# **Final Project**

## **Replication Timing**

Replication timing refers to the order in which segments of DNA along the length of a chromosome are duplicated. DNA replication errors increase genetic instability, and may be a causative factor in diseases such as cancer and neuronal disorders. Replication in eukaryotes initiates from discrete genomic regions, termed origins, according to a strict, often tissue-specific, temporal program. The genetic program that controls activation of replication origins in mammalian cells has still not been elucidated.

Recent technology advancement can now allow measurements of replication timing genomewide. One such a technology is developed by Hiratani et al. (2008)(1). Briefly, cells were pulse labeled with BrdU and separated into early and late S-phase fractions by flow cytometry (Figure 1A). BrdU-substituted DNA from each fraction was immunoprecipitated with an anti-BrdU antibody, differentially labeled, and cohybridized to a whole-genome oligonucleotide microarray (Figure 1A). The ratio of the abundance of each probe in the early and late fraction ["replication timing ratio" = log2(Early/Late)] was then used to generate a replication-timing profile for the entire genome (Figure 1B). Biologists are often interested in partitioning the genome according to the obtained replication profiles to early replication domains, late replication domains and timing transition regions (TTR). The circular binary segmentation (CBS) algorithm (2) is often used to segment the genome(1;3;4). Early and late replication domains then are called as segments with log2(Early/Late)<0 and log2(Early/Late)>0, respectively (Figure 1B). The TTRs are called by first loess-smoothing the profile and identification of the regions with large positive and negative slopes(3) (Figure 1C). This method is not satisfactory both statistically and biologically. The replication domains and the transition regions should have no overlap and this cannot be guaranteed by this method. The choice of the slope to determine the TTRs is arbitrary (e.g. Ryba et al. 2010 used +/-68e-7 RT/bp) and the slope estimates could be influenced by the parameters used in loess smoothing. The aim of this project is to develop a Bayesian method that can improve the aforementioned method for detecting early/late replication domains and TTRs.

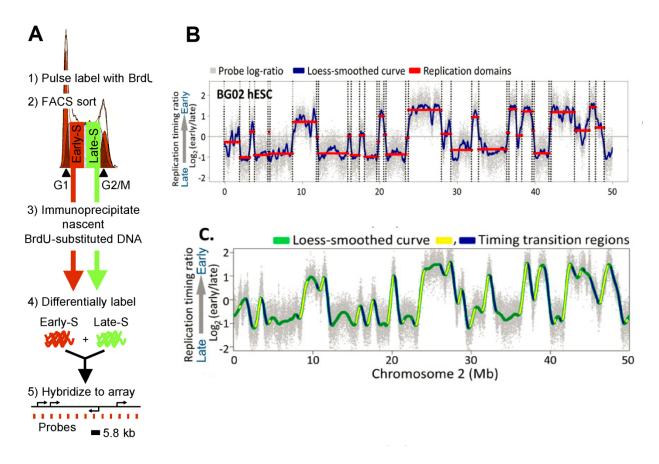


Figure 1 (A) Protocol for genome-wide replication timing analysis using oligonucleotide microarrays. (B) Replication timing profile across a 50-Mb segment of human chromosome 2. Data shown are the average of two replicate hybridizations (dye-swap) for hESC line BG02. DNA synthesized early vs. late during S phase was hybridized to an oligonucleotide microarray, and the log2 ratio of early/late signal for each probe (probe spacing 1.1 kb) across the genome was plotted on the y-axis vs. map position on the x-axis. (Gray dots) Rawdata. (Blue line) Loess-smoothed data. Replication domains (red lines) and

boundaries (dotted lines) were identified by circular binary segmentation (CBS algorithm). (C) Identification of timing transition regions (TTRs; blue and yellow highlight alternating TTRs) from loess-smoothed RT profiles. (Green) BG02 hESC.

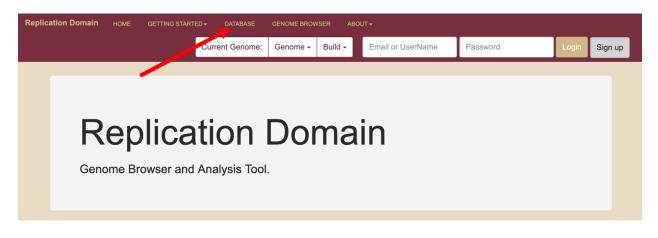
### **Project Task**

- 1. Develop a new Bayesian method for detecting early/late replication domains and TTRs. You need to clearly present your Bayesian model and explain why your model is appropriate. You should also develop the corresponding algorithm to estimate the parameters of your model and show that your method works well or even works better than the aforementioned method.
- 2. Apply your method to the data in Ryba et al. 2010 (3) and perform data analysis to show whether or not you can achieve similar conclusions about the replication domains and TTRs that you find. The data in Ryba et al. 2010 (3) can be downloaded from http://www.replicationdomain.org. Note: Since Ryba et al. 2010 (3) have a large amount of analyses, you do NOT need to perform every analysis in this paper, but you will be rewarded (specifically, in your final scores) if you perform more. If you have analysis results that do not agree with Ryba et al. 2010(3) and you can logically and convincingly explain why your results are correct, you will be significantly rewarded in your final score.

#### Remark about the data

Suppose that you want to download the replication timing data for the human embryonic stem cell line (hESC) BG01. After you open <a href="http://www.replicationdomain.org">http://www.replicationdomain.org</a> in your browser, click **Database** (shown

as the following figure).



In the next page, select Homo sapiens genome (i.e. human) and reference genome assembly hg38 in

eplication Domain	HOME	GETTING STARTED + DATABASE GENOME BROWSER ABOUT +										
			Current G	enome	: Homo sapiens -	hg38 <del>-</del>	Email or UserName	Password	Login Sign up			
HOME / DATABASE												
Data Type				Data		Differentiation						
RT	186		Accession#	Тур	e Cell Line	State	File Name		Chip/FlowCell ID			
GE	6	Select	Ext58122002	🛓 RT	10-212	Patient Leukemia	RD_X10_212_R1_Sm300_100913.	txt	441205A03_10-212_R1			
N-Domain	2				Sample							
RepOris	2	Select	Ext38079077	± RT	10-356	Patient	RD_X10_356_R1_Sm300_100913.	txt	441205A02_10-356_R1			
Hi-C	0					Leukemia Sample						
Cell Line		Select	Int71645241	Ł RT	10-771	Patient Leukemia Sample	RD_X10.771.R1_Sm300_02-03-2012.txt		471084A01_10-771.R1			
BG02	19											
CyT49	19	Select	Int90241999	Ł RT	10-799	Patient	RD_X10.799.R1_Sm300_05-23-20	11.txt	462966A01_10-799_R1			
H9	15					Leukemia Sample						
IMR90	9	(Data st.)	Int17906556	1 07	10-799	Patient	PD V10 700 P2 6m200 05 04 00	11 tut	462066402 10 700 02			
BG01	8	Select	11117900556	🛃 KI	10-199	Patient Leukemia	RD_X10.799.R2_Sm300_05-31-20	11.04	462966A03_10-799_R2			

the selection box (see below).

Under the Data Type menu list, you could select **RT** (replication timing) and under the cell line you could select **BG01**. You may choose to download ESC (embryonic stem cells) or NPC (neural precursor cells).

			Current Ge	enome:	Homo sapie	ens - hg38 -	Email or UserName	Password Login S
HOME / DATABAS	E							
Data Type			Accession#	Data Typ	e Cell Line	Differentiation State	File Name	Chip/FlowCell ID
RT	186	Select	Int93941761	🛓 RT	BG01	NPC	RD_BG01NPave_Sm300_090127.1	xt BG01NPC_UGA_15473302&15473602CGH
GE	6	Select	Int49279242	± RT	BG01	ESC	RD_BG01hES2_Sm300_090127.tx	t BG01_2_UGA_15707202CGH
N-Domain	2	Select	Int70021030	Ł RT	BG01	NPC	RD_BG01NP2_Sm300_090127.txt	BG01NPC_2_UGA_15473602CGH
RepOris	2	Select	Int61646100	Ł RT	BG01	ESC	RD_BG01hESave_Sm300_090127	.txt BG01_UGA_15473002&15707202CGH
Hi-C	0	Select	Int84203916	± RT	BG01	NPC	RD_BG01NP1_Sm300_090127.txt	BG01NPC_1_UGA_15473302CGH
Cell Line		Select	Int73895503	Ł RT	BG01	ESC	RD_BG01hES1_Sm300_090127.tx	t BG01hESC_1_UGA_15473002CGH
BG02	19							
CyT49	19							
Н9	15							
IMR90	9							
BG01	8							
Show More	75							

## **Reference List**

- 1. 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.
- 2. Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657-663.
- Ryba,T., Hiratani,I., Lu,J., Itoh,M., Kulik,M., Zhang,J., Schulz,T.C., Robins,A.J., Dalton,S. and Gilbert,D.M. (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*, 20, 761-770.
- 4. Hiratani,I., Ryba,T., Itoh,M., Rathjen,J., Kulik,M., Papp,B., Fussner,E., Bazett-Jones,D.P., Plath,K. and Dalton,S. (2010) Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome research*, **20**, 155-169.