SUPPLEMENTAL INFORMATION FOR

SYNCHRONIZED REPLICATION OF GENES ENCODING THE SAME PROTEIN COMPLEX IN FAST-PROLIFERATING CELLS

Y. Chen, K. Li, X. Chu, L. B. Carey, and W. Qian

Supplemental information includes Supplemental Tables S1-3 and Supplemental Figures S1-7.

SUPPLEMENTAL TABLES

Table S1. Cell-cycle phase-specific genes used in this study

Gene	Phase
ENSG00000119640	G_1/S
ENSG00000169188	G_1/S
ENSG00000156802	G_1/S
ENSG00000105173	G_1/S
ENSG00000094804	G_1/S
ENSG00000164649	G_1/S
ENSG00000167670	G_1/S
ENSG00000092853	G_1/S
ENSG00000159147	G_1/S
ENSG00000143476	G_1/S
ENSG00000101412	G_1/S
ENSG00000131153	G_1/S
ENSG00000112312	G_1/S
ENSG00000119969	G_1/S
ENSG00000116679	G_1/S
ENSG00000073111	G_1/S
ENSG00000104738	G_1/S
ENSG00000100297	G_1/S
ENSG00000132780	G_1/S
ENSG00000132646	G_1/S
ENSG00000145604	G_1/S
ENSG00000163950	G_1/S
ENSG00000163781	G_1/S
ENSG00000276043	G_1/S
ENSG00000092470	G_1/S
ENSG00000165724	G_1/S
ENSG00000105011	S
ENSG00000156802	S
ENSG00000012048	S
ENSG00000175455	S
ENSG00000146670	S
ENSG00000100162	S
ENSG00000228716	S
ENSG00000106462	S
ENSG00000187741	S
ENSG00000168496	S
ENSG00000119969	S

ENSG00000197061	S
ENSG00000006576	S
ENSG00000127564	S
ENSG00000111247	S
ENSG00000101773	S
ENSG00000049541	S
ENSG00000117748	S
ENSG00000167325	S
ENSG00000171848	S
ENSG00000131747	S
ENSG00000176890	S
ENSG00000077152	S
ENSG00000162607	S
ENSG00000122952	S
ENSG00000011426	G_2/M
ENSG00000104442	G_2/M
ENSG00000178999	G_2/M
ENSG00000154473	G_2/M
ENSG00000145386	G_2/M
ENSG00000111665	G_2/M
ENSG00000134690	G_2/M
ENSG00000170312	G_2/M
ENSG00000123080	G_2/M
ENSG00000120334	G_2/M
ENSG00000136108	G_2/M
ENSG00000136485	G_2/M
ENSG00000132002	G_2/M
ENSG00000101447	G_2/M
ENSG00000188486	G_2/M
ENSG00000164104	G_2/M
ENSG00000138182	G_2/M
ENSG00000137807	G_2/M
ENSG00000182481	G_2/M
ENSG00000143815	G_2/M
ENSG00000113368	G_2/M
ENSG00000164109	G_2/M
ENSG00000165304	G_2/M
ENSG00000129534	G_2/M
ENSG00000137804	G_2/M
ENSG00000166965	G_2/M
ENSG00000164105	G_2/M

ENSG00000113810	G_2/M
ENSG00000137203	G_2/M
ENSG00000120802	G_2/M
ENSG00000131747	G_2/M
ENSG00000116830	G_2/M
ENSG00000196230	G_2/M
ENSG00000188229	G_2/M
ENSG00000175063	G_2/M
ENSG00000011426	M
ENSG00000143401	M
ENSG00000170540	M
ENSG00000087586	M
ENSG00000089685	M
ENSG00000145386	M
ENSG00000157456	M
ENSG00000117399	M
ENSG00000101224	M
ENSG00000004897	M
ENSG00000111665	M
ENSG00000115163	M
ENSG00000117724	M
ENSG00000138180	M
ENSG00000136108	M
ENSG00000175216	M
ENSG00000173207	M
ENSG00000123975	M
ENSG00000024526	M
ENSG00000139734	M
ENSG00000114346	M
ENSG00000111206	M
ENSG00000075218	M
ENSG00000072571	M
ENSG00000109971	M
ENSG00000142945	M
ENSG00000170759	M
ENSG00000118193	M
ENSG00000138182	M
ENSG00000143815	M
ENSG00000129534	M
ENSG00000148773	M
ENSG00000204899	M

ENSG00000010292	M
ENSG00000110713	M
ENSG00000137804	M
ENSG00000168078	M
ENSG00000166851	M
ENSG00000068489	M
ENSG00000116560	M
ENSG00000171241	M
ENSG00000013810	M
ENSG00000149483	M
ENSG00000088325	M
ENSG00000071539	M
ENSG00000188229	M
ENSG00000100526	M/G_1
ENSG00000088986	M/G_1
ENSG00000170759	M/G_1
ENSG00000133703	M/G_1
ENSG00000145220	M/G_1
ENSG00000123562	M/G_1
ENSG00000122140	M/G_1
ENSG00000069275	M/G_1
ENSG00000108256	M/G_1
ENSG00000111845	M/G_1
ENSG00000168078	M/G_1
ENSG00000113575	M/G_1
ENSG00000198901	M/G_1
ENSG00000164611	M/G_1
ENSG00000132341	M/G_1
ENSG00000169371	M/G_1
ENSG00000112081	M/G_1
ENSG00000149483	M/G_1
ENSG00000198900	M/G_1

Table S2. Data sources of 19 cell lines/types used in this study[1].

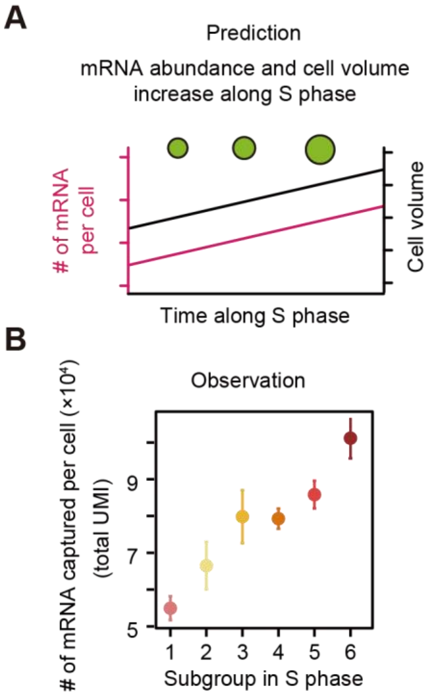
Cell line/type	Replication timing	Gene expression	FACS data
ESC (BG01)	Int61646100	GSE72923	
ESC (BG02)	Int87960943	GSE72923	
ESC (CyT49)	Int83608519	GSE63592	Yes
ESC (H1)	Int79529114	GSE67915	
ESC (H7)	Int97847336	GSE52657	
ESC (H9)	Ext29405702	GSE67915	
Immature hepatocyte	Int54607190	GSE63592	Yes
Hepatoblast	Int79325459	GSE63592	Yes
Primitive hepatocyte	GSE63428	GSE63592	Yes
Primitive gut tube	GSE63428	GSE63592	Yes
Posterior foregut	GSE63428	GSE63592	Yes
Pancreatic endoderm	Int56654336	GSE63592	Yes
Smooth muscle	Int16405711	GSE63592	
Mesothelium	Int51173422	GSE63592	Yes
Adenocarcinoma (MCF−7)	Int67688290	GSE33480	
Colon carcinoma (HCT116)	Int90617792	GSE33480	
Cervical carcinoma (HeLa S3)	Int95117837	GSE33480	
Neuroblastoma (SK-N-SH)	Int67184500	GSE90265	
Hepatocellula carcinoma (HepG2)	Int49277082	GSE87999	

[1] Int* and Ext* are accession numbers in the ReplicationDomain database (www.replicationdomain.org). GSE* are accession numbers in the NCBI GEO (www.ncbi.nlm.nih.gov/geo/).

Table S3. Meta-PCNA genes used in this study.

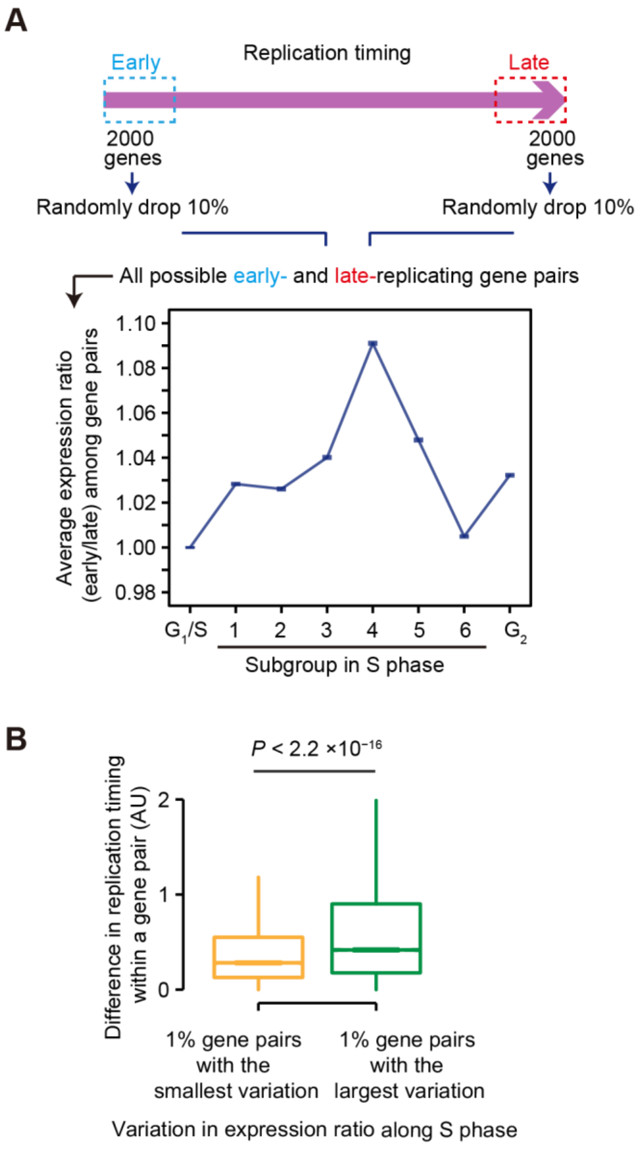
Name
AURKA
BIRC5
CDC20
CDCA4
CKLF
FEN1
GINS1
GTPBP2
LBR
LSM6
MAD2L1
MCM4
MKI67
NUSAP1
PCNA
PSMD9
RFC3
RFC4
RRM2
SMC4
SNF8
SNRPB
TACC3
TCF3
TFDP1
TROAP
TYMS
UBE2C
VRK1
ZWINT

SUPPLEMENTAL FIGURES



Supplemental Figure S1. The mRNA abundance per cell increases along S phase. (A) Both cell volume (black) and RNA content (magenta) were reported to increase through S phase.

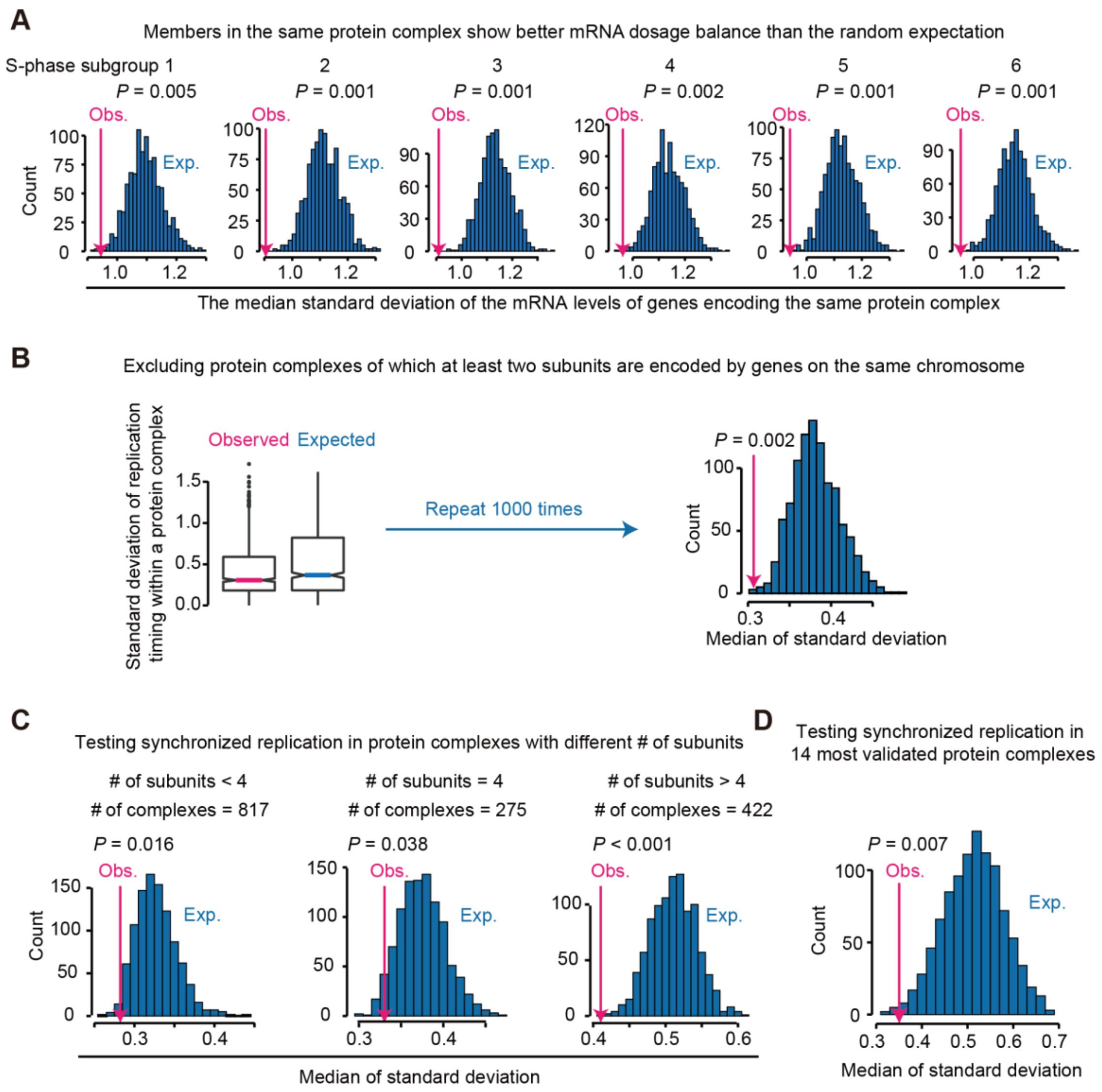
(B) The mRNA content per cell increases as the S phase progresses in our scRNA-seq data. The total number of unique molecular identifiers (UMI) of all genes in each cell is used to represent the mRNA content. Error bars represent the standard errors of the mean.



Supplemental Figure S2. Early-replicating genes are overexpressed during S phase, and *vice versa*.

(A) The same analysis was performed as in Fig. 2B when 10% of the 2000 early- or late-replicating genes are randomly dropped. Error bars represent the standard deviation of 1000 replicates.

(B) All possible gene pairs between two random genes expressed in S phase are generated. The expression ratio within each gene pair was calculated in each of the six subgroups in S phase, of which the coefficient of variation (CV) among subgroups was calculated. The gene pairs with the largest CV (ranking top 1%) exhibit a greater difference in replication timing than those with the smallest CV (ranking bottom 1%), indicating that the variation of the expression ratio between two genes through S phase is associated with their difference in replication timing. Outliers in the box plots are omitted.



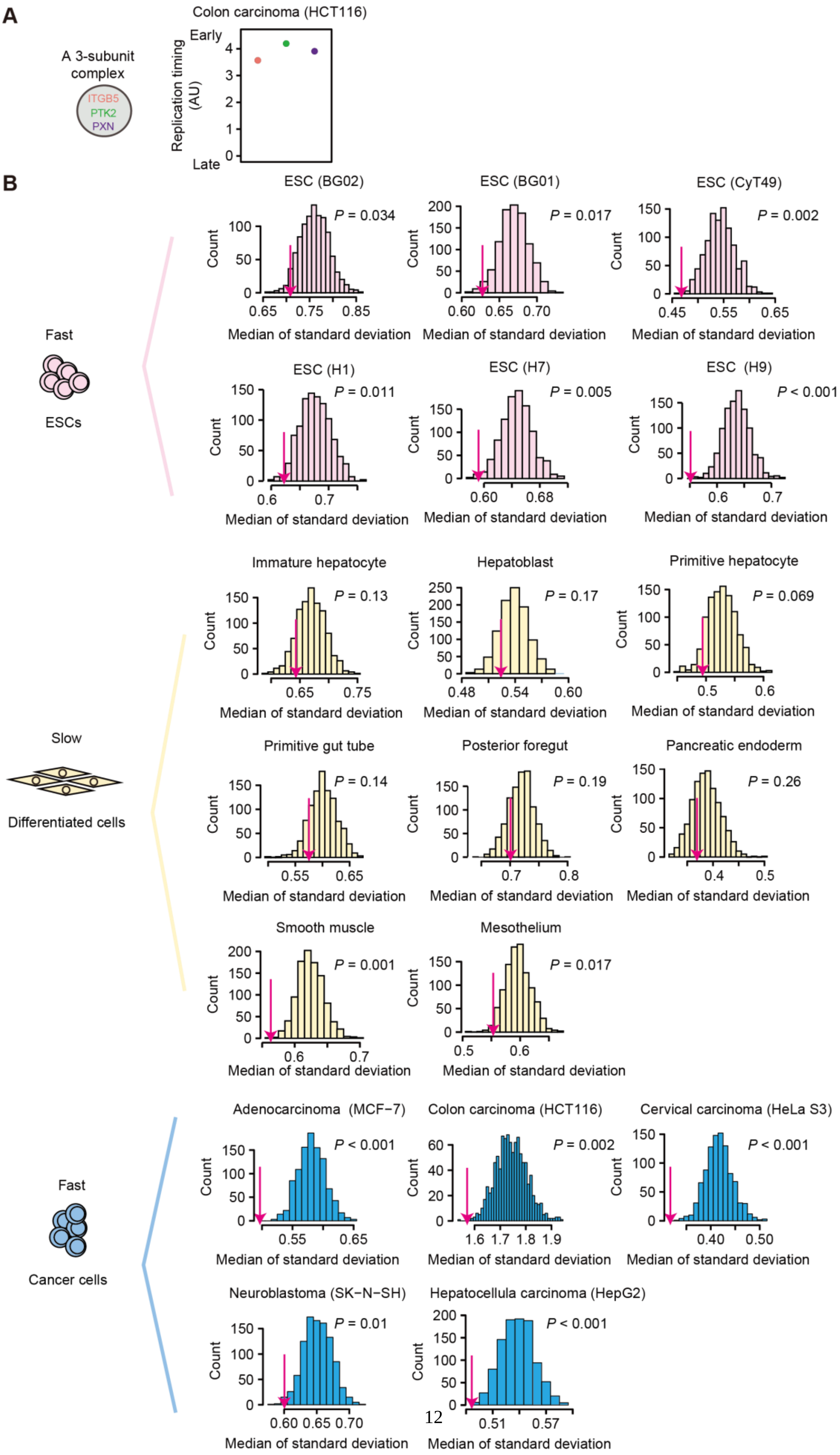
Supplemental Figure S3. The synchronized replication remained after controlling for confounding factors.

(A) For each subgroup in S phase, the standard deviation of the expression level ($\log_2$(RPKM+1)) of genes encoding the same protein complex was calculated. The expression levels of the genes encoding protein complexes were shuffled 1000 times to generate the random expectation. The median of the observed standard deviation (the pink arrow) was significantly smaller than the random expectation (blue) in each of the six subgroups. Note that we assumed genes encoding the same protein complex having a stoichiometric relationship of 1:1 since the stoichiometric data of protein complexes are not entirely available. Nevertheless, the test here was conservative because the observed standard deviation would be overestimated if the actual stoichiometric relationship was not 1:1.

(B) The median of the observed standard deviations of replication timing within a protein complex is significantly smaller than the random expectation after we excluded protein complexes of which at least two subunits are encoded by genes on the same chromosome.

(C) Protein complexes are separated into three classes according to the subunit number. The synchronized replication was tested in each class by only shuffling genes in this class. Note that we did not further separate complexes into 2-subunit and 3-subunit classes because the former has only 23 protein complexes.

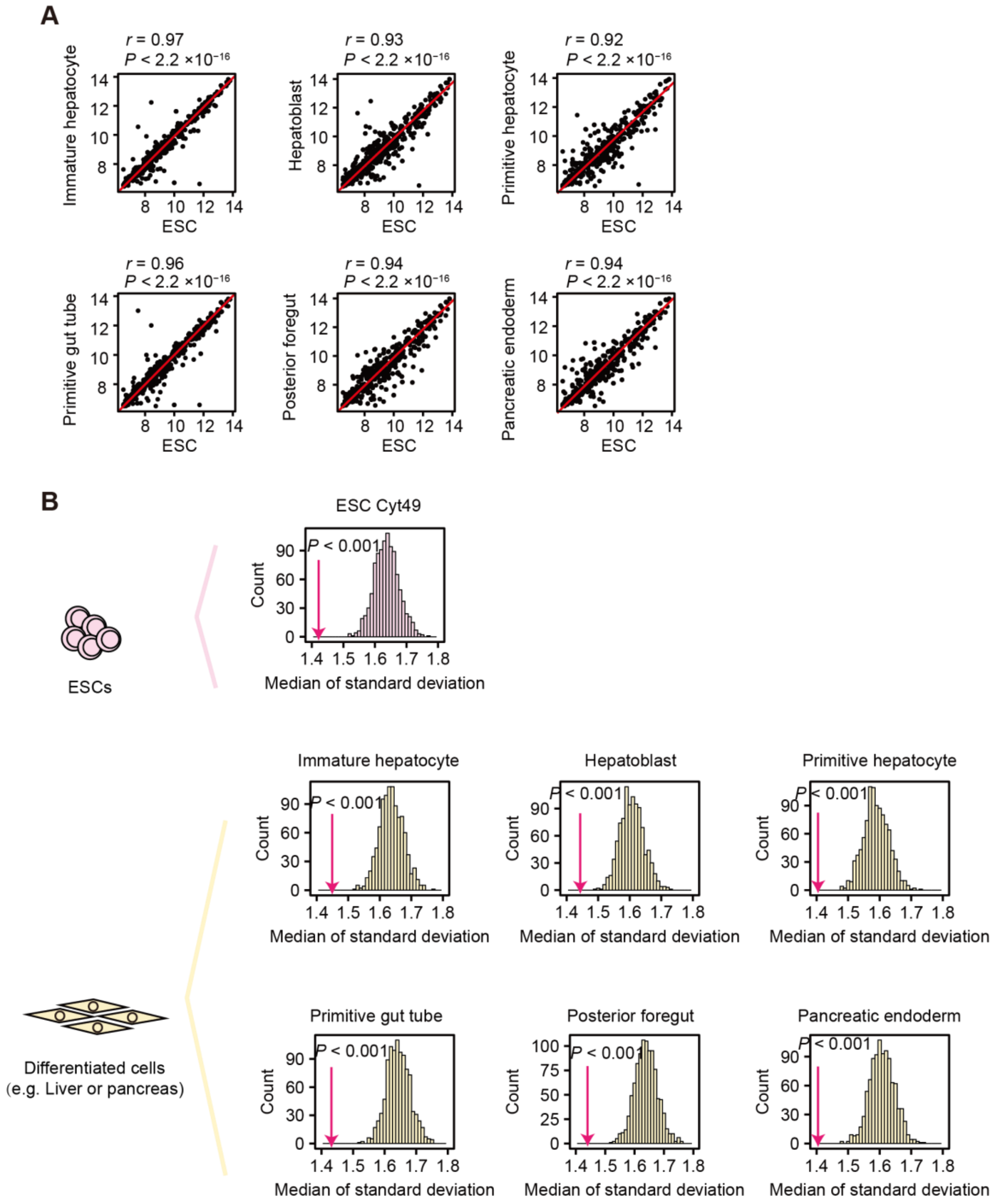
(D) Test of synchronized replication in 14 most validated protein complexes. These protein complexes were identified by at least four methods in DIP, including transcription factor II D complex (TF$_{II}$D, # of subunits = 12), exon junction complex (# of subunits = 4), and cyclin-dependent kinase 8 subcomplex (CDK8, # of subunits = 4).


A
A 3-subunit complex
ITGB5
PTK2
PXN
Colon carcinoma (HCT116)
Replication timing (AU)
Early
Late
0
1
2
3
4
B
Fast
ESCs
Slow
Differentiated cells
Fast
Cancer cells
ESC (BG02)
P = 0.034
ESC (BG01)
P = 0.017
ESC (CyT49)
P = 0.002
ESC (H1)
P = 0.011
ESC (H7)
P = 0.005
ESC (H9)
P < 0.001
Immature hepatocyte
P = 0.13
Hepatoblast
P = 0.17
Primitive hepatocyte
P = 0.069
Primitive gut tube
P = 0.14
Posterior foregut
P = 0.19
Pancreatic endoderm
P = 0.26
Smooth muscle
P = 0.001
Mesothelium
P = 0.017
Adenocarcinoma (MCF−7)
P < 0.001
Colon carcinoma (HCT116)
P = 0.002
Cervical carcinoma (HeLa S3)
P < 0.001
Neuroblastoma (SK−N−SH)
P = 0.01
Hepatocellula carcinoma (HepG2)
P < 0.001
Count
Median of standard deviation

Supplemental Figure S4. Test of synchronized replication in each of the 19 cell lines/types.

(A) Replication timing of three genes encoding a 3-subunit protein complex in HCT116 cell line.

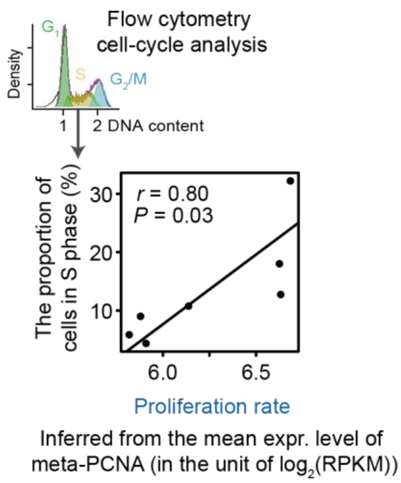
(B) The observed median of the standard deviation of replication timing within a protein complex for each cell type (magenta arrow) is shown on the distribution of the random expectation of 1000 shuffling.



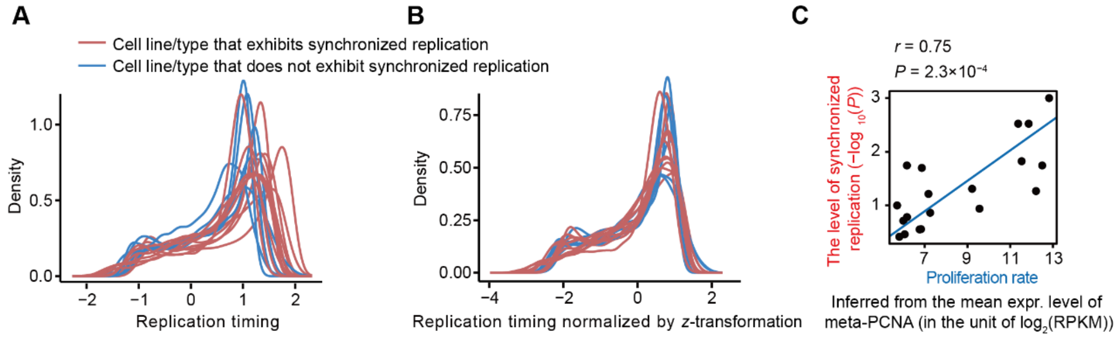
Supplemental Figure S5. The expression levels of genes encoding protein complexes in the liver and pancreas cells.

(A) Expression levels are highly correlated between ESC (Cyt49) and each of the six types of differentiated cells that do not exhibit synchronized replication. Each dot represents a protein complex-encoding gene. Pearson's correlation coefficients *r* and the corresponding *P* values are shown.

(B) Test of mRNA dosage balance among members of the same protein complex in the liver and pancreas cells as well as in ESC (Cyt49).



Supplemental Figure S6. The fraction of cells in S phase estimated by flow-cytometry can be predicted from the transcriptome-based proliferation rate. Therefore, the transcriptome-based proliferation rate can be used as an indicator of the fraction of time in S phase (S%) in the rest of this study.



Supplemental Figure S7. The S% dependent synchronized replication was observed after the *z*-transformation of the replication-timing programs.

(A) The distribution of replication timing of all genes in 18 cell lines/types. HCT116 is not shown because it significantly deviates from others in replication timing).

(B) The distribution of replication timing after the *z*-transformation of the replication-timing programs.

(C) The level of synchronized replication remained positively correlated with the fraction of time in S phase after the *z*-transforming. Pearson's correlation coefficient and the corresponding *P*-value are shown.