Transcription profiling during the cell cycle shows that a subset of Polycomb-targeted genes is upregulated during DNA replication

Javier Peña-Diaz[1], Siv A. Hegre[1,2], Endre Anderssen[1], Per A. Aas[1], Robin Mjelle[1], Gregor D. Gilfillan[3], Robert Lyle[3], Finn Drabløs[1], Hans E. Krokan[1] and Pål Sætrom[1,4,*]

[1]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway, [2]St. Olavs Hospital, NO-7006 Trondheim, Norway, [3]Department of Medical Genetics and Norwegian Sequencing Centre, Oslo University Hospital, NO-0407 Oslo, Norway and [4]Department of Computer and Information Science, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway

ABSTRACT

Genome-wide gene expression analyses of the human somatic cell cycle have indicated that the set of cycling genes differ between primary and cancer cells. By identifying genes that have cell cycle dependent expression in HaCaT human keratinocytes and comparing these with previously identified cell cycle genes, we have identified three distinct groups of cell cycle genes. First, housekeeping genes enriched for known cell cycle functions; second, cell type-specific genes enriched for HaCaT-specific functions; and third, Polycomb-regulated genes. These Polycomb-regulated genes are specifically upregulated during DNA replication, and consistent with being epigenetically silenced in other cell cycle phases, these genes have lower expression than other cell cycle genes. We also find similar patterns in foreskin fibroblasts, indicating that replication-dependent expression of Polycomb-silenced genes is a prevalent but unrecognized regulatory mechanism.

INTRODUCTION

Genome-wide studies of gene expression throughout the cell division cycle have revealed several genes that are differentially expressed (1–9), but have also indicated that the set of cycling genes differs between primary and cancer cells (3). Primary cells are, however, inherently difficult to synchronize, for example, only 40–50% of foreskin fibroblast cells in culture can be synchronized by serum starvation or double thymidine block (3). Although sophisticated statistics may partially overcome lack of synchronization (3), a large population of asynchronous or arrested cells results in high background gene expression noise. Consequently, more cycling genes can be detected in a highly synchronous culture than in a culture where at most 50% of the cells are synchronized. Moreover, as the only human cell line—in addition to primary fibroblasts (1,3,4)—profiled for cell cycle expression so far is the cervical cancer cell line HeLa (2,5), it is unclear to what extent cell type-specific factors affect reported differences in cycling genes.

We have used the human keratinocyte cell line HaCaT to address this question. Specifically, by measuring the gene expression profiles of double thymidine synchronized HaCaT cells, we identified three major groups of cycling genes. First, a set of genes with housekeeping characteristics, strong enrichment for known cell cycle functions and overlap with previously identified cell cycle genes. Second, a set of genes with cell type-specific characteristics, enrichment for HaCaT-specific functions and poor overlap with previously identified cell cycle genes. Third, a set of genes that has the mark for Polycomb silencing: histone H3 lysine 27 tri-methylation (H3K27me3). We show that this third set of genes is expressed in a replication-dependent manner, as the genes are upregulated during S phase in a pattern related to DNA replication timing. Consistent with being epigenetically silenced in other cell cycle phases, these genes are generally lower expressed than are other cell cycle expressed genes. We also find

similar patterns in foreskin fibroblasts synchronized by serum starvation, indicating that replication-dependent expression of Polycomb-silenced genes is a prevalent but unrecognized regulatory mechanism.

MATERIALS AND METHODS

HaCaT cell culture and synchronization

HaCaT cells were plated at 10% confluence (1×10^6 cells) in 150-mm tissue culture dishes in Dulbecco's modified Eagle's medium (DMEM) with 10% fetal bovine serum (FBS). Cells were arrested in the interphase G_1/S by double thymidine block; briefly, cells were treated with 2 mM of thymidine for 18 h, released from the arrest for 10 h and arrested a second time with 2 mM of thymidine for additional 18 h. After treatment, media was replaced, and cells were collected at 3-h intervals for up to 33 h, covering approximately two cell cycles. Synchrony was monitored by flow cytometry analysis of propidium iodide-stained cells and by cell counting. Quantification of cells in each phase was done with the MultiCycle DNA cell cycle analysis software (Phoenix Flow Systems Inc., San Diego, CA, USA) combined with the cell counting results.

HeLa cell culture and synchronization

Adherent HeLa cells were plated in 150-mm culture dishes in DMEM with 10% of FBS, 2 mM of glutamine, 0.1 mg/ml of gentamicin and 1.25 µg/ml of fungizone. Cells at 60–70% confluence were arrested in the G_2/M transition with 100 ng/ml of nocodazole for 17 h. The mitotic cells were then collected by manual shake-off, washed twice and re-plated in fresh DMEM to progress through the cell cycle. Cells were harvested from culture dishes by trypsinization every 30 min for the first 2 h and then every 3 h from 3 to 24 h after release. Phosphate-buffered saline containing 3% of FBS was added to inactivate the trypsin.

HeLa cells were pelleted and resuspended in 100 µl of RNAlater (Applied Biosystems/Ambion, Austin, TX, USA). All pellets were kept at 4°C overnight and were stored at −80°C before use. Verification of the cell cycle stage was determined by analysing the DNA content of propidium iodide-stained cells by a BD FACSAria (BD Biosciences, San Jose, CA, USA) flow cytometer. Quantification of cells in each phase was done with FlowJo (Tree Star Inc., OR, USA).

cRNA synthesis and microarray hybridization—HaCaT

Total RNA was extracted using the High Pure RNA Isolation Kit (Roche Applied Science, Indianapolis, IN, USA) and the manufacturer's protocol. RNA from synchronous cells was reverse transcribed into cDNA (cDNA synthesis Kit, Invitrogen, Carlsbad, CA, USA), which was used as a template for the RNA polymerase Enzo (Affymetrix, Santa Clara, CA, USA) to synthesize dUTP–dCTP biotinylated cRNA. The labelled cRNA was hybridized to Affymetrix oligonucleotide arrays (HG-U133 Set) under conditions specified by the manufacturer.

Microarray analyses—HeLa

Total RNA was prepared using the *mir*Vana miRNA Isolation Kit (Ambion) according to the manufacturer's protocol. The integrity and stability of RNA samples were assessed by Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). We used the Illumina TotalPrep RNA amplification Kit (Ambion) to amplify RNA for hybridization on Illumina BeadChips. Total RNA from each sample was used to synthesize first strand cDNA by reverse transcription. After the second strand cDNA synthesis and cDNA purification steps, the *in vitro* transcription to synthesize cRNA was prepared overnight for 12 h. The gene expression profiles were measured using Illumina HumanHT-12 v3 Expression BeadChip (Illumina, San Diego, CA, USA).

Identification of periodically expressed genes—statistical analysis

Normalization across arrays was performed by robust multi-array average (RMA) from the Bioconductor package affy (10). As there are no probes specific for splice variants in the arrays used, for probes that hybridized with mRNAs corresponding to the same gene, only the probe having the highest variation was used for further analyses. The normalized data were then used to identify genes that showed cyclic behaviour. Genes that matched a cell cycle profile were identified using partial least squares (PLS) regression. Specifically, the gene expression matrix was regressed on to a sine and cosine function with periods equal to the estimated duration of the cell cycle. False discovery rates were computed using a modified Hotelling's T-square statistic and resampling. Resampling ($n = 1000$) was done by assigning phase angles randomly to the time-steps, thus preserving the correlation within replicates of the same time-steps.

Assigning cell cycle phases

Assignment of cell cycle phases used the following protocol. First, each gene in the PLS model was assigned a phase angle by computing the arctangent of the gene's first and second principle component. Second, we used a list of genes previously published to be predominantly expressed in G_1/S, S, G_2, G_2/M and M/G_1 phase [(2): Table 2 and the four M/G_1 associated genes from Figure 2] and identified which of these genes were significant ($q < 0.05$) in our PLS model (Supplementary Figure S3A). Third, for each set of significant G_1/S, S, G_2, G_2/M and M/G_1 phase genes, we computed the set's average phase angle and the corresponding standard deviation and used these values as parameters in a normal distribution to construct a phase angle probability density function for each of the five cell cycle phases (Supplementary Figure S3A). Fourth, we used these probability density functions to assign the most probable phase IDs to each of the significant ($q < 0.05$) genes in the PLS model. Fifth, if necessary, the standard deviation parameters in the phase angle probability density functions were

increased to ensure that the phase assignments had the same order as the average phase angles for the five groups. Specifically, if any out of order phase assignments were detected, the lowest of the five standard deviations was increased by 25%, and the phase assignments were redone. This process was repeated until all phase assignments were in the order defined by the averages.

Promoter CpG analyses

Gene annotations were downloaded from the refGene table in the University of California, Santa Cruz (UCSC) Genome Browser database (human genome assembly version hg18) (11). Promoter regions were defined as the 1000 nt upstream of annotated transcription start sites. Overlapping regions corresponding to genes with multiple transcripts having identical or alternative start sites were merged into one region to create a non-redundant set of promoter regions. Based on this non-redundant set, CpG ratios and CpG promoter classes were calculated and assigned as previously described (12).

Promoter histone analyses

Processed peak data from chromatin immunoprecipitation sequencing (ChIP-seq) experiments of H3K4me3 and H3K27me3 in seven cell lines (GM12878, HUVEC, K562, NHEK, H1-hESC, HMEC and NHLF) were downloaded from UCSC (hg18) (13,14). Non-redundant promoter regions that contained overlapping H3K4me3 and H3K27me3 peaks in all seven cell lines were classified as H3K4me3 and H3K27me3 enriched, respectively; promoter regions that contained both marks in all seven cell lines were classified as bivalent; and promoter regions that did not contain any of the marks or only contained the marks in some of the cell lines were classified as having no or inconsistent marks.

Processed peak data from ChIP-seq experiments of H3K4me3 and H3K27me3 in HeLa and normal adult dermal fibroblasts (NHDF-Ad) were downloaded from UCSC (hg19; Broad Histone track) and remapped to hg18 by the UCSC liftOver tool.

HaCaT ChIP-seq

ChIP, library preparation and sequencing were done as described in the Supplementary Methods. The resulting input chromatin and H3K4me3 and H3K27me3 IP libraries contained ~64 million (M), ~42 M and ~52 M sequence reads, respectively. We used cutadapt (15) to trim adapter sequences, bowtie (16) to align sequence reads from the human genome (version hg18) and discarded all reads that aligned to more than one genomic location. The ~50 M, ~32 M and ~39 M aligned reads for the input, H3K4me3 and H3K27me3 libraries, respectively, were processed by SICER (17) to identify genomic regions that were significantly enriched for H3K4me3 and H3K27me3 in HaCaT (SICER used the input chromatin library as control to identify significant regions; see Supplementary Dataset S3 and S4).

Gene ontology and Kyoto Encyclopedia of Genes and Genomes analyses

Gene ontology (GO) term and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment was assessed by GOstats (18). For GO terms, the test was conditioned on the structure of the GO graph.

Data access

The microarray data from this publication have been submitted to the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/) and were assigned the identifiers E-MTAB-454 (HaCaT) and E-TABM-1152 (HeLa). HaCaT ChIP-seq data are available from http://tang.medisin.ntnu.no/~palsat/HaCaT-ChIP.tgz; bed-files with significant H3K4me3 and H3K27me3 regions are available as Supplementary Dataset S3 and S4.

RESULTS

Partial least squares regression identifies cell cycle-regulated transcripts

To overcome the problem of low synchronization of primary cells and to investigate potential cell type-related differences in cycling genes, we turned to the human keratinocyte cell line HaCaT—a non-tumorigenic, spontaneously immortalized cell line that exhibits apparently normal differentiation (19). Using a double thymidine block to arrest and, subsequently, release HaCaT cells at the G_1/S boundary resulted in ~90% of the cells re-entering the cell cycle and continuing as a uniform cohort through the cell cycle (Supplementary Figure S1). Based on the DNA content profile, we estimated the cell cycle duration to be ~20 h. Although the cells gradually lost synchrony, such that 70% instead of 90% of the cells were synchronously entering the second S phase, the results were reproducible across three independent experiments. We, therefore, considered synchronized HaCaT cells a good starting point for gene expression analyses.

We were primarily interested in genes that showed consistent cyclic expression with a period equal to the estimated cell cycle duration of 20 h. We, therefore, used PLS (20) to identify genes that followed a sine or cosine pattern with a period of 20 h. Conceptually, this PLS analysis projects the time profile for all genes onto a plane described by the sine and cosine of the cell cycle. As a result, genes that are periodically expressed with high amplitude will be far from the origin, whereas genes that vary little or randomly will be close to the origin. Moreover, the phase angle of the cyclic variation will determine the direction from the origin such that genes that have the same temporal expression patterns will lie in the same direction from the origin on the plane.

By profiling RNA isolated from the synchronized cells at 3 h intervals for 33 h, and combining PLS analyses with resampling to compute false discovery rates (21), we identified 1249 Entrez genes that had significant periodic expression patterns during the HaCaT cell cycle (Figure 1, Supplementary Dataset S1). The synchronization experiments' trajectory through the PLS cell cycle plane
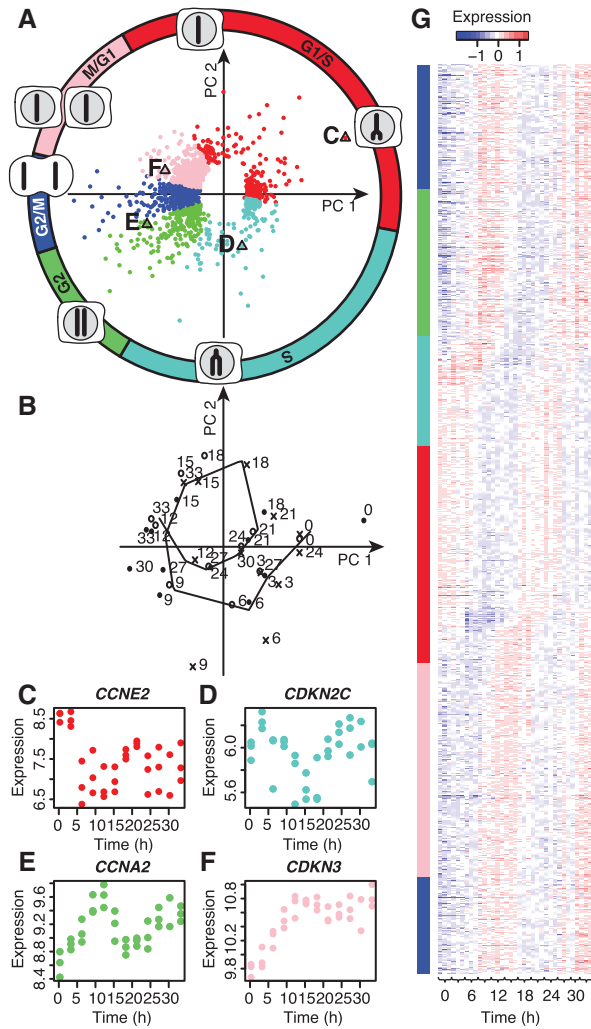
Figure 1. PLS regression identifies cyclic gene expression patterns. The loadings (**A**) and scores (**B**) plots show the contribution of each gene (loadings) to the location of each sample in the subspace of the PLS model (20) (scores; here, sine and cosine, period of 20 h). (A) Points show significant genes [*q* < 0.05; permutation test (21)] and their contribution to the PLS cell cycle model's first and second components (PC1 and PC2). Colours show the cell cycle phase where the gene is predominantly expressed (see 'Materials and Methods' section). The circle and cell diagrams indicate early and late DNA replication, G₂, mitosis and G₁ as determined by the scores plot (B) and flow cytometry (Supplementary Figure 1). (B) Points show the location of the samples from the three replicates (point, circle and cross) within the PLS model's first and second components; the line shows the replicates' average trajectory; numbers show the sample time (h). (**C–F**) Expression profiles for selected genes from the loadings plot [labelled triangles in (A)]. (C) Cyclin E2 (*CCNE2*); (D) Cyclin-dependent kinase inhibitor 2C (*CDKN2C*); (E) Cyclin A2 (*CCNA2*); and (F) Cyclin-dependent kinase inhibitor 3 (*CDKN3*). (**G**) Heatmap showing the expression changes of the cell cycle genes relative to their median expression. The profiles are ordered according the genes' phase angle in the PLS model in (A). Margin colours show the genes' assigned cell cycle phase.

confirmed that the three replicates gave similar and reproducible results (Figure 1B). The trajectory also indicated the cells' gradual loss of synchrony, as the samples were gradually approaching the origin—especially at later time points.

To further analyse the cell cycle genes, we used existing annotations (2) to subdivide the genes into five main groups that represented phases G₁/S, S, G₂, G₂/M and M/G₁ of the cell cycle (Figure 1; see 'Materials and Methods' section). The assigned phases showed good correspondence with previously published phases (Supplementary Table S1); the genes missing from our list had expression patterns that were either inconsistent between replicates, not cyclic, or consistently expressed in the first cell cycle only (Supplementary Figures S2–S4; Supplementary Discussion S1). We, therefore, concluded that our PLS model could robustly identify genes that showed consistent and reproducible cyclic expression patterns within the synchronized HaCaT cells.

HaCaT cell cycle genes include both generally active housekeeping and keratinocyte-specific genes

Having established a set of genes with cell cycle-dependent expression, we set out to characterize these genes further. As cell cycle control is a central housekeeping function, we expected our set of genes to be enriched for high-CpG (HCG) promoters—a hallmark of developmentally regulated genes and housekeeping genes (12,22). We also expected the cell cycle genes to be enriched for histone H3 lysine 4 tri-methylation (H3K4me3) within their promoters, as this epigenetic mark is associated with actively transcribed genes (23,24). Indeed, although the complete set of genes followed a bimodal distribution of HCG and low-CpG (LCG) promoters (12,22), the distribution for the cell cycle genes was shifted, such that a larger percentage had HCG promoters (Figure 2A and B; *P* = 2e-11, binomial test). Similarly, ChIP-seq of H3K4me3 and H3K27me3 marks in the HaCaT cells (see 'Materials Methods' section; Supplementary Dataset S2 and S3) showed that the cell cycle genes described here were enriched for H3K4me3 (Figure 2C). Moreover, the genes were also enriched for H3K4me3 marks that were present in and common for all of seven other cell lines (13,14) (Figure 2D), which is consistent with these genes having housekeeping functions common to many cell types.

The cell cycle genes that showed inconsistent or no histone modifications within their promoters were highly enriched among LCG genes (Figure 2E and F). Tissue-specific genes are commonly enriched among LCGs, which suggested that this subset of cell cycle expressed genes that have LCG promoters with no or inconsistent H3K4me3 modifications might be cell type-specific for HaCaT. To test this possibility, we first investigated whether all the human LCG genes without consistent H3K4me3 within the seven cell lines (LCG/NoH3K4me3) were enriched within GO categories or KEGG pathways (18). As expected, the set of all LCG/NoH3K4me3 genes was highly enriched for tissue specific functions, whereas the set of all HCG/H3K4me3 genes was enriched for housekeeping categories and pathways (Supplementary Figure S5, Supplementary Dataset S4). Second, by accounting for such background enrichment within the nine subgroups, we found both the HCG/H3K4me3 cell cycle genes and the intermediate CpG (ICG)/H3K4me3 cell cycle genes to be specifically enriched for cell cycle-related terms (Figure 2G). Moreover, the LCG/NoH3K4me3 cell cycle genes were
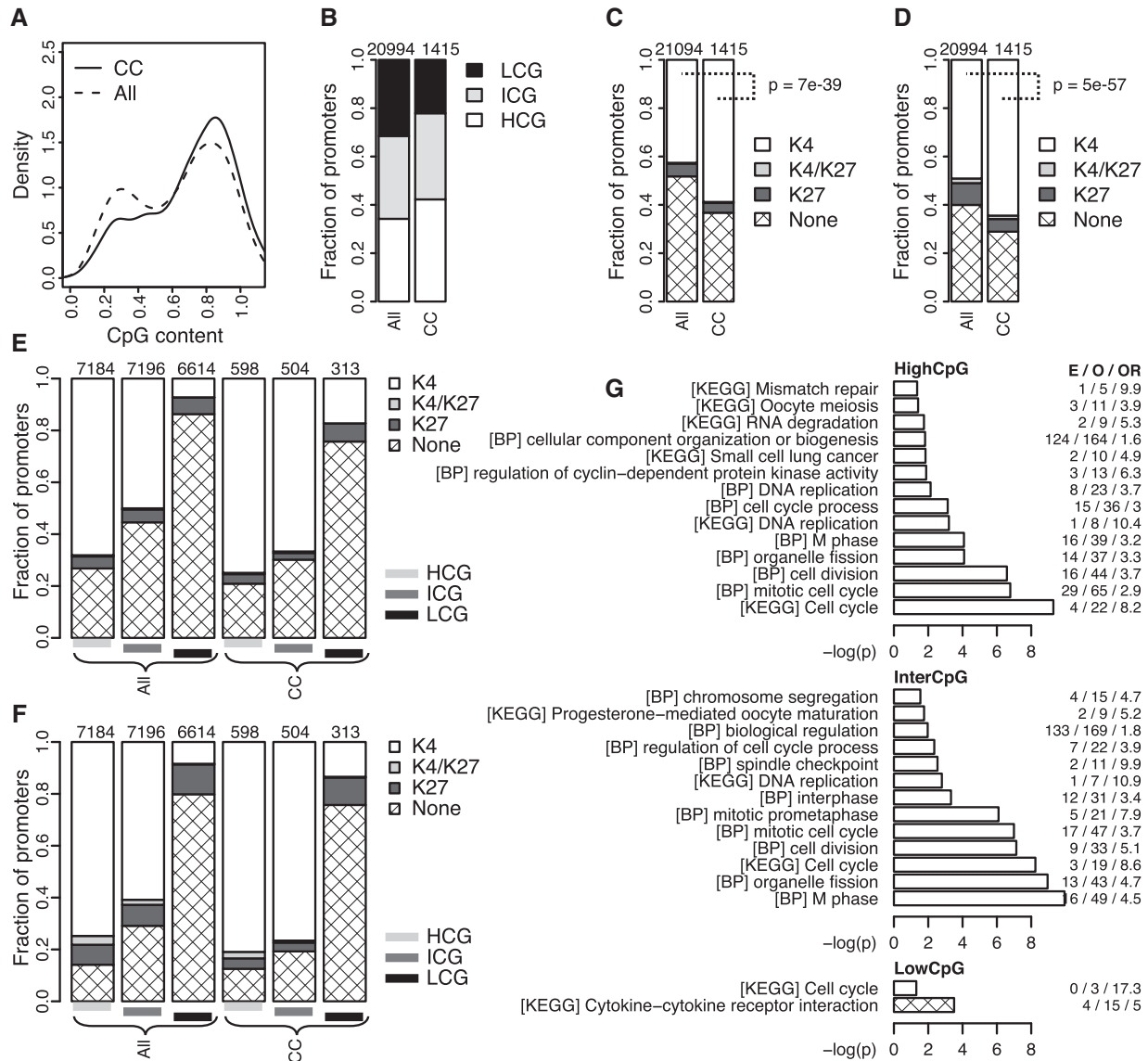
Figure 2. HaCaT cell cycle genes include both generally active housekeeping and keratinocyte-specific genes. (**A**) Distributions of CpG content in promoters for all genes included in the PLS model (all, dashed line) and cell cycle expressed genes (CC, solid line). The lines are the kernel density estimates of the two CpG content distributions. (**B**) Fraction of LCG (black), ICG (grey) and HCG (white) promoters for the two groups in (A). Numbers above bars are the total number of non-redundant promoter regions. (**C and D**) Fraction of promoters that in HaCaT (C) or in seven different cell lines (D) contain exclusive H3K4me3 or H3K27me3 marks (K4, white; K27, dark grey), bivalent K4 and K27 marks (K4/K27, light grey), or no or inconsistent marks (None, diamonds). *P*-values are the results of binomial tests comparing the fraction of H3K4me3-marked genes in the two sets. (**E and F**) Fractions from (C: E) and (D: F) subdivided into LCG, ICG and HCG (black, dark grey and light grey lines) promoters. (**G**) GO biological process (BP) terms or KEGG pathways significantly enriched (*P* < 0.05) among HCG, ICG or LCG cell cycle genes subdivided into promoters with consistent H3K4me3 (white) or H3K27me3 (no significant terms) marks, or no or inconsistent marks (diamonds). Horizontal axes show Benjamini–Hochberg corrected *P*-values (–log$_{10}$); numbers on the right show the expected (E) and observed (O) number of genes and the corresponding odds ratios (OR).

specifically enriched for the KEGG cytokine–cytokine receptor interaction pathway, which is consistent with cytokines being important regulators of wound healing in general and keratinocyte proliferation in particular (25). Thus, at least some of the cell cycle genes seem to be cell type-specific for HaCaT cells.

Distinct cell types have exclusive cell cycle regulated genes related to the cells' specific characteristics and functions

To further determine whether cell type-specific factors could explain differences in reported cycling genes, we

analysed the overlap among the cell cycle-regulated genes identified in our and previous human studies (Figure 3). Although our study identified more cell cycle genes than did the foreskin fibroblast (FF) (3) and HeLa (2) studies (Figure 3A), the cell cycle genes we identified had clear cycling patterns in HaCaT. In contrast, the genes identified in the two previous studies but missed in our study had weak or inconsistent patterns in HaCaT; compare Figure 3B and 3C.

Some of the differences between the studies were likely related to differences in microarray technologies and study
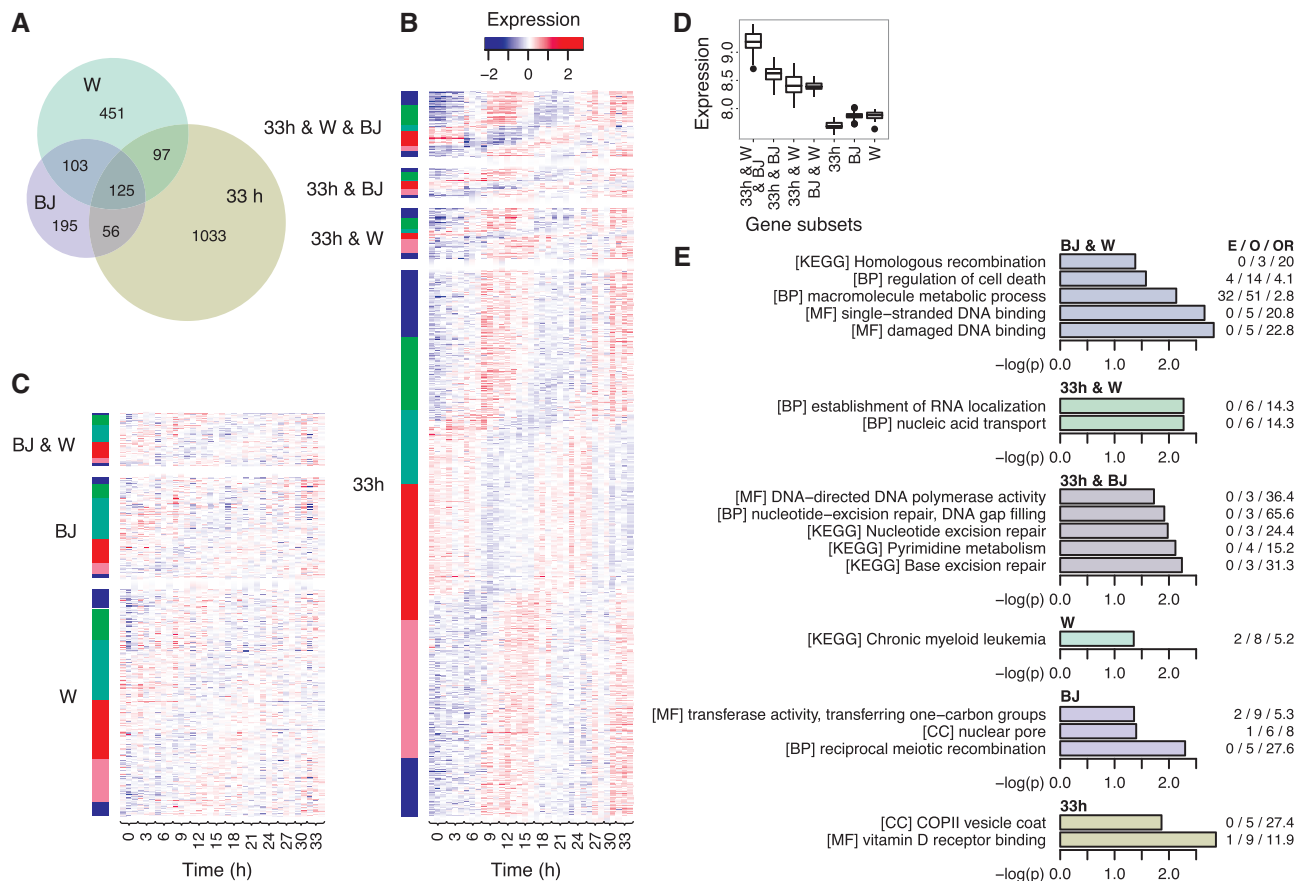
Figure 3. Cell cycle genes that are common or exclusive for foreskin fibroblasts (FF) (3), HeLa (2) and HaCaT reveal characteristics that are specific for the three cell types. (**A**) Venn diagram showing the overlap in cell cycle genes between FF (BJ), HeLa (W) and HaCaT (33 h). (**B**) Heatmap showing the expression changes of the HaCaT cell cycle genes relative to their median expression. The genes are grouped according to their overlap with the cell cycle genes from the FF and HeLa studies and are ordered according to the genes' phase angle (from $-\pi$ to π) in the HaCaT 33 h PLS model. Margin colours show the genes' assigned cell cycle phase (red, G_1/S; turquoise, S; green, G_2; blue, G_2/M; and pink, M/G_1). (**C**) Heatmap showing the expression changes in HaCaT for genes identified as cell cycle regulated in FF or HeLa but not in HaCaT. The genes are grouped according to their overlap between the two studies and are ordered according to the genes' phase angle in the HaCaT 33 h PLS model. (**D**) Median expression level in HaCaT of the genes common and exclusive for the three studies. The genes are grouped according to their overlap between the three studies. Boxes, horizontal black line, and circles show the first and third quartiles, the median and outliers; whiskers show the most extreme data points up to 1.5 times the interquartile range from the box. (**E**) KEGG pathways and GO molecular function (MF), biological process (BP) and cellular component (CC) terms that are overrepresented ($P < 0.05$) and unique for six of the seven subsets of cell cycle genes common or exclusive for FF, HeLa and HaCaT; see Supplementary Table S2 for pathways and terms overrepresented among cell cycle genes common for FF, HeLa and HaCaT.

designs. Microarray results are generally reproducible across platforms and laboratories for genes with medium and high expression (26), and the genes common for the three studies were generally highly expressed in HaCaT (Figure 3D). These genes were also strongly enriched for cell cycle functions (Supplementary Table S2), indicating that these common genes represent a core set of cell cycle genes. As for design, although previous studies have either not included or used poorly matched biological replicates, we included three direct biological replicates. This, combined with the reproducibly high level of synchronization for the HaCaT cells, gave our study better statistical power to identify cyclically expressed genes.

Although the cell cycle genes common for all three studies represented a core set of human cell cycle genes, the other subsets of cell cycle genes were at least partially related to cell type-specific functions and characteristics (Figure 3E). To illustrate, the GO term 'vitamin D

receptor binding' was overrepresented among the cell cycle genes exclusive for HaCaT, which is consistent with vitamin D and the vitamin D receptor's important functions in keratinocytes (27). In contrast, and consistent with HeLa's cancer origin, 'chronic myeloid leukaemia' was the only term exclusive for the HeLa cell cycle genes. Further supporting this conclusion, the KEGG pathways 'nucleotide excision repair' and 'base excision repair' were exclusively overrepresented among the cell cycle genes common for FF and HaCaT, which suggests that these two DNA repair pathways are partially dysregulated in HeLa. Although we cannot exclude that technical differences underlie some of the differences in gene expression (see 'Discussion' section), distinct cell types do seem to have distinct cell cycle regulated genes. For some cells, such as HeLa and likely other cancers (3), these distinct cell cycle genes may reflect that the cells are in an abnormal dysregulated state. Nevertheless, for more

normal non-tumorigenic cells, such as HaCaT, at least some of the distinct cell cycle genes are related to the cells' specific functions.

Cell cycle genes marked with H3K27me3 are transcribed during DNA replication

Although the cell cycle genes were generally enriched for H3K4me3, a fraction of the cell cycle genes were marked with H3K27me3 (Figure 2B)—the mark for genes silenced by Polycomb group proteins (28,29). Replication opens compacted chromatin, and although H3K27me3 marks are copied to daughter chromatin, the marks may initially be diluted, such that the normally silenced DNA is transiently accessible for transcription during this replication process. If this model was correct, we would expect that cell cycle genes marked with H3K27me3 were enriched among genes upregulated during DNA replication.

Indeed, the H3K27me3 marks in the HaCaT cells were differentially distributed among the cell cycle genes (Figure 4A). Of genes predominantly expressed in the G_1/S and S phases, 37% had H3K4me3 marks and 12% had H3K27me3 marks. In contrast, of genes expressed in G_2, G_2/M and M/G_1, 80% contained H3K4me3, whereas only 1% contained H3K27me3 marks. In total, 87% of the H3K27me3 marked cell cycle genes were expressed in G_1/S or S ($P = 5e-17$; one-tailed binomial test). Similarly, a larger percentage of the G_1/S and S phase genes was LCG (Figure 4B). The enrichment of H3K27me3 was not related to CpG content, as the G_1/S and S phase genes showed similar enrichment for H3K27me3 independent of CpG content (Figure 4C). These patterns were also consistent within the previously published data from seven different cell lines (13,14) (Supplementary Figure S6).

Genes upregulated in early S-phase tend to be replicated earlier than late S-phase genes

Replication occurs in distinct domains largely characterized by differences in GC content, and although these domains are partly reorganized during development, domains with high and low GC content are generally replicated early and late in S phase, respectively (30–35). If DNA replication facilitated transient transcription of silenced genes, we would, therefore, expect the genes upregulated during G_1/S to generally have a higher GC content than the S phase and other cell cycle genes. Indeed, the G_1/S genes had the highest overall GC content of the cell cycle genes (Figure 4D; P-value of 7e-3, two-sided un-paired Student's t-test with unequal variance of difference in mean between G_1/S and S phase genes) and replication-timing data from three different cell lines (30,33) indicated that the G_1/S genes on average tended to be replicated earlier than the S phase genes (Figure 4E, Student's t-test P-value of 1e-3; Supplementary Figure S7). Importantly, 63% of the cell cycle regulated genes reside in genomic regions previously shown to have stable replication timing across multiple cell types (31), and these genes had the same pattern in replication timing as the complete set of cell cycle genes (Supplementary Figure S8).

Consistent with some of the genes residing in silenced chromatin and being transiently expressed during replication, the G_1/S and S genes had a general lower expression level than the genes expressed in the other phases (Figure 4F). Moreover, the genes marked with H3K27me3 in HaCaT had the lowest expression level of the G_1/S and S phase genes (Figure 4G). The differences in expression between H3K27me3-marked genes and other subgroups were significant for all G_1/S phase groups (P-values of 2e-16, 2e-9, 3e-5 and 0.01 for K4, HCG, ICG and LCG groups, respectively; two-sided un-paired Student's t-test with unequal variance) and for the K4 and HCG S phase groups (P-values of 7e-4, 0.02, respectively; P-values of 0.1 and 0.3 for ICG and LCG groups, respectively).

Foreskin fibroblasts show similar replication-related expression of H3K27me3 marked genes

To investigate whether this replication-related cell cycle expression pattern could be found in other cell types, we first used our computational framework to reanalyse the previously published primary foreskin fibroblast data (3) (Supplementary Figure S9). As in the HaCaT cells, the majority (64%) of the genes that had significant cell cycle expression and were marked with H3K27me3 in fibroblasts were upregulated during DNA replication (Figure 5A); the mark was especially enriched for the G_1/S genes ($P = 0.009$; one-tailed binomial test for enrichment). In contrast to HaCaT, the majority of the fibroblast cell cycle genes marked with H3K27me3 was also marked with H3K4me3. The fibroblast cell cycle genes also showed a similar trend in GC content as the HaCaT genes, although the difference between the genes expressed early (G_1/S) and late (S) during DNA replication was not significant (Figure 5B; $P = 0.1$, Student's t-test).

Second, we measured the gene expression profiles in HeLa cells synchronized in the G_2/M transition by nocodazole and mitotic shake-off and used PLS to identify genes with cell cycle-dependent expression (Supplementary Figures S10 and S11). Again, the majority (52%) of the cell cycle genes that were marked with H3K27me3 in HeLa was upregulated during DNA replication, but the differences between the phases were not significant (Figure 5C; $P = 0.2$). Moreover, the cell cycle genes again showed a similar trend in GC content as the HaCaT, such that the G_1/S genes had a higher GC content than the S genes (Figure 5D; $P = 0.01$, Student's t-test). Thus, replication-related expression of H3K27me3-marked genes does not seem to be a strong characteristic of all cell types. It is relevant to note, however, that the characteristic is absent in the most abnormal of the three cells, whereas HaCaT and the primary fibroblasts both have significant replication-related expression of H3K27me3-marked genes.

DISCUSSION

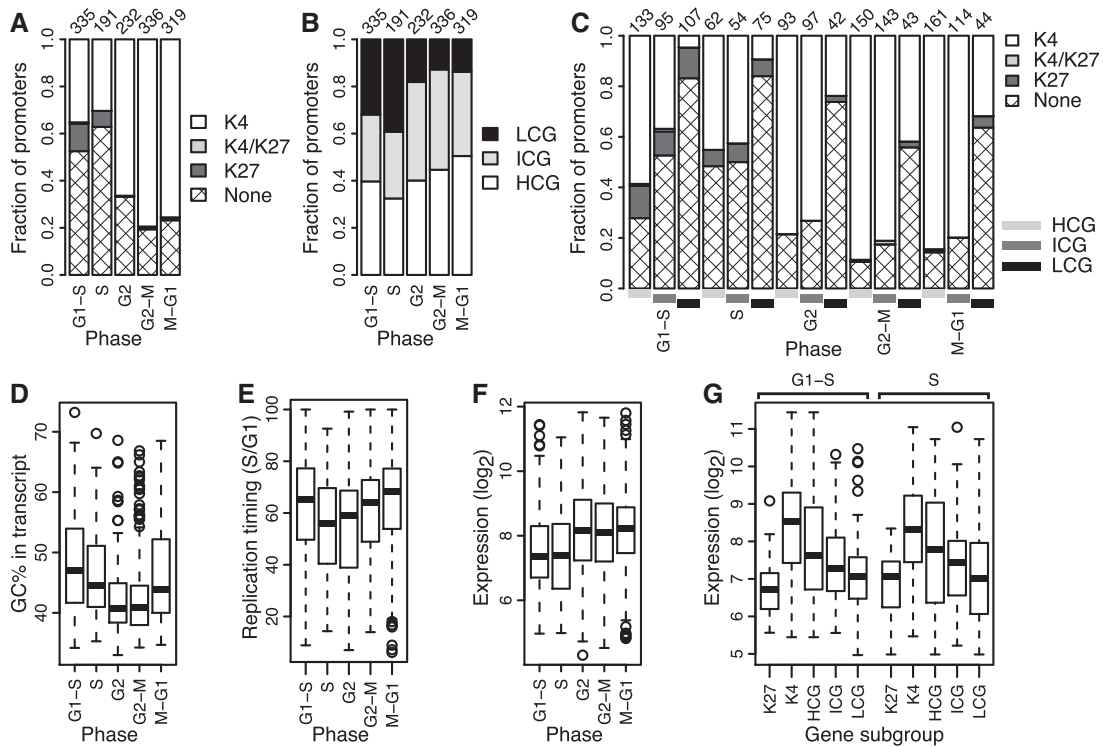Previous studies of the gene expression patterns during the human cell division cycle have indicated marked

Figure 4. G_1/S and S phase genes are enriched for H3K27me3 marks and show replication-related expression. (**A**) Distribution of HaCaT H3K4me3 and H3K27me3 marks in promoters of genes upregulated in G_1/S, S, G_2, G_2/M and M/G_1 phases; see Figure 2. (**B**) Fractions of LCG, ICG and HCG promoters for genes upregulated in the five cell cycle phases. (**C**) Distribution of histone marks from (A) subdivided by promoter CpG content. (**D–F**) GC% (D), replication timing (E) and expression level (F) of genes expressed in the five cell cycle phases. GC% is the GC% in the region defined by the gene's annotated transcription start and end sites; replication timing is the median S/G_1 ratio within embryonic stem cells (33) of regions overlapping each gene, such that values close to 100 and 0 represent early and late replication, respectively; expression is the gene's average microarray expression. Boxes, horizontal black line, and circles show the first and third quartiles, the median and outliers; whiskers show the most extreme data points up to 1.5 times the interquartile range from the box. (**G**) Expression levels of specific subgroups of G_1/S and S phase genes.

differences between primary and cancer cells (3), but our results show that distinct cell types have distinct cell cycle expressed genes that are related to the cells' functions. Specifically, by profiling the gene expression in synchronized human HaCaT keratinocytes, we have identified a set of cell cycle expressed genes containing known marks for cell type-specific expression, such as low CpG content in their promoters and few or inconsistent promoter histone marks. These genes were enriched for functions important for keratinocyte proliferation, such as cytokine–cytokine receptor interaction.

Cell cycle regulation is an important housekeeping function and accordingly, most of the cell cycle expressed genes identified in our study have high CpG promoters and the active chromatin mark H3K4me3 in multiple cell lines. Nevertheless, we have also shown that some genes with low CpG promoters or the silent chromatin mark H3K27me3 have cell cycle dependent expression. These genes are mostly expressed during S phase in a replication-related pattern, which suggests that DNA replication can open some epigenetically silenced loci for transcription.

Replication-dependent transcription is a mechanism that can achieve precise transcriptional regulation during the cell cycle, as normally silenced genes will only be transcribed when their loci are replicated in S-phase.

Compared with the other cell cycle genes, the Polycomb-regulated cell cycle genes in HaCaT were significantly enriched for the GO terms 'GO:0007267 cell–cell signalling' and 'GO:0046903 secretion' (Benjamini–Hochberg-corrected *P*-values of 0.007 and 0.02, respectively). In HaCaT cells, replication-dependent transcription may, therefore, be a regulatory mechanism for coordinating cell–cell signalling with the cell cycle.

Lanzuolo *et al.* (36) have earlier reported that two Polycomb-repressed homeotic genes show replication-dependent expression in *Drosophila* S2 cells. Our results extend this observation to humans and show that many other genes with Polycomb-repressive marks have replication-dependent expression. We also note that cis-encoded RNAs do play a role in Polycomb-mediated targeting and regulation (37–39). Given that Polycomb regulates all transcripts in targeted loci, replication-dependent expression may be a mechanism for expressing such cis-encoded Polycomb-associated RNAs at existing Polycomb-regulated loci.

When comparing our list of HaCaT cell cycle genes with the cell cycle genes from the previous HeLa (2) and primary foreskin fibroblast studies (3), we found several differences, such that only 125 genes were common for all three studies (Figure 3A). Several technical factors, such as differences in microarray technology, analytical
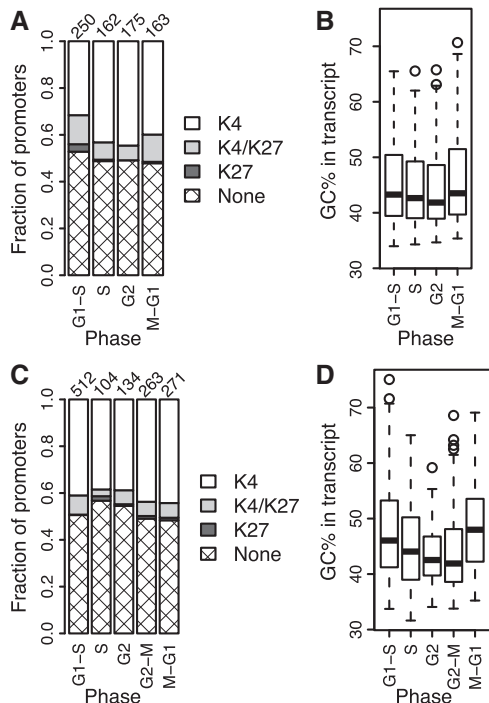
Figure 5. H3K27me3 marks are enriched in genes upregulated during DNA replication in primary foreskin fibroblasts. (**A**) Distribution of H3K4me3 and H3K27me3 marks in promoters of genes upregulated in G_1/S, S, G_2 and M/G_1 phases in primary foreskin fibroblasts synchronized by serum starvation [expression data from (3); ChIP-seq data from normal adult dermal fibroblasts (13,14)]. (**B**) GC% of the genes from (A). (**C**) Distribution of H3K4me3 and H3K27me3 marks in promoters of genes upregulated in G_1/S, S, G_2, G_2/M and M/G_1 phases in HeLa cells synchronized by nocodazole and mitotic shake-off [ChIP-seq data from HeLa (13,14)]. (**D**) GC% of the genes from (C). See Figures 2 and 4 for details on graphs.

approach and experimental design, likely contributed to these differences, for example, the genes that were common for all three studies were highly expressed, whereas the genes that were exclusive for our study were lowly expressed in HaCaT. We could detect these HaCaT-specific genes because our three matched biological replicates gave increased statistical power to identify weak, but consistent, cyclic expression patterns.

In addition to these technical differences, our results also show that underlying biological differences between the cells can explain some of the differences in cell cycle expressed genes. Specifically, our results show that in addition to the core genes that have cell cycle-dependent expression in multiple cell types, some genes related to the cell type's specific functions and characteristics also have cyclic expression. This is in contrast to a previous report, which indicated that such cell-type specific expression mostly reflects differences between primary and cancer cells (3). These cancer-related differences are evident in our analyses as well, but we also find that genes and pathways important to keratinocyte function, such as cytokine–cytokine receptor interaction (25) and vitamin D receptor binding, (27) have cell cycle-dependent expression. Both functions are important for normal

keratinocyte proliferation, which makes it unlikely that these results are an artifact of HaCaT's transformed state or technical differences between the studies. Instead, the cell cycle-dependent expression of these genes likely reflects that processing of proliferative signals from the extracellular environment should be coordinated with the cell cycle.

We should note that as we have measured expression changes in total cellular RNA, some of the expression changes we have detected could potentially be explained by cell cycle-dependent changes in RNA processing or RNA stability instead of RNA transcription. Moreover, our approach of chemically synchronizing cells can likely explain some of the changes in gene expression in our data, as the synchronization procedure likely triggers stress-related responses that could both give false-positive cyclic profiles and mask the cell-cycle-dependent expression of some true cell cycle genes. This masking effect could potentially explain why we did not detect the known cell-cycle genes CDKN1A and VEGFC to have cell cycle-dependent expression in HaCaT (Supplementary Figure S2C and D). Our analyses, including the good overlap between the genes we identified and known cell-cycle-expressed genes, strong enrichment for cell-cycle-related functions and overlap with previous studies, do suggest, however, that the majority of the genes we identified are true cell cycle genes.

In summary, by combining robust statistics with the power of matched biological replicates, we have identified a set of genes that show consistent cell cycle-dependent expression in HaCaT cells. These genes include both general cell cycle genes and genes specific for HaCaT function. Some of the cell cycle genes also have H3K27me3 marks—a histone modification commonly associated with Polycomb-silenced genes. We show that in both HaCaT and primary foreskin fibroblasts, these genes are upregulated during DNA replication, which is consistent with a mechanism where DNA replication can open some Polycomb-repressed loci for transcription.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2, Supplementary Figures 1–11, Supplementary Methods, Supplementary Datasets 1–4 and Supplementary References [2,3,18,30,31,33,40].

REFERENCES

1. Cho,R.J., Huang,M., Campbell,M.J., Dong,H., Steinmetz,L., Sapinoso,L., Hampton,G., Elledge,S.J., Davis,R.W. and Lockhart,D.J. (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet.*, **27**, 48–54.
2. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell.*, **13**, 1977–2000.
3. Bar-Joseph,Z., Siegfried,Z., Brandeis,M., Brors,B., Lu,Y., Eils,R., Dynlacht,B.D. and Simon,I. (2008) Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl Acad. Sci. USA*, **105**, 955–960.
4. Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C., Trent,J.M., Staudt,L.M., Hudson,J. Jr, Boguski,M.S. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.
5. Beyrouthy,M.J., Alexander,K.E., Baldwin,A., Whitfield,M.L., Bass,H.W., McGee,D. and Hurt,M.M. (2008) Identification of G1-regulated genes in normally cycling human cells. *PLoS One*, **3**, e3943.
6. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
7. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
8. Laub,M.T., McAdams,H.H., Feldblyum,T., Fraser,C.M. and Shapiro,L. (2000) Global analysis of the genetic network controlling a bacterial cell cycle. *Science*, **290**, 2144–2148.
9. Rustici,G., Mata,J., Kivinen,K., Lio,P., Penkett,C.J., Burns,G., Hayles,J., Brazma,A., Nurse,P. and Bahler,J. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.*, **36**, 809–817.
10. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
11. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
12. Weber,M., Hellmann,I., Stadler,M.B., Ramos,L., Paabo,S., Rebhan,M. and Schubeler,D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457–466.
13. Rosenbloom,K.R., Dreszer,T.R., Pheasant,M., Barber,G.P., Meyer,L.R., Pohl,A., Raney,B.J., Wang,T., Hinrichs,A.S., Zweig,A.S. *et al.* (2010) ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.*, **38**, D620–D625.
14. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
15. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.*, **17**, 10–12.
16. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
17. Zang,C., Schones,D.E., Zeng,C., Cui,K., Zhao,K. and Peng,W. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.
18. Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.
19. Boukamp,P., Petrussevska,R.T., Breitkreutz,D., Hornung,J., Markham,A. and Fusenig,N.E. (1988) Normal keratinization in a spontaneously immortalized aneuploid human keratinocyte cell line. *J. Cell Biol.*, **106**, 761–771.
20. Johansson,D., Lindgren,P. and Berglund,A. (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, **19**, 467–473.
21. Gidskehaug,L., Anderssen,E., Flatberg,A. and Alsberg,B.K. (2007) A framework for significance analysis of gene expression data using dimension reduction methods. *BMC Bioinformatics*, **8**, 346.
22. Saxonov,S., Berg,P. and Brutlag,D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA*, **103**, 1412–1417.
23. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
24. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
25. Werner,S. and Grose,R. (2003) Regulation of wound healing by growth factors and cytokines. *Physiol. Rev.*, **83**, 835–870.
26. Kuo,W.P., Liu,F., Trimarchi,J., Punzo,C., Lombardi,M., Sarang,J., Whipple,M.E., Maysuria,M., Serikawa,K., Lee,S.Y. *et al.* (2006) A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat. Biotechnol.*, **24**, 832–840.
27. Cianferotti,L., Cox,M., Skorija,K. and Demay,M.B. (2007) Vitamin D receptor is essential for normal keratinocyte stem cell function. *Proc. Natl Acad. Sci. USA*, **104**, 9428–9433.
28. Beisel,C. and Paro,R. (2011) Silencing chromatin: comparing modes and mechanisms. *Nat. Rev. Genet.*, **12**, 123–135.
29. Margueron,R. and Reinberg,D. (2011) The Polycomb complex PRC2 and its mark in life. *Nature*, **469**, 343–349.
30. Woodfine,K., Fiegler,H., Beare,D.M., Collins,J.E., McCann,O.T., Young,B.D., Debernardi,S., Mott,R., Dunham,I. and Carter,N.P. (2004) Replication timing of the human genome. *Hum. Mol. Genet.*, **13**, 191–202.
31. Hansen,R.S., Thomas,S., Sandstrom,R., Canfield,T.K., Thurman,R.E., Weaver,M., Dorschner,M.O., Gartler,S.M. and Stamatoyannopoulos,J.A. (2010) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA*, **107**, 139–144.
32. Farkash-Amar,S., Lipson,D., Polten,A., Goren,A., Helmstetter,C., Yakhini,Z. and Simon,I. (2008) Global organization of replication time zones of the mouse genome. *Genome Res.*, **18**, 1562–1570.
33. Desprat,R., Thierry-Mieg,D., Lailler,N., Lajugie,J., Schildkraut,C., Thierry-Mieg,J. and Bouhassira,E.E. (2009) Predictable dynamic program of timing of DNA replication in human cells. *Genome Res.*, **19**, 2288–2299.
34. Costantini,M. and Bernardi,G. (2008) Replication timing, chromosomal bands, and isochores. *Proc. Natl Acad. Sci. USA*, **105**, 3433–3437.

35. Hiratani,I., Ryba,T., Itoh,M., Yokochi,T., Schwaiger,M., Chang,C.W., Lyou,Y., Townes,T.M., Schubeler,D. and Gilbert,D.M. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.*, **6**, e245.

36. Lanzuolo,C., Lo Sardo,F., Diamantini,A. and Orlando,V. (2011) PcG complexes set the stage for epigenetic inheritance of gene silencing in early S phase before replication. *PLoS Genet.*, **7**, e1002370.

37. Plath,K., Fang,J., Mlynarczyk-Evans,S.K., Cao,R., Worringer,K.A., Wang,H., de la Cruz,C.C., Otte,A.P., Panning,B. and Zhang,Y. (2003) Role of histone H3 lysine 27 methylation in X inactivation. *Science*, **300**, 131–135.

38. Zhao,J., Sun,B.K., Erwin,J.A., Song,J.J. and Lee,J.T. (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, **322**, 750–756.

39. Pandey,R.R., Mondal,T., Mohammad,F., Enroth,S., Redrup,L., Komorowski,J., Nagano,T., Mancini-Dinardo,D. and Kanduri,C. (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell*, **32**, 232–246.

40. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.