **556**, 191–203 (2009).
15. Farkash-Amar, S. & Simon, I. Genome-wide analysis of the replication program in mammals. *Chromosome Res.* **18**, 115–125 (2009).
16. Sasaki, T. *et al.* The Chinese hamster dihydrofolate reductase replication origin decision point follows activation of transcription and suppresses initiation of replication within transcription units. *Mol. Cel. Biol.* **26**, 1051–1062 (2006).
17. Gilbert, D.M. Replication origin plasticity, taylor-made: inhibition vs recruitment of origins under conditions of replication stress. *Chromosoma* **116**, 341–347 (2007).
18. Anglana, M. *et al.* Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell* **114**, 385–394 (2003).
19. Gilbert, D.M. Evaluating genome-scale approaches to eukaryotic DNA replication. *Nat. Rev. Genet.* **11**, 673–684 (2010).
20. Gilbert, D.M. Temporal order of replication of *Xenopus laevis* 5S ribosomal RNA genes in somatic cells. *Proc. Natl. Acad. Sci. USA* **83**, 2924–2928 (1986).
21. Gilbert, D.M. & Cohen, S.N. Bovine papilloma virus plasmids replicate randomly in mouse fibroblasts throughout S phase of the cell cycle. *Cell* **50**, 59–68 (1987).
22. Hansen, R.S. *et al.* Association of fragile X syndrome with delayed replication of the FMR1 gene. *Cell* **73**, 1403–1409 (1993).
23. Yokochi, T. *et al.* G9a selectively represses a class of late-replicating genes at the nuclear periphery. *Proc. Natl. Acad. Sci. USA* **106**, 19363–19368 (2009).
24. Lu, J. *et al.* G2 phase chromatin lacks determinants of replication timing. *J Cell Biol.* **189**, 967–980 (2010).
25. Pollack, J.R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23**, 41–46 (1999).
26. Acevedo, L.G. *et al.* Genome-scale ChIP-chip analysis using 10,000 human cells. *BioTechniques* **43**, 791–797 (2007).
27. Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).
28. Yang, Y.H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
29. Bolstad, B.M. *et al.* A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
30. Core, R.D. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2008).
31. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
32. Ihaka, R. & Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).
33. Spector, P. *Data Manipulation with R* (Springer Publishing Company, 2008).
34. Chambers, J.M. *Software for Data Analysis: Programming with R* (Springer Publishing Company, 2008).
35. Crawley, M.J. *The R Book* (Wiley, 2007).
36. Wettenhall, J.M. & Smyth, G.K. limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics* **20**, 3705–3706 (2004).
37. Hansen, R.S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. USA* **107**, 139–144 (2010).
38. Pombo, A. & Gilbert, D.M. Nucleus and gene expression: the structure and function conundrum. *Curr. Opin. Cell Biol.* **22**, 269–270 (2010).
39. O'Geen, H. *et al.* Comparison of sample preparation methods for ChIP-chip assays. *BioTechniques* **41**, 577–580 (2006).
40. Peng, S. *et al.* Normalization and experimental design for ChIP-chip data. *BMC Bioinformatics* **8**, 219 (2007).
41. Reimers, M. & Weinstein, J.N. Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases. *BMC Bioinformatics* **6**, 166 (2005).
42. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
43. Dellinger, A.E. *et al.* Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res.* **38**, e105 (2010).
44. Willenbrock, H. & Fridlyand, J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084–4091 (2005).

45. Lai, W.R. *et al.* Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763–3770 (2005).
46. Olshen, A.B. *et al.* Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
47. Eisen, M.B. *et al.* Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
48. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006).
49. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
50. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
51. Rosenbloom, K.R. *et al.* ENCODE whole-genome data in the UCSC genome browser. *Nucleic Acids Res.* **38**, D620–D625 (2010).
52. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
53. Barrett, T. *et al.* NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.* **37**, D885–D890 (2009).
54. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
55. Salcedo-Amaya, A.M. *et al.* Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of Plasmodium falciparum. *Proc. Natl. Acad. Sci. USA* **106**, 9655–9660 (2009).
56. Guenther, M.G. *et al.* A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88 (2007).
57. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput. Biol.* **5**, e1000566 (2009).
58. Aladjem, M.I. *et al.* Replication initiation patterns in the beta-globin loci of totipotent and differentiated murine cells: evidence for multiple initiation regions. *Mol. Cel. Biol.* **22**, 442–452 (2002).
59. Woodfine, K. *et al.* Replication timing of the human genome. *Hum. Mol. Genet.* **13**, 191–202 (2004).