2014

Modeling and integrative analysis with applications to DNA replication, cancer, and epigenetics

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

MODELING AND INTEGRATIVE ANALYSIS WITH APPLICATIONS TO DNA REPLICATION, CANCER, AND EPIGENETICS

by

YEVGENIY GINDIN

M.S., The George Washington University, 2007
B.A., The George Washington University, 2006
B.S., Syracuse University, 2002

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2014

Approved by

First Reader

Simon Kasif, Ph.D.
Professor of Biomedical Engineering, Bioinformatics and Computer Science, Boston University; Staff Scientist, Childrens Hospital Informatics Program,

Second Reader

Paul S. Meltzer, M.D, Ph.D.
Chief, Genetics Branch Center for Cancer Research, National Cancer Institute

. . . As the weeks passed
into spring and the plane trees in the courtyard
of the ancient hospital burst into new green,
I decided one morning to test sobriety,
to waken at dawn to sparrow chirp and dark clouds
blowing seaward from the Bultaco factory,
to inhale the particulates and write nothing,
to face the world as it was. Everything
was actual, my utterances drab, my lies
formulary and unimaginative.
For the first time in my life I believed
everything I said. Think of it: simple words
in English, Spanish or Yiddish, words
that speak the truth and no more, hour after
hour, day after day without end, a life
in the kingdom of candor, without fire or wine.

(Levine, 2011)

Acknowledgments

I am enormously grateful to my research advisor, Dr. Paul S. Meltzer for allowing me to draw upon his great expertise and invaluable guidance while pursuing exciting research in his laboratory. I am indebted to my academic advisor, Dr. Simon Kasif, whose insightful comments have shaped my research. I would like to thank my thesis committee, Dr. Marc E. Lenburg, Dr. Calin A. Belta, Dr. Thomas D. Tullius, and Dr. Scott Mohr for invaluable advice and dedication that they have shown to my education.

None of the work presented here would have been possible without talented scientists at the Meltzer Lab. I would like to acknowledge Drs. Sarah L. Anzick and Sean Davis for the work on the Polish Breast Cancer Study; Dr. Sven Bilke for the work on DNA replication timing and Drs. Princy Francis and Yuan Jiang for the work to study the effects of microRNA expression in osteosarcoma.

This dissertation would not have succeeded without encouragement and support from my family: my brother, parents and grandparents. I would like to especially thank my wife, Dr. Mariel Gindin, for support in life as well as science.

Finally, I would like to acknowledge everyone who has read and commented on my dissertation. Any errors that remain are entirely my responsibility.

MODELING AND INTEGRATIVE ANALYSIS WITH APPLICATIONS TO DNA REPLICATION, CANCER, AND EPIGENETICS

(Order No.)

YEVGENIY GINDIN

Boston University, Graduate School of Arts and Sciences

and

College of Engineering, 2014

Major Professors: Simon Kasif, Ph.D., Professor of Biomedical Engineering, Bioinformatics and Computer Science, Boston University
Paul S. Meltzer, M.D. Ph.D., Chief, Genetics Branch, Center for Cancer Research, National Cancer Institute

ABSTRACT

Biological organisms have evolved complex epigenetic mechanisms to tailor their gene expression programs to specific needs. These adaptations allow cells, that otherwise have identical genomes, to carry out specialized functions. In this work I develop and use data-integrative techniques to examine the mechanisms and consequences of epigenetic processes.

To better understand the changes in DNA methylation landscape that accompany breast cancer molecular subtypes, I integrated DNA methylation and gene expression data from 208 breast cancer samples obtained from a Polish population-based case-control study. Using a weighted correlation network approach, I identified gene co-methylation modules and asked if the genes in these modules are preferentially methylated and silenced in a breast cancer subtype-specific manner. This approach identified two non-overlapping gene co-methylation modules. The first module is silenced in Basal breast cancers, while the second is silenced in Luminal B breast cancers. Gene-set enrichment analysis suggests that epigenetic silencing of these modules interferes with processes that maintain cellular dif-

ferentiation, and that the methylation status of the Luminal B module is associated with disease prognosis.

To uncover the determinants of the temporal order of metazoan genome replication, I used a reductionist model of DNA replication to test the ability of hundreds of epigenetic marks to predict replication timing. My work showed that DNA replication timing can be completely predicted from locations of DNase I hypersensitive sites. I further demonstrated the robust emergent character of DNA replication that could be understood without invoking a complex regulatory mechanism.

To determine the underlying cause of cell de-differentiation in osteosarcoma, I examined the relationship between microRNA expression and the bone-cell differentiation program. Focusing on the inhibitory role of miR-23a in bone differentiation, I analyzed the effect of its over-expression in osteosarcoma cells. Extensive computational analysis led me to propose that a major mechanism by which miR-23a exerts its effect is by interfering with expression of *GJA1*, which encodes a gap junction channel essential for intercellular communication and external stimuli sensing in bone cells. Follow-up experiments indicate that *GJA1* is sharply up-regulated during bone cell differentiation and that *GJA1* inhibition significantly delays the onset of differentiation.

Together, this work uses data integrative techniques to provide new insights into the decisive role of epigenetic processes in cellular differentiation.

Contents

List of Tables

List of Figures

List of Abbreviations

EMT	.....	Epithelial to mesenchymal transition
ERK	.....	Extracellular-signal-regulated kinase
HOS	.....	Human osteosarcoma cells
HS	.....	Hypersensitive
IPLS	.....	Initiation Probability Landscape
K-L	.....	Kullback Leibler
kb	.....	kilo-base
miR	.....	Micro RNA
mRNA	.....	Messenger RNA
OS	.....	Osteosarcoma
PRC2	.....	Polycomb repressive complex 2
qPCR	.....	Quantitative polymerase chain reaction
TF	.....	Transcription factor
TFBS	.....	Transcription factor binding site
UTR	.....	Untranslated region

Chapter 1

Introduction

In multi-cellular organisms, cell function is determined not by genome sequence, which is identical for every cell in an organism, but by a complex and multi-layered control mechanism, generally referred to as *epi*genetic, literally meaning *above* the genome. The earliest use of the term was used to describe the relationship between genotype and phenotype during development (Waddington, 1957). As interest in epigenetics grew, Andrew Riggs and colleagues expanded the definition beyond the realm of developmental biology to cover all heritable changes that could not be explained by DNA sequence alone (Russo et al., 1996). In the years since, the common usage of epigenetics has expanded to cover, in addition to DNA methylation, chromatin and its marks, the spatial organization of the chromosomes within the nucleus and effects of RNA interference. Many of these processes, however, are short-lived and are likely not inheritable – making Riggs' definition obsolete. In light of this discrepancy, Adrian Bird re-defined epigenetics as "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" (Bird, 2007).

This work is focused on analysis of dysfunction of epigenetic processes in disease. This is commonly referred to as *disease* epigenetics, described as "disruption of phenotypic plasticity – the ability of cells to change their in response to internal or external environmental cues" (Feinberg, 2007). The thesis is structured into three parts, each addressing a different aspect of epigenetic control of cell phenotypic plasticity. The first part tackles DNA methylation patterns in a population-based breast cancer dataset, uncovering dysregulated cell differentiation pathways relevant to breast cancer biology. The second part

reveals that the temporal order of DNA replication in metazoan cells, which is cell lineage specific, can be completely predicted using an intuitive approach that relies on locations of DNase hypersensitive sites. The third and final part identifies links between dysregulation of microRNA-23a and abnormal differentiation of bone cells in osteosarcoma.

1.1 DNA Methylation in Breast Cancer

Genome scale expression profiling has led to the development of novel classifications of breast cancer with growing importance for targeted treatment. However, the global DNA methylation changes that accompany breast cancer expression-based subtypes are incompletely understood. Integration of DNA methylation profiles with gene expression data may improve classification, diagnosis, and management of breast cancer. Accordingly, I analyzed DNA methylation and gene expression data from 208 breast cancer samples obtained from a Polish population-based case-control study. I used a well-established weighted correlation network analysis approach to identify clusters of correlated CpG island methylation probes, which I termed DNA methylation modules. I found two gene co-methylation modules strongly correlated with Basal (hypermethylated in Basal) and Luminal B (hypermethylated in Luminal B) breast cancer subtypes. Functional analysis of the Basal gene co-methylation module shows that it significantly overlaps with genes silenced during epithelial to mesenchymal transition. On the other hand, the Luminal B gene co-methylation module significantly overlaps with gene targets of the Polycomb repression complex 2. Kaplan-Meier survival analysis revealed poor prognosis associated with hypermethylation of the Luminal B gene co-methylation module. This study offers new biological perspectives on the role of DNA methylation in breast carcinogenesis and may have value in refining molecular classifications of breast cancer.

1.2 DNA Replication Timing

The metazoan genome is replicated in precise cell lineage specific temporal order. However, the mechanism controlling this orchestrated process is poorly understood as no molecu-

lar mechanisms have been identified that actively regulate the firing sequence of genome replication. Here I develop a mechanistic model of genome replication capable of predicting, with accuracy rivaling experimental repeats, observed empirical replication timing program in humans. In this model, replication is initiated in an uncoordinated (time-stochastic) manner at well-defined sites. The model contains, in addition to the choice of the genomic landmark that localizes initiation, only a single adjustable parameter of direct biological relevance: the number of replication forks. I find that DNase hypersensitive sites are optimal and independent determinants of DNA replication initiation. I demonstrate that DNA replication timing program in human cells is a robust emergent phenomenon that, by its very nature, does not require a regulatory mechanism determining a proper replication initiation firing sequence.

1.3 MicroRNAs and Differentiation

Osteosarcoma is the most common type of bone cancer in children and adolescents. Impaired differentiation of osteoblast cells is a distinguishing feature of this aggressive disease. As improvements in survival outcomes have largely plateaued, better understanding of the bone differentiation program may provide new treatment approaches. To this end, I carried out a large-scale integrative computational analysis of mRNA and miRNA expression as well as genome copy number aberrations of sixteen osteosarcoma and two osteoblast cell lines. This work identified copy number gain and over-expression of miRNA cluster miR-23a∼27a∼24-2 in a substantial fraction of osteosarcoma samples. Previous studies identified interactions between the microRNAs in this cluster, particularly miR-23a, and select genes important for bone development. However, global changes in gene expression associated with functional gain of this cluster have not been fully explored. Experimental results show that over-expression of miR-23a delays ossification and calcification in osteosarcoma (HOS) cells. Downstream bioinformatic analysis identified miR-23a target gene connexin-43 (*Cx43/GJA1*), a mediator of intercellular signaling critical to osteoblast development, as acutely affected by miR-23a levels. Connexin-43 is up-regulated in the

course of HOS cell differentiation and is down-regulated in cells transfected with miR-23a. Analysis of gene expression data, housed at Gene Expression Omnibus, reveals that Cx43 is consistently up-regulated during osteoblast differentiation. Suppression of Cx43 mRNA by miR-23a was confirmed *in vitro* using a luciferase reporter assay. This work demonstrates novel interactions between microRNA expression, intercellular signaling and bone differentiation in osteosarcoma.

Chapter 2

Methylation portraits of breast cancer: results from a population-based case-control study

2.1 Introduction

Clinical, pathological and epidemiological heterogeneity among breast cancers has long been recognized (Rakha et al., 2008). Breast cancer can be classified into one of the five generally accepted subtypes by gene expression profiling. Luminal subtypes express the estrogen receptor and are the most common type of breast cancer, accounting for two-thirds of all cases (Carey, 2010). Luminal cancers are further divided into Luminal A and Luminal B, with Luminal B being more proliferative. Basal subtypes are highly proliferative and are characterized by low expression of estrogen receptor. The Normal-like subtype is the least well characterized; it is distinguished by low expression of other intrinsic clusters. Importantly, categorization of breast cancers according to distinct gene expression profiles has been relevant to clinical management (Sø rlie et al., 2001; van de Vijver et al., 2002; Paik et al., 2004; Gruvberger et al., 2001).

Luminal and Basal breast cancers may originate from distinct cell lineages (Li and Durbin, 2009). Since lineage commitment is driven by epigenetic processes, a number of recent studies explored a possible connection between DNA methylation patterns in breast cancer and cell types of origin (Dedeurwaerder et al., 2011). DNA methylation is often observed in the presence of H3K27me3, which is associated with the action of the Polycomb repressive complex 2 (PRC2) (Ohm et al., 2007; Schlesinger et al., 2007; Widschwendter et al., 2007). Holm and colleagues observed breast cancer subtype-specific patterns of DNA methylation of PRC2 gene targets (Holm et al., 2010). Specifically, in Basal breast cancer,

these genes are silenced via PRC2-induced trimethylation of H3K27; whereas in Luminal B cancers, the same genes are silenced via DNA methylation of the promoter regions. However, the relationships between Polycomb proteins, DNA methylation-mediated gene silencing and their impact on survival remain to be fully explored.

The purpose of this study is to derive a unified understanding of the contribution of DNA methylome to breast cancer biology. To this end, we broadly explored the CpG island DNA methylation landscapes obtained by studying a subset of breast cancers collected in a population-based case-control study conducted in Poland (García-Closas et al., 2006). In studying DNA methylation patterns, we used a weighted correlation network approach, identifying clusters of highly correlated CpG island methylation markers termed gene co-methylation modules (Horvath et al., 2012). A distinguishing feature of our approach is that gene co-methylation modules were constructed without the use of expression data, gene annotation, or other data reduction techniques such as variance filters. We examined a number of features of these modules including methylation and expression patterns, reproducibility in an independent dataset, gene functional enrichment, and prognostic properties.

2.2 Results

2.2.1 Sample classification

We profiled gene expression and DNA methylation of 208 breast cancer samples (Table 1) using DNA microarray hybridization. We determined the distribution of these samples across established breast cancer categories (Sø rlie et al., 2001) by applying a random forests classifier (Breiman, 2001), a well-established class prediction method (Díaz-Uriarte and Alvarez de Andrés, 2006), on the gene expression data. As in previous studies, although these categories were clearly apparent, individual samples varied in how well they could be separated from neighboring clusters. Overall frequency of breast cancer subtypes observed in these 208 samples was: Luminal A (42.3%), Luminal B (20.2%), Basal (11.1%), HER2 (11.5%), and Normal-like (14.9%) (Table 2.1).

Table 2.1: Patient Cohort Characteristics (N=208)

Age	Median	55
	Interquartile range	49-63
Diagnosis – no. (%)	In situ invasive	151 (73%)
	Invasive only	57 (27%)
Tumor type – no. (%)	Ductal	115 (55%)
	Lobular	34 (16%)
	Medullary	2 (1%)
	Mixed (ductal and lobular)	35 (17%)
	Mucinous (colloid)	1 (0%)
	Papillary	1 (0%)
	Tubular cribriform	7 (3%)
	Tubulolobular	10 (5%)
	Other primary	3 (1%)
Tumor differentiation – no. (%)	Poor	54 (26%)
	Moderate	119 (57%)
	Well	35 (17%)
Estrogen Receptor Status – no. (%)	Positive	127 (61%)
	Negative	64 (31%)
	Undetermined	17 (8%)
Progesterone Receptor Status – no. (%)	Positive	106 (51%)
	Negative	85 (41%)
	Undetermined	17 (8%)
Tumor Diameter – no. (%)	0.6-1.0 cm	6 (3%)
	1.1-2.0 cm	92 (44%)
	2.1-5.0 cm	102 (49%)
	>5.1 cm	8 (4%)
Predicted Molecular Subtype – no. (%)	Basal	23 (11%)
	Her2	24 (12%)
	Luminal A	88 (42%)
	Luminal B	42 (20%)
	Normal-like	31 (15%)

2.2.2 Correlation network approach identifies gene co-methylation modules correlated with Luminal B and Basal breast cancer subtypes

Using a network correlation approach (see section 2.4), we examined the possible contribution of DNA methylation in the establishment of gene expression-derived breast cancer categories. We defined gene co-methylation modules (Langfelder and Horvath, 2008; Horvath et al., 2012) as sets of highly correlated methylation probes, uncovering 27 such modules (Fig. 2·1). Out of the initial modules identified, we found two dominant gene co-methylation modules that show significant correlations to breast cancer subtype annotations (see 2.4.6) based on mRNA profiling (Fig. 2·1). The average methylation profile of probes in these modules is significantly correlated with sample breast cancer subtype annotations. The first (Basal) gene co-methylation module (1094 probes corresponding to 943 genes) is significantly correlated with Basal breast cancer subtype ($P = 4 \times 10^{-3}$), while the second (Luminal B) gene co-methylation module (755 probes corresponding to 505 genes) is significantly correlated with the Luminal B breast cancer subtype ($P = 9 \times 10^{-5}$).

2.2.3 Basal and Luminal B gene co-methylation modules exhibit subtype-specific expression and methylation patterns

We next sought to define the relationships between expression and methylation in the Basal and Luminal B gene co-methylation modules. We found that the Basal and Luminal B gene co-methylation modules exhibit inversely correlated expression and methylation patterns that are breast cancer subtype specific (Fig. 2·2). The Basal gene co-methylation module generally exhibits a greater degree of DNA methylation in Basal breast cancer samples compared to samples characterized as Normal-like or Luminal (Fig. 2·2). Moreover, the expression of the Basal gene co-methylation module is significantly down-regulated in Basal breast cancer subtypes (Fig. 2·2). Similar to the Basal gene co-methylation module, the Luminal B gene co-methylation module shows an inverse expression and methylation pattern. The Luminal B gene co-methylation module is characterized by hypermethylation (Fig. 2·2) and reduced expression (Fig. 2·2) in Luminal B samples.

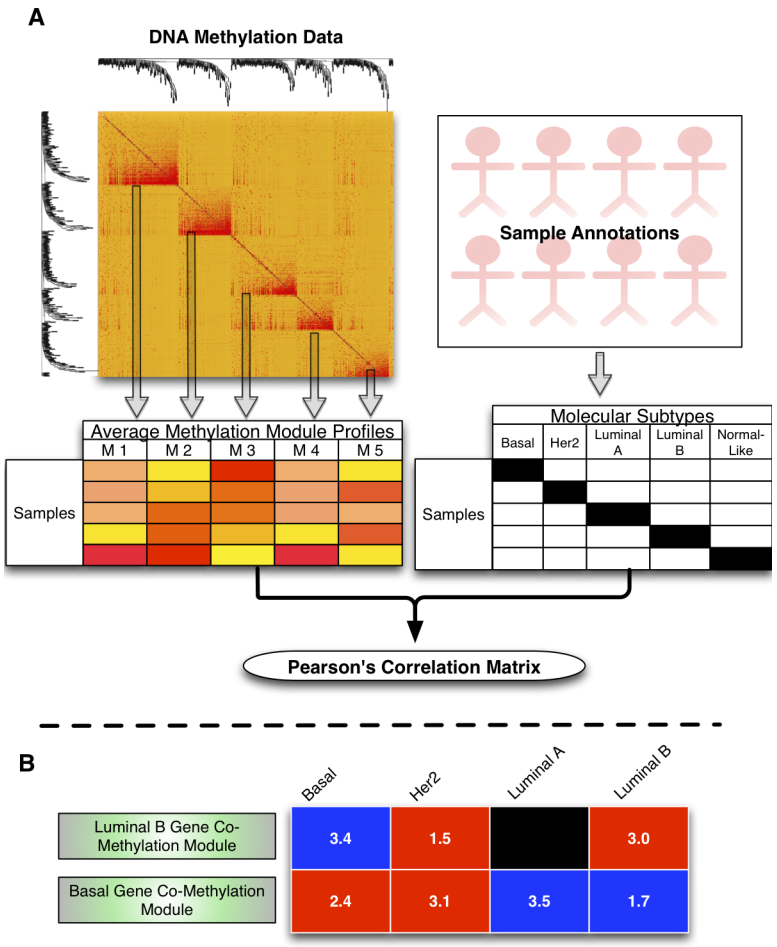

Figure 2·1: Gene co-methylation module discovery and selection. **(A)** Schema representing our approach. Starting with CpG island methylation data, we identify sets of highly correlated methylation probes evidenced by red squares along the diagonal and branches of the accompanying hierarchical clustering tree. In this simulated example, five gene co-methylation modules were uncovered. Next, for each module, we calculate an average methylation profile. Separately, we extract gene expression-derived breast cancer molecular subtype annotations from our dataset. As a final step, we calculate Pearson's correlation coefficients between methylation module profiles and molecular subtype classifications with the goal of identifying methylation modules that are significantly correlated with molecular subtypes. **(B)** We identified two gene co-methylation modules highly correlated with expression-derived breast cancer molecular subtypes. Numbers indicate P values for Pearson's correlation expressed in negative log; red signifies positive correlations; blue signifies negative correlations; black signifies no significant correlation.

To ensure that grouping for of methylation probesets is not due to a probe sequence artifact, we calculated Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) for every probe pair on the DNA methylation microarray. While K-L divergence is often used to differences between two probability distributions, it has proven useful when the goal is to quantify the differences between a pair of short texts (see for instance (Pinto et al., 2007)). We find that the mean K-L divergence for methylation probes within the Basal gene co-methylation module equals to that of the Luminal gene co-methylation module at 0.19. Therefore, we conclude that DNA methylation probe sequence similarity is not a contributing factor in our analysis.

2.2.4 The Basal and Luminal B gene co-methylation modules are preserved in an independent dataset

To determine if the Basal and Luminal B gene co-methylation modules are robust and reproducible, we complemented our analysis with an external dataset recently published by Dedeurwaerder and colleagues (Dedeurwaerder et al., 2011). Dedeurwaerder et al. analyzed DNA methylation profiles of 248 breast cancer tissues with a subtype composition similar to this study using the same microarray platform. When reanalyzing their dataset with our methods, we find that the Basal and Luminal B gene co-methylation modules are highly preserved. The P value of the preservation statistic (Langfelder and Horvath, 2008; Langfelder et al., 2011) for the Luminal B and Basal modules in the two datasets is less than 1×10^{-20}, suggesting that these gene co-methylation modules are stable and are not due to a data artifact.

2.2.5 The Basal gene co-methylation module is enriched for gene signatures related to the process of epithelial to mesenchymal transition

To determine the likely function of the Basal gene co-methylation module, we performed a functional gene enrichment analysis of the genes in this module. We find that the Basal gene co-methylation module is highly enriched for gene signatures that relate to epithelial to mesenchymal transition (EMT) in cancer (Table 2.2). The most enriched signature within

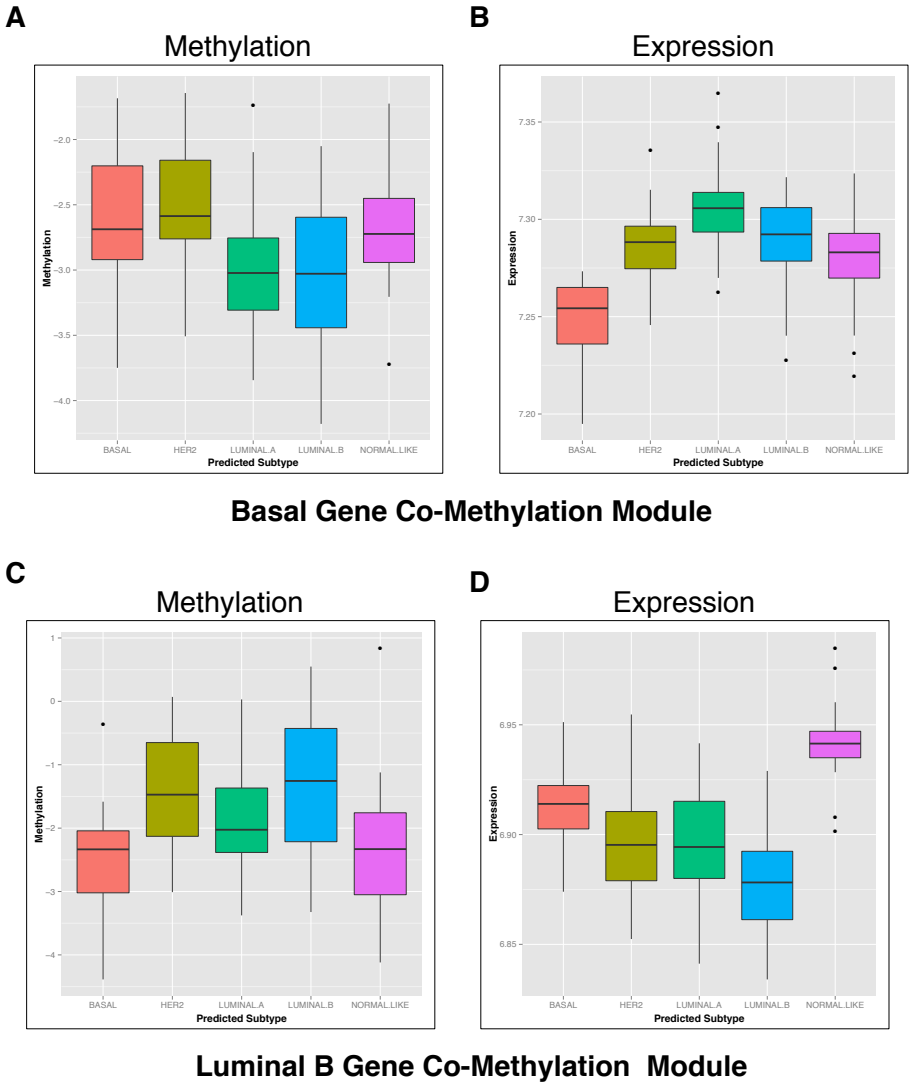

Figure 2·2: Average expression and methylation levels of Basal and Luminal B gene co-methylation modules. **(A)** Average CpG island methylation of Basal gene co-methylation module. **(B)** Average gene expression of Basal gene co-methylation module. **(C)** Average CpG island methylation of Luminal B gene co-methylation module. **(D)** Average gene expression of Luminal B gene co-methylation module.

Table 2.2: Top Enriched Basal Gene Co-Methylation Module Gene Signatures

Signature	**Enrichment P value**
Charafe breast cancer luminal vs mesenchymal up	6.26×10^{-14}
Doane breast cancer ESR1 up	3.19×10^{-8}
Gozgit ESR1 targets dn	5.80×10^{-8}
Smid breast cancer basal dn	1.29×10^{-7}
Aigner ZEB1 targets	5.21×10^{-7}
Onder CDH1 targets 2 dn	8.81×10^{-7}
Ohm methylated in adult cancers	1.34×10^{-6}
Lien breast carcinoma metaplastic vs ductal dn	2.41×10^{-6}
Wamunyokoli ovarian cancer lmp up	2.60×10^{-6}
Charafe breast cancer luminal vs basal up	2.92×10^{-6}

the Basal gene co-methylation module is that of genes down-regulated in mesenchymal-like breast cancer cell lines compared to luminal-like breast cancer cell lines (Subramanian et al., 2005; Charafe-Jauffret et al., 2006) with a P value of 6.26×10^{-14}. Furthermore, there is a strong enrichment ($P = 8.81 \times 10^{-7}$) of genes within the Basal gene co-methylation module that were observed to be down-regulated after knockdown of the E-cadherin (*CDH1*) gene (Onder et al., 2008; Subramanian et al., 2005).

The expression and methylation profiles of EMT and E-cadherin signatures within the Basal gene co-methylation module are inversely correlated and breast cancer subtype specific (Fig. 2·2). These signatures are consistently hypermethylated in Basal breast cancer samples relative to other subtypes. The expression patterns follow a reverse trend: gene silencing in Basal breast cancer subtypes and up-regulation in remaining subtypes.

In light of these findings, we examined whether the Basal gene co-methylation module is enriched for targets of transcription factors known to play a role in EMT. Transcription factor target enrichment analysis using a manually curated database (Ingenuity Systems; www.ingenuity.com) uncovered that the Basal gene co-methylation module is enriched ($P = 1.9 \times 10^{-4}$) with targets of *SNAI1* a potent repressor of E-cadherin and an inducer of EMT (Peinado et al., 2007). Among the genes identified, as under the influence of SNAI1 in the

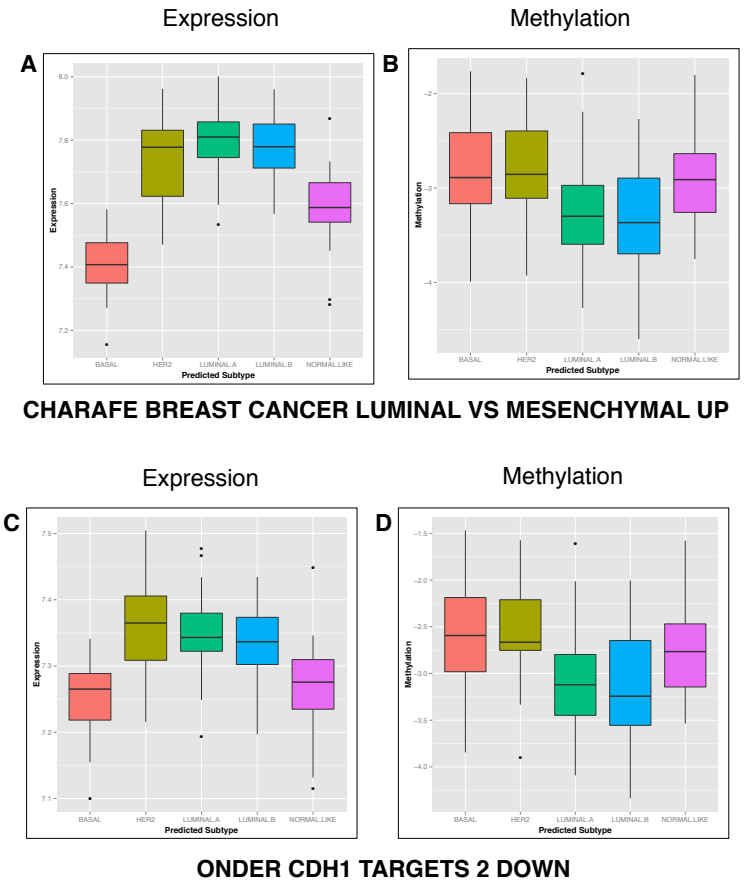

Figure 2·3: Average expression and methylation levels of gene signatures associated with the epithelial to mesenchymal transition within the Basal gene co-methylation module. **(A)** Average gene expression and **(B)** CpG island methylation of a gene signature up-regulated in mesenchymal breast cancers relative to Luminal ones. **(C)** Average gene expression **(D)** and CpG island methylation of a gene signature with knock-down of the E-cadherin gene.

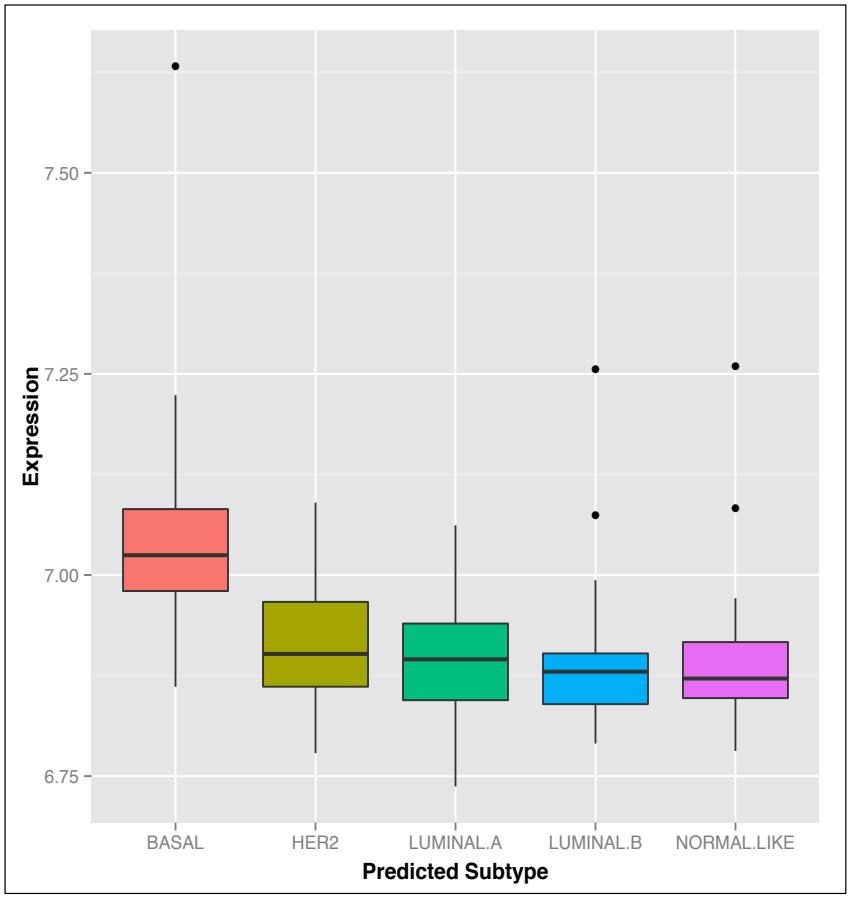

Figure 2·4: Expression of transcription factor Snail1 across the PBCS dataset.

Basal gene co-methylation module, is the E-cadherin gene as well as *CCND1*, *OCLN*, *MUC1*, and *PTEN*. Moreover, our results show that *SNAI1* is overexpressed in Basal breast cancer subtypes (Fig. 2·4).

Further examination of Basal gene co-methylation module reveals that it contains a number of genes that have been described as hypermethylated in cancer (Heyn and Esteller, 2012; Esteller, 2007). These genes include cell-cell adhesion gene *CDH1*; estrogen receptor *ESR1*; detoxification enzyme gene *GSTP1*; DNA repair gene *MGMT*; retinoic acid receptor gene *RARB*; cell cycle control gene *SEPT9*; cell growth control gene *SNCG*; extracellular matrix protein gene *THBS1* and TSH receptor gene *TSHR*.

2.2.6 The Luminal B gene co-methylation module is enriched for gene signatures related to Polycomb targets

To determine the likely function of the Luminal B gene co-methylation module, we performed an enrichment analysis of the genes in this module. The enrichment analysis shows that a significant number of genes of the Luminal B gene co-methylation module have been characterized previously as targets of histone mediated transcriptional silencing (Table 2.3). The Luminal B gene co-methylation module is highly enriched ($P = 3.40 \times 10^{-71}$) for gene targets of Polycomb repressive complex 2 (PRC2) (Ben-Porath et al., 2008; Subramanian et al., 2005).

The enriched PRC2 signature shows signs of differential methylation and expression across the breast cancer subtypes (Fig. 2·5). The genes in this signature are most methylated in the Luminal B samples and are least methylated in Basal samples. Moreover, the PRC2 expression signature is down-regulated in Luminal B subtypes relative to the remainder of the dataset.

The above results lead us to examine whether *EZH2*, which codes for the histone methyltransferase subunit of PRC2, is differentially expressed across the breast cancer subtypes in our dataset. We found *EZH2* to be overexpressed in Basal breast cancer subtypes (Fig. 2·6), in line with previous reports (Holm et al., 2010). Expression of the

Table 2.3: Top Enriched Luminal B Gene Co-Methylation Module Gene Signatures

Signature	**Enrichment P value**
Benporath EED Targets	1.52×10^{-93}
Benporath SUZ12 targets	1.12×10^{-85}
Benporath PRC2 targets	3.40×10^{-71}
KEGG neuroactive ligand receptor interaction	5.65×10^{-18}
Schlesinger H3K27me3 in normal and methylated in cancer	1.77×10^{-17}
Reactome GPCR ligand binding	2.18×10^{-17}
Hatada methylated in lung cancer up	1.32×10^{-14}
Reactome downstream events in gpcr signaling	6.84×10^{-13}
rReactome class a1 rhodopsin like receptors	3.68×10^{-11}

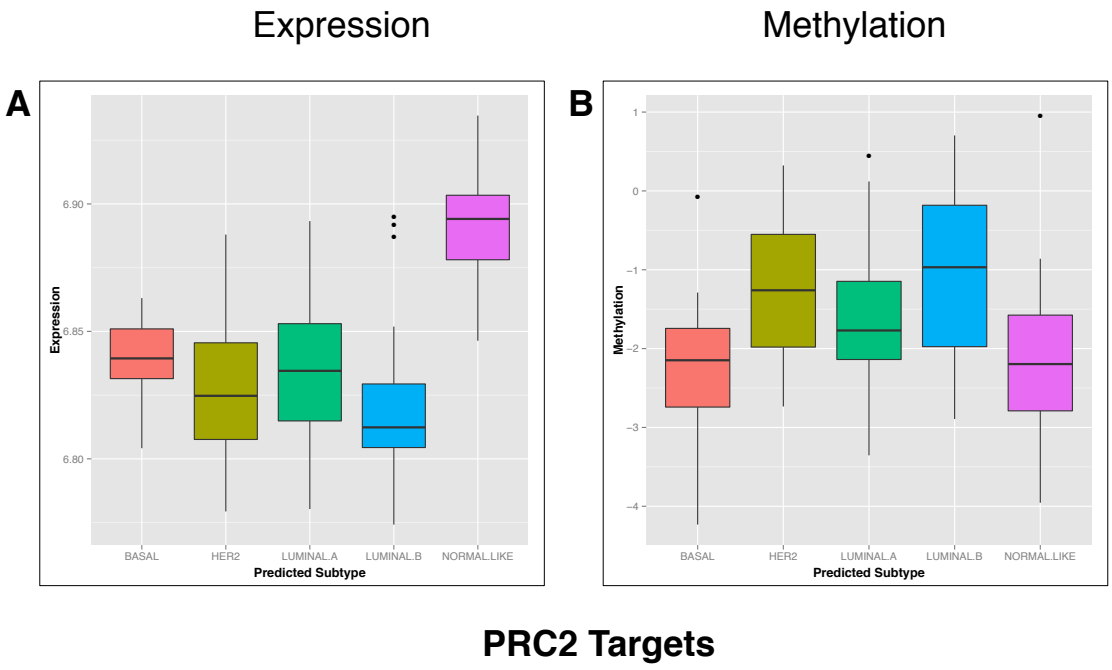

Figure 2·5: Average expression **(A)** and methylation levels **(B)** of a gene set identified as under the control of the PRC2 complex within the Luminal B gene co-methylation module.

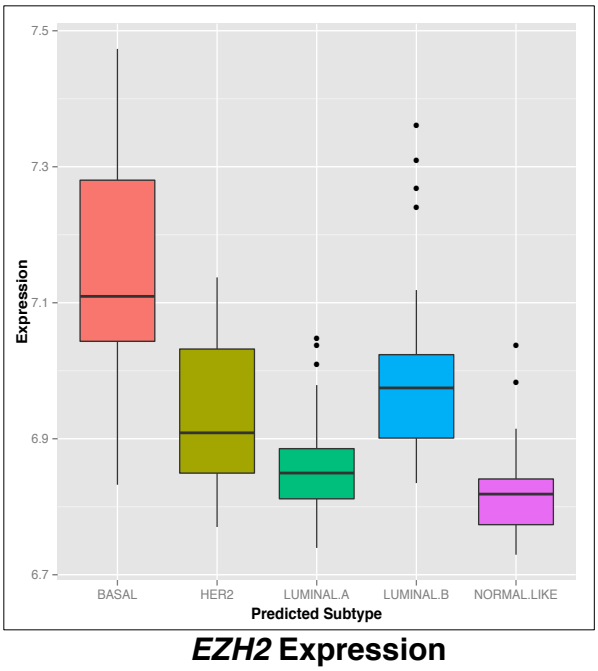

Figure 2·6: Expression of PRC2 complex member EZH2 across the PBCS dataset.

remaining subunits of PRC2, namely *EED* and *SUZ12*, was found not to vary across breast cancer subtypes (results not shown).

2.2.7 Methylation of the Luminal B gene co-methylation module is associated with poor prognosis

In a previous work, we identified a methyl-deviator breast cancer epigenotype associated with high-grade, poor-survival, estrogen receptor positive cancers (Killian et al., 2011). The Luminal B gene co-methylation module significantly overlaps with the genes of the methyl-deviator epigenotype ($P = 1.75 \times 10^{-29}$). Therefore, we examined whether the Luminal B gene co-methylation module may be predictive of clinical outcome.

Using Kaplan-Meier analysis, we find that the methylation status of the Luminal B gene co-methylation module is associated with significantly worse outcome within Luminal samples ($P = 0.002$; Fig. 2·7). Multivariate analysis shows that the Luminal B gene co-methylation module has prognostic value independent of commonly used clinical indicators

including age and tumor size (Table 2.4). A Gene Set Enrichment Analysis-inspired test (see 2.4), to verify the prognostic value of the Luminal B gene co-methylation module, shows that it is overrepresented at the top of the list of methylation probes ranked by association with clinical outcome (Fig. 2·7). These results demonstrate that DNA methylation of Luminal B gene co-methylation module is associated with poor prognosis.

Table 2.4: Multivariate Odds Ratios and 95% Confidence Intervals for Risk of Death among the Patient Cohort with Luminal Breast Cancer

Variable	Odds Ratio (95% CI)	P Value
Age	1.03 (0.98-1.11)	0.27
ER Status (Positive vs Negative)	0.82 (0.22-3.02)	0.77
Tumor Size	0.66 (0.39-1.11)	0.11
Methylation Status of the Luminal B Gene Co-Methylation Module (Hypomethylation vs Hypermethylation)	0.25 (0.08-0.79)	0.02

2.3 Discussion

In this study, we used an unsupervised network analysis of DNA methylation data to reveal a number of key biological insights into the relationship between DNA methylation, gene expression, and epigenetic programming with relevance to survival outcomes. Thus far, these relationships have largely been studied using data-reduction approaches that limit the number of genes considered based on their prior defined role, such as those thought to be important in cancer, or their ability to discriminate between prior defined sample groups, or their variance across the dataset. The current study overcomes many of these limitations by making few assumptions about the data.

Our results show (Fig. 2·5) that there is a significant hypermethylation and silencing of PRC2 gene targets (Table 2.3) in Luminal B subtypes, which form the Luminal B gene co-methylation module. We also observed an accompanying downregulation of *EZH2* in Luminal B tumors (Fig. 2·6). Conversely, in Basal tumors, PRC2 targets are hypomethylated (Fig. 2·5) and *EZH2* is expressed at significantly higher levels compared to other

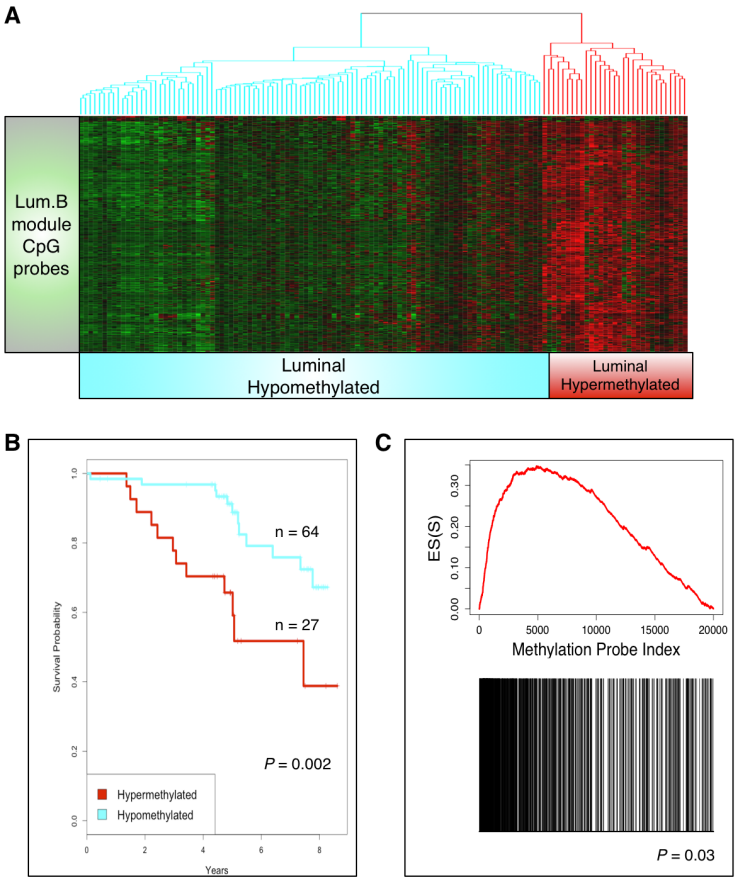

Figure 2·7: Survival analysis based on the methylation status of the Luminal B gene co-methylation module applied to Luminal samples. **(A)** Luminal breast cancer samples were split into two groups according to an unsupervised hierarchical clustering analysis. **(B)** Samples where the Luminal B gene co-methylation is hypermethylated (red) are associated with worse prognosis compared to the remainder of the data set (magenta). **(C)** Gene Set Enrichment Analysis, where methylation probes are ranked by their association with clinical outcome, shows that the methylation probes that belong to the Luminal B gene co-methylation module are overrepresented at the top of the ranking.

breast cancer subtypes (Fig. 2·6). The expression pattern of PRC2 in Basal tumors may reflect the presence of bivalent chromatin marks. Such a chromatin state is thought to maintain low transcription of genes responsible for lineage commitment (Bracken and Helin, 2009). While our study did not query chromatin states directly, the results presented here support previous observations that the methylation and expression profiles of Basal breast tumors share a number of similarities with embryonic stem cells; namely that PRC2 activity is required to maintain cell fate plasticity (Ben-Porath et al., 2008; Holm et al., 2010; Bloushtain-Qimron et al., 2008; Dedeurwaerder et al., 2011).

Our results further show (Fig. 2·3) hypermethylation and silencing of genes involved in maintenance of cell-cell junctions and cell polarity in Basal tumors, which form the Basal gene co-methylation module. DNA methylation of the E-cadherin (*CDH1*) promoter, which belongs to the Basal gene co-methylation module, has been shown previously to be related to epithelial-to-mesenchymal transition in breast cancer (Lombaerts et al., 2006). A clue as to the relationship between PRC2 activity, which is increased in Basal tumors, and repression of *CDH1* activity comes from a previous observation that PRC2 complex is required for *CDH1* repression (Herranz et al., 2008). Specifically, Herranz et al. demonstrated that SNAI1, a transcriptional repressor, is responsible for PRC2 recruitment to the *CDH1* promoter. Our own results show increased expression of *SNAI1* in Basal tumors (Fig. 2·4). Additionally, the Basal gene co-methylation module is enriched for likely down-stream targets of SNAI1.

Our laboratory (Killian et al., 2011) and others (Roll et al., 2008; Fang et al., 2011; Van der Auwera et al., 2010) demonstrated that DNA methylation patterns could be indicative of disease outcome. Killian et al. devised a methyl deviation index (MDI), calculated relative to terminal ductal-lobular unit baseline, and used it to show that ER+ high-grade and short-survival cancers had a significantly higher MDI score compared to ER+ low-grade and long-survival cancers. Work by Fang et al., Roll et al., and Van der Auwera et al., attempted to characterize a CpG island methylator phenotype (CIMP), which was first described in colorectal tumors (Toyota et al., 1999). While work by Roll et al., and Van der

Auwera et al. concluded that the CIMP phenotype correlates with poor prognosis, work by Fang et al. suggests the opposite. Intriguingly, a significant number of genes identified by Fang et al. as part of their CIMP signature are part of the Luminal B gene co-methylation module described here. These apparent discrepancies should be addressed in future studies with higher density DNA methylation profiling platforms.

In the current study, we show that methylation of the Luminal B gene co-methylation module is associated with poor prognosis (Fig. 2·7). While the biological mechanism that leads to this phenomenon remains to be elucidated, here, as in previous studies (Ku et al., 2008), we present evidence of a relationship between Polycomb-group protein targets and CpG island methylation.

In summary, our study paints a comprehensive portrait of CpG island methylation in breast cancer. Importantly, the results highlighted here were achieved using an unsupervised systems approach applied exclusively to DNA methylation data. This approach, based on analysis of weighted gene networks (Horvath et al., 2006), enabled us to reduce the multiple hypothesis testing problem inherent to these types of datasets while making as few assumptions as possible.

2.4 Materials and Methods

2.4.1 Study population

The study population has been described previously (García-Closas et al., 2006), and is derived from a population-based breast cancer case-control study conducted in Poland. Eligible subjects included women between ages 20 and 74 years who resided in Warsaw or Łódź, Poland, in 2000-2003, and a total of 2,386 cases (79% of eligible) and 2,502 controls (69% of eligible) participated in the study. Information on breast cancer risk factors was elicited through a personal interview. Histopathologic features, including histology, grade, tumor size and axillary lymph node metastases, were assessed using surgical pathology reports and independent evaluation by the study pathologist. At the Warsaw Cancer Center only, an additional set of breast tumor tissues were collected and snap frozen in

liquid nitrogen. In this study, 226 frozen tumors were selected from cases who had invasive breast cancer, had no treatment prior to surgery, and had tumor tissues constructed on TMAs. Compared to cases not included in the analyses, the 226 cases were more frequently larger and node positive tumors; other tumor characteristics (histology and grade) and breast cancer risk factors (age, age at menarche, age at menopause, parity, family history of breast cancer, BMI, previous breast disease, mammogram screening, etc.) were not significantly different. Of the original 226 samples, 208 were of sufficient quality for DNA methylation and gene expression microarray analysis.

2.4.2 Sample handling

All tumor samples were stored in liquid nitrogen (-196 C). Derivative RNA and DNA samples were bar coded, tracked, and inventoried using a barcoding system linked to a queryable, Research Information Management System, (LabMatrix, BioFortis, Inc).

2.4.3 RNA isolation, labeling, and microarray hybridization

RNA was isolated from frozen tumors using TRIzol reagent (Invitrogen, Carlsbad, CA) and Qiagen RNAeasy Mini columns. Approximately 30 mg of frozen tissue was shaved into 350 microliters TRIzol solution and then homogenized 2 x 3 minutes at 25 Hz using the Retsch Qiagen TissueLyser (Qiagen, Valencia, CA). Following TRIzol RNA isolation using the manufacturer‘s instructions, the samples were purified using Qiagen RNAeasy Mini columns. Briefly, following centrifugation of the supernatant through the Qiagen mini column, 700 μl of RW1 buffer was added to the column and centrifuged 15 seconds at 10,000 x g followed by two washes with 500 μl of RPE buffer. The RNA was eluted with RNase-free water. RNA quantity and integrity was assessed using Nanodrop Spectrophotometry (Thermo Scientific, Waltham, MA) and 6000 Nano LabChip Kit on Agilent 2100 BioAnalyzer (Agilent, Santa Clara, CA), respectively. Two hundred fifty nanograms of input RNA was amplified and labeled using the Illumina TotalPrep RNA Amplification kit (Applied Biosystems/Ambion, Austin, TX), following the manufacturer‘s recommended protocol.

The biotin-labeled cRNAs were quantitated using RiboGreen RNA Quantitation reagent (Molecular Probes, Eugene, OR) and 750 ng was hybridized to Illumina HumanRef-8 v2 Expression BeadChip microarrays (Illumina, San Diego, CA). BeadChips were scanned in an Illumina scanner. Data are deposited with NCBI under GSEXXXXX.

2.4.4 DNA isolation

DNA was isolated following standard phenol/chloroform procedures. Briefly, 30 mg of frozen tissue was shaved into 400 μl digestion buffer (0.1 M NaCl, 0.01 M Tris, pH 8.0, 0.025 M EDTA, pH 8.0, and 0.5% SDS) containing 0.1 mg/ml proteinase K. Samples were incubated overnight at 50 C with gentle rocking and purified using phenol:chloroform:isoamyl alcohol (25:24:1) and PhaseLock Gel (heavy) tubes. DNA quality and quantity was assessed using agarose gel electrophoresis and PicoGreen dsDNA Quantitation Kit (Molecular Probes, Eugene, OR). Upon bisulfite conversion, DNA methylation status of 25,578 CpG probes was assayed using the Illumina Methylation27 bead-array.

2.4.5 Gene expression microarray data analysis

Expression data were processed in the R statistical environment. Data were normalized using the lumi package (Du et al., 2008). Classification of expression categories was done using the random forests method. A subset of 40 samples, which closely fit the Sorlie categories (Sø rlie et al., 2001), was identified by hand-curation using 20 members of the "intrinsic" gene list which performed well on the Illumina arrays. These were used to train the randomForest algorithm which was then used to classify the entire data set.

2.4.6 DNA gene co-methylation network modules

Methylation data were processed in the R statistical environment (v. 3.0.2). CpG probes located outside CpG islands, as annotated by the manufacturer, were excluded from the analysis. A Pearson's correlation matrix was constructed from gene methylation profiles of 25,006 CpG methylation probes across 208 tumor samples. Gene methylation network

construction and module detection were achieved automatically using the blockwiseModules function, available in R package WGCNA (v. 1.34) (Langfelder and Horvath, 2008). Default values were used as inputs for blockwiseModules function, with the following exceptions: the TOMType was set to "none", mergeCutHeight was set to 0.25, and maxBlockSize was set to 27000. Gene co-methylation modules were related to sample annotations by calculating a Pearson's correlation between an average methylation profile of each module (mean methylation level of module probes across the sample dataset) and sample annotation. Categorical sample annotations were converted to binary variables. The significance of the correlation was estimated using a Student asymptotic p-value (Langfelder and Horvath, 2008). Module preservation statistics were calculated using R function 'modulePreservation', part of the WGCNA package (v. 1.34), using default parameters (Langfelder and Horvath, 2008).

2.4.7 Gene enrichment

Curated gene sets were downloaded from the Molecular Signature Database / C2 collection (Subramanian et al., 2005). Hypergeometric distribution was used to model the null distribution of the number of genes of interest belonging to a gene category of interest, given the total number of genes and the total number of genes in the gene category of interest. The P-value associated with enrichment was calculated according to a one-sided test using the mid-P-value minimum-likelihood approach (Rivals et al., 2007).

2.4.8 Kaplan-Meier survival analysis

Methylation data were median polished and samples were separated into two groups according to the top-level branches produced by the application hierarchical clustering onto the methylation data. Kaplan-Meier analysis was performed using the R 'survival' package (v. 2.36-14).

2.4.9 K-L divergence

K-L divergence D_{KL} for every probe pair P and Q was calculated according to

$$D_{KL}(P \parallel Q) = D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P), \tag{2.1}$$

where, for every nucleotide i

$$D_{KL}(P \parallel Q) = \sum_i \log\left(\frac{P(i)}{Q(i)}\right) P(i) \tag{2.2}$$

and

$$D_{KL}(Q \parallel P) = \sum_i \log\left(\frac{Q(i)}{P(i)}\right) Q(i) \tag{2.3}$$

2.4.10 Gene set enrichment survival analysis

We fitted a Cox proportional hazards regression model to every probe on the methylation array, ranking the probes based on the Wald statistic P value (R 'survival' package v. 2.36-14). We then performed an analysis inspired by GSEA (Subramanian et al., 2005; Mootha et al., 2003) calculating enrichment score and a permutation-based P value of the Luminal B gene co-methylation module. Permutation P value was calculated as fraction of 1000 permutations where the enrichment score was greater than observed. At each permutation, follow-up time and status indicator were randomly shuffled among samples in the dataset.

Chapter 3

A chromatin structure based model accurately predicts DNA replication timing in human cells

(**Note: Chapter adapted from:**

Gindin, Y., Valenzuela, M.S., Aladjem, M.I., Meltzer, P.S. (2014) A chromatin structure based model accurately predicts DNA replication timing in human cells. (In press at Molecular Systems Biology.))

3.1 Introduction

In eukaryotes, DNA replication is a tightly regulated process that follows a strict temporal program (Masai et al., 2010; Taylor, 1960). This timing program is intimately associated with key aspects of cell biology, including cell differentiation (Hansen et al., 2010; Hiratani et al., 2010; Hiratani et al., 2004), cancer progression (Donley and Thayer, 2013; Fritz et al., 2013; Ryba et al., 2012), the 3D conformation of cellular DNA (Moindrot et al., 2012; Ryba et al., 2010; Ryba et al., 2012) and the formation of cytogenetic aberrations (De, 2011). Whereas the genome-wide replication program in eukaryotes appears nearly deterministic, the individual replication initiation events display a large degree of stochasticity (Bechhoefer and Rhind, 2012). An important step in resolving this apparent discrepancy was to recognize a formal analogy between DNA replication and nucleation in one dimension (Jun and Bechhoefer, 2005; Kolmogorov, 1937), which serves as the foundation for most of today's mathematical models of DNA replication. But while the molecular components of DNA replication modeled in this formalism are mostly conserved across the domains of life, one finds that the mechanism of recognition and regulation of initiation sites varies

greatly, even between lower and higher eukaryotes (Aladjem, 2007).

Particularly amendable to modeling are extreme examples of initiation site recognition: random and well-characterized. *Xenopus Laevis* is a representative of random initiation site selection. Modeling efforts for this organism, which need not take into account locations of initiation sites, have helped to provide theoretical answers to the so-called random completion problem (Blow et al., 2001; Herrick et al., 2002; Yang and Bechhoefer, 2008), and the global increase of the replication initiation rate throughout the S-phase suggested as one possible solution has later been confirmed experimentally and described as a universal feature across eukaryotic replication (Goldar et al., 2009). *Saccharomyces cerevisiae* occupies the other end of the initiation site recognition spectrum. Its quite well-characterized and efficient replication initiation sequence have helped to extract a number of parameters relevant for modeling efforts, such as the average and the variance of the firing time distribution for individual initiation sites. Based on such estimates, mathematical models were able to reproduce the global timing program found in yeast (de Moura et al., 2010; Lygeros et al., 2008; Yang et al., 2010) thus demonstrating how the deterministic timing program emerges from individually stochastic initiation events. Initiation site selection in metazoan genomes lies somewhere between these extreme cases. While metazoan replication initiation occurs at discrete sites in the genome, the metazoan replicator remains relatively poorly characterized, as even the most efficient sites fire in only a fraction of cell cycles (Martin et al., 2011; Valenzuela et al., 2011). This makes it more difficult to directly observe location and amplitudes of initiation (Besnard et al., 2012; Martin et al., 2011) or to extract this information from replication timing data (Baker et al., 2012), contributing to the dearth of timing models for metazoan cells.

Beyond the technical difficulties of obtaining a comprehensive set of robust parameters for metazoan replication timing models, a model built around tuning a large number of parameters (at least one for each of the 100,000 estimated initiation sites (Pope et al., 2013) in human cells) would remain somewhat unsatisfactory. It would sidestep the question of what factors determine replication timing and could therefore not explain timing

plasticity. Moreover, parameters for such a model would have to be re-determined for every cell-state. To address these challenges, we built a minimal model of metazoan DNA replication and identified a genomic marker that can be utilized to predict, rather than reproduce, genome-scale DNA replication timing profiles at high resolution. Our replication timing model for the human genome predicts timing with an accuracy (Pearson's r=0.92) rivaling that of experimental repeats (r=0.94) performed in different laboratories. We use our model to demonstrate that (a) the replication timing program can be accounted for by the approximate location of initiation sites alone, regardless of other factors such as exact initiation probabilities and that (b) initiation sites are optimally localized by DNase hypersensitive (HS) sites.

3.2 Results

3.2.1 Mechanistic model of DNA replication

The focus of this study was to understand and predict the dynamic DNA replication timing program of human cells. Here we took a reductionist modeling approach, including only essential components while omitting all features not required to model the timing program. In the resulting model (Figures 3·1, 3·3 and 3·2), a number N of rate-limiting factors independently select genomic locations and initiate replication (if the location has not yet been replicated) with probabilities specified for that location by an initiation probability landscape (IPLS). Thus, the probability of replication initiation at a given genomic location x is the product of the probability of a rate-limiting factor selecting one of the unreplicated competent initiation sites at time t, the initiation probability assigned to that location by the IPLS and the number of available (unengaged) rate-limiting factors at time t.

Since the result of each simulation is determined by the choice of the input IPLS, the biological question of what determines the DNA replication timing program can be addressed by identifying the IPLS that most accurately predicts experimentally observed data. Here, human replication timing data published in (Hansen et al., 2010) and (Ryba et al., 2012) were used for this benchmark. Both datasets report the average behavior of cell

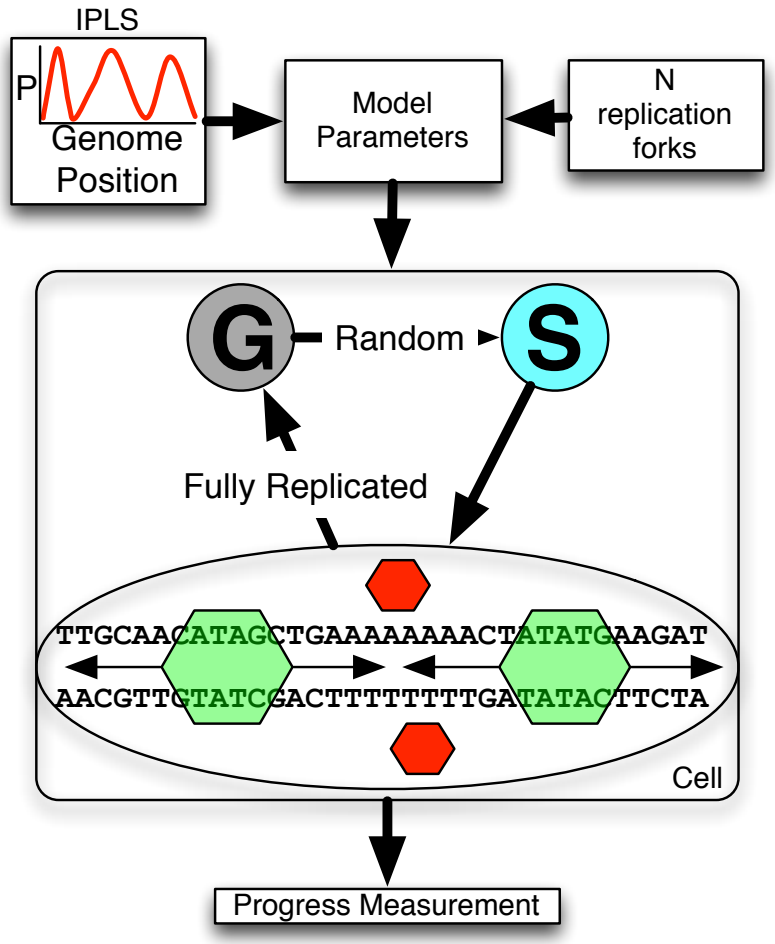

Figure 3·1: Mechanistic model inputs are Initiation Probability Landscape (IPLS) and the number N of replication forks. The DNA replication program is executed on a simulated cell population (a single cell is depicted). Simulated cells can be either in a non-replicating state (denoted as "G") or a replicating state ("S"). At the start of the simulation all cells are in the G state. Transition from G to S occurs randomly. When in the S state, free (red) rate limiting forks select a random location and bind with a probability set by the IPLS or remain unengaged otherwise. Once engaged (green), replication occurs bi-directionally until forks collide returning to their unengaged state, restarting the process until the genome is replicated. The model periodically queries each cell's replication progress. Once the genome is replicated, the cell enters G state, repeating the process until simulation is terminated.

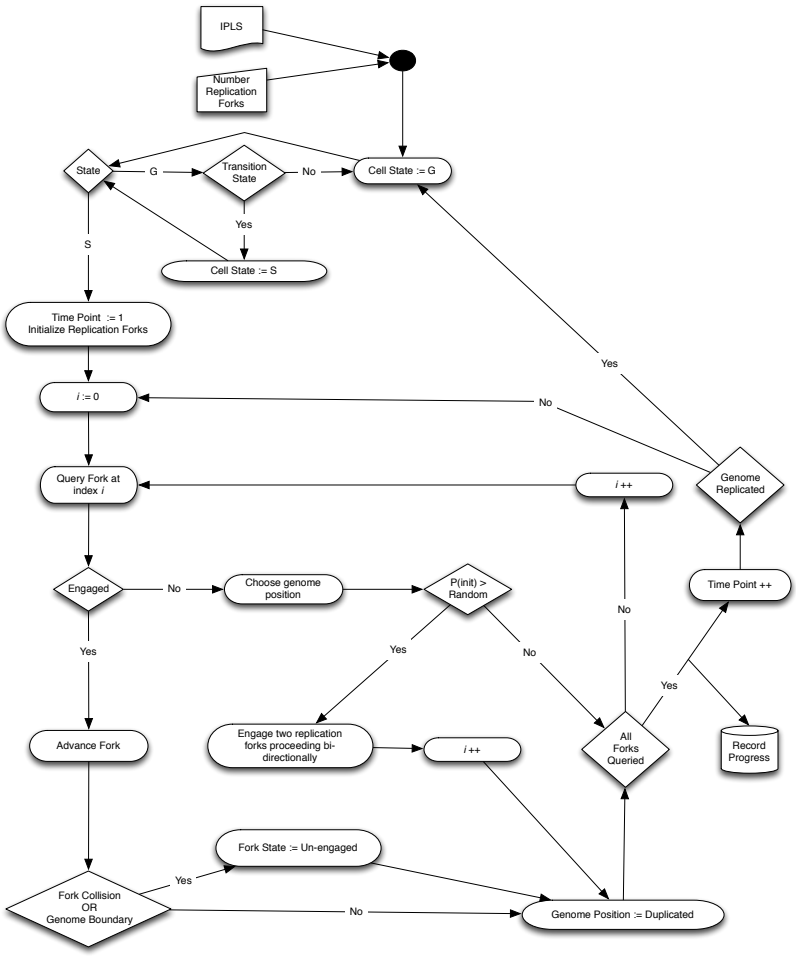

Figure 3·2: Model inputs are the number of replication forks and the IPLS to specify the probability of initiation of replication at given positions in the genome. The model transitions from G state (resting) to S state (replicating) at random. Once in the S state, the state of the replication forks are queried in an arbitrary sequential order. If the fork is not engaged, the model selects a random un-replicated position on the chromosome and initiates replication at that position with a probability that is assigned by the IPLS for that position. If DNA replication is initiated, then two replication forks are engaged and assigned to move in opposite directions. All forks move one step at a time, disengaging when they collide or reach the chromosome boundary. Model's progress is recorded periodically. Model exits the S state when all DNA is replicated.

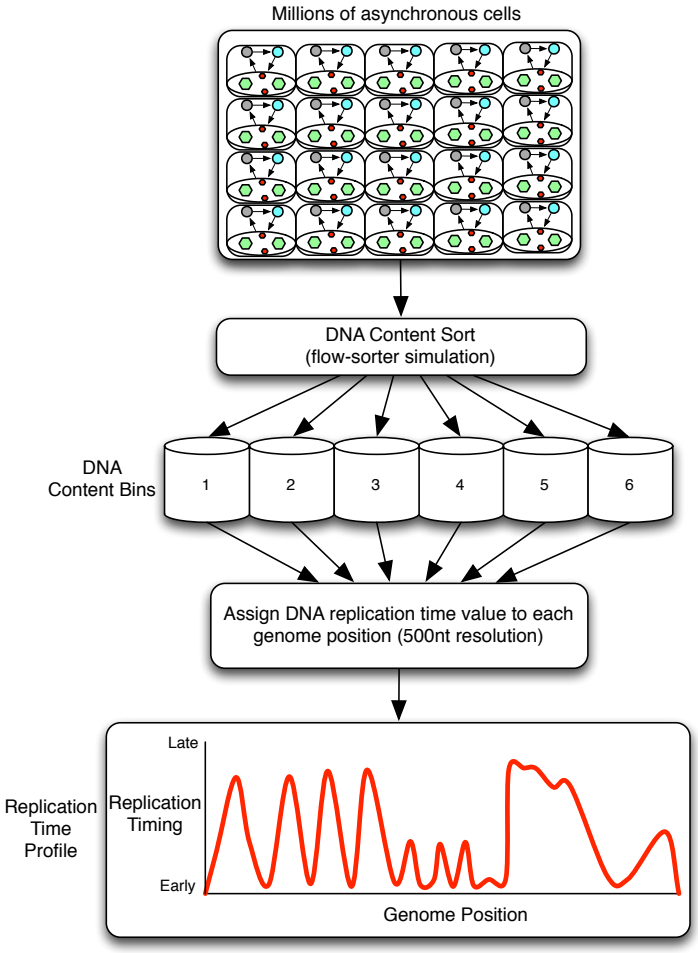

Figure 3·3: The simulation consists of millions of asynchronous cells. In order determine the DNA replication timing profile, the cell population is first separated according to DNA content into one of six bins. DNA replication time is calculated for each genomic position (500bp resolution) by taking the average of the product of DNA content bin number and the number of occurrences of the genome coordinate in each bin. For visualization, DNA replication timing is plotted as a function of genome position.

populations. We compared our model's prediction, averaged over millions of Monte Carlo simulated cell cycles, to these benchmark datasets. The concordance between predictions generated by the optimal model (see below) and the experimental data is striking at the 500bp resolution used in our simulations (Figures 3·4 and 3·5), recapitulating peaks and valleys of replication timing on a chromosome-wide scale (Figure 3·6).

3.2.2 Predictive power of static genomic features

Earlier studies (Cayrou et al., 2011; Martin et al., 2011; Valenzuela et al., 2011) had indicated that DNA replication initiation is more likely to occur in the vicinity transcription start sites (TSSs). Thus, as a starting point, we tested the predictive capacity of an IPLS where we assigned a constant, time-independent high initiation probability to all TSSs annotated in RefSeq (Pruitt et al., 2005) and low probabilities everywhere else (see Supplement for further discussion). Despite the simplicity of these assumptions, the resulting timing prediction is quite similar (average r=0.69 across four cell lines Figure 2A) to the experimental data. Testing other sequence features that were previously associated with replication initiation generates similarly good predictions: CpG islands (Meyer et al., 2013) (r=0.64), GC content (Meyer et al., 2013) (r=0.58), and predicted G4-quadruplexes (Besnard et al., 2012) (r=0.55) (Figures 2A and S3). Remarkably, an IPLS based on a structural feature of the DNA molecule, namely its solvent-accessible surface (Greenbaum et al., 2007), produced profiles (r=0.51) only slightly less predictive than some of the other, more commonly discussed factors (Figures 3·7 and 3·8). However, such invariant properties of the genome cannot account for timing plasticity observed across cell types (Hansen et al., 2010). We therefore hypothesized that dynamic genomic landmarks would generate models better suited to capture differentiation lineage-specific timing plasticity.

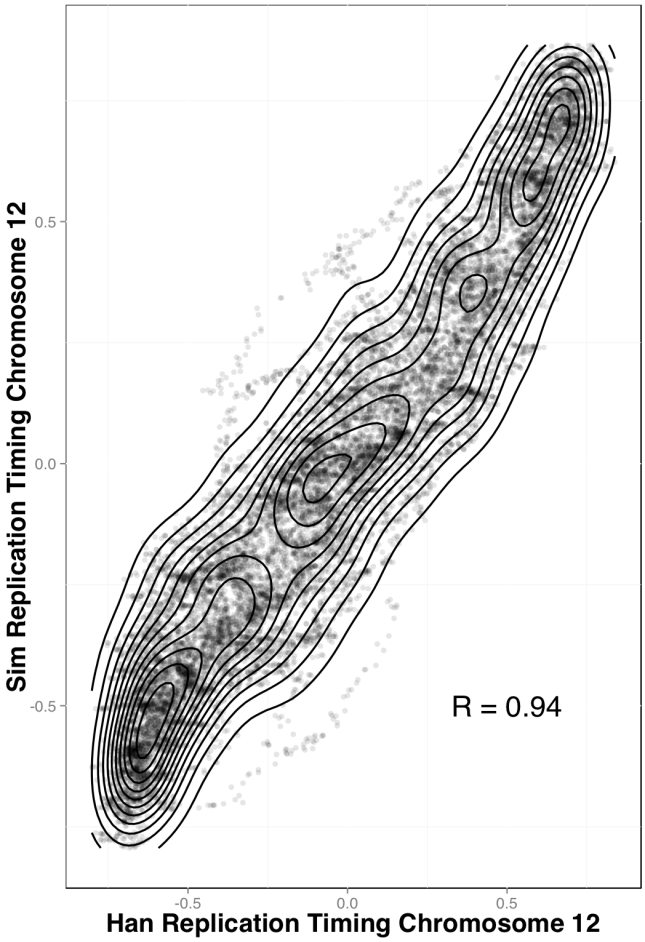

Figure 3·4: Simulated and empirical DNA replication timing are highly correlated. Each point in the contour plot represents a replication time assignment for a 500nt bin on chromosome 12 of GM06990 cells. Simulated replication timing assignment is given on the y-axis and the experimentally derived assignment is given on the x-axis. Contour lines are meant to aid in interpretation. R value represents Pearson's correlation between simulated and empirical data.

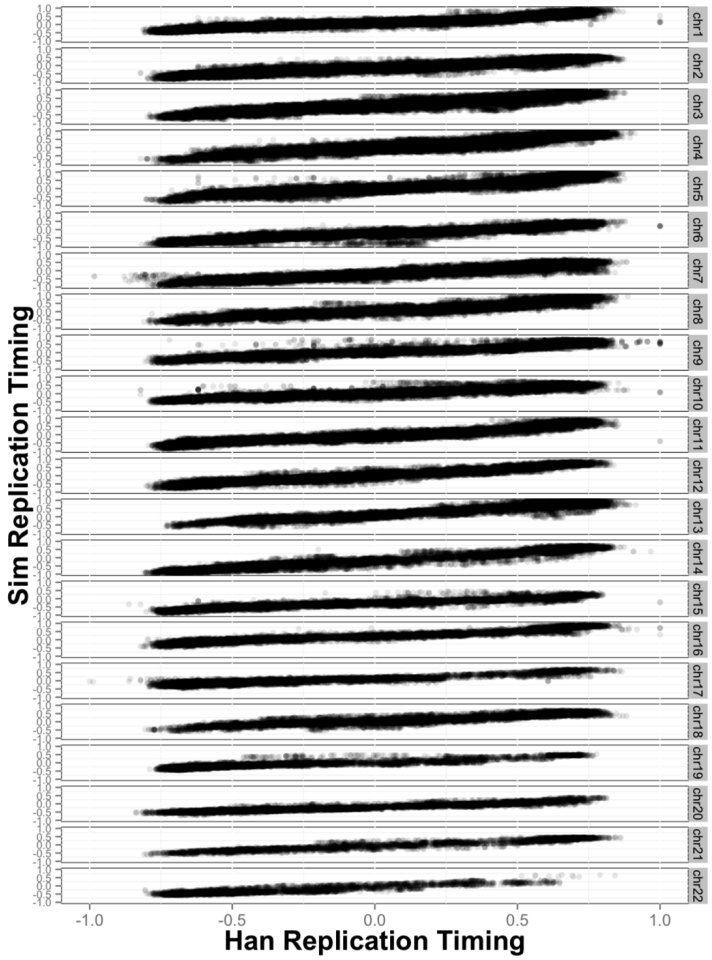

Figure 3·5: Each point represents a replication time assignment for a 500 nt bin for 22 autosomal chromosomes of GM06990 cells. Simulated replication timing assignment is given on the y-axis and the experimentally derived assignment is given on the x-axis. Contour overlay is meant to aid in interpreting plot density.

Table 3.1: Top DNA Replication Timing Predicting IPLS Sources

IPLS Source	Average Correlation
DnaseDgf	0.865
CCNT2	0.855
JunD	0.855
FaireSeq	0.854
ZNF384	0.849
COREST	0.849
CEBPB	0.847
MAZ	0.842
TBLR1	0.839
eGFP-JunD	0.835
ZNF-MIZD-CP1	0.834
H3K9acB	0.829
H3K4me2	0.829
HCFC1	0.828
UBTF	0.828
HMGN3	0.828
BHLHE40	0.827
TBP	0.827
DnaseSeq	0.825
H3K4me1	0.824

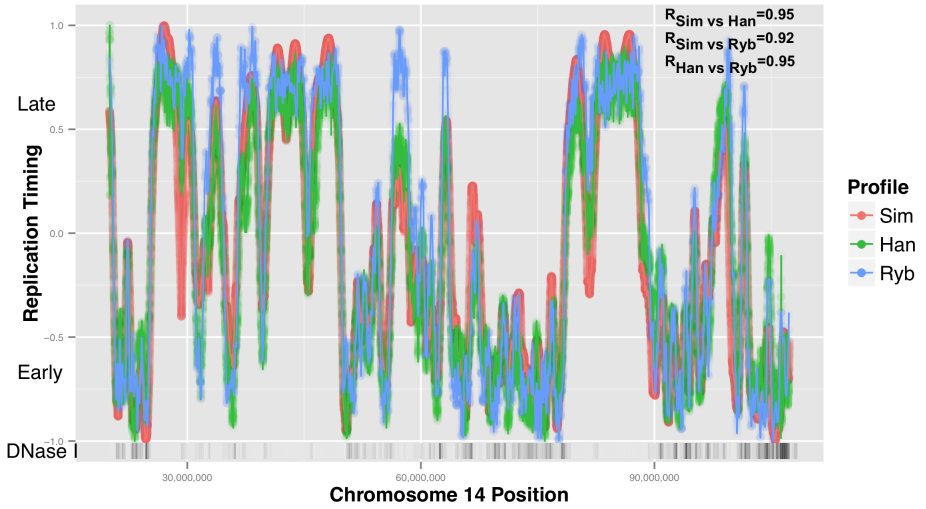

Figure 3·6: Simulation based on DNase HS sites produces high-fidelity replication timing predictions. The simulated timing program (red) generally lies on or between two experimental datasets plotted on the same axes, the Hansen (Han) dataset (red) and the Ryba (Ryb) dataset (blue). The stated correlation R values are specific to chromosome 14.

3.2.3 DNase hypersensitive sites are the main determinants of DNA replication timing

Utilizing the recently published (Bec, 2011; Rosenbloom et al., 2013) ENCODE data, we generated IPLSs from all 167 cell-specific datasets available for the cell lines in the Hansen data by assigning an initiation probability proportional to the ENCODE amplitude, simulated the timing patterns and compared the results to the empirical DNA replication timing data for corresponding cells. Nearly one-half (77 out of 167) of the probed ENCODE marks produce better predictive models compared to the best (TSS based) static model (Table 3.1). Notably, the gene expression based model (AffyExonArray, r=0.75) did not show a measurably improved accuracy in comparison to the static TSS model (r=0.69). The top-ranking model (r=0.87) is based on an IPLS derived from DNase HS sites. This is followed by models derived from activating chromatin marks such as H3k9ac (r=0.83), H3k4me2 (r=0.83) or transcription factor binding (e.g. JunD r=0.86).

We hypothesized that the ability of more than one epigenetic mark to predict DNA

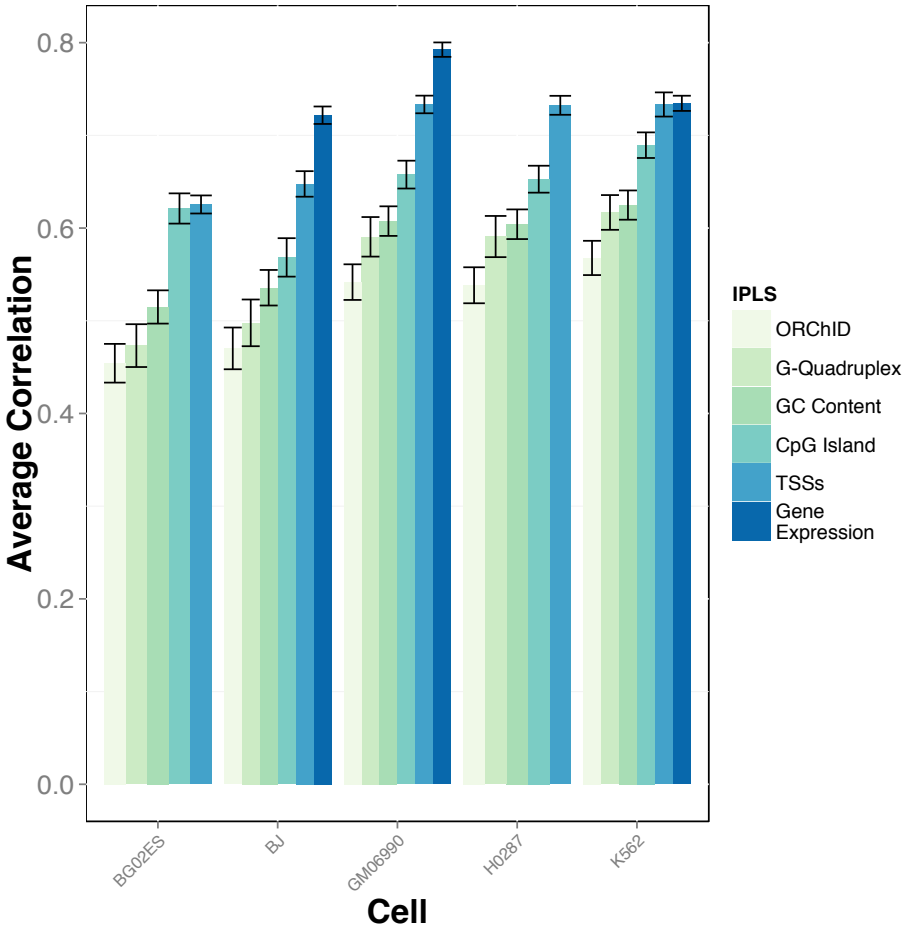

Figure 3·7: Simulations based on genome sequence features (GC content, CpG islands), or local genome conformation (ORChID, G-quadruplex), RefSeq annotated transcription start sites (TSS) and gene expression levels (where available) in five cell lines. Shown is the correlation with the Hansen (Han) dataset averaged over 22 autosomal chromosomes, error bars represent the standard error of the mean.

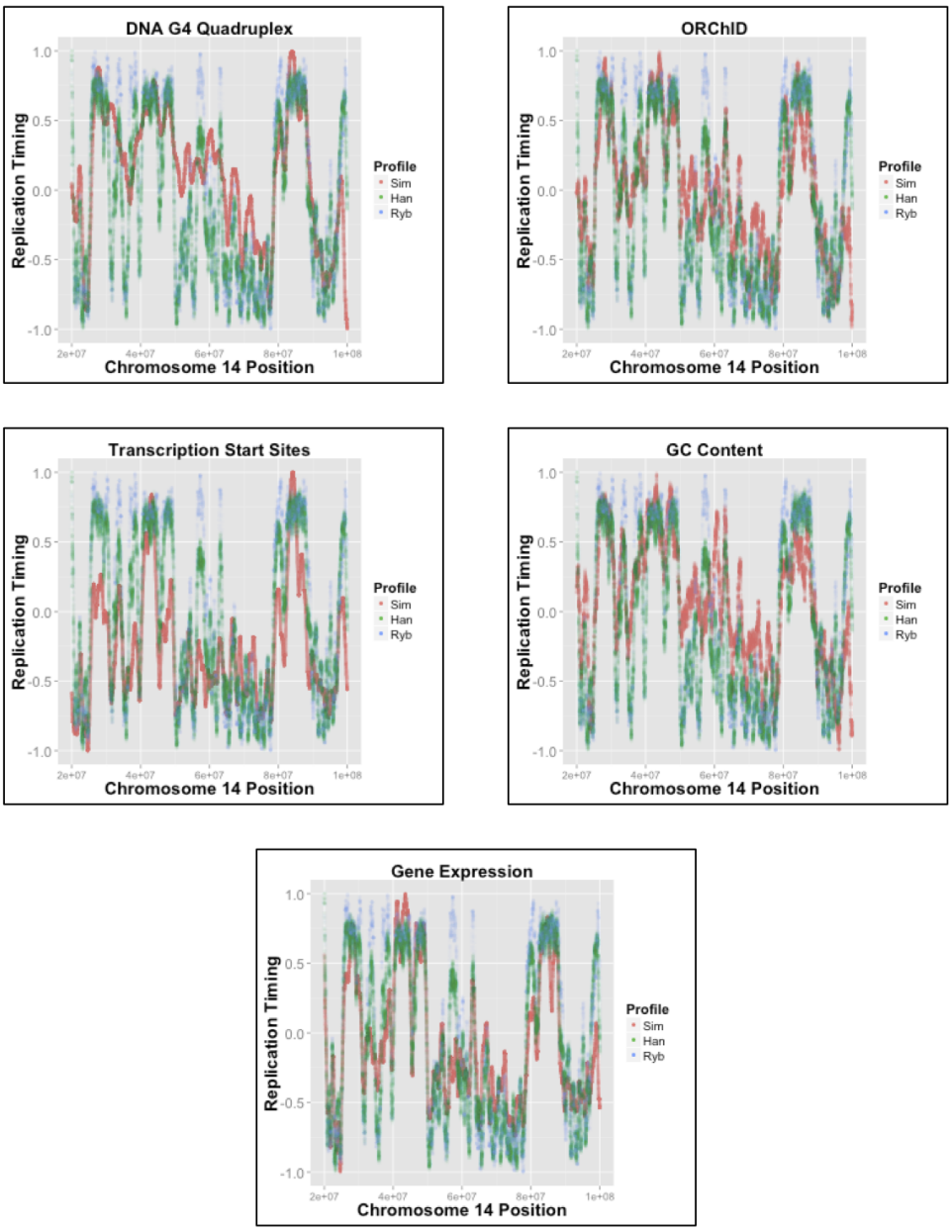

Figure 3·8: Experimental and simulated timing on chromosome 14 in GM06990 for sequence features and gene expression. Simulations based on genome sequence features (GC content, CpG islands), on local genome conformation (ORChID, G-quadruplex), and on organism-wide transcription start sites generate timing patterns that correlate well with replication timing in five cell lines. Simulated profiles (red) are plotted on the same axes as two empirical datasets: the Hansen (Han) dataset (red) and the Ryba (Ryb) dataset (blue).

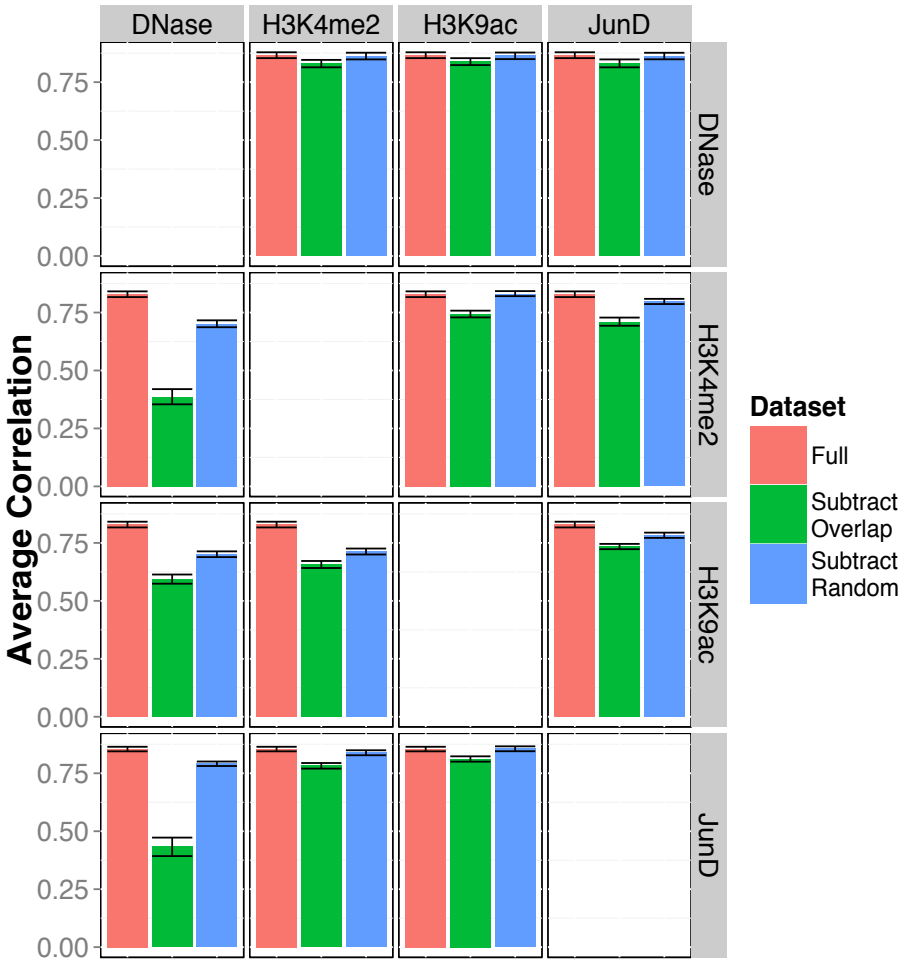

Figure 3·9: Mutual independence of representative top-ranking ENCODE marks (DNase, JunD, H3K4me2, H3K9ac) is probed by eliminating co-localized genomic marks in pairwise comparisons. The results of these 4 (datasets) $\times$ 3 (overlaps) = 12 sets of simulations are presented in a 4x4 matrix format: rows indicate the dataset that was used to generate the IPLS, columns indicate the subtracted dataset. Each panel plots the correlation to the experimental timing data in K652 cells (the only set for which all annotations were available) for the full dataset (red), the non-co-localized marks (green) and a random dataset (blue) from which the same number of (not necessarily overlapping) marks was removed. Error bars represent standard error of the mean.

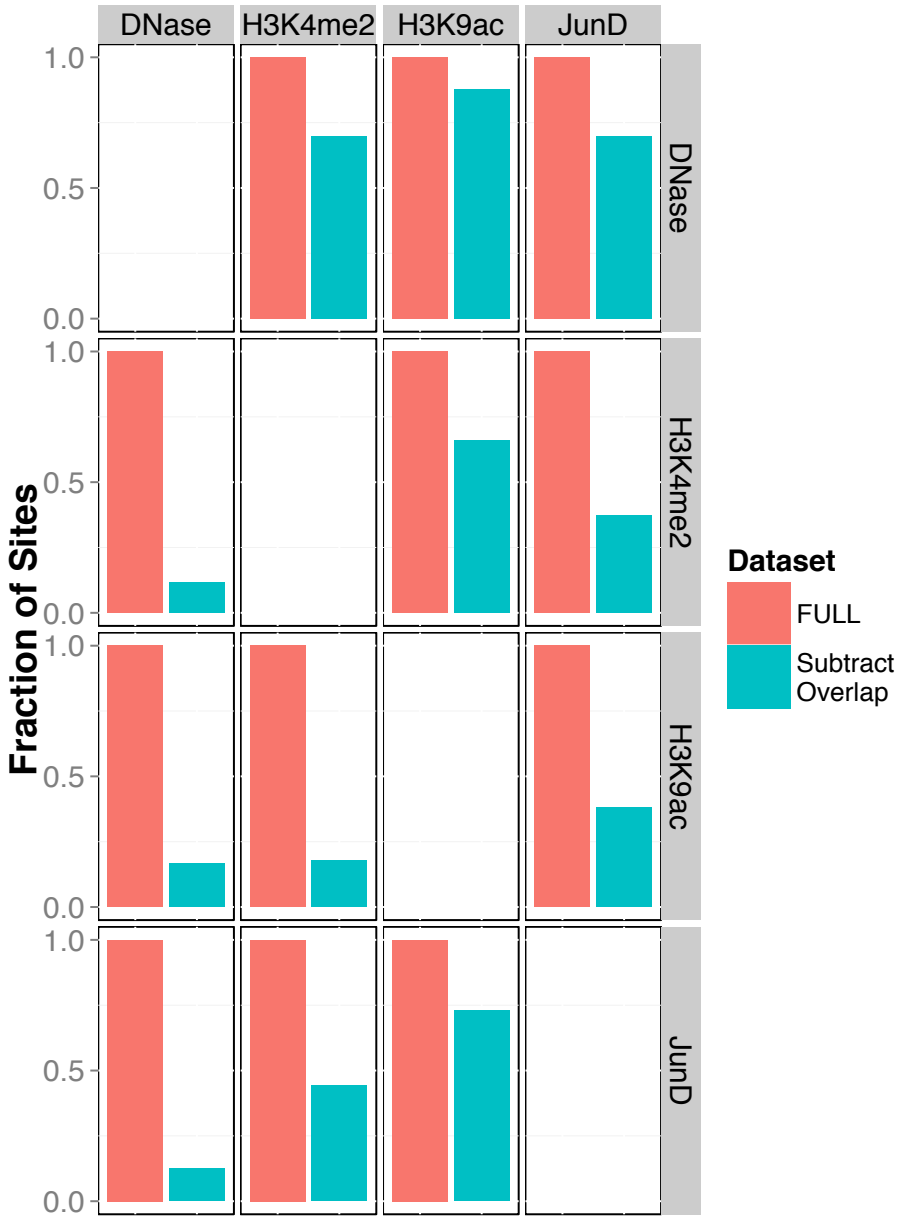

Figure 3·10: Using the datasets presented in Figure 3·9, rows indicate the dataset and columns indicate the effect of subtracting a dataset. Each vertical bar represents the fraction of genomic regions in the dataset: original dataset (red) and reduced dataset (blue).

replication timing with high fidelity is a consequence of the fact that many chromatin marks tend to co-localize (Thurman et al., 2012) and that, in isolation, some marks would lose much of their predictive value. To test this possibility, we performed simulations based on reduced sets, where mutually co-localized marks were removed (Figures 3·9 and 3·10). Remarkably, among the tested top-ranking genomic marks selected for this analysis (histone H3K4me2, H3K9ac, transcription factor JunD and DNase HS sites) only DNase HS sites fully retained their ability to predict replication timing in all pairwise comparisons. For all other marks, accuracy of the timing prediction was substantially reduced when removing overlaps with DNase HS sites, even when accounting for the reduced set size. We further explored whether these same marks co-localize with empirically determined DNA replication initiation sites (Besnard et al., 2012). Our results show that JunD, H3K4me2, and H3K9ac sites overlap DNA replication origins only so long as they also overlap DNase HS sites (Figure S6). We therefore conclude, based on the available data, that DNase HS is the main independent determinant of replication timing. This conclusion is further supported by observing that almost half of the DNase HS sites in HeLa ($47\%, p < 10^{-6}, OR = 5.0$) and IMR90 ($47\%, p < 10^{-6}, OR = 3.8$) cells are located within 500 bases of empirically determined initiation sites (Besnard et al., 2012). Also, the non-trivial distribution of initiation sites across chromosomes described in (Besnard et al., 2012) is closely recapitulated by the distribution of DNase HS sites (Figures 3·11 and 3·13).

We show that most DNA replication timing-predictive ENCODE marks did not maintain a high level of precision when marks overlapping with DNase HS sites were removed. From this result, it is reasonable to expect to see differences in how strongly the complete and the DNase subtracted sets overlap with empirically determined replication initiation events. Besnard and co-workers (Besnard et al., 2012) determined the global probability of replication initiation events by large-scale sequencing of nascent DNA. The read-tag density for these data show (Figure 3·12)) clearly discernible maxima at the location of H3K9ac, H3K4me2 and JunD genomic marks. However, after removing marks overlapping with DNase HS sites, the signal is significantly diminished. Moreover, there is evidence

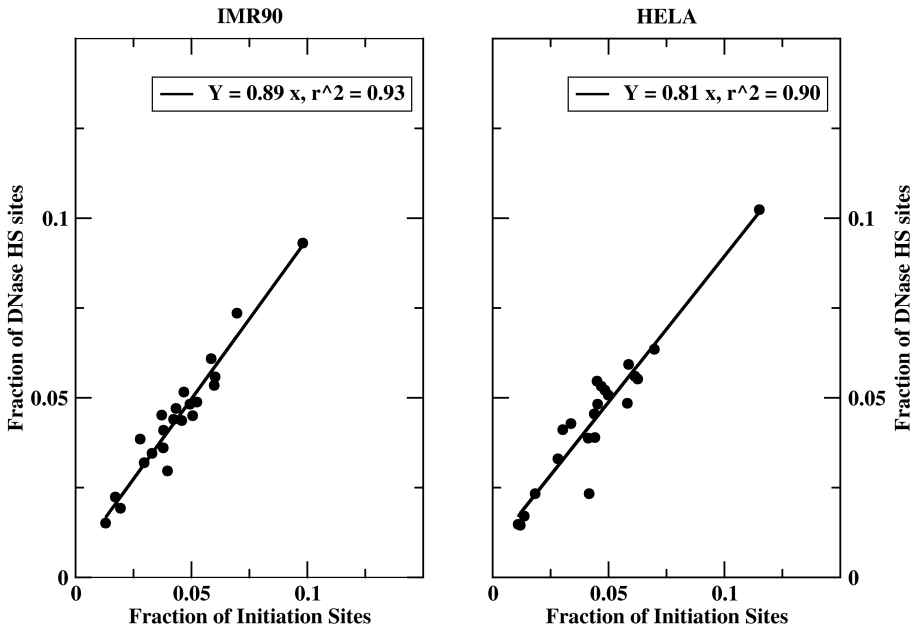

Figure 3·11: The number of initiation sites had been shown earlier to be non-trivially distributed across chromosomes (Besnard et al., 2012). Comparison of the number of DNase HS sites in IMR90 and HeLa with the number of initiation sites on each chromosome reveals a tight correlation between the two. Each data-point in the plot represents the fraction (sum = 1) of initiation and DNase HS sites, respectively, on a autosomal chromosome (see also Figure3·12).

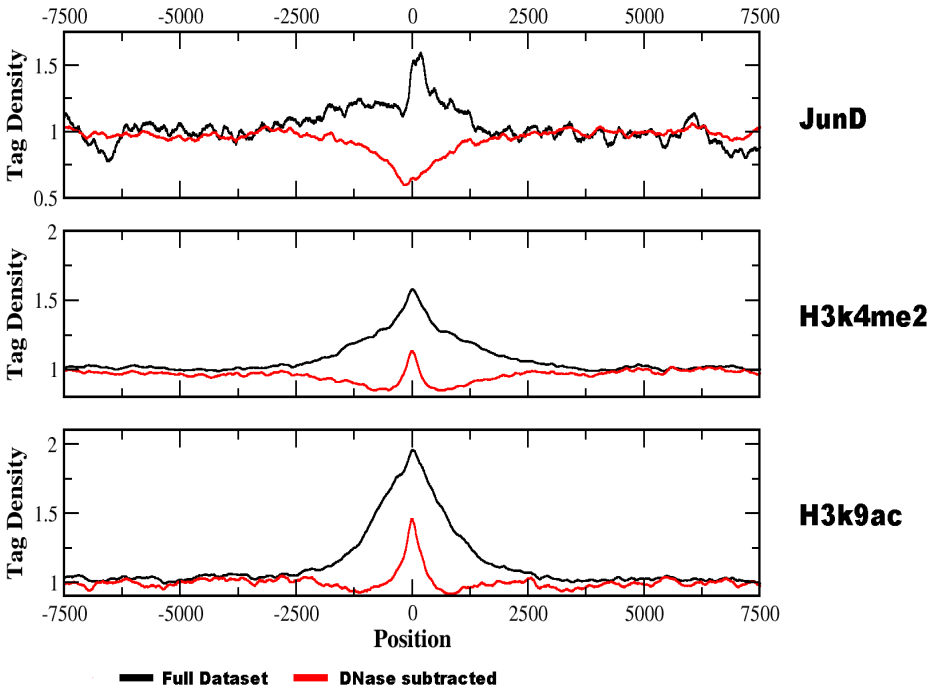

Figure 3·12: The tag density of initiation sequencing reads around JunD, H3k4me2 and H3k9ac sites are much lower for sites not overlapping DNase sites (red) than those that do (black curve). Plots are normalized to 1 at large distances. A value below 1, observed in the vicinity of all non-DNase overlapping marks, indicates suppression of replication initiation compared to the average genomic location.

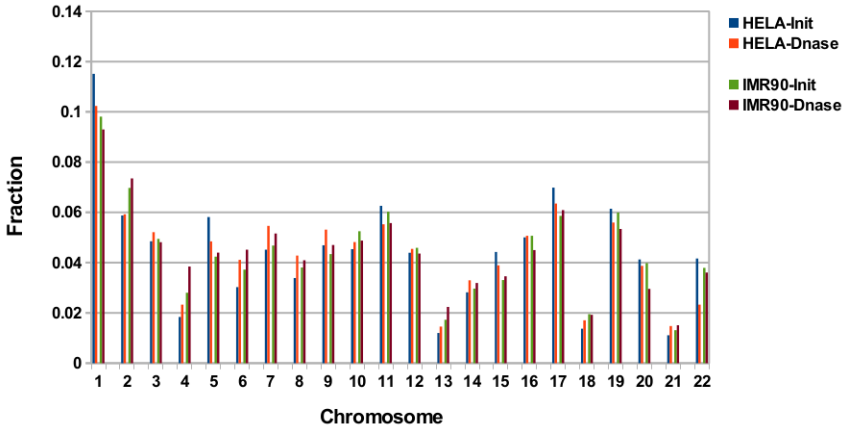

Figure 3·13: Shown is the fraction of the total number of DNase hypersensitive sites and replication initiation peaks (derived from (Besnard et al., 2012)) across all autosomal chromosomes.

of a weak suppression of initiation events at the remaining genome marks. These results reinforce our findings that DNase HS sites are the main independent determinants of DNA replication timing.

3.2.4 DNA replication timing plasticity across cell lineages and species and its alteration as a result of chromosomal fusions

Replication timing shows remarkable plasticity across differential lineages (Donley and Thayer, 2013) and in cancer cells (Ryba et al., 2012). Utilizing DNase HS data for three cell lines (BJ, GM06990, K562), for which matching experimental timing and DNase HS data were available, we performed DNA replication simulations and hierarchical clustering of the simulated and experimental data (Figure 3·14). The model predictions tightly cluster with the experimental data for the matching cell and also recapitulate the closer relatedness of GM06990 and K562 cells (both of hematopoietic origin) in comparison to BJ (fibroblast). Using stringent parameters (see Methods), we identified 60 genes in regions with replication timing variable regions between GM06990 and K562 cells and found a significant enrichment for interferon and haemoglobin complexes (DAVID (Huang et al., 2009) P-value 3.3×10^{-12} and 2.3×10^{-10}, respectively), including the human β-globin

locus (Figure 3·15) in line with phenotypic properties of these cells.

The accuracy of our model predictions in human cells suggested that the same mechanism will likely work in other mammalian cells. Currently, the lack of simultaneous availability of both, replication timing and DNase HS data for the same cells, limits the ability for a broader analysis. To test the applicability of our model to mouse embryonic fibroblast cells, we compared replication timing predictions generated from DNase HS sites in NIH/3T3 cells (ENCODE Project Consortium and others, 2011) to observed timing data in MEF cells (Hiratani et al., 2010). The average Pearson correlation between model prediction and experimental replication data is 0.85 (Figure 3·16), confirming that our model can be extended to other metazoan cells.

Recurrent chromosomal fusions are found in many cancers (Delattre et al., 1992; Rowley, 1973; Tomlins et al., 2005). In acute lymphoblastic leukemia, the well-characterized t(12;21)(p13;q22); *ETV6-RUNX1* fusion is accompanied by an abrupt change in DNA replication timing near the fusion site (Ryba et al., 2012). Our model reproduces this behavior when inducing (see Supplement) an in silico t(12;21)(p13;q22) translocation in GM06990 lymphoblastoid cells (Figure 3·17). This behavior is also reproduced when comparing replication timing at the in-silico induced breakpoint in GM06990 cells with observed replication timing in REH cells, which harbor the translocation (Figure 3·18). The results show that replication timing is not determined at the site of the breakpoint. Instead, the timing pattern arises from the combined influence of the DNase HS sites situated on either side of the break. The discontinuity, observed experimentally and reproduced in the simulation, is the result of mapping physical coordinates of the rearranged chromosome 12 onto the normal genome.

3.2.5 Modeling parameters

The proposed model has remarkably few parameters. In addition to an IPLS and an optional technical variable (see below) there is only one adjustable parameter, namely the maximum number (N) of replication forks that can be active simultaneously. As N is

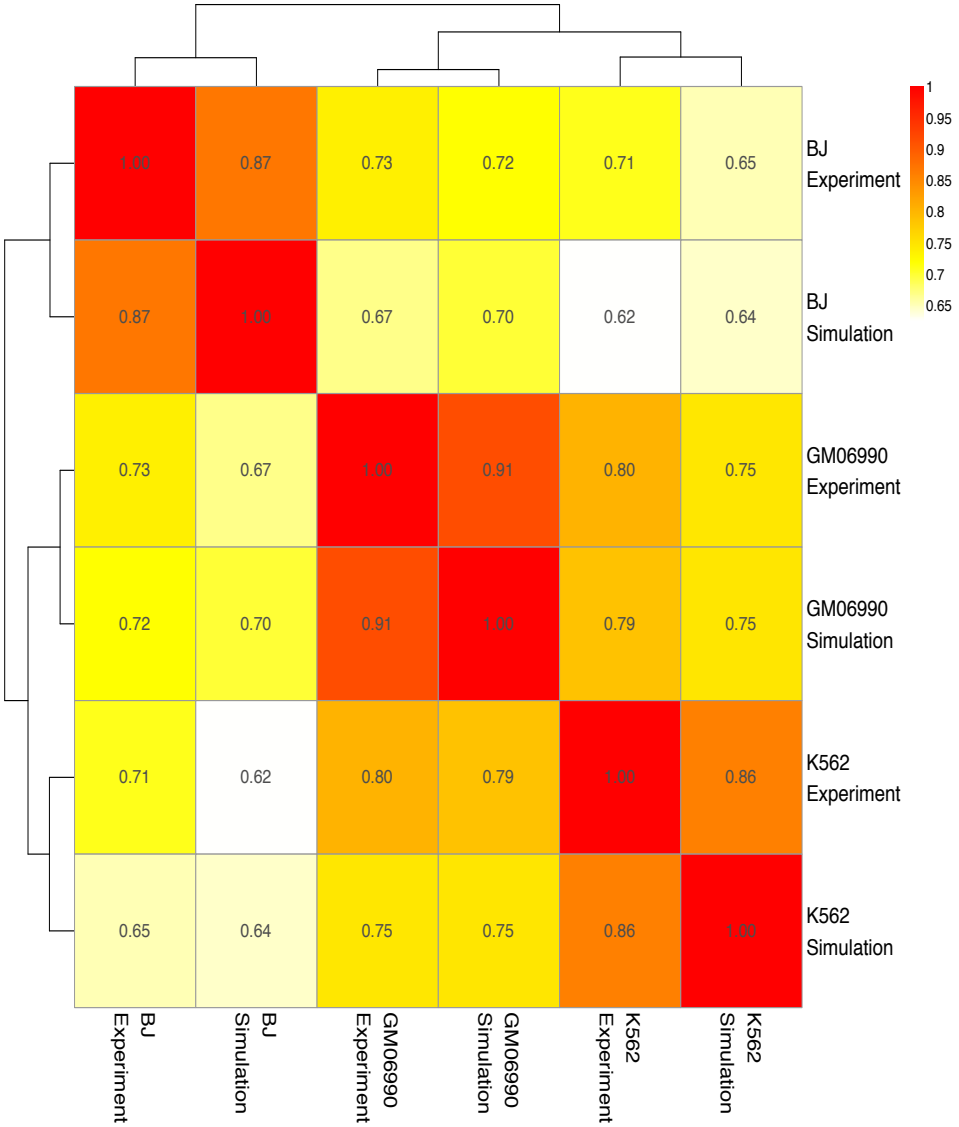

Figure 3·14: Hierarchical clustering and correlations heatmap of simulated and empirical data. Individual correlations (Pearson's) are noted in the matrix for every dataset pair. Simulations are consistently placed closest to the associated experimental data.

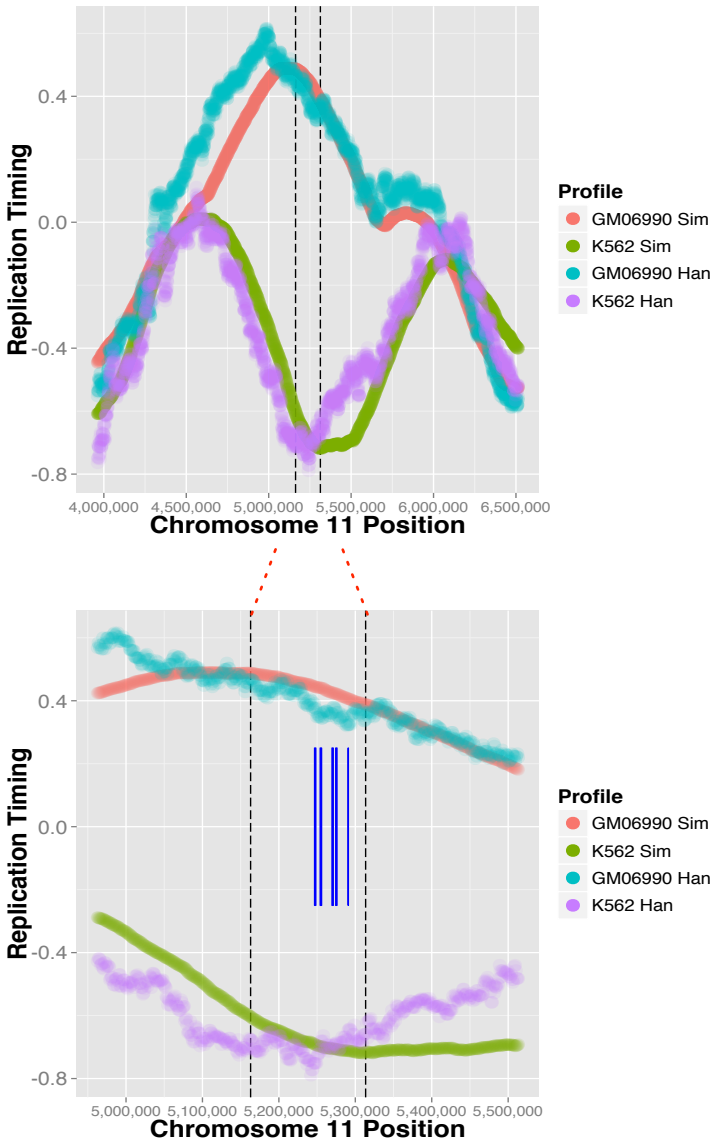

Figure 3·15: Analysis of timing plasticity between simulated GM06990 and K562 cells identified, among other regions, differential timing in the β-globin locus (indicated by dashed lines, genes marked in blue). Hansen dataset (Han) is shown for reference.

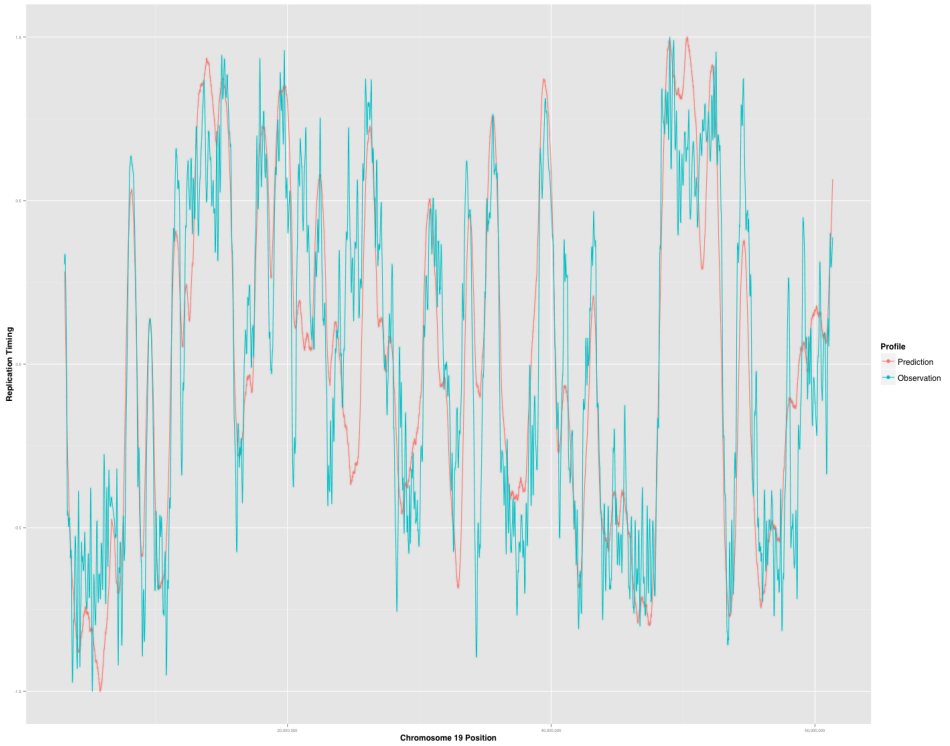

Figure 3·16: Plotted are simulated (red) and observed (blue) DNA replication timing (y-axis) profiles for chromosome 19 (x-axis) of mouse embryonic fibroblast cells. Simulated profile is based on DNase HS data for NIH/3T3 cells. Observed data is derived from MEF cells.

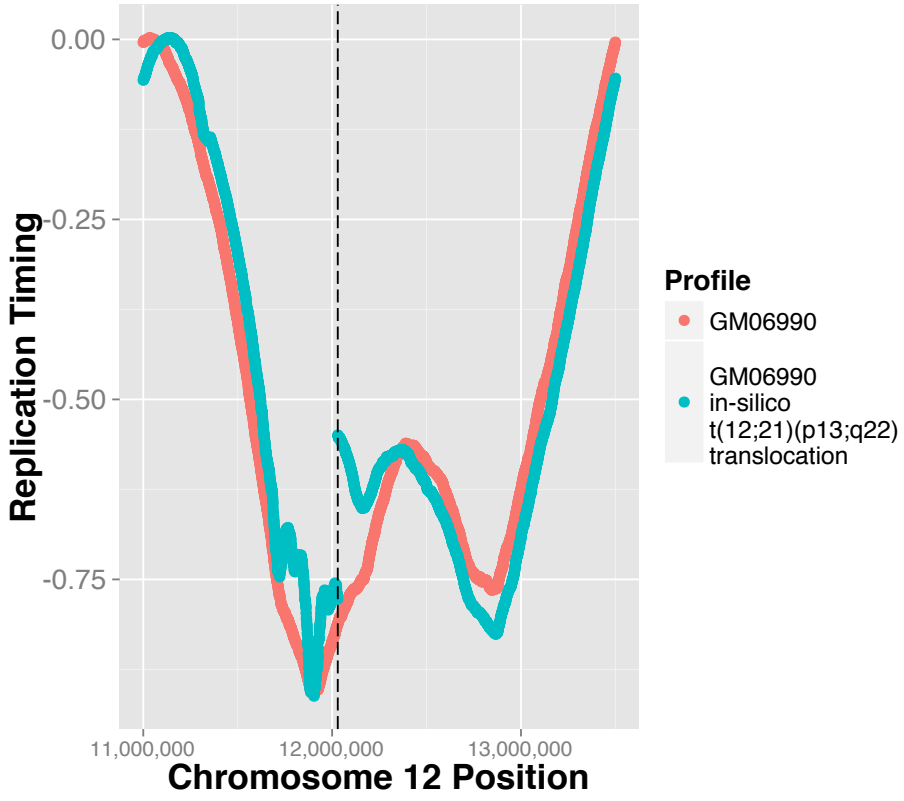

Figure 3·17: A translocation event simulated in silico in GM06990 cells qualitatively reproduces the timing discontinuity observed (Wiemels et al., 2000) at a *TEL-AML1* translocation in ALL. Replication profile of translocated (blue line) and normal (red line) are shown on the same genomic axis, the dashed line signifies the translocation coordinate.

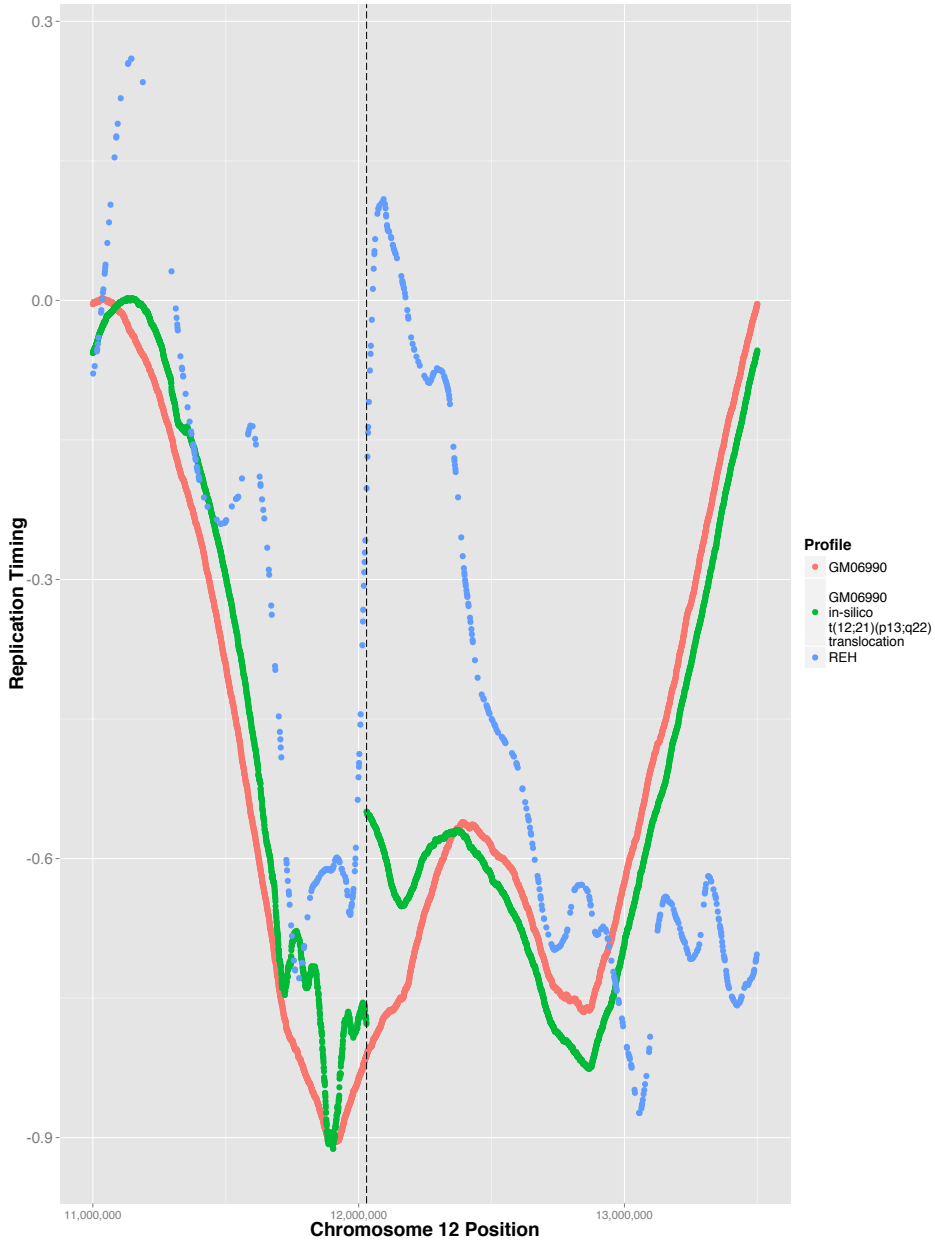

Figure 3·18: Simulated replication profile of in silico translocated (green line) normal (red line) are plotted on the same set of coordinates. Also shown is the experimentally observed replication timing profile of REH cells, which harbor the translocation (blue dotted line). The REH and GM06990 in silico translocated profiles show an abrupt change in replication timing from early to late at the breakpoint site (dotted vertical line).

not set a priori (Figure 3·19) we performed a series of simulations identifying, for each chromosome, the optimal N that generates the closest match to the experimental data. We find that the optimal N grows linearly with chromosome length at a rate of 1 fork per 1.3 mega bases (Figure 3·20), compatible with the assumption that the stochastic process governing replication does not substantially differ between chromosomes. Subsequently, we used the estimate from the linear regression curve in this experiment for N. With this setting, the predicted median length of the S-phase is 5965 (mean=6134) simulation steps (Figure 3·21) in GM06990 cells. In real human cells, replication forks move at a speed of about 50 bases per second. With a 500bp model-resolution (and two forks moving in opposite directions in each simulation step) the predicted median wall-clock time for the S-phase is t = 5965 steps * 500 bases / (50 bases / second) / (2 step) = 8.3 hours, in line with the experimentally observed duration of 6-10 hours. The optional technical variable mentioned previously governs the simulation of a flow-sorter, which in laboratory experiments divides asynchronously replicating cells into S-phase fractions (Ryba et al., 2011). As the actual gate settings used were not available(Hansen et al., 2010), numerical optimizations (see below) were used, further improving the similarity of the best (DNase HS based) simulated models from r=0.89 (with 5 equidistant gates in GM06990 cells) to r=0.92, a level approaching the limit of experimental noise (r=0.94 between experiments performed in different laboratories).

In the laboratory, replication timing is observed by BrdU labeling of un-synchronized cells followed by flow-sorting with respect to DNA content in order to assign a cell-cycle time (Ryba et al., 2011) to each cell. In one such work, Hansen and co-workers (Hansen et al., 2010) used 6 pre-defined time-slices termed G_1, S_1, S_2, S_3, S_4, G_2 (in increasing temporal order) separated by five adjustable gates (in addition to 0 and 1) during flow sorting. Naturally, these parameters, as well as the response curve of the flow-sorter, impact the estimate of the average replication times. In order to reduce the non-biological influence of the choice for gating parameters, we performed a simulated annealing optimization of flow-sorter gate settings for the top predictive model based on DNase HS sites. It

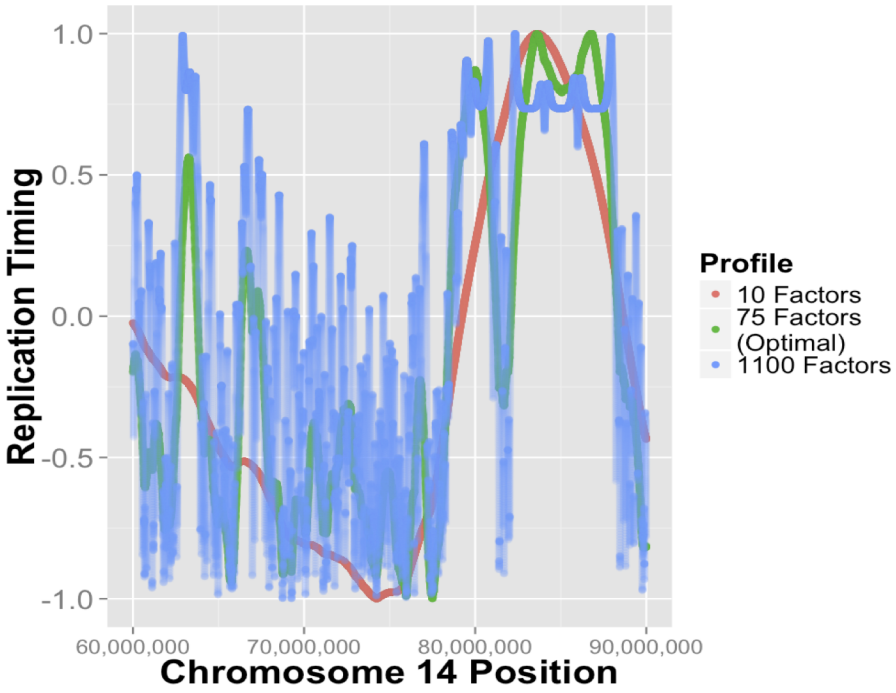

Figure 3·19: DNA replication timing was simulated for GM06990 cells based on DNase data using a range of the maximal number N of replication 'factors' for chromosome 14. Simulations with few (10 factors; red profile) produce smooth DNA replication timing profiles, owing to the long distance that each factor travels. Simulations with many (1100 factors; blue profile) produce rough landscapes owing to the short distance that each factor travels. The number of replication factors that produce best correlations (75 factors; green profile) grows linearly with chromosome size.

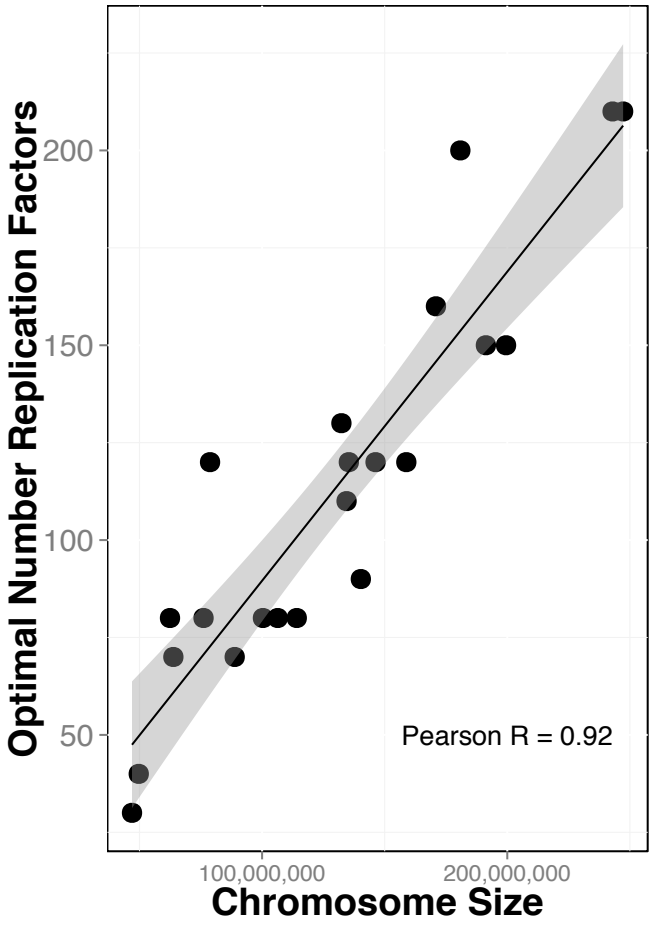

Figure 3·20: DNA replication timing profiles were simulated using IPLSs derived from GM06990 DNase HS data, noting the number of replication factors that produced highest correlation for each chromosome. Solid line represents a linear fit (shading area denotes the 95 confidence interval). The linear regression curve estimates that the number of forks per megabase is given by N = 10.24+7.9x10E-7x, where x is chromosome length. The Pearson correlation between the optimal number of replication forks and chromosome length is 0.92.

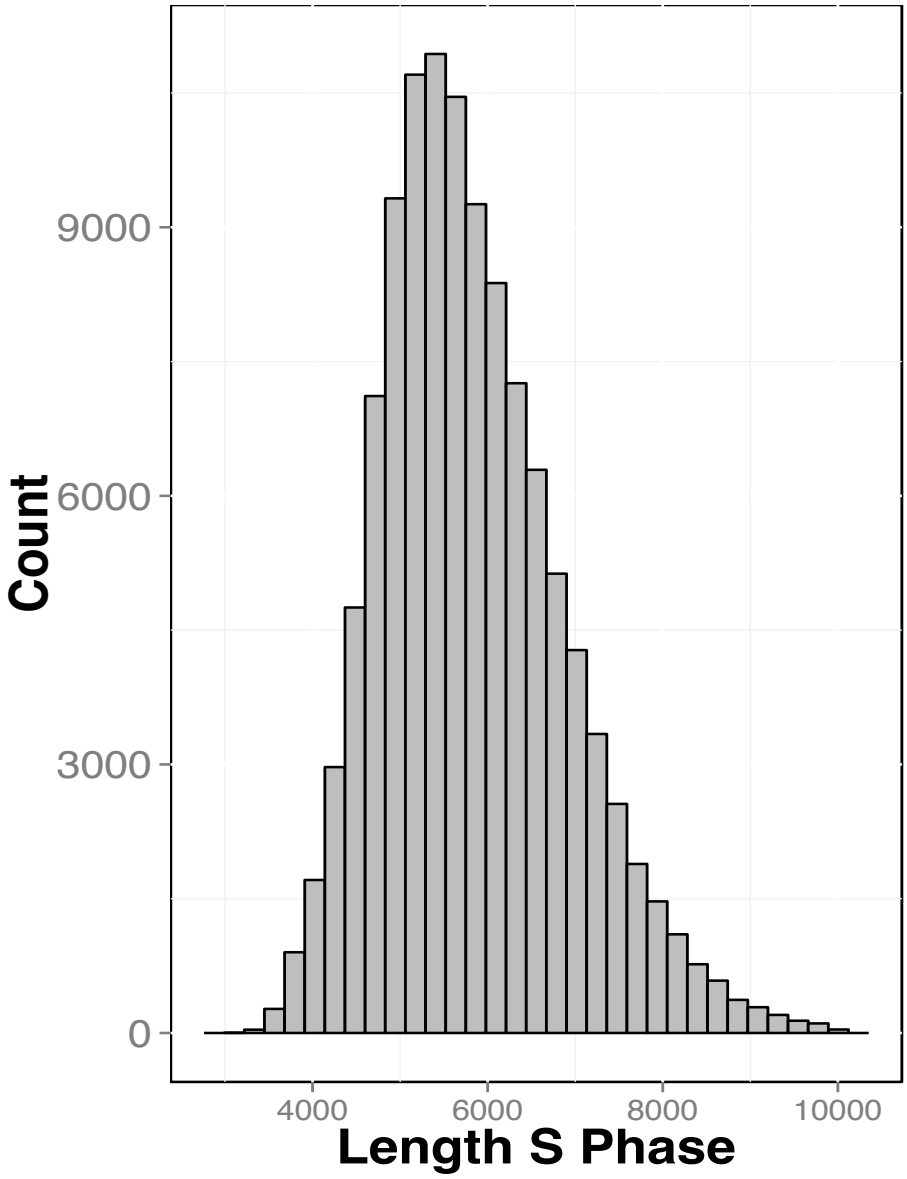

Figure 3·21: Histogram illustrating the distribution of the lengths of the S-phase in a simulated asynchronous cycling population of GM06990 cells.

improved the average correlation between experiment and simulation by 3% (from r=0.89 to r=0.92). At this level, our prediction approaches the reproducibility levels between experiments: we found that timing measurements for same cells line (GM06990) performed in different laboratories (Hansen (Hansen et al., 2010) and Gilbert (Ryba et al., 2012) datasets, respectively) were correlated at a level r=0.94, only marginally better than our best model prediction.

3.2.6 DNA replication timing is highly robust

With only a single adjustable biological parameter, and thus no real risk of over-fitting the data, the accuracy of our model predictions is exceptionally high, suggesting a high degree of robustness of the proposed model. One potential limitation is the completeness of genomic annotations. To test its importance, we built a series of models by randomly sub-sampling DNase HS annotations. The predictions were essentially unchanged despite removing up to 75% of DNase HS sites with the accuracy degrading gradually beyond this point (Figure 3·22). We conclude that the local replication timing program emerges from the collective contribution of adjacent initiation sites, and, as a systems phenomenon, it is largely independent from individual sites.

The model also shows a large degree of insensitivity with respect to specific modeling choices. We wondered how strongly the specifics of assigning probabilities in the IPLS based on ENCODE amplitudes affects the simulation results. For the simulations presented so far, the local initiation probability was set to be proportional to the ENCODE amplitude. We tried alternate assignment functions (Figure 3·23) which resulted in only insignificant changes in the accuracy of the model prediction (linear r=0.86, square r=0.86, square-root r=0.84) and even when assigning the same constant value to all sites (r=0.86). On the molecular level in real cells, this implies that, once a site is competent to initiate, the probability that it is going to do so does not substantially affect the global replication timing program. We conclude that the relevant information provided by the ENCODE data, with regard to DNA replication timing, is location while amplitude is irrelevant. In

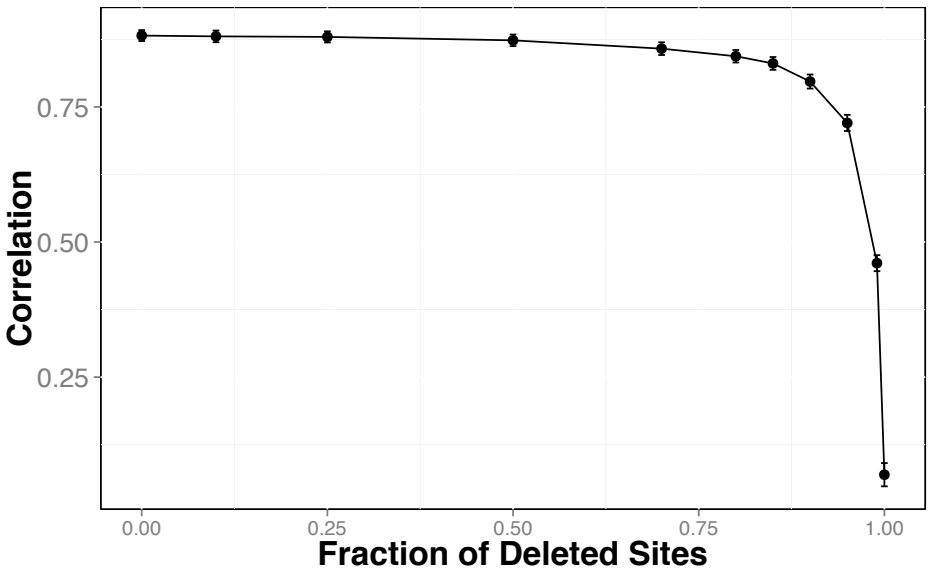

Figure 3·22: Randomly eliminating an ever larger fraction (x-axis) of DNase HS marks from the simulation of replication timing in GM06900 cells reduces the correlation (y-axis) of the model prediction with the empirical data only marginally when removing up to 50% of DNase sites. Beyond that, point, the correlation degrades gradually. Error bars represent standard error of the mean resulting from averaging the correlation value across 22 autosomal chromosomes

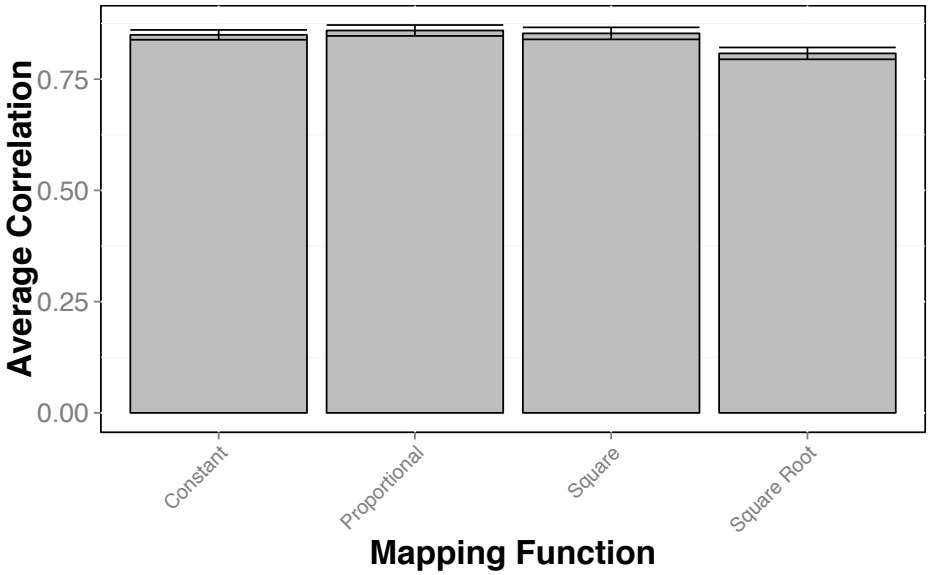

Figure 3·23: DNA replication timing was predicted for K562 and GM06990 cells based on cell-specific DNase data. Different functions were used to mapping the amplitudes a(x) of the DNase signal at a location x in the ENCODE database to the probability p(x) of replication initiation at this DNase mark. No significant differences were seen between assigning a constant value p = 1 or amplitude dependent values p(x) a(x), p(x) a(x)2, p a(x)1/2

our early simulations, we had included a small background initiation rate outside of the high efficiency initiation sites demarcated by DNase HS sites. This choice, too, was found to not affect the accuracy of the model prediction (see below), even when assigning a zero background initiation rate, i.e. when initiation exclusively occurred at high efficiency sites (Figure 3·24).

In addition to the high efficiency initiation sites, all IPLSs used in the current work contain a low probability ($P = 10 \times 10^{-4}$) replication initiation background. This step was thought to be necessary to avoid the random completion problem (i.e. to ensure complete replication in a finite time). Towards the end of our study we found that this step is, in fact, not required: setting this background probability to zero did not, overall, affect the shape of the timing prediction and lead to a slight increase of the predicted length of the S-phase from 8.3h to 8.7h. Other variants with a non-zero background probabilities did

not affect the accuracy of the model predictions: assigning P=0.001, 0.01 and 0.1 to the background did not significantly change the results (3·23).

3.2.7 Simplified DNase HS density based model produces less accurate predictions

One may wonder if it is possible to built a simpler model based on the local density of DNase HS sites alone, rather than using the more complex diffusion/collision model presented in this work (Figures 1A and S1). DNase HS have been associated (Aladjem, 2007) with early replication in several earlier studies (for a review see (Aladjem, 2007)). Re-analysis of DNase HS ENCODE data and experimental replication timing data confirms that more than 90% of DNase HS sites are located in early or medium replication timing domains (Figure 3·26). However, when trying to utilize this information to built a predictive quantitative replication timing model, one also has to answer the question of the timing for regions without DNase sites, in particular how far does the early timing information propagate around DNase sites? In our diffusion model (Figures 3·1 and 3·2) this question is answered dynamically by collision with other replication forks. A simplified model might use a fixed-size window instead, assigning a timing to each location depending on the number and intensity of DNase HS sites around the location. We simulated this model and systematically tried several fixed window-sizes (Figure 3·25). While this simple model correlates well with the experimental timing data, the correlation never exceeds 0.71, which is significantly lower compared to when the same data are input into the mechanistic model.

3.3 Discussion

Designed in a reductionist spirit, we attempted to omit all details from the model that are not required to understand the timing program (Figures 3·2 and 3·3). We wondered, if the replication fork collision mechanism in the model, which dynamically determines the distance a fork travels, could be removed by instead using the density of DNase HS sites in the vicinity (see Supplement) to assign a replication time. All such models produced

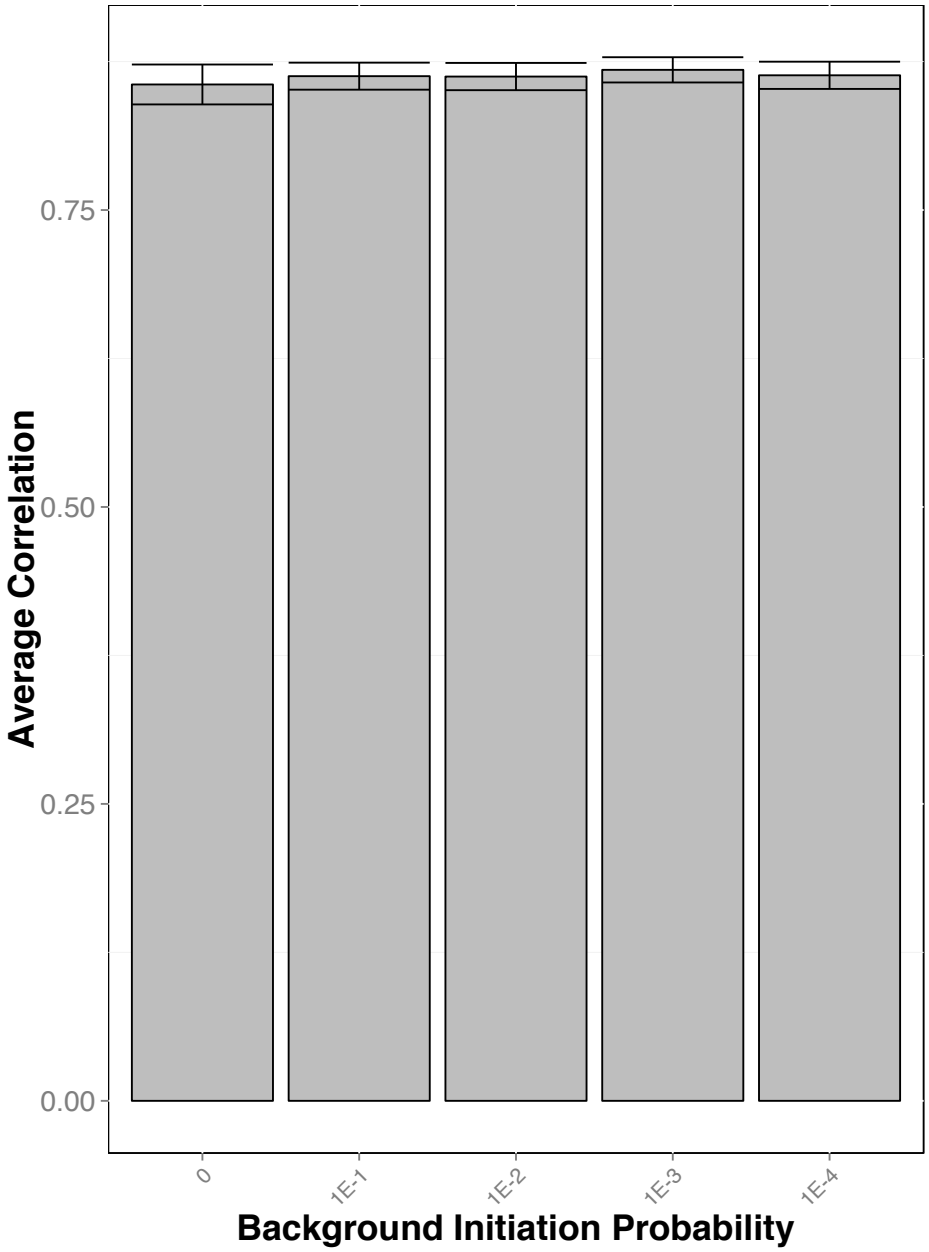

Figure 3·24: Background initiation probability does not affect the accuracy of model's predictions. DNA replication timing was predicted using DNase HS IPLS model as input for a range of background initiation probabilities (x-axis). In each case the average correlation with empirical data (y-axis) is only marginally affected. Error bars are standard error of the mean arising from averaging correlations across 22 autosomal chromosomes.

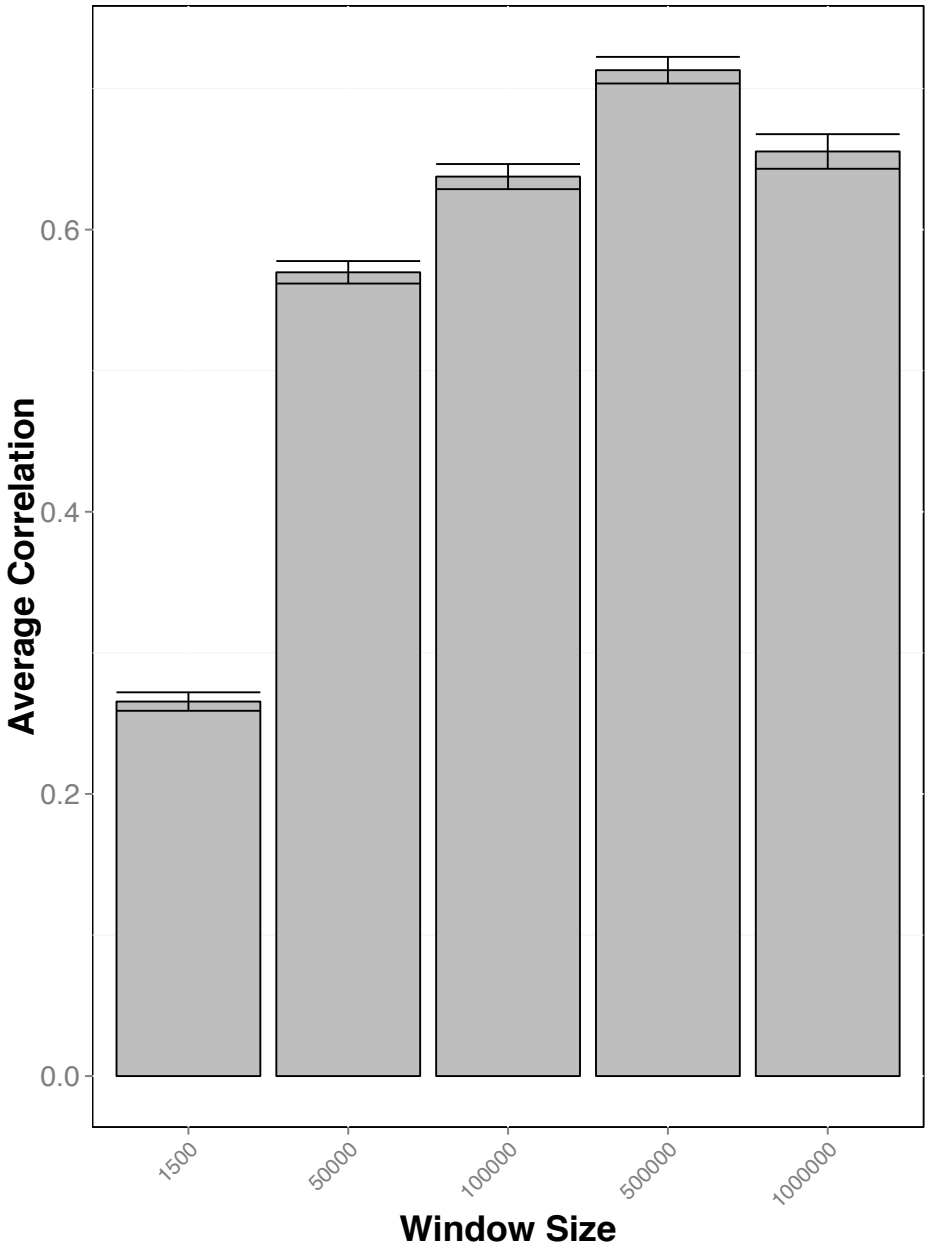

Figure 3·25: The number of DNase HS were summed for various sliding genome widows (x-axis) and correlated with empirical DNA replication timing data for same cells (GM06990). The best correlation (r=0.71) was obtained using the 0.5MB window. Error bars represent the standard error of the mean resulting from averaging correlation across 22 autosomal chromosomes.

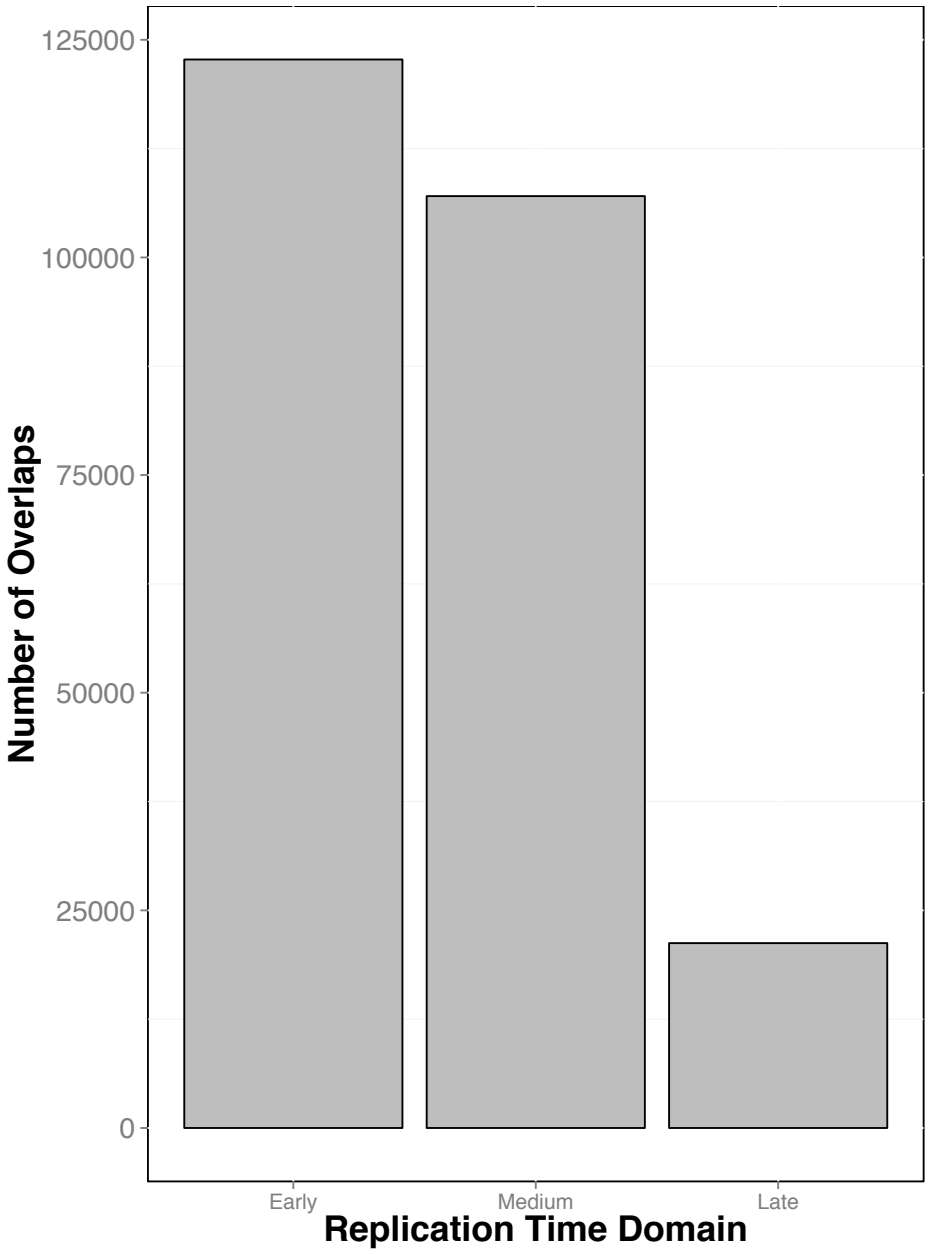

Figure 3·26: DNase HS were assigned into one of three DNA replication timing domain bins (x-axis) by consulting empirical data for matching cells (GM06990). The number of DNase sites in each bin were then counted (y-axis).

substantially worse predictions (Figure 3·25 and Supplemental material) indicating that the collision mechanism is a required aspect of the model.

An explicit separation of replication into licensing and initiation steps proved also to be unnecessary. This separation is known to be an essential molecular mechanism to avoid over-replication: licensing occurs exclusively in late M / early G1 by assembly of the so-called pre-replication complex (PreRC) at potential initiation sites, with initiation occurring later in the S-phase by conversion of the PreRC into bi-directional replication forks through phosphorylation and recruitment of other factors (Machida et al., 2005). In our model, the IPLS subsumes these two steps (over-replication itself is prevented by explicitly keeping track of replicated regions), the initiation probability at a given site represents the product of the biological probabilities to first assemble and later activate the PreRC. The above described intrinsic robustness of the model with respect to the assignment of probabilities in the IPLS is remarkable in this context. It implies that the factor dominating the timing program is the selection of the location of PreRC assemblies. Our model predicts (Figure 3·23) that the empirical timing pattern will emerge even if all PreRCs, once assembled, have the same, constant probability of being subsequently activated unless the site is passively replicated. While, to our knowledge, this possibility has not been tested in metazoan cells, it has broadly been shown to be the case in yeast (Yang et al., 2010), where a majority of initiation sites were demonstrated to have a "potential initiation efficiency", with the initiation probability remaining larger than 0.9 after correcting for passive replication.

Remarkably, we were not required to introduce a time-dependent IPLS in order to precisely predict the global replication timing pattern. Earlier models (Hyrien and Goldar, 2010; Yang et al., 2010) used location and explicit time-dependent initiation rates $I(x,t)$ to force individual initiation sites to fire, on average, at the right time to reproduce the global timing pattern. While these approaches elegantly reconcile the orchestrated global replication timing program with the stochastic nature of individual initiation events, they do not ultimately address what determines the local initiation timing. Instead, they reproduce,

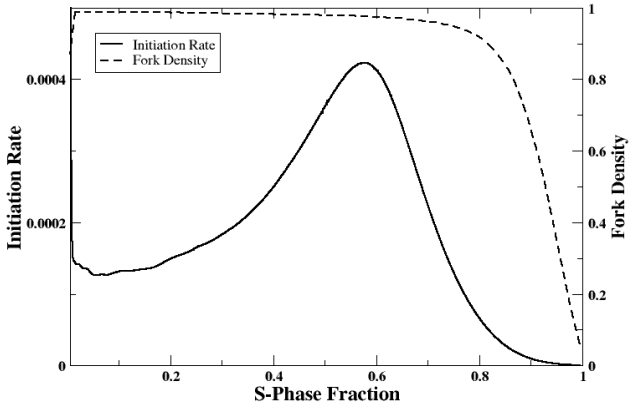

Figure 3·27: The initiation rate (number of initiation events per kilobase) and fork density (fraction of engaged replication forks) were averaged for simulated cell population and plotted as a function of S-phase fraction (time along the S-phase).

but not predict, replication timing. This is because these models rely on existing timing data, for each cell type, in order to fit a large number of variables, one or more for each initiation site. In contrast, in the model presented here, the global timing program results from the spatial distribution of initiation sites, a determination of individual firing rates was therefore not necessary. Once the genomic landmark that optimally locates initiation sites, DNase HS, was determined, timing could be predicted for all cell types. We expect that the basic mechanism described here will also work in other metazoan cells. Indeed, we found an excellent agreement between our model prediction and experimental timing data in mouse embryonic fibroblast cells (Figure 3·16).

Another important reason to use a time dependent, globally increasing initiation rate throughout S-phase in earlier models is to stabilize the S-phase length, thus avoiding the random completion problem (Blow et al., 2001; Herrick et al., 2002; Yang and Bechhoefer, 2008). These predictions were confirmed by a recent analysis (Goldar et al., 2009), uncovering a universal behavior of the global initiation firing rate across a number of species. How does this reconcile with the time-independent IPLS presented here? The firing rate in our

model depends not only on the explicitly time-independent IPLS, but also on the number of unengaged rate-limiting factors, which dynamically changes over time, as well as on the search time to find unreplicated origins, which differs between early and late replicating regions as a result of the difference in the density of initiation sites. A numerical analysis of the global initiation rate (Figure 3·27), shows a remarkable qualitative similarity to the universal patterns described in (Goldar et al., 2009). It will be interesting to see if it is necessary to extend our model by including a more detailed replication factor diffusion process, such as the sub-diffusive model discussed in (Gauthier and Bechhoefer, 2009), in order to obtain a quantitative match with experimentally determined global initiation rates in human cells.

We identified DNase hypersensitivity as the optimal IPLS predicting the DNA replication timing in metazoan cells. This suggests that DNA replication timing is largely determined mechanistically: locally by DNA accessibility as the dominant factor modulating the likelihood of forming competent initiation complexes and globally by the process of colliding replication forks a reduced representation of the known molecular processes. This interpretation implies a causal relationship, where the distribution of accessible genome regions determines DNA replication timing. Recently, a tight correlation, although significantly weaker (Pearson's r=0.8) compared to the best models tested here, between replication timing and the first eigenvector of the HiC contact probability matrix has been reported (Ryba et al., 2010), suggesting that the 3D genome organization may play a prominent role in DNA replication timing for example via replication factories or by determining the boundaries of replication domains (Baker et al., 2012). It may, therefore, seem surprising that our accurately predictive model does not require any reference to the spatial genomic organization. It could be that both phenomena, the distribution of DNase HS and 3D conformation, have a common cause. Yet, it is generally believed that DNase HS sites are established by transcription factors dislocating and/or limiting the movement of histones (Felsenfeld et al., 1996). It therefore seems reasonable to speculate that the distribution of DNase HS sites contributes to the control of the genomic conformation.

In summary, provided with a proper "initiation probability landscape" a mathematical construct that encodes the location information, the model predicts the replication timing program and recapitulates cell-specific timing patterns, including abnormal timing behavior in cancer cells. These results strongly support the concept that replication timing is a stochastic process ultimately determined by chromatin structure, which itself is a consequence of the topological organization of genes and functional regulatory elements on the chromosome as encoded in the DNA sequence.

3.4 Materials and Methods

3.4.1 Software implementation

The custom-written software (Replicon) is capable to simulating genome replication and recording various associated measurements, such as DNA replication timing. Replicon is written in C++ and can be executed in a multi-threaded mode. In our experiments, a typical simulation of a human genome-wide DNA replication profile took about 15 minutes when executed in parallel: 22 simulations each running on a 4-core, 2.93 GHz Linux node.

3.4.2 Simulated replication time assignment to genome coordinates

The simulation consists of millions of simulated asynchronous cells. The assignment of replication time to genome coordinates starts by first separating the cell population, according to each cell's DNA content, into one of six bins (akin to a flow-sorter sort). The replication time is calculated for each genome coordinate (500nt resolution) by taking the average of the product of the bin number (1 through 6) and the number of times the genome coordinate in question was observed in each bin.

3.4.3 Flow sorter gating optimization

We used a simulated annealing algorithm to approximate DNA flow-sorter bin boundaries with the objective to minimize the Euclidean distance between simulated and experimentally derived DNA replication timing profile. Starting from a state where flow-sorter bin

boundaries were randomized, replication timing was simulated based on DNase DGF data for GM06990 cells. The neighboring state was calculated by perturbing a randomly chosen bin boundary. The new boundary value was chosen from Normal distribution, where was set to the old boundary and to a value of 1.

3.4.4 IPLS generation

Utilizing a 500nt resolution, the probability of initiating replication at any given genomic location was set to either a scaled value of an attribute of interest or to a background frequency of 1E-4, whichever was greater. Scaling was achieved using the formula x / max(x), where x is the attribute of interest. All DNA replication initiation landscapes, unless stated otherwise, were generated from a local copy of the UCSC ENCODE database 24, where the data attribute 'score' was used as the attribute of interest. For GC-content IPLS, the probability of DNA replication initiation was scaled to the 'sumData' attribute of the 'gc5Base' annotation table. For CpG island IPLS, the probability of DNA replication initiation was scaled to the 'obsExp' attribute of the 'cpgIslandExt' annotation table. For DNA G-quadruplex (G4) IPLS, the probability of DNA replication initiation was scaled to the length of the G4 motif. The G4 motifs were identified using a regular expression as described in (Todd et al., 2005). ORChID IPLS was based on 'wgEncodeBuOrchidV1.bigWig' annotation file available at the UCSC Genome Browser (http://genome.ucsc.edu/), where the intensities of hydroxyl radical accessibility were averaged over non-overlapping 500nt windows. The transcription start site (TSS) IPLS, was set to a constant probability of 1.0 for every genomic region annotated as 'txStart' in the 'refGene' table.

3.4.5 Generation of Reduced-Model IPLSs

For each set of genome annotations in a pair-wise comparison, we identified and removed co-localized genome regions, generating the 'Subtract Overlap' reduced model for each model in the comparison. The 'Subtract Random' model was generated by removing randomly chosen genome regions from each model in the comparison, such that the number of regions

in 'Subtract Overlap' and 'Subtract Random' models were equal.

3.4.6 In-silico ETV6-RUNX1 Translocation

We generated t(12;21)(p13;q22) chromosomal translocation in-silico by joining GM06990 DNase DGF data for chromosomes 12 and 21 producing an ETV6-RUNX fusion gene using molecularly mapped breakpoint coordinates (Wiemels et al., 2000). We then simulated replication timing for two fused chromosome products and compared simulated replication timing data for translocated and un-translocated chromosome 12.

3.4.7 Robustness

The effect of deleting DNase HS sites was investigated using DNase DGF data available for GM06990 cells. At each iteration of the algorithm, we erased an ever-increasing fraction of DNase sites and generated a corresponding DNA replication initiation landscape.

3.4.8 DNA Replication Plasticity Regions

DNA replication plasticity regions were identified using custom-developed software. First, a DNA replication difference profile was derived for a given pair of DNA replication timing profiles by subtracting one DNA replication profile from another for matching genome coordinates. The distribution of differences was observed to follow Normal distribution. Using the Normal distribution, a P-value was assigned to every 500nt non-overlapping genome bin (the resolution of our model) in the difference profile. A DNA replication plasticity region was identified as such if at least 3 consecutive bins were assigned a P-value of 0.001 or less.

Chapter 4

MicroRNA-mediated differentiation impairment in osteosarcoma

4.1 Introduction

Osteosarcoma is the most common primary bone malignancy and occurs most frequently in adolescents (Mirabello et al., 2009). Osteosarcoma tumors most often arise in the long bones of the skeleton, with more than half presenting around the knee (Broadhead et al., 2011), and is less common in axial skeleton (Martin et al., 2012). At diagnosis, 20% of osteosarcoma patients present with lung metastases with an additional 40% developing metastases at later stage (Martin et al., 2012). Survival rates for localized osteosarcoma are at 60-70% (Mirabello et al., 2009; Longhi et al., 2006), while the five-year survival for osteosarcoma patients with metastases is 20% (PosthumaDeBoer et al., 2011). Despite intense research efforts, the survival rates for osteosarcoma have remained essentially unchanged for over two decades (Mirabello et al., 2009; Longhi et al., 2006). Contributing to the challenge of understanding and ultimately developing effective treatments for osteosarcoma is its complex karyotype and high level of chromosomal instability (Helman and Meltzer, 2003). With the increasing understanding of tumor biology, osteosarcoma is often regarded as a disease of cell differentiation (Tang et al., 2008; Thomas and Kansara, 2006). Therefore the path to developing new treatment approaches for osteosarcoma lies through an improved understanding of the dysregulation of the bone differentiation program in this devastating disease.

Discovery of small (about 22nt in length) non-coding RNA species, termed microRNAs (miRNAs), has, in many ways, revolutionized the understanding of gene expression

regulation. It is now recognized, for instance, that miRNAs contribute to many biological processes (Ambros, 2004) and that their expression patterns can be used to classify cancers (Lu et al., 2005; Bloomston et al., 2007), suggesting that miRNAs play the roles similar to tumor suppressors and oncogenes (Esquela-Kerscher and Slack, 2006; Dalmay and Edwards, 2006). MicroRNAs play an integral role in controlling cell differentiation by suppressing genes that maintain plasticity (Yi et al., 2008), or by suppressing genes that inhibit cell-lineage commitment (Li et al., 2008) or through a combination of the two (Forrest et al., 2010).

MiRNAs play a paramount role in bone differentiation. (Hassan et al., 2012; Inose et al., 2009; Kobayashi et al., 2008; Wang et al., 2008; Sugatani and Hruska, 2007). Recent studies identified miRNA biomarkers relevant to therapy response and identification of therapeutic targets (Jones et al., 2012; Lulla et al., 2011; Cai et al., 2013; Maire et al., 2011). Much attention has been devoted to the role of miR-23a in bone differentiation, primarily via its targeting (both direct and indirect) of transcription factors essential to osteoblastogenesis such as TRPS1, RUNX2 and SATB2 (Hassan et al., 2010; Zhang et al., 2012; Zhang et al., 2011). In the current work, we study the effects of miR-23a expression in HOS cells, which are distinguished from other human osteosarcoma cells by their ability to undergo a bone cell lineage differentiation program (Siggelkow et al., 1998).

4.2 Results

4.2.1 Mir-23a inhibits differentiation in osteosarcoma cells

To confirm that HOS cells are amendable to bone differentiation induction (Siggelkow et al., 1998), we treated these cells with L-ascorbic acid, which induces the formation of collagenous extracellular matrix and brings about osteoblast-specific gene expression program in osteogenic lineage cells (Franceschi et al., 1994). We then monitored HOS cell culture for the presence of calcium deposits, which serve as a marker of bone mineralization, via Alizarin Red staining. Our results indicate that HOS cells undergo osteoblast-like differentiation upon stimulation with L-ascorbic acid. HOS cells exhibit intense Alizarin

Red staining on day 7 post differentiation induction (Figure 4·1). We confirmed this result by monitoring the expression of collagen Ia1 (*COL1A1*) – a gene marker of bone differentiation (Figure 4·2). We have observed a 2-fold increase in *COL1A1* mRNA levels between initial and terminal differentiation time-points.

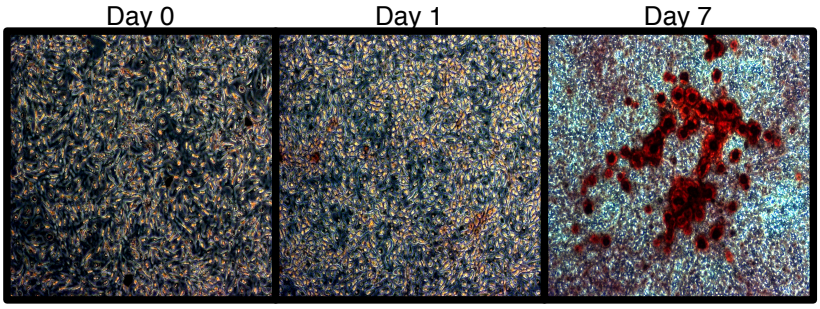

Figure 4·1: Alizarin red staining of HOS cells during the differentiation time course (7 days). Red staining is indicative of calcium deposits.

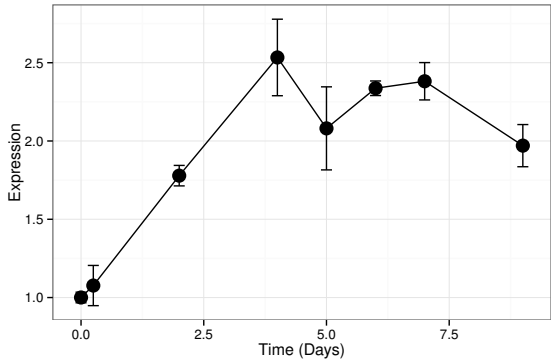

Figure 4·2: Relative expression levels (assayed with qPCR) of COL1A1 during HOS differentiation time course. Expression levels are normalized to the initial time point. Error bars represent standard deviation.

4.2.2 MicroRNA-23a targets genes involved in bone differentiation

To study the effect of miR-23a on the gene expression program in osteosarcoma, we set out to identify genes that are transcriptionally repressed by miR-23a in HOS cells. To that end, we first transfected HOS cells with a miR-23a mimic and compared their gene expression profile with mock-transfected cells using Illumina microarray platform. Our analysis shows

that 1,530 genes are down-regulated in miR-23a transfected cells versus mock-transfected cells. These genes are significantly enriched for predicted (Lewis et al., 2005) miR-23a targets (262 overlapping genes; $P_{overlap} = 4.96 \times 10^{-46}$).

Having established the on-target effects of miR-23a over-expression in HOS cells, we next asked if the genes affected by miR-23a overlap significantly with genes that change in expression during HOS differentiation. Given miR-23a's role as a dampener of bone differentiation gene expression program (Hassan et al., 2010), we focused on genes that are up-regulated during HOS differentiation time course. To that end, we employed a gene expression Illumina microarray to compare mRNA expression levels in HOS cells prior to differentiation induction with that of HOS cells that display phenotypic properties of bone cells post differentiation induction. Our analysis shows that 3,065 genes increase in expression during HOS cell differentiation. Of those, 466 genes are down-regulated upon miR-23a transfection ($P_{overlap} = 1.21 \times 10^{-10}$). To identify genes of interest that are under miR-23a control and are relevant to HOS cell differentiation, we identified 77 genes that meet the following criteria: (i) are computationally predicted miR-23a targets; (ii) are down- regulated on miR-23a over-expression and; (iii) are up-regulated during HOS cell differentiation time course.

Transcription factor binding site enrichment analysis (Loots et al., 2002) reveals that more than one-half of the 77 genes contain an SP1 transcription factor motif within 2kb of their transcription start site (42 genes; $P_{enrichment} = 1 \times 10^{-23}$). Members of the SP1 transcription factor family, which includes Osterix (SP7), are bone lineage specific and are required for osteoblast differentiation and bone formation (Nakashima et al., 2002). These results suggest that over-expression of miR-23a interferes with the bone differentiation program by counteracting the action of osteoblast lineage inducing transcription factor SP1. In order to narrow down the list of likely miR-23a targets that are involved in bone differentiation we looked for gene signature enrichments, as curated in Molecular Signatures Database (Subramanian et al., 2005), among the 42 high-quality miR23a with an upstream SP1 binding site. The top-enriched signature is that of genes up-regulated

upon *EZH2* knock-down in prostate cancer cells (Nuytten et al., 2008) (Table 4.1; $P_{overlap} = 3.61 \times 10^{-9}$). EZH2, the catalytic subunit of the PRC2 repressive complex, is commonly associated with silencing of pro-differentiation genes (Simon and Lange, 2008), which makes its function analogous with that of miR-23a in bone differentiation. This observation is particularly interesting as prostate metastatic tumors are often osteoblastic (Logothetis and Lin, 2005) and phosphorylation of EZH2 by CDK1 is critical for osteogenic differentiation of human bone-marrow-derived mesenchymal cells (Wei et al., 2011).

Table 4.1: Mir-23a Target Genes Relevant to HOS Differentiation

Gene Symbol	Gene Name
CAB39	calcium binding protein 39
CLDN12	claudin 12
DCBLD2	discoidin, CUB and LCCL domain containing 2
FAM46A	family with sequence similarity 46, member A
GJA1	gap junction protein, alpha 1, 43kDa (connexin 43)
IRF1	interferon regulatory factor 1
MARCKS	myristoylated alanine-rich protein kinase C substrate
RAB8B	RAB8B, member RAS oncogene family
TNFAIP3	tumor necrosis factor, alpha-induced protein 3
UBL3	ubiquitin-like 3

4.2.3 GJA1 is a major target of miR-23a

In order to identify a specific miR-23a target related to bone differentiation, we examined the gene list in Table 4.1 for genes involved in sensing extracellular environment and intercellular communication, which are essential in bone formation that are also present in aforementioned gene signatures. Expert-based examination of the gene list led a gene whose product is connexin(Cx)-43 (also known as *GJA1*). GJA1 is a member of the gap junction family and is the most abundant gap junction expressed in bone (Loiselle et al., 2013), where it facilitates response to extracellular mechanical (Jiang et al., 2007), pharmacologic and hormonal stimuli (Plotkin and Bellido, 2013) and is required for signal transduction among bone lineage cells (Civitelli, 2008). Crucially, GJA1 is essential for osteoblast differentiation in humans and animals in vivo (Stains and Civitelli, 2005a).

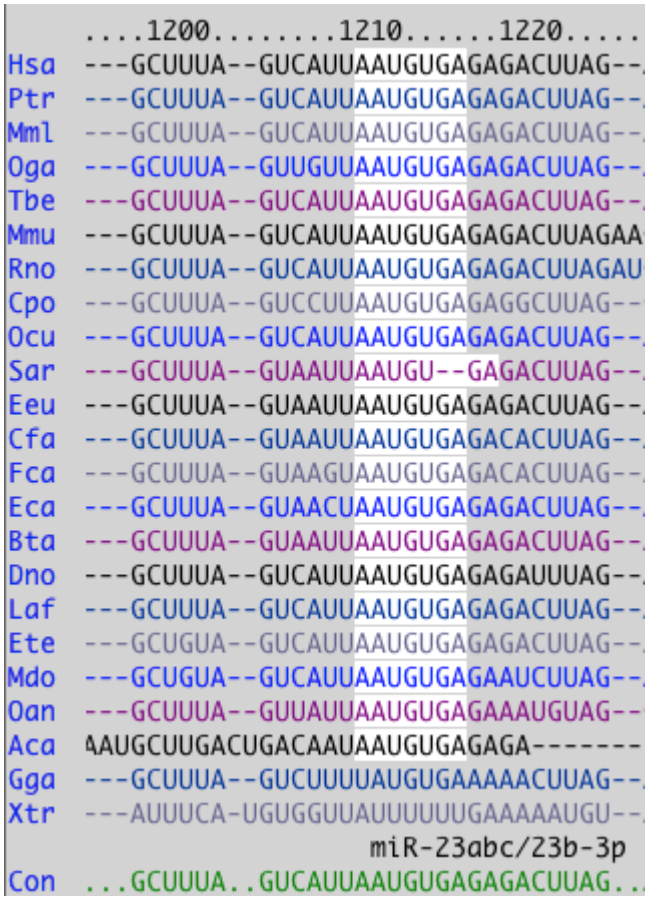

Figure 4·3: Conservation (sequence alignment) of miR-23a binding site on 3' UTR of the GJA1 gene (highlighted in white). Each sequence row identifies a different species with a conserved (Con) sequence included in the last row.

We verified that the miR-23a binding site is well-conserved in 3'UTR of the GJA1 gene (Figure 4·3). We next sought to verify miR23a:*GJA1* interaction in vitro. To that end we carried out a reporter assay where the 3'UTR of *GJA1* was cloned into the 3'UTR of a luciferase gene. We find that miR-23a significantly reduces luciferase *GJA1* reporter activity (Figure 4·4). These results confirm that *GJA1* is a bona fide miR-23a target.

We then set out to elucidate the expression pattern of *GJA1* during osteoblast cell differentiation. To that end, we identified, in Gene Expression Omnibus (GEO) (Barrett et al., 2013), a dataset used in two recently published studies (Nabavi et al., 2012; Pustylnik

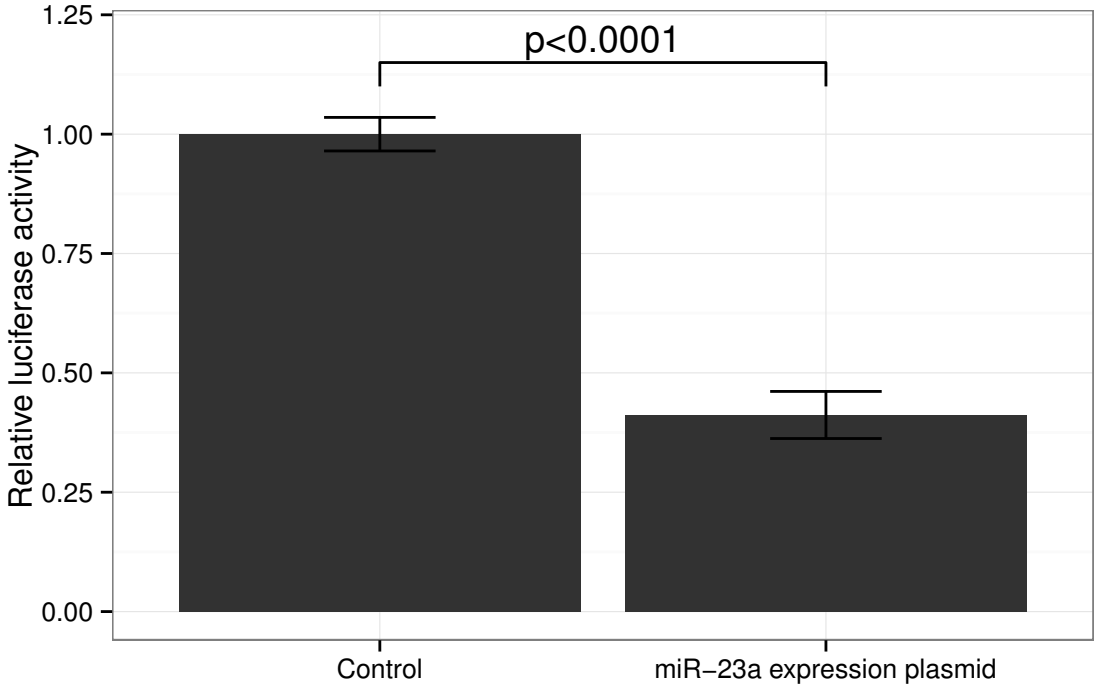

Figure 4·4: Luciferase activity in HOS cells transfected with either control (renilla) or reporter (firefly) plasmid. Data were normalized to the luciferase activity of the control plasmid. Error bars represent standard deviation of technical repeat experiments (n=3). P value calculated by Student's t-test.

et al., 2013) that assayed gene expression in mouse MC3T3-E1 osteoblast cells following differentiation induction by L-ascorbic acid. Neither study explicitly addressed gap junction expression in osteoblast differentiation. Our analysis of the data deposited in GEO shows that *GJA1* expression increases over 500-fold (6-probe average; $P < 1.0 \times 10^{-6}$) in mouse osteoblast cells following differentiation.

Next, we asked whether *GJA1* expression levels increase during HOS cell differentiation and how this expression pattern may be related to miR-23a expression. To that end, we induced differentiation in HOS cells and analyzed mRNA expression with quantitative (q)PCR. Our results show (Figure 4·5) that *GJA1* levels increase as HOS cells begin to display phenotypical hallmarks of osteblast cells. *GJA143* levels reach their maximum on day 7 post differentiation induction, at which point HOS cells display the phenotypic properties

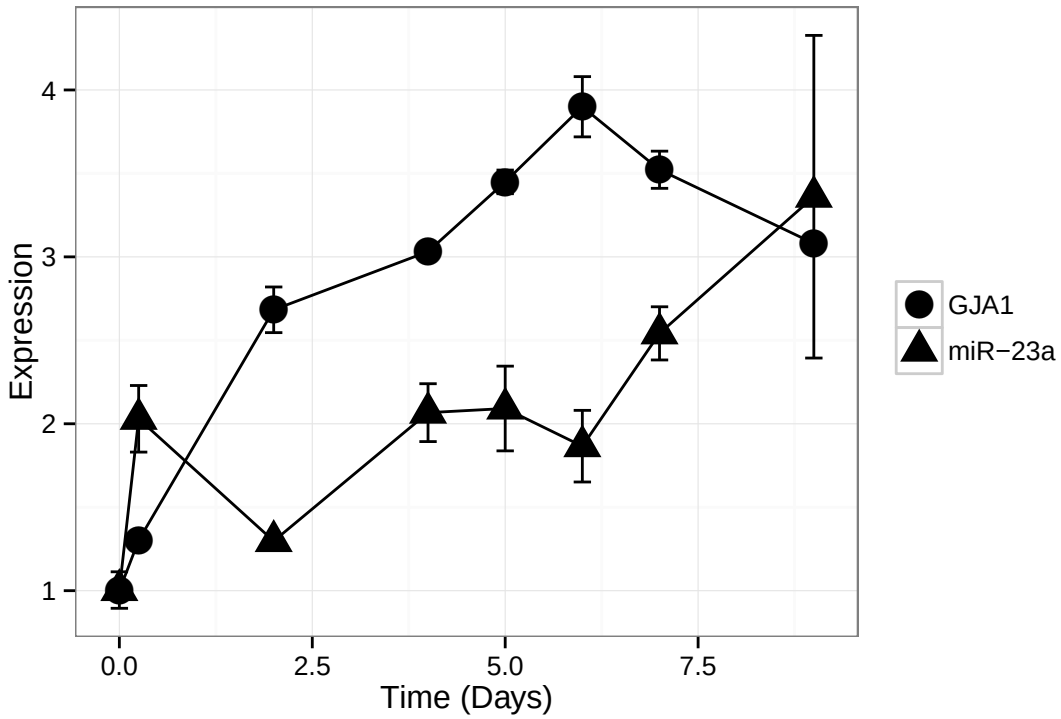

Figure 4·5: Relative expression of GJA1 and miR-23a during HOS cell differentiation time course. HOS cells were induced to differentiate and mRNA aliquots were isolated at selected time points (x-axis). Relative expression of GJA1 and miR-23a were normalized, separately, to their basal levels (Day 0). Error bars represent standard deviation due to technical repeats (n=3).

of bone differentiation (Figure 4·1) and *COL1A1* levels have reached their peak (Figure 4·2). Importantly, miR-23a levels are inversely related to *GJA1* expression: reaching minimum, when *GJA1* levels are at their peak and Alizarin red staining is at its maximum; then increasing gradually past day 4 when *GJA1* levels off and begins to decrease.

4.3 Discussion

A number of recent studies have identified miRNAs that are differentially expressed between normal bone and osteosarcoma (see recent summaries in (Miao et al., 2013) and (Zhou et al., 2013)). Importantly, miR-23a has been shown to control bone differentiation (Hassan et al., 2010; Zhang et al., 2012; Zhang et al., 2011). However, it is not clear what role, if any, miR-23a has within the realm of osteosarcoma. Studies that focused on the function of miR-23a in bone differentiation have been restricted to the examination of miR-23a and expression of transcription factors that are paramount to bone biology such as RUNX2 (Zhang et al., 2012; Zhang et al., 2011) and/or SATB2 (Hassan et al., 2010), which were chosen *a priori*. Here, we profiled miR-23a expression in a comprehensive panel of osteosarcoma cell lines, identifying a subset of cells that overexpress miR-23a relative to osteoblast cells. We find that an osteosarcoma cell line with lowest miR-23a levels, HOS, is capable of undergoing bone-like differentiation, as reported earlier (Siggelkow et al., 1998). Overexpression of miR-23a in HOS cells inhibits their ability to differentiate. We find that miR-23a acts to inhibit differentiation, at least in part, by blocking the expression of connexin 43 (GJA1) a key protein required of cell-cell communication and osteoblast differentiation.

In this study we examined the relationships between miR-23a and bone differentiation within the context of osteosarcoma. Previous studies demonstrated the relationship between miR-23a and transcription factors that are central to bone differentiation program (Zhang et al., 2012; Hassan et al., 2010). Separately, Inose and colleagues (Inose et al., 2009) have shown that bone differentiation is negatively impacted in mice by miR-206-mediated silencing of *GJA1*. We could not detect miR-206 expression in HOS nor human osteoblast cells (data not shown). This points to redundant pathways that fine-tune *GJA1*

expression during bone differentiation.

Loss of gap junctional communication delays osteoblast differentiation and reduces the ability of these cells to form mineralized extracellular matrix (Lecanda et al., 1998; Schiller et al., 2001). A clue as to how this effect arises came from an observation that loss of GJA1 function is accompanied by diminished extracellular-signal-regulated kinase (ERK) activity (Stains and Civitelli, 2005b). In the proposed mechanism, a GJA1 gap junction allows for passage of a 2^{nd} messenger activating ERK/PI3K signaling cascades that would in turn recruit transactivator SP1 to promoter regions of genes associated with the osteoblastic gene expression program, such as osteocalcin and *COLIA1* (Stains and Civitelli, 2005b; Stains et al., 2003). The loss of GJA1 gap junctions diminishes ERK activity resulting in preferential recruitment of SP3 repressor to osteocalcin and *COLIA1* gene promoters (Stains and Civitelli, 2005b; Stains et al., 2003). Here, we show that miR-23a gene targets in HOS cells are enriched for SP1 binding site within 2kb of their transcription start site, which suggests that miR-23a may function by counteracting the effects of the SP1 transcription factor.

4.4 Materials and Methods

4.4.1 Cell culture

All cell lines were obtained from ATCC. The cells were grown in DMEM media with 10% fetal bovine serum and supplemented with 1% penicillin and streptomycin. HOS cells were grown to got 100% confluence, followed by differentiation at 7-9 days induced by bone inducing agents, that include L-ascorbic acid 50ug/ml and beta-glycerophosphate 5mM (Hassan et al., 2006). Cells were harvested at indicated times for mRNA and protein extraction or fixed with 10% neutral-buffered formalin (NBF) for detection of calcium deposits by Alizarin Red staining.

4.4.2 Protein and mRNA analyses

Total RNA was isolated using Trizol reagent (Invitrogen), treated with DNase I (Ambion) and reverse transcribed using "iScript Reverse Transcription Supermix for RT-qPCR" (BIO-RAD). *GJA1* and *COL1A1* gene expression qRT-PCR were performed using the TaqMan Gene Expression Assays (ABI/ Life Technologies). mRNA levels were normalized to housekeeping gene ACTB. miRNA-23a was quantified in triplicate using the TaqMan MicroRNA Assay (ABI/ life technologies) and normalized to U6. mRNA levels were assayed for relative expression using procedure described in (Livak and Schmittgen, 2001).

4.4.3 Immunoblot

Whole cell lysates from transfected HOS cells were prepared using RIPA buffer. Proteins were analyzed by SDS PAGE, transferred to nitrocellulose membranes and probed with GJA1 antibody (ab11370 Abcam). Western Blots were quantified by densitometry.

4.4.4 Luciferase reporter assay

HOS cells were co-transfected in 24 well-plates using Lipofectamine 2000 (Invitrogen) with 20nM miR-23a mimic or control miRNA mimic and 100ng of psiCHECK2- 3UTR (Promega) vector containing the GJA1-3UTR cloned into the multiple cloning site of Renilla luciferase. After 48hr of transfection luciferase activity was measured using the Dual Luciferase Assay System (Promega). The experiment was performed in triplicate. Results were normalized to those obtained in cells transfected with an empty vector. Data were normalized to Firefly luciferase and results from 3 independent experiments were compared. GJA1 sequences were cloned into psiCHECK-2 by annealing complementary oligomers matching each GJA1 sequence with overhanging ends complementary to the XhoI and NotI sites of psiCHECK-2.

4.4.5 Transfection assay

HOS cells were differentiated as described above. Two days after differentiation, cells were transfected using Lipofectamine RNAiMAX Reagent (Invitrogen) with ON-TARGETplus-siGJA1-pool, siGJA1-05, siGJA1-06 (Thermo Scientific L-011042-00-0005) at a final concentration of 100pmol. After 72hrs transfection, the cells were harvested for mRNA and protein assays or fixed with 10% NBF for detection of calcium deposits by Alizarin Red Staining.

4.4.6 Data analysis

All statistical analyses were carried out using R statistical environment version 3.0 Microarray data were analyzed using limma package (Smyth, 2005). Data from GEO were obtained using the GEOquery package (Davis and Meltzer, 2007). MicroRNA-seq data were normalized using using the DESeq package (Anders and Huber, 2010).

Chapter 5

Conclusion and Further Work

Availability of extensive high-quality biological datasets have made it possible to draw exciting connections between multi-layered observations. While it is important to guard against data artifacts (Bilke and Gindin, 2012), data integration techniques have proven to be an invaluable resource for understanding of systems-level phenomena.

In Chapter 2, I explored links between genome-wide methylation patterns in breast cancer, their ability to stratify breast cancer subclasses, and their effect on gene expression and disease outcome, yielding a number of discoveries. My analysis demonstrated that DNA methylation levels, alone, are sufficient to recapitulate expression-based breast cancer molecular subtypes (section 2.2.2) and serve as a prognostic marker for survival (section 2.2.7). DNA methylation is inherently more chemically stable than mRNA, requiring fewer handling safeguards, making it more amendable to a clinical setting (Heyn and Esteller, 2012). Therefore, there is a 'translational' advantage to develop DNA methylation based clinical assays (Martens et al., 2009). The work presented here that is relevant to cancer classification (section 2.2.2) and prognosis (section 2.2.7) could serve as very early steps toward real-world applications. Furthermore, findings outlined in sections 2.2.5 and 2.2.6 point to specific biological pathways that are hypermethylated and silenced in Basal and Luminal breast cancer subtypes pointing to potential therapeutic targets. This research opens up a number of possibilities that could be explored in future work.

Breast cancer is a heterogeneous disease with distinct chromosomal aberrations (van Beers and Nederlof, 2006). A number of recent studies have explored relationships between DNA methylation, copy-number gains and losses in breast cancer. Tang and colleagues

(Tang et al., 2012) found that differentially methylated DNA regions tended to co-localize with sites of copy-number aberrations and Alu repeats, while Aure and colleagues (Aure et al., 2013) identified a set of miRNAs whose expression is altered by DNA methylation and copy-number aberrations. It would be telling to see if the hypermethylated regions of the Basal or the Luminal gene co-methylation modules are proximal to copy number loss regions and whether those regions encompass tumor suppressors.

In Chapter 3, I have shown that an intuitive model of DNA replication is able to predict DNA replication timing in human cells with extraordinary accuracy (Pearson's r=0.92 between prediction and experimentally determined data). This model of DNA replication timing, based on the location of DNase HS sites, contains a single tunable parameter: the number of simulated replication forks. The model predictions were able to identify regions of replication timing plasticity between cell lineages and provide an explanation for the often-observed rapid changes in replication timing around the sites of chromosomal fusions (section 3.2.4). The model has proven to be highly robust – generating accurate predictions even in cases were most of the DNase HS sites were deleted (section 3.2.6). A major finding of the work is that location of replication initiation sites is the main component driving the DNA replication timing program rather then the strength of individual sites. Figure 3·23 illustrates that the replication timing program is unchanged even if all of the initiation sites have the same probability of firing. There are a number of ways that this study could be expanded.

The development of the DNA replication model has been guided by the reductionist approach. Yet, certain modifications may be made to the model to study various biological scenarios. In the current work, the replication fork velocity has been set to a constant 3kb per minute, which is in good agreement with widely-reported results (Alberts et al., 2008). However, the actual replication fork velocity may vary by as much as 10-fold (from 0.5 to 5.0 kb/min) (Conti et al., 2007; Hyrien and Goldar, 2010; Guilbaud et al., 2011). The current model may be adapted to study, for instance, replication fork stalling, or slowing (Labib and Hodgson, 2007). This could be done by adopting, analogous to IPLS,

an additional genome landscape parameter that maps replication fork velocity to genome coordinates.

In the current study, I have been limited to cell line data for which both replication timing and DNase HS site datasets were available. While this is adequate to establish the accuracy of the model, it is overly restrictive as replication timing is available for only a handful of cell lines. Having established the validity of using DNase HS data to predict DNA replication timing with the DNA replication model described in Chapter 3, it is now reasonable to work with predicted timing data directly. Using the vast amount of DNase HS data produced by the ENCODE project for a wide variety of cell lines it would be rather straight-forward to simulate DNA replication timing for each cell line. One potential application would be to characterize cancer cells by matching their replication timing profile with that of cells isolated from normal tissues.

In Chapter 4, I have shown that at least part of the differentiation delay effect of miR-23a is through its suppression of GJA1 expression. By comparing gene expression of HOS cells transfected with miR-23a with untransfected cells, I found a significant enrichment on an SP1 TFBS among miR-2a targets. Using literature analysis, I identified *GJA1* as a major miR-23a target involved in bone differentiation (section 4.2.3) as its expression, which is induced by SP1 enhancer and repressed by SP3 repressor, has been shown to be crucial for bone differentiation (Stains et al., 2003). Experiments were carried to show a knockdown of *GJA1* expression delays HOS cell differentiation, confirming its role in bone-like differentiation. Confirmatory analysis of gene expression data (Nabavi et al., 2012; Pustylnik et al., 2013) deposited in GEO, revealed that *GJA1* expression significantly increases in differentiating osteoblast cells, which was confirmed in HOS cells.

The work presented in Chapter 4 lays the groundwork for a number of new research directions. A potentially interesting aspect to explore would be the presence of copy-number aberrations encompassing miR-23a or *GJA1*. Gains on chromosome 19 may explain why miR-23a is highly expressed in some osteosarcoma cells.

References

(2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology*, 9(4):e1001046.

Aladjem, M. I. (2007). Replication in context: dynamic regulation of DNA replication patterns in metazoans. *Nature Reviews. Genetics*, 8(8):588–600.

Alberts, B., Wilson, J. H., and Hunt, T. (2008). *Molecular Biology of the Cell*. Garland Science, 5th edition.

Ambros, V. (2004). The functions of animal microRNAs. *Nature*, 431(7006):350–5.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

Aure, M. R., Leivonen, S.-K., Fleischer, T., Zhu, Q., Overgaard, J., Alsner, J., Tramm, T., Louhimo, R., Alnæ s, G. I. G., Perälä, M., Busato, F., Touleimat, N., Tost, J., Bø rresen Dale, A.-L., Hautaniemi, S., Troyanskaya, O. G., Lingjæ rde, O. C., Sahlberg, K. K., and Kristensen, V. N. (2013). Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors. *Genome biology*, 14(11):R126.

Baker, A., Audit, B., Chen, C.-L., Moindrot, B., Leleu, A., Guilbaud, G., Rappailles, A., Vaillant, C., Goldar, A., Mongelard, F., D'Aubenton-Carafa, Y., Hyrien, O., Thermes, C., and Arneodo, A. (2012). Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *PLoS computational biology*, 8(4):e1002443.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(Database issue):D991–5.

Bechhoefer, J. and Rhind, N. (2012). Replication timing and its emergence from stochastic processes. *Trends in genetics : TIG*, 28(8):374–81.

Ben-Porath, I., Thomson, M. W., Carey, V. J., Ge, R., Bell, G. W., Regev, A., and Weinberg, R. a. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nature genetics*, 40(5):499–507.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., and Lemaitre, J.-M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature structural & molecular biology*, 19(8):837–44.

Bilke, S. and Gindin, Y. (2012). Analyzing the association of SCNA boundaries with replication timing. *Nature biotechnology*, 30(11):1043–5; author reply 1045–6.

Bird, A. (2007). Perceptions of epigenetics. *Nature*, 447(7143):396–8.

Bloomston, M., Frankel, W. L., Petrocca, F., Volinia, S., Alder, H., Hagan, J. P., Liu, C.-G., Bhatt, D., Taccioli, C., and Croce, C. M. (2007). MicroRNA expression patterns to differentiate pancreatic adenocarcinoma from normal pancreas and chronic pancreatitis. *JAMA : the journal of the American Medical Association*, 297(17):1901–8.

Bloushtain-Qimron, N., Yao, J., Snyder, E. L., Shipitsin, M., Campbell, L. L., Mani, S. A., Hu, M., Chen, H., Ustyansky, V., Antosiewicz, J. E., Argani, P., Halushka, M. K., Thomson, J. A., Pharoah, P., Porgador, A., Sukumar, S., Parsons, R., Richardson, A. L., Stampfer, M. R., Gelman, R. S., Nikolskaya, T., Nikolsky, Y., and Polyak, K. (2008). Cell type-specific DNA methylation patterns in the human breast. *Proceedings of the National Academy of Sciences of the United States of America*, 105(37):14076–81.

Blow, J. J., Gillespie, P. J., Francis, D., and Jackson, D. a. (2001). Replication origins in Xenopus egg extract Are 5-15 kilobases apart and are activated in clusters that fire at different times. *The Journal of cell biology*, 152(1):15–25.

Bracken, A. P. and Helin, K. (2009). Polycomb group proteins: navigators of lineage pathways led astray in cancer. *Nature reviews. Cancer*, 9(11):773–84.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Broadhead, M. L., Clark, J. C. M., Myers, D. E., Dass, C. R., and Choong, P. F. M. (2011). The molecular pathogenesis of osteosarcoma: a review. *Sarcoma*, 2011:959248.

Cai, H., Lin, L., Cai, H., Tang, M., and Wang, Z. (2013). Prognostic evaluation of microRNA-210 expression in pediatric osteosarcoma. *Medical oncology (Northwood, London, England)*, 30(2):499.

Carey, L. A. (2010). Through a glass darkly: advances in understanding breast cancer biology, 2000-2010. *Clinical breast cancer*, 10(3):188–95.

Cayrou, C., Coulombe, P., Vigneron, A., Stanojcic, S., Ganier, O., Peiffer, I., Rivals, E., Puy, A., Laurent-Chabalier, S., Desprat, R., and Méchali, M. (2011). Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome research*, 21(9):1438–49.

Charafe-Jauffret, E., Ginestier, C., Monville, F., Finetti, P., Adélaïde, J., Cervera, N., Fekairi, S., Xerri, L., Jacquemier, J., Birnbaum, D., and Bertucci, F. (2006). Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene*, 25(15):2273–84.

Civitelli, R. (2008). Cell-cell communication in the osteoblast/osteocyte lineage. *Archives of biochemistry and biophysics*, 473(2):188–92.

Conti, C., Saccà, B., Herrick, J., Lalou, C., Pommier, Y., and Bensimon, A. (2007). Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Molecular biology of the cell*, 18(8):3059–67.

Dalmay, T. and Edwards, D. R. (2006). MicroRNAs and the hallmarks of cancer. *Oncogene*, 25(46):6170–5.

Davis, S. and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)*, 23(14):1846–7.

De, S. (2011). DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nature Biotechnology*, advance on.

de Moura, A. P. S., Retkute, R., Hawkins, M., and Nieduszynski, C. a. (2010). Mathematical modelling of whole chromosome replication. *Nucleic acids research*, 38(17):5623–33.

Dedeurwaerder, S., Desmedt, C., Calonne, E., Singha, S. K., Haibe-Kains, B., Defrance, M., Michiels, S., Volkmar, M., Deplus, R., Luciani, J., Lallemand, F., Larsimont, D., Toussaint, J., Haussy, S., Rothé, F., Rouas, G., Metzger, O., Majjaj, S., Saini, K., Putmans, P., Hames, G., Baren, N. V., Coulie, P. G., Piccart, M., Sotiriou, C., and Fuks, F. (2011). DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO molecular medicine*, pages 1–16.

Delattre, O., Zucman, J., Plougastel, B., Desmaze, C., Melot, T., Peter, M., Kovar, H., Joubert, I., de Jong, P., and Rouleau, G. (1992). Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature*, 359(6391):162–165.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7:3.

Donley, N. and Thayer, M. J. (2013). DNA replication timing, genome stability and cancer: late and/or delayed DNA replication timing is associated with increased genomic instability. *Seminars in cancer biology*, 23(2):80–9.

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics (Oxford, England)*, 24(13):1547–8.

ENCODE Project Consortium and others (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology*, 9(4):e1001046.

Esquela-Kerscher, A. and Slack, F. J. (2006). Oncomirs - microRNAs with a role in cancer. *Nature reviews. Cancer*, 6(4):259–69.

Esteller, M. (2007). Epigenetic gene silencing in cancer: the DNA hypermethylome. *Human molecular genetics*, 16 Spec No:R50–9.

Fang, F., Turcan, S., Rimner, A., Kaufman, A., Giri, D., Morris, L. G. T., Shen, R., Seshan, V., Mo, Q., Heguy, A., Baylin, S. B., Ahuja, N., Viale, A., Massague, J., Norton, L., Vahdat, L. T., Moynahan, M. E., and Chan, T. a. (2011). Breast cancer methylomes establish an epigenomic foundation for metastasis. *Science translational medicine*, 3(75):75ra25.

Feinberg, A. P. (2007). Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447(7143):433–40.

Felsenfeld, G., Boyes, J., Chung, J. A. Y., Clarkt, D., and Studitsky, V. (1996). Chromatin structure and gene expression. *Proceedings of the National Academy of Sciences*, 93(September):9384–9388.

Forrest, A. R. R., Kanamori-Katayama, M., Tomaru, Y., Lassmann, T., Ninomiya, N., Takahashi, Y., de Hoon, M. J. L., Kubosaki, A., Kaiho, A., Suzuki, M., Yasuda, J., Kawai, J., Hayashizaki, Y., Hume, D. A., and Suzuki, H. (2010). Induction of microRNAs, mir-155, mir-222, mir-424 and mir-503, promotes monocytic differentiation through combinatorial regulation. *Leukemia*, 24(2):460–6.

Franceschi, R. T., Iyer, B. S., and Cui, Y. (1994). Effects of ascorbic acid on collagen matrix formation and osteoblast differentiation in murine MC3T3-E1 cells. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*, 9(6):843–54.

Fritz, A., Sinha, S., Marella, N., and Berezney, R. (2013). Alterations in replication timing of cancer-related genes in malignant human breast cancer cells. *Journal of cellular biochemistry*, 114(5):1074–83.

García-Closas, M., Brinton, L. a., Lissowska, J., Chatterjee, N., Peplonska, B., Anderson, W. F., Szeszenia-Dabrowska, N., Bardin-Mikolajczak, a., Zatonski, W., Blair, a., Kalaylioglu, Z., Rymkiewicz, G., Mazepa-Sikora, D., Kordek, R., Lukaszek, S., and Sherman, M. E. (2006). Established breast cancer risk factors by clinically important tumour characteristics. *British journal of cancer*, 95(1):123–9.

Gauthier, M. and Bechhoefer, J. (2009). Control of DNA Replication by Anomalous Reaction-Diffusion Kinetics. *Physical Review Letters*, 102(15):158104.

Goldar, A., Marsolier-Kergoat, M.-C., and Hyrien, O. (2009). Universal temporal profile of replication origin activation in eukaryotes. *PloS One*, 4(6):e5899.

Greenbaum, J. a., Pang, B., and Tullius, T. D. (2007). Construction of a genome-scale structural map at single-nucleotide resolution. *Genome research*, 17(6):947–53.

Gruvberger, S., Ringnér, M., Chen, Y., Panavally, S., Saal, L. H., Borg A, Fernö, M., Peterson, C., and Meltzer, P. S. (2001). Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer research*, 61(16):5979–84.

Guilbaud, G., Rappailles, A., Baker, A., Chen, C.-L., Arneodo, A., Goldar, A., D'Aubenton-Carafa, Y., Thermes, C., Audit, B., and Hyrien, O. (2011). Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. *PLoS computational biology*, 7(12):e1002322.

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. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):139–44.

Hassan, M. Q., Gordon, J. a. R., Beloti, M. M., Croce, C. M., van Wijnen, A. J., Stein, J. L., Stein, G. S., and Lian, J. B. (2010). A network connecting Runx2, SATB2, and the miR-23a˜27a˜24-2 cluster regulates the osteoblast differentiation program. *Proceedings of the National Academy of Sciences of the United States of America*, 107(46):19879–84.

Hassan, M. Q., Maeda, Y., Taipaleenmaki, H., Zhang, W., Jafferji, M., Gordon, J. a. R., Li, Z., Croce, C. M., Van Wijnen, A. J., Stein, J. L., Stein, G. S., and Lian, J. B. (2012). miR-218 Directs a Wnt Signaling Circuit to Promote Differentiation of Osteoblasts and Osteomimicry of Metastatic Cancer Cells. *The Journal of biological chemistry*, 2.

Hassan, M. Q., Tare, R. S., Lee, S. H., Mandeville, M., Morasso, M. I., Javed, A., van Wijnen, A. J., Stein, J. L., Stein, G. S., and Lian, J. B. (2006). BMP2 commitment to the osteogenic lineage involves activation of Runx2 by DLX3 and a homeodomain transcriptional network. *The Journal of biological chemistry*, 281(52):40515–26.

Helman, L. J. and Meltzer, P. (2003). Mechanisms of sarcoma development. *Nature reviews. Cancer*, 3(9):685–94.

Herranz, N., Pasini, D., Díaz, V. M., Francí, C., Gutierrez, A., Dave, N., Escrivà, M., Hernandez-Muñoz, I., Di Croce, L., Helin, K., García de Herreros, A., and Peiró, S. (2008). Polycomb complex 2 is required for E-cadherin repression by the Snail1 transcription factor. *Molecular and cellular biology*, 28(15):4772–81.

Herrick, J., Jun, S., Bechhoefer, J., and Bensimon, A. (2002). Kinetic model of DNA replication in eukaryotic organisms. *Journal of molecular biology*, 320(4):741–750.

Heyn, H. and Esteller, M. (2012). DNA methylation profiling in the clinic: applications and challenges. *Nature reviews. Genetics*, 13(10):679–92.

Hiratani, I., Leskovar, A., and Gilbert, D. M. (2004). Differentiation-induced replication-timing changes are restricted to AT-rich/long interspersed nuclear element (LINE)-rich isochores. *Proceedings of the National Academy of Sciences of the United States of America*, 101(48):16861–6.

Hiratani, I., Ryba, T., Itoh, M., Rathjen, J., Kulik, M., Papp, B., Fussner, E., Bazett-Jones, D. P., Plath, K., Dalton, S., Rathjen, P. D., and Gilbert, D. M. (2010). Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome research*, 20(2):155–69.

Holm, K., Hegardt, C., Staaf, J., Vallon-Christersson, J., Jönsson, G., Olsson, H. k., Borg, A., and Ringnér, M. (2010). Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast cancer research : BCR*, 12(3):R36.

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., Laurance, M. F., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, a. C., Liau, L. M., Wu, H., Geschwind, D. H., Febbo, P. G., Kornblum, H. I., Cloughesy, T. F., Nelson, S. F., and Mischel, P. S. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proceedings of the National Academy of Sciences of the United States of America*, 103(46):17402–7.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., van den Berg, L. H., and Ophoff, R. a. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. *Genome biology*, 13(10):R97.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57.

Hyrien, O. and Goldar, A. (2010). Mathematical modelling of eukaryotic DNA replication. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 18(1):147–61.

Inose, H., Ochi, H., Kimura, A., Fujita, K., Xu, R., Sato, S., Iwasaki, M., Sunamura, S., Takeuchi, Y., Fukumoto, S., Saito, K., Nakamura, T., Siomi, H., Ito, H., Arai, Y., Shinomiya, K.-i., and Takeda, S. (2009). A microRNA regulatory mechanism of osteoblast differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(49):20794–9.

Jiang, J. X., Siller-Jackson, A. J., and Burra, S. (2007). Roles of gap junctions and hemichannels in bone cell functions and in signal transmission of mechanical stress. *Frontiers in bioscience : a journal and virtual library*, 12:1450–62.

Jones, K. B., Salah, Z., Del Mare, S., Galasso, M., Gaudio, E., Nuovo, G. J., Lovat, F., LeBlanc, K., Palatini, J., Randall, R. L., Volinia, S., Stein, G. S., Croce, C. M., Lian, J. B., and Aqeilan, R. I. (2012). miRNA signatures associate with pathogenesis and progression of osteosarcoma. *Cancer research*, 72(7):1865–77.

Jun, S. and Bechhoefer, J. (2005). Nucleation and growth in one dimension. II. Application to DNA replication kinetics. *Physical Review E*, 71(1):011909.

Killian, J. K., Bilke, S., Davis, S., Walker, R. L., Jaeger, E., Killian, M. S., Waterfall, J. J., Bibikova, M., Fan, J.-B., Smith, W. I., and Meltzer, P. S. (2011). A methyl-deviator epigenotype of estrogen receptor-positive breast carcinoma is associated with malignant biology. *The American journal of pathology*, 179(1):55–65.

Kobayashi, T., Lu, J., Cobb, B. S., Rodda, S. J., McMahon, A. P., Schipani, E., Merkenschlager, M., and Kronenberg, H. M. (2008). Dicer-dependent pathways regulate chondrocyte proliferation and differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):1949–54.

Kolmogorov, A. (1937). On the statistical theory of crystallization in metals. *Bulletin of the Academy of Sciences of the USSR. Physical Series*, 1:355–359.

Ku, M., Koche, R. P., Rheinbay, E., Mendenhall, E. M., Endoh, M., Mikkelsen, T. S., Presser, A., Nusbaum, C., Xie, X., Chi, A. S., Adli, M., Kasif, S., Ptaszek, L. M., Cowan, C. A., Lander, E. S., Koseki, H., and Bernstein, B. E. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS genetics*, 4(10):e1000242.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Labib, K. and Hodgson, B. (2007). Replication fork barriers: pausing for a break or stalling for time? *EMBO reports*, 8(4):346–53.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559.

Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is My Network Module Preserved and Reproducible? *PLoS Computational Biology*, 7(1):e1001057.

Lecanda, F., Towler, D. A., Ziambaras, K., Cheng, S. L., Koval, M., Steinberg, T. H., and Civitelli, R. (1998). Gap junctional communication modulates gene expression in osteoblastic cells. *Molecular biology of the cell*, 9(8):2249–58.

Levine, P. (2011). Black Wine. *New Yorker*, 87(25):36.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, Z., Hassan, M. Q., Volinia, S., van Wijnen, A. J., Stein, J. L., Croce, C. M., Lian, J. B., and Stein, G. S. (2008). A microRNA signature for a BMP2-induced osteoblast lineage commitment program. *Proceedings of the National Academy of Sciences of the United States of America*, 105(37):13906–11.

Livak, K. J. and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods (San Diego, Calif.)*, 25(4):402–8.

Logothetis, C. J. and Lin, S.-H. (2005). Osteoblasts in prostate cancer metastasis to bone. *Nature reviews. Cancer*, 5(1):21–8.

Loiselle, A. E., Jiang, J. X., and Donahue, H. J. (2013). Gap junction and hemichannel functions in osteocytes. *Bone*, 54(2):205–12.

Lombaerts, M., van Wezel, T., Philippo, K., Dierssen, J. W. F., Zimmerman, R. M. E., Oosting, J., van Eijk, R., Eilers, P. H., van de Water, B., Cornelisse, C. J., and Cleton-Jansen, A.-M. (2006). E-cadherin transcriptional downregulation by promoter methylation but not mutation is related to epithelial-to-mesenchymal transition in breast cancer cell lines. *British journal of cancer*, 94(5):661–71.

Longhi, A., Errani, C., De Paolis, M., Mercuri, M., and Bacci, G. (2006). Primary bone osteosarcoma in the pediatric age: State of the art. *Cancer Treatment Reviews*, 32(6):423–436.

Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I., and Rubin, E. M. (2002). rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome research*, 12(5):832–9.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R., and Golub, T. R. (2005). MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–8.

Lulla, R. R., Costa, F. F., Bischof, J. M., Chou, P. M., de F Bonaldo, M., Vanin, E. F., and Soares, M. B. (2011). Identification of Differentially Expressed MicroRNAs in Osteosarcoma. *Sarcoma*, 2011:732690.

Lygeros, J., Koutroumpas, K., Dimopoulos, S., Legouras, I., Kouretas, P., Heichinger, C., Nurse, P., and Lygerou, Z. (2008). Stochastic hybrid modeling of DNA replication across a complete genome. *Proceedings of the National Academy of Sciences of the United States of America*, 105(34):12295–300.

Machida, Y. J., Hamlin, J. L., and Dutta, A. (2005). Right place, right time, and only once: replication initiation in metazoans. *Cell*, 123(1):13–24.

Maire, G., Martin, J. W., Yoshimoto, M., Chilton-MacNeill, S., Zielenska, M., and Squire, J. a. (2011). Analysis of miRNA-gene expression-genomic profiles reveals complex mechanisms of microRNA deregulation in osteosarcoma. *Cancer genetics*, 204(3):138–46.

Martens, J. W., Margossian, A. L., Schmitt, M., Foekens, J., and Harbeck, N. (2009). DNA methylation as a biomarker in breast cancer. *Future oncology (London, England)*, 5(8):1245–56.

Martin, J. W., Squire, J. a., and Zielenska, M. (2012). The genetics of osteosarcoma. *Sarcoma*, 2012:627254.

Martin, M. M., Ryan, M., Kim, R., Zakas, A. L., Fu, H., Lin, C. M., Reinhold, W. C., Davis, S. R., Bilke, S., Liu, H., Doroshow, J. H., Reimers, M. a., Valenzuela, M. S., Pommier,

Y., Meltzer, P. S., and Aladjem, M. I. (2011). Genome-wide depletion of replication initiation events in highly transcribed regions. *Genome research*, 21(11):1822–32.

Masai, H., Matsumoto, S., You, Z., Yoshizawa-Sugata, N., and Oda, M. (2010). Eukaryotic chromosome DNA replication: where, when, and how? *Annual review of biochemistry*, 79:89–130.

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., Raney, B. J., Pohl, A., Malladi, V. S., Li, C. H., Lee, B. T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R. A., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B. M., Fujita, P. A., Dreszer, T. R., Diekhans, M., Cline, M. S., Clawson, H., Barber, G. P., Haussler, D., and Kent, W. J. (2013). The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research*, 41(Database issue):D64–9.

Miao, J., Wu, S., Peng, Z., Tania, M., and Zhang, C. (2013). MicroRNAs in osteosarcoma: diagnostic and therapeutic aspects. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, 34(4):2093–8.

Mirabello, L., Troisi, R. J., and Savage, S. a. (2009). Osteosarcoma incidence and survival rates from 1973 to 2004: data from the Surveillance, Epidemiology, and End Results Program. *Cancer*, 115(7):1531–43.

Moindrot, B., Audit, B., Klous, P., Baker, A., Thermes, C., de Laat, W., Bouvet, P., Mongelard, F., and Arneodo, A. (2012). 3D chromatin conformation correlates with replication timing and is conserved in resting cells. *Nucleic acids research*, 40(19):9470–81.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrå le, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–73.

Nabavi, N., Pustylnik, S., and Harrison, R. E. (2012). Rab GTPase mediated procollagen trafficking in ascorbic acid stimulated osteoblasts. *PloS one*, 7(9):e46265.

Nakashima, K., Zhou, X., Kunkel, G., Zhang, Z., Deng, J. M., Behringer, R. R., and de Crombrugghe, B. (2002). The novel zinc finger-containing transcription factor osterix is required for osteoblast differentiation and bone formation. *Cell*, 108(1):17–29.

Nuytten, M., Beke, L., Van Eynde, A., Ceulemans, H., Beullens, M., Van Hummelen, P., Fuks, F., and Bollen, M. (2008). The transcriptional repressor NIPP1 is an essential player in EZH2-mediated gene silencing. *Oncogene*, 27(10):1449–60.

Ohm, J. E., McGarvey, K. M., Yu, X., Cheng, L., Schuebel, K. E., Cope, L., Mohammad, H. P., Chen, W., Daniel, V. C., Yu, W., Berman, D. M., Jenuwein, T., Pruitt, K.,

Sharkis, S. J., Watkins, D. N., Herman, J. G., and Baylin, S. B. (2007). A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature genetics*, 39(2):237–42.

Onder, T. T., Gupta, P. B., Mani, S. A., Yang, J., Lander, E. S., and Weinberg, R. A. (2008). Loss of E-cadherin promotes metastasis via multiple downstream transcriptional pathways. *Cancer research*, 68(10):3645–54.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J., and Wolmark, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine*, 351(27):2817–26.

Peinado, H., Olmeda, D., and Cano, A. (2007). Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? *Nature reviews. Cancer*, 7(6):415–28.

Pinto, D., Benedí, J.-M., and Rosso, P. (2007). *Computational Linguistics and Intelligent Text Processing*, volume 4394 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Plotkin, L. I. and Bellido, T. (2013). Beyond gap junctions: Connexin43 and bone cell signaling. *Bone*, 52(1):157–66.

Pope, B. D., Aparicio, O. M., and Gilbert, D. M. (2013). SnapShot: Replication Timing. *Cell*, 152(6):1390–1390.e1.

PosthumaDeBoer, J., Witlox, M. a., Kaspers, G. J. L., and van Royen, B. J. (2011). Molecular alterations as target for therapy in metastatic osteosarcoma: a review of literature. *Clinical & experimental metastasis*, 28(5):493–503.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 33(Database issue):D501–4.

Pustylnik, S., Fiorino, C., Nabavi, N., Zappitelli, T., da Silva, R., Aubin, J. E., and Harrison, R. E. (2013). EB1 levels are elevated in ascorbic Acid (AA)-stimulated osteoblasts and mediate cell-cell adhesion-induced osteoblast differentiation. *The Journal of biological chemistry*, 288(30):22096–110.

Rakha, E. a., El-Sayed, M. E., Reis-Filho, J. S., and Ellis, I. O. (2008). Expression profiling technology: its contribution to our understanding of breast cancer. *Histopathology*, 52(1):67–81.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics (Oxford, England)*, 23(4):401–7.

Roll, J. D., Rivenbark, A. G., Jones, W. D., and Coleman, W. B. (2008). DNMT3b overexpression contributes to a hypermethylator phenotype in human breast cancer cell lines. *Molecular cancer*, 7:15.

Rosenbloom, K. R., Sloan, C. A., Malladi, V. S., Dreszer, T. R., Learned, K., Kirkup, V. M., Wong, M. C., Maddren, M., Fang, R., Heitner, S. G., Lee, B. T., Barber, G. P., Harte, R. A., Diekhans, M., Long, J. C., Wilder, S. P., Zweig, A. S., Karolchik, D., Kuhn, R. M., Haussler, D., and Kent, W. J. (2013). ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic acids research*, 41(Database issue):D56–63.

Rowley, J. (1973). A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, 243(5404):290–3.

Russo, V. E. A., Martienssen, R. A., and Riggs, A. D. (1996). *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor monograph series. Cold Spring Harbor Laboratory Press.

Ryba, T., Battaglia, D., Chang, B. H., Shirley, J. W., Buckley, Q., Pope, B. D., Devidas, M., Druker, B. J., and Gilbert, D. M. (2012). Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. *Genome research*, 22(10):1833–44.

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(6):761–70.

Ryba, T., Hiratani, I., Sasaki, T., Battaglia, D., Kulik, M., Zhang, J., Dalton, S., and Gilbert, D. M. (2011). Replication Timing: A Fingerprint for Cell Identity and Pluripotency. *PLoS Computational Biology*, 7(10):e1002225.

Schiller, P. C., D'Ippolito, G., Balkan, W., Roos, B. A., and Howard, G. A. (2001). Gap-junctional communication is required for the maturation process of osteoblastic cells in culture. *Bone*, 28(4):362–9.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B. E., Bergman, Y., Simon, I., and Cedar, H. (2007). Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nature genetics*, 39(2):232–6.

Siggelkow, H., Niedhart, C., Kurre, W., Ihbe, a., Schulz, a., Atkinson, M. J., and Hüfner, M. (1998). In vitro differentiation potential of a new human osteosarcoma cell line (HOS 58). *Differentiation; research in biological diversity*, 63(2):81–91.

Simon, J. A. and Lange, C. A. (2008). Roles of the EZH2 histone methyltransferase in cancer epigenetics. *Mutation research*, 647(1-2):21–9.

Smyth, G. K. (2005). *Limma: linear models for microarray data*. Springer, New York.

Sø rlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lønning, P., and Bø rresen Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–74.

Stains, J. P. and Civitelli, R. (2005a). Gap junctions in skeletal development and function. *Biochimica et biophysica acta*, 1719(1-2):69–81.

Stains, J. P. and Civitelli, R. (2005b). Gap junctions regulate extracellular signal-regulated kinase signaling to affect gene transcription. *Molecular biology of the cell*, 16(1):64–72.

Stains, J. P., Lecanda, F., Screen, J., Towler, D. a., and Civitelli, R. (2003). Gap junctional communication modulates gene transcription by altering the recruitment of Sp1 and Sp3 to connexin-response elements in osteoblast promoters. *The Journal of biological chemistry*, 278(27):24377–87.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50.

Sugatani, T. and Hruska, K. A. (2007). MicroRNA-223 is a key factor in osteoclast differentiation. *Journal of cellular biochemistry*, 101(4):996–9.

Tang, M.-H., Varadan, V., Kamalakaran, S., Zhang, M. Q., Dimitrova, N., and Hicks, J. (2012). Major chromosomal breakpoint intervals in breast cancer co-localize with differentially methylated regions. *Frontiers in oncology*, 2(December):197.

Tang, N., Song, W.-X., Luo, J., Haydon, R. C., and He, T.-C. (2008). Osteosarcoma development and stem cell differentiation. *Clinical orthopaedics and related research*, 466(9):2114–30.

Taylor, J. H. (1960). Asynchronous duplication of chromosomes in cultured cells of Chinese hamster. *The Journal of biophysical and biochemical cytology*, 7:455–64.

Thomas, D. and Kansara, M. (2006). Epigenetic modifications in osteogenic differentiation and transformation. *Journal of cellular biochemistry*, 98(4):757–69.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B.,

Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. a., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. a. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Todd, A. K., Johnston, M., and Neidle, S. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic acids research*, 33(9):2901–7.

Tomlins, S. a., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. a., and Chinnaiyan, A. M. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (New York, N.Y.)*, 310(5748):644–8.

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B., and Issa, J. P. (1999). CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 96(15):8681–6.

Valenzuela, M. S., Chen, Y., Davis, S., Yang, F., Walker, R. L., Bilke, S., Lueders, J., Martin, M. M., Aladjem, M. I., Massion, P. P., and Meltzer, P. S. (2011). Preferential localization of human origins of DNA replication at the 5'-ends of expressed genes and at evolutionarily conserved DNA sequences. *PloS one*, 6(5):e17308.

van Beers, E. H. and Nederlof, P. M. (2006). Array-CGH and breast cancer. *Breast cancer research : BCR*, 8(3):210.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England journal of medicine*, 347(25):1999–2009.

Van der Auwera, I., Yu, W., Suo, L., Van Neste, L., van Dam, P., Van Marck, E. a., Pauwels, P., Vermeulen, P. B., Dirix, L. Y., and Van Laere, S. J. (2010). Array-based DNA methylation profiling for breast cancer subtype discrimination. *PloS one*, 5(9):e12616.

Waddington, C. H. (1957). *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. Allen & Unwin.

Wang, X., Dalkic, E., Wu, M., and Chan, C. (2008). Gene module level analysis: identification to networks and dynamics. *Current opinion in biotechnology*, 19(5):482–91.

Wei, Y., Chen, Y.-H., Li, L.-Y., Lang, J., Yeh, S.-P., Shi, B., Yang, C.-C., Yang, J.-Y., Lin, C.-Y., Lai, C.-C., and Hung, M.-C. (2011). CDK1-dependent phosphorylation of EZH2 suppresses methylation of H3K27 and promotes osteogenic differentiation of human mesenchymal stem cells. *Nature cell biology*, 13(1):87–94.

Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D. J., Campan, M., Young, J., Jacobs, I., and Laird, P. W. (2007). Epigenetic stem cell signature in cancer. *Nature genetics*, 39(2):157–8.

Wiemels, J. L., Alexander, F. E., Cazzaniga, G., Biondi, A., Mayer, S. P., and Greaves, M. (2000). Microclustering of TEL-AML1 translocation breakpoints in childhood acute lymphoblastic leukemia. *Genes, chromosomes & cancer*, 29(3):219–28.

Yang, S. C.-H. and Bechhoefer, J. (2008). How Xenopus laevis embryos replicate reliably: investigating the random-completion problem. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 78(4 Pt 1):041917.

Yang, S. C.-H., Rhind, N., and Bechhoefer, J. (2010). Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Molecular systems biology*, 6:404.

Yi, R., Poy, M. N., Stoffel, M., and Fuchs, E. (2008). A skin microRNA promotes differentiation by repressing 'stemness'. *Nature*, 452(7184):225–9.

Zhang, Y., Xie, R.-L., Croce, C. M., Stein, J. L., Lian, J. B., van Wijnen, A. J., and Stein, G. S. (2011). A program of microRNAs controls osteogenic lineage progression by targeting transcription factor Runx2. *Proceedings of the National Academy of Sciences of the United States of America*, 108(24):9863–8.

Zhang, Y., Xie, R.-L., Gordon, J., LeBlanc, K., Stein, J. L., Lian, J. B., van Wijnen, A. J., and Stein, G. S. (2012). Control of mesenchymal lineage progression by microRNAs targeting skeletal gene regulators Trps1 and Runx2. *The Journal of biological chemistry*, 287(26):21926–35.

Zhou, G., Shi, X., Zhang, J., Wu, S., and Zhao, J. (2013). MicroRNAs in osteosarcoma: from biological players to clinical contributors, a review. *The Journal of international medical research*, 41(1):1–12.

Curriculum Vitae